Observing How Students Program with ChatGPT: A Quantitative Eye-Tracking Assesment of Visual Expertise Acquisition

Anonymous ACL submission

Abstract

The proliferation of language models is revolutionizing the process of Human-AI Interaction (HAI), offering users a conversational interface to accomplish various tasks and access information. Understanding how these models affect the way students learn the skill of computer programming remains an unstudied area of research. This paper presents an experiment designed to investigate the interaction dynamics of undergraduate students with varying computer programming abilities when utilizing ChatGPT, as an AI-assisted tool to accomplish coding tasks. Eye-tracking technology is employed to capture participants' gaze patterns and visual attention during their interactions with the language model. The paper presents the analysis of a total of 120 eye tracking cases. Using the Kruskal-Wallis statistical test to assess whether students selectively accord attention to programming tasks based on their perceived importance and complexity, we find that significant differences ($p < .001$) across the 'hit time', 'time to the first fixation' and the 'areas of interest duration' eye tracking features. The results shed light on differences in visual attention patterns, the utilization of AI-generated suggestions, code comprehension strategies, and preferences for interacting with ChatGPT during coding tasks.

1 Introduction

The advent and public availability of mainstream Large Language Models (LLMs), at either very low-cost or no-cost has trivialized their use. Popular LLMs such as ChatGPT are being prompted daily by active users; with an estimated updake of 100 million monthly active users in January of 2023, just two months after its launch, making it the fastest-growing consumer application in history (Hu, 2023). The profileration of LLMs has resulted in uses for both professional and personal miscellaneous tasks. Whether the LLM prompt is mundane or complex, the average person seems to be prioritizing its use it as an alternative to web search (Ibrahim et al., 2023). Due to the rise in LLM usage, an understudied area of research is the impact of LLMs in higher education and its use by students in the learning process (Zumwalt et al., 2014; Maldonado et al., 2023).

1.1 LLMs in higher education

The latest studies suggest that ChatGPT and similar models perform equal to and sometimes than university students in a diverse set of courses (Ibrahim et al., 2023). However, the question that remains is to what extent is the output provided by ChatGPT translated by the student from an output text towards the student's own assimilation and learning. Moreover, LLMs, generally, and ChatGPT especially are now capable of generating functional programming code (Acher et al., 2023). This capability is expected to change the ways and modalities developers, coding enthusiasts and students learn and interact with programming tasks (Yilmaz and Yilmaz, 2023). Thus, the aim of the study is to understand how a student navigates the task of completing a programming exercise given the availability of LLM tools like ChatGPT. We perform this assessment with eye tracking technology.

1.2 Visual expertise acquisition

Eye tracking has been widely used to uncover the process of learning visually (Gegenfurtner and van Merriënboer, 2017; Davies, 2018), across different domains, including medicine (Zammarchi and Conversano, 2021), strategic and algorithmic thinking (Reingold and Sheridan, 2011), and natural language processing (Salicchi et al., 2021; Barrett et al., 2016; Bolotova et al., 2020). This study lays the foundation for examining visual expertise acquisition in persons utilizing AI aids to complete technical coding tasks, a growing domain of research within Human-AI Interaction. (Langner et al., 2023; MacKenzie, 2012).

2 Research Question and Hypothesis

Our aim is to quantify the visual attention behavior of undergraduate students towards ChatGPT when using it to accomplish programming tasks. We use eye tracking data for this quantification. We present the following two research questions: (RQ1) Can we recognize any eye-tracking patterns within the ChatGPT interface when programming students use it to solve a programming exercise or task? (RQ2) In addition to eye tracking data, what are other suitable parameters that need to to be analyzed to have a better understanding of the student's assimilation of a programming task when using ChatGPT as a supporting tool?

Our hypothesis is that the attention, depicted through eye fixations and the duration for which students look at specific areas of the ChatGPT interface directly affect their perceived complexity and time required to understand and solve programming problems.

3 Materials and Methods

3.1 Materials

Stududents were required to solve a set of four programming tasks for the experiment. These tasks were designed to mimic academic programming exercises, where ChatGPT could function as a supportive coding tool. The tasks varied in complexity to capture a wide spectrum of user interactions and challenges. These tasks were analyzed across five areas of interests discussed in section 4. The experiment was approved by the institution's IRB committee. To support extensions and reproducibility of this work, the collected data is made available upon request.

Each of the four programming tasks were composed of:

1. *Looping constructs*: The problem requires the student's understanding of 'for' loops and 'nested for' loops

2. *Data structures manipulations*: The problem requires the student's familiarity with data structures in order to solve a searching and sorting problem.

