INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 15, Issue 4, (2026) E-ISSN: 2226-6348

Integrating Multi-Modal Machine Learning into Immersive Pedagogy: A Conceptual Framework for AI-Driven Adaptive Scaffolding in VR/AR English Language Learning

[1]Kellie Kai. L. Ai Hon, [2]Arulanandam Benedict. Valentine

[1]Department of Education, Faculty of Education and Humanities, Nilai University, Negeri Sembilan, [2]Department of Business, Faculty of Business, Nilai University, Negeri Sembilan

DOI Link: http://dx.doi.org/10.6007/IJARPED/v15-i4/29138

Published Online: 03 November 2026

Abstract

While Virtual and Augmented Reality (VR/AR) technologies have advanced English language learning, existing adaptive systems rely on fragmented analyses of learner interactions. Current environments process speech, gaze, or gesture in isolation, leading to static instructional designs, limited personalization, and reduced pedagogical responsiveness. To address this gap, this paper proposes an AI-driven multi-modal adaptive framework that integrates Natural Language Processing (NLP), Reinforcement Learning (RL), and multi-modal behavioural fusion for personalized English language acquisition. Grounded in Self-Determination Theory, Embodied Cognition Theory, Scaffolding Theory, and Cognitive Load Theory, the conceptual architecture dynamically synthesizes concurrent behavioural streams to drive real-time, responsive pedagogical interventions. The proposed model overcomes single-channel limitations by translating multi-modal learner data into adaptive instructional decisions. Theoretically, this framework delivers superior intelligent scaffolding, fosters learner autonomy by adapting to user agency, and manages cognitive load through optimized content delivery within immersive environments. Ultimately, this framework shifts VR/AR language learning from static interfaces to highly responsive, personalized ecosystems, offering a robust blueprint for next-generation educational technologies, future empirical validation, and higher education implementation.

Keywords: Artificial Intelligence, Virtual Reality, Augmented Reality, Natural Language Processing, Reinforcement Learning

Introduction

Background of the Study

Artificial Intelligence (AI), Machine Learning (ML), Virtual Reality (VR), and Augmented Reality (AR) — collectively immersive technologies — may have contributed to a demonstrable transformation of digital education by enabling immersive and personalized learning experiences. Within the domain of English language education, immersive technologies proffer authentic communicative contexts, which appear to foster enhanced learner engagement and a diminution of language-related anxiety (Huang et al., 2021). Nevertheless,

a substantial proportion of existing VR/AR systems may largely remain content-driven, thereby typically furnishing identical learning pathways irrespective of the learners' individual needs (Syed et al., 2025, p. 2).

Recent developments in Natural Language Processing (NLP), Reinforcement Learning (RL), and multi-modal learning analytics may facilitate opportunities for intelligent learning environments capable of adapting instruction according to learners' linguistic performance and behavioural cues (Syed et al., 2025, p. 2). However, the frequent independent implementation of these technologies may result in the fragmentation of adaptive systems (Ironsi, 2023, p. 3). It should be acknowledged that recent scholarship has begun to address this isolation; for instance, Hajji et al. (2025) have proposed an architecture for intelligent tutoring in VR that integrates LLM-based non-playable characters with multi-modal interaction through hand tracking, gaze detection, and haptic feedback. Likewise, Kishor et al. (2025) have demonstrated the feasibility of an Agentic AI–enabled VR framework that autonomously adjusts pedagogical content based on multi-modal feedback. Nevertheless, the systematic mapping by Lampropoulos (2025) indicates that such integrations remain uneven and concentrated primarily within narrow, simulated user populations. Accordingly, the present conceptual paper extends this nascent trajectory by proposing a more comprehensive, theoretically grounded, and culturally-adaptable unification of NLP, RL, and multi-modal behavioural fusion for adaptive English language learning in immersive VR/AR environments.

This conceptual paper proposes an integrated Artificial Intelligence (AI)-driven multi-modal framework that could synthesize NLP, RL, speech analysis, gaze tracking, gesture recognition, and immersive VR/AR environments to facilitate adaptive English language learning.

Problem Statement

These four core challenges may represent the current boundary between technological implementation and substantive pedagogical optimization in immersive educational technology. By deconstructing these limitations, the potential inadequacy of a purely empirical approach becomes apparent; consequently, the necessity for a unified conceptual architecture may be established.

First, the issue of uniform instructional pathways in VR environments stems from a historical reliance on static, branching scripts (Ipsita et al., 2024). While VR offers a high degree of spatial presence, the underlying educational design often mimics traditional "Choose-Your-Own-Adventure" books. Every learner, regardless of their native language, prior vocabulary size, or processing speed, is forced through identical predetermined milestones. The rigidity of this design appears to neglect the individual Zone of Proximal Development (ZPD), potentially precipitating a bifurcated failure: advanced learners experience boredom and disengagement, while learners manifesting difficulty may encounter cognitive barriers that impede skill acquisition.

Second, the exclusion of behavioural and biometric indicators create a profound sensory blindness within current AI tutors. When a human instructor facilitates language acquisition, they do not merely attend to linguistic output; rather, they observe physiological markers, including ocular dilation indicative of confusion or hesitations in gestural expression. By

relying exclusively on textual or speech inputs, modern AI architectures appear to disregard these multifaceted, high-frequency non-verbal data streams. A student might correctly pronounce a word but require 10 seconds of anxious cognitive processing to do so; this delay, current speech-recognition systems may interpret as a success, thereby potentially missing a crucial indicator of cognitive overload.

Third, the fragmentation of ML techniques within educational platforms often reflects an instance of engineering compartmentalization. Developers often implement a powerful NLP transformer for conversational dialogue, or an isolated computer vision model for tracking, but these components rarely interact within a unified pedagogical loop. Without an orchestrating layer—such as a RL engine that dynamically weighs both linguistic performance and emotional state—the system may struggle to effect real-time and holistic decisions. The outcome appears to be a disjointed user experience, wherein the technology functions as an assemblage of discrete tools rather than a cohesive, intelligent learning partner.

Finally, there is a distinct theoretical vacuum pertaining to the processes by which AI-driven multi-modal adaptation effectively translates into language acquisition. Much of the current literature tends to focus heavily on system metrics (e.g., accuracy rates or latency speeds), while often neglecting the intricate mechanisms by which real-time algorithmic adjustments interact with established psychological principles, such as Self-Determination Theory (SDT) or Embodied Cognition. Without the establishment of a rigorous conceptual model to elucidate the dynamics underlying the maintenance of a learner's psychological flow by micro-scaffolding, or the methods through which physical gestures, when integrated with AI feedback, may deepen semantic memory, the field notably appears to risk the development of advanced technologies that ultimately lack foundational pedagogical substance.

Research Gap

The Isolated Strengths of Existing Technologies

To establish the necessity of a unified framework, the independent capabilities of these four technological domains may warrant preliminary recognition:

- **Immersive VR:** Current scholarship appears to document VR's capacity to induce spatial presence and psychological immersion (Chun et al., 2022). By placing learners in authentic, high-fidelity cultural and linguistic contexts (e.g., a simulated boardroom or a medical clinic), VR may lower affective filters, elicit situational interest, and facilitate embodied cognition through spatial interaction.

- **NLP:** In Computer-Assisted Language Learning (CALL), NLP models and Automated Speech Recognition (ASR) engines have demonstrated the potential for significant accuracy. They may analyse phonetic outputs, detect phonemic mispronunciations, and provide instantaneous corrective feedback on syntax, rhythm, and lexical choice.

- **RL:** In adaptive educational systems, RL algorithms (such as Q-learning or Deep Q-Networks) appear to excel at dynamic pathing. By treating instructional delivery as a sequential decision problem, RL maps out personalized learning trajectories, optimizing the sequence of content to balance challenge and skill.

- **Multi-modal Analytics:** Emerging research in learning analytics suggests that the monitoring of nonverbal cues, such as gaze duration, facial expressions, and gestural velocity—may provide a rich window into user engagement, attention allocation, and impending cognitive fatigue.

The Fragmentation Problem (The Core Gap)

Despite these individual advances, current educational platforms appear to operate under a paradigm of technological isolation. Virtual Reality environments frequently retain a deterministic character, relying on static, hard-coded branching logic that necessitates a monolithic treatment of the user. This rigid design limits the platform's ability to respond to individual needs, essentially reducing the immersive experience to a sequence of predetermined outcomes that overlook the learner's cognitive variability.

