CTE Workshop Proceedings, 2026, Vol. **13**, pp. 256–270 10.55056/cte.1415

Designing the content component of immersive blended science learning: a CAMIL-grounded framework refined and pre-validated by a simulated multi-model Delphi panel

Oleksandra M. Sokolyuk

Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine

Abstract. Immersive technologies – virtual reality (VR), augmented reality (AR) and 360° video – are spreading through school science, yet teachers lack research-informed criteria for deciding *what* content to deliver immersively and how to embed it in blended lessons. Building on a prior conceptual analysis, this paper develops the content component of an immersive-learning methodology into an explicit design framework grounded in the Cognitive Affective Model of Immersive Learning (CAMIL). The framework was refined and stress-tested with a novel instrument: a *simulated Delphi panel* of fourteen large language models (LLMs) from eight model families and three providers, each adopting an expert persona, across three rounds (generative development, rating, and re-rating after anonymised feedback). The panel expanded the framework from eleven to fourteen content-design criteria and a VR/AR/360° modality-fit mapping. We report the exercise transparently, including its limits. Relevance ratings sat at a ceiling – every item was endorsed – so the content-validity indices are non-discriminating, and the panel, which also generated the items it rated, offers *refinement and pre-validation*, not independent validation. The informative signal lay in feasibility, where the panel discriminated sharply and, on re-rating, became markedly *more* pessimistic: full VR, differentiation and teacher orchestration were judged least feasible in real classrooms. We contribute (i) the refined framework and (ii) the simulated multi-model Delphi as a fast, fully logged but explicitly *synthetic* pre-validation method, whose affordances and limits for educational design research we analyse. No human participants or classroom data are involved; the LLM panel complements rather than replaces human expertise.

Keywords: immersive technologies, blended learning, science education, content design, CAMIL, virtual reality, augmented reality, simulated Delphi, large language models

1. Introduction

The digital transformation of education, characterised by the spread of virtual, augmented and mixed reality, is reshaping how science is taught and learned. Immersive technologies have repeatedly been named among the key educational technologies of the coming decade, and systematic reviews now document a decade of empirical work on VR and AR across K–12 and higher education [4, 18, 20]. At the same time, blended learning – the deliberate combination of online and face-to-face study – has moved from an option to a necessity, a shift felt acutely in Ukraine, where distance and blended formats have kept schooling running through prolonged disruption. The convergence of these two trends raises a practical question that the present paper addresses: when a science teacher plans a blended lesson, *what* content should be delivered immersively, in which modality, and how should it be sequenced and assessed?

Much of the immersive-learning literature answers an adjacent but different question. It studies hardware, presence, and the comparative effectiveness of VR or AR as interventions [7, 12, 16]. Far less attention has been paid to the *content* decision – the selection, didactic transformation and sequencing of curricular material – even though this is precisely where a teacher's pedagogical content knowledge is exercised and where immersion most easily degrades into edutainment. A

0000-0002-5963-760X (O. M. Sokolyuk)
sokolyuk@iitlt.gov.ua (O. M. Sokolyuk)

Received 2026-02-17 | Accepted 2026-03-20 | Published 2026-03-21 | Version of record 2026-03-21

recent conceptual analysis of the "content aspect" of an immersive-learning methodology [22] argued that the content component is the pivotal one, but left it described rather than operationalised: the components were enumerated, not turned into criteria a teacher or designer could apply and a researcher could test.

This paper takes up that gap. It develops the content component into an explicit, operationalised design framework and then subjects the framework to a form of validation. Because the study is deliberately *conceptual* – it involves no classroom intervention and no human participants – we validate the framework not with student outcomes but with expert judgement of its content validity. The novelty is in how that judgement is obtained: through a *simulated Delphi panel* composed of large language models. Recent work has asked whether LLMs can stand in for human respondents in social-science research, with both encouraging demonstrations [2] and pointed warnings [6, 8, 24]. We adopt the method in full view of those warnings: the panel is a transparent, reproducible *synthetic pre-validation*, not a substitute for a human-expert Delphi or classroom evidence.

This study relates directly to, and extends, the earlier Ukrainian-language article [22], which it cites as its conceptual basis. It is not a translation of that work: it adds (i) a formalised, operationalised criteria framework, (ii) a simulated multi-model Delphi refinement and pre-validation, (iii) a reflexive methodological contribution on using LLMs as a synthetic expert panel, and (iv) corrected conceptual definitions and an updated international literature base. Our research questions are:

RQ1. What content-design criteria for immersive blended secondary-science learning follow from CAMIL and the recent literature?

RQ2. To what extent does a heterogeneous panel of LLM "experts" judge these criteria, and an accompanying VR/AR/360° modality-fit mapping, relevant, clear and feasible, and where do their judgements converge or diverge across three Delphi rounds?

RQ3. What are the affordances and limitations of a multi-model LLM panel as a *simulated* Delphi for validating educational design frameworks?

2. Theoretical framework

2.1. Immersive technologies and the reality–virtuality continuum

Precise terminology matters, because instructional affordances differ by modality. Following the reality–virtuality continuum [13], *virtual reality* replaces the physical environment with a computer-generated one, usually experienced through a head-mounted display; *augmented reality* overlays digital information on the physical environment in real time, typically through a phone, tablet or glasses; and *mixed reality* denotes environments in which physical and digital objects are spatially registered and interact – it is *not* merely "an analogue of augmented reality." *Extended reality* is an umbrella term spanning these, together with 360° video, which provides spatial viewing with limited learner agency and should not be equated with fully interactive VR. The three modalities most used for instruction – VR, AR and 360° video – differ in immersion depth, agency, infrastructure and cognitive-load profile, and these differences are what a content framework must exploit.

As a pedagogical tool, immersive technology is simultaneously an *educational technology* – a system of actions realising learning goals – and a *means of learning*, an interactive instrument that models objects and phenomena. By reproducing the appearance and behaviour of real objects through information modelling, it reinforces the principle of visualisation, giving the learner a near-real personal experience of perception and action (figure 1).