3. *Creative problem solving using algorithms*: This problem contains a form of ChatGPT hallucination, where the student is required to verify the correctness of the reasoning provided by ChatGPT.

4. *Testing sample input/output*. This problem requires the students to analyze a provided sample input and output for a programming exercise to have a complete understanding of the algorithmic problem.

The experiment employed a screen-based eye tracker, SmartEye AI-X, with a frequency of 60 Hz, with iMotions software version 9.3 to record and capture participants' gaze patterns and visual attention. This allowed for tracking participants' eye movements and fixation points as they interacted with the ChatGPT interface and solved the programming exercises.

3.1.1 Methods

Participants were initially briefed on the study's objectives and were also informed about the type of data that would be collected. Before the experiment, participants received concise instructions on how to interact with the ChatGPT interface and were given time to familiarize themselves with the system and calibrate their eyes with the eye tracker. Participant were then assigned the four coding tasks to complete using ChatGPT. The participants were asked to record their answers on Google Colab, an online Python Notebook IDE. It is important to note that the participants were required to be familiar with Python as a programming language and be enrolled in a Computer Science or related undergraduate program to participate in the study. Screen recordings as well as eye tracking, mouse tracking, textual input, the language model's responses, and supplementary demographics information were recorded. Moreover, we collected responses from participants through a post-study semi-structured interview and questionnaire to gain insights into participants' thought processes before and after the tasks completion. The objective behind these tasks include investigating why, how, and when participants misidentify hallucinations.

4 Analysis

The aim of our data analysis was to understand participants' interaction with ChatGPT using a number of different features, mainly areas of interest (AOIs) (i.e. how much time eyes were focused in a specific area of the screen), fixation count (i.e. number of times eyes were focused on a specific point), fixation duration, time to first fixation (TTFF), and fixation revisitations. We defined five AOI types, informed by work conducted to find the optimal

bias-free AOI distribution (Sqalli et al., 2021) as well as on the requirements of the experiment.

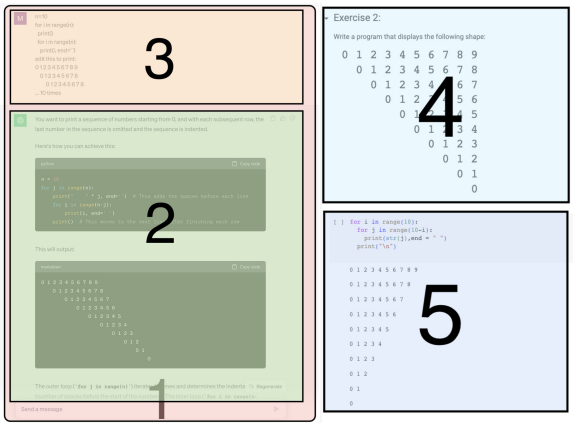

Figure 1: *Pred-defined Areas of Interest (AOIs) for the ChatGPT experiment.* AOI 1: ChatGPT interface (red). AOI 2: ChatGPT response (green). AOI 3: Student participant's prompt (yellow). AOI 4: Programming exercise task (light blue). AOI 5: Participant's function code implementation and test via Google Colab (dark blue).

The five defined AOIs represent the five eye movement hotspots that correspond to the experiment research questions. These are shown in figure 1. Below is a description for each AOI: 1. The ChatGPT interface (highlighted in red). 2. The answer provided by ChatGPT (highlighted in green). 3. The questions that the students prompt ChatGPT with (highlighted in yellow). All these aforementioned AOIs are from the ChatGPT interface. The following two AOIs are from the Google Colab interface where students read the programming problem, and provide their code as an answer. 4. The programming exercise prompt (highlighted in light blue) and the answer that the student provides, and finally 5. the answer that the students run, test and provide as their final answer to the programming exercise. We use all of those AOI distributions in our analysis to examine the behavior of students when solving a programming exercise. We also correlate some of the eye tracking findings with the results of post-study semi-structured interviews conducted after the end of the experiment.

5 Results

The aim is to understand how eye tracking behavior reflects students' use of the ChatGPT interface. Also, how this interaction affects their perceived complexity and time required to understand and solve programming problems. The results are based on the analysis of a set of 120 eye tracking cases. These cases were gathered through the participation of 6 undergraduate students majoring in Computer Science. Four students were in their senior year, and two were in their junior year. The eye tracking metrics were calculated based on the duration of the experiment that spanned across a mean of 17 minutes and 43 seconds for all the six participants.