Conversely, powerful NLP engines and RL personalization models are typically confined to flat, two-dimensional web or mobile interfaces; such systems may fail to utilize the rich, embodied spatial context that VR potentially affords.

Furthermore, multi-modal analytics are frequently relegated to the utilization of post hoc diagnostic procedures for data analysis subsequent to the conclusion of a learning session, rather than their deployment as live, operational telemetry (Yan et al., 2024).

Consequently, there is an absence of an architecture that enables these distinct streams to communicate simultaneously. A NLP engine might identify a pronunciation error; however, without multi-modal behavioural integration, the system may struggle to determine if the error resulted from a deficiency in lexical acquisition, visual distraction (tracked via gaze), or elevated cognitive anxiety (tracked via physiological or gestural pacing).

How This Paper Addresses the Gap

This study seeks to address this fragmentation by proposing a Closed-Loop, Multi-modal Intelligent Scaffolding Framework. Rather than the compartmentalization of immersion, linguistic analysis, algorithmic pathing, and behavioural tracking, this framework appears to facilitate the synthesis of these components into a singular, responsive, and adaptive architecture:

By unifying these technologies, the framework facilitates an emergent property unattainable by isolated systems: **context-aware, real-time pedagogical adaptation**. The system continuously monitors the learner's linguistic performance alongside their behavioural and cognitive states.

If the RL engine detects cognitive overload through a combination of stuttered speech metrics (NLP) and erratic gaze fixation (multi-modal analytics) within the VR, it immediately adjusts the environment. It may dynamically deploy visual AR overlays, introduce supportive AI avatars, or downgrade the immediate linguistic complexity of the task.

Ultimately, this paper appears to shift the academic discourse from the isolated functional capacities of individual technologies to the manner in which a unified, multi-agent AI architecture could enable the realization of a truly continuous, self-optimizing ZPD for the

learner. By integrating these disparate modalities into a cohesive, data-driven ecosystem, this research addresses existing deficiencies in system generalizability and methodological rigidity (Sajja et al., 2025, p. 4).

Objectives of the Paper

Theoretical Synthesizer (Integrating Educational Theories with ML)

The primary value of this conceptual paper resides in its capacity for reconciliation between advanced ML and human cognitive architectures. In traditional educational technology development, software engineers construct sophisticated systems absent pedagogical grounding, whereas educators articulate pedagogical methodologies without a comprehensive understanding of algorithmic capabilities. This framework potentially serves to integrate these distinct domains:

- **Dynamic Scaffolding Meets RL:** Vygotsky's ZPD dictates that instructional support must fade as competence increases (Ludwig-Hardman & Dunlap, 2003, p. 3). The framework operationalizes this construct by mapping the ZPD onto a RL engine, such as Q-learning. The algorithm may treat the scaffolding level as a variable state, dynamically deploying or withdrawing instructional assistance, such as visual hints, translation overlays, or conversational adjustments—based on the learner's real-time performance vector.

- **Embodied Cognition Meets Multi-modal Biometrics:** Embodied Cognition Theory (ECT) argues that learning is intrinsically tied to physical actions and spatial presence (Koning & Tabbers, 2011, p. 503). By integrating gaze tracking, spatial gesture tracking, and vocal metrics, this framework interprets a learner's physical movements inside VR and AR not merely as stochastic actions, but as significant cognitive data streams that may inform the NLP engine to assess comprehension.

- **SDT Meets Adaptive Personalization:** To foster intrinsic motivation, a learner needs a sense of Autonomy, Competence, and Relatedness (Deci et al., 1991, p. 327). The framework may achieve this by ensuring the artificial intelligence avoids abrupt shifts in difficulty. Rather, it appears to facilitate the analysis of biometric anxiety or boredom indicators for the purpose of modulating immersion variables, thereby preserving the learner's sense of competence and facilitating a sustained psychological state of "flow."

Operational Guide for Instructional Designers (The Blueprint)

For instructional designers, current immersive technologies present considerable implementation challenges due to a perceived absence of rigorous design principles. This framework delineates a structured **Three-Tier Architectural Model** that may standardize the implementation of immersive lessons:

1) **The Telemetry Tier (Inputs):** This tier facilitates the establishment of precise sensory criteria, which designers must meticulously consider; such criteria notably transcend the limitations of simplistic quiz scores by integrating high-frequency data streams (e.g., speech latency, gaze fixation duration on object overlays, gesture velocity).

2) **The Orchestration Tier (Processing):** Defines the programmatic rules for how an AI tutor evaluates those inputs. This structure appears to provide designers with a robust logic

matrix: If [Linguistic Performance = Low] AND [Gaze Tracking = Erratically Spread], trigger [Visual Micro-Scaffolding Overlay].

3) **The Experience Tier (Outputs):** This tier typically delineates the mechanisms through which the virtual environment dynamically adapts—a process that could involve changes in ambient audio, alterations in the conversational complexity of AI avatars, or the projection of contextual AR text—thereby accommodating the learner's discernible cognitive state.

Literature Review

The synthesis of immersive pedagogical design and algorithmic cognitive tracking remains a burgeoning area of cross-disciplinary inquiry. Current research primarily bifurcates into studies emphasizing either the technical affordances of extended reality or the isolated efficacy of computer-assisted language learning software.

Evolution of VR/AR in English Language Learning

The integration of immersive technologies in foreign language education has progressed through distinct historical paradigms, evolving alongside computational capacity and pedagogical theory. As delineated by Yan and Lowell (2025), this evolution spans four operational phases: from early two-dimensional (2D) text-based virtual environments (Multi-User Dungeons (MUDs) and Multi-User Dungeons, Object-Oriented (MOOs)), to three-dimensional (3D) desktop-based virtual worlds like Second Life, progressing to high-fidelity 3D Head-Mounted Display (HMD) immersive VR, and culminating in modern AI-enhanced 3D HMD ecosystems. Throughout this trajectory, the primary affordance of VR and AR has transitioned from simple visual novelty to the creation of high-presence, high-context situated learning environments. Modern XR platforms appear to simulate authentic sociocultural and communicative spaces, potentially facilitating the mitigation of learners' affective barriers, the reduction of language anxiety, and the fostering of situational motivation (Kaplan-Rakowski & Gruber, 2023, p. 1). However, a persistent limitation across these evolutionary stages is the deterministic nature of environmental scripts. Traditional VR/AR setups often appear to operate on rigid, pre-programmed branching logic, which may force diverse learners down identical instructional pathways. Consequently, while current VR/AR systems maximize spatial presence, they appear to lack the capacity to adjust to real-time variations in student language ability or cognitive load.

AI for Adaptive Learning

AI appears to have shifted the educational paradigm from passive content consumption to active, adaptive personalization. In Second Language Acquisition (SLA) contexts, AI-driven adaptive learning systems leverage data processing and pattern-recognition algorithms to continuously evaluate learner interactions and optimize individual learning trajectories. By adjusting variables such as task complexity, vocabulary density, and scaffolding delivery, these systems ensure that instruction aligns with the learner's dynamic competencies. From an educational psychology perspective, this real-time adaptability may operationalize the Constructivist principle of maintaining instruction within the learner's ZPD. When optimized, adaptive AI functions as a scalable equalizer, the mitigation of cognitive burnout in struggling students while simultaneously sustaining engagement for advanced learners. Despite these capabilities, the implementation of AI adaptivity in contemporary platforms appears to be

restricted to two-dimensional interfaces, manifesting primarily within mobile and web-based applications. The field appears to lack robust conceptual frameworks that might elucidate how this digital adaptivity can be natively embedded within three-dimensional spatial environments to manipulate physical, auditory, and structural variables simultaneously.

Natural Language Processing in Language Education

Within Computer-Assisted Language Learning (CALL) and intelligent tutoring contexts, NLP and Automatic Speech Recognition (ASR) function as the primary mechanism for the evaluation of linguistic output. Modern Large Language Models (LLMs) and fine-tuned transformers may facilitate the granular, semantic, and phonetic analysis of student speech in real time (Chen & Dang, 2025). These algorithms parse complex syntax, evaluate lexical sophistication, detect phonemic mispronunciations, and compute structural fluency metrics. Unlike traditional human-led assessments, NLP agents may provide instant, formative, and objective feedback, allowing learners to iteratively refine their output without the social anxiety often associated with face-to-face instruction (Aziz et al., 2025). However, current NLP applications in language education appear to suffer from severe sensory isolation. Because these models operate strictly on textual or acoustic signals, they are blind to the environmental and behavioural contexts of communication. An NLP tutor can identify a stutter or a grammatical mistake, but the system may lack the capacity to ascertain whether the error is attributable to an actual vocabulary deficit, visual distraction in the immediate environment, or physiological hesitation.