2.2. The Cognitive Affective Model of Immersive Learning

Our theoretical anchor is the Cognitive Affective Model of Immersive Learning (CAMIL) [11], which explains *why* immersive media can support learning. CAMIL distinguishes *technological immersion* –

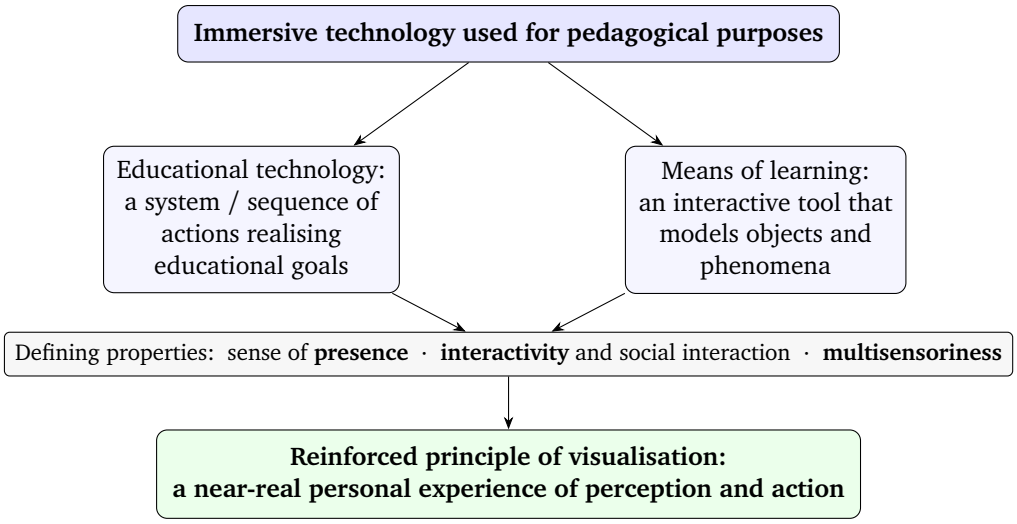

Figure 1: Immersive technology in education is simultaneously an educational technology (a system of actions realising learning goals) and an interactive means of learning. Its defining properties – presence, interactivity and multisensoriness – reinforce the principle of visualisation by giving the learner a near-real personal experience. Redrawn and developed after Sokolyuk [22].

an objective property of the system – from *presence*, the learner's subjective sense of "being there," and *agency*, the sense of control over events. These psychological states are shaped by media affordances (representational fidelity, immersion, interaction) and, in turn, drive a set of affective and cognitive factors – interest, intrinsic motivation, self-efficacy, embodiment, cognitive load and self-regulation – that predict factual, conceptual and procedural learning outcomes (figure 2). Two implications guide our framework. First, immersion is not self-justifying, and CAMIL's predictions are not uniformly borne out: adding immersive VR to a science simulation has been found to raise presence yet *reduce* learning [12]. Because presence can raise extraneous cognitive load for novices, content must be designed to direct working memory to the science rather than the interface. Second, the same affordances that raise presence can isolate learners socially, so collaboration must be designed in rather than assumed – a concern CAMIL itself does not centrally address.

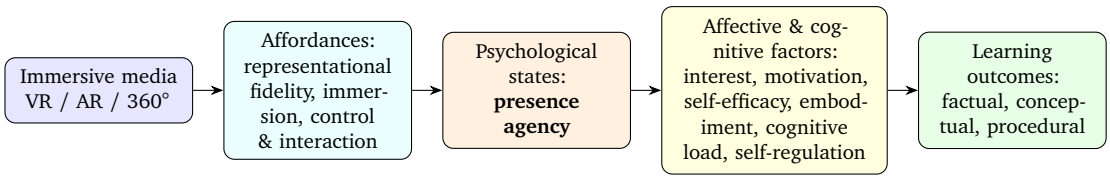

Figure 2: The Cognitive Affective Model of Immersive Learning: immersive media afford fidelity, immersion and interaction, which raise psychological *presence* and *agency*; these drive affective and cognitive factors that, in turn, predict learning outcomes. Redrawn and adapted from Makransky and Petersen [11]; technological immersion (a property of the system) and presence (the learner's experience) are kept distinct.

2.3. Blended learning and a content-design lens

Blended learning is more than the co-presence of online and offline study; it is the orchestration of activity across phases – typically pre-class preparation, in-class work and post-class consolidation [3, 14]. Immersion can sit in any phase, but its role differs: a low-agency 360° preview suits pre-class orientation, whereas a high-agency VR or AR inquiry belongs in class. The design problem is therefore not "should we use VR?" but "which content, in which modality, in which phase, to what end?" – a question of pedagogical content knowledge. Cognitive load theory [23] and the multimedia-learning tradition [15] supply the design discipline: segment, signal, and scaffold so that the limited capacity

of working memory is spent on the science, not on navigating the interface. The framework in section 3 operationalises exactly these commitments for the content component.

3. The content-component framework

A methodology for using immersive technologies in blended school science can be described through six components: theoretical–methodological, target (goal-setting), content (subject), technological, procedural (organisational–methodical) and diagnostic (evaluative) [22]. The present paper zooms in on the *content* component, which determines which concepts, phenomena and processes are worth presenting immersively, and how they are transformed for an immersive medium. We operationalise it as a set of content-design criteria, each carrying five facets: a definition, a link to the CAMIL mechanism it serves, a guiding question for the teacher, a science example, and an observable quality indicator.

The initial set comprised eleven criteria derived from [22] and the literature of section 2. After the Delphi process described in section 4, the refined framework comprises fourteen criteria (table 1). Conceptually they fall along a design pipeline (figure 3): a curriculum content element (C1–C3) is matched to an immersive affordance (C4, C7), which engages a learning mechanism (C5, C6, C9, C10 – CAMIL for C5/C6, and collaborative-learning and adaptive-design theory for C9/C10), is placed in a blended phase (C8) and yields assessment evidence (C11); three further criteria (C12–C14) are cross-cutting feasibility-and-equity gates – enabling preconditions rather than CAMIL mechanisms – that any design must pass.

