International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Supporting Relational Knowledge Construction Through LLM-Assisted Concept Mapping in Virtual Reality

Fangyuan Cheng, Bingyi Su & Karen B. Chen

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2719119

Supporting Relational Knowledge Construction Through LLM-Assisted Concept Mapping in Virtual Reality

Fangyuan Cheng, Bingyi Su and Karen B. Chen

Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA

ABSTRACT

Virtual reality (VR) is increasingly used to enhance learning, yet its effectiveness in supporting conceptual understanding, particularly relational knowledge construction, remains underexplored. Relational knowledge construction involves identifying meaningful connections among concepts to form coherent representations that support knowledge organization, guide interpretation, and facilitate learning transfer. Based on a review of the challenges in relational knowledge construction, we proposed conceptual, relational, and metacognitive scaffolds, upon which we developed CmapVR, an immersive learning environment with LLM-supported features to implement the proposed scaffolds. A within-subject study (N = 16) compared CmapVR with three alternative learning conditions in a science concept learning task. Results showed that CmapVR supported more extensive and higher-quality relational knowledge construction than the other conditions, resulted in lower extraneous cognitive load, and was preferred by participants over browsing in VR. We discussed factors that may have contributed to these outcomes and provided insights for integrating LLMs into VR-based learning environments.

KEYWORDS

Virtual reality; knowledge construction; large language models; concept map

1. Introduction

Virtual reality has shown effectiveness in enhancing learning, owing to its unique learning affordances, including immersive three-dimensional visualizations, interactive experiences, and embodiment (Kilteni et al., 2012; McGowin et al., 2021). These affordances make VR particularly effective for visualizing learning objects that are otherwise difficult to perceive, such as scale and sizes (Cheng et al., 2024) and microscopic phenomena (Ferretti et al., 2021), or for providing embodied learning experiences that promote understanding and engagement, such as virtual field trips and simulated laboratories (McGowin et al., 2021). A growing body of research has also begun to examine how VR can support conceptual understanding. Conceptual understanding refers to the active construction of meaning from information. It often involves higher-order cognitive processes, including reasoning, inference generation, and knowledge integration. For example, Lisle et al. (2021) demonstrated that VR can support sensemaking, where individuals actively construct coherent interpretations from distributed information.

In line with this direction, this study focused on relational knowledge construction as a critical component of conceptual understanding. Relational knowledge construction refers to the process by which learners identify, articulate, and organize meaningful connections among concepts, transforming isolated facts into coherent knowledge structures (Halford et al., 2010; Jonassen et al., 1993). This process is particularly salient in science learning, where understanding how concepts are causally, hierarchically, and functionally interconnected is essential for explaining phenomena, drawing inferences, and transferring knowledge to novel problems (National Research Council, 2007). As such relational connections accumulate and become organized over time, they form stable, long-term mental representations known as schemas (Anderson, 1994; Rumelhart, 1980). The richness and coherence of individuals' schemas

CONTACT Karen B. Chen kbchen2@ncsu.edu Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27606, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2719119.

can shape their ability to perform higher-order cognitive tasks (Anderson, 2018; D. Liu et al., 2025b; Varga et al., 2020). Both schemas and the process of their formation can be externalized through concept maps (J. D. Novak & Cañas, 2006; Urdiales-Ibarra et al., 2018), which are visual representations where concepts are depicted as nodes and their relationships as labeled links (J. D. Novak & Gowin, 1984). Concept maps thus serve as an assessment tool for knowledge construction, providing measurable indicators of the quantity and quality of relational connections learners have formed.

The affordances of VR, such as spatial representation and embodied interaction, position it as a promising medium for facilitating relational knowledge construction. First, by situating information in a three-dimensional environment, VR can leverage learners' spatial memory to help encode, organize, and retrieve conceptual relationships. Prior research indicates that spatial experience can strengthen memory traces by providing additional retrieval cues (Ekstrom & Hill, 2023; Maidenbaum et al., 2024; Mastrogiuseppe et al., 2019). Second, VR's immersive and interactive nature supports embodied cognition by enabling learners to physically navigate and manipulate conceptual representations, which prior research suggests can enhance understanding and knowledge retention (Johnson-Glenberg et al., 2016; Khorasani et al., 2023; Skulmowski & Rey, 2018).

Despite these potential benefits, conceptual understanding tasks in VR also introduce additional challenges, such as elevated cognitive load that reduced learners' ability to draw inferences (Juliano et al., 2022) and perceptual distractions that interfere with conceptual integration (Makransky et al., 2021). These findings suggest that deliberate scaffolding mechanisms and information design may be necessary to facilitate such tasks effectively. One promising approach to addressing these challenges is the integration of large language models (LLMs), which can provide targeted scaffolding to help learners access and process information. LLMs, trained on large text corpora, demonstrate strong capabilities in natural language understanding, relational reasoning, and structured information generation (Ali et al., 2024). By embedding LLM-based scaffolding within immersive environments, learners can receive context-sensitive support for clarifying concepts, identifying potential relationships, and evaluating connections (García-Méndez et al., 2025). Such support may reduce the cognitive demands of information retrieval and relational reasoning. However, limited research has examined how LLM-based scaffolding can be integrated into VR to support knowledge construction, particularly with regard to helping learners identify, comprehend, and retain conceptual relationships.

This study is guided by two aims. The first aim is to design and implement a VR-based learning environment that integrates LLM-based scaffolding to support relational knowledge construction. The second aim is to evaluate the proposed system by addressing two research questions (RQs): **RQ1.** How does the proposed system affect relational knowledge construction performance, cognitive load, usability, and user preference compared to existing learning tools? **RQ2**. How do users interact with and respond to the proposed system? To address these questions, the research team developed CmapVR, a VR-based learning environment with integrated LLM-based scaffolding and conducted a within-subject study comparing it with three existing learning tools. Findings indicated that CmapVR supported learners in constructing more extensive and higher-quality conceptual relationships than the comparison tools, while also imposing lower extraneous cognitive load. Drawing on these findings, this study highlights the factors underlying these outcomes and offers design insights for integrating LLMs into VR-based learning environments, as well as key usability challenges of such systems.

2. Related works

2.1. Virtual reality for conceptual understanding

Conceptual understanding refers to the process of constructing meaningful interpretations of information and the relationships among ideas. It typically transcends perceptual-level cognition and involves more complex mental activities, such as drawing inferences, identifying relationships, and integrating information (Alexander, 2016; Kyllonen, 2020; Radvansky & Copeland, 2004). Conceptual understanding underlies various tasks in immersive environments, such as sensemaking (Lisle et al., 2021), data analysis (Skarbez et al., 2019), and problem solving (Araiza-Alba et al., 2021).

Prior studies have demonstrated that VR can support conceptual understanding. For example, Lui et al. (2023) demonstrated that immersive simulations can support scientific reasoning by enabling learners to iteratively generate, test, and revise scientific models more effectively than in desktop-based environments. Nakić et al. (2025) found that VR-based adaptations of Bebras logic tasks, which are short problem-solving puzzles, improved students' logical and computational reasoning by spatially externalizing problem structures. Ghali et al. (2017) found that VR can foster intuitive inductive reasoning by helping learners detect relational patterns through assistance strategies. In addition, VR can support conceptual understanding by fostering cognitive or motivational factors such as spatial memory (Cimadevilla et al., 2023; Hong et al., 2026), attention allocation (C. Liu et al., 2025a), and self-efficacy (Ahn & Noh, 2025), or by eliciting affective states conducive to cognitive processing, such as curiosity and flow (Schutte, 2020).

However, other studies suggest that VR can impair conceptual understanding performance (Mayer et al., 2023), due in part to heightened distraction, elevated cognitive load, or a mismatch between perceptual and conceptual structures. For example, Juliano et al. (2022) found that VR visualizations can elevate cognitive load, reducing learners' ability to draw inferences. Makransky et al. (2021) reported that immersive VR can introduce perceptual distractions that interfere with tasks requiring conceptual integration and analytical reasoning. Furthermore, compared to traditional displays, VR can impede the processing of textual information due to limitations in field-of-view, visual stability, and display ergonomics (Kazemi & Lee, 2024; Novotny & Laidlaw, 2024; Rzayev et al., 2021; Sultana & Alam, 2024), which may in turn affect cognitive task performance. Despite these findings, no clear consensus has emerged on VR's effectiveness in supporting conceptual understanding. This suggests that VR's impact may be shaped by specific design features, such as how immersive environments structure information, support interaction, and scaffold learners' information processing, rather than by the medium alone.

2.2. Relational knowledge, schemas, and concept maps

Relational knowledge, a fundamental component of conceptual understanding, refers to how concepts, events, or phenomena are connected within a domain. Relational knowledge enables learners to construct coherent interpretations of complex information. Through cognitive processes such as comprehension and reasoning, learners construct relational knowledge by linking isolated pieces of information into interconnected knowledge structures (Kintsch, 1998). Over time, these structures stabilize into schemas, mental representations that organize and coordinate patterns of information or action (Iran-Nejad, 1987; McVee et al., 2005; Varga et al., 2020). Given this study's focus on declarative knowledge, we adopt a working definition of schemas as organized mental representations that capture how information, such as concepts, is interconnected within a domain. Schemas play a pivotal role in mediating cognitive and learning processes (Iran-Nejad, 1987). They can help learners attain deeper understanding, organize knowledge, draw inferences, and transfer knowledge across contexts by applying abstract structures (Hajian, 2019; Meylani, 2024). The concepts, relationships, and overall organization within a schema can collectively reveal how learners construct and integrate domain knowledge. Because schemas are internal mental representations, they are often studied through external representations (Ausubel, 1963). Concept maps are particularly well suited for this purpose, as they externalize schema structures through nodes and labeled links. They enable researchers to evaluate learners' understanding of a topic by examining both the quantity and quality of the relationships constructed in the maps (Markham et al., 1994). Beyond their use as an assessment tool, concept maps can also support knowledge construction by helping learners visually organize conceptual relationships (Moon et al., 2011; J. D. Novak, 1990b, 1995; Tan et al., 2017) and recognize the broader context of individual ideas (J. D. Novak & Cañas, 2006). This approach is exemplified by digital concept-mapping environments such as CmapTools (Ferreira et al., 2012).

2.3. LLMs in supporting conceptual understanding

Large language models (LLMs) are increasingly applied to tasks requiring conceptual understanding, helping users transform large volumes of unstructured information into organized and interpretable forms. For example, they can summarize complex content, reorganize scattered ideas, generate explanations for abstract concepts, and prompt users to compare, connect, and elaborate on ideas (Lam et al., 2024; Suh et al., 2023), thereby functioning as cognitive scaffolds for integrating information into coherent knowledge structures. Prior studies have explored how LLMs can facilitate conceptual understanding through summarization (Suh et al., 2023; Yen & Zhao, 2024), reorganization (Y. Liu et al., 2025c; Suh et al., 2023), and content generation (Jiang et al., 2023; Wu et al., 2022; R. Zhang et al., 2024; Z. Zhang et al., 2023). A growing body of research has also examined the integration of LLMs with VR to support conceptual understanding. Existing work has primarily focused on adaptive explanations and guidance (e.g., Domenech et al., 2024), agent-based interaction (e.g., Liu et al., 2024), and personalized feedback (e.g., Gianni et al., 2025). However, it remains unclear how LLM-based scaffolding can support learners in identifying, organizing, and evaluating conceptual relationships in VR.

3. System design

To inform the system design, this study drew on literature documenting recurring challenges in relational knowledge construction. The identified challenges were used to derive scaffolding needs, which were subsequently instantiated as LLM-driven features in the proposed system.