Reinforcement Learning for Personalized Learning

RL treats the instructional pathway as a sequential decision-making problem under uncertainty (Bulathwela et al., 2025, p. 16). Within an educational architecture, an RL agent learns an optimal teaching policy by interacting with the student (the environment) and receiving feedback in the form of mathematical rewards (e.g., student mastery gains, retention scores, or engagement metrics) (Ruan & Lu, 2025). Using algorithms such as Q-learning or Deep Q-Networks (DQN), the system maps current learner states to specific pedagogical actions, such as micro-scaffolding deployment or content sequencing. Over time, the policy self-optimizes, uncovering non-obvious instructional pathways tailored to individual cognitive profiles. In the context of global online distance education—where physical teachers cannot monitor individual student trajectories—RL may provide an automated mechanism for the management of learning velocity and content distribution. Nevertheless, the majority of current educational RL architectures may optimize paths based primarily on discrete, historical inputs—such as assessment metrics—thereby potentially elucidating a significant constraint regarding the capacity of these architectures to assimilate fluid, continuous, and high-frequency spatial behavioural data.

Multi-modal Learning Analytics

Multi-modal Learning Analytics (MMLA) represents a critical methodological pivot from macro-level performance metrics to high-frequency, continuous behavioural tracking. By capturing diverse data streams—including eye-tracking (gaze duration and fixation frequency), computer vision (facial expressions and gestural kinematics), and audio telemetry (vocal pitch, latency, and hesitation patterns)—MMLA facilitates a comprehensive delineation of the user's cognitive and affective states. In digital educational settings, MMLA

tools have successfully mapped eye movements to visual attention allocation and translated facial micro-expressions into affective states such as confusion, frustration, or boredom (Skaramagkas et al., 2021). Yet, across contemporary literature, a substantial operational gap persists: the utilization of MMLA data appears predominantly restricted to post-hoc diagnostic inquiries, as opposed to the prospective implementation of real-time system mediation. Researchers frequently employ multi-modal data to facilitate a comprehension of student performance deficits post-session, rather than the integration of such analytics into a live system to mitigate these occurrences.

Intelligent Tutoring Systems(ITS)

ITS appear to provide a structured foundation for computer-mediated pedagogy by utilizing three core interconnected components: the **Domain Model** (the curriculum/knowledge structure); the **Student Model** (the record of individual student knowledge, history, and psychological states); and the **Pedagogical Model** (the instructional strategy rules). In language acquisition, ITS frameworks have historically facilitated the didactic delivery of linguistic content, emulating individualized pedagogical intervention through the provision of corrective feedback and explicit rule elucidation (Kurup et al., 2022; Son et al., 2023, p. 101) While these systems are highly structured and reliable, they are architecturally rigid(Aleven et al., 2016, p. 245). Traditional pedagogical models within an ITS exhibit a substantial reliance upon predetermined, expert-authored heuristic logic(Chowdhury et al., 2024, p. 6; Pappas & Drigas, 2016, p. 12) This deterministic foundation may render it exceedingly difficult for the system to adapt fluidly to the chaotic, unstructured, and multi-layered realities of conversational, experiential speech in immersive settings. To evolve, the traditional ITS blueprint may necessitate the integration of more flexible, probabilistic machine learning agents that could dynamically adjust parameters without relying solely on predefined rule matrices.

Embodied Learning in Virtual Environments

Embodied Cognition Theory (ECT) posits that cognitive processes, semantic memory formation, and language acquisition are fundamentally grounded in the physical body's interactions with its environment. In virtual and augmented reality spaces, learning is not merely a passive mental exercise; it is an active, corporeal experience driven by spatial movement, object manipulation, deictic gesturing, and localized audio cues. When a language learner physically picks up a virtual object while hearing and pronouncing its target name, the cognitive system activates overlapping neural networks responsible for sensory, motor, and linguistic processing. This multi-sensory reinforcement appears to accelerate semantic mapping and may enhance long-term vocabulary retention. While ECT provides a robust philosophical and psychological justification for using immersive technologies, current educational implementations infrequently capture or adequately integrate these bodily actions as input data. The physical movements that define embodied learning may often be construed by systems as mere graphical outputs, rather than intentional behavioural inputs that should dictate the adaptive behaviour of the digital world.

Current Research Gaps

A rigorous synthesis of the literature suggests a profound technological and theoretical fragmentation at the intersection of XR, AI, and SLA. While individual studies comprehensively validate that VR improves psychological immersion, NLP enhances structural pronunciation

feedback, Reinforcement Learning effectively path-allocates, and MMLA provides granular user diagnostics, limited research appears to integrate these technologies into a singular, closed-loop adaptive instructional framework. This absence consequently precipitates three critical research lacunae:

- **The Multi-Modal Behavioural Fusion Lacuna:** Current immersive environments appear to demonstrate a relative insensitivity to the intricate array of non-verbal, physical telemetry (e.g., gaze, gesture, affect), thereby impeding their capacity for the diagnosis of complex cognitive states, such as spatial overload or affective disengagement.

- **The Algorithmic Orchestration Gap:** Existing adaptive systems are characterized by a tendency to conceptualize NLP, computer vision, and personalization models as discrete, isolated software silos, consequently impeding the operationalization of holistic, multi-agent pedagogical strategies.

- **The Theoretical Vacuum:** The educational technology field appears to remain heavily focused on engineering and system-level performance indicators (e.g., latency, frame rate, and model accuracy), resulting in a dearth of comprehensive conceptual models that elucidate the extent to which real-time, artificial-intelligence-driven multi-modal modifications may translate into the psychological parameters of language acquisition, such as cognitive load management, learner agency, and intrinsic motivation. This conceptual paper appears to address these combined deficiencies.

Theoretical Foundations

The proposed conceptual framework operates at the intersection of computational engineering and cognitive psychology. To move beyond a mere technical architecture, the design must be anchored in established pedagogical paradigms that delineate the mechanisms through which the human brain processes language, perceives immersive stimuli, and sustains psychological drive. This section delineates the four theoretical pillars that may ground the system’s closed-loop adaptation mechanism.

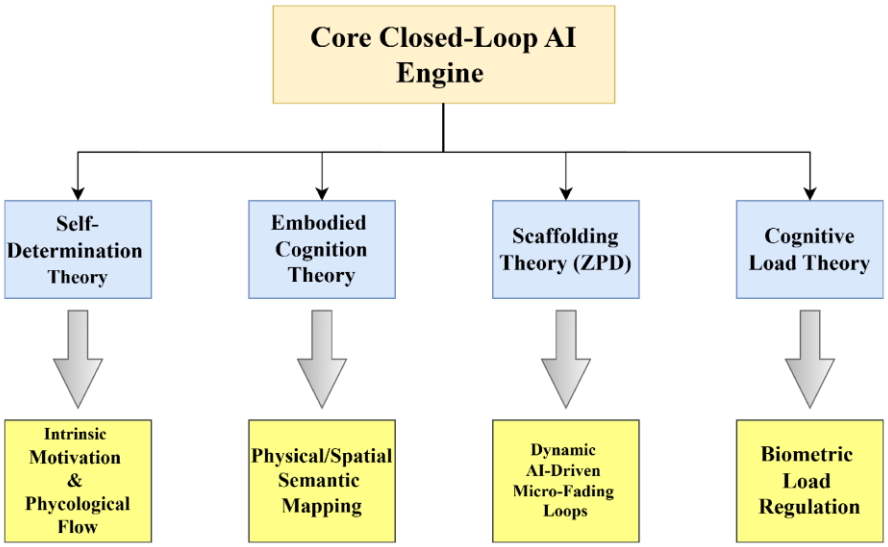


Figure 1 Theoretical Framework

Self-Determination Theory (SDT)

SDT posits that human motivation operates along a continuum from extrinsic compliance to intrinsic internalization, driven entirely by the satisfaction of three universal psychological needs: autonomy, competence, and relatedness. In online distance higher education, these needs are chronically compromised due to rigid digital curricula and geographic isolation. Traditional adaptive software may frequently attempt to ameliorate this condition by explicitly prompting a student to downgrade their difficulty level when failing; such an intervention may undermine user agency and damage perceived competence.

