What if AI Could Revolutionize Literature Reviews in Virtual Reality and Mental Health?

Paulo Guedes*, Pedro Abrantes†, Joao Marcelo Teixeira‡, and Veronica Teichrieb§
Universidade Federal de Pernambuco
Email: *pog@cin.ufpe.br, †paao@cin.ufpe.br, ‡jmxnt@cin.ufpe.br, §vt@cin.ufpe.br

***Abstract*—The rapid advancement of Large Language Models (LLMs) has opened new avenues for automating and enhancing the process of systematic literature reviews. This study investigates the effectiveness of three state-of-the-art LLMs — ChatGPT4o, LLaMA, and Gemini — in conducting literature reviews on the intersection of virtual reality and mental health, focusing on works by a renowned author in the field, Mel Slater. We defined two objective and two subjective questions to evaluate the performance of these models against a ground truth dataset. Our findings reveal significant insights into the accuracy, reliability, and limitations of each model, providing a comprehensive analysis of their potential and constraints. This study aims to guide future applications of LLMs in academic research, highlighting the transformative potential of these technologies in conducting systematic**

***Keywords*—Systematic Literature Review; Large Language Models; Virtual Reality and Mental Health.**

I. Introduction

The process of conducting systematic literature reviews is a cornerstone of academic research, offering a structured approach to synthesizing existing knowledge and identifying gaps within a specific field. Traditionally, this process is time-consuming and labor-intensive, requiring meticulous effort to ensure comprehensive coverage and accurate analysis. The advent of Large Language Models (LLMs) presents a transformative potential for this crucial aspect of research, promising to streamline and enhance the efficiency and effectiveness of literature reviews.

LLMs, such as ChatGPT-4, LLaMA, and Gemini, have demonstrated remarkable capabilities in understanding and generating human-like text, making them suitable candidates for automating aspects of systematic literature reviews. These models leverage advanced natural language processing techniques to analyze vast amounts of text, extract relevant information, and even provide insightful summaries. However, the extent to which these models can accurately and reliably perform literature reviews, particularly in specialized domains, remains an open question.

This study focuses on exploring the effectiveness of three prominent LLMs in conducting literature reviews within the context of Virtual Reality (VR) and mental health, a burgeoning area of research with significant implications for both technology and healthcare. We selected a set of 30 articles authored by Mel Slater, a renowned expert in this field, to serve as the basis for our evaluation. By defining two objective and two subjective questions, we systematically compared the performance of ChatGPT-4, LLaMA, and Gemini against a ground truth dataset meticulously curated by domain experts.

Our primary objective is to assess how well these models can replicate the depth and accuracy of human-conducted reviews. We aim to identify the strengths and limitations of each model, providing a comprehensive analysis that highlights their potential and areas for improvement. Through this study, we seek to offer valuable insights into the practical applications of LLMs in academic research and to pave the way for future innovations in the automation of systematic literature reviews.

In the following sections, we will detail our methodology, present the results of our evaluations, discuss the implications of our findings, and conclude with reflections on the future of LLMs in enhancing the landscape of academic research.

II. Methodology

Our initial goal was to conduct a comprehensive survey of research papers related to VR and mental health published in recent years within the main track of the IEEE VR conference, the premier event in the VR field globally. However, during our initial data collection, we encountered a surprisingly small number of relevant studies. This observation led us to realize that such works are more frequently published in the satellite workshops of the main event rather than in the main conference track.

Given the limited number of relevant papers in the main track, we shifted our approach. We opted to focus on a well-known author in the domain of VR and mental health, Mel Slater[1]. By selecting Mel Slater, who has established numerous collaborations over the years, we ensured a representative

[1]https://www.youtube.com/watch?v=fzS72LbJUxUhttps://www.youtube.com/watch?v=fzS

sample of authors and publication venues (both journals and conferences).

A. Mel Slater

Mel Slater (Figure 1) is a distinguished figure in the realm of VR and human-computer interaction, with a career marked by profound contributions to the scientific understanding and technological advancement of immersive environments. Holding a professorship at the University of Barcelona, he is a leading member of the Event Lab, a renowned research group specializing in the psychological and neuroscientific aspects of VR.

Fig. 1. Photo of Mel Slater during an interview[3] in November 2016, when he talked about VR's growing potential and how it was being used for social good.

Slater's academic journey began with a Ph.D. in computer science, after which he dedicated himself to exploring the interface between VR technology and human perception. His early work set the stage for foundational principles in virtual embodiment and presence, concepts that are now integral to VR research. Virtual embodiment, a core focus of his studies, examines how individuals perceive and interact with virtual bodies as though they were their own, offering significant insights into body image, social interaction, and therapeutic interventions.

Central to Slater's research is the exploration of presence, the sensation of being physically present in a virtual environment, and plausibility, the degree to which virtual scenarios are perceived as real. These theoretical frameworks have guided numerous experiments and applications, providing a robust understanding of how VR can influence human cognition, emotion, and behavior. His investigations into virtual body ownership, for example, have demonstrated how altering the appearance of a virtual body can impact users' attitudes and behaviors, revealing profound implications for psychological and rehabilitative therapies.

