Global Education Journal
Vol. 4, 1 (January-April, 2026), pp. 691-706
EISSN: 2986-6529, DOI: https://doi.org/10.59525/gej.2039

LLM-Driven Cognitive Digital Twins in Education: A Systematic Review and Conceptual Framework

Fadliyani Nawir

Institut Bisnis dan Keuangan Nitro, Makassar, Indonesia

ARTICLE INFO

Keywords:

large language models;
cognitive digital twin;
learner modeling;
trustworthy AI;
retrieval-augmented generation

Article history:

Received 2026-02-26
Revised 2026-03-28
Accepted 2026-04-30

ABSTRACT

Within educational technology, the adoption of Large Language Models (LLMs) and Generative AI (GenAI) has accelerated notably, yielding applications spanning intelligent tutoring systems, adaptive assessment pipelines, and automated formative feedback. Yet the fusion of these generative capabilities with Cognitive Digital Twin (CDT) architectures has proceeded unevenly, leaving theoretical and practical gaps unaddressed. This study reports a systematic literature review conducted in accordance with PRISMA 2020 principles. After screening 1,400 records from ScienceDirect and IEEE Xplore, 111 Core and Supporting studies were included in the qualitative synthesis. A narrative thematic synthesis was then carried out by mapping the included studies to five research questions and five thematic clusters. The five thematic clusters comprise CDT and learner modeling, LLM and GenAI tutoring, adaptive assessment and feedback, ethics and explainability, and learning analytics. The analysis reveals that automated feedback and assessment constitute the most densely evidenced application domains, whereas learner modeling and CDT remain underdeveloped as sustained personalization mechanisms. The paper's primary contribution is a five-layer LLM-driven CDT framework comprising a Learner Data Layer, a Learner Digital Twin Layer, an LLM Intelligence Layer, an Adaptive Learning Service Layer, and a cross-cutting Trust, Ethics, and Governance Layer. Taken together, these layers constitute a learner-aware, adaptive, explainable, and human-centered architecture intended to ground and direct future empirical inquiry.

Corresponding Author:
Fadliyani Nawir
Institut Bisnis dan Keuangan Nitro, Makassar, Indonesia; nfadliyani@gmail.com

1. INTRODUCTION

Generating coherent explanations, sustaining open-ended instructional dialogue, producing formative feedback, and flexibly adapting output to pedagogical prompts are among the documented capabilities that modern LLMs bring to educational contexts (Brown et al., 2020; Kasneci et al., 2023). Alongside these generative technologies, a parallel body of work has developed mechanisms for

modeling and representing learner states, encompassing learning analytics, adaptive learning systems, and educational digital twin proposals (Kinsner, 2021; Yamin et al., 2023). Within this review, a Cognitive Digital Twin is understood as a continuously updated computational representation of an individual learner that encodes the learner's current knowledge state, identifiable misconceptions, historical progress, behavioral event traces, and relevant cognitive or affective signals (Aboulsafa et al., 2023; Yamin et al., 2023).

Despite their individual maturity, these two streams have yet to be brought into productive alignment. Session-bound LLM deployments predominate in practice, meaning that no auditable record of the learner's history is preserved across interactions (Makharia et al., 2024). Conversely, digital twin implementations seldom incorporate the generative and explanatory affordances that LLMs provide (Yamin et al., 2023). Furthermore, reliability concerns and the risk of hallucinated feedback continue to shadow deployed assessment systems (Ji et al., 2023), while governance and ethical considerations are routinely positioned as external policy constraints rather than first-class elements of system design (European Parliament and Council of the European Union, 2024). It is in response to these persistent structural gaps that this review undertakes a systematic synthesis aimed at producing an integrated LLM-driven CDT framework for educational deployment.

