Examining the Role of LLM-Driven Interactions on Attention and Cognitive Engagement in Virtual Classrooms

Süleyman Özdel
Human-Centered Technologies for Learning
Technical University of Munich, Germany
ozdelsuleyman@tum.de

Can Sarpkaya
Human-Centered Technologies for Learning
Technical University of Munich, Germany
can.sarpkaya@tum.de

Efe Bozkir
Human-Centered Technologies for Learning
Technical University of Munich, Germany
efe.bozkir@tum.de

Hong Gao
School of Future Science and Engineering
Soochow University, China
gaohong@suda.edu.cn

Enkelejda Kasneci
Human-Centered Technologies for Learning
Technical University of Munich, Germany
enkelejda.kasneci@tum.de

ABSTRACT

Transforming educational technologies through the integration of large language models (LLMs) and virtual reality (VR) offers the potential for immersive and interactive learning experiences. However, the effects of LLMs on user engagement and attention in educational environments remain open questions. In this study, we utilized a fully LLM-driven virtual learning environment, where peers and teachers were LLM-driven, to examine how students behaved in such settings. Specifically, we investigate how peer question-asking behaviors influenced student engagement, attention, cognitive load, and learning outcomes and found that, in conditions where LLM-driven peer learners asked questions, students exhibited more targeted visual scanpaths, with their attention directed toward the learning content, particularly in complex subjects. Our results suggest that peer questions did not introduce extraneous cognitive load directly, as the cognitive load is strongly correlated with increased attention to the learning material. Considering these findings, we provide design recommendations for optimizing VR learning spaces.

Keywords

Virtual classroom, eye tracking, cognitive load, human-computer interaction, large language models, educational technologies, AI in education

1. INTRODUCTION

Education is undergoing a significant digital transformation, accelerated by technological advancements and further driven by the COVID-19 pandemic, which necessitated a shift from in-person to digital learning environments [103, 15]. Virtual reality (VR) technologies have become increasingly prevalent in this transformation. Advances in VR technology have made head-mounted displays (HMDs) more affordable and accessible, leading to their widespread application across various fields, including healthcare [39, 76, 50], entertainment [19, 6], and education [82, 23, 13, 24, 36, 25]. In education, VR is revolutionizing traditional teaching methods by transitioning them into a dynamic digital landscape. Institutions like Stanford University have begun conducting entire classes in VR, showcasing VR's potential to transform conventional teaching [90]. Virtual environments enable immersive experiences and enhanced visualizations, which can lead to more effective and engaging learning.

In the digital transformation of education, large language models (LLMs), which are powerful AI systems trained on vast amounts of text data to understand and generate human-like language, are increasingly being applied [46, 1, 102, 40]. In educational settings, integrating LLMs with VR environments enables a more interactive learning experience by allowing students to engage in realistic simulations, ask questions, and receive immediate, contextually accurate responses [38, 58, 44, 27]. These advancements allow for natural conversations and personalized interactions, enhancing the ability to create tailored learning experiences that adapt to individual student needs [66, 46, 65]. By combining VR with LLMs, educational environments can offer more interactive and flexible learning experiences that accommodate diverse learning styles and preferences, improving student engagement and retention.

Building on advancements in VR and AI-driven systems like LLMs, educational settings have increasingly enabled adaptable and tailored learning experiences that address individual student needs and preferences. Such personalized environments allow students to progress independently and at their own pace, aligning their learning journeys with their unique goals [77, 88, 89]. However, despite offering substantial flexibility and autonomy, these individualized learning settings often lack the collaborative dynamics characteris-

Süleyman Özdel, Can Sarpkaya, Efe Bozkir, Hong Gao, and Enkelejda Kasneci. Examining the Role of LLM-Driven Interactions on Attention and Cognitive Engagement in Virtual Classrooms. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 155–169. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870207

(a) View 1 from the virtual classroom.

(b) View 2 from the virtual classroom.

Figure 1: Views from the LLM-driven virtual classroom environment.

tic of traditional classroom environments. Peer interactions, particularly question-asking behaviors, significantly enhance the learning experience by capturing students' attention, fostering deeper engagement, and encouraging cognitive elaboration [48, 60, 5]. These interactions function as instructional signals, effectively guiding learners toward critical instructional content, focusing their attention, and enhancing comprehension [94, 64, 29]. Thus, incorporating peer questions into virtual learning environments can not only direct student attention to crucial concepts but also enrich engagement and foster meaningful interactions within the learning process.

Addressing these challenges and building on technological advancements, our study focuses on integrating LLMs into virtual learning environments to enhance the realism and effectiveness of VR-based education. While previous research has explored the use of VR in education [13, 24, 56, 82, 41], particularly in simulating real-world environments and improving engagement, the integration of LLMs to create more interactive and personalized experiences has not been widely studied. Few studies [58] have examined how LLM-driven interactions within VR classrooms can mirror the dynamics of traditional classrooms, particularly in peer-to-peer or teacher-student exchanges. Our work takes a first step toward addressing this gap by simulating a fully LLM-driven virtual environment (see Figure 1), where an LLM-powered teacher delivers content based on provided slides, complemented by LLM-powered peer interactions. This setup aims to create an immersive learning experience that mirrors real-world educational settings. To evaluate the impact of this integration, we test two conditions: LLM-driven Peer Interaction with Questions and Answers (Peer-QnA), where LLM-driven peers (students) ask questions to the teacher, creating a more interactive classroom environment, and LLM-driven Peer without Question and Answer Interaction (Peer-NoQnA), where LLM-driven peers do not ask questions, leaving the participant as the only entity interacting with the teacher. To understand the effects of these two settings, we evaluated factors including student engagement, cognitive load, and eye-tracking behaviors within the VR.

The contributions of this work are fivefold: (1) we designed a fully LLM-driven virtual classroom environment and collected data from 19 participants, demonstrating the effectiveness of LLM-driven interactions in immersive educational settings; (2) we evaluated two distinct interaction conditions, Peer-QnA and Peer-NoQnA, and found significant differences in student attention, with the Peer-QnA condition leading to increased attention and engagement with the learning content; (3) we analyzed cognitive load using NASA Task Load Index (NASA-TLX) assessments, supplemented by eye-tracking data, revealing that the Peer-QnA condition resulted in higher cognitive load, as evidenced by increased pupil diameter, which strongly correlated with student attention on crucial content; (4) peer questions led to more targeted attention, resulting in longer mean fixation duration and shorter average saccade amplitude; and (5) in more complex subjects, these changes in cognitive load and visual attention were more noticeable, emphasizing the importance of subject complexity.

2. RELATED WORK

In the evolving educational technology landscape, immersive VR and LLMs have become transformative tools that significantly impact learning environments. In the following subsections, we review the literature in these two areas, highlighting how each contributes to advancements in educational technology.

2.1 VR Environments in Education

Virtual learning spaces are increasingly integrated into educational settings [78, 82], providing valuable insights from both teacher and student perspectives [72, 68]. For teachers, VR allows them to simulate complex, real-world scenarios, allowing them to practice classroom management and lesson delivery in a controlled environment [74, 37]. This improves their confidence and teaching strategies before entering a real classroom. From the student perspective, VR enhances engagement by creating immersive, interactive environments that support experiential learning, enabling students to explore concepts in a more hands-on way than traditional methods [104, 67, 20].

In addition to offering general classroom preparation, particularly for teachers in training, VR environments provide a unique opportunity to simulate complex classroom scenarios that can help develop essential teaching and classroom management skills. To this end, Westphal et al. [100] found that

student teachers who focused on self-reflection using first-person pronouns in a VR environment experienced higher stress levels, which caused increased stress in subsequent teaching sessions. This highlights the psychological impact of self-focused reflection in VR settings and the importance of preparing teachers to manage stress effectively. Similarly, Huang et al. [37] examined how complex classroom environments, characterized by multiple and overlapping disruptions, impact student teachers' ability to detect and respond to those disruptions. Their study reveals that higher complexity reduces the likelihood of noticing and effectively addressing disruptions, underscoring the potential of VR to simulate challenging teaching environments that can better prepare teachers for real-world scenarios. Additionally, Huang et al. [35] explored the effect of class size on stress levels in pre-service teachers. The authors find that larger class sizes in a VR classroom significantly increase heart rate and perceived stress, indicating that VR effectively simulates classroom management challenges, helping educators develop the skills needed to handle real-world classroom demands.

