Analyzing Student Use of Spacing and Interleaving Strategies in Interactions with GenAI-Powered Chatbots in Programming Courses

Rodrigo Prestes Machado[1][a], Carlos Alario-Hoyos[2][b], Patricia Callejo[2][c],
Iria Estévez-Ayres[2][d] and Carlos Delgado Kloos[2][e]
[1] *Department of Informatics, Instituto Federal de Educação, Ciência e Tecnologia, Porto Alegre, Brazil*
[2] *Department of Telematics Engineering, Universidad Carlos III de Madrid, Leganés (Madrid), Spain*
rodrigo.prestes@poa.ifrs.edu.br, {calario, pcallejo, ayres, cdk}@it.uc3m.es

Keywords: Programming, Generative Artificial Intelligence, Chatbots, Spacing, Interleaving and Metacognition.

Abstract: This study aimed to analyze prompts of programming students with a chatbot powered by OpenAI's GPT-3.5, enhanced with the Retrieval Augmented Generation (RAG) technique, within the context of a Java programming course. The focus was on students using two metacognitive strategies: interleaving and spacing. Student prompts were categorized into eight categories along with their respective study topics. Findings revealed that the categories and markers of spacing and interleaving were important in identifying study sessions with the chatbot. However, students showed limited intentional application of these learning strategies. These results highlight the need for more comprehensive guidance on leveraging AI tools to improve learning outcomes.

1 INTRODUCTION

Recent advances in Generative Artificial Intelligence (GenAI) have created new opportunities in both professional programming and education (Puryear and Sprint, 2022). In programming courses, students can employ GenAI tools to improve their understanding, receive personalized feedback, and access detailed explanations. For example, GitHub Copilot, a tool that is integrated into development environments, assists in providing real-time suggestions and accelerating code writing, which could help streamline the learning process. Educational chatbots, on the other hand, try to guide students through coding challenges, answer questions, and try to foster a deeper understanding of programming concepts.

The study conducted by Chan and Hu (2023) revealed that both undergraduate and postgraduate students have positive attitudes towards the use of GenAI. A systematic review organized by Lo et al. (2024) demonstrated that students can effectively learn from ChatGPT, resulting in improved comprehension and academic achievement (Callejo et al., 2024). Furthermore, it was observed that ChatGPT may allow students to regulate their learning pace and can potentially support self-regulated learning, particularly for students with prior technical and disciplinary knowledge (Xia et al., 2023).

However, researchers have also expressed concerns about the impact of these tools on students. The systematic review by Vargas-Murillo et al. (2023) indicated that the use of ChatGPT could lead to overreliance on the tool. Chan and Hu (2023) noted that overdependence on ChatGPT could lead to a decrease in critical thinking, as students can make decisions based solely on the information provided by the tool. As a potential consequence, Bastani et al. (2024) observed in a study that when programming students lost access to ChatGPT, those who had previously relied on it saw their performance drop by 17%. In contrast, students who had never used the tool were unaffected and outperformed their peers.

The confidence of students in these tools is well founded, as shown by Puryear and Sprint (2022), who found that GitHub Copilot can generate solutions for student assignments with precision rates ranging from 68% to 95%. However, Sun et al. (2024) observed that using ChatGPT without a structured learning approach did not provide significant improvement over traditional self-directed learning methods in programming. This raises concerns that students may become

[a] https://orcid.org/0000-0003-0428-6387
[b] https://orcid.org/0000-0002-3082-0814
[c] https://orcid.org/0000-0001-6124-6213
[d] https://orcid.org/0000-0002-1047-5398
[e] https://orcid.org/0000-0003-4093-3705

Machado, R. P., Alario-Hoyos, C., Callejo, P., Estévez-Ayres, I. and Kloos, C. D.
Analyzing Student Use of Spacing and Interleaving Strategies in Interactions with GenAI-Powered Chatbots in Programming Courses.
DOI: 10.5220/0013272000003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 273-280
ISBN: 978-989-758-746-7; ISSN: 2184-5026

overly dependent on generative AI tools, potentially hindering their learning progress and missing opportunities to develop a deeper understanding of fundamental concepts.

