

**AUTHOR ACCEPTED MANUSCRIPT
ITS 2024**

**20th International Conference on Intelligent Tutoring Systems
Generative Intelligence and ITS**

**A Generative Artificial Intelligence empowered chatbot:
System usability and student teachers' experience**

Stavros Nikou¹, Arjun Guliya², Suraj Van Verma³, Maiga Chang⁴

¹ Strathclyde Institute of Education, University of Strathclyde, Glasgow, Scotland, UK
stavros.nikou@strath.ac.uk

² School of Computing and Information Systems, Athabasca University, Edmonton, Canada

³ McGill University, Canada

⁴ School of Computing and Information Systems, Athabasca University, Edmonton, Canada
maiga.chang@gmail.com

Abstract. Generative Artificial Intelligence empowered conversational agents (chatbots) seem to be increasingly used in various settings including education. While student teachers are key stakeholders in supporting and improving education, not many studies exist in student teachers' views on the educational use of chatbots. The current study performs a usability evaluation and explores student teachers' views on the academic use of the VIP-Bot, an advanced academic Discord chatbot, which leverages the OpenAI's gpt-3.5-turbo-instruct model. Student teachers, within the context of the formative task of writing a literature review, interacted with the chatbot and self-reported their experiences through an online survey. The usability evaluation returned a relatively high SUS score (76.36) for the chatbot. Moreover, student teachers view on the chatbot acceptance, effectiveness and motivation were positive. The chatbot can be helpful in developing ideas and initiating further engagement with the literature. Academic misconduct concerns have been expressed if the chatbot is not used properly. The study, as a usability evaluation, is an essential step in further chatbot development and, as an investigation of student teachers' views, it is an essential step on the chatbot employment in teaching and learning.

Keywords: chatbot, Discord bot, Generative Pre-trained Transformer (GPT), gpt-3.5-turbo-instruct, system usability, acceptance, effectiveness, motivation.

1 Introduction

Chatbots, also known as Conversational User Interfaces (CUI), are software applications with the capacity to conduct online conversations with users via text or speech,

acting as virtual assistants [1]. Conversational user interfaces can be based on various underlying technologies such as rule-based systems, Natural Language Processing (NLP), Machine Learning models include Generative Pre-trained Transformer (GPTs) in Generative AI.

Generative AI is a type of artificial intelligence that train models with large amount of data in order to generate new digital content (text, images, video, or audio clips) [2]. One example of Generative AI is the GPT (Generative Pre-trained Transformer) models and the first GPT was introduced by OpenAI [3]. GPT models are based on Large Language Models (LLMs) that autonomously learn from text data and can generate human-like text in response to a human-provided prompt. These models can be embedded in a variety of applications. In particular, they can offer dynamic and accessible forms of online interactions with chatbots providing human-like conversational experiences to users [4].

Generative AI empowered conversational agents seem to be increasingly used in various settings including education [5], [6]. However, related research is still at an early stage [2]. Moreover, with few exceptions (e.g., [7]) not many empirical studies examining the use of chatbots in student teachers' education exist [8]. Student teachers can be the ambassadors of any educational change such as the use of Generative AI in the educational context, transferring their skills and knowledge into their teaching. To the best of our knowledge, not many studies have explored the use of Generative AI empowered chatbots to assist student teachers in their formative essay-type assessments. The current study performs a usability evaluation and explores student teachers' views on the academic use of the VIP-Bot, an advanced academic Discord chatbot, which leverages the OpenAI's gpt-3.5-turbo-instruct model. The current study is aiming to answer the following research questions:

- 1) What is the System Usability score of the Discord academic chatbot, VIP-Bot?
- 2) What is student teachers' experience with using the VIP-Bot in terms of its acceptance, perceived effectiveness, and motivation?

2 VIP-Bot, Generative AI Empowered Discord Chatbot

The VIP-Bot is an advanced academic chatbot on Discord, which leverages the OpenAI API to access gpt-3.5-turbo-instruct model (See Figure 1) to provide assistance ranging from general conversational advice. It makes use of /chat command and public thread to achieve this, giving its users control over their conversation and focus on the chatbot interactions and given responses.

