Orchestrating Generative AI in an Immersive Learning Environment: A Participatory Design Study in Higher Education

Riem Yasin[1] and Carsten Lanquillon[2]

[1]Heilbronn University of Applied Sciences, Germany, riem.yasin@hs-heilbronn.de
[2]Heilbronn University of Applied Sciences, Germany, carsten.lanquillon@hs-heilbronn.de

Abstract

Research on generative AI in higher education has concentrated on individual tools that students operate on private screens, while a different design question stays open. How should several AI functions act together once they become part of a shared physical learning space? This paper reports a participatory design study in three stages that carries student input from preference to a working prototype. Stage one elicited conceptions of an immersive-cognitive learning environment from seven students through modified-Delphi card sorting. Stage two asked twelve students to distribute a weighting budget across twenty AI functions, where support for understanding code, text summarisation, and speech transcription passed a majority-agreement threshold. Guided by the first two stages, stage three evaluated a Unity prototype that coordinates a GPT-4 conversational guide, image generation, speech recognition, and voice synthesis behind two animated characters on a 98-inch multitouch display with six students. Speech input earned the most positive judgements, whereas five of six confused the input fields of the two generative components, and a haptic wearable added no perceived value despite earlier demand. The paper contributes this finding, a replicable process for prioritising AI functions, and design guidance grounded in a working prototype, open to confirmation at scale.

Keywords: *generative AI, higher education, immersive learning environment, participatory design, conversational agent, gamification*

ISSN 2991-9371

1. Introduction

Generative AI has reached students at scale. During the initial wave of widespread AI adoption (October 2023-February 2024) a global survey collected 23,218 responses from 109 countries and named brainstorming, summarising, and academic writing among the common uses of ChatGPT in higher education [1]. Systematic and meta reviews map a research field that grows quickly yet leans toward individual applications and toward the perspective of institutions rather than learners [2], [3]. Most of this evidence describes single tools on private screens. How AI should behave once it becomes part of a designed physical learning space has drawn far less attention.

Immersive learning environments sharpen the question. The Rensselaer Mandarin Project illustrates such designs: language students converse with AI agents inside a panoramic mixed-reality theatre [4]. Installations of this kind combine several AI functions, among them dialogue, speech recognition, and content generation, into one experience. Research on pedagogical conversational agents, by contrast, still concentrates on single text-based systems [5]. Which functions students actually want in such a space, and what happens once several of them run together, remains largely unexamined.

A second gap concerns who defines the requirements. Co-design work shows that students voice differentiated expectations of generative AI once they act as designers rather than as test subjects [6], and participatory methods improve the relevance of AI-supported systems in education [7]. These co-design studies, though, usually stop at storyboards and low-fidelity prototypes [6]. Requirement elicitation for immersive AI installations, however, rarely starts from the student perspective and even more rarely reaches a working build.

This paper addresses both gaps through a design study at a German university of applied sciences. The guiding research question asks how existing AI technologies can reshape the learning experience of a conventional lecture inside an immersive setting. Three sub-questions structure the work: which forms of AI-based technology suit this purpose, how they can be combined coherently, and whether their use extends the learning experience. From this related work the study derives five design hypotheses (H1 to H5), states them before the empirical work, and revisits each against the findings. Three empirical stages answer them in turn: an elicitation of environment preferences with seven students, a weighted requirement analysis of AI functions with twelve students, and an evaluation of a prototype that coordinates four AI services inside a purpose-built immersive space with six students. The term immersive-cognitive learning environment denotes a physical setting that combines sensory immersion, through display, sound, and haptics, with tasks that ask the learner to reason and self-assess rather than merely watch.

The study makes three contributions. First, it surfaces a legibility problem specific to multi-model installations; once several generative services share one interface, learners lose track of which model handles their input, which undercuts the critical evaluation that AI literacy expects [8]. Second, it demonstrates a replicable three-stage process for deciding which AI functions belong in an immersive learning space, joining card sorting, cumulative voting from requirements engineering [9], and evaluation of a working prototype. Unlike co-design that stops at storyboards [6] or single-agent chatbot studies [5], the work carries student input all

the way to an orchestrated multi-service system tested in a physical room. Third, it shows that students, given a scarce budget, rank assistive functions above generative ones. The samples are small, so these results guide design and invite confirmation at larger scale rather than close the questions they open.