Table 1. Research Questions and Motivation

RQ	Question	Motivation
RQ 1	How are LLMs/GenAI used for intelligent tutoring?	Map tutors, chatbots, assistants, scaffolding agents
RQ 2	How are LLMs/GenAI used for adaptive assessment?	Identify grading, quiz generation, short-answer/programming assessment
RQ 3	How are LLMs/GenAI used for feedback generation?	Map formative, corrective, peer, and programming feedback
RQ 4	How are learner modeling and CDT represented?	Assess personalization, knowledge tracing, digital twin
RQ 5	What ethics, trust, and explainability concerns emerge?	Identify privacy, bias, hallucination, integrity, overreliance, XAI

Three distinct contributions follow from this work. First, the paper delivers a focused SLR synthesis drawing on 111 Core and Supporting studies. Second, it constructs an evidence-grounded gap map that traces the relationships among LLM services, learner modeling, assessment and feedback mechanisms, and governance considerations. Third, it proposes a five-layer conceptual framework through which LLM-driven CDT may be systematically realized within educational settings.

1.1. Conceptual Background and Related Work

LLMs and GenAI in Education

Survey literature on GenAI and LLMs in educational settings has concentrated primarily on applied domains: tutoring dialogue, writing assistance, programming support, assessment automation,

and the implications for academic integrity (Hadi Mogavi et al., 2024; Kasneci et al., 2023; Vargas-Murillo et al., 2023). Although these accounts usefully map where adoption is occurring, they rarely examine how LLM-generated outputs might be anchored in, or fed back into, persistent representations of the learner or digital twin architectures (Hadi Mogavi et al., 2024; Kasneci et al., 2023).

Learner Modeling, Digital Twins, and Adaptive Learning

Research on educational digital twins has been oriented largely toward constructing digital representations that enable simulation and personalization at the individual learner level (Aboulsafa et al., 2023; Kinsner, 2021; Yamin et al., 2023). Related work in learning analytics and adaptive learning has established predictive modeling, recommendation, and self-regulated learning support as core functional goals (Alfredo et al., 2024; Gašević et al., 2015). What is conspicuously absent across both bodies of work is a systematic treatment of how LLM-enabled tutoring, assessment generation, and automated feedback might interact with or update a maintained learner digital twin state.

Positioning of This Review

Table 2. Positioning Against Related Review Streams

Review Stream	LLM	CDT/LM	Ass./FB	Ethics	LLM-CDT
GenAI/ChatGPT in education	High	Low	Med	Med	No
Educational digital twin	Low	High	Low	Low	No
Learning analytics	Low	Med	Med	Low	No
XAI/trustworthy	Med	Low	Low	High	No
This review	**High**	**High**	**High**	**High**	**Yes**

2. METHODS

2.1. Protocol, Databases, and Search Strategy

Methodologically, the study adhered to the PRISMA 2020 reporting framework in conjunction with the systematic literature review guidelines established by Kitchenham and colleagues (Kitchenham, 2007; Page et al., 2021). All database searches and eligibility screening were carried out during June 2026, drawing exclusively on ScienceDirect and IEEE Xplore. A Boolean query structure was employed, combining terminology drawn from the LLM and GenAI domain with education-related and digital twin or learner model vocabulary, with scope restricted to English-language, peer-reviewed contributions published between 2019 and 2026.

To support reproducibility, Table III reports the database-specific search configuration, including the search fields, the simplified Boolean string applied to each source, and the filters used.

Table 3. Search Strategy per Database

Database	Fields	Search String (simplified)	Filters
ScienceDirect	Title, abstract, keywords	("large language model" OR "LLM" OR "generative AI") AND ("education" OR "tutoring" OR "assessment") AND ("digital twin" OR "learner model")	2019–2026; English; articles and reviews
IEEE Xplore	Metadata, abstract, index terms	("large language model" OR "LLM" OR "generative AI") AND ("education" OR "tutoring" OR "feedback") AND ("digital twin" OR "learner modeling")	2019–2026; English; conferences and journals