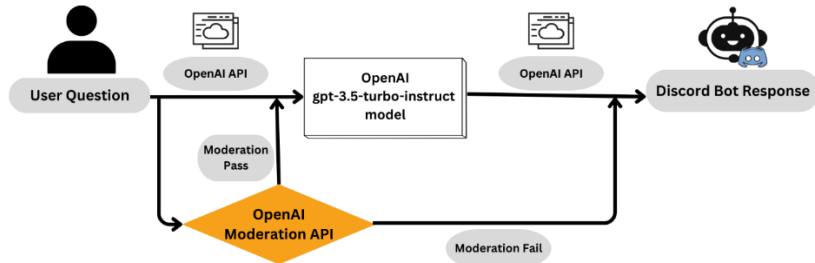


Fig. 1. Flowchart showing the system Workflow of the /chat command.

The /chat command creates a public thread, where each message is used as a prompt to get a response from the OpenAI gpt-3.5-turbo-instruct model, utilizing the message history to create context for the model. Along with the message and context, personality and behaviour instructions are also passed into the model, which are completely customizable using a separate config file. The default personality is set to imitate a typical discord user, complete with regular slang and internet shorthand usage.

The GPT model's response is returned through the API, and captured by the bot, which is then written out directly in the user-created thread. When the context limit or the maximum message count for the model is reached, VIP-Bot automatically closes the thread. Additionally, Since /chat communicates directly with OpenAI's model, each message is first moderated using the OpenAI Moderation API to assess the content of each message. If a message triggers any moderation flags then it is either directly blocked or flagged, and not passed onto the OpenAI API.

When a user initiates interactions with the bot via the primary command: /chat (See Figure 2), VIP-Bot sees the command, it creates a new public thread to allow the user and itself to have a conversation without interfering with the original channel's activity. It also allows multiple users to have simultaneous conversations with the bot and allows the bot to differentiate between each conversation. During the initial thread creation process the user also passes in a 'message' argument which serves as the initiating text with the chatbot.

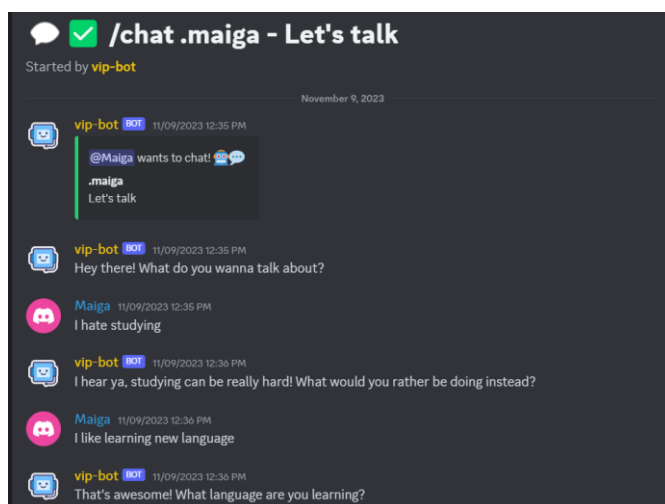


Fig. 2. Screenshot showing an example use case of the /chat command. The chatbot can hold a conversation with the user even when prompted with questions not related to coursework.

3 Methodology

3.1 Participants and Procedures

Data collected based on a convenience sampling. The participants were 10 student teachers, 4 males (40%) and 6 females (60%), enrolled in a post-graduate course in Technology Enhanced learning at the Department of Education of a UK University. Student teachers were five primary school teachers, four secondary school teachers and one higher education lecturer. Participants had not had any previous experience in using chatbots for learning. The study has been granted Ethics approval by the Ethics Committee of the University and has been conducted during the fall semester 2023.

Student teachers have been asked to write a brief literature review type essay on the challenges and benefits of using a specific educational technology of their choice in their own educational context. In alignment with one of the course learning objectives, e.g., students to explore new technologies and adapt as necessary, all students had access to the VIP-Bot on Discord. Therefore, student teachers have been encouraged to interact with the VIP-Bot to help them to prepare for and complete their formative type of assignment.