3.1. Scaffolds with conceptual, relational, and metacognitive support

Recurring challenges in relational knowledge construction have been identified in the concept mapping literature. Concept mapping requires learners to identify concepts, establish meaningful relationships among them, and organize these into a coherent structure, making it a representative task for relational knowledge construction. First, learners often struggle to understand concept meanings, particularly when identifying core concepts, interpreting unfamiliar terms, or distinguishing between closely related ideas. Insufficient conceptual clarity increases cognitive load and leads to vague or inaccurate propositions (A. Cañas et al., 2015). Second, learners frequently face difficulties constructing relational structures, especially when managing multiple connections simultaneously (J. D. Novak & Cañas, 2007). This relational complexity can cause learners to rely on superficial associations or omit important connections (J. D. Novak, 1990a; Schwendimann, 2023). Third, concept map construction places substantial metacognitive demands on learners, who must monitor coherence, identify errors or missing links, and assess the completeness of their representations (A. J. Cañas & Novak, 2006; J. D. Novak & Cañas, 2004). Without external feedback or self-assessment skills, learners may fail to detect gaps or inaccuracies in the map (Kinchin et al., 2000).

To address these challenges, we proposed a scaffolding framework comprising conceptual, relational, and metacognitive supports tailored to relational knowledge construction (Table 1). Specifically, conceptual scaffolds provide definitions, examples, and clarifying prompts that strengthen learners' understanding of key concepts and reduce the cognitive effort required to interpret unfamiliar information. Relational scaffolds help learners establish accurate connections by automatically identifying potentially related concepts and offering multiple relationship types through which two concepts may be linked. These supports reduce the burden of relational search and support learners in constructing coherent and rich propositional networks. Finally, metacognitive scaffolds, including focus questions and

Table 1. Scaffolding types and design features.

Scaffolding types	Description	Design features
Conceptual scaffold	Provides essential conceptual information to help learners understand key ideas before forming relationships.	• Definitions of the concept • Examples or clarifying explanations of the concept
Relational scaffold	Supports learners in identifying and constructing accurate relationships between concepts.	• Automatically generates links between related concepts • Multiple relational framings
Metacognitive scaffold	Helps learners monitor and refine their maps through reflection and evaluation.	• Focus questions to guide learning process • Self-evaluation questions to assess understanding

self-evaluation prompts, guide learners in monitoring their understanding, identifying inconsistencies or missing links, and refining the overall structure of their map.

3.2. System description

Building on these scaffolds, this study introduces CmapVR, an immersive VR environment that supports relational knowledge construction through LLM-enhanced concept mapping. The system integrates a spatial canvas for concept map construction with a learning dashboard (Figure 1).

3.2.1. LLM-based features

The scaffolds described earlier are operationalized through a set of LLM-based features. **Conceptual scaffolds** take the form of LLM-generated "*Definitions*" and "*Examples*" attached to each concept node (ⓐ), providing learners with essential background knowledge. **Relational scaffolds** support the construction and refinement of connections through two mechanisms. First, the system automatically connects concept pairs with strong semantic associations based on predefined criteria (ⓑ). Second, upon manually linking two concepts, a relationship panel presents four structured prompts (ⓒ): "*Relationship*" (R), "*Compare*"(C), "*Integrate*" (I), and "*Infer*" (F), guiding learners to explore different relational interpretations and strengthen propositional coherence. Each prompt draws on fine-tuned LLM outputs aligned with its intended cognitive process. For example, selecting "*Compare*" generates a structured comparison of the two concepts across predefined dimensions. Generated content is integrated into the current map context (ⓓ). **Metacognitive scaffolds** prompt learners to monitor, evaluate, and regulate their understanding throughout map construction. Focus questions (ⓔ) tailored to the topic encourage learners to reflect on the clarity and completeness of their emerging knowledge structures (e.g., How do signal transduction pathways influence gene regulation and downstream cellular processes?). Self-evaluation questions (24D5) guide learners to assess individual propositions, identify missing or incorrect connections, and evaluate the overall coherence of their map (e.g., Which statement correctly describes the relationship between DNA structure and RNA processing? A. RNA processing modifies RNA molecules that are transcribed from DNA. B. RNA processing alters the chemical structure of DNA. C. DNA structure is determined by RNA processing.). Additionally,

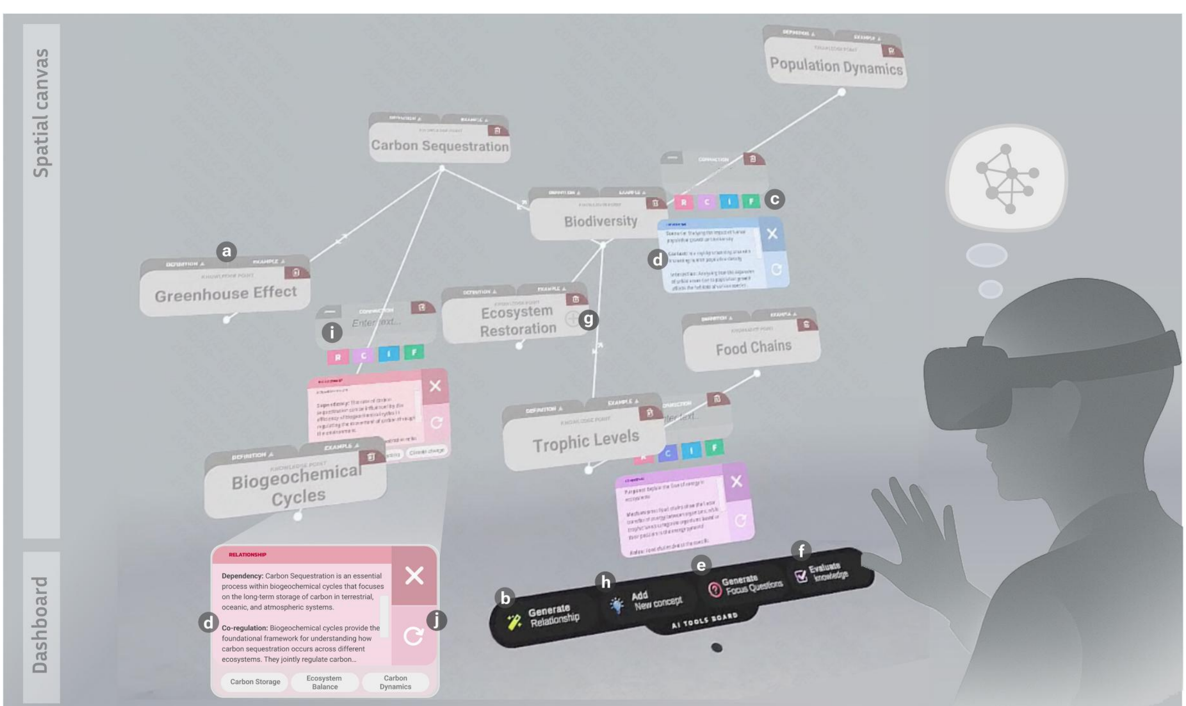

Figure 1. The user interface of CmapVR with a conceptual user silhouette overlaid. A user engages with concept map construction, with the support of a set of LLM-based features (ⓐ-ⓙ).

learners can expand the map by generating concepts related to a selected concept (ⓖ), or to all existing concepts (ⓗ). It also supports content refresh (ⓙ), allowing LLM-generated output to be updated in real time.

3.2.2. Interaction design

The system utilizes hand-based gesture interactions. Learners can use pinch gestures to select panels and activate buttons, and touch-and-release gestures to reposition panels. Text input via a virtual keyboard allows learners to enter relationships directly into the panel (ⓘ).

4. User study

A within-subject experiment was conducted to examine the effects of CmapVR on the relational knowledge construction of science concepts by comparing it with three other learning conditions.

4.1. Experimental design

4.1.1. Experimental conditions

This study compared CmapVR with three alternative conditions: (1) an immersive concept map construction environment with a web browser in VR (**VR_browse;** Figure 2a); (2) CmapTools with an LLM-driven conversational interface (ChatGPT-4o, OpenAI) on a desktop computer (**D_chat;** Figure 2b); and (3) CmapTools with a web browser on a desktop computer (**D_browse;** Figure 2c). CmapTools, a concept mapping software developed by the Institute for Human and Machine Cognition (Ferreira et al., 2012) was selected for the desktop conditions for its widespread use in educational contexts and core concept mapping capabilities, such as connecting nodes and specifying propositions through labeled links. VR_browse replicated CmapVR's layout and interaction design but replaced LLM-based features with web browsing for information retrieval. To minimize distractions from ads and irrelevant content, DuckDuckGo Lite (https://lite.duckduckgo.com/lite/) served as the search engine in both VR_browse and D_browse conditions. To ensure each condition adequately supported the task, D_chat permitted multi-turn interactions with the AI for query refinement, while participants in VR_browse and D_browse could freely search and explore various web pages to gather relevant information.

4.1.2. Task and procedure

Upon providing written informed consent, participants completed a demographic questionnaire. The task required participants to construct a concept map for a given scientific topic using their assigned tool ("condition"). Scientific concepts were used as learning materials to contextualize the relational knowledge construction task. To ensure comparability across conditions, eight concepts were pre-

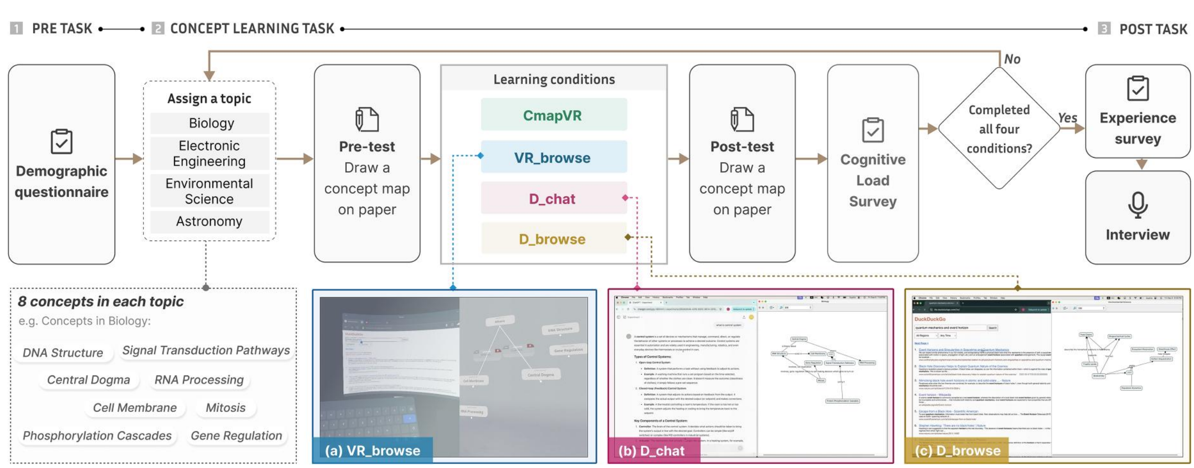

Figure 2. Overview of the experiment procedure.

selected for each topic from four distinct scientific disciplines to minimize cross-topic interference (Appendix A). The order of topics and conditions was counterbalanced across participants using a Latin square design. Pre- and post-tests of relational knowledge construction were administered before and after each condition, along with a cognitive load survey. In each test, participants diagramed relationships among the assigned concepts on paper. Participants were instructed to establish as many meaningful connections as possible and to explicitly describe each relationship rather than simply linking concepts. Lastly, participants completed a user experience survey and a semi-structured interview after all conditions (Figure 2).