Regardless of teachers' preferences or beliefs, preliminary surveys conducted by Dickey et al. (2024) indicate that more than 54.5% of students are already using GenAI for homework, likely due to their perception of this technology as advantageous and intuitive (Sun et al., 2024). All of these studies highlight the need to increase the understanding of how students interact with these tools and how they can be used to improve learning, as emphasized by Lo et al. (2024).

In response to the growing need for deeper insight into the use of GenAI tools in educational settings, this study aims to help address this gap by analyzing the interactions between students in a Java programming course and an educational chatbot powered by OpenAI gpt-3.5, enhanced with the Retrieval-Augmented Generation (RAG) technique. Specifically, it focuses on examining these interactions within the context of two metacognitive strategies: spacing (Carvalho et al., 2020) and interleaving (Firth et al., 2021). To achieve this objective, we formulate three research questions:

- RQ1 - What distribution patterns emerged in the classified student interaction prompts with the educational chatbot?
- RQ2 - Was there spacing between student prompts? Which category led to the most rapid and consistent interactions with the chatbot?
- RQ3 - Do students' interactions with the chatbot alternate between study topics to suggest interleaving?

This paper is organized as follows. Section 2 presents the theoretical framework, Section 3 describes the material and methods used in the study, Section 4 presents the results and discussion, and Section 5 concludes the paper and outlines future work.

2 THEORETICAL FRAMEWORK

This section outlines the theoretical framework that is the basis of the study. We begin by introducing the concept of metacognition, along with the spacing and interweaving strategies applied in this research. Next, we present related works organized into two subsections: research involving GenAI and studies focusing on metacognition.

2.1 Metacognition

Given that GenAI often produces highly accurate and automatic responses (Puryear and Sprint, 2022), it is essential that students use these tools within an active learning process to enhance their learning outcomes. This active learning process can be further understood through the lens of metacognition, which focuses on awareness of one's mental processes. Flavell (1979) proposed a model of metacognitive monitoring that includes four interrelated phenomena: Metacognitive Knowledge, Metacognitive Experience, Metacognitive Goals, and Metacognitive Actions. These processes do not occur in isolation, but they influence each other, altering cognitive progress over time.

Metacognitive knowledge includes beliefs about variables that affect the outcomes of cognitive activities. It is divided into three types: beliefs about personal abilities, perceived difficulty in the task, and previously used strategies. Metacognitive Experience refers to the feelings that arise before, during, and after cognitive activity, such as frustration, confusion, satisfaction, and others. Metacognitive Goals are the key to regulating thought as they relate to the goals the individual seeks to achieve, directly influencing the actions taken. For example, if a student's goal is to complete a task quickly, they may adopt a more passive approach to learning. Lastly, Metacognitive Actions involve the planning, monitoring, and evaluation of strategies used to achieve the goals. In terms of planning, students can determine how to approach a task, such as spacing their study sessions (Carvalho et al., 2020), interleaving topics (Firth et al., 2021), utilizing retrieval practice (Larsen, 2018), and other possible strategies.

Spacing and interleaving are two metacognitive strategies that have been shown to enhance learning outcomes and support the research questions of this study. Spacing refers to the practice of distributing study sessions over time, which has been shown to improve long-term retention and understanding of the material (Carvalho et al., 2020). Interleaving involves mixing different topics or problems within a single learning session, which has been shown to also enhance long-term learning and the application of student's knowledge in other contexts and situations (Firth et al., 2021) .

2.2 Related Work

Marguliex et al. (2024) conducted a study on how undergraduate students in introductory programming courses used generative AI to solve programming problems in a naturalistic setting. The research focused on examining the relationship between AI usage and students' self-regulation strategies, self-efficacy, and fear of failure in programming. Furthermore, the study explored how these variables interacted with the characteristics of the learners, the perceived usefulness of AI, and academic performance. The findings revealed that students with higher self-efficacy, lower fear of failure, or higher prior grades tended to use AI less frequently or later in the problem-solving process and perceived it as less useful compared to their peers. However, no significant relationship was found between students' self-regulation strategies and their use of AI.