The remainder proceeds as follows. Section 2 states the design hypotheses, Section 3 situates the study in related work, Section 4 describes the three stages, Section 5 the prototype, and Section 6 the results. Section 7 revisits the hypotheses against the findings, Section 8 discusses limitations and future work, and Section 9 concludes.

2. Design hypotheses

Each hypothesis states an expectation and its basis; the three stages then examine it, and the discussion reports how far the findings support it. The first carries the widest reach.

H1: A shared interface for several generative models must make each component legible. Component legibility produced the most frequent failure in stage three, where five of six users lost track of which field feeds which model. AI-literacy frameworks expect learners to judge output against the system that produced it [8], [10], so an interface that blurs component boundaries removes the basis for that judgement. Labels, distinct visual zones, and per-component affordances are the obvious candidates to test first.

H2: Assistive support should precede generative spectacle. Where students spend a scarce budget, functions that help them understand a task should take priority, and generative functions should earn their place only where they serve that task. An installation that leads with impressive generation may invert the priority its users express.

H3: Reliability, latency, and cost behave as design material. Quotas, character limits, per-request pricing, and response times shape the pedagogical experience as directly as any didactic choice. Planning an installation means budgeting these constraints from the start.

H4: Playful elements work as invitations, not decoration. Discoverable rewards can give students a reason to explore a physical space and tie it to the AI functions. This observation aligns with the meta-analytic finding that gamification effects depend on design [11], as the easter eggs engaged students only when their discovery triggered a surprise image.

H5: Stated preferences earn their place only through testing. Elicited preferences select candidates, and only situated testing decides what stays. This asymmetry motivates the staged design itself, since a single elicitation study could commit to a feature that fails in use.

3. Background

3.1 AI in higher education and immersive learning

Roll and Wylie separate an evolutionary path of AI in education, which refines current classroom practice, from a revolutionary path that redesigns the learning situation itself [12]. Survey and review work locates most applications on the evolutionary side, where profiling, prediction, adaptive systems, and assessment support dominate while the perspective of educators and learners stays underrepresented [2], [3]. Immersive learning environments belong to the revolutionary path. Allen et al. show how ambitious such designs can be. The Rensselaer Mandarin Project combines several AI functions in physical space, where a blackbox theatre with a 360-degree panoramic screen, markerless motion tracking, and spatial audio lets students practise a foreign language with AI agents through speech and gesture [4]. Borrowing this ambition at a smaller scale, the present study adds a participatory route to the requirements of such a space.

3.2 Conversational and embodied agents

Pedagogical conversational agents form the most mature strand of AI in education. A survey of 92 studies in higher education organises the field into a conceptual framework of pedagogical purposes and technological functions, and names richer, more personalised interaction as a central direction for future work [5]. Embodiment changes the interaction. Cai et al. join a generative model with augmented reality into an embodied agent and report gains in learners' motivation, inquisitiveness, and critical thinking [13]. Such findings encourage the use of animated characters as the visible surface of an AI system, yet they also raise a question that this study takes up. How does a learner keep track of what stands behind the character once several models contribute to one experience?

3.3 Gamification and playful elements

Game elements promise engagement. Meta-analytic evidence supports a positive overall effect of gamification on academic performance and, in the same breath, shows that the effect depends on design and context [11]. Hidden bonus content, known from games as easter eggs, carries this promise into spatial designs. Whether such elements add value inside an AI-supported learning installation is an empirical question that stage three examines.

3.4 AI literacy, transparency, and responsible use

Frameworks of AI literacy expect learners to evaluate AI output critically and to grasp, at a working level, what a system does with their input [8]. UNESCO extends this expectation into a curricular competency framework that ties AI use to human agency and transparency [10]. A recent meta review of AI in higher education calls for stronger attention to ethics across the field [3]. These demands gain practical weight in multi-component installations. A learner who cannot tell which model processes an input can neither judge the output critically, the ability that AI literacy places at its centre [8], nor easily notice that speech routed to external services leaves the institution's control, a data-handling point Section 4.4 records. Section 7 develops the legibility requirement.

Taken together, prior work supplies mature single agents, evidence for immersive AI settings, and normative frameworks. What it lacks is an account that starts from student preferences, turns them into a prioritised function set, and tests the resulting orchestration in a physical space. The present study supplies such an account.

4. Study design

Participants were students at a German university of applied sciences. Stage one recruited seven master's students, stage two twelve master-level students, and stage three six students at bachelor and master level from several disciplines. The evaluation tasks in stage three drew on two computer-science topics, finite automata and large language models.