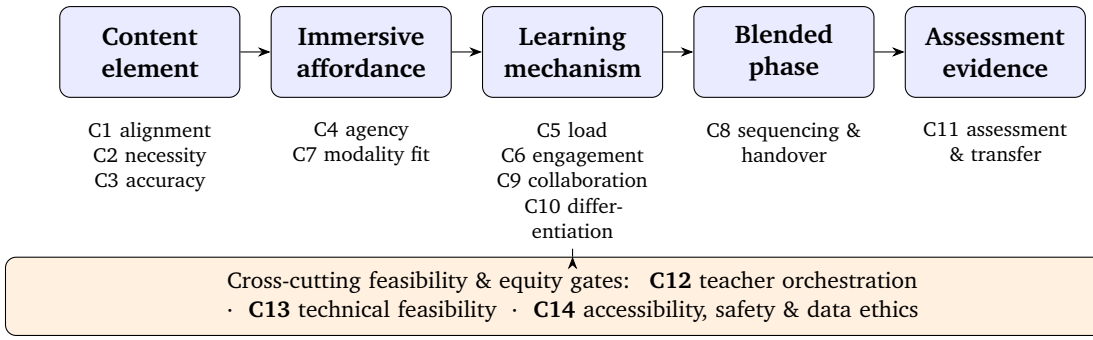

Figure 3: The framework as a design pipeline. A curriculum content element is matched to an immersive affordance, which engages a learning mechanism (CAMIL and, for C9/C10, complementary theory), is placed in a blended phase, and yields assessment evidence. Criteria C1–C11 attach to the five stages; the three feasibility-and-equity criteria (C12–C14) are cross-cutting gates – enabling preconditions, not CAMIL mechanisms – that any design must pass.

Alongside the criteria, the framework includes a *modality-fit mapping* (table 2) that matches the immersion level to the pedagogical goal, and a blended-phase sequencing scheme distinguishing the role of immersive content before, during and after class. Mixed reality, defined in section 2, is excluded from the mapping: it remains rare in schools, so we restrict the framework to the three modalities most used for instruction. Both the mapping and the sequencing scheme were rated by the panel as distinct instruments (rows M1–M3 and S1–S3 in section 5).

4. Method: the simulated Delphi panel

4.1. Rationale and stance

The Delphi technique builds expert consensus through iterative, anonymous rounds with controlled feedback, and is a standard way to validate frameworks and develop content-validity evidence [5, 9]. Recruiting and retaining a large, diverse human expert panel is, however, slow and costly, and this study is explicitly conceptual. We therefore use a *simulated* Delphi: the panelists are large

Table 1
The refined CAMIL-grounded content-design framework: 14 criteria for selecting, transforming and sequencing immersive content in blended secondary science. All criteria nominally met the consensus rule (I-CVI ≥ 0.78, IQR ≤ 1) in both rounds, but on relevance this is a degenerate ceiling (section 5), not evidence of validity. Criteria marked † were flagged by the panel as *overloaded* (candidates for splitting in a future human-expert Delphi); those marked ‡ carry the lowest feasibility ratings.

ID	Criterion	CAMIL link or enabling rationale	Quality indicator
C1	Curriculum alignment	Pre-condition for all CAMIL learning outcomes	Activity traceable to a numbered curriculum objective
C2	Representational necessity	Externalising an invisible phenomenon can reduce visualisation load (immersion may add load – see C5)	Documented rationale for why immersion (not a diagram) is required
C3†	Scientific accuracy and epistemic fidelity	Accurate fidelity is a precondition for valid conceptual learning	Content checked against curricular science and known misconceptions
C4	Learner agency and manipulability	Agency (a central CAMIL factor) supports self-efficacy and active processing	At least one student-controllable variable with visible feedback
C5	Cognitive-load management and scaffolding	Manages extraneous load; scaffolds redirect working memory to schema construction	Explicit segments and a scaffold-fading plan
C6	Affective and motivational engagement	Interest, affect and motivation deepen processing	Opens with a question/anomaly and an authentic purpose
C7	Modality fit	Immersion levels afford different presence/agency profiles	Chosen modality justified against the alternatives
C8	Blended-phase sequencing and handover	Self-regulation and transfer depend on what surrounds immersion	Each activity in a named phase with a handover artefact
C9	Social and collaborative learning	Social/collaborative learning (beyond CAMIL's scope); offsets VR isolation	At least one structured peer/collaborative interaction
C10†‡	Differentiation and personalisation	Learner characteristics moderate CAMIL pathways	One adaptation plus a reflection/self-monitoring prompt
C11†	Assessment integration and transfer	Closes the loop on CAMIL learning outcomes	Each activity has an aligned assessment, including a transfer task
C12‡	Teacher orchestration and readiness	Teacher orchestration mediates whether CAMIL mechanisms are realised	Stated prep time, required competencies and an orchestration plan
C13	Technical and logistical feasibility	Reliable delivery is a precondition for any mechanism	Device/throughput plan, session-length fit and a contingency
C14†	Accessibility, equity, safety and data ethics	Safeguards the conditions for all learners	A non-immersive fallback, a safety note and a data-privacy check

language models. This choice is deliberate and bounded. On the one hand, LLMs can approximate aspects of human survey responses [2] and show non-trivial agreement with human evaluators [25]; a heterogeneous, multi-provider panel is fast, fully logged and re-runnable. On the other hand, LLMs are not experts, can flatten the diversity of human viewpoints [24], and substituting them for

Table 2
The modality-fit mapping: matching the immersion level to the pedagogical goal. Median panel feasibility (Round 2 / Round 3) is shown in the last row; 360° video (M3) drew the lowest relevance endorsement intensity (Aiken's $V = 0.74$), and full VR (M1) the lowest feasibility.

	M1: Virtual reality	**M2: Augmented reality**	**M3: 360° video**
Immersion depth	Full (place illusion)	Partial (augmented real space)	Partial (spatial viewing)
Device	Head-mounted display	Phone, tablet, PC, AR glasses	Phone/cardboard, tablet, PC
Interaction	Embodied, 6-DoF manipulation	Overlay anchored to real objects	Look-around; little manipulation
Agency	High	Medium	Low
Infrastructure / cost	Headsets, space; higher cost	Mostly existing devices; lower	Minimal; very low
Cognitive-load risk	Higher for novices	Can lower load, but split attention is a risk	Low (guided viewing)
Suitable science tasks	Field immersion, virtual labs, hazardous experiments	Overlaying invisible quantities on real apparatus	Virtual site visits, orientation, previews
Blended phase	Mainly in-class	In-class and post-class	Pre-class (and replay)
Key limitations	Cost, cybersickness, throughput, isolation	Tracking stability, screen-mediated	Low agency; not a VR substitute
Median feasibility (R2 / R3)	2 / 2	4 / 3	4 / 4

participants invites well-catalogued fallacies [6, 8]. We therefore treat the exercise as a *synthetic pre-validation* whose purpose is to stress-test and refine the framework before any human Delphi or classroom study – never as evidence of classroom effectiveness. This stance is itself the object of RQ3.