The study of Sun et al. (2024) examined the effects of ChatGPT-assisted programming on university students' behaviors, performance, and perceptions. A quasi-experimental research was conducted with 82 students divided into two groups: one with ChatGPT-assisted programming (CAP) and the other with self-directed programming (SDP). The analysis included behavioral logs, performance evaluations, and interviews. Students in the CAP group engaged more actively in debugging and feedback review activities. Although they achieved slightly higher scores, there was no statistically significant difference in performance compared to the SDP group. Nevertheless, perceptions of ChatGPT improved significantly, highlighting greater perceived usefulness, ease of use, and intention to use the tool in the future.

Bastani et al. (2024) investigated the impact of GenAI, specifically GPT-4, on human learning, with a focus on mathematics education in a high school. The problem addressed was how the use of generative AI could affect the acquisition of new skills, which is crucial for long-term productivity. The method involved a controlled experiment with approximately one thousand students, who were exposed to two GPT-4-based tutors: a simple tutor (GPT Base) and another with safeguards designed to promote learning (GPT Tutor). The results showed that while access to GPT-4 improved performance on practice exercises (48% with GPT Base and 127% with GPT Tutor), the removal of access to GPT Base led to a 17% decrease in student performance on exams. This suggests that unrestricted use of *GPT Base* could hinder learning. However, the GPT Tutor was able to mitigate this negative effect.

3 MATERIAL AND METHOD

This section outlines the tools and procedures used in this study. CharlieBot served as the GenAI tool for educational purposes, and the method was structured into three distinct phases, each designed to systematically assess its interaction with students.

3.1 CharlieBot

CharlieBot is an educational chatbot powered by ChatGPT 3.5 and enhanced with Retrieval-Augmented Generation (RAG) (Chen et al., 2024). RAG is an AI technique that integrates information retrieval with generative models. It first retrieves relevant documents or data from a knowledge base or external source using a retrieval model. Then, a generative model uses this retrieved information to generate more accurate, context-aware responses. This dual-step process enhances the quality and reliability of the chatbot's answers. A prior study on CharlieBot's performance revealed that most students found its responses appropriate for the Java programming course (Alario-Hoyos et al., 2024).

3.2 Method

The study comprised three phases: data collection, categorization, and analysis. During the data collection phase, students enrolled in a second-semester Java course at the University Carlos III of Madrid (UC3M) were introduced to CharlieBot and allowed to use it without following any prescribed educational methodology. All data were collected anonymously to ensure that interactions could not be traced back to individual students or linked to their academic performance.

During the categorization phase, the students' prompts were initially classified into eight distinct categories using the Claude.ai (Anthropic, 2023) tool. Subsequently, the authors of this study manually reviewed these classifications to ensure accuracy and alignment with the research objectives.

Initially, Ghimire and Edwards (2024) proposed four categories: Debugging Help (DH), Conceptual Question (CQ), Code Snippet (CS), and Complete Solution (CSO). However, the data collected indicated the need for additional categories, leading to the inclusion of four more: Multiple Questions (MQ), Student Corrections (SC), Language Change (LC), and Uncategorized (U). Unlike study Ghimire and Edwards (2024), which analyzed students' behavior using artificial intelligence for specific programming tasks, our study allowed participants to explore the chatbot as they preferred. This approach enabled the emergence of additional categories of analysis, broadening the understanding of the tool's uses and interactions. Table 1 presents these categories, their descriptions, and an example. Besides that, the topics cov-

Table 1: Categories - adapted from Ghimire and Edwards (2024).