The study proceeded in three independently conducted stages, summarised in Table 1. Some participants took part in more than one stage. Each stage addressed a distinct analytical purpose and was evaluated separately. All three stages followed an exploratory qualitative design. Reported counts describe the respective samples and do not represent population estimates.

Stage	Purpose	Instrument	Sample
1	Elicit conceptions of an immersive-cognitive learning environment	Modified-Delphi card sorting with concurrent think aloud, guided questionnaire	7 students (2 female, 5 male)
2	Prioritise AI functions for a learning application	Cumulative voting (20 units across 20 functions in 4 categories), think aloud	12 students (6 female, 6 male)
3	Evaluate prototype and environment	Task-based sessions with questionnaire before and after use; oral discussion	6 students (3 female, 3 male)

Table 1 Overview of the three study stages.

4.1 Stage one: environment preferences

The first stage elicited what students picture when they hear the term immersive-cognitive learning environment. They worked through a modified-Delphi card sort, a variant in which a seed participant builds an initial model and every following participant revises the inherited one. Paul reports that this variant needs fewer participants than open card sorting and yields more useful results, and demonstrates it with a study of eight participants [14]. Six categories structured the card set: environment design, immersive technologies, feedback modality, audio, self-assessment method, and interaction form, each with several options and a free column for additions. Seven master's students (n = 7), whose studies had already exposed them to several lecture formats, took part in individual sessions of 40 to 60 minutes inside a lecture room and verbalised their reasoning under a concurrent think-aloud protocol. A guided questionnaire framed each session with warm-up, model-fixation, and closing questions.

4.2 Stage two: weighted AI-function requirements

To turn preferences into prioritised requirements, the second stage used a purpose-built instrument that presented twenty simplified descriptions of AI functions across four categories: text processing, image processing, audio processing, and data-based prediction. Each of the twelve participants (n = 12) received a fixed budget of twenty units and spread it over at most ten functions, with at least one function per category. This budget forces trade-offs, since whoever wants everything must still decide what matters most. The procedure follows cumulative voting, also called the budget or hundred-dollar test, an established requirements-engineering technique for prioritisation [9]. A function passed the priority threshold when more than half of the cohort, at least seven of twelve, chose it with an identical weight, an agreement heuristic instead of a statistical test. Where no function cleared the threshold in a category, the highest-weighted entry entered as the category representative. Sessions ran individually over a video-conferencing system with think-aloud, and ten were master's students and two were working professionals, all with a completed bachelor's degree.

4.3 Stage three: prototype evaluation

For the third stage, the prototype moved into a purpose-built immersive space. Usability research indicates that small samples surface the majority of usability problems, with five users as a common working figure for formative testing [15]. Six students of both bachelor and master level across disciplines (n = 6) took part, three female and three male. A 98-inch multitouch display, black drapes for visual seclusion, a stool at roughly 1.5 metres, headphones, and a haptic wearable that turns the audible frequency spectrum into vibration formed the physical setup. Figure 1 shows a participant interacting with the prototype in this space. Each session opened with entry questions on AI experience and expectations, continued with the choice of one predefined task, and gave the participant fifteen minutes of unaccompanied interaction. During the interaction, screen recordings documented the sessions. Afterward, a notepad supported

a brief discussion of the interaction and the factual plausibility of the generated content. The sessions concluded with questions on design, interaction, sound, and playful elements, followed by open feedback, and lasted 50 to 60 minutes. To reduce load-related delays in the AI services, all sessions took place in the evening.



Figure 1. Setup and interaction with the prototype in the immersive space.

4.4 Ethics and data protection

Participation in all three stages was voluntary and based on informed consent. Before each session, participants received information about the study purpose, procedure, and data handling. They were informed that the evaluation focused on the design artefacts, not on their individual performance. No directly identifying data were stored, and all research data were pseudonymised. Consent covered the collection and analysis of questionnaire responses, screen recordings, and audio and video data.