In addition to his theoretical contributions, Slater has been instrumental in applying VR to address social and psychological issues. His work on using VR to foster empathy and understanding through perspective-taking exercises, such as simulating experiences of racial discrimination or physical disabilities, has shown significant promise in reducing biases and improving social behaviors. Similarly, his development of VR applications for mental health treatment, including therapies for PTSD, anxiety disorders, and eating disorders, highlights the practical benefits of his research.

Mel Slater's prolific output includes over 300 scientific papers and articles, many of which are highly cited across multiple disciplines. His research has been recognized with numerous awards and honors, underscoring his impact on both academic knowledge and practical applications. He is a frequent speaker at international conferences and serves on the editorial boards of several leading journals in VR and psychology.

Slater's work has not only advanced the scientific understanding of VR but also paved the way for its application in diverse fields such as medicine, education, and social sciences. By integrating advanced VR technologies with psychological research, he has established himself as a preeminent scholar and innovator, continually pushing the boundaries of how virtual environments can enhance human experience and well-being. His contributions continue to inspire new studies and innovations, ensuring his lasting legacy in the field of virtual reality.

B. Paper selection and processing

We compiled a list of the 30 most recent papers authored by him (Table I), ensuring they were in English and pertinent to VR and mental health. Figure 2 illustrates the distribution of those papers per year.

With our corpus of 30 papers selected, the next step was to formulate a set of questions to guide the information extraction process. We defined four questions—two objective and two subjective. The objective questions were:

1) How many people took part in the experiment described in the article?
2) What are the ages of the participants mentioned in the study?

The subjective questions were:

TABLE I
LIST OF THE 30 MOST RECENT PAPERS FROM MEL SLATER REGARDING VIRTUAL REALITY AND MENTAL HEALTH.

Year	Title	Reference
2019	Body ownership increases the interference between observed and executed movements	[1]
2019	Decreasing pain ratings in chronic arm pain through changing a virtual body: different strategies for different pain types	[2]
2019	An experimental study of a virtual reality counselling paradigm using embodied self-dialogue	[3]
2019	Effect of Observing a Virtual Double on Paranoia in Social Virtual Environments: Experiment Preliminary Presentation	[4]
2019	It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory	[5]
2019	Automated psychological therapy using virtual reality (VR) for patients with persecutory delusions: study protocol for a single-blind parallel-group randomised controlled trial	[6]
2020	Manipulating the perceived shape and color of a virtual limb can modulate pain responses	[7]
2020	Which body would you like to have? The impact of embodied perspective on body perception and body evaluation in immersive virtual reality	[8]
2020	An embodied perspective as a victim of sexual harassment in virtual reality reduces action conformity in a later milgram obedience scenario	[9]
2020	”First-person virtual embodiment modulates the cortical network that encodes the bodily self and its surrounding space during the experience of domestic violence”	[10]
2020	Being the victim of intimate partner violence in virtual reality: first-versus third-person perspective	[11]
2020	Virtual body ownership and its consequences for implicit racial bias are dependent on social context	[12]
2021	Being the victim of virtual abuse changes default mode network responses to emotional expressions	[13]
2021	A Virtual Reality tool using embodiment and body swapping techniques for the treatment of obesity: A pilot usability study	[14]
2021	Bystander affiliation influences intervention behavior: A virtual reality study	[15]
2021	The influence of embodiment as a cartoon character on public speaking anxiety	[16]
2021	The golden rule as a paradigm for fostering prosocial behavior with virtual reality	[17]
2021	Self-observation of a virtual body-double engaged in social interaction reduces persecutory thoughts	[18]
2022	Encouraging bystander helping behaviour in a violent incident: a virtual reality study using reinforcement learning	[19]
2022	Impact of virtual embodiment and exercises on functional ability and range of motion in orthopedic rehabilitation	[20]
2022	Clinical efficacy of a virtual reality tool for the treatment of obesity: study protocol of a randomised controlled trial	[21]
2022	A separate reality: An update on place illusion and plausibility in virtual reality	[22]
2023	Haptic feedback in a virtual crowd scenario improves the emotional	[23]
2023	”Domestic violence from a child perspective: impact of an immersive virtual reality experience on men with a history of intimate partner violent behavior”	[24]
2023	Virtual self-conversation using motivational interviewing techniques to promote healthy eating and physical activity: A usability study	[25]
2023	Imperceptible body transformation in virtual reality: Saliency of self representation	[26]
2023	”Automated virtual reality cognitive therapy versus virtual reality mental relaxation therapy for the treatment of persistent persecutory delusions in patients with psychosis . . .”	[27]
2024	Multisensory experiences of affective touch in virtual reality enhance engagement, body ownership, perceived pleasantness, and arousal modulation	[28]
2024	Assessing the Clinical Efficacy of a Virtual Reality Tool for the Treatment of Obesity: Randomized Controlled Trial	[29]
2024	Virtual reality for mental health and in therehabilitation of violent behaviours	[30]

3) How effective was the result of the solution presented to the problem in question?
4) What was the main problem tackled by using VR?

These questions were designed to evaluate the LLMs' ability to extract both factual data (objective questions) and interpretative insights (subjective questions) from the research papers.