Category	**Description**	**Example of prompt**
Debugging Help (DH)	Prompts that seek help to identify, fix errors, or understand the provided code snippet.	*Would this code be ok? {code}*
Conceptual Question (CQ)	Prompts that are more about understanding concepts than specific code.	*What does it mean for a method to be static?*
Student Correction (SC)	Prompts where the student corrects the chatbot.	*The correct answer is B*
Code Snippet (CS)	Prompts that ask for a specific part of the code, like a function or a segment.	*A class inherits from another write this code*
Complete Solution (CSO)	Prompts that request an entire solution or a complete code snippet.	*Give me the code for a selection sort*
Multiple Question (MQ)	Prompts where the user wants to solve a multiple choice exercise (Quiz).	*A heap is a data structure appropriate for: {options}*
Language Change (LC)	Prompts that request a change of idiom.	*In Spanish*
Uncategorized (U)	Prompts that do not fit into any of the above categories.	*Thanks*

ered in the students' prompts were categorized based on the course issues, which include (1) Java, (2) Object Orientation, (3) Testing, (4) Recursion, (5) Data Structures, and (6) Sorting and Searching Algorithms.

The analysis phase involved using Python/Pandas scripts to extract information from previously classified data.

4 RESULTS AND DISCUSSION

A total of 625 student messages were categorized in 81 conversations, with an average of 7.72 messages per conversation. This data offers insights into how users engage with CharlieBot and the types of prompts they submit. The results of the analysis are presented in this section, addressing the research questions outlined in the introduction. Figure 1 shows the distribution of the messages in the categories. The following sections present the results and discussions of each research question.

4.1 Categorization of Messages

About the first research question: *RQ1 - What distribution patterns emerged in the classified student interaction prompts with the educational chatbot?*

According to Figure 1 messages classified as Conceptual Question account for 35.8% of the total. Sequences of Conceptual Questions often start from a practical perspective of a given code, as illustrated in conversation 4 of Figure 2, which presents examples of students' interactions with the chatbot. In other cases, the conversation consists entirely of messages classified as Conceptual Question, such as in conversation 15 of Figure 2.

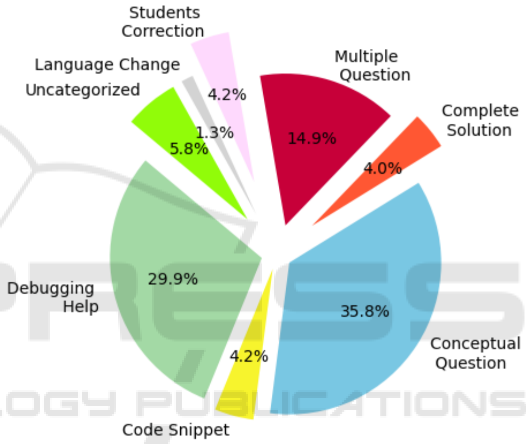

Figure 1: Categorization of messages.

As shown in Figure 1, approximately 29.9% of the messages were classified as Debugging Help. Debugging help messages are likely to provide students with a practical approach to understanding code, identifying errors, and building skills to enhance their debugging abilities for future tasks.

Most of the messages are located in the categories Conceptual Question and Debugging Help corroborate the findings of Ghimire and Edwards (2024) which showed that the questions are also localized in the same categories.

Multiple Question exercise resolutions represented 14.9% of the responses. These prompts typically involved students submitting one or more exercises to CharlieBot and requesting solutions. The conversation 13 in Figure 2 illustrates a sequence of multiple choice questions.