2.2. Inclusion, Exclusion, and Quality Appraisal

Eligibility was determined by four classes of criterion. At the document-type level, inclusion was limited to peer-reviewed journal articles and refereed conference papers. Topically, studies were required to address at least one of the following: LLMs or GenAI, intelligent tutoring, learner assessment, feedback generation, learner modeling, CDT, or ethics in educational contexts, and to contribute relevant evidence for one or more of RQ1 through RQ5. Evidence types accepted spanned empirical investigations, conceptual proposals, system-design reports, and structured reviews. For quality appraisal, each included study was rated on five dimensions: clarity of objective, methodological appropriateness, specificity of educational context, adequacy of evaluation or evidence, and transparency regarding limitations and risks. Scores between 4.0 and 5.0 indicated high quality, scores from 2.5 to 3.5 indicated moderate quality. To address the inherent limitations of single-reviewer screening, four procedural safeguards were maintained throughout: structured decision logging, iterative consistency re-checks, rubric anchoring against pre-defined examples, and full workbook transparency.

Applying these criteria, the appraisal classified 65 studies as high quality and 46 as moderate quality; no included study fell below the moderate threshold, so none was excluded on quality grounds. Table IV summarizes this distribution.

Table 4. Quality Appraisal Results

Quality Level	Score Range	Studies
High	4.0–5.0	65
Moderate	2.5–3.5	46
Low	0–2.0	0

2.3. PRISMA Flow and Data Extraction

Database searching returned 1,400 records in total, distributed across ScienceDirect (n = 602) and IEEE Xplore (n = 798). Removal of within-database duplicates reduced the pool to 1,185 unique records. Screening at the title and abstract stage eliminated a further 1,060 records, leaving 125 that proceeded to full-text review. Following full-text assessment, 111 studies entered the qualitative synthesis: 69 classified as Core contributions and 42 as Supporting. Data extraction for each study encompassed bibliographic metadata, study type, educational context, specific AI technology, mapping to one or more research questions, primary thematic cluster, main findings, research method, reported evaluation metrics, acknowledged limitations, and the study's particular contribution to the proposed framework.

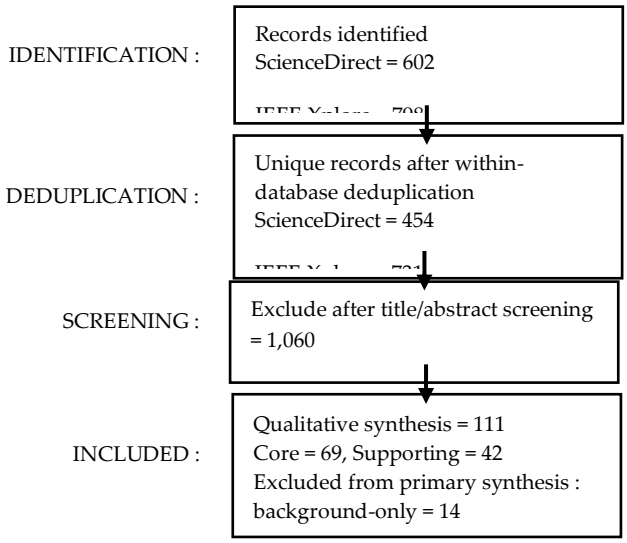

Figure 1. PRISMA flow diagram for ScienceDirect and IEEE Xplore searches

3. FINDINGS AND DISCUSSION

3.1. Bibliometric and Descriptive Results

Among the 111 included studies, 58 originated from ScienceDirect and 53 from IEEE Xplore, with 69 designated as Core contributions and 42 as Supporting. A pronounced temporal concentration is observable: 87 studies, representing 78.4 percent of the corpus, were published in 2024 alone, a pattern consistent with the surge in scholarly interest that followed the public availability of ChatGPT and GPT-4 (Brown et al., 2020; Hadi Mogavi et al., 2024; Kasneci et al., 2023). By thematic cluster, assessment and feedback activities attracted the greatest volume of primary evidence, with Cluster C3 accounting for 49 studies. Cluster C1, covering CDT and learner modeling, comprised 25 primary studies, while Cluster C2, focused on LLM-based tutoring, included 16. Cluster C4 (ethics) yielded a primary count of 12 studies. However, this figure understates the actual ethical coverage because ethics functioned as a cross-cutting concern embedded within studies whose primary classification fell elsewhere, a pattern that accounts for the substantially higher RQ5 evidence count of 75 studies.