4.1.3. Measures

The study measured performance, cognitive load, preference, usability, interaction behavior, and user feedback. Relational knowledge construction performance was assessed by comparing pre- and post-test concept maps. Two relationship types were used to evaluate performance (Figure 3): "a" denotes a proposition (a labeled link) reflecting reasoning between concepts; "b" denotes a crosslink connecting different branches of the map, reflecting knowledge integration across subtopics. For instance, "biogeochemical cycles" may be linked to "carbon sequestration" within one branch, while "greenhouse effect" may be linked to "population dynamics" within another. A crosslink connecting "carbon sequestration" and "population dynamics" across these branches requires learners to recognize that long-term carbon storage affects atmospheric carbon levels, which in turn influences ecosystem-wide population changes. Based on the literature, both the quantity and quality of relationships (propositions and crosslinks) are valid indicators of knowledge construction outcomes (Wallace & Mintzes, 1990). Established criteria (Schwendimann, 2023; Appendix A) were applied to assess relationship quantity and quality.

Cognitive load was measured using a questionnaire adapted from Krieglstein et al. (2023; Appendix B), assessing intrinsic, extraneous, and germane load. User experience comprised preference, assessed with a 7-point Likert scale (7 = highest rating), and perceived usability, assessed with the System Usability Scale (SUS; Lewis, 2018; Appendix B).

Participants' interaction behavior in CmapVR was analyzed to examine scaffold feature usage. Interaction logs were automatically recorded throughout each CmapVR session via a C# logging script tracking button-click behaviors. A semi-structured interview captured users' perceptions of each tool and challenges encountered (Appendix C).

4.1.4. Apparatus

A head-mounted display (Meta Quest 3, Meta) was used for CmapVR and VR_browse conditions. Both virtual environments were developed and rendered using Unity Engine (2022.3.19f) with hand tracking via the Meta XR Interaction SDK. In both desktop conditions, participants used a MacBook Pro with 32GB of memory. All LLM-based conditions used ChatGPT-4o (OpenAI) with identical system prompt configurations. CmapVR additionally employed scaffold-specific task prompts (See full prompts in Supplemental Materials), accessed via the OpenAI API with outputs delivered within the corresponding interface components.

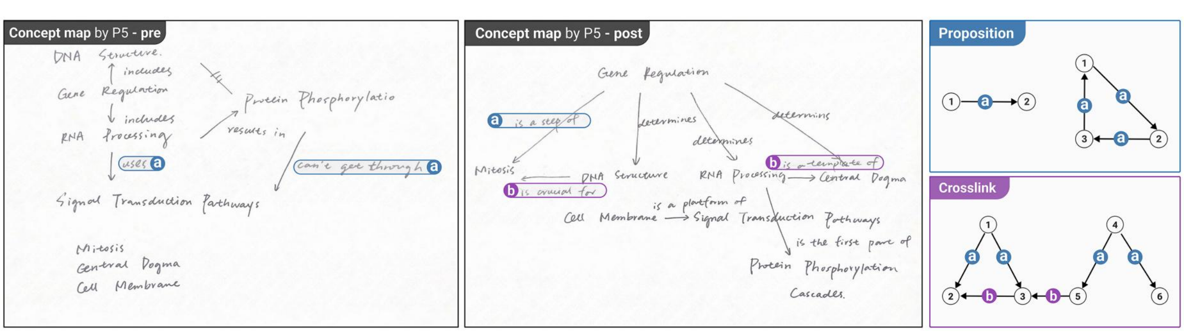

Figure 3. Example concept maps constructed by a participant (P5) in the pretest (left) and post-test (center). The right panels provide definitions of propositions and crosslinks, with representative examples highlighted in the maps.

4.1.5. Participants

Sixteen higher education students (9 females, 7 males) aged between 22 and 31 years ($M = 26.31$, $SD = 3.33$) participated in the study. Informed consent, approved by the Institutional Review Board of North Carolina State University (IRB protocol number 28025), was obtained from all participants. Participants reported their familiarity with VR ($M = 2.46$, $SD = 0.97$) on a 5-point scale (1 = very unfamiliar, 3 = neither familiar nor unfamiliar, 5 = very familiar), indicating a moderate level of unfamiliarity. They also reported usage frequency ($M = 1.4$, $SD = 0.97$) on a 7-point scale (1 = once a year, 4 = once a week, 7 = more than once per day), suggesting infrequent use. Participants' familiarity with LLMs was higher ($M = 3.23$, $SD = 1.01$), and their usage frequency was $M = 5.36$ ($SD = 1.36$), indicating usage around "several times a week."

4.2. Data analysis

4.2.1. Hypotheses

The following hypotheses were developed to guide the analysis: CmapVR is expected to outperform the other three conditions in relational knowledge construction (**H1**), as its scaffolds provide structured support while VR's spatial and embodied affordances further facilitate information processing; CmapVR's structured scaffolding is expected to result in lower cognitive load than VR_browse (**H2**); CmapVR is expected to receive higher preference ratings than the other three conditions (**H3**); CmapVR is expected to yield higher usability than VR_browse (**H4.a**), while the desktop conditions are expected to outperform both VR conditions given participants' greater familiarity with desktop interfaces (**H4.b**).

4.2.2. Data analysis methods

Two researchers independently counted and scored the propositions and crosslinks in the pre- and post-tests, blind to condition assignment. Inter-rater reliability was assessed using Cohen's kappa for each performance measure (Lombard et al., 2002), as both the scoring criteria and procedure are condition-independent, yielding: proposition count ($\kappa = 0.94$), crosslink count ($\kappa = 0.91$), proposition quality ($\kappa = 0.83$), and crosslink quality ($\kappa = 0.83$), indicating almost perfect agreement (Landis & Koch, 1977). Any remaining discrepancies were resolved through discussion. For the collected data (performance, cognitive load, preference ratings, and SUS scores), normality was assessed via the Shapiro-Wilk test, followed by one-way repeated-measures ANOVA with Tukey's post hoc tests or the Friedman test with Nemenyi-Wilcoxon comparisons for normally and non-normally distributed data, respectively. To control for familywise error, Holm-Bonferroni correction was applied across the four performance outcomes and across the three cognitive load subscales separately (Bender & Lange, 2001). Sensitivity analysis for repeated-measures ANOVA indicated the design was powered to detect effects of at least Cohen's $f = 0.305$ ($\eta^2 = 0.085$), suggesting adequate sensitivity for medium-sized but not smaller effects. Effect sizes were reported as partial eta squared (η_p^2) for ANOVA and Kendall's W for Friedman tests, with 95% confidence intervals (*CI*) estimated via the noncentral F distribution and bootstrapping (1,000 samples), respectively. Effect size benchmarks followed Cohen (2013) for η_p^2 (0.01, 0.06, 0.14), and Tomczak and Tomczak (2014) for Kendall's W (0.10, 0.30, 0.50), indicating small, medium, and large effects, respectively. To examine whether usability ratings may have confounded performance, repeated-measures correlations (Bakdash & Marusich, 2017) were conducted between SUS scores and each of the four performance outcomes.

For interaction behaviors, the total number of button clicks was logged and categorized according to the feature activated: *Definition* and *Example* (ⓐ; conceptual scaffold), *Generating Relationship* and *Relational Framings* (ⓑ and ⓒ; relational scaffold), and *Focus Questions* and *Self-evaluation Questions* (ⓓ and ⓔ; metacognitive scaffold), and *Others*. Others captured operational interactions, including closing panels, navigating back, and deleting. Clicks included both entry-level feature activation and subsequent interactions within feature panels, such as refreshing content or selecting options. Each feature's usage was expressed as a proportion of the participant's total button clicks. Due to a technical issue, one participant's log data was lost, leaving 15 participants for this analysis.

Semi-structured interview responses were analyzed using a thematic coding approach. Prior to coding, a codebook was developed based on the interview protocol, identifying an initial set of themes grounded in the key topics covered during the interviews. The transcripts were pre-segmented into units prior to coding, with each segment defined as a complete response to a single interview question. Each researcher then independently coded all interview transcripts using the codebook, working from the same pre-segmented units. Inter-rater reliability was assessed using Cohen's kappa (κ) separately for each theme, yielding $\kappa = 0.78$ overall (range across themes: [0.62, 1.00]), which indicates substantial agreement.

5. Results

5.1. Performance

There was a statistically significant difference across conditions in the increase in the number of propositions from pretest to post-test ($\chi^2(3) = 19.10$, $p < 0.001$, $W = 0.42$, 95% *CI* [0.27, 0.68]; Figure 4). CmapVR ($M = 5.40$, $SD = 2.50$) had significantly more increase in proposition count, outperforming VR_browse ($M = 1.00$, $SD = 2.03$, $p < 0.001$), D_chat ($M = 1.80$, $SD = 2.42$, $p < 0.01$), and D_browse ($M = 1.93$, $SD = 2.46$, $p < 0.01$). There was also a significant difference in the increase in crosslink counts ($\chi^2(3) = 22.93$, $p < 0.001$, $W = 0.51$, 95% *CI* [0.28, 0.79]). CmapVR ($M = 2.13$, $SD = 1.55$) had significantly more increase than VR_browse ($M = -0.47$, $SD = 1.25$, $p < 0.001$) and D_chat ($M = 0.33$, $SD = 1.05$, $p < 0.05$). Although numerically greater than D_browse ($M = 0.53$, $SD = 0.92$), the difference did not reach statistical significance ($p = 0.053$). These results indicate that CmapVR fostered more concept relationships and stronger knowledge connections across subtopics.

There was a significant difference across conditions in the increase in the proposition's quality scores ($F\ (3,42) = 18.52$, $p < 0.001$, $\eta_p^2 = 0.57$, 95% *CI* [0.39, 1.00]). CmapVR ($M = 17.40$, $SD = 9.01$) showed significantly greater increase than VR_browse ($M = 1.87$, $SD = 4.55$, $p < 0.001$), D_browse ($M = 4.33$, $SD = 6.18$, $p < 0.001$), and D_chat ($M = 4.40$, $SD = 6.21$, $p < 0.01$). There was also a significant difference across conditions in the increase in crosslink quality scores ($\chi^2(3) = 19.21$, $p < 0.001$, $W = 0.43$, 95% *CI* [0.23, 0.69]), with CmapVR ($M = 5.13$, $SD = 5.69$) showed significantly greater increase than VR_browse ($M = -1.87$, $SD = 4.03$, $p < 0.01$) and D_chat ($M = 0.13$, $SD = 2.26$, $p < 0.05$). This suggests that CmapVR helped users construct fewer generic relationships between concepts.