It has been also communicated to student teachers that, while this could be seen as an opportunity to re-think assessment in ways that could be transformative, under no circumstances the use of AI-generated content be permitted verbatim without clear indication and acknowledgement. Indicatively, among other topics, students have chosen to write about augmented, virtual reality, learning analytics etc. After engaging with

the VIP-Bot, student teachers asked to complete an online survey with closed and open-ended questions self-reporting their experience in using the VIP-Bot.

3.2 Instruments

The survey consisted of two parts. The first part intended to evaluate the system usability and the second part to explore student teachers' acceptance levels, perceived effectiveness, motivation, and concerns. Usability is a pragmatic attribute that refers to the fulfilment of users' functional goals and therefore it is important to be measured [9]. Usability, as defined in ISO 9241-11:2018 (Sect. 3.1.1) [10] as 'the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.'

For the system usability of the VIP-Bot we have used the System Usability Scale (SUS) [11] because of its validity and reliability [12] and its wide acceptance and easy administration. The questionnaire consists of 10 items that are answered using a 5-point Likert scale ranging from "strongly disagree" to "strongly agree", resulting in a single score between 0 and 100 (in 2.5 points increments) where higher scores indicate better usability.

To explore chatbot's acceptance, effectiveness, and student teachers' motivation we have used a 10-item questionnaire. For the acceptance we adopted 3 items from [13], for perceived effectiveness we used a 4-items self-developed scale and for motivation we adopted 3 items from the intrinsic motivation inventory [14]. Sample items for acceptance are "I intend to use the Discord chatbot in the future" and "I find the Discord bot chatbot easy to use", for perceived effectiveness "The Discord bot chatbot provides me with helpful responses relevant to my queries", "The Discord bot chatbot engaged me in meaningful conversations" or "The Discord bot chatbot can have an impact on student learning outcomes and educational practices" and for motivation "I enjoy using the Discord bot chatbot" and "I would describe interacting with Discord bot chatbot as very interesting".

Cronbach's alpha tests were applied to examine the reliability of the instrument with the results to show acceptable (> 0.70) levels of internal consistency. Cronbach's value for acceptance was 0.95, for perceived effectiveness was 0.77 and for motivation 0.81. For an external validation of the scale properties, we compared the findings with measurements from similar studies [15], [16]. To further capture student teachers' experience, we have used open-ended questions focusing on the acceptance and the effectiveness of the VIP-Bot and participants' motivation. Moreover, student teachers' concerns on the use of the chatbot have been explored.

4 Data Analysis and Results

4.1 System Usability

To answer the first research question regarding system's usability, we have used the overall System Usability Score (SUS) questionnaire. A total of ten responses were

collected. The number of participants is within the usability study's general rule 10 ± 2 for optimal sample size. The overall SUS of the VIP-Bot, representing the composite measure of the overall usability of the system [11], was found 76.36. Based on [17],[18] the result found is above average and it is considered as **Good**. This indicates that the VIP-Bot and its functionality is good for using it. SUS is a unidimensional instrument with its questionnaire items better not considered individually [11], however we have reported the responses on the individual questionnaire items along with the median, the mean and standard deviation in order to highlight each one questionnaire item (Table 1).

4.2 Student teachers' experience

To answer the second research question regarding student teachers' experience in using the VIP-Bot, we have used an online survey with a) quantitative data collected from a survey on chatbot's acceptance, effectiveness and participants' motivation and b) qualitative data from open-ended questions on the above constructs.

For the quantitative data, participants responded online to Likert-type questions related to their acceptance level toward the chatbot, perceived effectiveness and their motivation. Table 2 presents the descriptive statistics for student teachers' experience in terms of their acceptance and their perceived effectiveness of the chatbot and their motivation in using it. Students self-reported a high-level of acceptance for the VIP-Bot (3.81, SD = 1.16) in the 5-point Likert scale. Student teachers also reported that they perceived the effectiveness of the chatbot as high (3.19, SD = 0.81). Moreover, student teachers self-reported their level of motivation while interacting with the chatbot as high (3.85, SD = 0.97).

Table 1. SUS questionnaire and statistics for each item.