From the student perspective, VR environments have been shown to enhance engagement, motivation, and overall learning outcomes. Liu et al. [57] demonstrate that primary school students in an immersive VR classroom achieved higher academic success and increased science motivation compared to those in traditional classrooms while also experiencing reduced cognitive load. This suggests that VR can create more engaging and less mentally taxing learning environments for students. Furthermore, Gao et al. [24] and Bozkir et al. [13] explore how various factors, such as student seating positions, visualization styles, and hand-raising behaviors of virtual peers, impact students' engagement and attention in a VR classroom. The findings reveal that students seated at the back of the classroom struggle to effectively extract information, while realistic visualization styles of avatars lead to better engagement with lectures.

Building on previous research that has examined student behaviors and interactions in VR classrooms, Hasenbein et al. [32] investigate how students interact with social comparison information, particularly in relation to peers' achievement-related behaviors. The authors' findings indicate that students who spend more time observing their peers' achievements tend to have lower self-evaluations, highlighting the psychological effects of peer interactions in virtual settings. Stark et al. [92] examine student interactions by using gaze entropy to identify and differentiate classroom discourse events. They are able to predict teacher-led activities and explanations with a high degree of accuracy, demonstrating the potential of gaze entropy as a tool for analyzing classroom participation and engagement in VR settings. While existing studies emphasize the immersive potential of VR in education, there is a research gap regarding fully LLM-driven, dynamic individual learning environments that closely resemble real-world educational experiences. The effects of these AI-driven interactions on student engagement, attention, and learning outcomes are still largely unexplored.

2.2 Large Language Models in Education

Recent advancements in LLMs significantly expand their applications across various fields, including healthcare [98], education [46], and beyond [105, 101]. In the education domain, these models are now playing a larger role in enhancing both teaching and learning experiences. These models provide personalized learning opportunities, helping students learn independently and adapt to their unique needs, thereby contributing to more equitable education [46, 102, 73, 31].

LLMs have demonstrated their versatility and effectiveness across various educational levels, from primary schools to universities. For instance, Yan et al. [102] identify 53 different application scenarios where LLMs are used to automate educational tasks at different levels, including assessment and grading, teaching support, and knowledge representation. This broad applicability highlights the potential of LLMs for innovating traditional educational processes. At the university level, Abd-alrazaq et al. [1] explore the potential of LLMs in medical education, noting their ability to innovate curriculum design, teaching methodologies, and student assessments. Similarly, in secondary education, Lieb and Goel [54] introduced NewtBot, a personalized tutor chatbot for physics students, which provides positive learning experiences and demonstrates the potential of LLMs as effective virtual tutors. Moreover, Lu and Wang [61] utilize these models to simulate student profiles for evaluating multiple-choice questions. The authors find that the simulated responses are consistent with real student answers, aiding in the refinement of question quality.

LLMs have also been utilized in enhancing interactive learning environments. Liu et al. [58] introduced ClassMeta, a GPT-4 driven agent that simulates an active student in a VR classroom. This integration of LLMs with VR significantly expands engagement and learning outcomes, providing a more immersive and effective educational experience. Similarly Izquierdo-Domenech et al. [38] combined VR and LLMs to create context-aware educational experiences. Participants using this integrated setup achieve significantly better learning outcomes compared to those using traditional methods, highlighting the potential of LLMs to enhance interactive learning.

Despite the promising advancements in integrating LLMs with VR environments [11], research in this area is still in its early stages. While there have been successful applications of LLMs for tasks like individual tutoring and knowledge representation, there is still much to explore the effects of interacting with AI-powered peers and teachers in immersive settings have not been fully investigated. This highlights the need for further research into how LLMs can enhance not only individual learning experiences but also create more engaging and interactive virtual classrooms. Our study takes an initial step toward addressing these gaps by exploring how LLMs can be integrated into VR to simulate more interactive and realistic classroom dynamics.

3. METHODOLOGY

As LLM-driven classrooms become more common, it is important to understand how students interact and learn in these new environments. The main purpose of this study is to evaluate student behaviors in fully LLM-driven classroom settings and to analyze the impact of LLM-driven peer questions on engagement, attention, cognitive load, and learning

outcomes. Understanding how student attention and cognitive load evolve in these virtual environments can guide the design of more effective and engaging LLM-driven learning spaces. In this section, we provide an overview of the participant details, apparatus, experimental design, procedure, measurement techniques, and data pre-processing steps.

3.1 Participants

The study included 19 participants with a mean age of 25.32 and standard deviation 8.57, with a gender distribution of 68.42% male ($n = 13$) and 31.58% female ($n = 6$). Educational backgrounds varied, with 63.16% holding a bachelor's degree, 26.32% having completed high school or equivalent, and 10.53% possessing a master's degree. Occupationally, 68.42% were students, while 31.58% were recent graduates employed in various industries. Most participants (68.42%) had prior experience with VR; however, only 5.26% had used VR in educational settings. Additionally, 94.74% had interacted with LLMs before, indicating a certain level of familiarity with AI technologies.

3.2 Apparatus

The study was conducted using a VR classroom environment designed in Unity3D (see Figures 1). The virtual classroom was equipped with avatars representing a teacher and students, all powered by LLMs to simulate real-time interactions. Specifically, we utilized "ChatGPT-4o" [69] to power the interactions within the classroom. For speech recognition, we employed OpenAI's Whisper API [70] for speech-to-text conversion and Amazon Polly [86] for text-to-speech synthesis. This environment was designed to mirror a realistic classroom, with all avatars equipped with animations to enhance realism. Student avatars featured both speaking and idle animations, while the teacher avatar included additional variations of idle and speaking animations. For question-asking behavior, student avatars raise their hands before speaking, while the teacher avatar uses both a selection gesture and verbal cue when calling on a student. Participants took on the role of a student seated at the center of a 3×3 grid, surrounded by eight desks assigned to LLM-driven peers. Participants could ask questions using the HTC Vive controller; pressing the trigger button initiated the speech input system, allowing them to verbally ask their question. The study was conducted with a Varjo XR-3 [96] mixed reality HMD paired with a desktop featuring a 13th Gen Intel Core i7-13700K processor, 32.0 GB of RAM, and an NVIDIA GeForce RTX 4080 GPU. Eye-tracking data was collected using the Varjo XR-3's built-in eye tracker, operating at the maximum sampling rate of 200Hz.

3.3 Experimental Design

In the study, we evaluated the VR classroom environment by addressing the effectiveness of the virtual setting, which is essential for understanding how well it replicates a real classroom and how immersive and engaging it is for individual participants. Additionally, two separate conditions were tested to examine the effects of interactive dynamics. In the first condition, the participant could address the teacher and ask questions, but the LLM-powered students did not interact. In the second condition, both the participant and the LLM-powered students could address the teacher by asking questions. In both conditions, the teacher presented a set of instructional slides, explaining the content of each one. After the explanation of each slide, a structured opportunity for questions followed. In the Peer-NoQnA condition, the participant could ask a question using a button on the controller, and the teacher responded accordingly. In the Peer-QnA condition, the participant could still ask a question using the same method. Additionally, one or two randomly selected AI student avatars also asked questions after each slide, regardless of whether the participant chose to ask one.

In this study, we employed a within-subjects experimental design, where each participant experienced two interaction conditions within a virtual reality classroom. For each participant, to avoid content repetition, each condition was associated with a different topic, and the order of topic presentation was counterbalanced across participants to mitigate potential order effects. This design ensured that any observed differences in engagement or learning outcomes were attributable to the interaction model rather than the sequence in which the topics were presented. Although we observed variation in participant behavior and outcomes across the two topics, topic complexity was not manipulated as an independent variable. Instead, topic assignment served solely as a counterbalancing mechanism. Nonetheless, the differences observed across topics provide valuable exploratory insights into how content complexity may affect attention, cognitive load, and learning outcomes.