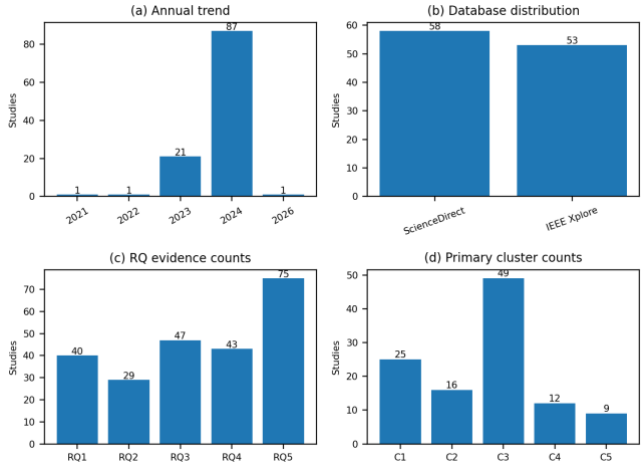

Figure 2. Bibliometric dashboard: publication year trend, database distribution, RQ evidence counts (multi-label), and primary cluster distribution across 111 included studies.

3.2. Thematic Synthesis: Answering RQ1–RQ5

Table 5. RQ-Based Synthesis Summary

RQ	N	Dominant Applications	Key Gap
RQ1	40	AI tutor, chatbot, course assistant, programming tutor, scaffolding agent	Most systems are session-based and weakly connected to persistent learner models.
RQ2	29	Quiz generation, grading, short-answer assessment, programming assessment	Reliability, fairness, and expert validation remain unresolved.
RQ3	47	Formative feedback, corrective feedback, peer feedback, programming feedback	Feedback may be generic, misleading, or hallucinated.
RQ4	43	Learner profiles, knowledge tracing, knowledge graphs, educational digital twins	LLM-driven CDT architectures remain rare and weakly validated.
RQ5	75	Privacy, bias, hallucination, academic integrity, explainability, overreliance.	Governance remains fragmented and rarely embedded into system design.

3.2.1. RQ1: Intelligent Tutoring

A growing range of tutoring modalities has been realized through LLMs, including conversational tutors, general-purpose learning assistants, domain-specific programming tutors, and dialogue-based scaffolding agents (Frank et al., 2024; Lui et al., 2024; Makharia et al., 2024). Among representative contributions, Makharia et al. (2024) couple an AI tutor with a deep knowledge tracing component, while Frank et al. (2024) report a quantitative evaluation of a GenAI tutoring system for the R programming language. A consistent limitation, however, runs through virtually the entire body of work: implemented systems remain session-bound, with no mechanism for preserving, updating, or querying a record of the learner's state across interactions. This architectural shortcoming constrains longitudinal adaptation and forestalls the evolution of AI tutoring toward a genuinely learner-aware digital twin service.

3.2.2. RQ2: Adaptive Assessment

Adaptive assessment constitutes one of the two most active applied domains in the corpus. Gopi et al. (2024) advance a retrieval-augmented generation approach to adaptive quiz construction, while Cagliero et al. (2024) assess ChatGPT's capacity for the automated correction of SQL exercises. Across this work, LLM-based assessment is shown to be technically feasible at scale; what remains unresolved is reliability and fairness, since grading consistency, rubric alignment, and expert validation are not yet systematically guaranteed (Cagliero et al., 2024; Ji et al., 2023).

3.2.3. RQ3: Feedback Generation

Feedback generation is the single most densely evidenced application area. Ouyang et al. (2024) conduct a comparative study of instructor-generated versus ChatGPT-generated feedback on collaborative programming tasks; Jürgensmeier and Skiera (2024) demonstrate how LLMs can deliver scalable feedback across multimodal assessment formats; and Lin and Crosthwaite (2024) examine the relative merits of teacher-led and GPT-assisted written corrective feedback. Taken together, these studies confirm feasibility but converge on a shared safety concern: hallucinated content, under-specified commentary, and inconsistent quality remain recurrent risks, and expert validation before learner exposure is frequently absent (Ji et al., 2023).