	Strongly Disagree					Strongly Agree		Median	Mean	SD
	1	2	3	4	5					
1. I think that I would like to use this system frequently.	0	0	3	6	2	4	3.91	0.67		
2. I found the system unnecessarily complex.	1	10	0	0	0	2	1.91	0.29		
3. I thought the system was easy to use.	0	0	1	7	3	4	4.18	0.57		
4. I think that I would need the support of a	0	6	5	0	0	2	2.45	0.49		

technical person to be able to use this system.	0	0	5	6	0	4	3.55	0.49
5. I found the various functions in this system were well integrated.								
6. I thought there was too much inconsistency in this system.	7	4	0	0	0	1	1.36	0.48
7. I would imagine that most people would learn to use this system very quickly.	0	0	2	8	1	4	3.91	0.51
8. I found the system very cumbersome to use.	4	6	1	0	0	2	1.73	0.61
9. I felt very confident using the system.	0	0	2	6	3	4	4.09	0.66
10. I needed to learn a lot of things before I could get going with this system.	4	7	0	0	0	2	1.64	0.48

Table 2. Student teachers' acceptance, perceived effectiveness, & motivation toward VIP-Bot.

	N	Minimum	Maximum	Mean	Std. Deviation
Acceptance	10	1.33	5.00	3.81	1.16
Effectiveness	10	2.00	4.25	3.19	0.81
Motivation	10	1.67	5.00	3.85	0.97

A visual representation of the above sub-scales is depicted in the following Figure 3 with box plots with the overall patterns of student teachers' responses. Acceptance and motivation have high medians (around 4) while the median for the perceived effectiveness of the chat is lower (approximately 3). Comparing the interquartile ranges, we identified a rather similar dispersion.

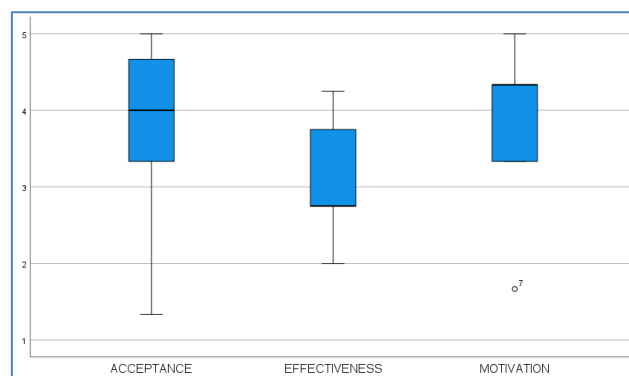


Fig. 3. Visual box-plot representation of the descriptive findings.

For the qualitative data, participants responded online to open-ended questions providing a few interesting insights. Qualitative data were analysed using thematic analysis [19] following a deductive approach using a pre-determined coding scheme based on the constructs of acceptance, effectiveness and motivation, rather allowing the themes to be determined by the data [20].

In terms of the chatbot acceptance, student teachers found the chatbot to be useful and easy to use and they “would like to use it in the future”. Participants expressed their willingness to embrace and integrate the chatbot as a valuable resource for learning and skill development, e.g., “I will use it to clarify terms and concepts as I am engaging with the literature” and “I would encourage my students to use it during their preparation of class” or “to improve their writing skills” since it “can provide useful language structures”. Moreover, “it is good tool for reflecting critically on its output”. The above statements imply student teachers' acceptance of the VIP-Bot and their willingness to incorporate it into the educational process.

In terms of the chatbot's effectiveness, it is prominent in terms of the evaluation of response quality, consideration of prompting techniques, and emphasis on clarity in questioning to achieve optimal outcomes. A participants mentioned “I received clear and conscience answers” however, another said, “some of the responses were not comprehensive”, while most participants reported that “clear and specific prompting obviously results in better results”. The influence of clear and specific prompting on the quality of responses has been acknowledged as “using straightforward words such as (simplify) to indicate the purpose of the prompt can have better results to my questions” and “following up questions to clarify and breaking down questions into smaller sub questions” is important.