In the experiment, participants were exposed to four cases involving two topics: the Double-Slit Experiment and the History of Video Games. We choose those topics to allow for an analysis of user behavior in both a technical and a less specialized or non-technical subject, providing deeper insights into how the virtual environment performs across different types of content. Each case varied based on whether questions were asked solely by the user or by the user and AI students. In Case 1, only the participant was allowed to ask questions during the Double-Slit Experiment, whereas both the participant and AI students could ask questions during the History of Video Games. Case 2 reversed this setup, with both the participant and AI students asking questions during the Double-Slit Experiment, followed by only participant questions during the History of Video Games. Case 3 and Case 4 mirrored the structure of Case 1 and Case 2, respectively, but with the order of topics reversed.

In the experiment, we collected eye-tracking data to analyze student behavior in the virtual reality classroom, following the approach used in other studies [79, 91, 33, 13]. This method provided valuable insights into how participants directed their attention and interacted with various elements of the virtual environment. Additionally, the cognitive load was assessed to measure the mental effort required, consistent with approaches in similar studies [57, 4]. The NASA Task Load Index (NASA-TLX), a widely recognized tool for evaluating cognitive load [21, 14], was used to gather subjective assessments of participants' mental demand, effort, and overall workload. These measures offer a comprehensive approach to understanding attention, engagement, and mental effort in the virtual learning environment, validating the use of eye-tracking and cognitive load as key tools for this evaluation. Additionally, we administer pre- and

post-questionnaires to gather participants' feedback on their experience in the virtual classroom environment.

3.4 Procedure

Upon arrival, participants were welcomed and asked to complete informed consent forms and a pre-questionnaire with demographic questions. Following this, they were introduced to the first condition of the experiment, conducted in a VR classroom environment. After completing the first condition, participants took the NASA-TLX test to assess the task load. They then proceeded to the second condition, again in the VR classroom, followed by the NASA-TLX test once more. Finally, participants completed a post-questionnaire that gathered general feedback on the overall experience. To ensure participants remained attentive to the lecture content in the virtual classroom, a set of questions related to the presented topic was asked after each condition, assessing their retention and engagement. Each VR session took approximately 15-18 minutes, and the total duration of the experiment was about 1 to 1.5 hours. Participants received compensation for their time and participation.

3.5 Measurements

Data collection in this study involved multiple methods to comprehensively assess participant experiences and outcomes.

3.5.1 Visual Scanpath

Eye-tracking data was collected using the Varjo XR-3 headset's built-in capabilities. From this data, we extract fixation points, where the gaze remains focused for a significant period, and saccades, which are rapid eye movements between fixation points. We analyze key metrics such as total fixation duration, mean fixation duration, saccade amplitude, and saccade velocity.

The eye-tracking data is crucial for understanding the behaviors of the students in the virtual classroom [24, 13, 26, 22]. Total fixation durations represent the user's attention to specific content and areas of interest, indicating how long the information is engaged with [30]. In our experiment, we normalize these durations by dividing them by the total fixation duration for each participant to evaluate attention. The mainboard and teacher were identified as the primary instructional content, and we designated them as the key areas of interest for analyzing how participants focused on the content. Mean fixation duration serves as an indicator of cognitive processing demands [43, 30]. Higher values generally suggest that deeper cognitive processing is required to process the information being viewed. It is engaged in more complex mental activities, such as understanding, analyzing, or integrating different pieces of information, which also supports meaningful learning [80, 87]. Saccade amplitudes provided insight into how broadly or narrowly participants scanned the environment, indicating their visual exploration patterns. Larger saccade amplitudes suggest participants were scanning across a wider area, while smaller amplitudes indicate higher cognitive load with a more concentrated focus on particular content [17]. Saccade velocities indicate how cognitive load and task demands impact attention and engagement. As cognitive load rises, saccade velocity tends to increase. This indicates deliberate focus, where participants take more time to process detailed information or engage more with the content [28] However, higher average saccade velocities are often associated with increased stress and reduced concentration during cognitive tasks [8, 52]. In addition to fundamental fixation and saccade metrics, we examined the total fixation duration on key objects, such as the mainboard and teacher, to assess students' attention. These objects were identified as the primary sources of instructional content, reflecting where the core learning material was delivered.

3.5.2 Cognitive Load

Cognitive Load Theory [94, 93] is a framework for understanding the mental effort involved in learning. Cognitive load is classified into three categories: intrinsic, extraneous, and germane. Intrinsic load is primarily related to the inherent difficulty of the material. Extraneous load arises from unnecessary complexity in the environment. Germane load results from processing information to support understanding [49, 99]. Effective instructional design focuses on minimizing extraneous load while enhancing germane load, enabling learners to concentrate on the essential material without being overwhelmed. In this study, we evaluate the impact of the fully LLM-driven virtual classroom on participants' cognitive load to understand how interacting with AI-driven peers and teachers influences mental effort. The NASA-TLX was used after each condition to assess participants' perceived workload. In addition to measuring cognitive load with NASA-TLX, pupil diameter was also utilized to assess the cognitive load experienced by participants during the experiment, serving as an objective indicator of cognitive effort [7, 51, 47, 12].

Additionally, we investigate the relationship between cognitive load and visual attention using Pearson correlation and linear regression analyses. The Pearson correlation measured the strength and direction of the association between cognitive load, as assessed by NASA-TLX, and normalized fixation duration on primary instructional elements. Following the correlation analysis, a linear regression was conducted to predict cognitive load based on normalized fixation duration on these key instructional areas. This regression analysis quantified how much variance in cognitive load could be explained by participants' attention to these primary content areas.

3.5.3 Questionnaires

We administered several questionnaires in the study, and their details are as follows.

Pre-Questionnaire includes demographic questions to capture participants' background information, including their age, gender, education level, prior experience with VR and AI technologies, and familiarity with the subject. Knowledge questionnaire has a multiple-choice test designed to assess their understanding and retention of the content presented during the VR sessions. To investigate the relationship between visual attention and learning outcomes, we conducted a regression analysis. Post-Questionnaire was administered to gather participants' overall impressions and feedback on their experience within the virtual classroom environment. This questionnaire was divided into five categories and employed a 5-point Likert scale with response options: "strongly disagree," "disagree," "neutral," "agree," and

"strongly agree."The questionnaire was structured into five categories. First, we focused on "Technical Challenges and Audiovisual Quality", addressing any technical issues participants faced that could have impacted their engagement. Second, we evaluated the Teacher-Student Interaction Quality, focusing on the clarity of the LLM-driven teacher's content delivery and the effectiveness of its responses during interactions. Third, we examined "Student Participation and Peer Influence", particularly how the question-and-answer dynamics between peers and the teacher affected participants' attention, engagement, and overall learning process. Fourth, we assessed "Assessment Quality and Relevance", collecting participants' feedback on the appropriateness and difficulty of the test questions used to evaluate their understanding of the lecture content. Finally, we collected feedback on Overall Experience and Satisfaction, reflecting participants' general impressions of the LLM-driven VR classroom environment.

3.6 Data Processing

Eye-tracking data was processed using the Identification by Velocity Threshold (I-VT) algorithm [84, 45] to identify fixations and saccades. The I-VT algorithm classifies eye movements by measuring gaze velocity, with slower movements being categorized as fixations and faster movements as saccades. We also incorporated the head movements in the fixation detection, as fixations were only counted when both eye and head movements were stable. The specific criteria used for detecting fixations and saccades, including velocity and duration thresholds [24, 2], are provided in Table 1. For the pupil diameter data, we applied a Savitzky-Golay filter [85] to smooth the data and remove noise. Following this, divisive baseline correction with a baseline duration of 1 second [63] was used to normalize the readings, ensuring a more precise analysis of cognitive load and engagement, as seen in similar studies [24, 10, 9]. Similarly, to analyze the total fixation duration on key objects (mainboard and teacher), we normalized this metric by dividing it by the overall fixation duration for each participant.

Table 1: Criteria for Fixation and Saccade Detection.