5.1 Eye Tracking Features

Table 1 presents the calculated mean for different eye tracking features across the five AOIs defined in section 4. The features are respectively: 1. The duration in milliseconds during which the student was looking at a specific AOI. 2. This duration converted to a percentage during which the student was looking at that specific AOI. 3. The dwell count at that specific AOI. Dwell time measures the duration a person's gaze remains fixed on a specific point of interest. 4. Hit time, in milliseconds refers to the amount of elapsed time before the participant dwells or fixates at a specific AOI. 5. A revisit count when the student looks at a specific AOI, leaves it and comes back to it through a fixation. 6. the Time To First Fixation (TTFF). It measure the time it takes for the student's eyes to fixate on a specific point of interest after being presented with a visual stimulus or display. 7. Saccades count referring to the number of saccades made by a student during the period when looking at a specifid AOI. Saccades are rapid, involuntary eye movements that shift the point of gaze from one location to another. 8. The duration of those saccades in milliseconds, and finally 9. The mouse click count across that specific AOI.

5.2 Statistical Significance of Results

To identify the significance of the eye tracking metrics found, we perform the Kruskal-Wallis statistical test (Wallis). This test was selected because it allows for comparison between the five defined AOIs across the experiment. Since we were comparing between more than two unpaired datasets (five defined AOIs) in a non-normal distribution of data that had the same shape, which satisfied the criteria for applying the Kruskal-Wallis test on the four experimental tasks. We used an α of .001 as the cutoff for significance.

The results from applying the test indicate that the AOI Duration, The Hit Time, and the Time to the First Fixation have a p-value less than the cutoff

Table 1: *Eye Tracking Features Per Area of Interest*.*Indicates that features were statistically significant according to the Kurk-Wallis test ($p < .001$) at distinguishing eye-tracking behavior between AOIs.

Information / AOI	ChatGPT	Task 1	Task 2	Task 3	Task 4	Task Q
AOI Duration* (ms)	193,926	71,100	12,396	82,206	92,012	12,749
Rel. AOI Duration (%)	20.4	7.5	13.0	8.6	9.7	1.3
Dwell Count	47	5	40	31	38	6
Hit Time* (ms)	13,791	3.2	5.5	10.4	14	1,442
Revisit Count	29	1	10	21	27	3
Fixation Count	317	222	249	117	211	32
TTFF* (ms)	13,799	0	0	18	0	1,484
Saccades Count	371	254	287	105	223	42
Saccades Duration (ms)	35	36	36	36	38	39
Mouse Click Count	8	1	22	1	10	0

for significance (p-value = .00034 $<$ .001), which achieves significant difference among the participants working on different exercises. Translating this to eye-tracking behavior, at the granular level, there is a significant difference in the proportion of fixations that the programming students accorded to different AOIs depending on the tasks at hand. The results indicate that the students' focus on different AOIs varied according to the programming exercise.

6 Discussion

The statistical significance for the AOI duration, the hit time and the TTFF indicate that eye-tracking patterns may be recognized while programming with ChatGPT (RQ1) as well as the specific parameters that may be utilized to interpret a subjects' visual attending behavior (RQ2).

Regarding the significant difference in the time spent across each AOI, it was observed that students fixate more than double their time on the ChatGPT AOI compared to the questions or the IDE to test their solutions. This was also reflected in both the dwell count, and the fixations count. Additionally, the statistically significant results found across the Hit Time and the TTFF features show an important visual attention trend. The average TTFF across the first three exercises was very low (zero in some cases), while in the ChatGPT AOI, it was very high. This reflects that the students focus and fixate on the exercises and the IDE more than the answers that ChatGPT provided (A TTFF value of 0 means that the student fixated in the area as soon as their eyes hit the AOI). This observation was not necessarily true in other AOIs, like the ChatGPT AOI, as the student can look at the area but fixate very late in the observation process. This indicates that the fixation dynamics of the students vary depending on the complexity of the problem they are faced with, since ChatGPT can generate rich responses.

In the interviews conducted after the experiment, students raised the point that by the time they reached the third exercise, they felt that ChatGPT could solve the exercises. Thus, they sought the solution through the interface, which was reflected through the low values in gaze and fixation features across the remaining two exercises (Tasks 3 and 4). Additionally some students mentioned prior to the start of the experiment for the need to use pen and paper to think about the solutions, however, by the end of the experiment, they felt that the availability of ChatGPT reduced the need for a pen and paper. The observation that the increased reliance and comfort in using ChatGPT (within the duration of the experiment alone) may potentially be explained by the "threshold of indignation" (Winograd, 1996). That is, ChatGPT requires a reduced-level of effort from the user (relative to functional coding and testing within an IDE), and correspondingly provides high value responses, such that the threshold of indignation to complete the task is reduced.