The proposed framework operationalizes SDT through invisible algorithmic orchestration. Rather than the reliance upon overt system alerts, the internal RL engine appears to employ dynamic performance vectors to implement subtle, peripheral environmental adjustments.

- **Autonomy** is preserved because the system adapts with the learner rather than the dictation of instructional directives to the learner. Micro-adjustments—such as introducing a supportive visual context clue or decelerating an avatar's speech rate—occur seamlessly in the background, potentially facilitating the learner's experience of enhanced environmental ownership.

- **Competence** is continuously reinforced by maintaining the learner in a psychological state of "flow." Through the computation of the individual's performance in real time, the AI ensures tasks are neither anxiety-inducing nor boring, thereby facilitating the provision of a consistent sequence of non-punitive, formative micro-successes.

- **Relatedness** is re-engineered by converting the AI from a cold assessment tool into an empathetic, responsive "intelligent learning partner." Because the virtual avatars dynamically adjust their behavioural expressions based on real-time biometric telemetry, they may project a sense of safe co-presence, potentially mitigating the social evaluation anxiety often associated with face-to-face language performance.

Embodied Cognition Theory (ECT)

Embodied Cognition Theory (ECT) challenges traditional Cartesian dualism by asserting that cognitive processes, semantic memory formation, and language acquisition are fundamentally grounded in the physical body's interaction with the environment. In standard vocabulary acquisition, words are treated as abstract symbols decoupled from physical reality. ECT may suggest that when a linguistic concept is learned in tandem with a physical action, sensorimotor simulation tracks across overlapping neural networks, accelerating long-term semantic mapping.

Within this framework, ECT serves as the justification for the utilization of high-frequency gesture and spatial tracking as an input modality rather than merely a graphical display output. In the proposed architecture, physical gestures inside the virtual space (such as grabbing an object, pointing, or shifting posture) are quantified as rich behavioural data streams.

When a learner physically picks up a virtual object while simultaneously hearing and pronouncing its target English word, the semantic representation of that word is deeply encoded across visual, motor, and auditory cortices (Fernandino et al., 2015, p. 2018). Furthermore, by tracking deictic and iconic gestures, the system's computer vision layer may enable an assessment of whether a learner's physical movements match their linguistic output. If a student correctly identifies a tool but exhibits hesitation or reaches for an incongruous object, the system may identify a conceptual mapping failure that a purely text- or audio-based artificial intelligence might fail to recognize, thereby triggering targeted contextual adjustments.

Scaffolding Theory and the ZPD

Vygotsky's Scaffolding Theory suggests that optimal learning takes place within the ZPD—the psychological distance between what a learner can accomplish independently and what they can achieve with guidance. A critical requirement of effective scaffolding appears to be its dynamic and transient nature: as the learner's mastery increases, instructional supports may require systematic attenuation to mitigate dependency and facilitate cognitive independence.

The proposed framework appears to map this pedagogical concept onto a multi-agent machine learning loop, converting a static, theoretical model into an automated, operational algorithm. The framework utilizes a fine-tuned NLP model and a RL pathing engine to continuously redefine the boundaries of the user's ZPD during conversational tasks.

Upon a learner's engagement with a complex communicative scenario (e.g., a simulated academic defence), the scaffolding architecture appears to initialize at a high-support state, projecting real-time translation text overlays, highlighting critical objects via AR, and slowing the speech rate of virtual agents. As the NLP model measures increases in grammatical accuracy, lexical richness, and pronunciation confidence, the RL agent initiates an automated micro-fading sequence. Text overlays gradually thin out to single keyword definitions, visual highlights dissolve, and virtual agents smoothly transition to natural native speech velocities. Should a decline in learner performance be observed, the system may gracefully reduce the

scaffolding layer, providing a friction-less, self-correcting safety net that scales perfectly to individual learning curves.

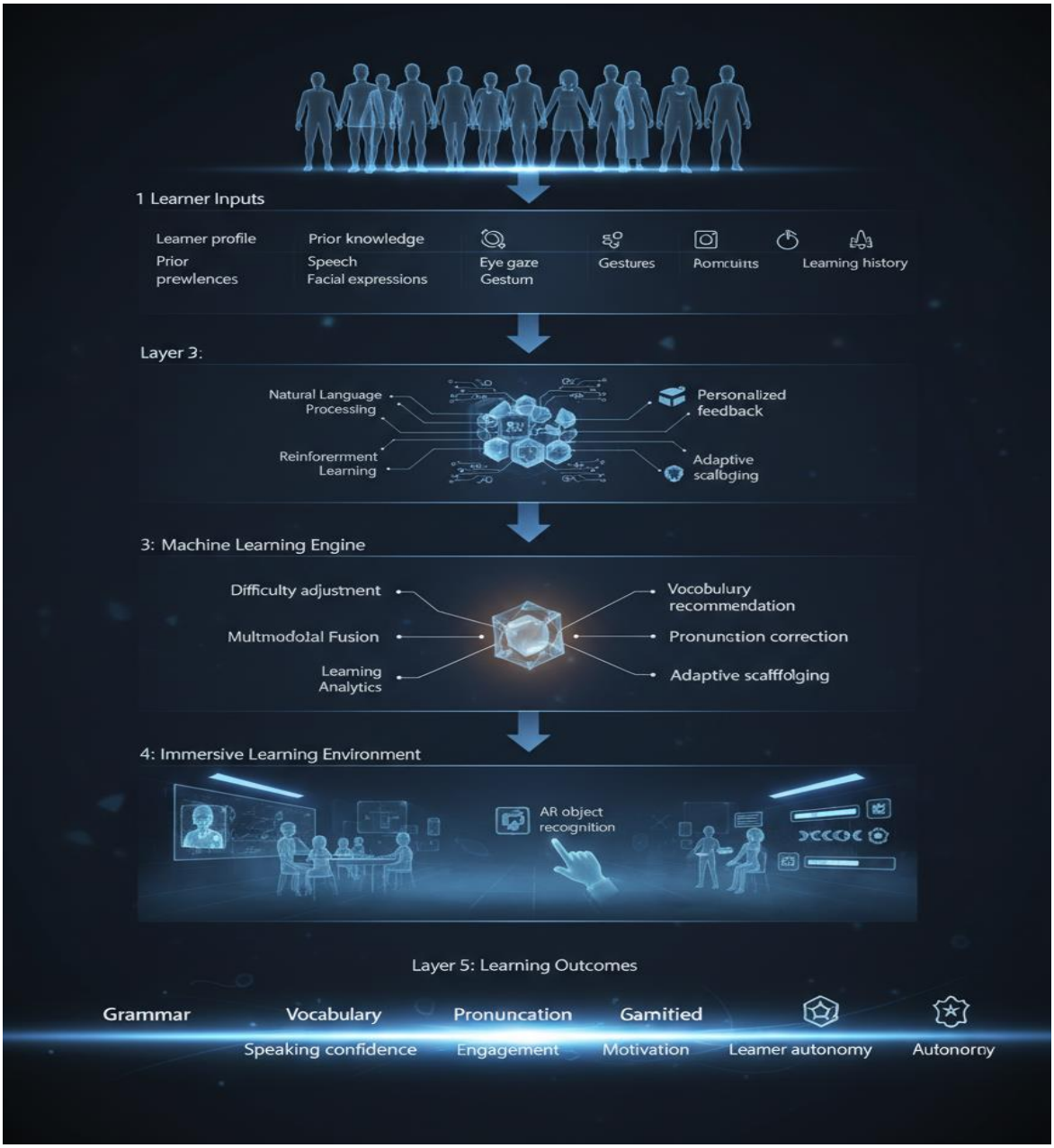
Cognitive Load Theory (CLT)

Cognitive Load Theory (CLT), developed by John Sweller, posits that human working memory has a strictly limited processing capacity (Sweller et al., 2019, p. 261). Total cognitive load consists of three vectors: **intrinsic load** (the inherent difficulty of the learning material), **extraneous load** (the mental effort wasted on poorly designed instructional delivery or environmental noise), and **germane load** (the productive mental energy dedicated to constructing cognitive schemas (Skulmowski & Xu, 2021, p. 173). Immersive three-dimensional environments present a profound operational paradox for CLT: while VR and AR facilitate the maximization of engagement, the overwhelming volume of visual and auditory stimuli could generate potentially excessive extraneous cognitive load, thereby inducing a "split-attention effect" that may impede language processing.