5. Prototype

5.1 Architecture

The prototype, developed in the Unity engine, orchestrates four distinct AI services behind a unified user interface. Figure 2 traces one full interaction and shows how the four services hand work to one another. Student speech enters through the microphone, and the speech-recognition service (Whisper) transcribes it to text. The transcript reaches the dialogue guide (ChatGPT-4), which answers in the persona of the on-screen character. A task code sets the guide's per-task prompt, so the same character adapts to the topic the student selected. The speech-synthesis service (ElevenLabs) then voices the answer, and the display presents both the spoken reply and the animated 2D character, allowing users to choose between a male and a female avatar. A separate path drives the visuals. Room codes, hidden in the space as easter eggs, select predefined prompts for the image model (DALL·E 3), so that finding a hidden code triggers a generated image on the display. Since all four services meet behind one interface, the learner faces a single surface while several models act underneath, which is where the component-legibility problem observed in stage three arises.

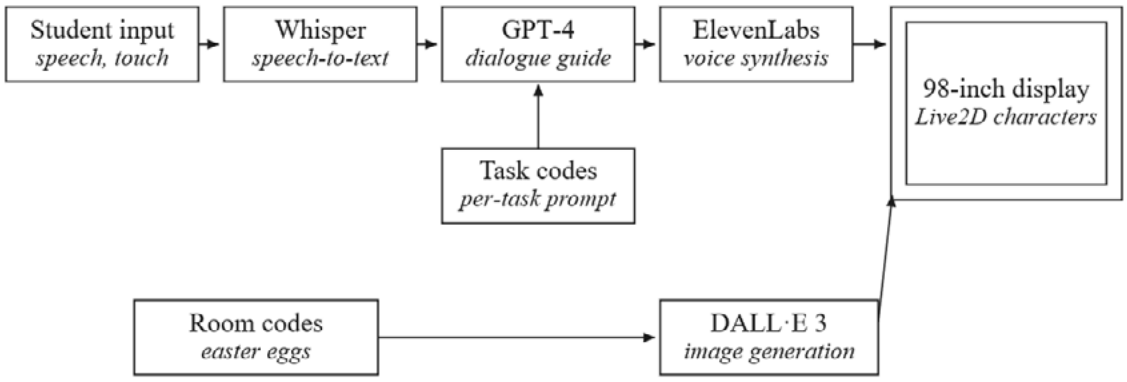


Figure 2. Orchestration of the four AI services.

5.2 Didactic design

A system prompt configures the conversational model as a courteous guide that avoids ready-made solutions, assesses learners' prerequisite knowledge, provides concise responses, supports German and English, and uses informal forms of address. Prior to the interaction, each student selected one of six prepared short tasks, comprising four tasks on finite automata and two on the fundamentals of large language models. This selection gave the session a defined purpose. A topic-specific code word, provided to the student in advance, initiated the session and indicated the subject to the guide. Consequently, the guide could direct the discovery dialogue from the first turn. In addition, separate code words were hidden throughout the room as easter eggs. Discovering one triggered a predefined image-generation prompt that served as a gamification element inspired by cheat systems in games. Section 8 revisits the associated risk.

6. Results

6.1 Stage one: a calm, story-driven, multi-modal space

All seven chose a bright and calm setting that tells a story through its design, and asked for hidden bonus content as easter eggs. Content-matched decoration and warm light followed, each with five of seven. Among immersive technologies, AI, mixed reality, and augmented reality drew all seven votes, whereas virtual reality drew a single one, several times with a reference to motion sickness. Visual and auditory feedback were unanimous choices, and haptic feedback reached six of seven. Quiet instrumental music led the audio category. Preferences for self-assessment scattered across runtime checks, multiple-choice tests, oral examination, written submission, and presentation, so no single method satisfied the cohort. Multitouch topped the interaction category with seven votes, ahead of robots as interaction elements and voice commands with six each. Five of the seven added a wish to learn together with others in such a space. Eight of ten prior expectations held. The two that failed were that everyone would want all three feedback modalities, and that fewer than half would care about the easter eggs.

6.2 Stage two: assistive functions outrank generative spectacle

Table 2 lists the prioritised functions. Support for writing and understanding code drew the highest aggregate weight, 43 units from eight students, and those who picked it framed its value as understanding, since a learning application should convey how programming works. Summarisation of long texts followed with 31 units from eleven students. Speech transcription drew a selection from all twelve and cleared the threshold as well. In image processing no function reached majority agreement, so object recognition entered as the category representative with the highest weight. The prediction category shows a revealing pattern. Personalised recommendation drew the highest raw weight there, 29 units from eight

students, yet the cohort never converged on an identical weight, so it missed the threshold and stayed out of the prioritised set. Functions with creative appeal but unclear learning value, among them music or avatar generation, drew minimal weight. Several students remarked that live demonstrations of the functions would have sharpened their choices, a limitation Section 8 takes up.