We selected three state-of-the-art LLMs for this study: ChatGPT-4, Gemini-1.5-Flash, and Llama-3-70b-Groq. Each model was tasked with answering the four predefined questions based on the content of the selected papers. The responses generated by the models were then compared against a ground truth dataset, which was meticulously curated by domain experts to ensure accuracy and reliability.

To assess the performance of each LLM, we established criteria for accuracy and reliability. Accuracy was measured by how well the LLM's responses matched the ground truth, relevance by the extent to which the responses were pertinent to the questions asked, and comprehensiveness by the degree to which the LLM covered all relevant aspects of the questions.

After collecting the responses from the three LLMs, we conducted a detailed analysis to identify which model provided the most accurate and relevant answers. We also examined the limitations of each model, focusing on areas where they struggled to provide correct or comprehensive answers. The final part of our methodology involved discussing the principal findings derived from the LLMs' responses to the four questions. This discussion aimed to highlight the strengths and weaknesses of each model, providing insights into their

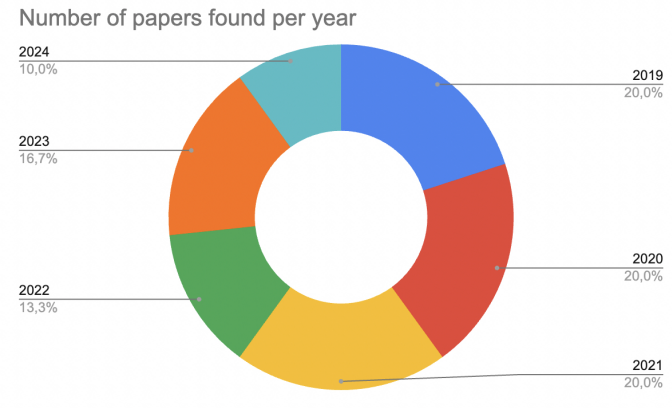

Fig. 2. Distribution of Mel Slater's work per year.

potential for automating systematic literature reviews.

III. Evaluation Metrics

The evaluation of the three selected LLMs—ChatGPT-4, Gemini-1.5-Flash, and Llama-3-70b-Groq—was carried out using a set of carefully defined metrics that aimed to measure the accuracy, relevance, and comprehensiveness of the responses provided by each model. These metrics were essential in determining how effectively each model could handle the tasks of extracting factual data and generating insightful summaries.

Accuracy was the first metric we focused on. This was defined as the degree to which the LLM's responses matched the ground truth, which was meticulously curated by domain experts from the selected papers. For the objective questions, accuracy was measured by comparing the exact numerical data provided by the models against the actual data from the papers. For instance, the question "How many people took part in the experiment described in the article?" required the models to identify and report participant numbers accurately. Similarly, the question "What are the ages of the participants mentioned in the study?" demanded precise extraction of age ranges and specific details.

The second metric, relevance, evaluated the extent to which the responses were pertinent to the questions asked. This metric was particularly crucial for the subjective questions, where the models needed to interpret and synthesize information rather than merely extracting it. For the question "How effective was the result of the solution presented to the problem in question?" relevance was measured by assessing whether the models' summaries accurately reflected the key outcomes and effectiveness measures discussed in the papers. Similarly, for the question "What was the main problem tackled by using VR?" relevance was gauged by how well the models identified and articulated the core issues addressed by the studies.

Comprehensiveness, the third metric, measured the degree to which the LLM covered all relevant aspects of the questions. This involved evaluating the breadth and depth of the responses, ensuring that the models did not omit critical details or oversimplify complex information. For the objective questions, comprehensiveness included providing complete data ranges and relevant context when applicable. For the subjective questions, it involved generating detailed and contextually rich summaries that captured the full scope of the studies' findings and objectives.

Each response from the LLMs was evaluated against these metrics by a panel of experts who provided scores based on predefined criteria. The scores for accuracy, relevance, and comprehensiveness were then averaged to provide an overall performance score for each model on each question. This systematic approach ensured a fair and thorough assessment of the LLMs' capabilities.

IV. Results

A. Performance of ChatGPT4o

ChatGPT4o demonstrated exceptional performance in our evaluation, achieving perfect scores across both objective and subjective questions. The model's ability to accurately and reliably extract information and generate insightful summaries highlights its potential as a powerful tool for conducting systematic literature reviews. Based on the evaluation, ChatGPT4o scored 100% in both objective and subjective questions, showcasing high precision in extracting factual data such as participant numbers and age ranges, and interpreting complex information. This underscores its reliability and potential for automating systematic literature reviews, making it an invaluable asset in academic research.

B. Performance of Gemini

Gemini performed admirably, nearly achieving perfect scores in both objective and subjective questions. For objective questions, Gemini scored 58 out of 60, demonstrating a high level of accuracy in extracting factual data. The model consistently identified participant numbers and age ranges accurately. In subjective questions, Gemini scored 59 out of 60, reflecting its strong ability to interpret and synthesize complex information. It generated summaries that closely matched key outcomes and measures of effectiveness, and effectively identified the main problems addressed by VR. Overall, Gemini's high scores in both accuracy and relevance underscore its potential as a valuable tool for academic research, enhancing the efficiency and effectiveness of systematic literature reviews.