Requests for the chatbot to generate a Code Snippet or Complete Solution accounted for 8.2% of student messages. These prompts reflected a desire for

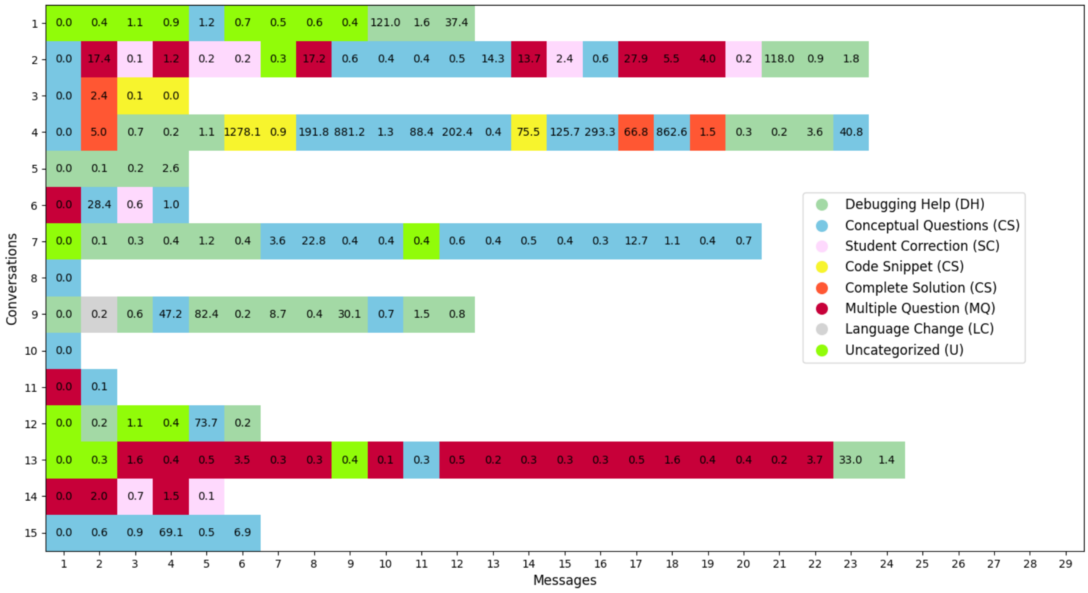

Figure 2: Examples of conversations.

more direct answers; however, it was observed that, after receiving these solutions, many students transitioned to a more active approach in seeking to understand the code or the underlying concepts. Thus, it can be inferred that the prompts within the Code Snippet or Complete Solution category were often used as a starting point for a more in-depth study.

Student corrections to chatbot responses represented 4.2% of the interactions. In most cases, these corrections occurred when the student already had the correct answer to an exercise and identified an error in the chatbot response, as illustrated in conversation 2 in Figure 2. These corrections highlight that, although the chatbot provides quick feedback, it is not always accurate.

Figure 1 also indicates that approximately 5.8% of the prompts written by the students were classified as Uncategorized. These messages include expressions of gratitude toward the chatbot, such as a simple '*thanks*', contextual statements like '*I'm reviewing object orientation*', as well as requests for additional exercises and summaries. These Uncategorized messages can be subdivided into new categories depending on the investigation. For instance, requests for new exercises or summaries could be interpreted as a metacognitive monitoring strategy.

Finally, 1.3% of the messages were classified as Language Change. These prompts were typically requests for the chatbot to switch languages (English to Spanish), as seen in Figure 2, conversation 9, message 2.

4.2 Spacing

About the first part of the second research question: *RQ2 - Was there spacing between student prompts?*

For the spacing analysis, we initially considered a minimum interval of 60 minutes as a reference to define the significant spacing (Uguina-Gadella et al., 2024). Based on this, our results showed that, of the 81 conversations analyzed, just 27, or around 34%, had at least one message with an interval more significant than 60 minutes. Figure 2 illustrates the elapsed time between messages (delta). For example, in conversation 15, the first message was sent at time zero, while the student sent the second message after a delta of 0.6 minutes. Additionally, message 4 in conversation 15, as shown in Figure 2, presents a delta of 69.1 minutes, which, according to our reference, characterizes spacing. The average number of subsections per conversation was 2.2 among conversations that exhibited spacing. Therefore, a conversation with one spacing is divided into two distinct study sessions.

This data suggests that there is indeed little spacing between study sessions. To create a new section in CharlieBot; students must close the browser tab and start a new conversation, which results in losing the previous chat history.