Event	Velocity (v)	Duration (Δ)
Fixation	$v_{head} < 7°/s$ $v_{gaze} < 30°/s$	$\Delta_{fixation} > 100\ ms$ $\Delta_{fixation} < 500\ ms$
Saccade	$v_{gaze} > 40°/s$	$\Delta_{saccade} > 20\ ms$ $\Delta_{saccade} < 100\ ms$

3.7 Analysis

We conducted a separate analysis of cognitive load, visual scanpath, and learning outcomes for each topic, the Double-Slit Experiment and the History of Video Games. For both topics, we compared the Peer-QnA and Peer-NoQnA conditions in terms of cognitive load, pupil diameters, fixations, saccades, and learning outcomes. We conducted an independent t-test for normally distributed samples. For distributions that did not conform to a normal distribution, we used the non-parametric Wilcoxon signed-rank test. Normality was evaluated using the Shapiro-Wilk test for sample sizes under 2000 [81, 83] and the Kolmogorov-Smirnov test [55] for larger samples. In all analyses, a significance level of $\alpha = 0.05$ was applied to determine statistical significance.

To analyze the questionnaire data, we calculated the mean and standard deviation for each question using the Likert scale, where "strongly agree" corresponds to 5 and "strongly disagree" corresponds to 1 [42]. A mean score above 3 indicates general agreement or a positive response, while a mean below 3 reflects disagreement or negative feedback. For reverse-worded questions, we adjusted the scoring by inverting the response values during analysis to maintain consistency. Additionally, we calculated internal consistency for each category using Cronbach's alpha [95]. This method evaluates how well the items within each category measure the same construct. A Cronbach's alpha value above 0.8 is considered high reliability, between 0.6 and 0.8 indicates moderate reliability, and below 0.6 suggests low reliability [71]. A high alpha value indicates strong internal consistency, meaning that the items within each category are reliably measuring the intended construct. We used Cronbach's alpha as a robust reliability indicator, although it can underestimate reliability when applied to a small number of items [16, 62, 95].

4. RESULTS

We presented the results for each topic's cognitive load analysis, visual-scanpath analysis, and learning outcomes separately. Following this, we provided the findings from a general questionnaire, where user feedback on their overall experience was collected.

4.1 Topic 1: Double-Slit Experiment

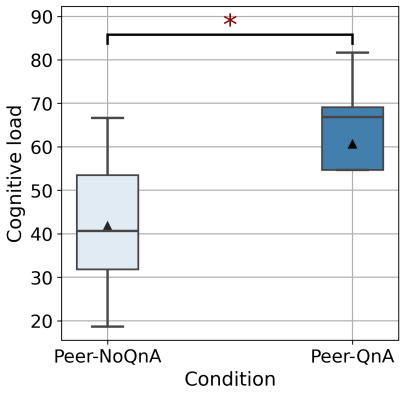
4.1.1 Cognitive Load Analysis

In the Double-Slit Experiment, the interactivity of LLM-driven peers significantly impacted cognitive load, as assessed by the NASA-TLX. We observed significantly higher cognitive load scores in the Peer-QnA condition ($M = 60.50$, $SD = 18.16$) compared to the Peer-NoQnA ($M = 41.73$, $SD = 15.36$). This difference was statistically significant, with a p-value of $p = .026$ ($p < .05$), as given in Figure 2 (a). This finding is further supported by pupil diameters, with significantly higher mean pupil diameters in the Peer-QnA condition ($M = .59, SD = .11$) compared to the Peer-NoQnA condition ($M = .51, SD = .19$), indicating a significant difference ($p < .001$), as shown in Figure 2 (b).

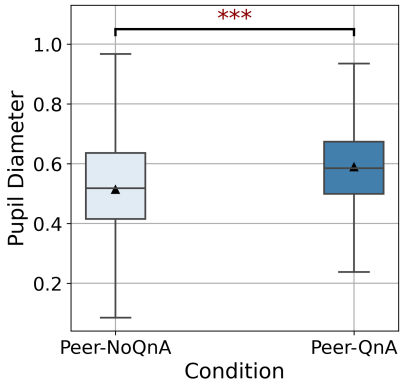
In our regression analysis, we identified a significant relationship between cognitive load and the normalized total fixation duration on the primary instructional content. The Pearson correlation coefficient was $r(18) = 0.60$, $p = .0067$, indicating a strong positive correlation between these variables. As the total fixation duration on the primary instructional content increased, cognitive load also increased. The regression model explained a significant portion of the variance in cognitive load, with $R^2 = .36$, adjusted $R^2 = .32$, $F(1, 18) = 9.53$, $p = .007$, indicating that 32.2% of the variance in cognitive load could be attributed to participants' fixation duration on the main instructional content.

4.1.2 Visual Scanpath Analysis

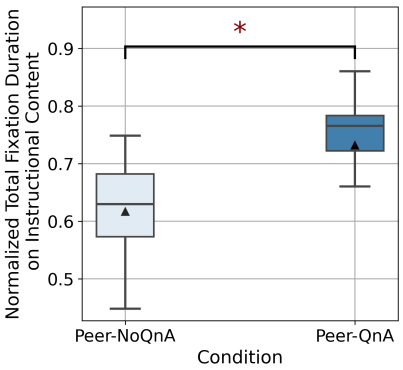
A significant difference in fixation durations is observed between the Peer-QnA and Peer-NoQnA groups, with the Peer-QnA condition showing slightly higher means. The mean fixation duration for the Peer-QnA condition is $M = 233ms$, $SD = 103ms$, compared to $M = 228ms, SD = 101ms$ for



(a) Cognitive load scores.



(b) Mean pupil diameters.



(c) Fixation duration on primary instructional content.

Figure 2: Results for the Double-Slit Experiment across Peer-QnA and Peer-NoQnA conditions.

the Peer-NoQnA condition, indicating a statistically significant difference ($p < .001$).

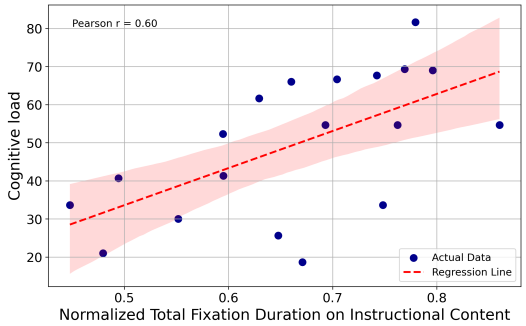
A significant difference is found between the Peer-QnA and Peer-NoQnA conditions in saccade amplitudes. The mean saccade amplitude for the Peer-QnA condition is lower, $M = 135.99°, SD = 68.74°$, compared to $M = 137.87°, SD = 69.89°$ for the Peer-NoQnA condition, with a statistically significant difference ($p = .015$, $p < .05$). There are no significant differences in saccade velocities between the Peer-QnA and Peer-NoQnA groups. The mean saccade velocity for the Peer-QnA condition is $M = 127.22°/s, SD = 68.70°/s$, while for the Peer-NoQnA condition, it is $M = 128.29°/s, SD = 69.34°/s$, with no statistically significant difference.

Additionally, we observed that participants gazed significantly more at the primary instructional content. The normalized total fixation duration is significantly higher in the Peer-QnA condition ($M = 0.72, SD = 0.11$) compared to the Peer-NoQnA condition ($M = 0.60, SD = 0.08$). The t-test results confirmed a significant difference between these two conditions in the Double-Slit Experiment ($p = .02$, $p < .05$), as shown in Figure 2 (c).

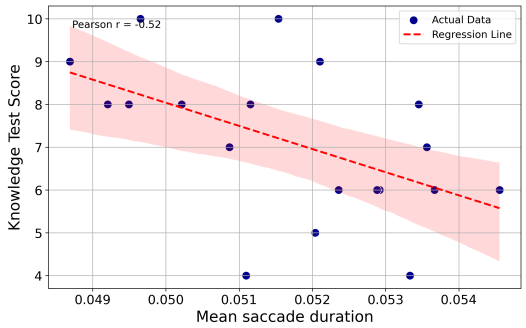
4.1.3 *Learning Outcome*

The results of the knowledge questionnaire indicate that in the Peer-QnA condition, the scores were higher compared to the Peer-NoQnA condition, although the difference was not statistically significant. For the Double-Slit Experiment, the mean score for the Peer-QnA condition was $M = 7.50, SD = 2.07$, while the mean score for the Peer-NoQnA condition was $M = 6.82, SD = 1.60$ as shown in Figure 4 (a).