In the work of [18], Gemini failed to answer the objective question related to the ages of the participants accurately. The ground truth was an age range of 18 to 33 with an average of 23.6 years. Gemini responded, *"The mean age of the participants in the Random group was 22.5 ± 4.50 and the mean age of the participants in the Targeted group was 24.7 ± 5.43. This information is provided in the 'Results' section."* While the model emphasized where the information was obtained from, its failure stemmed from the fact that this information related only to the control group, conducted in the first section of the experiment, and not to all participants. This led to erroneous information as the model could not verify that there was more than one set of information about the participants' ages in the project.

In another article [22], Gemini also failed to detect objective information accurately. This work discusses various experiments and more than one group of participants, and not all of them are explicitly mentioned, leading to confusion. Gemini's response was: *"The article mentions a pilot study involving a Dire Straits concert VR scenario. 20 participants were involved in this pilot study."* While Gemini didn't fail to provide this information, the response was somewhat flawed because it didn't indicate that there was more than one study or mention that this number of participants was related to only one of the experiments, implying that there could be more. In contrast, ChatGPT's response emphasized this fact: *"The specific number of participants involved in the experiments described in the article is not explicitly mentioned in the excerpts provided. However, multiple experiments and studies are referenced, implying that a significant number of participants were involved across various studies."*. Leading us to highlight that Gemini did not observe that there was more than one study in the article. Even in the subjective questions, Gemini focused only on one of the studies and did not emphasize that there could be another study.

C. Performance of LLaMA

LLaMA showed solid performance, with particularly strong results in subjective questions. For objective questions, LLaMA scored 49 out of 60, generally identifying participant numbers and age ranges correctly but occasionally missing some details. This indicates that while LLaMA is capable of handling numerical and factual information, it may require further refinement to achieve higher precision. In subjective questions, LLaMA scored 59 out of 60, demonstrating a strong ability to interpret and synthesize complex information. It produced summaries that accurately reflected key outcomes and effectively identified core issues addressed by VR. Overall, LLaMA's performance shows it is a valuable tool for academic research, particularly in synthesizing and summarizing complex information. However, improvements in its accuracy for factual data extraction could further enhance its reliability and effectiveness in conducting systematic literature reviews.

TABLE II
NUMBER OF CORRECT GUESSES (OUT OF 60) FOR EACH OF THE ASSESSED MODELS (CHATGPT4O, GEMINI AND LLAMA.

Question type	ChatGPT4o	Gemini	Llama
Objective	60 (100%)	58 (96.6%)	49 (81.67%)
Subjective	60 (100%)	59 (98.3%)	59 (98.3%)

LLaMA left much to be desired when it came to answering objective questions. Although the model failed both objective questions in only one article, specifically [23], where it could not explicitly state either the number of participants or their ages, the model showed great difficulty in finding the average or range of participants' ages in other articles. In the articles [6], [12], [14], [15], [18], [19], [22], [24], [28], [29], the model managed to find the number of participants in the experiments but always indicated that the age information was not explicitly mentioned in the article, consistently responding: *"The article does not explicitly mention the ages of the participants."* This led us to conclude that LLaMA's performance in more objective questions, which require a greater understanding of the presented text, resulted in the model hallucinating and providing false information, even when the information was present in the text. It is worth noting that LLaMA had the same problem as Gemini in the articles [18], [22], as mentioned in the previous section.

Regarding the subjective questions, the model performed very well, achieving an accuracy of 98.3%. The only failure occurred in the study [17], where the model did not find a solution presented for the article's problem, responding: *"The article does not present a specific solution or experiment with results, so it is not possible to assess the effectiveness of a solution."* Upon analyzing its responses, we concluded that a possible cause of this error was the fact that the article contained more than one study, which may have led the model to become confused in its response, even though the authors reached the same conclusion. Consequently, we evaluate that when the model is asked to perform a complex analysis, its response is often conditioned to inform that the information does not exist.

D. Comparative analysis

Overall, according to Table II, the comparative analysis reveals that ChatGPT-4 leads in both accuracy and interpretation, making it the most reliable tool among the three for conducting systematic literature reviews. Gemini closely follows with strong performance in both extracting factual data

and interpreting information, while LLaMA, despite needing improvements in accuracy, shows significant promise in synthesizing and summarizing complex information. This analysis provides valuable insights into the capabilities and limitations of each model, guiding future applications and developments in the use of large language models for systematic literature reviews.

At the time of writing this article, ChatGPT-4o has a monthly cost of USD $20/month, while Gemini and Llama were tested in their free versions. We used the paid version of ChatGPT-4o, which does not have a limit on messages per period, providing greater flexibility. In contrast, the Gemini and Llama models were used through the Poe platform[4], where there is a limitation of 40 messages per day. This limitation may vary depending on the platform used.

In addition to their direct functionalities, all models offer the possibility of developing tools and applications via APIs. The cost per request for ChatGPT-4o, using the OpenAI API, can vary based on the number of tokens processed. Typically, OpenAI charges for the use of its models, while Google and Meta offer free versions. Google provides Gemini-1.5-Flash with a monthly limit of 1,048,576 tokens/words, and Facebook allows Llama-3 to be downloaded and used locally on your computer at no additional cost, facilitating the development and implementation of customized solutions.