5.2. Cognitive load

There was a significant main effect of conditions on overall cognitive load ($F\ (3,42) = 9.37$, $p < 0.001$, $\eta_p^2 = 0.40$, 95% *CI* [0.19, 1.00]; Figure 5). Post-hoc comparisons revealed that VR_browse ($M = 5.20$, $SD = 1.0$) induced significantly higher overall cognitive load than CmapVR ($M = 4.53$, $SD = 0.85$, $p < 0.05$) and D_chat ($M = 4.04$, $SD = 0.69$, $p < 0.01$). D_chat induced significantly lower cognitive load

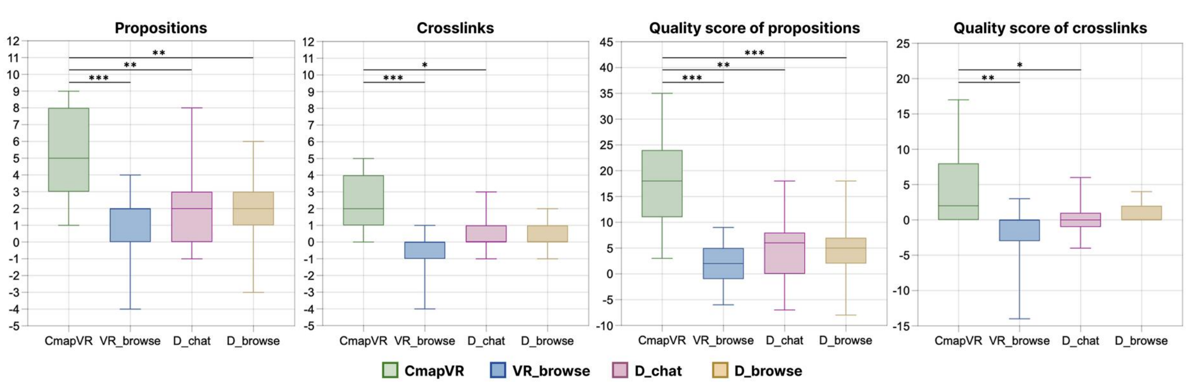

Figure 4. Results of relational knowledge construction performance across four conditions. In each plot, the box represents the interquartile range (IQR), the center line indicates the median, and the plus symbol (+) denotes the mean. Whiskers indicate the minimum and maximum values. Significant pairwise differences are denoted by asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

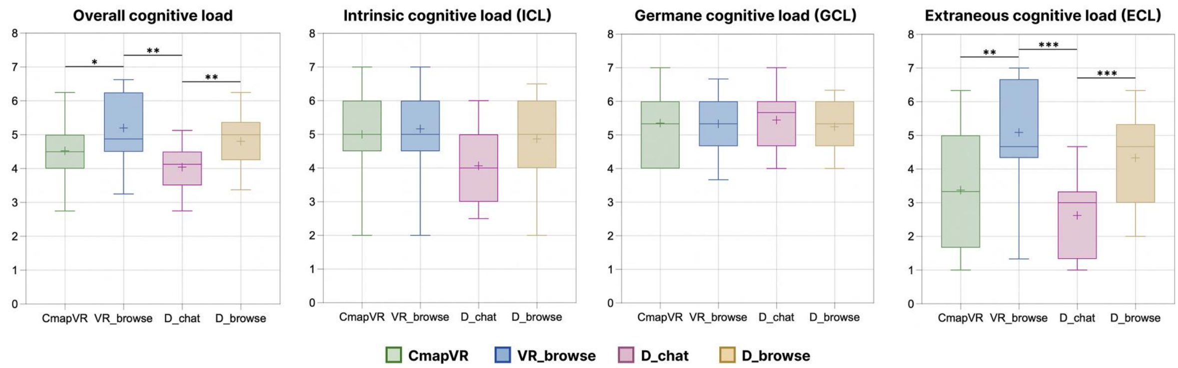

Figure 5. Results of cognitive load across four conditions. In each plot, the box represents the interquartile range (IQR), the center line indicates the median, and the plus symbol (+) denotes the mean. Whiskers indicate the minimum and maximum values. Significant pairwise differences are denoted by asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

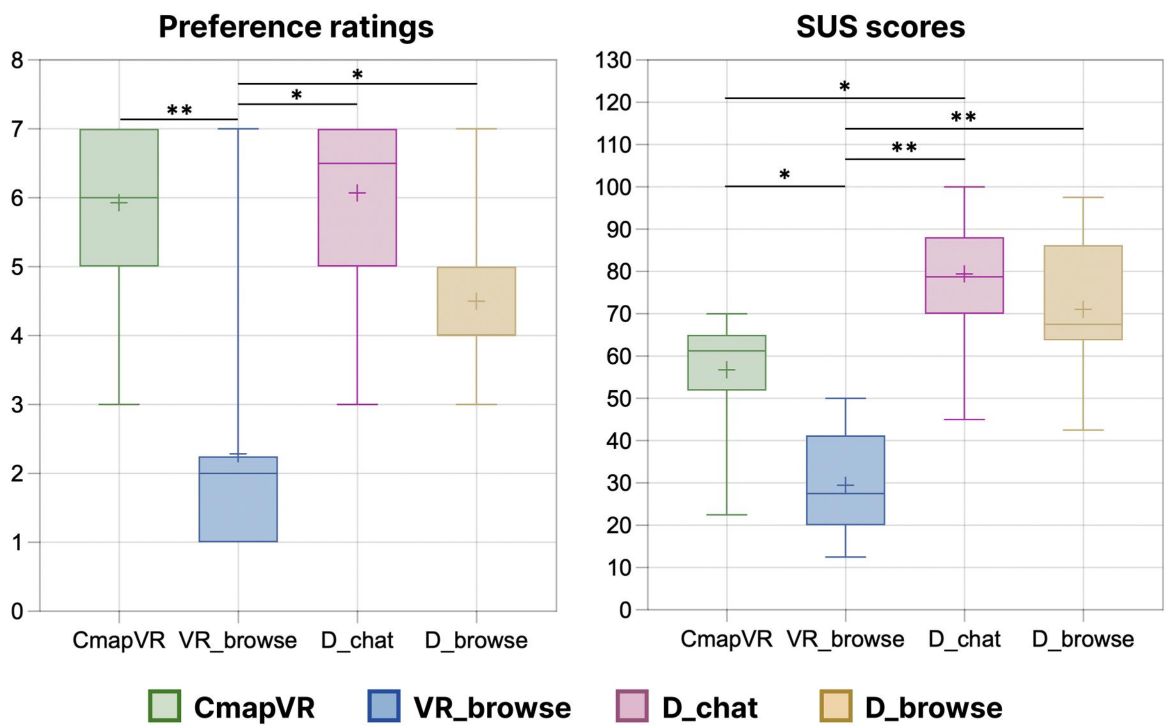

Figure 6. Results of user experience across four conditions. In each plot, the box represents the interquartile range (IQR), the center line indicates the median, and the plus symbol (+) denotes the mean. Whiskers indicate the minimum and maximum values. Significant pairwise differences are denoted by asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

than D_browse ($M = 4.81$, $SD = 0.82$, $p = 0.01$). For the subscales, significant differences in extraneous load were found across conditions (F (3,42) = 12.33, $p < 0.001$, η_p^2 = 0.47, 95% CI [0.26, 1.00]). Post-hoc comparisons revealed that VR_browse ($M = 4.76$, $SD = 1.72$) induced significantly higher extraneous load than CmapVR ($M = 3.58$, $SD = 1.94$, $p < 0.01$) and D_chat ($M = 2.62$, $SD = 1.10$, $p < 0.001$). D_chat induced significantly lower extraneous load than D_browse ($M = 4.33$, $SD = 1.39$, $p < 0.001$). Although intrinsic cognitive load differed significantly across conditions (F (3,42) = 3.82, $p < 0.05$, η_p^2 = 0.21, 95% CI [0.03, 1.0]), post-hoc comparisons revealed no significant pairwise differences. No significant differences in germane load were observed across conditions (F (3,42) = 0.33, $p = 0.80$).

5.3. Preference

A Friedman test revealed a significant effect of condition on user preference ratings ($\chi^2(3)$ = 28.42, $p < 0.001$, $W = 0.63$, 95% CI [0.51, 0.86]; Figure 6). Post-hoc Wilcoxon signed-rank tests (Bonferronicorrected) showed that participants preferred CmapVR ($M = 5.93$, $SD = 1.14$) significantly more than VR_browse ($M = 2.29$, $SD = 1.73$, $p < 0.01$). Participants also preferred D_chat ($M = 6.07$, $SD = 1.21$) over VR_browse ($p < 0.05$), and D_browse ($M = 4.50$, $SD = 1.09$) over VR_browse ($p < 0.05$).

5.4. Usability

A Friedman test revealed significant differences in usability across conditions ($\chi^2(3) = 36.55$, $p < 0.001$, $W = 0.76$, 95% CI [0.50, 0.96]). D_chat ($M = 79.5$, $SD = 14.2$) was rated significantly higher than VR_browse ($M = 29.5$, $SD = 12.3$, $p < 0.01$) and CmapVR ($M = 56.8$, $SD = 13.0$, $p < 0.05$). CmapVR was rated significantly higher than VR_browse ($p = 0.05$). Additionally, D_browse ($M = 71.1$, $SD = 15.9$) was rated significantly higher than VR_browse ($p < 0.01$). No significant correlations were found between usability and the increase in proposition count ($r(41) = 0.061$, $p = 0.696$, 95% CI [−0.244, 0.355]), proposition quality ($r(41) = 0.017$, $p = 0.915$, 95% CI [−0.285, 0.315]), crosslink count ($r(41) = 0.076$, $p = 0.626$, 95% CI [−0.229, 0.368]), or crosslink quality ($r(41) = 0.053$, $p = 0.737$, 95% CI [−0.252, 0.348]). This suggested that no meaningful relationship between usability and performance was detected in this sample.

5.5. Interaction logs

To examine participants' engagement with CmapVR's scaffold features, interaction logs were analyzed to capture scaffold usage frequency (Figure 7). Relational scaffolds accounted for the largest share of interactions ($M = 51.3\%$), with relational framings ($M = 34.7\%$, $SD = 15.8\%$) being the most frequently used feature, followed by automatically generated relationships ($M = 16.6\%$, $SD = 7.0\%$). Conceptual scaffolds were the second most used ($M = 14.7\%$), driven primarily by Definition interactions ($M = 12.8\%$, $SD = 10.8\%$), while Example use was minimal ($M = 1.8\%$, $SD = 2.6\%$). Metacognitive scaffolds were the least used ($M = 8.9\%$), with self-evaluation questions ($M = 7.3\%$, $SD = 16.1\%$) used more

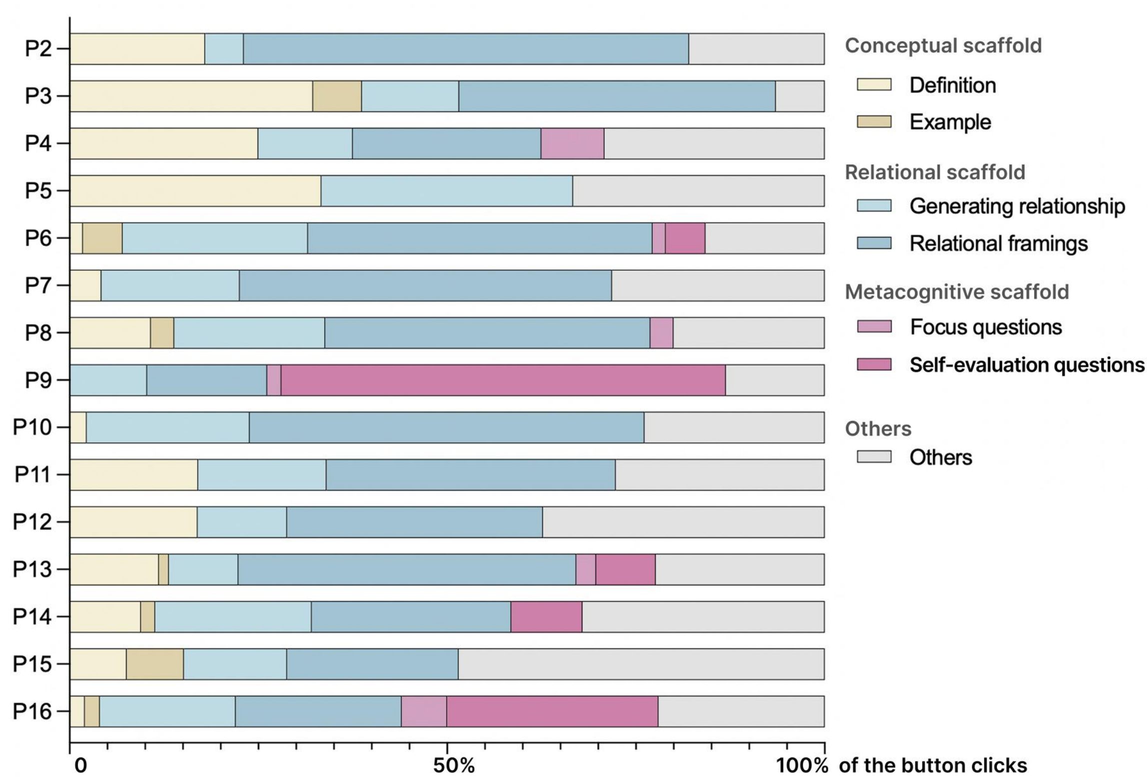

Figure 7. Distribution of interaction counts across features in CmapVR for each participant, expressed as a percentage of total button clicks.

frequently than focus questions ($M = 1.6\%$, $SD = 2.6\%$). The remaining interactions fell under Others ($M = 25.2\%$, $SD = 10.4\%$). Substantial individual variation was evident. Notably, P9 devoted 58.9% of interactions to self-evaluation questions while, eight participants (P2–3, P5, P7, P10–12, P15) recorded no metacognitive scaffold use throughout the session. P2 and P10 had the highest proportions of relational framing interactions (59.0% and 52.3%, respectively), while P5 recorded no relational framing interactions, distributing evenly across definitions, auto-generated relationships, and others (33.3% each).