Function	Category	Aggregate weight	Selections	Basis
Write and understand code from text input	Text	43	8	Majority-agreement threshold
Compress long texts into summaries	Text	31	11	Majority-agreement threshold
Convert spoken words into text	Audio	18	12	Majority-agreement threshold
Recognise specific objects in images	Image	10	6	Category representative

Table 2. Prioritised AI functions after stage two (n = 12, budget of 20 units per participant).

6.3 Stage three: reliability and legibility shape the experience

All six asked for a full or partial redesign of the classic lecture, which echoes the strongest signal of stage one in a separate sample. Speech transcription, the function every stage-two student had selected, ran without a single failure and drew the most positive comments, and the text stayed readable throughout. Five of six judged the anime-style characters as fitting for a learning application, and five reported general interest in AI. The clearest usability failure lay in component legibility. Five of six confused the input fields of the conversational and the image model during the session. Four of the six enjoyed hunting the hidden code words in the room, which confirmed the easter-egg expectation from stage one. Four preferred a dark interface mode. Worn by everyone, the haptic wearable convinced nobody, and four of six saw no added value for a learning application, although six of seven votes had favoured haptic feedback in stage one. Students asked for more visual feedback during waiting times. A recording window that proved too short for voice input rose to twenty seconds afterward, a value aligned with reported voice-message lengths. Operating conditions shaped the experience in material ways. Throttling of the image service forced a rebuild, the character limit of the voice service required a paid extension, per-request costs accumulated with use, and the switch to GPT-4 shortly before the evaluation visibly improved adherence to the guide persona and the code-word protocol. Of five prior expectations, three held, namely easter eggs, instrumental music, and demand for more visual feedback, and two failed, namely dissatisfaction with the recording window and image generation as the preferred entry point.

7. Discussion: revisiting the design hypotheses

The findings provide preliminary support for all five design hypotheses, although the strength of the evidence varies across them.

H1 (component legibility) receives the clearest support. Confusion between the dialogue and image-generation fields indicates that a unified interface can obscure the roles of individual models. Component legibility therefore emerges as a prerequisite for critically evaluating outputs in multi-model learning environments.

H2 (assistive before generative) is reflected in stage two, where students favoured functions that supported comprehension and information processing over generative features with less explicit learning value. This pattern suggests that perceived pedagogical utility outweighed technological novelty.

H3 (reliability, latency, and cost as design material) is borne out in stage three, where immersion broke under service throttling and waiting times, although speech recognition performed flawlessly.

H4 (playful elements as invitation) received partial support, as four of the six participants enjoyed searching for the hidden code words, particularly when each discovery triggered a surprise image.

H5 (stated preferences require situated testing) is illustrated by the haptic wearable, which won six of seven votes in stage one yet lost four of six users in the concrete experience. Two further prior expectations failed, namely dissatisfaction with the recording window and image generation as the preferred entry point.

8. Limitations and future work

The findings reflect a specific study context of small, self-selected samples from one German university of applied sciences and tasks from two computer-science topics. They therefore offer design-oriented insights for comparable settings, and future work can test their transferability across institutions, disciplines, and larger samples.

Two procedural factors invite caution. The author was present during the card sorting and think-aloud sessions, which can encourage socially desirable answers, and the modified-Delphi seed model may anchor later participants to earlier choices. Stage three captured one fifteen-minute session per participant and assessed experience and usability; establishing a pedagogical effect on learning therefore remains the natural next step. The study did not systematically assess the factual accuracy of the guide, and since a model can state incorrect explanations with unwarranted confidence, a classroom version should add a factual-accuracy check that future work can measure.

In stage two, textual descriptions stood in for live demonstrations, which may have compressed the perceived difference between familiar and unfamiliar capabilities. The prototype reflects the models available at build time; current multimodal models fold several of these functions into one system and reduce orchestration overhead, yet constraints of legibility, cost, service quotas, and data protection persist across model generations, and the participatory process stays applicable to different technical architectures. Promising directions for future work include testing the design hypotheses with larger and more diverse samples, adding pre- and post-measures of learning, comparing interfaces that differ only in component legibility, and examining the collaborative use requested by stage-one students.