3.2.4. RQ4 and RQ5: Learner Modeling, CDT, and Governance

The representational infrastructure for learner-aware systems draws on a set of foundational contributions in learner modeling and educational digital twins (Aboulsafa et al., 2023; Abu-Rasheed et al., 2025; Corbett & Anderson, 1994; Yamin et al., 2023). Abu-Rasheed et al. (2025) illustrate how knowledge graphs can serve as context sources for LLM-driven learning recommendations, Yamin et al. (2023) explore the applicability of digital twin concepts to lifelong learning scenarios, and Corbett and Anderson (1994) supply the classical probabilistic knowledge tracing model that underpins much of the field. Questions of ethics, trust, and explainability surface repeatedly across the corpus, though predominantly as secondary considerations attached to studies whose primary focus lies elsewhere (Ji et al., 2023; Khosravi et al., 2022; Susnjak, 2024). Susnjak (2024) offers a notable exception, extending XAI principles toward prescriptive analytics in educational settings, while the regulatory landscape is substantially shaped by the EU AI Act (European Parliament and Council of the European Union, 2024).

3.3. Cluster-Based Synthesis

Cluster	Theme	N	Key Refs	Framework Contribution
C1	CDT & learner modeling	25	(Aboulsafa et al., 2023; Abu-Rasheed et al., 2025; Corbett & Anderson, 1994; Yamin et al., 2023)	Learner Modeling / Digital Twin Layer
C2	LLM/GenAI tutoring	16	(Frank et al., 2024; Lui et al., 2024; Madhubhashana et al., 2024; Makharia et al., 2024)	LLM Intelligence Layer (RAG)
C3	Assessment & feedback	49	(Cagliero et al., 2024; Gopi et al., 2024; Jürgensmeier & Skiera, 2024; Lin & Crosthwaite, 2024; Ouyang et al., 2024)	Adaptive Learning Service Layer
C4	Ethics, trust, XAI	12	(European Parliament and Council of the European Union, 2024; Ji et al., 2023; Khosravi et al., 2022; Susnjak, 2024)	Trust, Ethics & Governance Layer
C5	Learning analytics	9	(Alfredo et al., 2024; Dahri et al., 2024; Gašević et al., 2015; Prasetyo & Muslaini, 2025)	Learner Data & Personalization Layer

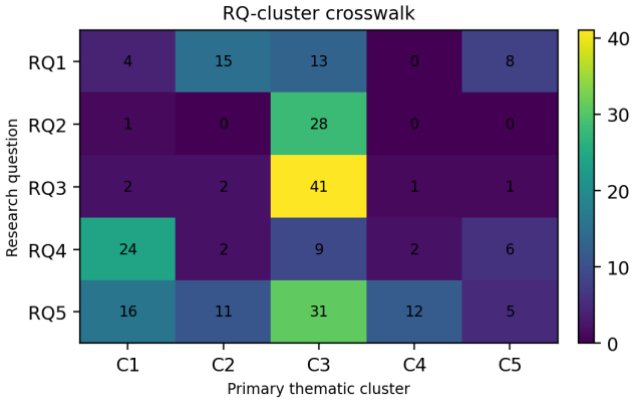

Figure 3. RQ-cluster crosswalk

3.4. Gap, Framework, Evaluation Protocol, and Roadmap

3.4.1. Evidence-Based Gap Map

Table 7. Evidence-Based Gap Matrix

Gap Area	Evidence Pattern	Remaining Gap	Opportunity
LLM tutoring	Chatbots and AI tutors are common	Tutors are often stateless	Integrate tutoring with learner digital twin
Adaptive assessment	Grading and quiz generation are active	Reliability and fairness remain weak	Use explainable, human-validated assessment
Feedback generation	Formative and corrective feedback are frequent	Feedback can hallucinate or mislead	Embed validation and hallucination checks
CDT / learner modeling	Learner models and digital twin components exist	LLM-driven CDT remains rare	Develop integrated learner-aware architecture
Ethics / governance	Privacy, bias, trust, XAI recur widely	Controls fragmented	Add cross-cutting governance layer

Table VII distills the evidence assembled through the RQ-based thematic analysis of Section V into five structured opportunity areas, each capturing what the literature has established, what it has left unresolved, and what architectural response the proposed framework targets. This three-column structure evidence pattern, residual gap, opportunity is designed to make explicit the logical chain that connects reviewed evidence to framework design decisions.