5.6. Interview responses

5.6.1. CmapVR facilitated spatial presentation, embodied interaction, immersion, and focus

Participants identified four ways CmapVR's VR features supported knowledge construction. First, VR enhanced task engagement (3/16; hereafter n/16). P3, P11, and P15 noted that VR's novelty and virtual setup made the experience more interesting and engaging. Second, VR's immersive nature helped participants focus on the task (4/16). P8 noted that "*VR makes it harder to get distracted because you're more immersed in the environment, whereas with a desktop the surrounding space is more open and it's easier to lose focus.*" P6 added that desktop's easy access to other web pages posed a greater distraction risk. Third, VR's three-dimensional layout helped participants build and retain a clearer mental image of their concept map (5/16). P5 noted that VR's spatial structure made the map easier to recall, and P3 described it was "*more intuitive than 2D.*" P9 also noted that "*[VR] somehow formed a sense of space in my memory. I could roughly recall what the map looked like, which words were in which positions, and which words they were connected to.*" P1 described using spatial positioning as a memory strategy, adjusting concept placements to build a mental image. Finally, participants valued VR's active, hands-on interaction (3/16). P1 noted that being in VR "*creates an urge to try rearranging things,*" whereas in a 2D interface, "*there's no such impulse.*" P15 added that interacting physically with content, such as "*wanting to grab and manipulate things,*" encouraged deeper engagement with the material. P2 emphasized that being able to "*continuously drag and move these concepts*" felt more direct and helped them lay out groupings naturally.

5.6.2. LLM-Based features freed up cognitive resources for knowledge construction

Participants reported that CmapVR's LLM-based features reduced cognitive effort by facilitating information retrieval, multi-concept management, and knowledge exploration (7/16). First, LLM-based features reduced the burden of information retrieval. P6 noted that CmapVR presented information in structured modules, eliminating the need to craft prompts: "*to get the same information on a desktop, I would probably need to prompt many times, whereas here I only need to click a button.*" P7 pointed out that browser search is inherently sequential, requiring users to look up one concept at a time and manually trace connections, making it ill-suited for integrating eight interrelated concepts simultaneously. By contrast, in CmapVR, "*AI has already done a lot of that for you.*" P5 and P9 noted that CmapVR's LLM-based features were well-aligned with the task goal: "*almost every part of the system directly provided relationships, so the information it gave was consistent with what the task actually needed.*" (P5). Beyond information retrieval, CmapVR's LLMs also helped participants discover previously unconsidered connections. P5 noted that when using ChatGPT directly, "*you tend to only ask about things you've already thought of,*" whereas a system that automatically generates related concepts "*leads you to learn more and broader content just by clicking through.*"

By offloading retrieval and organization to the AI, participants felt they could devote more cognitive resources to understanding and retaining the conceptual structure (3/16). P3 noted that AI "*skips the filtering step*" that browser search requires. P13 noted that LLM-generated connections reduced the effort of relationship construction, allowing them to focus on evaluating and supplementing the provided links rather than generating them from scratch. P2 similarly commented that with CmapVR, "*all you need to do is to find the relationships between concepts and remember what it looks like.*"

5.6.3. Trust judgment of information across tools

Participants' trust in retrieved information varied across tools and was shaped by source visibility, information consistency, and presentation format. Several participants expressed greater trust in LLM-generated content (5/16), primarily due to AI's ability to synthesize information. P7 described AI output as "*a prepared meal*" compared to browser results, which are "*raw ingredients that still need to be cooked*." P10 and P15 observed that while browsers may present conflicting accounts, AI unified perspectives by synthesizing across sources. P2 noted that browser-based maps were more error-prone, given the difficulty of piecing together fragmented information under time pressure. Others expressed higher trust in browser-based sources due to their visibility and traceability (3/16). P6 and P11 noted that browsers allow users to judge source credibility directly, whereas LLMs may hallucinate. P13 reasoned that browser content is "*less likely to be obviously wrong*" given its grounding in established websites.

Several participants adopted a more calibrated stance (5/16). For some, trust was contingent on task purpose. P5 argued that for initial concept learning, broad understanding matters more than precision, making AI sufficiently trustworthy. P3 described a pragmatic approach, searching only when in doubt and treating the browser as a verification tool for "*confirming whether a connection makes sense.*" Presentation format also shaped participants' trust. P6 noted that CmapVR's structured, complete presentation suggested a well-designed system, reducing hallucination concerns and making information feel more trustworthy. P9 echoed this, noting that CmapVR's "*relatively fixed and clearly defined outputs*" felt more dependable than more variable responses from the chatbot in D_chat.

5.6.4. Learning strategies across tools

Different conditions shaped participants' engagement with the conceptual material, leading them to adopt different learning strategies.

Sequential, definition-first lookup in browser-based conditions (7/16). Participants typically began by looking up each concept individually before attempting to identify relationships in browser-based conditions. P11 described the process: "*first look up the definition of each term separately, then search for pairs of concepts together to see if they appear in the same results.*" P13 and P15 reported that after clarifying definitions, they drew on prior knowledge to hypothesize relationships, searching further only when a connection seemed implausible. However, participants noted limitations of this strategy when managing eight interrelated concepts. P5 and P7 noted that browser search is "*inherently one concept at a time,*" making it difficult to build complex, multi-directional relationships. The VR_browse condition introduced additional friction: P2 noted that navigating web pages in VR was slow, making the process even more laborious. Faced with the same challenge, P8 narrowed their strategy to focus on terms they were least certain about.

Structure-led, relation-first exploration in CmapVR (6/16). Rather than building understanding bottom-up from definitions, participants began by surveying the relational structure as a whole. P2, P11, and P15 noted that they first used CmapVR's automatically generated relationships to connect concepts into a draft network, which served as a scaffolded starting point for further exploration. P13 echoed this, noting that: "*I would first look at the connections the system provided, and if they seemed reasonable, I would note them down and then supplement them with my own understanding.*" P5 described using the density of connections as an orienting heuristic: "*I would look at which concepts had the most connections, as those tended to be the central ones that linked everything else together, so from the very beginning I knew which concepts were most important.*" P5 further noted that generated connections made unfamiliar concepts feel more approachable by linking them to known ones.

Batch input and conversational synthesis in D_chat. The conversational AI interface on the desktop enabled a streamlined retrieval strategy (3/16). P8 described inputting all eight concepts at once and requesting a unified explanation of their relationships. P9 noted this allowed them to focus on broader theoretical connections rather than information retrieval. However, P6 noted that the chat interface's open-ended nature meant output quality depended heavily on query framing, requiring more judgment to direct the conversation toward task-relevant information. Additionally, receiving multiple concept comparisons in a single output could feel overwhelming (P5).

5.6.5. Interactional and physical burdens in VR

Participants in both VR conditions reported interactional and physical burdens that disrupted knowledge structuring. First, the virtual keyboard was the most commonly cited source of frustration (6/16). P2, P5-7, and P11 noted that typing errors were frequent and difficult to correct. Second, participants found window switching and information comparison cumbersome in VR_browse (3/16). P11 pointed out that "*if you want to compare two concepts or their relationships, you have to switch tabs back and forth, and it's nowhere near as fast as using a mouse.*" P1 also noted that cursor imprecision in VR made tab switching and dragging effortful. Third, participants encountered difficulties with gesture-based selection and object movement (5/16). P5 attributed this to VR's three-dimensional nature, noting that "*dragging things back and forth and selecting targets were both difficult, and a lot of time was spent on basic operations rather than searching for knowledge itself.*" P15 similarly remarked that "*the overall operational feel was still not very clear,*" while P7 noted that limited VR familiarity added difficulty. Lastly, physical discomfort also contributed to reduced fluency (3/16). P8 and P9 reported neck discomfort from the headset's weight during extended use, and P15 noted hand fatigue from continuous movement.

5.6.6. Suggestions for system improvement and visions of an ideal learning tool

Participants offered suggestions for CmapVR (5/16), primarily regarding information density and input efficiency. P15 recommended progressive information disclosure to reduce reading fatigue: "*start by showing less content, and let users expand to see more if they want.*" P1 suggested automatic spatial structure generation to reduce manual arrangement effort. P6 proposed a curved display wall to improve peripheral visibility without requiring physical movement. P2, P5, P7, and P13 requested an integrated chat interface for queries beyond the system's existing prompts. Participants also anticipated richer interaction (P2, P4) and media modalities (P7).

Participants envisioned ideal systems for concept learning tasks (8/16), with VR–AI integration regarded as a promising direction (6/16). P6 described an ideal in which "*VR handles the intuitive presentation of the overall structure and relationship map, while LLM provides a convenient way to quickly input, double-check, or view more details.*" Notably, some participants distinguished contexts in which different tools would be most appropriate (3/16). P9 noted that for familiar topics, desktop-based AI would be sufficient, but for "*a set of terms with no prior concept at all,*" CmapVR would be preferable because "*it presents relationships and explanations in a more integrated and accessible way.*" P11 echoed this, stating that for unfamiliar topics, "*the structured support provided by VR combined with AI would be more helpful.*" P5 further suggested that CmapVR may be particularly well-suited for informal learning contexts such as science outreach or self-directed exploration.

6. Discussion

6.1. Factors contributing to the enhanced performance in CmapVR

CmapVR outperformed the three alternative conditions across most performance metrics, yielding greater gains in proposition and crosslink quantity and quality. The crosslink advantage over D_browse was not statistically significant but followed the same direction, and the overall pattern supports **H1**. These results suggest that CmapVR not only facilitated broader conceptual exploration but also supported more precise and complex relational representations. The following analysis examines the factors underlying these findings.

6.1.1. Enhanced information access supported broader relational knowledge exploration

Participants generated more propositions in CmapVR, indicating enhanced exploration and retention of inter-conceptual relationships. First, CmapVR facilitated learners' broader structural exploration as they built upon an existing relational foundation as opposed to constructing connections from scratch. Automatically generated connections served as an orienting heuristic, allowing participants to use connection density and distribution to identify central concepts and navigate unfamiliar material (P11, P15). This also supported the discovery of unanticipated connections: observing the network's structural properties encouraged learners to explore previously unconsidered relationships, such as linking isolated concepts to established clusters (P5). This process ultimately broadened the scope of conceptual exploration.

Furthermore, CmapVR's interface design and LLM-driven information generation enhanced information access, helping learners construct more conceptual relationships. Unlike other tools, which require users to formulate queries, navigate tabs and pages, and extract relevant content, CmapVR's LLM-based features, particularly structured relational framing, enabled learners to surface connections through direct, single-click interactions. Information retrieval with minimal effort enabled learners to more quickly identify more relationships across concepts. This is particularly valuable when constructing an integrated map across eight interrelated concepts (P7), where cumulative retrieval costs become substantial, and lowering this cost at each step helps learners to sustain exploratory engagement.