In terms of motivation, student teachers enjoyed the interaction with the VIP-Bot (e.g., “I enjoyed using the chatbot a lot”), and they would like to use it because it can be helpful in supporting them in their study. A participant said that the VIP-Bot can help to “discover the key thinkers/writers and their works in relation to a field of study” implying a motivation to deepen one's understanding and knowledge base. Additionally, participants mentioned that the VIP-Bot “is good additional for initial research, ideas, or guidance”, by providing “initial explanations of concepts, before exploring these in greater depth in the literature” and “it is good additional for initial research, ideas, or guidance”. The above statements imply an underlying motivation to further explore and engage with scholarly literature to inform and enhance one's own work.

However, student teachers reported a few concerns on the use of the VIP-Bot. Concerns have been expressed about students simply copying the answers without verifying their accuracy or engaging in critical thinking, e.g., “taking the answers as 'fact' and using these as the basis for assignments, which could be limiting and sometimes incorrect.” Also, few concerns that students might view the VIP-Bot as a complete alternative to doing their own work, potentially undermining academic integrity, e.g., “my only concern would be that students use it as a complete alternative to do all the work for them.”

5 Discussion and Conclusion

Interest in Generative AI empowered chatbots development is growing. Moreover, its employment in education is promising [21], [2]. The current study is a usability evaluation and an investigation of student teachers' views on the academic use of the VIP-Bot, an advanced academic Discord chatbot, which leverages the OpenAI's gpt-3.5-turbo-instruct model. The study examines usability, acceptance, effectiveness, and motivation as main indicators of how well users can learn and use chatbots and how satisfied and motivated users are during the interaction. Similar metrics have been identified by a recent systematic overview of various chatbots usability studies [15]. Student teachers used the VIP-Bot to assist them in completing their formative assessment and self-reported their views on the potential use of the VIP-Bot in education for similar purposes.

Study findings found for the SUS of the VIP-Bot to be 76.36 which is considered as Good (B+) based on [11]. Considering the structure of the SUS questionnaire of having two factors i.e., usability (items #1, #2, #3, #5, #6, #7, #8 and #9) and learnability (items #4 and #10) [22], student teachers' responses indicated that the system was easy to learn (the median values for items 4 and 10 were quite low) and highly usable.

Since student teachers found the chatbot useful and easy to use they intend to use it in the future. The effectiveness of the chatbot attributed mainly to its ability to provide responses relevant to user queries offering opportunities for meaningful conversations. Students self-reported also that they enjoyed using the chatbot which is in agreement with similar studies exploring students' intrinsic motivation while interacting with chatbots [21], [23]. Student teachers also acknowledged the educational value of the VIP-Bot if used appropriately. They agreed that generative AI supported chatbots pose both opportunities and challenges to education.

While they can be helpful in supporting essay writing and assignments completions, several challenges have been identified as well. "Students use it as a complete alternative to do all the work for them" has been identified as the main concern of the participants. However, student teachers agreed that if "a discussion would have taken place either prior to use, or after use to reflect/discuss critically on the use of VIP-Bot, how it works, what it can do, it's limitations" would be useful. Participants seem to agree that it is not always the case that using Generative AI in educational settings is considered academic misconduct if this resource is used critically. The chatbot can be a "positive tool to generate and develop ideas and for developing critical thinking" as long as "both students and teachers depend in original sources".

Research provides evidence that Generative AI empowered chatbots can introduce new ways of teaching and learning transforming such education [2], [24]. Our study is significant because it provides an evaluation of a generative AI empowered chatbot with potential in education and in particular to assist students with their formative assessments. As a usability evaluation, the study is an essential step in further chatbot development. As an investigation of student teachers' views, the study is an essential step on the chatbot employment in teaching and learning. Our study has limitations. One limitation is the small sample size and especially for eliciting quantitative data. Future research will use larger cohorts and moreover it will develop a more structured

instructional design based on the use of the VIP-bot, the Generative AI empowered Discord chatbot.