V. Discussion

A. Key findings from the objective questions

The three LLMs were able to accurately describe specific points in the text and find details located within a single line throughout the article. However, LLaMA showed slight limitations in this aspect and did not perform as well.

Among the 30 articles analyzed, three were systematic reviews [17] [22] [30] and one was a preliminary presentation of an experiment [4], hence they did not involve any participants. For the remaining 24 articles that reported on experimental studies, the average number of participants was 43.80, with a standard deviation of 25.57. The experiment with the smallest sample size included only six participants [14], whereas the experiment with the largest sample size involved 96 participants [21].

Regarding the second objective question, which focused on the age of participants, we were able to extract data on the minimum and maximum ages of participants, as well as their average age. Unfortunately, these details were not consistently available across all evaluated articles. Specifically, the average age was reported in 15 (50%) of the selected articles, the minimum age was mentioned in 23 (76.66%), and the maximum age was provided in 19 (63.33%) of the articles.

Some studies only specified the minimum age of participants (e.g., 16 or 18 years old) [6]. Indeed, 19 of the 23 papers with minimum age information used 18 as the minimum age for the volunteers. Among the 15 articles that included the average age, the mean age of participants was 28.98 years, with a standard deviation of 9.89. For the 23 articles that reported the minimum age, the mean minimum age was 19 years, with a standard deviation of 4.62. The youngest minimum age reported was 16 years, found in a single study [6]. This suggests that Mel Slater's research typically focuses on an adolescent and adult population, with no representation of children. The highest maximum age reported was 73 years, also found in a single study [27]. Figure 3 illustrates the distribution of the minimum, maximum, and average ages of participants in the analyzed experiments.

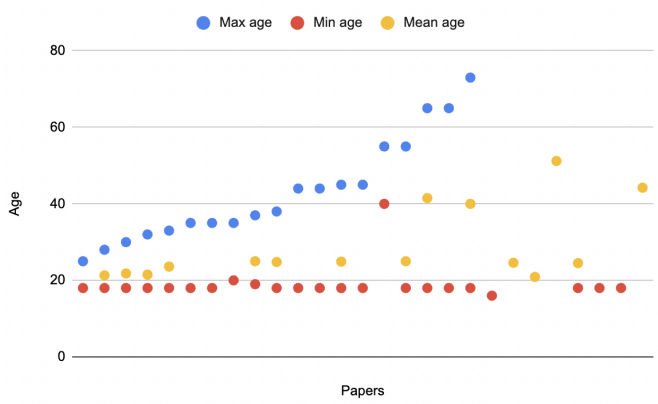

Fig. 3. Distribution of min, max and mean values for the volunteer numbers in the experiments analyzed.

B. Key findings from the subjective questions

Similar to the findings from the Objective Questions, we can see that the current LLM models are very good at summarizing the subjective aspects of articles or even other documents. This demonstrates their strength in interpreting and synthesizing the information presented throughout the article and responding with quality.

Based on the answers provided for all 30 papers evaluated, we can categorize the main findings into six main themes according to the effectiveness of the VR solutions and also to the main problem tackled:

A) Therapeutic Effectiveness and Behavioral Changes:
- Body swapping vs. virtual Freud [3].

[4]https://poe.com/

- Measurement of the effectiveness of VR in reducing the degree of conviction in persecutory delusions [6].
- Reduction in harmful behaviors with VR embodiment [9].
- Evaluation of participant satisfaction and acceptance of a VR tool compared to a control group [14].
- Effectiveness of VR solutions in enhancing readiness to change lifestyle habits [21].
- Successful applications of VR methods like sentiment analysis and configuration transitions in measuring engagement and plausibility [23].
- VR cognitive therapy vs. VR mental relaxation [27].
- ConVRself platform for lifestyle changes [29].

B) Emotional and Cognitive Impact:
- Neural modulation in domestic violence scenarios [10].
- Comparison of participant reactions and physiological responses in VR scenarios depicting domestic violence [11].
- Brain activity modulation with emotional stimuli [13].
- Influence of haptic feedback on emotional responses and interaction in VR scenarios [23].
- Emotional recognition improvement in IPV survivors [24].
- Change blindness and the immersive experience in VR [26].
- Visuo-tactile feedback for VR interactions [28].

C) Physical Rehabilitation and Functional Improvement:
- The study demonstrated how body ownership affects motor performance in VR scenarios [1].
- VR for functional recovery and joint movements [20].
- ConVRself for readiness to change and lifestyle habits [25].

D) Social Behavior and Interaction:
- Social identity influence on intervention likelihood [15].
- Empathy enhancement through Golden Rule Embodiment Paradigm [17].
- Impact of VR conditions on helping interventions [19].
- Influence of social identity and bystander intervention in VR scenarios [30].

E) Pain Management and Body Perception:
- Effectiveness of VR conditions on pain ratings [2].
- Improved body satisfaction through VR perspectives [8].

F) Anxiety, Stress Response, and Bias:
- Reduction in anxiety and paranoia levels post-VR exposure [4].
- VR-induced stress responses without affecting memory performance [5].
- Induction of stress responses in VR scenarios without impacting working memory performance [7].
- VR's impact on racial bias depending on social context [12].
- Impact of embodiment as a cartoon character on anxiety reduction [16].
- Reduction in persecutory thoughts with virtual body-double [18].