4.2.1 Interaction Time per Category

About the second part of the second research question: *RQ2 - Which category led to the most rapid and consistent interactions with the chatbot?*

The objective of this research question was to study the time a student takes to re-engage with the bot after sending a message for each category. This analysis can offer insights into students' behavior: faster interactions might with the chatbot might suggest a more passive learning posture.

To calculate the average time and standard deviation for messages within a category initially was first observed conversations in which message timing was influenced by previous messages. Pearson's correlation analysis revealed that 28.40% of the conversations exhibited a strong correlation, which, in this study, was defined as values outside the range of -0.3 to 0.3. In this context, a strong correlation suggests that the response time for a message may depend on the time of the preceding message or multiple prior messages. Consequently, these 28.40% of conversations were excluded from the average time and standard deviation calculation, leaving 56 out of the initial 86 conversations. Table 2 presents the means and standard deviations for each category.

When a student requests a Code Snippet (CS), it is generally predictable that they will summarize interaction with the chatbot within a short period, as indicated by the mean (1.34) and standard deviation (1.95) of the Code Snippet category in Table 2. This result is noteworthy because, although some requests are simple and lead to quick responses, others may be more complex and require more reflection time. However, the low average response time and minimal variability may indicate that students adopt a more passive learning posture when requesting the chatbot to generate pieces of code.

Despite occasional variation (3.59), prompts classified as Uncategorized exhibit a relatively low average time (1.70) for students to resume interaction with the chatbot. This variation occurs partly because of situations where students end the conversation with an Uncategorized message like *"thanks"* but later decide to resume the interaction.

When a student requests the chatbot to Change the Language (LC) in a response, in our context English to Spanish, the new interaction occurs quickly (2.23) with slight variation in response time (3.71), meaning that it is predictable that the student will continue interacting.

Although still below the overall average, students take moderate time (2.47) to resume interaction with the chatbot after requesting the solution to a Multiple Question (MQ) exercise. The variation (5.56) suggests that some responses may lead to a slower follow-up interaction, possibly indicating a moment of deeper reflection on the part of the students. However, since the mean and standard deviation are below the overall average, this could indicate a more passive behavior, as students obtain ready-made answers from the chatbot.

The average time of the Student Correction (SC) category is close to the overall average (4.07), although the high standard deviation (8.42) indicates significant variation. In some cases, students know the answer to an exercise and only accomplish a quick correction from the chatbot; in others, they detect flaws in the chatbot's responses and need to examine the subsequent answer to identify potential errors, which suggests a more active behavior.

A message classified as Debugging Help (DH) has both a long average response time (5.30) and high variation (10.80). Consequently, the behavior in this category is more time-consuming and unpredictable. A possible explanation is that some DH messages are simple, such as '*what does x++ do?*'. In contrast, others involve requests for analysis of more complex code, like *'explain this piece of code (accompanied by a complete method).'*

Conceptual Questions (CQ) generally result in a high average time (5.47) for students to reengage with the chatbot. Moreover, the high standard deviation (10.46) indicates an unpredictable variation in response times. This unpredictability may be attributed to the fact that, in some cases, the complexity of the question and answer demands more reflection time from the student, whereas in others, lesser complexity allows for a quicker interaction with the chatbot.

Finally, a message classified as a Complete Solution request (SCO) exhibits the longest average time (7.65) for students to resume interaction with the chatbot. Additionally, the high standard deviation (11.03) highlights significant unpredictability in response time. After requesting a complete solution, students may take longer to return due to the complexity of the problems involved, such as the implementation of sorting algorithms such as Heapsort.

4.3 Interleaving

About the third research question: *RQ3 - Do students' interactions with the chatbot alternate between study topics to suggest interleaving?*

Switching between topics can help students identify distinct concepts inside problems. The results show that approximately 50.8% of the conversations include interleaving; consequently, 49.2% are related to a single study topic. To gain a deeper comprehension of this interleaving, conversations were analyzed by separating those without spacing from those that included spacing.