References

1. Luo, X., Tong, S., Fang, Z., Qu, Z.: *Frontiers: Machines Vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases*. *Marketing Science*, 38(6), 913-1084. (2019) <https://doi.org/10.1287/mksc.2019.1192>
2. Chiu, T.K.F.: *Future research recommendations for transforming higher education with generative AI*, *Computers and Education: Artificial Intelligence*, 6 (2024) DOI: 10.1016/j.caeai.2023.100197
3. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. : *Improving Language Understanding by Generative Pre-Training* (2018) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
4. Skjuve, M., Følstad, A., & Brandtzaeg, P.B.: *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 1–10. (2023) DOI: 10.1145/3571884.3597144
5. Adeshola I., Adepoju, A.P.: *The opportunities and challenges of ChatGPT in education*, *Interactive Learning Environments*, (2023) DOI: 10.1080/10494820.2023.225385
6. Mariani, M.M., Hashemi, N., Wirtz, J.: *Artificial intelligence empowered conversational agents: A systematic literature review and research agenda*, *Journal of Business Research*, 161, 113838 (2023) DOI: 10.1016/j.jbusres.2023.113838
7. Nikou, S.A., Chang, M.: *Learning by Building Chatbot: A System Usability Study and Teachers' Views About the Educational Uses of Chatbots*. In: Frasson, C., Mylonas, P., Troussas, C. (eds) *Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science*, vol 13891. Springer, Cham (2023) https://doi.org/10.1007/978-3-031-32883-1_31
8. Hwang, G. J., Chang, C. Y.: *A review of opportunities and challenges of chatbots in education*. *Interactive Learning Environments*, 31(7), 4099-4112. (2023) DOI: 10.1080/10494820.2021.1952615
9. Hassenzahl, M.: *User experience and experience design*. *The Encyclopedia of Human-Computer Interaction*. Interaction Design Foundation (2013) Access: <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/user-experience-and-experience-design>
10. ISO.2018. 9241-11:2018.: *Ergonomics of Human-System Interaction–Part11: Usability: Definitions and concepts*. International Standardization Organization (ISO) (2018)
11. Brooke, J.: In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) *SUS: A 'quick and dirty' usability scale in Usability evaluation in industry*, pp. 189-194. Taylor and Francis, London/UK (1996)
12. Bangor, A., Kortum, P., Miller, J.: *An empirical evaluation of the system usability scale*. *Int. J. Human-Computer Interaction* 24(6), 574–594 (2008). <https://doi.org/10.1080/10447310802205776>
13. Davis, F.D.: *Perceived usefulness, perceived ease of use, and user acceptance of information technology*. *MIS Quarter*, 13(3), 319-340 (1989). Access: <http://www.jstor.org/stable/249008>

14. McAuley, E., Duncan, T., Tammen, V. V.: Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60, 48-58 (1987)
15. Ren, R., Zapata, M., Castro, J. W., Dieste, O., Acuña, S. T.: Experimentation for Chatbot Usability Evaluation: A Secondary Study, in *IEEE Access*, 10, 12430-12464 (2022) DOI: 10.1109/ACCESS.2022.3145323
16. Casas, J., Tricot, M.-O., Khaled, O.A., Mugellini, E., Cudré-Mauroux, P.: Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, pp. 280-286 (2021). DOI: 10.1145/3395035.3425319
17. Sauro, J.: 5 Ways to Interpret a SUS Score. *Measuring U* (2018). Access: <https://measuringu.com/interpret-sus-score/>
18. Sauro, J., Lewis, J.: *Quantifying the user experience: Practical statistics for user research*". Amsterdam; Waltham, MA: Elsevier/Morgan Kaufmann (2016)
19. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101 (2006) <https://doi.org/10.1191/1478088706qp063oa>
20. Saldana, J. (2015). *The Coding Manual for Qualitative Researchers*, Third Edition, Sage Publications Ltd
21. Lai, C.Y., Cheung, K.Y., Chan, C.S.: Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model. *Computers and Education: Artificial Intelligence*, 5, 100178, (2023) DOI: 10.1016/j.caeai.2023.100178.
22. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International*, pp. 94-103 (2009). DOI: 10.1007/978-3-642-02806-9_12
23. Chiu, T.K., Moorhouse, B.L., Chai, C.S., Ismailov, M.: Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments* (2023) DOI: 10.1080/10494820.2023.2172044
24. Wollny S., Schneider J., Di Mitri D., Weidlich J, Rittberger M., Drachsler H.: Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*. 4:654924 (2021) DOI: 10.3389/frai.2021.654924