Studies in the "Therapeutic Effectiveness and Behavioral Changes" category evaluate the effectiveness of VR interventions in inducing behavioral changes, improving therapeutic outcomes, and promoting positive behavior modifications. This includes applications aimed at psychological treatments, behavior modification, and readiness for lifestyle changes.

The "Emotional and Cognitive Impact" category encompasses research that explores the emotional responses, cognitive impacts, and perceptual changes induced by VR experiences. Studies here focus on how VR affects emotional processing, cognitive functioning, and the immersion levels of participants.

Papers in the "Physical Rehabilitation and Functional Improvement" category investigate the use of VR for physical rehabilitation, enhancing functional abilities, and improving motor skills. Research here often examines VR applications in rehabilitation therapies, recovery from physical injuries, and mobility enhancement.

Research in the "Social Behavior and Interaction" category explores how VR influences social behaviors, interactions, and prosocial behaviors. This includes studies on social identity within VR environments, bystander interventions, and empathy enhancement through VR experiences.

The "Pain Management and Body Perception" category focuses on studies that assess the effectiveness of VR in managing pain, altering body perception, and improving body satisfaction. Research here examines VR applications in pain therapy, body image enhancement, and perception alterations through immersive experiences.

Studies in the "Anxiety, Stress Response, and Bias" category investigate the impact of VR on anxiety levels, stress responses, and biases. This includes research on stress induction in VR scenarios, anxiety reduction techniques using VR, and VR's influence on implicit biases in various social contexts.

C. Potential improvements and future directions

This study presents a significant step forward in utilizing LLMs to automate systematic literature reviews, particularly in the domain of virtual reality and mental health. However, there are several potential improvements and future directions to consider for enhancing this work.

One of the primary areas for improvement is the refinement of LLMs to increase their accuracy in extracting factual data. While the models demonstrated strong capabilities in handling objective questions, occasional errors and omissions highlight the need for more precise algorithms. Future research could focus on developing models with improved contextual understanding to accurately interpret and extract numerical and specific details from complex texts.

Additionally, expanding the dataset used for training these models could significantly enhance their performance. The study utilized 30 papers by Mel Slater, which provided a robust foundation, but incorporating a more extensive and diverse range of articles could improve the models' generalizability and robustness. This would allow the LLMs to handle a wider variety of research papers and topics, further proving their utility in systematic reviews.

Integrating advanced natural language processing techniques, such as improved entity recognition and disambiguation methods, could also help in better identifying and linking relevant information across different sections of research papers. This enhancement would address issues where models struggle to provide comprehensive answers due to fragmented information spread throughout the text.

Furthermore, the study's evaluation metrics could be expanded to include additional dimensions such as temporal consistency and adaptability to evolving research trends. Incorporating real-time updates and continuous learning mechanisms would enable the models to stay current with the latest research developments, thereby increasing their relevance and applicability.

Another promising direction is the development of hybrid models that combine the strengths of different LLMs. For instance, leveraging the interpretative strengths of models like ChatGPT-4 with the factual accuracy of Gemini could result in a more balanced and effective tool for literature reviews. Collaborative filtering and ensemble learning techniques could be explored to merge outputs from multiple models, enhancing overall performance.

Lastly, user interface and experience improvements are crucial for practical applications. Developing user-friendly platforms that allow researchers to interact with and fine-tune the LLMs' outputs could bridge the gap between automated reviews and human oversight. These platforms could include features for feedback loops, where researchers can provide corrections and adjustments, enabling the models to learn from their mistakes and improve over time.

VI. Conclusion

This study demonstrates the potential of LLMs in automating and enhancing systematic literature reviews, particularly within the niche but burgeoning field of virtual reality and mental health. The evaluation of three state-of-the-art models—ChatGPT-4, Gemini, and LLaMA—revealed significant insights into their capabilities and limitations. ChatGPT-4 emerged as the most reliable tool, excelling in both objective and subjective questions, thus proving its high precision in extracting factual data and interpreting complex information. Gemini also performed admirably, showcasing strong accuracy and relevance, albeit with occasional lapses in handling fragmented data. LLaMA, while showing promise in synthesizing complex information, displayed a need for refinement in accuracy.

The findings underscore the immense potential of LLMs in streamlining the literature review process, making it more efficient and comprehensive. These models can significantly reduce the time and effort required for systematic reviews, allowing researchers to focus on higher-level analytical tasks. However, the study also highlights areas for improvement, particularly in enhancing the models' accuracy and contextual understanding, expanding the training datasets, and integrating advanced natural language processing techniques.

Looking ahead, the future of LLMs in systematic literature reviews appears bright. Continued advancements in AI and machine learning will likely address current limitations, further enhancing the accuracy, reliability, and utility of these models. The integration of real-time updates and continuous learning mechanisms will enable LLMs to stay current with the latest research developments, increasing their relevance. Additionally, developing hybrid models that combine the strengths of different LLMs could offer a more balanced and effective tool for literature reviews. User-friendly platforms that facilitate interaction between researchers and models will be crucial in bridging the gap between automated processes and human oversight.