The proposed framework may resolve this paradox through the utilization of Multimodal Learning Analytics (MMLA) as a live regulatory valve for working memory capacity. The architecture features an operational telemetry layer that continuously processes eye-gaze tracking (fixation durations, saccadic regressions) and audio-speech latency metrics.

Through the analysis of these synchronized data streams, the system may facilitate the mathematical calculation of cognitive friction in real time. For instance, if a student's gaze data suggests erratic, hyper-frequent saccadic jumps across the virtual room while speech patterns demonstrate prolonged hesitation, the system may diagnose imminent cognitive overload.

The system may affect an immediate adjustment to the environment to facilitate the clearance of the learner's working memory; it systematically suppresses extraneous visual elements, simplifies the background scenery, and narrows the linguistic task's intrinsic difficulty. Conversely, if the telemetry registers low cognitive friction paired with rapid linguistic execution, the system may safely raise the task complexity. By dynamically balancing this cognitive equation, the framework may ensure that maximum mental energy is channelled exclusively into germane processing, thereby accelerating language schema construction without exceeding the threshold of mental exhaustion.



Different types of Layers

Layer One: Learner Inputs (The Telemetry & Profiling Layer)

Layer one appears to serve as the sensory and historical foundation of the entire architecture, facilitating the categorization of incoming data into static baselines and high-frequency dynamic telemetry streams.

- **Static Baselines:** Upon system initialization, the framework facilitates the compilation of Learner Profile, Prior Knowledge Maps (via pre-diagnostic language testing), Learning Preferences (e.g., visual versus auditory pacing variations), and historical records (Learning History from past sessions). This may establish the student's initial coordinate within the curriculum matrix.

- **Dynamic Telemetry Streams:** Once the user enters the virtual environment, the system initiates real-time data collection at high sampling rates. The Speech channel appears to track vocal acoustic waves, calculating features such as pitch variations, phonemic structures, and response latency. Simultaneously, the Eye Gaze layer utilizes internal Head-Mounted Display (HMD) eye-tracking cameras to measure fixation duration, saccadic regression, and visual coordinates. The Gestures stream uses spatial and computer vision controllers to quantify bodily kinematics (e.g., reach velocity, deictic pointing accuracy, and spatial posture), while Facial Expressions are mapped via micro-expression vectors to decode emerging emotional states such as frustration or boredom. This combination may ensure that the user's cognitive and physiological states are fully represented.

Layer Two: Machine Learning Engine (The Cross-Modal Fusion Layer)
Layer two appears to function as the core computational stratum, potentially responsible for the transformation of raw data streams into actionable insights. Rather than processing inputs in isolated silos, this layer executes **cross-modal data fusion** using specialized machine learning sub-agents:

- **NLP:** Deep transformer networks may facilitate the evaluation of syntactic complexity, lexical selection, and phonemic deviations through the analysis of acoustic and textual inputs in real time.

- **Multi-modal Fusion & Learning Analytics:** This subsystem may facilitate the resolution of temporal alignment across distinct sensory inputs; The integration of linguistic data from the Natural Language Processing engine with physiological and behavioural telemetry (gaze duration and gesture hesitation) into a shared latent space may mitigate cross-modal confounding, thereby permitting the system to distinguish between physical distraction and actual vocabulary deficits.

- **RL:** Operating on the integrated feature vectors, an RL policy engine (e.g., Deep Q-Networks) calculates the next optimal pedagogical strategy. It appears to treat the current state of the student as an environmental variable, and the evaluation of potential instructional moves suggests a maximization of long-term linguistic retention and engagement.

Layer Three: Adaptive Decision Engine (The Pedagogical Orchestration Layer)
Layer three appears to facilitate the translation of high-level computational policies from Layer two into real-time instructional actions, functioning as an automated syllabus manager that may continuously adjust variables across five distinct areas:

- **Difficulty Adjustment:** Involves the modulation of systemic complexity of conversational scenarios based on real-time competence metrics.

- **Personalized Feedback:** Entails the construction of context-aware corrective prompts tailored to the individual's learning profile.

- **Vocabulary Recommendation:** Involves the dynamic injection of target lexical units into conversational options based on recorded gaps in the student's prior knowledge model.

- **Pronunciation Correction:** Necessitates the generation of immediate, localized acoustic adjustments, shifting the phonetic baseline needed to progress.

- **Adaptive Scaffolding:** Governs the automatic fading and deployment of instructional aids. If cognitive load limits are approached, the system may instruct the environment to display visual hints or translate textual elements; as language competence stabilizes, a reduction in these prompts could cultivate user independence.

Layer 4: Immersive Learning Environment (The Rendering & Experience Layer)
Layer four appears to constitute the interactive interface wherein pedagogical adjustments are manifested for the student. This layer facilitates the mapping of rendering directives from Layer three into a three-dimensional spatial canvas:

- **Virtual Classroom / Spaces**: The system appears to manipulate the surrounding visual environment. It may facilitate the reduction of background visual clutter to lower extraneous cognitive load, adjust lighting schemes, or dynamically modify setting variables—such as shifting from a casual coffee shop to an intensive academic boardroom—to potentially assess situational confidence.

- **AI Tutor / Avatars**: Virtual conversational partners may dynamically undertake the adjustment of their behaviours. They appear to modulate speech velocity, select varying syntactic complexities, and alter facial animations to project empathy, while potentially exhibiting supportive gestures—such as pointing to an object of interest when gaze diagnostics suggest potential user uncertainty.

- **AR Object Recognition & Gamified Activities**: When deployed in augmented reality configurations, the interface appears to facilitate the superimposition of context-aware textual overlays onto real-world items, potentially converting physical surroundings into an interactive linguistic domain.

Layer 5: Learning Outcomes (The Evaluation & Validation Layer)
Layer 5 appears to represent the target destination of the entire architectural pipeline, through the characterization of system efficacy across both hard linguistic performance indicators and broader psychological metrics:

- **Linguistic Gains:** Evaluation may be conducted through the quantification of structural improvements in grammatical accuracy, lexical breadth, phonological precision, and conversational self-efficacy during conversational activities.

- **Psychological & Motivational Transformations:** Grounded in Self-Determination Theory and Flow Theory, this vector appears to evaluate long-term behavioural changes. It may track shifts in Motivation—moving from extrinsic compliance to intrinsic drive—increases in task Engagement time, and the development of Learner Autonomy.

As Layer Five metrics are logged, they undergo integration into the Learning History of Layer One. This creates a fully enclosed, self-correcting loop that ensures the intelligent immersive learning system evolves alongside the user's growing mastery.

Research Propositions

- **P1:** The implementation of AI-driven adaptive virtual reality and augmented reality environments may positively influence English language learning outcomes through the provision of personalized instructional scaffolding.
- **P2:** Multimodal behavioural fusion appears to enhance adaptive learning with greater efficacy than single-modal machine learning approaches.
- **P3:** The utilization of real-time Natural Language Processing feedback may significantly improve the acquisition of pronunciation and grammatical proficiency.
- **P4:** Reinforcement learning techniques could facilitate the optimization of individualized learning pathways, potentially resulting in increased learner engagement.
- **P5:** Biometric adaptation might facilitate the mitigation of extraneous cognitive load through the dynamic adjustment of instructional complexity.
- **P6:** Adaptive scaffolding appears to facilitate the enhancement of learner autonomy and foster intrinsic motivation.

Theoretical Contributions

The proposed conceptual framework may offer significant theoretical contributions to the field of educational technology, artificial intelligence in education (AIED), and technology-enhanced language learning through the integration of multifaceted theoretical perspectives with contemporary machine learning methodologies.

Integrating Educational Theories into an AI-Driven Learning Framework

The initial contribution of this framework appears to reside in the integration of four complementary educational theories—Self-Determination Theory, Embodied Cognition Theory, Scaffolding Theory, and Cognitive Load Theory—within a unified artificial intelligence-driven instructional model; whereas previous research has typically applied these theories in isolation to delineate learner motivation, physical interaction, instructional support, or cognitive processing, the proposed framework suggests how these perspectives may operate synergistically within an adaptive immersive learning environment.