Table 2: Average and Standard Deviation of Different Categories (in minutes).

Category	**Average**	**Standard Deviation**
Code Snippet	1.34	1.95
Uncategorized	1.70	3.59
Language Change	2.23	3.71
Multiple Question	2.47	5.66
Overall	*3.78*	*6.95*
Student Correction	4.07	8.42
Debugging Help	5.30	10.80
Conceptual Questions	5.47	10.46
Complete Solution	7.65	11.03

Conversations without spacing comprise around 66% of the total, with 34.2% showing a topic change, indicating interleaving (equivalent to 22.8% of all conversations). Qualitative analysis of these conversations without spacing revealed that about 7% of the conversations presented interleaving caused by students' at least one request to solve quizzes. However, the speed with which students interact with the chatbot to solve quizzes may suggest a more superficial reflection on the answers received (see section 4.2.1). On the other hand, 8.8% of the conversations showed no spacing but included topic changes, suggesting that students effectively might employ a metacognitive interleaving strategy to distinguish concepts around a problem. Finally, 7% of the conversations featured interaction pauses lasting between 20 and 59 minutes but fell short of the study's definition of spacing, which required pauses of over 60 minutes. This subset of conversations revealed a natural interleaving of topics, reflecting a chatbot usage pattern.

Conversations with spacing comprise approximately 34% of the total, with 84.2% indicating interleaving (equivalent to 28% of all conversations). Pauses longer than 60 minutes revealed similar behavior to those shorter than 60 minutes, with students using the chatbot primarily for quick consultations, asking questions, and returning at least one hour late with a query about a different study topic. The argument that some students use the chatbot primarily as a tool for quick queries is supported by data showing an average of 4.18 messages per study session. Interestingly, only 6% of the conversations involved students pausing their interaction with the chatbot for more than 60 minutes before sending a message related to the exact topic of study.

Therefore, although approximately 50.8% of the conversations contain interleaving, the data shows that this topic switching is limited in terms of using a conscious metacognitive strategy.

5 CONCLUSION AND FUTURE WORK

This study aimed to analyze the interactions of Java programming students with an educational chatbot developed using the RAG technique, employing two metacognitive study strategies: Spacing and Interleaving.

The analysis categories, spacing, and interleaving markers provided insights into students' interactions with the chatbot. With the concept of spacing, it was possible to define students' study sessions, which, in this work, consisted of a few interaction messages with the chatbot. Consequently, a significant portion of the observed interleaving occurred due to spacing, that is, in the transition between study sessions. These findings indicate that the intentional use of learning strategies, such as spacing and interleaving, is still limited. Therefore, encouraging the deliberate application of these practices can enhance students' learning outcomes using generative artificial intelligence (GenAI) tools.

Messages from the Code Snippets and Multiple Question (Quiz) categories stood out for involving consistently quick interactions with the bot. This pattern suggests some level of engagement with the chatbot but also raises the possibility of less in-depth reflection on the content received, especially when students request small code snippets or the resolution of exercises.

It is fundamental to consider how GenAI tools impact students in various demographic groups, academic disciplines, cultural backgrounds, and levels of previous experience (Catalán et al., 2021) (Neo, 2022). Consequently, this study is limited to a specific group of students and focuses on using a single GenAI tool. As a result, the findings may not be generalizable to other populations or tools.

There are ways to deepen this study. While 60 minutes was the standard spacing measure, future research will explore different intervals using quantitative variables. Additional metacognitive strategies, such as the Monitoring strategy—where students seek clarification on unclear concepts—will also be examined. Moreover, investigating the relationship between student profiles, GenAI metacognitive strategies, and learning outcomes is essential.

ACKNOWLEDGMENT

This work was supported by grants PID2023-146692OB-C31 (GENIELearn project) and PID2020-112584RB-C31 (H2O Learn project) funded by MICIU / AEI / 10.13039 / 501100011033, MCIN / AEI / 10.13039 / 501100011033 and ERDF, EU. Further support has been received from the UNESCO Chair "Scalable Digital Education for All" at UC3M.