4.2. Panel

The panel comprised fourteen LLM panelists spanning eight model families (Anthropic Claude, OpenAI gpt-oss, Google Gemma, Nvidia Nemotron, Alibaba Qwen, MiniMax, DeepSeek and Z.ai GLM) and three access channels: the Anthropic API, the Ollama cloud, and an OpenAI-compatible gateway (table 3). Each panelist was assigned one of seven expert personas – secondary science teacher, educational-technology researcher, curriculum specialist, learning scientist, VR/AR developer, assessment specialist and inclusive-education specialist – with two panelists per persona, so that the expertise a real panel would need was represented. Decoding temperature (the sampling-randomness parameter) was fixed at 0.3, with a seed where the channel supported it. Four Claude, three MiniMax and two GLM endpoints appear across versions or gateways; such same-family endpoints are *not* statistically independent, which we acknowledge but do not correct for (section 5). Five of the eight families are represented by a single endpoint each, so "family" here denotes model lineage, not a sample of models.

4.3. Procedure

The panel completed three rounds (figure 4). In **Round 1** (generative), each panelist, in persona, critiqued the eleven-criterion draft and proposed additions, merges, deletions and rewrites, and named missing constructs. We consolidated the proposals into a revised instrument, applying a change supported by at least one third of the panel (an integer threshold of $\geq 5/14$); a few smaller

Table 3
The simulated expert panel: 14 large-language-model panelists spanning eight model families and three access channels, each assigned one of seven expert personas (two panelists per persona).

#	Model family	Version	Access channel	Expert persona
1	Claude	Opus 4.8	Anthropic API	Learning scientist
2	Claude	Sonnet 4.6	Anthropic API	Ed-tech researcher
3	Claude	Haiku 4.5	Anthropic API	Assessment specialist
4	Claude	Sonnet 4.5	OmniRoute	Curriculum specialist
5	GPT-OSS	gpt-oss 120B	Ollama	Science teacher
6	Gemma	Gemma 4 31B	Ollama	Inclusive-ed specialist
7	Nemotron	Nemotron 3 Super	Ollama	VR/AR developer
8	Qwen	Qwen3 Coder Next	Ollama	Ed-tech researcher
9	MiniMax	MiniMax M3	Ollama	Science teacher
10	MiniMax	MiniMax M2.5	OmniRoute	Curriculum specialist
11	GLM	GLM 4.7	Ollama	Learning scientist
12	GLM	GLM 5	OmniRoute	Assessment specialist
13	DeepSeek	DeepSeek 3.2	OmniRoute	VR/AR developer
14	MiniMax	MiniMax M2.1	OmniRoute	Inclusive-ed specialist

strengthenings – folding scaffolding, transfer and metacognition into existing criteria – were author decisions below that threshold rather than panel consensus, and we label them as such in section 5. In **Round 2** (rating), each panelist rated every criterion and every modality and sequencing row on three four-point scales – relevance, clarity and feasibility (1 = not, 4 = highly). In **Round 3** (re-rating), each panelist received, for every item, the anonymised group median, inter-quartile range and percentage endorsement from Round 2, together with a digest of dissenting comments, and then re-rated. System prompts explicitly licensed critical, discriminating judgement and discouraged agreement for its own sake. All prompts, raw responses and code are archived (section 8).

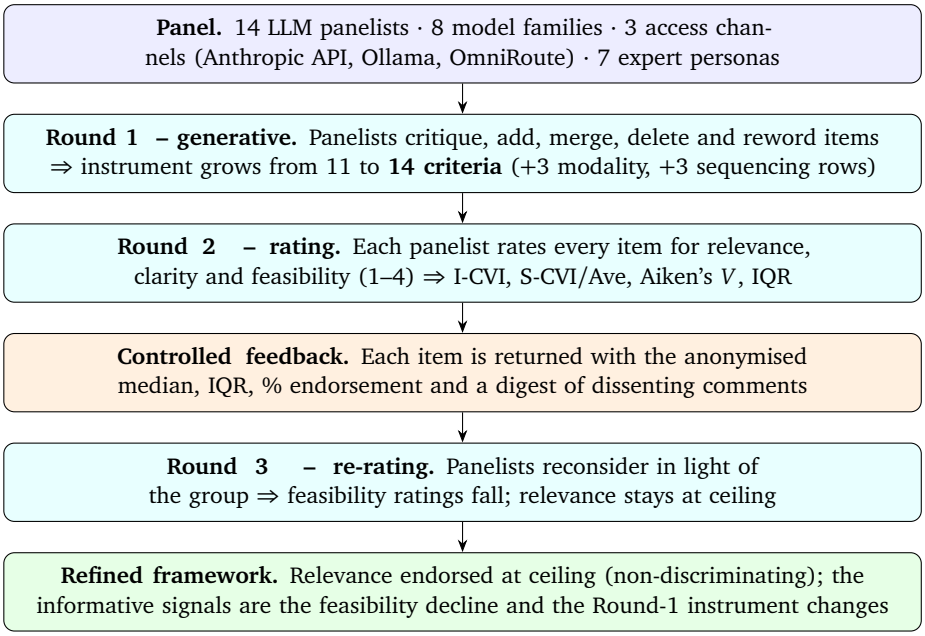

Figure 4: The three-round simulated Delphi. A heterogeneous panel of large language models first co-develops the instrument (Round 1), then rates it (Round 2), receives anonymised group feedback, and re-rates it (Round 3). The relevance dimension was endorsed at ceiling and is non-discriminating; the informative signals are the feasibility decline on re-rating and the Round-1 instrument changes.

4.4. Analysis

For each item we computed the item-level content validity index (I-CVI) – the proportion of panelists rating relevance ≥ 3 – the modified κ (chance-corrected I-CVI), and the scale-level S-CVI/Ave (the mean of the I-CVIs) [9, 17]. To capture the *intensity* of endorsement beyond the dichotomy underlying I-CVI, we also computed Aiken's V [1], with the median and inter-quartile range (IQR) of each dimension. Consensus was defined a priori as I-CVI ≥ 0.78 [9] with IQR ≤ 1, and convergence as the change in V and IQR from Round 2 to Round 3. We flag at the outset that this apparatus proved largely degenerate: because relevance ratings sat at ceiling, I-CVI, κ and S-CVI/Ave are all 1.00 and carry no validation weight, and the Lynn threshold in any case presumes ten or more *independent* experts, which our correlated endpoints do not supply. We therefore lean the analysis on the feasibility dimension, where ratings varied. All statistics were implemented as small, unit-tested pure functions.