Self-Determination Theory elucidates the mechanisms through which AI-enabled personalization may promote learner autonomy, competence, and intrinsic motivation via individualized learning pathways. Embodied Cognition Theory offers a foundational perspective for understanding how physical interactions—including speech, gestures, and gaze behaviours—may become meaningful sources of learning data within immersive VR and AR environments. Scaffolding Theory elucidates how instructional support might be continuously adjusted according to learners' evolving needs; concurrently, Cognitive Load Theory may guide the optimization of instructional complexity to mitigate learner overload. Integration of these four theories into a unified conceptual model suggests that the framework offers a holistic elucidation of how intelligent technologies might simultaneously support learners' cognitive, behavioural, emotional, and motivational development.

Extending Scaffolding Theory toward Intelligent Adaptive Scaffolding

The second contribution involves the theoretical extension of traditional Scaffolding Theory to the context of intelligent adaptive learning environments; this conceptual transition may facilitate the reconciliation of established pedagogical principles with contemporary technological paradigms. Conventional scaffolding assumes that instructional support is primarily provided by teachers or carefully designed learning materials and gradually withdrawn as learners become more competent (Chen et al., 2003, p. 348). While this principle remains relevant, recent advances in artificial intelligence may facilitate the evolution of scaffolding toward dynamic, data-driven, and responsive interaction modalities.

The proposed framework introduces the concept of Intelligent Adaptive Scaffolding, in which artificial intelligence appears to continuously analyse multi-modal learner data—including speech performance, eye gaze, gestures, behavioural patterns, and learning history—to determine the appropriate level of instructional support. The optimization of instructional decisions through Reinforcement Learning algorithms may be facilitated by the selection of learning activities that correspond with individual learner profiles, while Natural Language Processing provides immediate feedback on pronunciation, grammar, and language use. Rather than following predetermined instructional sequences, the learning environment adapts continuously to learners' progress, allowing scaffolding to evolve dynamically throughout the learning process. This reconceptualization extends classical educational theory by demonstrating how artificial intelligence can function as an intelligent instructional partner capable of maintaining learners within their Zone of Proximal Development.

Introducing Multi-modal Behavioural Fusion into Immersive Language Learning

A third contribution of the framework entails the incorporation of multi-modal behavioural fusion into the landscape of immersive English language learning; this synthesis appears to address extant limitations in current automated assessment methodologies. Existing artificial intelligence-supported language learning systems often rely on single sources of learner information, such as textual responses, quiz performance, or speech recognition. Although these approaches improve automated assessment, they appear to provide only a limited representation of learners' cognitive and behavioural states.

The proposed framework is posited to address this limitation by integrating multiple behavioural modalities, including speech, eye gaze, gestures, facial expressions, learning history, and interaction patterns within virtual environments. Through multi-modal machine learning, the synthesis of these heterogeneous data streams may be leveraged to produce a more comprehensive understanding of learner engagement, cognitive load, and learning progress. This richer representation facilitates the provision of more accurate instructional decisions, timely feedback, and personalized learning experiences. Consequently, the framework may be understood as extending current approaches to immersive language learning by positioning multi-modal behavioural analytics as a central mechanism for intelligent personalization rather than as a supplementary analytical tool.

Bridging Educational Psychology and Machine Learning

The final theoretical contribution is the development of an interdisciplinary bridge between educational psychology and machine learning. A significant portion of the extant research within artificial intelligence-driven education frequently emphasizes algorithmic

performance, predictive accuracy, or technological innovation, yet it concurrently allocates comparatively limited consideration to established theories of learning and human development. Conversely, educational psychology has traditionally concentrated its efforts upon learner cognition, motivation, and instructional design, often without a comprehensive interrogation of the capabilities inherent in contemporary artificial intelligence.

The proposed framework may facilitate the convergence of these domains by embedding pedagogical principles within AI decision-making processes. Machine learning techniques such as Natural Language Processing, Reinforcement Learning, and multi-modal fusion are conceptualized not merely as computational tools but as mechanisms for operationalizing educational theories in real-time instructional contexts. Through this integration, adaptive learning systems are enabled to render pedagogically informed decisions, which concurrently support learner motivation, the optimization of cognitive processing, and facilitate meaningful interaction within immersive environments. Consequently, the framework may be interpreted as contributing to the nascent field of human-centered artificial intelligence in education by demonstrating the mechanisms through which intelligent systems could be designed to enhance both technological sophistication and pedagogical effectiveness.

Overall, the proposed conceptual framework may advance existing knowledge by providing a theoretically grounded and technologically integrated model for Artificial Intelligence (AI)-supported immersive English language learning. It establishes a foundation for future empirical research while offering practical guidance for the development of next-generation intelligent tutoring systems capable of delivering personalized, adaptive, and learner-centered educational experiences.

Implications of the Study

Implications for Educators

The framework appears to fundamentally redefine the role of the instructor in immersive, technology-mediated language classrooms, shifting their primary responsibility from basic administrative monitoring to targeted, high-value pedagogical guidance.

In traditional digital settings, a single instructor cannot simultaneously orchestrate tailored, real-time learning trajectories for dozens of students (Herawati, 2023, p. 169). The framework functions, in essence, as an automated instructional co-pilot. Through the facilitation of dynamic environmental parameter modification and the adjustment of scaffolding levels contingent upon individual cognitive states, the implementation may enable immediate micro-personalization. This allows educators to step back from rigid lesson-pacing management and instead focus on providing nuanced, human-centric mentoring tailored to each student's unique learning profile.

Instead of relying on delayed, summative test instruments that only uncover student deficiencies after a learning block has ended, the integrated NLP and cross-modal engine provides teachers with immediate diagnostics. Instructors receive automatically synthesized, granular performance matrices that detail precise phonemic articulation errors, syntactic patterns, and biometric indicators of cognitive anxiety during conversational speech. This capability may enable highly localized and timely remedial interventions.

One of the most persistent bottlenecks in modern language education is the exhaustive workload associated with grading conversational fluency, speech rhythms, and free-form dialogues (Yusuf et al., 2025). By delegating iterative pronunciation drilling, routine vocabulary testing, and initial conversational micro-scaffolding to the closed-loop AI engine, the framework may substantially mitigate administrative exhaustion. Consequently, educators may reallocate their capacity towards the cultivation of more profound student-teacher relationships and the development of intricate collaborative assignments.

Implications for Universities and Higher Education Institutions

For institutional leaders navigating the competitive landscape of internationalized higher education, the framework may afford a structured blueprint for upgrading digital infrastructure and expanding academic reach.

- **The Evolution of Smart Language Laboratories:** Traditional university language labs may often be limited to passive desktop workstations running static software (Al-Jarf, 2021, p. 74; Brenes, 2011, p. 3) This framework offers a strategic model for the next generation of "Smart Language Laboratories." By equipping physical facilities with modern spatial hardware integrated with responsive multi-modal AI agents, universities may facilitate the evolution of their campuses into advanced centres for experiential learning and automated linguistic assessment.

- **AI-Supported Distance Learning Parity:** In global online distance education, remote students often face severe isolation and restricted access to real-time, interactive language practice (Sukirman, 2023, p. 9; Sun et al., 2023, p. 1289577) This architectural configuration directly addresses this extant equity gap. By embedding an empathetic, intelligent learning partner inside accessible virtual spaces, institutions may thereby provide online distance learners with conversational engagement comparable to the quality of face-to-face, on-campus instruction. This advancement could effectively surmount geographical impediments to academic success.

- **Scalable and Cost-Effective Language Education:** Recruiting and retaining specialized, native-speaking instructors to manage small conversational groups is financially unfeasible at scale (Galvis, 2018, p. 27). This framework, consequently, may present an infinitely scalable alternative. The deployment of the unified adaptive framework across massive open online courses (MOOCs) or expansive international programs may enable a university to deliver highly personalized, context-aware instruction to thousands of global students simultaneously without incurring unsustainable faculty expansion costs.

Implications for Software Developers and System Engineers

For the technical architects tasked with the development of next-generation educational software, the framework appears to translate abstract learning theories into explicit, structured engineering priorities.

- **Design Principles for Intelligent Tutoring Systems (ITS):** Modern spatial developers may lack coherent, pedagogically grounded design frameworks when building educational software, a deficit which appears to result in applications prioritizing gamified visual gimmicks over the optimization of learning outcomes. This framework appears to

establish explicit, multi-layered design rules for immersive ITS architectures. It appears to delineate how telemetry inputs may map directly onto algorithmic decision tiers and rendering directives, thereby potentially standardizing the creation of education-first software.