6.1.2. Structured relational scaffolding supported more precise proposition construction

CmapVR also yielded greater gains in proposition quality, indicating that learners formed more precise and well-articulated relational statements. One likely contributor was the structured information provided by LLM-based relational framings. Their task relevance was evidenced by interview data (P5–6, P9) and interaction logs, which recorded the highest activation rate among all scaffold types. These framings helped mitigate oversimplification by generating relational statements that captured conceptual nuances, preventing incomplete or overly generic propositions such as "related to" (Schwendimann, 2023). In browser-based conditions, learners often stopped short of articulating the precise nature of a confirmed relationship, as doing so required additional searches and cognitive effort to synthesize information across sources. Moreover, relational framings exposed learners to multiple relational perspectives across comparative, integrative, and inferential dimensions. Unlike D_chat, where relational information was generated in response to user-initiated queries and thus constrained by how questions were posed, CmapVR systematically presented multiple relational framings for the same concept pair in parallel. These framings could be complementary, each illuminating a distinct aspect of the relationship, enabling participants to draw selectively across them to construct richer, more precisely articulated propositions. Interview data further revealed that participants valued the system's capacity to pre-organize relational content into well-formed propositional statements (P2-3, P5, P13), which were more readily processed than the raw information returned by browser search or the extended responses in D_chat.

6.1.3. Systems-Level relational reasoning through integrated scaffold, spatial, and embodied support

CmapVR yielded the greatest gains in crosslink quantity and quality, suggesting CmapVR supported not only local relational reasoning between concept pairs but also systems-level reasoning across conceptual subtopics. This performance may be explained by two mechanisms suggested by participant responses. First, participants used automatically generated relationships to survey the conceptual landscape through a top-down, graph-based approach (P2, P11, P13, P15). This approach appeared to support their identification of connections across branches of the concept map, such as using the structural overview to locate conceptual clusters and explore potential links between them (P5). Once a potential crosslink was identified, CmapVR allowed learners to connect any two concepts and access LLM-generated relational content for that pair, perceived as reducing the effort required to articulate cross-topic relationships.

Second, CmapVR situated concept nodes within a three-dimensional spatial environment, enabling learners to perceive the overall map structure as a physical landscape. Several participants reported that this spatial presentation made the global organization of concepts more salient (P1, P3, P5, P9), helping them hold multiple conceptual regions in view and notice structural gaps between distant clusters that might otherwise have gone undetected. Furthermore, working in the immersive environment appeared to evoke a stronger impulse in some participants to bridge disconnected conceptual regions, which they associated with eventually forming crosslinks (P1-2, P15). Physically navigating and manipulating concept nodes was also perceived as deepening their sense of the connective tension between concepts, which may have reinforced their mental representation of inter-cluster relationships.

6.1.4. Supporting retention of relational knowledge

CmapVR reduced the need for users to retrieve, filter, or reorganize content, freeing cognitive resources for the retention of inter-conceptual relationships. This was reflected in the lower extraneous cognitive load reported in CmapVR, partially supporting **H2**. Extraneous cognitive load refers to mental effort arising from the design and presentation of the learning environment rather than from the learning content itself. Accordingly, information quality, structure, and interface design were likely key factors driving extraneous load differences across conditions. In CmapVR, task-aligned, pre-organized relational content reduced the need to filter information, likely lowering extraneous load compared to VR_browse, where participants navigated unstructured web content to identify relevant relationships. Similarly, D_chat's conversational interface enabled retrieval of synthesized, query-specific responses in a single interaction, reduced information-seeking effort compared to D_browse and VR_browse. Additionally, the lower cognitive load in D_chat relative to VR_browse may partly reflect participants' greater familiarity with conversational LLM interfaces ($M = 3.23$) compared to VR ($M = 2.46$). However, CmapVR's LLM-based features may have offset extraneous load from VR unfamiliarity, as its structured interface reduced the need for website navigation and text input, both of which participants found particularly demanding in VR. By minimizing irrelevant information management, CmapVR enabled learners to dedicate more cognitive resources to encoding and consolidating relational structures, likely contributing to greater proposition and crosslink retention in the post-test.

Furthermore, by lowering information access costs, CmapVR provided sufficient conceptual content for spatial memory mechanisms to engage. By contrast, the absence of such content in VR_browse left learners with little relational material to spatially organize or manipulate, which may have limited the extent to which VR's affordances could be leveraged. In CmapVR, participants reported that the three-dimensional layout supported formation of a clearer mental image of the concept map (P3, P5, P9), with some actively using spatial positioning as a deliberate memory strategy (P1). This pattern is consistent with research suggesting that three-dimensional environments can strengthen memory encoding by binding content to spatial location cues (Ekstrom & Hill, 2023; Maidenbaum et al., 2024; Mastrogiuseppe et al., 2019). Moreover, manipulable conceptual content may also have encouraged embodied interaction: participants described an urge to rearrange concepts, reposition nodes, and organize groupings through direct manipulation (P1, P15). Grounded in embodied cognition theory (Castro-Alonso et al., 2024), such interactions with external representations may reinforce formation of corresponding mental structures, potentially contributing to greater retention of relationships.

6.2. Using LLMs to support conceptual understanding in VR

Participants reported significantly higher preference ratings and usability scores for CmapVR than VR_browse, partially supporting **H3** and fully supporting **H4.a**, suggesting LLM-based scaffolding's potential to improve both learning performance and user experience in VR. The design and evaluation of CmapVR revealed four considerations regarding the use of LLMs in VR.

6.2.1. Modularizing scaffolds as LLM-based functions in VR

A key design principle of CmapVR is modularizing cognitive scaffolds into discrete, task-aligned LLM-based functions. Rather than providing a general-purpose conversational interface, CmapVR decomposed the scaffolding process into targeted functions, each addressing a specific cognitive demand: definitions and examples for conceptual understanding, structured relational framings for proposition construction, and focus and self-evaluation questions for metacognitive regulation. Based on these findings, this modular approach offered two advantages in VR. First, it reduced interaction overhead in prompt formulation, as learners could trigger scaffolding through direct manipulation rather than natural language queries. This reduced the need for text entry and may have lowered the cognitive load of prompt formulation. Second, these functions could be flexibly embedded within the spatial environment as interactive modules, enabling richer interface structures that could shape learners' cognitive processes and strategies. In

CmapVR, this enabled AI-generated relationships to be displayed directly along concept links, supporting a relation-first exploration strategy for navigating the conceptual landscape. These findings suggest that embedding LLMs in VR can regulate cognitive effort and shape learners' cognitive processes.

However, the LLM-based functions were not used evenly. Metacognitive scaffolding accounted for only 8.9% of interactions, and eight of fifteen participants with available logs recorded no metacognitive use, indicating that the present findings primarily support the effectiveness of the conceptual and relational scaffolds. This pattern may reflect the self-regulatory demands of metacognitive engagement, which some learners may not spontaneously initiate in immersive environments. How to better support metacognitive engagement in immersive contexts remains an open question for future research. One possibility is to implement context-aware functions that proactively deliver metacognitive support when learners may benefit from reflective engagement.

6.2.2. Influence of interface design on bias toward LLM-generated content

Participants' bias toward LLM-generated content depended not only on information source but on how that information was presented within the interface. CmapVR's structured, modular presentation, which delivered relational content in clearly defined, context-specific formats, reduced hallucination concerns and increased perceptions of the system as reliable and well-designed (P6, P9). This contrasts with open-ended conversational interfaces such as D_chat, where variable response length and format introduced greater uncertainty about information quality. These findings suggest that interface design may shape users' perception and trust of AI-generated content. While structured presentation can foster productive engagement, it may also inadvertently suppress critical evaluation of AI-generated content. Future designs should therefore incorporate explicit prompts encouraging users to verify AI-generated content, ensuring that interface-induced trust does not come at the cost of epistemic vigilance.

6.2.3. LLM-Backed functions as a prerequisite for realizing VR's spatial and embodied affordances

The findings suggest that VR's spatial and embodied affordances, while theoretically beneficial for knowledge construction, are contingent on the availability of accessible information. VR_browse's comparatively poorer performance suggests that immersion and spatial layout alone are insufficient to support relational knowledge construction, especially when learners must expend significant effort on information retrieval and organization before constructing relational representations. In this context, ready access to LLM-backed functions within the spatial interface was a prerequisite for realizing VR's spatial and embodied affordances. This implies that the cognitive value of immersive environments emerges from the interplay between their affordances and the quality of content they are designed to deliver.

6.3. Usability constraints in VR

Both VR-based conditions showed lower usability than the desktop-based condition, supporting **H4.b**. This can be attributed to two key factors. First, participants reported relatively low familiarity with VR ($M = 2.46$, $SD = 0.97$) and infrequent usage ($M = 1.40$, $SD = 0.97$). In contrast, participants were more comfortable with conversational LLMs, reporting higher familiarity ($M = 3.23$, $SD = 1.01$) and more frequent usage ($M = 5.36$, $SD = 1.36$). This likely introduced a learning curve despite the pre-experiment practices. Second, the most criticized element was mid-air typing, described as "chunky" and disruptive to task flow (P6). The difficulty of text entry in VR has been well documented in prior research (Z. Liu et al., 2025d; Yildirim & Osborne, 2020). Together, these results suggest that VR still presents nontrivial usability challenges, and that reducing the learning curve associated with VR interaction and improving text-entry mechanisms may enhance the usability of VR-based tools.

Notably, lower usability ratings in VR-based conditions did not correspond to poorer relational knowledge construction performance. In CmapVR, interactional friction from VR unfamiliarity and text entry did not appear to negate the cognitive benefits of its spatial affordances and LLM-based scaffolding. This pattern suggests that usability and performance capture distinct aspects of the learning experience and should not be treated as interchangeable indicators of system effectiveness. It also suggests that usability constraints, while consequential for user experience, may be addressable through design iteration without compromising CmapVR's learning advantages.

7. Limitations

Several limitations should be acknowledged. First, the learning tasks focused on structured, factual STEM content. Caution is therefore warranted when extending these results to less structured or more interpretive domains, such as history, ethics, or social science. Future research could further examine how CmapVR performs across a wider range of content types to better understand the scope of its applicability. Second, the relatively small, college-student sample limits both statistical power and generalizability to broader learner populations. Third, although family-wise correction was applied within the performance outcomes and cognitive load subscales defined by research hypotheses, the experiment-wise error rate across all tested hypotheses was not controlled. Fourth, the study did not independently manipulate VR modality and LLM-based support, and the use of different information sources, namely web browsers and LLM-based features, may also have introduced differences in the quality, accuracy, and ease of use of the information available to participants. Therefore, the findings should be viewed as exploratory evidence of CmapVR's potential benefits rather than as a direct test of the separate or interactive effects of VR and LLMs. Future research with larger samples and factorial designs should examine how VR, LLM-based scaffolding, and their specific forms of integration contribute to relational knowledge construction. Lastly, this study measured immediate post-test performance, and observed gains may therefore partly reflect recall of AI-supplied content from LLM-based features such as generated links and relational framings, rather than independently constructed knowledge. Whether these gains reflect durable understanding and whether relational knowledge persists, generalizes, or develops over time warrants longitudinal investigation.

8. Conclusion

This study investigated the effects of a VR-based learning environment augmented with LLM-based scaffolding on supporting relational knowledge construction. This work introduced CmapVR, a VR-based learning environment integrating conceptual, relational, and metacognitive scaffolds, and evaluated it through a science concept learning experiment with 16 participants, comparing it against three alternative learning conditions. Results showed that CmapVR led to greater improvements in both the quantity and quality of constructed concept relationships than all other conditions, while also inducing lower extraneous cognitive load and eliciting stronger user preference compared to browsing in VR. This study identified several factors underlying CmapVR's performance advantages, including enhanced information access, structured relational scaffolding, and the combined support of spatial and embodied affordances for systems-level reasoning and knowledge retention. It further discussed how modularizing LLM-based scaffolds as discrete functions shaped learners' knowledge construction processes and served as a prerequisite for realizing VR's affordances. Despite usability constraints related to VR familiarity and text entry, this study demonstrated that LLM-supported VR environments show promise for supporting relational knowledge construction while reducing unnecessary cognitive effort.