4.5. Reproducibility and threats to validity

The study is *fully logged and re-runnable in principle* rather than bit-for-bit reproducible: every prompt and raw response is stored and the analysis is scripted end-to-end, but the endpoints are non-deterministic. A temperature of 0.3 and a seed were recorded for all panelists, yet only the Ollama channel honours the seed – for the Anthropic and gateway channels it is inert – and the run was single-shot ($n = 1$), so rating volatility across repeat runs is unquantified. Two further threats matter and are revisited in section 7. Same-family endpoints produce *correlated* judgements that inflate apparent agreement, which we acknowledge but do not statistically correct. And LLMs are prone to *agreeableness* – a tendency to endorse whatever is presented; we used anti-agreeableness prompting, but, as the relevance ceiling shows, it did not work on that dimension, so our positive reading rests only on feasibility and on the Round-1 changes. (The gateway endpoint uses a local development credential, not a public service, so that channel is not externally reproducible.)

5. Results

5.1. Round 1: from eleven to fourteen criteria

All fourteen panelists returned usable structured responses in every round. Round 1 produced 43 addition proposals, 20 merge proposals, 6 deletion proposals and 46 notes on missing constructs. The convergent signal was strong. Thirteen of fourteen panelists proposed merging the affective and motivational criteria, which we combined into a single criterion (C6). Four new criteria crossed the one-third threshold: *teacher orchestration and readiness* (proposed in some form by roughly thirteen panelists), *social and collaborative learning* (≈ 11), *technical and logistical feasibility* (≈ 7) and *scientific accuracy and epistemic fidelity* (≈ 5). Data-privacy and health concerns (raised by about six panelists) were folded into an expanded accessibility criterion (C14); scaffolding was folded into the cognitive-load criterion (C5), and transfer into sequencing and assessment. No item reached the deletion threshold. The four additions crossed the integer threshold ($\geq 5/14$); the foldings of scaffolding, transfer and metacognition fell below it and were editorial decisions rather than panel consensus. We also note that merging the affective and motivational criteria, though endorsed by thirteen panelists, collapses two factors that CAMIL keeps distinct (interest/positive affect and intrinsic motivation) – a panel-driven simplification a finer framework might resist. The instrument thus grew from eleven to fourteen criteria, alongside the three modality and three sequencing rows.

5.2. Rounds 2–3: a relevance ceiling, and a feasibility decline

Every item met the consensus rule in both rounds, and S-CVI/Ave was 1.00 in each round – but this is a degenerate result, not a finding. Relevance never fell below 3 for any panelist on any item, so every I-CVI, every modified κ, and therefore S-CVI/Ave are pinned at 1.00 by construction (table 4);

the IQR component of the consensus rule is likewise trivially satisfied (16 of 20 items had relevance IQR = 0 in both rounds). The relevance ceiling is partly genuine – the criteria are sensible – but partly an artefact of LLM agreeableness that the anti-agreeableness prompting did not break. We therefore draw no validation weight from the relevance-based indices, and read the informative signal from feasibility and from the Round-1 instrument changes (table 4, figure 5).

Table 4
Simulated-Delphi results (14 panelists). The relevance-based content-validity apparatus is *degenerate*: every item has I-CVI = 1.00 (and modified $\kappa = 1.00$) in both rounds, so S-CVI/Ave = 1.00 is forced and the relevance dimension does not discriminate. The informative signal is feasibility, which *fell* on re-rating (criteria median $3 \rightarrow 2$). Aiken's V (relevance) is reported for completeness; it rose slightly but moved off ceiling for only four items.

ID	Item	V (R2)	V (R3)	Feas. R2	Feas. R3
C1	Curriculum alignment	1.00	1.00	4	4
C2	Representational necessity	1.00	1.00	3	3.5
C3	Scientific accuracy and epistemic fidelity	1.00	1.00	3	2
C4	Learner agency and manipulability	0.95	1.00	3	3
C5	Cognitive-load management and scaffolding	1.00	1.00	3	2
C6	Affective and motivational engagement	0.86	1.00	3	3
C7	Modality fit	0.95	0.98	3	2
C8	Blended-phase sequencing and handover	0.98	0.98	3	3
C9	Social and collaborative learning	0.88	1.00	3	2
C10	Differentiation and personalisation	0.86	0.98	2	2
C11	Assessment integration and transfer	1.00	1.00	3	2
C12	Teacher orchestration and readiness	1.00	1.00	2	2
C13	Technical and logistical feasibility	1.00	1.00	2.5	2
C14	Accessibility, equity, safety and data ethics	1.00	1.00	3	2
M1	VR (full immersion)	0.98	1.00	2	2
M2	AR (overlay)	1.00	1.00	4	3
M3	360° video	0.71	0.74	4	4
S1	Pre-class sequencing	0.98	1.00	3	4
S2	In-class sequencing	1.00	1.00	3	2
S3	Post-class sequencing	0.86	1.00	3	3

The *feasibility* dimension did discriminate, and it moved (figure 5a). In Round 2 the criteria median feasibility was 3 of 4, with three criteria already at the floor (median 2): differentiation (C10), teacher orchestration (C12) and, just above it, technical logistics (C13, median 2.5); full VR (M1) was the least feasible modality. Crucially, the controlled-feedback round did not produce bland convergence here – it made the panel *more* pessimistic. On re-rating, the criteria median feasibility fell from 3 to 2: nine of twenty items dropped (C3, C5, C7, C9, C11, C13, C14 and S2 from 3 to 2; M2 from 4 to 3) and only two rose (C2, S1). The panel's consistent narrative was that the criteria name the right targets, but that real classrooms are constrained by teacher workload, device throughput and time – and that hearing the group's feasibility concerns sharpened, rather than softened, that judgement.

On the relevance dimension, by contrast, there was little to converge (figure 5b). The movement was concentrated in the four items that had been clearly below ceiling in Round 2 – C6, C9, C10 and S3 (V between 0.86 and 0.88, IQR = 1) – which closed to $V \geq 0.98$ with IQR = 0. These four are the panel's *only* relevance-IQR changes, so the small "mean $|\Delta\text{IQR}| = 0.2$" reflects four items, not panel-wide movement; the remaining items shifted by at most a few hundredths of V, and we do not claim reflective convergence on relevance. One item held a stable minority position: 360° video (M3) retained the lowest V ($0.71 \rightarrow 0.74$), consistent with its low-agency profile.