- **Operationalizing Human-Centrred AI:** In alignment with modern design principles, this framework appears to move past traditional "black-box" artificial intelligence implementations that prioritize algorithmic speed or accuracy at the expense of the human experience. Instead, it appears to embed **Human-Centred AI** directly into the core engine architecture. By explicitly weighting the system's reinforcement learning reward functions to balance user cognitive capacity (Cognitive Load Theory [CLT]) and protect emotional autonomy (Self-Determination Theory [SDT]), it may create an engineering precedent where algorithmic optimization is fundamentally tethered to human well-being.

- **Ethical Multimodal Analytics and Privacy-by-Design:** The constant collection of high-frequency biometric streams—such as eye gaze, facial movements, and vocal patterns—may present significant ethical challenges regarding user privacy and data security. This framework addresses these challenges through the advocacy of a strict Privacy-by-Design methodology. Its provision of an ethical blueprint delineates parameters for implementation, thereby detailing specifications for secure, localized on-device processing of eye-tracking and spatial telemetry. By transforming raw biometric signals into anonymized cognitive state vectors before routing data to the central engine, the system appears to ensure compliance with global data protection mandates while preventing the exploitation of sensitive personal telemetry.

Future Research

The proposed AI-driven, multimodal, adaptive VR and AR framework delineates several significant avenues for future research. Although the conceptual model is grounded in established educational theories and contemporary machine learning approaches, its practical value warrants examination through systematic empirical investigation. Future studies may, therefore, concentrate upon the validation of the framework, the assessment of its instructional efficacy, and the exploration of the broader ethical and contextual conditions that may impinge upon its implementation in language learning environments.

Empirical Validation of the Proposed Framework

The empirical validation of the proposed framework within authentic educational settings constitutes a primary research imperative. Conceptual models provide theoretical direction, but their efficacy necessitates demonstration through rigorous quantitative, qualitative, or mixed-methods studies. Future researchers may therefore examine whether the integration of NLP, RL, and multi-modal behavioural analytics demonstrably enhances learner engagement, language performance, and adaptive feedback quality in VR/AR-based English learning environments. Experimental and quasi-experimental designs could, therefore, be utilized for the comparison of the proposed framework with conventional digital language-learning approaches, thereby ascertaining its relative instructional value.

In addition, future inquiries might elucidate how learners and educators perceive the usability, relevance, and pedagogical effectiveness of the framework. Evaluations rooted in User-Centred Design (UCD), interviews, and classroom observations would provide valuable insights into the system's functional efficacy in operational settings, alongside the potential necessity of design refinements. Such empirical validation appears essential for the progression of the framework from theoretical proposition to evidence-based educational innovation.

Examining Long-Term Language Retention

An additional trajectory for future inquiry involves the examination of long-term language retention. While many technology-enhanced learning studies focus on immediate performance gains, there exists a paucity of empirical validation regarding the sustainability of these gains over time (Mukherjee, 2026). Future studies might therefore endeavour to assess whether learners who use the proposed adaptive VR and AR framework retain vocabulary, grammar, pronunciation, and communicative skills more effectively than those who learn through traditional or non-adaptive digital methods.

Longitudinal research designs would be particularly valuable in this regard. Through the tracking of learner progress across extended periods, it may be determined whether intelligent adaptive scaffolding contributes not only to short-term achievement but also to durable language development. Such studies would help clarify whether immersive multi-modal learning environments appear to support deeper cognitive processing and more stable knowledge retention.

Comparing Different Machine Learning Algorithms

Future research might also necessitate a comparative analysis of machine learning algorithms; such an undertaking could ascertain the approaches that prove most effective for adaptive language learning. The proposed framework is amenable to implementation via a range of computational techniques, including supervised learning, reinforcement learning, deep learning, and hybrid multi-modal models. However, each algorithm may differ in its ability to process learner data, predict performance, and generate personalized instructional responses (Zawacki-Richter et al., 2019, p. 11).

The execution of comparative studies appears necessary to evaluate the strengths and limitations of these algorithms within educational contexts. For example, researchers could examine whether reinforcement learning might facilitate more efficacious adaptive sequencing compared to rule-based systems or whether multi-modal deep learning models provide more accurate learner profiling than single-modality approaches. Such investigations could potentially contribute to both educational technology research and applied machine learning by identifying the most suitable computational strategies for intelligent language instruction.

Investigating Ethical AI and Algorithmic Fairness

As AI systems increasingly influence educational decision-making, future research might address ethical concerns and algorithmic fairness. The proposed framework appears to rely on the collection and analysis of sensitive learner data, including speech, gaze, gestures, and behavioural interactions. This may raise important questions about privacy, informed

consent, transparency, and the assurance of data security; indeed, such concerns appear foundational. Future studies should therefore examine how ethical principles can be embedded into the design and deployment of multi-modal educational systems.

Researchers might also investigate whether adaptive algorithms produce fair outcomes across different learner groups. The emergence of bias appears probable if machine learning models are trained upon unrepresentative datasets or if they interpret behavioural cues disparately across linguistic, cultural, or demographic contexts. Future work should assess whether the framework performs equitably for learners with diverse backgrounds, abilities, and language proficiency. Addressing these issues is essential for ensuring that Artificial Intelligence-supported education remains inclusive, trustworthy, and socially responsible.

Exploring Multi-modal Adaptation Across Cultures

Another promising area for future research is the exploration of multi-modal adaptation across cultures. Learner behaviour, communication styles, and interpretations of gestures, gaze, and facial expressions may vary significantly across cultural contexts (Endra & André, 2014, p. 227). As a result, multimodal AI systems developed in one setting may not function identically in another. The examination of the extent to which cultural variance influences the accuracy and efficacy of behavioural analytics within immersive language learning environments may, therefore, be warranted.

Cross-cultural research could investigate whether the framework requires localization or adaptation to support learners from different linguistic and cultural backgrounds. Such studies might elucidate how cultural norms shape learner interaction with VR/AR environments and how adaptive systems can be designed to respect diversity while maintaining pedagogical effectiveness. This line of inquiry may prove particularly vital for the global deployment of AI-driven educational technologies.

Evaluating Implementation in Higher Education

The evaluation of the implementation of the proposed framework within higher education settings may also be warranted. Universities may represent a particularly relevant context for this model because they increasingly rely on digital learning infrastructures, international student populations, and scalable instructional solutions **(**Ortiz et al., 2021, p. 168**)**. Future studies should examine how the framework performs in undergraduate, postgraduate, and continuing education programs, especially in English as a Foreign Language (EFL) and English as a Second Language (ESL) contexts.

Implementation studies might explore institutional readiness, technological infrastructure, faculty acceptance, and student engagement. Researchers could also assess the extent to which the framework might be integrated into existing curricula, language laboratories, and online learning platforms. Such evaluations would likely provide practical evidence regarding feasibility, scalability, and sustainability in real university environments. They could also assist in the identification of the organizational and pedagogical conditions necessary for successful adoption.

Conclusion

This conceptual paper delineates a potentially novel AI-driven multi-modal adaptive framework for English language acquisition in immersive VR and AR environments. Unlike existing models that employ isolated technologies, the proposed framework may facilitate the integration of Natural Language Processing, Reinforcement Learning, multi-modal behavioural fusion, and adaptive instructional scaffolding into a unified architecture. Grounded in Self-Determination Theory, Embodied Cognition Theory, Scaffolding Theory, and Cognitive Load Theory, the framework elucidates how intelligent systems may personalize learning based on learners' linguistic performance and behavioural cues. This paper may advance current educational technology literature through the provision of a theoretically grounded blueprint for the design of next-generation intelligent immersive learning environments and may establish a foundation for future empirical validation.