Additionally, we applied regression analysis to identify the relationship between visual scanpath metrics and the knowledge questionnaire scores for the Double-Slit Experiment. The analysis revealed a significant negative correlation between mean saccade duration and knowledge scores, with Pearson correlation $r(17) = -.52$, $p = .024$, indicating that shorter saccade durations were associated with higher questionnaire scores, as shown in Figure 3 (b). Linear regression analysis further demonstrated that mean saccade duration explained a significant portion of the variance in the questionnaire scores, $R^2 = .27$, adjusted $R^2 = .22$, $F(1, 17) = 6.19$, $p = .024$. When we considered only the mean saccade



(a) Cognitive Load vs Normalized Total Fixation Duration on Primary Instructional Content.



(b) Knowledge Questionnaire Scores vs Mean Saccade Duration.

Figure 3: Linear regression results for the Double-Slit Experiment.

duration while participants focused on the primary instructional content, the relationship became more pronounced. The Pearson correlation was $r(18) = -.61$, $p = .006$, indicating a stronger negative correlation. Mean saccade duration explained 33.6% of the variance in knowledge questionnaire scores, with $R^2 = .37$, adjusted $R^2 = .34$, and the regression model showing statistical significance, $F(1, 18) = 10.12$, $p = .005$.

4.2 Topic 2: History of Video Games

4.2.1 Cognitive Load Analysis

In the History of Video Games, cognitive load metrics were generally lower compared to the Double-Slit Experiment. When comparing the Peer-QnA and Peer-NoQnA conditions, no significant differences in cognitive load were found. The mean cognitive load for the Peer-QnA condition was $M = 43.58, SD = 15.38$, and for the Peer-NoQnA condition, it was $M = 42.57, SD = 17.04$. Similarly, normalized pupil diameters showed no significant variation between the two conditions, with $M = 0.64, SD = 0.11$ for Peer-QnA and $M = 0.62, SD = 0.11$. These results indicate that, within the context of the video games topic, the level of interactivity had no statistically significant effect on cognitive load or pupil responses. Similarly, correlation and regression analyses revealed no significant relationship between total fixation duration on the primary instructional content and cognitive load.

4.2.2 Visual Scanpath Analysis

Similar to the Double-Slit Experiment, a significant difference in fixation durations was found between the two conditions. The mean fixation duration for the Peer-QnA condition was slightly higher ($M = 237ms, SD = 101ms$) compared to the Peer-NoQnA condition ($M = 235ms, SD = 103ms$), with the difference being statistically significant, $p = .017$ ($p < .05$). Saccade amplitudes were similar across both conditions. However, there was a significant difference in saccade mean velocities. The mean velocity for the Peer-QnA condition was slightly higher ($M = 1.28°/s, SD = 0.71°/s$) compared to the Peer-NoQnA ($M = 1.26°/s, SD = 0.67°/s$), with a statistically significant difference, $p = .040$ ($p < .05$). This suggests that participants in the Peer-QnA condition exhibited more rapid saccadic movements compared to those in the Peer-NoQnA condition. We did not observe any significant difference in the normalized total fixation duration on the primary instructional content.

4.2.3 Learning Outcome

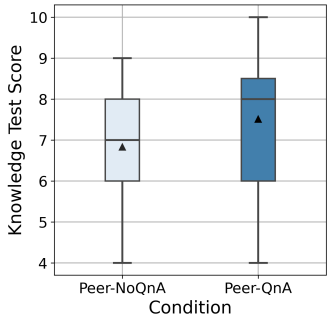
A difference in knowledge questionnaire scores is observed between the conditions. The mean score for the Peer-QnA condition is $M = 7.36, SD = 1.21$, while the Peer-NoQnA condition has a mean score of $M = 5.88, SD = 1.96$. Although the difference is not statistically significant, it approached the threshold ($p = .056$), indicating a potential trend towards improved performance in the Peer-QnA condition, as shown in Figure 4 (b). In the regression analysis, no significant relationship was found between visual scanpath metrics and knowledge questionnaire scores.

4.3 General Analysis

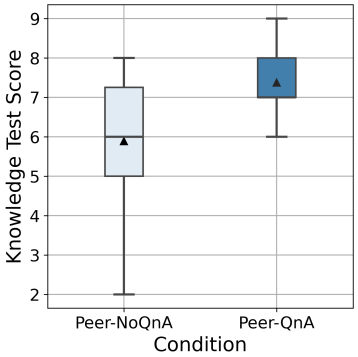
The results of the questionnaire are summarized in Table 2, which includes the mean and standard deviation for each question. For each category, Cronbach's alpha has been calculated to assess the internal consistency of the items, and the corresponding reliability level is provided to indicate the strength of this consistency. Additionally, questions that are reverse-worded are marked with "(R)".

5. DISCUSSION

This section discusses the impact of LLM-driven peer interactions on cognitive load, attention, and learning outcomes.



(a) Knowledge questionnaire scores for Double-Slit Experiment.



(b) Knowledge questionnaire scores for History of Video Games.

Figure 4: Knowledge questionnaire scores for Peer-QnA and Peer-NoQnA conditions.

We explore how peer-driven questions and content complexity influence these factors on attention, engagement, cognitive load, and learning outcomes. Then, user feedback on the LLM-driven VR classroom environment provides insights into the user experience and design improvements.

5.1 Cognitive Load, Attention, and Peer Interactions

The results from the Double-Slit Experiment indicate that the Peer-QnA condition significantly increases cognitive load compared to the Peer-NoQnA condition, as reflected in both NASA-TLX scores and pupil diameter. This increase in cognitive load is attributed to participants' directed attention toward the primary instructional content, supported by the positive correlation between cognitive load and the normalized total fixation duration on key instructional elements. Peer questions effectively direct participants' attention to the lecture material, increasing the cognitive effort required as learners engage more deeply with the primary instructional content. These questions do not directly cause extraneous cognitive load; rather, the observed increase in cognitive load is primarily related to the enhanced focus on the lecture material. Furthermore, no significant increase in cognitive load was observed in the History of Video Games topic, also suggesting that peer questions did not introduce extraneous load. This indicates that peer questions do not inherently increase cognitive load; instead, the increase depends on the directed attention to the main lecture content and intrinsic complexity of the material.

Table 2: The means and standard deviations of each question, along with reliability statistics for each category.

Items	M	S.D.
Technical Challenges and Audiovisual Quality (Cronbach's α = .660, Acceptable)		
I experienced the latency, and it was disturbing. (R)	4.4737	0.6967
I experienced technical issues during the session. (R)	3.7368	1.3267
The audio quality was clear, allowing me to understand the teacher and other students.	4.3684	0.7609
Interaction Quality (Cronbach's α = .875, High)		
The teacher's slide presentations and explanations were understandable and effective.	4.0526	0.7799
The content of the teacher's responses was satisfactory.	3.2632	1.1945
The responsiveness of the teacher was satisfactory.	3.5263	1.0203
The teacher's responses to my questions were adequate and helpful.	3.5789	1.1698
The teacher made clear explanations.	3.3684	1.1648
Students' questions were realistic.	3.8421	0.8342
The length of the questions was appropriate.	3.6316	1.0651
Interaction between the teacher and students seemed natural and fluid.	3.6316	1.0116
Student Participation and Peer Influence (Cronbach's α = .708, Moderate)		
The peer interactions make the environment more engaging.	3.9474	0.8481
The presence of active students in the VR environment enhanced my learning experience.	4.0000	0.8165
Other students' questions added value to my learning experience.	3.8947	0.9366
Seeing other students ask questions encouraged me to ask questions as well.	3.5789	0.9612
Seeing other students ask questions helped maintain my focus on the subject matter.	4.0000	0.6667
The questions asked by other students in the VR classroom were distracting and made it difficult to maintain focus. (R)	4.2105	0.7873
Assessment Quality and Relevance (Cronbach's α = .826, High)		
Multiple-choice questions regarding the lecture content were comprehensive and relevant.	4.2632	0.6534
The level of difficulty of the multiple-choice questions was appropriate for my level of understanding.	3.9474	0.8481
Overall Experience and Satisfaction (Cronbach's α = .794, Moderate)		
I felt comfortable interacting in the VR environment.	3.5789	1.1213
I believe the immersive nature of VR classrooms enhances the learning experience.	3.9474	0.8481
The VR environment made the subject matter more interesting.	3.3684	1.0116
VR classroom enhanced my understanding of the material.	3.2632	0.6534
Using VR technology/experiments changed my perspective on virtual learning positively.	4.1053	0.7375
I would recommend VR classroom experiences to others.	4.1579	0.7647
Overall, I was satisfied with my VR classroom experience.	4.1053	0.5671