The application of LLMs in systematic literature reviews marks a significant step forward in academic research. By automating the labor-intensive aspects of literature reviews, these models hold the promise of not only enhancing efficiency but also improving the depth and breadth of reviews. As the technology evolves, LLMs will become indispensable tools in the research toolkit, paving the way for more informed and timely scientific discoveries. The journey towards fully realizing the potential of LLMs in this domain is ongoing, but the initial results are promising and point towards a future where AI-driven literature reviews are the norm, driving advancements across various fields of study.

Acknowledgements

We extend our deepest gratitude to Professor Mel Slater for his groundbreaking contributions to the fields of virtual reality

and human-computer interaction. His pioneering research on virtual embodiment, presence, and the therapeutic applications of VR has been instrumental in shaping our understanding and approach to leveraging artificial intelligence in the development of solutions for mental health within virtual environments. Professor Slater's extensive body of work not only provides a robust theoretical foundation but also inspires ongoing innovation and exploration in these interdisciplinary fields. His commitment to advancing knowledge and practical applications has significantly influenced this study, and his insights have been invaluable in guiding our research endeavors.

References

[1] D. Burin, K. Kilteni, M. Rabuffetti, M. Slater, and L. Pia, "Body ownership increases the interference between observed and executed movements," *PloS one*, vol. 14, no. 1, p. e0209899, 2019.

[2] M. Matamala-Gomez, A. M. D. Gonzalez, M. Slater, and M. V. Sanchez-Vives, "Decreasing pain ratings in chronic arm pain through changing a virtual body: different strategies for different pain types," *The Journal of Pain*, vol. 20, no. 6, pp. 685–697, 2019.

[3] M. Slater, S. Neyret, T. Johnston, G. Iruretagoyena, M. Á. d. l. C. Crespo, M. Alabèrnia-Segura, B. Spanlang, and G. Feixas, "An experimental study of a virtual reality counselling paradigm using embodied self-dialogue," *Scientific reports*, vol. 9, no. 1, p. 10903, 2019.

[4] G. Gorisse and M. Slater, "Effect of observing a virtual double on paranoia in social virtual environments: Experiment preliminary presentation," in *ACM Symposium on Applied Perception 2019*, 2019.

[5] M. A. Martens, A. Antley, D. Freeman, M. Slater, P. J. Harrison, and E. M. Tunbridge, "It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory," *Journal of Psychopharmacology*, vol. 33, no. 10, pp. 1264–1273, 2019.

[6] D. Freeman, R. Lister, F. Waite, L.-M. Yu, M. Slater, G. Dunn, and D. Clark, "Automated psychological therapy using virtual reality (vr) for patients with persecutory delusions: study protocol for a single-blind parallel-group randomised controlled trial (thrive)," *Trials*, vol. 20, pp. 1–8, 2019.

[7] M. Matamala-Gomez, B. Nierula, T. Donegan, M. Slater, and M. V. Sanchez-Vives, "Manipulating the perceived shape and color of a virtual limb can modulate pain responses," *Journal of clinical medicine*, vol. 9, no. 2, p. 291, 2020.

[8] S. Neyret, A. I. Bellido Rivas, X. Navarro, and M. Slater, "Which body would you like to have? the impact of embodied perspective on body perception and body evaluation in immersive virtual reality," *Frontiers in Robotics and AI*, vol. 7, p. 492886, 2020.

[9] S. Neyret, X. Navarro, A. Beacco, R. Oliva, P. Bourdin, J. Valenzuela, I. Barberia, and M. Slater, "An embodied perspective as a victim of sexual harassment in virtual reality reduces action conformity in a later milgram obedience scenario," *Scientific reports*, vol. 10, no. 1, p. 6207, 2020.

[10] A. W. de Borst, M. V. Sanchez-Vives, M. Slater, and B. de Gelder, "First-person virtual embodiment modulates the cortical network that encodes the bodily self and its surrounding space during the experience of domestic violence," *Eneuro*, vol. 7, no. 3, 2020.

[11] C. Gonzalez-Liencres, L. E. Zapata, G. Iruretagoyena, S. Seinfeld, L. Perez-Mendez, J. Arroyo-Palacios, D. Borland, M. Slater, and M. V. Sanchez-Vives, "Being the victim of intimate partner violence in virtual reality: first-versus third-person perspective," *Frontiers in psychology*, vol. 11, p. 507601, 2020.

[12] D. Banakou, A. Beacco, S. Neyret, M. Blasco-Oliver, S. Seinfeld, and M. Slater, "Virtual body ownership and its consequences for implicit racial bias are dependent on social context," *Royal Society open science*, vol. 7, no. 12, p. 201848, 2020.

[13] S. Seinfeld, M. Zhan, M. Poyo-Solanas, G. Barsuola, M. Vaessen, M. Slater, M. V. Sanchez-Vives, and B. de Gelder, "Being the victim of virtual abuse changes default mode network responses to emotional expressions," *cortex*, vol. 135, pp. 268–284, 2021.

[14] D. Anastasiadou, B. Spanlang, M. Slater, J. V.-D. SEBASTIAN, J. A. Ramos-Quiroga, G. P. Puig, A. Ciudin, M. Comas, and P. Lusilla-Palacios, "A virtual reality tool using embodiment and body swapping techniques for the treatment of obesity: A pilot usability study," *ANNUAL REVIEW OF CYBERTHERAPY AND TELEMEDICINE 2021*, vol. 105, 2021.