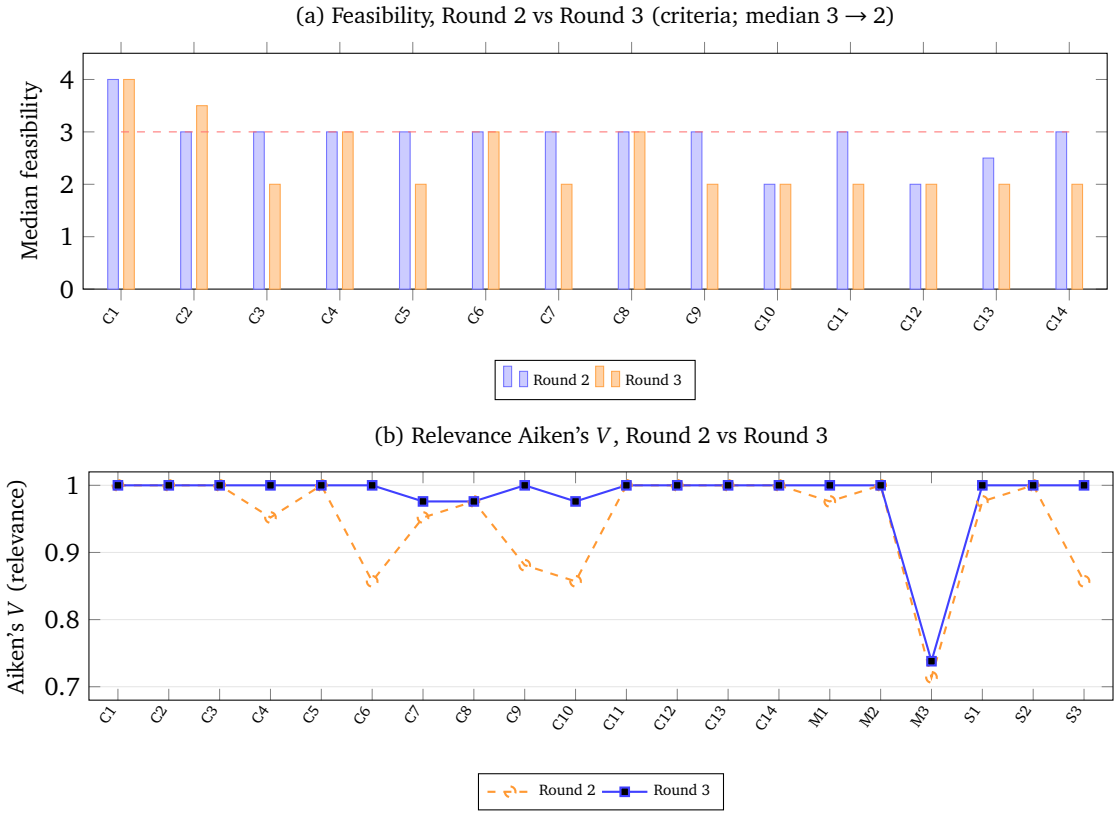

Figure 5: Simulated-Delphi results. (a) Relevance was at ceiling, so the informative signal is feasibility. The dashed line marks the Round-2 criteria median (3); on re-rating after feedback the panel became *more* pessimistic and the criteria median fell to 2 (nine of twenty items dropped; only C2 and S1 rose). The least feasible items are full VR (M1), differentiation (C10) and teacher orchestration (C12). (b) Relevance Aiken's V rose slightly between rounds but moved off ceiling for only four items; 360° video (M3) retained the lone, stable minority position.

5.3. Cross-family analysis

We computed a per-family mean rating to check whether the relevance agreement was an artefact of a single lineage, but the check is weak and we draw little from it. At the relevance ceiling there is almost no variance to analyse; five of the eight "families" (DeepSeek, gpt-oss, Gemma, Nemotron, Qwen) are represented by a *single* endpoint, and the entire spread of family means is 3.83–3.98 on a four-point scale – within rating granularity. The single most critical panelist overall was in fact a Claude endpoint, and a MiniMax endpoint tied for the most lenient, so no conclusion about "lineage flattery" is supportable. The substantive point is the dependence itself: with Claude ×4, MiniMax ×3 and GLM ×2, the effective number of independent voices is well below fourteen. We acknowledge but do not statistically correct for this dependence; the consensus statistics should be read as agreement among fourteen *correlated model endpoints*, not fourteen independent experts, and a sensitivity check collapsing each family to a single vote leaves the relevance ceiling (and hence the uninformativeness of the relevance indices) unchanged.

6. Worked example: Grade-8 "Electric phenomena"

To show the framework in use, we apply it to a Grade-8 physics topic from the model science programme – *Electric phenomena* – building on prior work that introduced VR/AR support for this topic [10, 21]. The design distributes immersion across the three blended phases (figure 6). *Pre-class*, a short 360° preview of everyday electrical phenomena and a prediction quiz activate prior knowledge and curiosity at low immersion (criteria C1, C6, sequencing S1). *In-class*, students work in pairs: AR

overlays the invisible quantities – current, voltage, resistance – on a real circuit they also wire by hand, while a brief VR visualisation makes the electric field manipulable (C2–C5, C7, C9, C13). They predict, change a variable, observe the consequence, and record it – the agency that C4 demands. *Post-class*, students reflect, explain Ohm's law in their own words, and solve a transfer problem on a circuit not seen in the simulation (C10, C11, sequencing S3). Each design decision is traceable to a criterion and a learning mechanism. The framework also *rejects* designs: an initial plan for a full-class, headset-based VR circuit lab fails C7 (modality fit) and C13 (throughput) – thirty headsets are neither available nor necessary when AR overlays the same quantities on real apparatus – so VR is pared back to a brief, rotated field visualisation and AR carries the inquiry. Even the resulting six-headset rotation, however, strains C12, the panel's least feasible criterion: it demands device-management and orchestration skill (for example, at the DigCompEdu "Expert" level [19]) and a charged, updated headset set, so the design ships with a ready lesson script, a device routine and a full AR/desktop fallback (C12–C14). The example thus shows the framework discriminating among designs, not merely blessing one.