In addition to enhancing attention to the main content in the Double-Slit Experiment, the peer questions in the Peer-QnA condition help direct participants to key points within the lesson. This is evidenced by longer mean fixation durations and shorter saccade amplitudes, indicating that participants' attention is not only more focused but also more precisely targeted. Longer mean fixation durations suggest that participants spend more time processing specific elements of the material, allowing for deeper cognitive engagement [75, 30]. Shorter saccade amplitudes, on the other hand, reflect more localized and deliberate eye movements, with participants scanning less broadly across the visual field and honing in on relevant instructional elements [17]. These shorter saccade amplitudes suggest a more efficient and concentrated visual processing strategy, where attention is focused on specific, relevant information without being distracted by peripheral content. Together, these visual attention patterns—longer fixation durations and shorter saccades—indicate better performance [18] and also imply that peer questions acted as signals, guiding participants to focus on critical aspects of the lesson. This aligns with signaling theory [29, 59, 64], which proposes that cues in the learning environment, such as peer questions, can direct learners' attention to the most important information, thereby enhancing the learning process. In this context, the LLM-driven peer questions in the Double-Slit Experiment function as effective signals, guiding participants to identify and concentrate on the most critical parts of the lesson. These questions contribute to a more focused and targeted learning experience, particularly in the more complex instructional environment.

Another important point is that the results from the History of Video Games topic show no significant differences in cognitive load, pupil diameter, or total fixation duration between the Peer-QnA and Peer-NoQnA conditions. This highlights the influence of content complexity on attention and cognitive load. The History of Video Games, being less technically demanding, likely does not require the same level of cognitive effort as the Double-Slit Experiment. As a result, the LLM-driven peer questions do not significantly impact attention on the primary content, but there was a noticeable trend toward improved learning outcomes in the Peer-QnA condition. This trend, though not statistically significant, suggests that peer interactions might still support learning even in less complex topics by fostering engagement and verbal processing rather than through increased cognitive load.

These findings have important implications for the design of LLM-driven virtual learning environments. In more complex subjects, like the Double-Slit Experiment, LLM-driven peer interactions can effectively direct attention toward key instructional elements and increase cognitive engagement. In less complex subjects, such as the History of Video Games, peer interactions may not significantly impact visual atten-

tion but can still provide educational benefits by promoting engagement and verbal processing, enhancing learning outcomes without introducing extraneous cognitive load.

These findings suggest that the effects of peer interactions may be influenced by content complexity. Although topic complexity was not manipulated as an independent variable, exploratory patterns indicate that peer questions in the more demanding topic, the Double-Slit Experiment, were associated with increased cognitive load, more focused visual attention, and improved learning outcomes. In contrast, for the less complex topic, the History of Video Games, peer interactions had no significant effect on cognitive load or attention metrics, though a trend toward improved learning outcomes was observed. These exploratory insights highlight the need for future research to explicitly manipulate topic complexity and examine interaction effects more rigorously.

The observed negative correlation between mean saccade duration and knowledge questionnaire scores indicates that shorter saccade durations are associated with higher learning outcomes. This suggests that participants exhibiting more rapid eye movements between fixations processed the instructional content more effectively. Shorter saccades typically reflect more focused and efficient visual scanning, facilitating quicker identification and engagement with key information. As a result, this visual processing efficiency likely contributed to improved retention and comprehension of the material. These findings emphasize the critical role of visual attention dynamics in enhancing learning outcomes within virtual learning environments.

5.2 User Feedback and Design Implications

The post-experiment questionnaire offers key insights into participants' experiences within the fully LLM-driven virtual classroom, revealing both positive aspects and areas for improvement. Participants generally rated the technical aspects of the VR environment positively. However, "I experienced technical issues during the session. (R)" ($M = 3.74$) is reported as relatively low by some participants, indicating some issues. These were primarily related to the speech-to-text functionality. This emphasizes the importance of system robustness, as those problems could impact the practical use of the technology in real-world educational settings. The interaction quality between LLM-driven peers and the teacher is mostly rated positively. However, some questions, such as "The content of the teacher's responses was satisfactory" ($M = 3.26$), score lower, suggesting room for improvement in response quality. The quality of responses could be enhanced by employing prompting strategies that provide more comprehensive content and by using advanced techniques such as retrieval-augmented generation (RAG), which can deliver more accurate information and reduce hallucinations [34, 53]. Despite this, the overall interaction quality received positive feedback, with positive scores across items like "The teacher made clear explanations" and "The teacher's responses to my questions were adequate and helpful." In general, Peer interactions were well-received, with most participants finding them engaging rather than distracting. For instance, "Peer questions encouraged me to ask questions as well" was rated positively ($M = 3.57$), indicating that peer involvement promotes engagement and participation. Participants also agreed that peer questions enhanced their focus and learning experience, as reflected in statements like "The presence of active students in the VR environment enhanced my learning experience" ($M = 4.0$). These findings suggest that incorporating peer interactions makes the learning environment more dynamic and engaging. However, there is room for improvement in exploring interaction strategies, such as peer-to-peer interaction, which could offer additional benefits. Furthermore, the effectiveness of these interactions may also depend on the individual learning styles of the participants [3, 97]. Regarding the quality of knowledge questionnaires, participants find the questions "comprehensive and relevant" ($M = 4.26$) and the difficulty level appropriate ($M = 3.95$). This feedback, supported by strong reliability scores, indicates that the assessments effectively aligned with the instructional content and measured participants' comprehension.

In terms of the overall experience, the lowest-rated statement is "The VR classroom enhanced my understanding of the material" ($M = 3.26$), reflecting varied perceptions. This suggests that while VR tools can be effective, their success may depend on factors such as the topic, environment design, and individual learning preferences [78, 82]. Despite this, the majority of participants expressed satisfaction with their VR classroom experience, with many agreeing to the statement, "I would recommend VR classroom experiences to others" ($M = 4.16$), indicating a generally positive perception.

Individual virtual learning environments can be designed similarly to traditional classroom settings, even in self-directed and personalized learning contexts. The integration of LLMs, which are currently highly effective and satisfactory for lecture presentations, also allows these environments to be tailored to individual needs. Incorporating peer interactions can help sustain attention and increase engagement. However, the impact of peer interactions may vary depending on the complexity of the subject matter. In more complex subjects, peer interactions are more effective in guiding attention, whereas, in less demanding subjects, their effects are less pronounced. Importantly, these interactions do not introduce excessive cognitive load and can still enhance learning outcomes.

5.3 Limitations and Future Work

While this study highlights the benefits of LLM-driven peer interactions, there are several limitations to consider. The sample size was relatively small, and there were gender imbalances. Expanding the range of subjects and increasing the participant pool would help strengthen the generalizability of the findings. Although avatar behaviors were animated to simulate natural interactions, the realism and quality of these animations and interactions were not formally assessed through user or expert evaluation. Future work could incorporate such assessments and further examine the quality of the AI-generated questions and answers. The study also focused on only two specific topics. Future research could explicitly investigate how topic complexity influences the suitability of virtual environments and identify which interaction settings are most effective for different content types. Additionally, examining long-term learning outcomes and the role of more active learning environments in fully

LLM-driven classroom settings could offer deeper insights into how to optimize these virtual learning experiences.

6. CONCLUSION

In this study, we designed an individual learning environment with a fully LLM-driven virtual classroom, where students could interact with LLM-driven teachers and engage in a classroom setting with LLM-driven peers who also interacted with the instructor. We investigated student behavior using eye-tracking data and cognitive load assessments across two interaction conditions: one where LLM-driven peers asked questions and interacted with the teacher and another where peer interactions were not present. Our findings reveal that LLM-driven peer interactions significantly enhanced student engagement and attention, particularly in complex subjects like the Double-Slit Experiment. The interaction of LLM-driven peers increased the duration of students' fixated time on the primary instructional content, promoting sustained attention and improving learning outcomes. Although cognitive load increased in complex subjects, this was primarily attributed to the heightened attention participants directed toward the learning material. In less complex subjects, peer interactions did not increase attention or cognitive load, yet they demonstrated the potential to enhance learning outcomes without introducing excessive cognitive load.