Turning to Gap Area 1, the tutoring literature (RQ1, Cluster C2) documents a rich array of LLM-based implementations: conversational tutors across subject domains, specialized programming tutors, and scaffolding agents designed to guide problem-solving sequences (Frank et al., 2024; Lui et al., 2024; Madhubhashana et al., 2024; Makharia et al., 2024). The shortfall is not one of empirical output but of architectural design. Without exception, deployed systems begin each learner interaction from a blank slate: no prior knowledge state, no record of previously identified misconceptions, no accumulation of longitudinal progress data informs the model's responses (Makharia et al., 2024). The practical consequence is that the LLM's capacity to calibrate explanation complexity, difficulty progression, or pedagogical approach to what a particular learner actually understands remains unrealized. Addressing this gap calls for a coupling mechanism by which tutoring interactions are both grounded in and subsequently update a persistent CDT, converting episodic dialogue sequences into a traceable and continuously personalized learning record.

Gap Area 2 pertains to adaptive assessment (RQ2, Cluster C3). Activity in this domain is considerable, spanning quiz generation through LLM prompting, automated essay and short-answer grading, and programming assignment evaluation (Cagliero et al., 2024; Du et al., 2024; Gopi et al., 2024). Where the literature falls short is in establishing the reliability and fairness of these outputs under realistic deployment conditions. Grading inconsistency across repeated assessments, marked sensitivity to minor variations in prompt formulation, inadequate alignment with expert rubric standards, and the near-complete absence of demographic fairness analysis represent recurring findings (Cagliero et al., 2024; Jürgensmeier & Skiera, 2024). The design opportunity is to treat these as system-level requirements embedded in the assessment pipeline from the outset, rather than as quality checks applied retrospectively after deployment.

Feedback generation (RQ3, Cluster C3) accounts for the densest concentration of primary evidence in this review, with documented applications covering formative academic feedback (Ouyang et al., 2024), written corrective feedback in language learning (Lin & Crosthwaite, 2024), scalable multimodal feedback delivery (Jürgensmeier & Skiera, 2024), and automated programming feedback (Sinha et al., 2024). The volume of production does not, however, indicate solved problems. A substantive safety

gap runs through the feedback literature: susceptibility to hallucination, tendency toward generic or under-specified commentary, and the production of pedagogically plausible but factually inaccurate responses all remain documented risks (Ji et al., 2023). Lin and Crosthwaite (2024) observe with particular clarity that GPT-assisted feedback does not reliably outperform instructor feedback for content requiring nuanced linguistic or disciplinary judgment. The opportunity here lies in treating hallucination detection, provenance attribution, and human review not as optional safeguards but as non-negotiable architectural features, particularly where feedback outcomes carry academic consequences.

Gap Area 4 concerns learner modeling and CDT integration (RQ4, Cluster C1). The existence proof for individual components is well-established: probabilistic knowledge tracing frameworks (Corbett & Anderson, 1994), knowledge-graph-based learning recommendation engines (Abu-Rasheed et al., 2025), conceptual digital twin proposals for educational settings (Aboulsafa et al., 2023; Yamin et al., 2023), and persistent learner profile schemes (Prasetyo & Muslaini, 2025) all appear in the reviewed literature. The problem is one of integration rather than invention: these components exist in relative isolation. Learner models are seldom connected to LLM generative services, digital twin implementations do not typically incorporate LLM-produced tutoring or feedback as inputs, and longitudinal empirical validation of any integrated system remains essentially absent (Aboulsafa et al., 2023; Yamin et al., 2023). The architectural response proposed by this framework positions the Learner Modeling and CDT layer as a persistent state store that conditions all downstream LLM outputs through RAG-mediated retrieval, thereby converting the LLM from a context-free responder into a learner-state-aware reasoning agent.