Acknowledgements

We would like to thank Emily Fang for serving as the second coder for the thematic coding of the interview responses, and Blake Boykin for serving as the second coder for counting and scoring the propositions and cross-links in the concept maps.

Author contributions

CRediT: **Fangyuan Cheng**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Bingyi Su**: Formal analysis, Writing – original draft; **Karen B. Chen**: Methodology, Project administration, Resources, Writing – review & editing. All authors agree to be accountable for all aspects of the work.

Declaration of generative AI use

The authors report that, based on the experimental design, generative AI (ChatGPT, OpenAI) was used to provide information in the CmapVR and D_Chat conditions. Additionally, it was used to check grammar and improve the flow of the manuscript. The authors take full responsibility for the content of the publication.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID

Fangyuan Cheng http://orcid.org/0009-0001-9306-8471

Data availability statement

The de-identified quantitative data and analysis materials supporting the findings of this study are available in the Open Science Framework at https://osf.io/syaqn/overview. Due to privacy and ethical restrictions related to participant consent, the raw interview responses are not publicly available.

References

Ahn, J., & Noh, G.-Y. (2025). Optimizing the educational potential of virtual reality: The effects of virtual reality and pop quizzes on spatial presence and efficacy perceptions. *International Journal of Human–Computer Interaction*, *41*(2), 1107–1118. https://doi.org/10.1080/10447318.2024.2313274

Alexander, P. A. (2016). Relational thinking and relational reasoning: Harnessing the power of patterning. *NPJ Science of Learning*, *1*(1), 16004. https://doi.org/10.1038/npjscilearn.2016.4

Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in teaching and learning: A systematic review. *Education Sciences*, *14*(6), 643. https://doi.org/10.3390/educsci14060643

Anderson, R. C. (1994). *Role of the reader's schema in comprehension, learning, and memory.* (p. 482) International Reading Association.

Anderson, R. C. (2018). Role of the reader's schema in comprehension, learning, and memory. In *Theoretical models and processes of literacy* (7th ed.). Routledge.

Araiza-Alba, P., Keane, T., Chen, W. S., & Kaufman, J. (2021). Immersive virtual reality as a tool to learn problem-solving skills. *Computers & Education*, *164*, 104121. https://doi.org/10.1016/j.compedu.2020.104121

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning* (pp. xiv, 255) Grune & Stratton.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how? *Journal of Clinical Epidemiology*, *54*(4), 343–349. https://doi.org/10.1016/s0895-4356(00)00314-0

Cañas, A. J., & Novak, J. D. (2006). Re-examining the foundations for effective use of concept maps. In *Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping* (Vol. 1, pp. 494–502).

Cañas, A., Novak, J. D., & Reiska, P. (2015). How good is my concept map? Am I a good Cmapper? *Knowledge Management & E-Learning*, *7*(1), 6–19.

Castro-Alonso, J. C., Ayres, P., Zhang, S., de Koning, B. B., & Paas, F. (2024). Research avenues supporting embodied cognition in learning and instruction. *Educational Psychology Review*, *36*(1), 10. https://doi.org/10.1007/s10648-024-09847-4

Cheng, F., Harper-Gampp, T., Planchart, R., Dunning, M., Peterson, M., Delgado, C., & Chen, K. B. (2024). Study of graphic armatures, multimodal cues, and numeric measures in virtual reality on learners' performance and workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 1229–1234. https://doi.org/10.1177/10711813241269236

Cimadevilla, J. M., Nori, R., & Piccardi, L. (2023). Application of virtual reality in spatial memory. *Brain Sciences*, *13*(12), 1621. https://doi.org/10.3390/brainsci13121621

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* routledge.

Domenech, J. I., Pellicer, J. L., & Molla, I. F. (2024). Virtual reality and language models, a new frontier in learning. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(5), 46–54. https://doi.org/10.9781/ijimai.2024.02.007

Ekstrom, A. D., & Hill, P. F. (2023). Spatial navigation and memory: A review of the similarities and differences relevant to brain models and age. *Neuron*, *111*(7), 1037–1049. https://doi.org/10.1016/j.neuron.2023.03.001

Ferreira, P. B., Cohrs, C. R., & De Domenico, E. B. L. (2012). [Software CMAP TOOLS ™ to build concept maps: An evaluation by nursing students]. *Revista da Escola de Enfermagem da U S P*, *46*(4), 967–972. https://doi.org/10.1590/s0080-62342012000400026

Ferretti, S., Bianchi, S., Frangipane, G., & Di Leonardo, R. (2021). A virtual reality interface for the immersive manipulation of live microscopic systems. *Scientific Reports*, *11*(1), 7610. https://doi.org/10.1038/s41598-021-87004-5

García-Méndez, S., de Arriba-Pérez, F., Somoza-López, M., & del, C. (2025). A review on the use of large language models as virtual tutors. *Science & Education*, *34*(2), 877–892. https://doi.org/10.1007/s11191-024-00530-2

Ghali, R., Abdessalem, H. B., Frasson, C. (2017). Improving intuitive reasoning through assistance strategies in a virtual reality game. In V. Rus & Z. Markov (Eds.), *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017* (pp. 382–387). AAAI Press. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15445

Gianni, A. M., Nikolakis, N., & Antoniadis, N. (2025). An LLM based learning framework for adaptive feedback mechanisms in gamified XR. *Computers & Education: X Reality*, *7*, 100116. https://doi.org/10.1016/j.cexr.2025.100116

Hajian, S. (2019). Transfer of learning and teaching: A review of transfer theories and effective instructional practices. *IAFOR Journal of Education*, *7*(1), 93–111. https://doi.org/10.22492/ije.7.1.06

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505. https://doi.org/10.1016/j.tics.2010.08.005

Hong, Y. A., Shi, X., & Dogan, S. (2026). A systematic review of virtual reality technology and human memory augmentation. *International Journal of Human–Computer Interaction*, *42*(4), 1971–2011. https://doi.org/10.1080/10447318.2025.2530080

Iran-Nejad, A. (1987). The schema: A long-term memory structure or a transient structural phenomena. In *Understanding readers' understanding*. Routledge.

Jiang, P., Rayan, J., Dow, S. P., & Xia, H. (2023). *Graphologue: Exploring large language model responses with interactive diagrams* [Paper presentation]. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–20. UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology. https://doi.org/10.1145/3586183.3606737

Johnson-Glenberg, M. C., Megowan-Romanowicz, C., Birchfield, D. A., & Savio-Ramos, C. (2016). Effects of embodied learning and digital platform on the retention of physics content: Centripetal force. *Frontiers in Psychology*, *7*, 1819. https://doi.org/10.3389/fpsyg.2016.01819

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge* (pp. ix,–265). Lawrence Erlbaum Associates, Inc.

Juliano, J. M., Schweighofer, N., & Liew, S.-L. (2022). Increased cognitive load in immersive virtual reality during visuomotor adaptation is associated with decreased long-term retention and context transfer. *Journal of Neuroengineering and Rehabilitation*, *19*(1), 106. https://doi.org/10.1186/s12984-022-01084-6

Kazemi, R., & Lee, S. C. (2024). Human factors/ergonomics (HFE) evaluation in the virtual reality environment: A systematic review. *International Journal of Human–Computer Interaction*, *40*(17), 4533–4549. https://doi.org/10.1080/10447318.2023.2227835

Khorasani, S., Victor Syiem, B., Nawaz, S., Knibbe, J., & Velloso, E. (2023). Hands-on or hands-off: Deciphering the impact of interactivity on embodied learning in VR. *Computers & Education: X Reality*, *3*, 100037. https://doi.org/10.1016/j.cexr.2023.100037

Kilteni, K., Groten, R., & Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, *21*(4), 373–387. https://doi.org/10.1162/PRES_a_00124

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, *42*(1), 43–57. https://doi.org/10.1080/001318800363908

Kintsch, W. (1998). *Comprehension: A paradigm for cognition* (pp. xvi, 461). Cambridge University Press.

Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, *35*(1), 9. https://doi.org/10.1007/s10648-023-09738-0

Kyllonen, P. C. (2020, July). *Reasoning abilities*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.878

Lam, M. S., Teoh, J., Landay, J. A., Heer, J., & Bernstein, M. S. (2024). *Concept induction: Analyzing unstructured text with high-level concepts using LLooM* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–28. https://doi.org/10.1145/3613904.3642830

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307
Lisle, L., Davidson, K., Gitre, E. J. K., North, C., & Bowman, D. A. (2021). *Sensemaking strategies with immersive space to think* [Paper presentation]. 2021 IEEE Virtual Reality and 3D User Interfaces (VR), 529–537. https://doi.org/10.1109/VR50410.2021.00077
Liu, C., Meng, S., Zheng, W., & Zhou, Z. (2025a). Research on the impact of immersive virtual reality classroom on student experience and concentration. *Virtual Reality*, *29*(2), 82. https://doi.org/10.1007/s10055-025-01153-w
Liu, D., He, P., & Yan, H. (2025b). The impact of cognitive schema on learning transfer ability and stability of classical Chinese poetry. *PloS One*, *20*(11), e0336135. https://doi.org/10.1371/journal.pone.0336135
Liu, Y., Wen, Z., Weng, L., Woodman, O., Yang, Y., & Chen, W. (2025c). SPROUT: An interactive authoring tool for generating programming tutorials with the visualization of large language models. *IEEE Transactions on Visualization and Computer Graphics*, *31*(9), 4761–4775. https://doi.org/10.1109/TVCG.2024.3410523
Liu, Z., Liu, X., & Wang, L. (2025d). CrossKeys: Text entry for virtual reality using a single controller via wrist rotation. *International Journal of Human–Computer Interaction*, *41*(9), 5150–5162. https://doi.org/10.1080/10447318.2024.2358456
Liu, Z., Zhu, Z., Zhu, L., Jiang, E., Hu, X., Peppler, K. A., & Ramani, K. (2024). *ClassMeta: Designing interactive virtual classmate to promote vr classroom participation* [Paper presentation]. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI, '24, 1–17. https://doi.org/10.1145/3613904.3642947
Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x
Lui, M., Chong, K.-Y., Mullally, M., & Mcewen, R. (2023). Facilitated model-based reasoning in immersive virtual reality: Meaning-making and embodied interactions with dynamic processes. *International Journal of Computer-Supported Collaborative Learning*, *18*(2), 203–230. https://doi.org/10.1007/s11412-023-09396-y
Maidenbaum, S., Kremen, V., Worrell, G., & Jacobs, J. (2024). Improved spatial memory for physical versus virtual navigation. PsyArXiv. https://doi.org/10.31234/osf.io/d5fra
Makransky, G., Andreasen, N. K., Baceviciute, S., & Mayer, R. E. (2021). Immersive virtual reality increases liking but not learning with a science simulation and generative learning strategies promote learning in immersive virtual reality. *Journal of Educational Psychology*, *113*(4), 719–735. https://doi.org/10.1037/edu0000473
Markham, K. M., Mintzes, J. J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching*, *31*(1), 91–101. https://doi.org/10.1002/tea.3660310109
Mastrogiuseppe, M., Bertelsen, N., Bedeschi, M. F., & Lee, S. A. (2019). The spatiotemporal organization of episodic memory and its disruption in a neurodevelopmental disorder. *Scientific Reports*, *9*(1), 18447. https://doi.org/10.1038/s41598-019-53823-w
Mayer, R. E., Makransky, G., & Parong, J. (2023). The Promise and pitfalls of learning in immersive virtual reality. *International Journal of Human–Computer Interaction*, *39*(11), 2229–2238. https://doi.org/10.1080/10447318.2022.2108563
McGowin, G., Fiore, S. M., & Oden, K. (2021). Learning affordances: Theoretical considerations for design of immersive virtual reality in training and education. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 883–887. https://doi.org/10.1177/1071181321651293
McVee, M. B., Dunsmore, K., & Gavelek, J. R. (2005). Schema theory revisited. *Review of Educational Research*, *75*(4), 531–566. https://doi.org/10.3102/00346543075004531
Meylani, R. (2024). Innovations with schema theory: Modern implications for learning, memory, and academic achievement. *International Journal For Multidisciplinary Research*, *6*(1), 2582–2160. https://doi.org/10.36948/ijfmr.2024.v06i01.13785
Moon, B., Hoffman, R. R., Novak, J., & Cañas, A. (2011). *Applied concept mapping: Capturing, analyzing, and organizing knowledge*. CRC Press.
Nakić, J., Rogulj, I., & Maleš, L. (2025). Rethinking the Bebras challenge in virtual reality: Implementation and usability study of a computational thinking game. *Multimodal Technologies and Interaction*, *9*(6), 60. https://doi.org/10.3390/mti9060060
National Research Council (2007). *Taking science to school: Learning and teaching science in grades K–8*. The National Academies Press. https://doi.org/10.17226/11625
Novak, J. D. (1990a). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, *27*(10), 937–949. https://doi.org/10.1002/tea.3660271003
Novak, J. D. (1990b). Concept maps and Vee diagrams: Two metacognitive tools to facilitate meaningful learning. *Instructional Science*, *19*(1), 29–52. https://doi.org/10.1007/BF00377984
Novak, J. D. (1995). Concept mapping to facilitate teaching and learning. *Prospects*, *25*(1), 79–86. https://doi.org/10.1007/BF02334286