9. Conclusion

The study combined three complementary stages. It asked students what a learning space should feel like, made them prioritise the AI functions inside it, then built and tested the result. Each step returned something the others could not. Elicitation produced a coherent target, a calm and story-driven space with visible and audible AI. Prioritisation revealed a preference for assistive over generative functions and a wish for personalisation that never settled into agreement. Evaluation exposed what neither earlier stage could show. Its firmest signal is clear and actionable. Once several generative models meet in one interface, legibility, not novelty, decides whether students can still evaluate what they see. The three stages deliver a working prototype, this finding, and a repeatable path from student voice to a design that later studies can build on.

For educators and designers who plan AI-supported immersive learning spaces, the findings suggest five concrete steps. Give each generative component a distinct label, visual zone, and input affordance, so that learners can tell which model handles their input. Lead with assistive functions such as summarisation and transcription, and add generative elements only where they serve the task. Budget for latency, quotas, character limits, and per-request cost from the start. Use playful discovery, such as a hidden prompt that triggers a meaningful response, as an invitation to explore. Finally, treat stated preferences as candidates and confirm them through situated testing before committing hardware such as haptic wearables.

10. Acknowledgements

This research was supported by the EUonAIR project, no. 101177370, ERASMUS-EDU-2024-EUR-UNIV-1, within the framework of the EUonAIR Centre of Excellence in Responsible AI in Education (CORE.AI), co-funded by the European Union.

11. Data availability

The anonymised questionnaires, coded results, and analysis files supporting the findings are available from the authors on reasonable request. The prototype includes sample content owned and copyrighted by Live2D Inc.

References

[1] D. Ravšelj et al., "Higher education students' perceptions of ChatGPT: A global study of early reactions", *PLOS ONE*, Volume 20, Issue 2, e0315011, 2025.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education - where are the educators?", *International Journal of Educational Technology in Higher Education*, Volume 16, Article 39, 2019.

[3] M. Bond et al., "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour", *International Journal of Educational Technology in Higher Education*, Volume 21, Article 4, 2024.

[4] D. Allen et al., "The Rensselaer Mandarin Project - a cognitive and immersive language learning environment", *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, Issue 1, pg. 9845-9846, 2019.

[5] H. Yusuf, A. Money and D. Daylamani-Zad, "Pedagogical AI conversational agents in higher education: a conceptual framework and survey of the state of the art", *Educational Technology Research and Development*, Volume 73, Issue 2, pg. 815-874, 2025.

[6] P. Prasad, R. Balse and D. Balchandani, "Exploring multimodal generative AI for education through co-design workshops with students", *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pg. 1-17, 2025.

[7] R. Alfredo et al., "Co-designing AI-powered learning analytics: bringing students and teachers together", *International Journal of Educational Technology in Higher Education*, Volume 22, Issue 1, Article 78, 2025.

[8] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pg. 1-16, 2020.

[9] P. Berander and A. Andrews, "Requirements prioritization", *Engineering and Managing Software Requirements*, A. Aurum and C. Wohlin (Eds.), Springer, Berlin, pg. 69-94, 2005.

[10] UNESCO, *AI competency framework for students*, UNESCO Publishing, Paris, 2024.

[11] J. Zeng, D. Sun, C.-K. Looi and A. C. W. Fan, "Exploring the impact of gamification on students' academic performance: A comprehensive meta-analysis of studies from the year 2008 to 2023", *British Journal of Educational Technology*, Volume 55, Issue 6, pg. 2478-2502, 2024.

[12] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education", *International Journal of Artificial Intelligence in Education*, Volume 26, Issue 2, pg. 582-599, 2016.

[13] S. Cai, H. Huang, X. Jiao and Y. Zhang, "Generative AI + AR: a novel embodied conversational agent in an interactive learning environment", *Interactive Learning Environments*, advance online publication, pg. 1-15, 2025.

[14] C. L. Paul, "A modified Delphi approach to a new card sorting methodology", *Journal of Usability Studies*, Volume 4, Issue 1, pg. 7-30, 2008.

[15] R. AlRoobaea and P. J. Mayhew, "How many participants are really enough for usability studies?", *Proceedings of the 2014 Science and Information Conference*, pg. 48-56, 2014.

THE USE OF AI TOOLS

During the preparation of this manuscript, generative AI tools were used solely to improve the language, clarity, and readability of the text, under the full oversight and responsibility of the authors. All figures and tables are the authors' own work. Figure 1 combines two photographs of the prototype in use, one taken professionally and one with a smartphone. No figure was generated by AI.