References

Aleven, V., McLaren, B. M., Sewall, J., Velsen, M. van, Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-Tracing Tutors: Intelligent tutor development for non-programmers. International Journal of Artificial Intelligence in Education, 26(1), 224–269. https://doi.org/10.1007/s40593-015-0088-2

Al-Jarf, R. (2021). Feasibility of digital multimedia language labs for interpreting instruction as perceived by interpreting instructors in Saudi Arabia. International Journal of Linguistics Literature & Translation, 4(4), 70–79. https://doi.org/10.32996/ijllt.2021.4.4.8

Aziz, N. A. N. B., Hanif, Z. S. A. B. M., Kasim, V. N. B. M., Baharin, N. N. B., & Manap, M. R. (2025). Tackling presentation anxiety among ESL university students. International Journal of Research and Innovation in Social Science, 9(22), 224–229. https://doi.org/10.47772/ijriss.2025.922ileiid0023

Brenes, C. A. N. (2011). The language laboratory and the EFL course. Actualidades Investigativas en Educación, 6(2). https://doi.org/10.15517/aie.v6i2.9209

Bulathwela, S., Niekerk, D. van, Shipton, J., Pérez-Ortiz, M., Rosman, B., & Shawe-Taylor, J. (2025). TrueReason: An exemplar personalised learning system integrating reasoning with foundational models. In ArXiv.org. https://doi.org/10.48550/arxiv.2502.10411

Chen, Y., Kao, T. C., & Sheu, J. (2003). A mobile learning system for scaffolding bird watching learning. Journal of Computer Assisted Learning, 19(3), 347–359. https://doi.org/10.1046/j.0266-4909.2003.00036.x

Chen, Y. T., & Dang, J. (2025). Developing a virtual reality system integrated with large language models for real-time evaluation and feedback to improve public speaking skills. International Journal of Research and Innovation in Social Science, 165–176. https://doi.org/10.5121/csit.2025.151713

Chowdhury, S. P., Zouhar, V., & Sachan, M. (2024). AutoTutor meets large language models: A language model tutor with rich pedagogy and guardrails. 5–15. https://doi.org/10.1145/3657604.3662041

Chun, D. M., Karimi, H., & Sañosa, D. J. (2022). Traveling by headset. CALICO Journal, 39(2), 129–149. https://doi.org/10.1558/cj.21306

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and Education: The Self-Determination Perspective. Educational Psychologist, 26(3), 325–346. https://doi.org/10.1207/s15326985ep2603&46

Endra, B., & André, E. (2014). Integration of cultural factors into the behavioral models of virtual characters. In OPUS (Augsburg University) (pp. 227–251). Augsburg University. https://doi.org/10.1017/cbo9780511844492.010

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2015). Concept representation reflects multimodal abstraction: A framework for embodied semantics. Cerebral Cortex, 26(5), 2018–2034. https://doi.org/10.1093/cercor/bhv020

Galvis, Á. (2018). Supporting decision-making processes on blended learning in higher education: literature and good practices review. International Journal of Educational Technology in Higher Education, 15(1). https://doi.org/10.1186/s41239-018-0106-1

Hajji, M. E., Baha, T. A., Berka, A., Nacer, H. A., Aouifi, H. E., & Es-Saady, Y. (2025). An architecture for intelligent tutoring in virtual reality: Integrating LLMs and multimodal interaction for immersive learning. Information, 16(7), 556–556. https://doi.org/10.3390/info16070556

Herawati, A. (2023). Implementing personalized learning in universities classrooms: Lecturers' challenges and perceptions. Humaniora, 14(2), 167–172. https://doi.org/10.21512/humaniora.v14i2.8890

Huang, X., Zou, D., Cheng, G., & Xie, H. (2021). A systematic review of AR and VR enhanced language learning. Sustainability, 13(9), 4639–4639. https://doi.org/10.3390/su13094639

Ipsita, A., Patel, M., Unmesh, A., & Ramani, K. (2024). Authoring instructional flow in VR learning units to promote outcome-oriented learning. Computers & Education X Reality, 5, 100074–100074. https://doi.org/10.1016/j.cexr.2024.100074

Ironsi, C. S. (2023). Investigating the use of virtual reality to improve speaking skills: insights from students and teachers. Smart Learning Environments, 10(1). https://doi.org/10.1186/s40561-023-00272-8

Kaplan-Rakowski, R., & Gruber, A. (2023). The impact of high-immersion virtual reality on foreign language anxiety. Smart Learning Environments, 10(1). https://doi.org/10.1186/s40561-023-00263-9

Kishor, I., Mamodiya, U., Almaayah, M., Alqutaish, A., Shehab, R., & Aldhyani, T. H. H. (2025). Agentic AI-Enhanced Virtual Reality for adaptive immersive learning environments. Mesopotamian Journal of Computer Science, 2025, 398–416. https://doi.org/10.58496/mjcsc/2025/026

Koning, B. B. de, & Tabbers, H. K. (2011). Facilitating understanding of movements in dynamic visualizations: An embodied perspective. Educational Psychology Review, 23(4), 501–521. https://doi.org/10.1007/s10648-011-9173-8

Lampropoulos, G. (2025). Augmented Reality, Virtual Reality, and Intelligent Tutoring Systems in Education and Training: A Systematic Literature review. Applied Sciences, 15(6), 3223. https://doi.org/10.3390/app15063223

Ludwig-Hardman, S., & Dunlap, J. C. (2003). Learner support services for online students: Scaffolding for success. The International Review of Research in Open and Distributed Learning, 4(1). https://doi.org/10.19173/irrodl.v4i1.131

Mukherjee, P. K. (2026). Re-examining technology integration in education: Cognitive foundations, pedagogical implications, and contextual constraints. International Journal of Novel Trends and Innovation, 4(3). https://doi.org/10.56975/ijnti.v4i3.231769

Ortiz, L. G., Clougher, D., Anderson, T., & Maina, M. F. (2021). IDEAS for transforming higher education: An overview of ongoing trends and challenges. The International Review of

Research in Open and Distributed Learning, 22(2), 166–184. https://doi.org/10.19173/irrodl.v22i2.5206

Pappas, M. A., & Drigas, A. (2016). Incorporation of Artificial Intelligence tutoring techniques in Mathematics. International Journal of Engineering Pedagogy (iJEP), 6(4), 12–12. https://doi.org/10.3991/ijep.v6i4.6063

Ruan, S., & Lu, K. (2025). Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization. Computers and Education Artificial Intelligence, 9, 100463–100463. https://doi.org/10.1016/j.caeai.2025.100463

Sajja, R., Sermet, Y., Cwiertny, D. M., & Demir, İ. (2025). Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education. Technology Knowledge and Learning. https://doi.org/10.1007/s10758-025-09897-9

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N. S., Tripoliti, E. E., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. IEEE Reviews in Biomedical Engineering, 16, 260–277. https://doi.org/10.1109/rbme.2021.3066072

Skulmowski, A., & Xu, K. M. (2021). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. Educational Psychology Review, 34(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Son, J., Ružić, N. K., & Philpott, A. (2023). Artificial intelligence technologies and applications for language learning and teaching. Journal of China Computer-Assisted Language Learning, 5(1), 94–112. https://doi.org/10.1515/jccall-2023-0015

Sukirman, S. (2023). Exploring the impacts of the covid-19 pandemic on English language teachers. Vision Journal for Language and Foreign Language Learning, 12(1), 1–16. https://doi.org/10.21580/vjv12i216998

Sun, L., Dong, H., & Zhang, X. (2023). Innovative solutions for language growth: the impact of problem-based learning via DingTalk on Chinese undergraduates' business vocabulary amid COVID-19. Frontiers in Psychology, 14, 1289575–1289575. https://doi.org/10.3389/fpsyg.2023.1289575

Sweller, J., Merriënboer, J. J. G. van, & Paas, F. (2019). Cognitive architecture and instructional design: 20 Years Later. Educational Psychology Review, 31(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Syed, T. A., Khan, S., Jan, S., & Nauman, M. (2025). Revolutionizing arabic learning: GNT and AI-enhanced Metaverse environments. Frontiers in Computer Science, 7. https://doi.org/10.3389/fcomp.2025.1626092

Yan, L., Echeverría, V., Jin, Y., Fernandez-Nieto, G. M., Zhao, L., Li, X., Alfredo, R., Swiecki, Z., Gašević, D., & Martínez-Maldonado, R. (2024). Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. British Journal of Educational Technology, 55(5), 1900–1925. https://doi.org/10.1111/bjet.13498

Yan, W., & Lowell, V. L. (2025). The Evolution of Virtual Reality in foreign language education: From Text-Based MUDs to AI-Enhanced Immersive Environments. TechTrends, 69(4), 853–858. https://doi.org/10.1007/s11528-025-01086-w

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Towards reducing teacher burden in Performance-Based assessments using aivaluate: an emotionally intelligent LLM-Augmented pedagogical AI conversational agent. Education and Information Technologies, 30(17), 24649–24693. https://doi.org/10.1007/s10639-025-13755-7

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1). https://doi.org/10.1186/s41239-019-0171-0