LLM-driven peer interactions in virtual learning environments not only replicate real-world classroom dynamics but also have the potential to improve learning by keeping students engaged and focused for longer periods. This approach could be particularly valuable in higher education or specialized training, where understanding difficult concepts is crucial. Additionally, incorporating LLM-driven peers allows educators to create more personalized and interactive virtual learning environments, making advanced educational opportunities accessible to a broader range of learners. These insights may also support the development of self-directed learning environments by helping to sustain learner attention and provide more effective, personalized learning experiences. Future research should focus on expanding these insights by exploring more diverse subject areas, different types of interactions, and the long-term impact on learning retention, to fully understand the potential of LLM-driven classrooms in supporting personalized and active learning experiences.

7. REFERENCES

[1] A. Abd-Alrazaq, R. AlSaad, D. Alhuwail, A. Ahmed, P. M. Healy, S. Latifi, S. Aziz, R. Damseh, S. Alabed Alrazak, and J. Sheikh. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education*, 9:e48291, 2023.

[2] I. Agtzidis, M. Startsev, and M. Dorr. 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1007–1015, 2019.

[3] B. Akkoyunlu and M. Y. Soylu. A study of student's perceptions in a blended learning environment based on different learning styles. *Journal of Educational Technology & Society*, 11(1):183–193, 2008.

[4] P. Albus, A. Vogt, and T. Seufert. Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, 166:104154, 2021.

[5] S. M. Alharbi, A. I. Elfeky, and E. S. Ahmed. The effect of e-collaborative learning environment on development of critical thinking and higher order thinking skills. *Journal of Positive School Psychology*, 6(6):6848–6854, 2022.

[6] S. Z. A. Ansari, V. K. Shukla, K. Saxena, and B. Filomeno. Implementing virtual reality in entertainment industry. In *Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021*, pages 561–570. Springer, 2022.

[7] J. Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276, 1982.

[8] M. Behroozi, A. Lui, I. Moore, D. Ford, and C. Parnin. Dazed: measuring the cognitive load of solving technical interview problems at the whiteboard. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 93–96, 2018.

[9] E. Bozkir, D. Geisler, and E. Kasneci. Assessment of driver attention during a safety critical situation in vr to generate vr-based training. In *ACM Symposium on Applied Perception 2019*, pages 1–5, 2019.

[10] E. Bozkir, D. Geisler, and E. Kasneci. Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 1834–1837. IEEE, 2019.

[11] E. Bozkir, S. Özdel, K. H. C. Lau, M. Wang, H. Gao, and E. Kasneci. Embedding large language models into extended reality: Opportunities and challenges for inclusion, engagement, and privacy. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–7, 2024.

[12] E. Bozkir, S. Özdel, M. Wang, B. David-John, H. Gao, K. Butler, E. Jain, and E. Kasneci. Eye-tracked virtual reality: a comprehensive survey on methods and privacy challenges. *arXiv preprint arXiv:2305.14080*, 2023.

[13] E. Bozkir, P. Stark, H. Gao, L. Hasenbein, J.-U. Hahn, E. Kasneci, and R. Göllner. Exploiting object-of-interest information to understand attention in vr classrooms. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 597–605. IEEE, 2021.

[14] J. A. Bueno-Vesga, X. Xu, and H. He. The effects of cognitive load on engagement in a virtual reality learning environment. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 645–652. IEEE, 2021.

[15] B. Bygstad, E. Øvrelid, S. Ludvigsen, and M. Dæhlen. From dual digitalization to digital learning space: Exploring the digital transformation of higher education. *Computers & Education*, 182:104463, 2022.

[16] S. N. Chakrabartty. Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education*, 3(1):1–8, 2013.

[17] S. Chen, J. Epps, N. Ruiz, and F. Chen. Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 315–318, 2011.

[18] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung. Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*, 74:61–72, 2014.

[19] A. Chirico, F. Lucidi, M. De Laurentiis, C. Milanese, A. Napoli, and A. Giordano. Virtual reality in health system: beyond entertainment. a mini-review on the efficacy of vr during cancer treatment. *Journal of cellular physiology*, 231(2):275–287, 2016.

[20] A. Christopoulos, M. Conrad, and M. Shukla. Increasing student engagement through virtual interactions: How? *Virtual Reality*, 22(4):353–369, 2018.

[21] S. Criollo-C, J. Cerezo, A. Guerrero-Arias, A. D. Samala, S. Rawas, and S. Luján-Mora. Analysis of the mental workload associated with the use of virtual reality technology as support in the higher educational model. *IEEE Access*, 2024.

[22] J. Ferdinand, H. Gao, P. Stark, E. Bozkir, J.-U. Hahn, E. Kasneci, and R. Göllner. The impact of a usefulness intervention on students' learning achievement in a virtual biology lesson: An eye-tracking-based approach. *Learning and Instruction*, 90:101867, 2024.

[23] L. Freina and M. Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The international scientific conference elearning and software for education*, pages 10–1007, 2015.

[24] H. Gao, E. Bozkir, L. Hasenbein, J.-U. Hahn, R. Göllner, and E. Kasneci. Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.

[25] H. Gao, L. Hasenbein, E. Bozkir, R. Göllner, and E. Kasneci. Evaluating the effects of virtual human animation on students in an immersive vr classroom using eye movements. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, VRST '22, New York, NY, USA, 2022. ACM.

[26] H. Gao, L. Hasenbein, E. Bozkir, R. Göllner, and E. Kasneci. Exploring gender differences in computational thinking learning in a vr classroom: Developing machine learning models using eye-tracking data and explaining the models. *International Journal of Artificial Intelligence in Education*, 33(4):929–954, 2023.

[27] H. Gao, H. Huai, S. Yildiz-Degirmenci, M. Bannert, and E. Kasneci. Datalivr: Transformation of data literacy education through virtual reality with chatgpt-powered enhancements. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 120–129, 2024.

[28] A. Gibaldi and S. P. Sabatini. The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis. *Behavior Research Methods*, 53:167–187, 2021.

[29] S. M. Glynn and F. J. Di Vesta. Control of prose processing via instructional and typographical cues. *Journal of Educational Psychology*, 71(5):595, 1979.

[30] L. Hahn and P. Klein. Eye tracking in physics education research: A systematic literature review. *Physical Review Physics Education Research*, 18(1):013102, 2022.

[31] A. Han, X. Zhou, Z. Cai, S. Han, R. Ko, S. Corrigan, and K. A. Peppler. Teachers, parents, and students' perspectives on integrating generative ai into elementary literacy education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[32] L. Hasenbein, P. Stark, U. Trautwein, H. Gao, E. Kasneci, and R. Göllner. Investigating social comparison behaviour in an immersive virtual reality classroom based on eye-movement data. *Scientific Reports*, 13(1):14672, 2023.

[33] L. Hasenbein, P. Stark, U. Trautwein, A. C. M. Queiroz, J. Bailenson, J.-U. Hahn, and R. Göllner. Learning with simulated virtual classmates: Effects of social-related configurations on students' visual attention and learning experiences in an immersive virtual reality classroom. *Computers in Human Behavior*, 133:107282, 2022.

[34] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[35] Y. Huang, E. Richter, T. Kleickmann, and D. Richter. Class size affects preservice teachers' physiological and psychological stress reactions: An experiment in a virtual reality classroom. *Computers & education*, 184:104503, 2022.

[36] Y. Huang, E. Richter, T. Kleickmann, and D. Richter. Virtual reality in teacher education from 2010 to 2020. *Bildung für eine digitale Zukunft*, pages 399–441, 2023.

[37] Y. Huang, E. Richter, T. Kleickmann, A. Wiepke, and D. Richter. Classroom complexity affects student teachers' behavior in a vr classroom. *Computers & Education*, 163:104100, 2021.

[38] J. Izquierdo-Domenech, J. Linares-Pellicer, and I. Ferri-Molla. Virtual reality and language models: A new frontier in learning. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 2024.

[39] M. Javaid and A. Haleem. Virtual reality applications toward medical field. *Clinical Epidemiology and Global Health*, 8(2):600–605, 2020.

[40] J. Jeon and S. Lee. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892, 2023.