[15] A. Rovira, R. Southern, D. Swapp, C. Campbell, J. J. Zhang, M. Levine, and M. Slater, "Bystander affiliation influences intervention behavior: A virtual reality study," *Sage Open*, vol. 11, no. 3, p. 21582440211040076, 2021.

[16] A. I. Bellido Rivas, X. Navarro, D. Banakou, R. Oliva, V. Orvalho, and M. Slater, "The influence of embodiment as a cartoon character on public speaking anxiety," *Frontiers in Virtual Reality*, vol. 2, p. 695673, 2021.

[17] M. Slater and D. Banakou, "The golden rule as a paradigm for fostering prosocial behavior with virtual reality," *Current Directions in Psychological Science*, vol. 30, no. 6, pp. 503–509, 2021.

[18] G. Gorisse, G. Senel, D. Banakou, A. Beacco, R. Oliva, D. Freeman, and M. Slater, "Self-observation of a virtual body-double engaged in social interaction reduces persecutory thoughts," *Scientific reports*, vol. 11, no. 1, p. 23923, 2021.

[19] A. Rovira and M. Slater, "Encouraging bystander helping behaviour in a violent incident: a virtual reality study using reinforcement learning," *Scientific reports*, vol. 12, no. 1, p. 3843, 2022.

[20] M. Matamala-Gomez, M. Slater, and M. V. Sanchez-Vives, "Impact of virtual embodiment and exercises on functional ability and range of motion in orthopedic rehabilitation," *Scientific reports*, vol. 12, no. 1, p. 5046, 2022.

[21] D. Anastasiadou, M. Slater, B. Spanlang, D. C. Porras, M. Comas, A. Ciudin, G. P. Puig, J. Vázquez-De Sebastián, J. A. Ramos-Quiroga, and P. Lusilla-Palacios, "Clinical efficacy of a virtual reality tool for the treatment of obesity: study protocol of a randomised controlled trial," *BMJ open*, vol. 12, no. 6, p. e060822, 2022.

[22] M. Slater, D. Banakou, A. Beacco, J. Gallego, F. Macia-Varela, and R. Oliva, "A separate reality: An update on place illusion and plausibility in virtual reality," *Frontiers in virtual reality*, vol. 3, p. 914392, 2022.

[23] R. Venkatesan, D. Banakou, and M. Slater, "Haptic feedback in a virtual crowd scenario improves the emotional response," *Frontiers in Virtual Reality*, vol. 4, p. 1242587, 2023.

[24] S. Seinfeld, R. Hortensius, J. Arroyo-Palacios, G. Iruretagoyena, L. E. Zapata, B. de Gelder, M. Slater, and M. V. Sanchez-Vives, "Domestic violence from a child perspective: impact of an immersive virtual reality experience on men with a history of intimate partner violent behavior," *Journal of interpersonal violence*, vol. 38, no. 3-4, pp. 2654–2682, 2023.

[25] D. Anastasiadou, P. Herrero, J. Vázquez-De Sebastián, P. Garcia-Royo, B. Spanlang, E. Álvarez de la Campa, M. Slater, A. Ciudin, M. Comas, J. A. Ramos-Quiroga *et al.*, "Virtual self-conversation using motivational interviewing techniques to promote healthy eating and physical activity: A usability study," *Frontiers in psychiatry*, vol. 14, p. 999656, 2023.

[26] G. Senel, F. Macia-Varela, J. Gallego, H. P. Jensen, K. Hornbæk, and M. Slater, "Imperceptible body transformation in virtual reality: Saliency of self representation," *Iscience*, vol. 26, no. 10, 2023.

[27] D. Freeman, R. Lister, F. Waite, U. Galal, L.-M. Yu, S. Lambe, A. Beckley, E. Bold, L. Jenner, R. Diamond *et al.*, "Automated virtual reality cognitive therapy versus virtual reality mental relaxation therapy for the treatment of persistent persecutory delusions in patients with psychosis (thrive): a parallel-group, single-blind, randomised controlled trial in england with mediation analyses," *The Lancet Psychiatry*, vol. 10, no. 11, pp. 836–847, 2023.

[28] W. Sun, D. Banakou, J. Świdrak, I. Valori, M. Slater, and M. Fairhurst, "Multisensory experiences of affective touch in virtual reality enhance engagement, body ownership, perceived pleasantness, and arousal modulation," 2024.

[29] D. Anastasiadou, P. Herrero, P. Garcia-Royo, J. Vázquez-De Sebastián, M. Slater, B. Spanlang, E. Álvarez de la Campa, A. Ciudin, M. Comas, J. A. Ramos-Quiroga *et al.*, "Assessing the clinical efficacy of a virtual reality tool for the treatment of obesity: Randomized controlled trial," *Journal of Medical Internet Research*, vol. 26, p. e51558, 2024.

[30] N. Barnes, M. TORAO-ANGOSTO, M. Slater, and M. V. SANCHEZ-VIVES, "Virtual reality for mental health and in therehabilitation of violent behaviours," *Fonseca, Journal of Communication*, no. 28, pp. 10–46, 2024.