REFERENCES

Alario-Hoyos, C., Kemcha, R., Kloos, C. D., Callejo, P., Estévez-Ayres, I., Santín-Cristóbal, D., Cruz-Argudo, F., and López-Sánchez, J. L. (2024). Tailoring your code companion: Leveraging llms and rag to develop a chatbot to support students in a programming course. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 1–8.

Anthropic (2023). Claude: Next-generation ai assistant. Accessed: January 16, 2025.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Özge Kabakcı, and Mariman, R. (2024). Generative ai can harm learning. Technical report, The Wharton School Research Paper.

Callejo, P., Alario-Hoyos, C., and Delgado-Kloos, C. (2024). Evaluating the impact of chatgpt on programming learning outcomes in a big data course. *International Journal of Engineering Education*, 40(4):863–872.

Carvalho, P. F., Sana, F., and Yan, V. X. (2020). Self-regulated spacing in a massive open online course is related to better learning. *npj Science of Learning*, 5(1):2.

Catalán, A. C., González-Castro, N., Delgado, K. C., Alario-Hoyos, C., and Muñoz-Merino, P. J. (2021). Conversational agent for supporting learners on a mooc on programming with java. *Computer Science and Information Systems*, 18(4):1271–1286.

Chan, C. K. Y. and Hu, W. (2023). Students' voices on generative ai: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1):43.

Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Dickey, E., Bejarano, A., and Garg, C. (2024). Ai-lab: A framework for introducing generative artificial intelligence tools in computer programming courses. *SN Comput. Sci.*, 5(6).

Firth, J., Rivers, I., and Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, 9(2):642–684.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Ghimire, A. and Edwards, J. (2024). Coding with ai: How are tools like chatgpt being used by students in foundational programming courses. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I., editors, *Artificial Intelligence in Education*, pages 259–267, Cham. Springer Nature Switzerland.

Larsen, D. P. (2018). Planning education for long-term retention: the cognitive science and implementation of retrieval practice. In *Seminars in neurology*, volume 38, pages 449–456. Thieme Medical Publishers.

Lo, C. K., Hew, K. F., and yung Jong, M. S. (2024). The influence of chatgpt on student engagement: A systematic review and future research agenda. *Computers and Education*, 219:105100.

Margulieux, L. E., Prather, J., Reeves, B. N., Becker, B. A., Cetin Uzun, G., Loksa, D., Leinonen, J., and Denny, P. (2024). Self-regulation, self-efficacy, and fear of failure interactions with how novices use llms to solve programming problems. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ITiCSE 2024, page 276–282, New York, NY, USA. Association for Computing Machinery.

Neo, M. (2022). The merlin project: Malaysian students'acceptance of an ai chatbot in their learning process. *Turkish Online Journal of Distance Education*, 23(3):31–48.

Puryear, B. and Sprint, G. (2022). Github copilot in the classroom: learning to code with ai assistance. *J. Comput. Sci. Coll.*, 38(1):37–47.

Sun, D., Boudouaia, A., Zhu, C., and Li, Y. (2024). Would chatgpt-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? an empirical study. *International Journal of Educational Technology in Higher Education*, 21(1):14.

Uguina-Gadella, L., Estévez-Ayres, I., Fisteus, J. A., Alario-Hoyos, C., and Kloos, C. D. (2024). Analysis and prediction of students' performance in a computer-based course through real-time events. *IEEE Transactions on Learning Technologies*, 17:1794–1804.

Vargas-Murillo, A. R., de la Asuncion Pari-Bedoya, I. N. M., and de Jesús Guevara-Soto, F. (2023). Challenges and opportunities of ai-assisted learning: A systematic literature review on the impact of chatgpt usage in higher education. *International Journal of Learning, Teaching and Educational Research*, 22(7):122–135.

Xia, Q., Chiu, T. K. F., Chai, C. S., and Xie, K. (2023). The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*, 54(4):967–986.