Gap Area 5 addresses the cluster of concerns grouped under ethics, trust, and governance (RQ5, Cluster C4). Privacy exposure, algorithmic bias, hallucination risk, academic integrity threats, demands for explainability, and learner overreliance on automated outputs are all catalogued across the reviewed literature (European Parliament and Council of the European Union, 2024; Hadi Mogavi et al., 2024; Ji et al., 2023; Khosravi et al., 2022; Sallam, 2023; Susnjak, 2024). What is conspicuously absent is any systematic treatment of governance as an architectural property of the systems under discussion. The prevailing pattern is that governance surfaces as a policy-level external constraint rather than a design element embedded within system components. The EU AI Act establishes binding regulatory requirements (European Parliament and Council of the European Union, 2024), yet the reviewed systems largely do not address how those requirements would be implemented in practice. Susnjak's (2024) demonstration of prescriptive XAI stands as an instructive exception but not as representative practice. The opportunity, taken up in this framework's Trust, Ethics, and Governance Layer, is to treat these considerations as active, system-integrated controls include monitoring, auditing, and intervening across all functional layers rather than leaving enforcement to post-deployment regulatory processes.

Read in sequence, the five gap areas articulate a mutually reinforcing set of problems that together motivate the unified framework presented in Section VI-B. Isolation of any single gap produces only partial solutions: a CDT-independent tutor cannot be rendered reliably adaptive no matter how sophisticated its generative capabilities, a feedback system without embedded hallucination detection cannot satisfy the safety requirements of academic deployment, assessment fairness cannot be systematically enforced without governance controls operating at the system level, and those governance controls cannot function meaningfully without the provenance metadata that RAG-grounded LLM outputs make available. The gap map thus serves a dual function as a diagnostic representation of the literature's current limits and as the structural justification for pursuing an integrated rather than piecemeal architectural response.

3.5. Proposed LLM-Driven CDT Framework

The proposed framework emerges directly from the convergence of three analytical outputs: the RQ-based thematic synthesis, the cluster-level analysis of the 111 included studies, and the five-area

gap matrix described above. Its architecture comprises four functionally distinct layers and one transversal governance layer, illustrated in Fig. 4. At the base, the Learner Data Layer (corresponding to Cluster C5 and RQ4) receives and structures the heterogeneous data streams generated by learner activity: behavioral event records, summative and formative assessment outcomes, interaction logs, self-regulated learning indicators, and relevant cognitive or affective signals. Above this, the Learner Modeling and Digital Twin Layer (C1/RQ4) applies probabilistic knowledge tracing methods (Corbett & Anderson, 1994) and profile inference procedures to transform raw data streams into a dynamic, continuously updated digital representation of the individual learner, encoding per-concept knowledge state estimates, identified misconception patterns, and affective state indicators.