Overall, the results indicate that programming students adapt their attention, through gaze and fixations, to the mental workload demanded of them. Finally, this work establishes a foundation to pursue higher-resolution studies of users' visual attending behavior using a larger set of diverse tasks, and to study how the nuances of LLM responses affect users' behaviors.

Ethics Statement

The institutional review board approval for this study was granted by the ethics board of [Anonumized] under the IRB Protocol Reference Number [Anonmyzied] prior to the commencement of the study.

References

Mathieu Acher, Jose Galindo Duarte, and Jean-Marc Jezequel. 2023. On programming variability with large language model-based assistant. In *Proceedings of the 27th ACM International Systems and Software Product Line Conference-Volume A*, pages 8–14.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584.

Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do people and neural nets pay attention to the same words: studying eye-tracking data for non-factoid qa evaluation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 85–94.

Alan Davies. 2018. *Examining Expertise Through Eye Movements: A Study of Clinicians Interpreting Electrocardiograms*. Ph.D. thesis, The University of Manchester.

Andreas Gegenfurtner and Jeroen J. G. van Merriënboer. 2017. Methodologies for studying visual expertise. *Frontline Learning Research*, 5(3):1–13.

Krystal Hu. 2023. ChatGPT sets record for fastest-growing user base - analyst note — reuters.com. [Accessed 10-09-2023].

Hazem Ibrahim, Fengyuan Liu, Rohail Asim, Balaraju Battu, Sidahmed Benabderrahmane, Bashar Alhafni, Wifag Adnan, Tuka Alhanai, Bedoor AlShebli, Riyadh Baghdadi, Jocelyn J. Bélanger, Elena Beretta, Kemal Celik, Moumena Chaqfeh, Mohammed F. Daqaq, Zaynab El Bernoussi, Daryl Fougnie, Borja Garcia de Soto, Alberto Gandolfi, Andras Gyorgy, Nizar Habash, J. Andrew Harris, Aaron Kaufman, Lefteris Kirousis, Korhan Kocak, Kangsan Lee, Seungah S. Lee, Samreen Malik, Michail Maniatakos, David Melcher, Azzam Mourad, Minsu Park, Mahmoud Rasras, Alicja Reuben, Dania Zantout, Nancy W. Gleason, Kinga Makovi, Talal Rahwan, and Yasir Zaki. 2023. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports*, 13(1).

Moritz Langner, Peyman Toreini, and Alexander Maedche. 2023. Leveraging eye tracking technology for a situation-aware writing assistant. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–2.

I MacKenzie. 2012. Human-computer interaction: An empirical research perspective.

Liam Richards Maldonado, Azza Abouzied, and Nancy W. Gleason. 2023. ReaderQuizzer: Augmenting research papers with just-in-time learning questions to facilitate deeper understanding. In *Computer Supported Cooperative Work and Social Computing*. ACM.

Eyal M. Reingold and Heather Sheridan. 2011. *Eye movements and visual expertise in chess and medicine*. Oxford University Press.

Lavinia Salicchi, Alessandro Lenci, and Emmanuele Chersoni. 2021. Looking for a role for word embeddings in eye-tracking features prediction: does semantic similarity help? In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 87–92.

Mohammed Tahri Sqalli, Dena Al-Thani, Mohamed B Elshazly, and Mohammed Al-Hijji. 2021. Interpretation of a 12-lead electrocardiogram by medical students: Quantitative eye-tracking approach. *JMIR Medical Education*, 7(4):e26675.

Kruskal Wallis. Getting Started with the Kruskal-Wallis Test | UVA Library — library.virginia.edu. `https://library.virginia.edu/data/articles/getting-started-with-the-kruskal-wallis-test`. [Accessed 15-10-2023].

Terry Winograd. 1996. *Bringing Design to Software*. Addison Wesley, Boston, MA.

Ramazan Yilmaz and Fatma Gizem Karaoglan Yilmaz. 2023. Augmented intelligence in programming learning: Examining student views on the use of chatgpt for programming learning. *Computers in Human Behavior: Artificial Humans*, 1(2):100005.

Gianpaolo Zammarchi and Claudio Conversano. 2021. Application of eye tracking technology in medicine: A bibliometric analysis. *Vision*, 5(4):56.

Ann C. Zumwalt, Arjun Iyer, Abenet Ghebremichael, Bruno S. Frustace, and Sean Flannery. 2014. Gaze patterns of gross anatomy students change with classroom learning. *Anatomical Sciences Education*, 8(3):230–241.