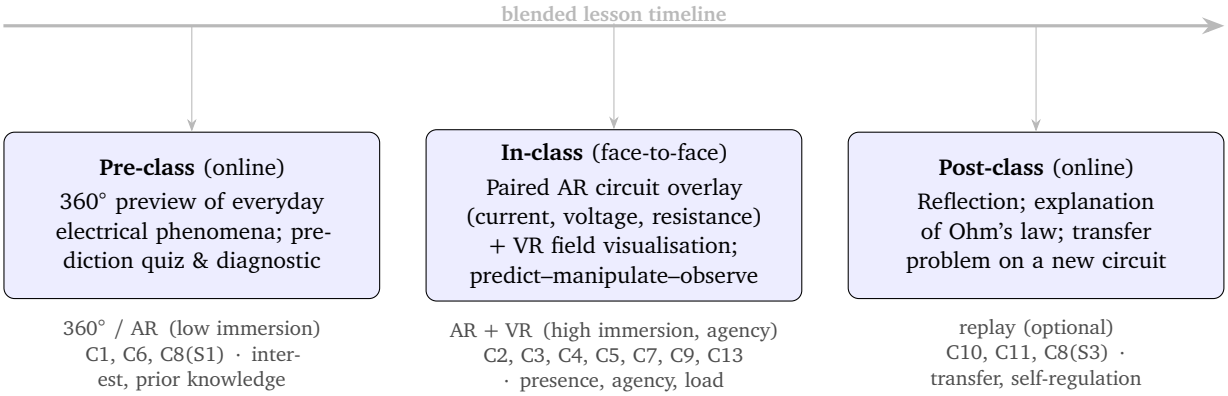

Figure 6: A worked application of the framework to the Grade-8 topic *Electric phenomena*. Immersion is concentrated in the in-class phase, where AR overlays the invisible quantities of a real circuit (C2–C4) and VR visualises the electric field; the low-immersion pre- and post-class phases activate prior knowledge and secure transfer. Tags name the criteria and the CAMIL mechanisms each phase exercises.

7. Discussion

7.1. Theoretical contribution

The framework operationalises CAMIL for the content decision. Where CAMIL explains how presence, agency and affect mediate immersive learning, the fourteen criteria translate those mechanisms into design commitments a teacher can act on and a researcher can audit: representational necessity (C2) and scientific accuracy (C3) make presence serve the science rather than spectacle; agency (C4), cognitive-load management (C5) and collaboration (C9) tune the mechanisms CAMIL identifies; and the feasibility-and-equity gates (C12–C14) acknowledge that no mechanism operates if the activity cannot be delivered safely and reliably. In CAMIL's own terms, the framework is a content-side specification of the model's antecedents – a contribution to pedagogical content knowledge for immersive science.

7.2. Methodological reflection: LLMs as a simulated Delphi panel (RQ3)

The second contribution is the method itself. Its *affordances* were real: a heterogeneous, multi-provider panel was assembled and run three times in hours, every judgement was logged, the instrument improved substantially in Round 1, and the panel revised its ratings after feedback as a human Delphi would. The generative round, in particular, produced criteria a designer would

want – teacher orchestration, collaboration, technical feasibility – suggesting that such panels can be a useful idea-generation and stress-testing tool early in a design cycle.

The *limitations* are equally real and must temper any reading of the numbers. The relevance ceiling is the clearest symptom of LLM agreeableness: a panel that endorses everything is, on that dimension, uninformative, and the useful signal had to be sought in feasibility. Same-family endpoints are correlated, so the effective number of independent "experts" is smaller than fourteen. A distinct circularity compounds the ceiling: the panel rated the very criteria it had generated in Round 1, so high relevance is partly an autocorrelation artefact – the panel judging its own proposals – not only sycophancy, and an independent panel rating the same instrument is the proper test. The panel also has no ecological validity – it has never met a Grade-8 class – and its judgements reproduce patterns in training data, which can flatten or misrepresent the perspectives real experts would bring [24] and cannot be relied on to catch subtle theoretical misattributions. And practical reproducibility is imperfect: two of three channels expose no seed, and provider rate limits intermittently delayed panelists during this very study. These are not reasons to discard the method but reasons to bound its claims. We therefore position the simulated Delphi as a *pre-validation* – a fast, transparent way to refine a framework and surface disagreement before committing scarce human-expert and classroom resources – and explicitly not as a replacement for them [6, 8].

7.3. Practical implications

For teachers and designers, the framework is usable as a checklist and the modality-fit table as a decision aid. The panel's feasibility verdict is itself actionable: the bottlenecks are not the pedagogical ideals but their delivery – teacher preparation time and competence (C12), device throughput and reliability (C13), and genuine differentiation (C10). This argues for investment in teacher professional development and in robust, low-resource fallbacks (C14) at least as much as in headsets, and it supports designs that concentrate full immersion where it is pedagogically necessary while using lower-cost AR and 360° video elsewhere.

7.4. Limitations and future work

Beyond the methodological limits above, the framework is pre-validated only synthetically and only against an English rendering of a Ukrainian curricular context; four criteria (C3, C10, C11, C14) were flagged by the panel as overloaded – bundling distinct constructs with different assessment implications – and cannot serve as unambiguous checklist items until split. The natural next steps are a human-expert Delphi to confront the synthetic results with practitioner judgement, and a design-based classroom study of the *Electric phenomena* sequence measuring conceptual understanding, transfer, cognitive load, presence and motivation. The simulated panel is best understood as the first, inexpensive stage of that longer programme.

8. Conclusion

This paper turned the content component of an immersive-learning methodology from a description into an explicit, refined design framework: fourteen criteria, a modality-fit mapping and a blended-phase sequencing scheme for selecting and orchestrating immersive content in secondary science. The refinement used a novel instrument – a three-round simulated Delphi of fourteen large language models across eight families and three providers – which expanded the framework from eleven to fourteen criteria and stress-tested it. We report the exercise without inflation: relevance was endorsed at a non-discriminating ceiling, so the weight rests not there but on the Round-1 changes and on the feasibility dimension, where the panel discriminated sharply and, on re-rating, hardened its verdict that the real obstacles are teacher orchestration, technical logistics and differentiation. The study is candid about what a synthetic panel can and cannot do: it is a transparent, fully logged pre-validation, not a substitute for human experts or classroom evidence. Used in that spirit, multi-model LLM

panels may become a useful early instrument in educational design research – and the framework they helped refine is now ready for the human-expert and classroom studies that must follow.

Funding

This research received no external funding.