Novak, J. D., & Cañas, A. J. (2006). *The theory underlying concept maps and how to construct them* (Technical Report IHMC CmapTools 2006-01). Florida Institute for Human and Machine Cognition. https://cmap.ihmc.us/publications/researchpapers/theoryunderlyingconceptmaps.pdf
Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press.
Novak, J., & Cañas, A. (2007). Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting Education*, *3*(1), 29–42.
Novak, N. J. D., & Cañas, A. J. (2004). Building on new constructivist ideas and CmapTools to create a new model for education. In *Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping* (Vol. 1, pp. 469–476).
Novotny, J., & Laidlaw, D. H. (2024). Evaluating text reading speed in VR scenes and 3D particle visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *30*(5), 2602–2612. https://doi.org/10.1109/TVCG.2024.3372093
Radvansky, G. A., & Copeland, D. E. (2004). Reasoning, integration, inference alteration, and text comprehension. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *58*(2), 133–141. https://doi.org/10.1037/h0085793
Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In *Theoretical issues in reading comprehension*. Routledge.
Rzayev, R., Ugnivenko, P., Graf, S., Schwind, V., & Henze, N. (2021). *Reading in VR: The effect of text presentation type and location* [Paper presentation]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, 1, ––10. https://doi.org/10.1145/3411764.3445606
Schutte, N. S. (2020). The impact of virtual reality on curiosity and other positive characteristics. *International Journal of Human–Computer Interaction*, *36*(7), 661–668. https://doi.org/10.1080/10447318.2019.1676520
Schwendimann, B. A. (2023). Concept maps as versatile learning, teaching, and assessment tools. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology* (pp. 631–693). Springer International Publishing. https://doi.org/10.1007/978-3-319-17461-7_86
Skarbez, R., Polys, N. F., Ogle, J. T., North, C., & Bowman, D. A. (2019). Immersive analytics: Theory and research agenda. *Frontiers in Robotics and AI*, *6*, 82. https://doi.org/10.3389/frobt.2019.00082
Skulmowski, A., & Rey, G. D. (2018). Embodied learning: Introducing a taxonomy based on bodily engagement and task integration. *Cognitive Research: principles and Implications*, *3*(1), 6. https://doi.org/10.1186/s41235-018-0092-9
Suh, S., Min, B., Palani, S., & Xia, H. (2023). *Sensecape: Enabling multilevel exploration and sensemaking with large language models* [Paper presentation]. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–18. https://doi.org/10.1145/3586183.3606756
Sultana, S., & Alam, A. (2024). *Exploring the impact of display field-of-view and display type on text readability and visual text search on large displays in virtual reality*.
Tan, S., Erdimez, O., & Zimmerman, R. (2017). Concept mapping as a tool to develop and measure students' understanding in science. *Acta Didactica Napocensia*, *10*(2), 109–122. https://doi.org/10.24193/adn.10.2.9
Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited: An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *21*(1), 19–25.
Urdiales-Ibarra, M. E., Lopez-Ramirez, E. O., Castro-Campos, C., Villarreal-Treviño, M. G., & Carrillo-Colon, J. E. (2018). Biology schemata knowledge organization and meaning formation due to learning: A constructive-chronometric approach to concept mapping usability. *Creative Education*, *09*(16), 2693–2706. https://doi.org/10.4236/ce.2018.916203
Varga, N. L., Morton, N. W., Preston, A. (2020). *Schema, inference, and memory*. https://www.semanticscholar.org/paper/Schema-%2C-Inference-%2C-and-Memory-Varga-Morton/84878e7693ca92c597b21e999914dee15c05e090
Wallace, J. D., & Mintzes, J. J. (1990). The concept map as a research tool: Exploring conceptual change in biology. *Journal of Research in Science Teaching*, *27*(10), 1033–1052. https://doi.org/10.1002/tea.3660271010
Wu, T., Terry, M., & Cai, C. J. (2022). *AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts* [Paper presentation]. CHI Conference on Human Factors in Computing Systems, 1–22. CHI '22: CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517582
Yen, R., & Zhao, J. (2024). *Memolet: Reifying the reuse of user-AI conversational memories* [Paper presentation]. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 1–22. https://doi.org/10.1145/3654777.3676388
Yildirim, C., & Osborne, E. (2020). Text entry in virtual reality: A comparison of 2D and 3D keyboard layouts. In C. Stephanidis, J. Y. C. Chen, & G. Fragomeni (Eds.), *HCI International 2020 – Late breaking papers: Virtual and augmented reality* (Vol. 12428, pp. 450–460). Springer International Publishing. https://doi.org/10.1007/978-3-030-59990-4_33

Zhang, R., Zhang, Z., Zhu, F., Zhou, J., & Rao, A. (2024). Mindalogue: LLM-powered nonlinear interaction for effective learning and task exploration. *arXiv*. https://doi.org/10.48550/ARXIV.2410.10570

Zhang, Z., Gao, J., Dhaliwal, R. S., & Li, T. J.-J. (2023). *VISAR: A human-AI argumentative writing assistant with visual programming and rapid draft prototyping* [Paper presentation]. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–30. UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology. https://doi.org/10.1145/3586183.3606800

About the authors

Fangyuan Cheng is a PhD student in Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She earned her master's degree from Tongji University in 2023 and her bachelor's degree from Hunan University in 2020. Her research focuses on human cognition in extended reality.

Bingyi Su received his PhD in Industrial and Systems Engineering from North Carolina State University. His research focuses on human factors, physiological computing, and human-centered intelligent systems, with particular interests in cognitive workload, stress, trust, and the design and evaluation of LLM-enhanced agents.

Karen B. Chen is an Associate Professor in Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. Chen received her PhD from Department of Biomedical Engineering at University of Wisconsin-Madison in 2015. Her interest is in human performance and behavior in VR and human-computer interaction.

Appendix A

Science Concepts and Scoring Rubric for Concept Maps

Table A1. Science concepts for the four topics.

Astronomy	Electronic engineering	Environmental science	Biology
Solar System	Circuit Design	Food Chains	DNA Structure
Gravity and Orbits	Control Systems	Biogeochemical Cycles	Central Dogma
Stars	Electromagnetic Fields	Ecosystem Restoration	Cell Membrane
The Sun's Structure	Power Electronics	Greenhouse Effect	Mitosis
Quantum Mechanics	Signal Processing	Biodiversity	Protein Phosphorylation Cascades
Stellar Evolution	Microcontrollers	Trophic Levels	Gene Regulation
General Relativity	Digital Systems	Population Dynamics	Signal Transduction Pathways
Event Horizons	Embedded Systems	Carbon Sequestration	RNA Processing

Table A2. Scoring rubric.

Accuracy	Score	Definition
Excellent	4	Outstanding proposition. Complete and correct. It shows a deep understanding of the relationship between the two concepts. Example: metamorphic rocks can be made from igneous rocks
Good	3	Complete and correct proposition. It shows a good understanding of the relationship between the two concepts. Example: climate exposes minerals
Average	2	Incomplete but correct proposition. It shows partial understanding of the relationship between the two concepts. Example: minerals help geologists
Below Average	1	Although valid, the proposition does not show understanding of the relationship between the two concepts. Example: water is earth materials
Invalid/inaccurate	0	The proposition is incorrect. Example: limestone is quartz

Appendix B

Survey Instruments

Table B1. Cognitive load questionnaire.

Index	Type of load	Item
Q1	ICL	For this task, many things needed to be kept in mind simultaneously.
Q2	ICL	This task was very complex.
Q3	GCL	I made an effort, not only to understand several details, but to understand the overall context.
Q4	GCL	My point while dealing with the task was to understand everything correctly.
Q5	GCL	The learning task consisted of elements supporting my comprehension of the task.
Q6	ECL	During this task, it was exhausting to find the important information.
Q7	ECL	The design of this task was very inconvenient for learning.
Q8	ECL	During this task, it was difficult to recognize and link the crucial information.

Note: The abbreviations in the table refer to ICL (Intrinsic Cognitive Load), GCL (Germane Cognitive Load), and ECL (Extraneous Cognitive Load).

Table B2. SUS questionnaire.

Index	Item
Q1	I am willing to frequently use this learning tool/environment in my daily studies.
Q2	I found the learning tool/environment unnecessarily complex.
Q3	I thought the learning tool/environment was easy to use.
Q4	I think that I would need the support of a technical person to be able to use this learning tool/environment.
Q5	I found the various functions in this learning tool/environment were well integrated.
Q6	I thought there was too much inconsistency in this learning tool/environment.
Q7	I would imagine that most people would learn to use this learning tool/environment very quickly.
Q8	I found the learning tool/environment very cumbersome to use.
Q9	I felt very confident using the learning tool/environment.
Q10	I needed to learn a lot of things before I could get going with this learning tool/environment.

Appendix C

Table C1. Semi-structured interview protocol.

Category	Item
Experience	Could you tell me what it was like to learn with these tools/environments in general? Which one do you like best?
Expectation	Before starting learning, what role did you expect the tools/environments to take? And what role did it actually take during the learning process? Did the tools/environments match your expectations?
Usefulness	Which function of these tools/environments did you find most useful? Why?
Confusion	Do you remember situations in which you found the behavior of any tool/environment confusing or irritating?
	Can you describe situations in which the communication worked well and not so well?
Role	What role do you want the large-language models (AI) to take during your learning process?
Trust	How much do you trust the results from the tools/environments?
Comparison	How was the learning experience with each tool/environment different from your workflow of working with a human teacher?
Comparison	How do you think the learning process in VR differs from the learning process on a desktop?
Supportiveness	Do you think the AI feature successfully understands your learning intent and provides sufficient support?
Learning process	How did you find that the tools/environments changed the way you learn new concepts? Did you ever adjust your learning process / strategy according to the tool's capabilities?
	Which situations of the learning process did require the most support and guidance?
Ideal tool	How would you describe an ideal learning tool/environment for learning scientific concepts?