[41] Q. Jin, Y. Liu, S. Yarosh, B. Han, and F. Qian. How will vr enter university classrooms? multi-stakeholders investigation of vr in higher education. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–17,

2022.

[42] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015.

[43] M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

[44] N. Kapadia, S. Gokhale, A. Nepomuceno, W. Cheng, S. Bothwell, M. Mathews, J. S. Shallat, C. Schultz, and A. Gupta. Evaluation of large language model generated dialogues for an ai based vr nurse training simulator. In *International Conference on Human-Computer Interaction*, pages 200–212. Springer, 2024.

[45] E. Kasneci, H. Gao, S. Ozdel, V. Maquiling, E. Thaqi, C. Lau, Y. Rong, G. Kasneci, and E. Bozkir. Introduction to eye tracking: A hands-on tutorial for students and practitioners. *arXiv preprint arXiv:2404.15435*, 2024.

[46] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education, 2023.

[47] P. Kiefer, I. Giannopoulos, A. Duchowski, and M. Raubal. Measuring cognitive load for map tasks through pupil diameter. In *Geographic Information Science: 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27-30, 2016, Proceedings 9*, pages 323–337. Springer, 2016.

[48] A. King. Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14(4):366–381, 1989.

[49] P. A. Kirschner. Cognitive load theory: Implications of cognitive load theory on the design of learning, 2002.

[50] M. M. Kouijzer, H. Kip, Y. H. Bouman, and S. M. Kelders. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. *Implementation science communications*, 4(1):67, 2023.

[51] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, 13(9):e0203629, 2018.

[52] A. Lewandowska, I. Rejer, K. Bortko, and J. Jankowski. Eye-tracker study of influence of affective disruptive content on user's visual attention and emotional state. *Sensors*, 22(2):547, 2022.

[53] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[54] A. Lieb and T. Goel. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[55] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

[56] X. P. Lin, B. B. Li, Z. N. Yao, Z. Yang, and M. Zhang. The impact of virtual reality on student engagement in the classroom–a critical review of the literature. *Frontiers in Psychology*, 15:1360574, 2024.

[57] R. Liu, L. Wang, T. A. Koszalka, and K. Wan. Effects of immersive virtual reality classrooms on students' academic achievement, motivation and cognitive load in science lessons. *Journal of Computer Assisted Learning*, 38(5):1422–1433, 2022.

[58] Z. Liu, Z. Zhu, L. Zhu, E. Jiang, X. Hu, K. Peppler, and K. Ramani. Classmeta: Designing interactive virtual classmate to promote vr classroom participation. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2024.

[59] N. L. Loman and R. E. Mayer. Signaling techniques that increase the understandability of expository prose. *Journal of Educational psychology*, 75(3):402, 1983.

[60] Y. Lou, P. C. Abrami, and S. d'Apollonia. Small group and individual learning with technology: A meta-analysis. *Review of educational research*, 71(3):449–521, 2001.

[61] X. Lu and X. Wang. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27, 2024.

[62] C. P. Malkewitz, P. Schwall, C. Meesters, and J. Hardt. Estimating reliability: A comparison of cronbach's α, mcdonald's ωt and the greatest lower bound. *Social Sciences & Humanities Open*, 7(1):100368, 2023.

[63] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods*, 50:94–106, 2018.

[64] P. D. Mautone and R. E. Mayer. Signaling as a cognitive guide in multimedia learning. *Journal of educational Psychology*, 93(2):377, 2001.

[65] J. Meyer, T. Jansen, R. Schiller, L. W. Liebenow, M. Steinbach, A. Horbach, and J. Fleckenstein. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199, 2024.

[66] E. R. Mollick and L. Mollick. Using ai to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*, 2023.

[67] S. Mystakidis. Distance education gamification in social virtual reality: A case study on student engagement. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*, pages 1–6. IEEE, 2020.

[68] I. Naimi-Akbar, M. Weurlander, and L. Barman.

Teaching-learning in virtual learning environments: a matter of forced compromises away from student-centredness? *Teaching in Higher Education*, pages 1–17, 2023.

[69] OpenAI. Hello gpt-4o, 2024. Accessed: 2024-08-30.

[70] OpenAI. Whisper api, 2024. Accessed: 2024-09-12.

[71] J. Pallant. *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge, 2020.

[72] G. Papaioannou, M.-G. Volakaki, S. Kokolakis, and D. Vouyioukas. Learning spaces in higher education: a state-of-the-art review. *Trends in Higher Education*, 2(3):526–545, 2023.

[73] H. Park and D. Ahn. The promise and peril of chatgpt in higher education: Opportunities, challenges, and design implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[74] M. Peterson-Ahmad. Enhancing pre-service special educator preparation through combined use of virtual simulation and instructional coaching. *Education Sciences*, 8(1):10, 2018.

[75] A. Poole and L. J. Ball. Eye tracking in hci and usability research. In *Encyclopedia of human computer interaction*, pages 211–219. IGI global, 2006.

[76] J. Pottle. Virtual reality and the transformation of medical education. *Future healthcare journal*, 6(3):181–185, 2019.

[77] M. P. Pratama, R. Sampelolo, and H. Lura. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of education, language teaching and science*, 5(2):350–357, 2023.

[78] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & education*, 147:103778, 2020.

[79] N. A. Rappa, S. Ledger, T. Teo, K. Wai Wong, B. Power, and B. Hilliard. The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review. *Interactive Learning Environments*, 30(7):1338–1350, 2022.

[80] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[81] N. M. Razali, Y. B. Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[82] M. A. Rojas-Sánchez, P. R. Palos-Sánchez, and J. A. Folgado-Fernández. Systematic literature review and bibliometric analysis on virtual reality and education. *Education and Information Technologies*, 28(1):155–192, 2023.

[83] J. P. Royston. An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):115–124, 1982.

[84] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78, 2000.

[85] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[86] A. W. Services. Amazon polly: Text-to-speech service, 2024. Accessed: 2024-09-12.

[87] T. Seufert. Supporting coherence formation in learning from multiple representations. *Learning and instruction*, 13(2):227–237, 2003.

[88] A. Shemshack and J. M. Spector. A systematic literature review of personalized learning terms. *Smart Learning Environments*, 7(1):33, 2020.

[89] J. M. Spector. The potential of smart technologies for learning and instruction. *International Journal of Smart Technology and Learning*, 1(1):21–32, 2016.

[90] Stanford News. New class among first taught entirely in virtual reality, 2021. Accessed: 2024-08-20.

[91] P. Stark. *Towards Effective Virtual Reality Learning Environments: Assessment of Information Processing and Learning through Eye Tracking*. PhD thesis, Universität Tübingen, 2024.

[92] P. Stark, A. J. Jung, J.-U. Hahn, E. Kasneci, and R. Göllner. Using gaze transition entropy to detect classroom discourse in a virtual reality classroom. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–11, 2024.

[93] J. Sweller. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22:123–138, 2010.

[94] J. Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

[95] M. Tavakol and R. Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[96] V. Technologies. Varjo xr-3: Mixed reality headset, 2024. Accessed: 2024-09-12.

[97] H. R. Tenenbaum, N. E. Winstone, P. J. Leman, and R. E. Avery. How effective is peer interaction in facilitating learning? a meta-analysis. *Journal of Educational Psychology*, 112(7):1303, 2020.

[98] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[99] J. J. Van Merriënboer and J. Sweller. Cognitive load theory in health professional education: design principles and strategies. *Medical education*, 44(1):85–93, 2010.

[100] A. Westphal, E. Richter, R. Lazarides, and Y. Huang. More i-talk in student teachers' written reflections indicates higher stress during vr teaching. *Computers & Education*, 212:104987, 2024.

[101] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

[102] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review, 1 2024.

[103] O. Zawacki-Richter. The current state and impact of covid-19 on digital higher education in germany. *Human Behavior and Emerging Technologies*, 3(1):218–226, 2021.

[104] H. Zhang, L. Yu, M. Ji, Y. Cui, D. Liu, Y. Li, H. Liu, and Y. Wang. Investigating high school students' perceptions and presences under vr learning environment. In *Cross Reality (XR) and Immersive Learning Environments (ILEs) in Education*, pages 97–117. Routledge, 2023.

[105] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.