Data availability statement

The complete simulated-Delphi pipeline – panel roster, prompts, per-panelist raw responses for all three rounds, the analysis code and its outputs (content-validity statistics, convergence and family-breakdown tables) – is provided as supplementary material in the GitHub repository accompanying this article (https://github.com/ssemerikov/delphi), together with a README documenting how to reproduce the study.

Conflicts of interest

The author declares no conflict of interest.

Declaration on Generative AI

This study uses generative AI as its object of method, and this is disclosed in full. The Delphi panel consisted entirely of large language models (the fourteen models listed in table 3); *no human experts participated* in the panel. The models were prompted to adopt expert personas and to rate the framework; their verbatim prompts and responses are archived (section 8). Anthropic Claude models (Opus and Sonnet) were additionally used to assist with drafting and editing the manuscript and with implementing the analysis code. The author reviewed and verified all content, takes full responsibility for it, and makes no claim that the simulated panel constitutes human-expert or empirical validation.

References

[1] Aiken, L.R., 1985. Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), pp.131–142. Available from: https://doi.org/10.1177/0013164485451012.

[2] Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C. and Wingate, D., 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), pp.337–351. Available from: https://doi.org/10.1017/pan.2023.2.

[3] Crawford, R. and Jenkins, L., 2017. Blended learning and team teaching: Adapting pedagogy in response to the changing digital tertiary environment. *Australasian Journal of Educational Technology*, 33(2), pp.51–72. Available from: https://doi.org/10.14742/ajet.2924.

[4] Di Natale, A.F., Repetto, C., Riva, G. and Villani, D., 2020. Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research. *P*, 51(6), pp.2006–2033. Available from: https://doi.org/10.1111/bjet.13030.

[5] Diamond, I.R., Grant, R.C., Feldman, B.M., Pencharz, P.B., Ling, S.C., Moore, A.M. and Wales, P.W., 2014. Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67(4), pp.401–409. Available from: https://doi.org/10.1016/j.jclinepi.2013.12.002.

[6] Dillion, D., Tandon, N., Gu, Y. and Gray, K., 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7), pp.597–600. Available from: https://doi.org/10.1016/j.tics.2023.04.008.

[7] Lehikko, A., Nykänen, M., Lukander, K., Uusitalo, J. and Ruokamo, H., 2024. Exploring interactivity effects on learners' sense of agency, cognitive load, and learning outcomes in immersive virtual reality: A mixed methods study. *Computers & Education: X Reality*, 4, p.100066. Available from: https://doi.org/10.1016/j.cexr.2024.100066.

[8] Lin, Z., 2025. Six Fallacies in Substituting Large Language Models for Human Participants. *Advances in Methods and Practices in Psychological Science*. Available from: https://doi.org/10.1177/25152459251357566.

[9] Lynn, M.R., 1986. Determination and Quantification of Content Validity. *Nursing Research*, 35(6), pp.382–386. Available from: https://doi.org/10.1097/00006199-198611000-00017.

[10] Lytvynova, S.H. and Sokolyuk, O.M., 2022. Criteria and indicators for assessing the quality of augmented reality educational objects in physics textbooks. *Information Technologies and Learning Tools*, 88(2), pp.23–37. Available from: https://doi.org/10.33407/itlt.v88i2.4870.

[11] Makransky, G. and Petersen, G.B., 2021. The Cognitive Affective Model of Immersive Learning (CAMIL): a Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educational Psychology Review*, 33, pp.937–958. Available from: https://doi.org/10.1007/s10648-020-09586-2.

[12] Makransky, G., Terkildsen, T.S. and Mayer, R.E., 2019. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, pp.225–236. Available from: https://doi.org/10.1016/j.learninstruc.2017.12.007.

[13] Milgram, P. and Kishino, F., 1994. A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information and Systems*, E77-D(12), pp.1321–1329. Available from: https://www.alice.id.tue.nl/references/milgram-kishino-1994.pdf.

[14] Müller, C. and Mildenberger, T., 2021. Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education. *Educational Research Review*, 34, p.100394. Available from: https://doi.org/10.1016/j.edurev.2021.100394.

[15] Parong, J. and Mayer, R.E., 2018. Learning science in immersive virtual reality. *Journal of Educational Psychology*, 110(6), pp.785–797. Available from: https://doi.org/10.1037/edu0000241.

[16] Parong, J. and Mayer, R.E., 2021. Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, 37(1), pp.226–241. Available from: https://doi.org/10.1111/jcal.12482.

[17] Polit, D.F. and Beck, C.T., 2006. The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, 29(5), pp.489–497. Available from: https://doi.org/10.1002/nur.20147.

[18] Poupard, M., Larrue, F., Sauzéon, H. and Tricot, A., 2025. A systematic review of immersive technologies for education: Learning performance, cognitive load and intrinsic motivation. *British Journal of Educational Technology*, 56(1), pp.5–41. Available from: https://doi.org/10.1111/bjet.13503.

[19] Punie, Y., ed., 2017. *European framework for the digital competence of educators – DigCompEdu*. Luxembourg: Publications Office of the European Union. Available from: https://doi.org/10.2760/159770.

[20] Radianti, J., Majchrzak, T.A., Fromm, J. and Wohlgenannt, I., 2020. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, p.103778. Available from: https://doi.org/10.1016/j.compedu.2019.103778.

[21] Sokolyuk, O.M., 2024. Immersive technologies and learning tools for conducting a school educational experiment in the conditions of blended learning. *Innovative pedagogy*, 77, pp.282–288. Available from: https://doi.org/10.32782/2663-6085/2024/77.56.

[22] Sokolyuk, O.M., 2025. The content aspect of the methodology for using immersive technologies to support blended learning in general secondary education institutions. *Innovative pedagogy*, 89, pp.338–344. Available from: https://doi.org/10.32782/ip/89.66.

[23] Sweller, J., van Merriënboer, J.J.G. and Paas, F., 2019. Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, 31, pp.261–292. Available from: https://doi.org/10.1007/s10648-019-09465-5.

[24] Wang, A., Morgenstern, J. and Dickerson, J.P., 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3), pp.400–411. Available from: https://doi.org/10.1038/s42256-025-00986-z.

[25] Watts, I., Gumma, V., Yadavalli, A., Seshadri, V., Swaminathan, M. and Sitaram, S., 2024. PARIKSHA: A Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data. In: Y. Al-Onaizan, M. Bansal and Y.N. Chen, eds. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp.7900–7932. Available from: https://doi.org/10.18653/v1/2024.emnlp-main.451.