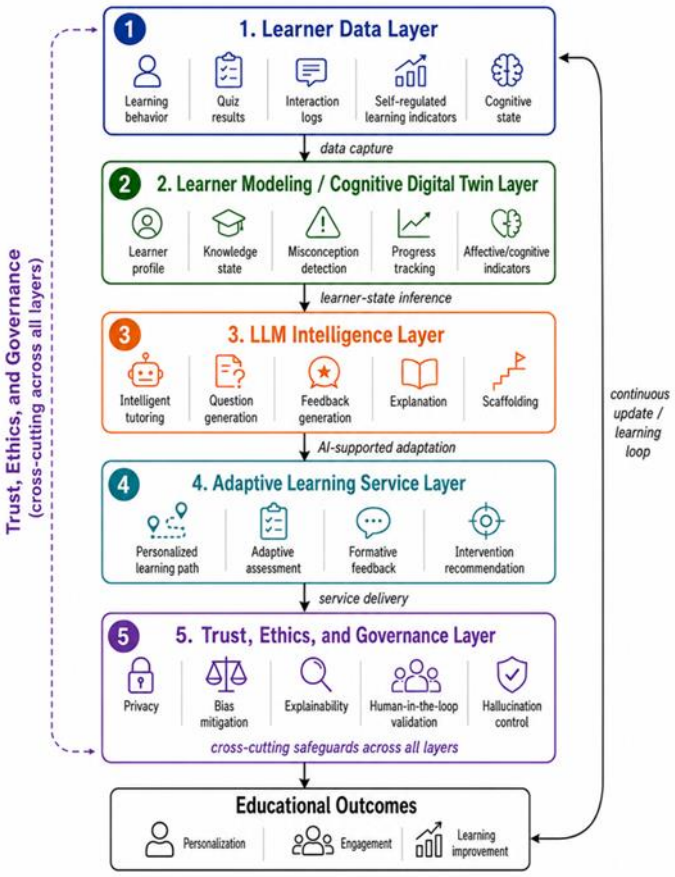

Figure 4. Proposed Five-Layer LLM-Driven Cognitive Digital Twin Framework

The LLM Intelligence Layer (C2/RQ1) is the mechanism through which generative capability and persistent learner knowledge are coupled. Upon each learner-initiated interaction, a retrieval module queries the digital twin to obtain the current knowledge state vector, a record of recently identified misconceptions, and a summary of relevant prior interactions. These elements are serialized into a structured context document and injected into the LLM prompt as grounding context, ensuring that generative output is conditioned on what is known about the learner rather than produced from a context-free prior. The Adaptive Learning Service Layer (C3/RQ2–RQ3) then operationalizes these conditioned outputs as concrete educational services: personalized tutoring dialogue, adaptively calibrated assessment tasks, formative feedback, and instructor-directed intervention recommendations, all tagged with provenance metadata linking each output to the learner state and retrieval context that produced it. Spanning all four functional layers, the Trust, Ethics, and Governance Layer (C4/RQ5) operates as a cross-cutting control plane, actively enforcing privacy constraints,

conducting bias auditing, applying hallucination detection checks, and maintaining human-in-the-loop review pathways.

3.6. Recommended Evaluation Protocol

Table VIII gives operational form to the five-layer framework by mapping each layer to a set of validation measures that constitute layer-level evidence of adequacy before integration testing commences. The protocol is staged by design: rather than demanding simultaneous validation of all components, it specifies what constitutes sufficient evidence at each layer, thereby offering a tractable research agenda for teams seeking to implement partial or complete versions of the proposed architecture.

Evaluation Level	Recommended Measures
Learner Modeling	Knowledge-state accuracy, misconception detection, profile update traceability, longitudinal stability. (Corbett & Anderson, 1994; Yamin et al., 2023)
LLM Tutoring	Response correctness, pedagogical quality, RAG retrieval precision, grounding coverage, teacher review rate. (Lui et al., 2024; Makharia et al., 2024)
Adaptive Assessment	Rubric alignment, grading consistency, inter-rater agreement, fairness across demographic groups. (Cagliero et al., 2024; Gopi et al., 2024)
Feedback Quality	Usefulness rating, specificity score, hallucination rate, revision improvement, human validation rate. (Jürgensmeier & Skiera, 2024; Lin & Crosthwaite, 2024; Ouyang et al., 2024)
Governance	Privacy compliance, bias audit report, XAI log coverage, human-in-the-loop review logs, academic integrity safeguard metrics. (European Parliament and Council of the European Union, 2024; Khosravi et al., 2022; Susnjak, 2024)

Validation at the Learner Modeling and Digital Twin layer centers on the question of representational fidelity. The primary concern is whether the maintained digital twin reliably tracks the learner's actual knowledge state over time, assessed by comparing twin-estimated knowledge state against independently obtained assessment scores across multiple temporal points, following the probabilistic knowledge tracing lineage formalized by Corbett and Anderson (1994). Misconception detection should be validated by systematic comparison with expert-annotated records of student error patterns, while profile update traceability requires that every state modification be linked to a specific, time-stamped interaction event. A validation dimension that has received insufficient attention in prior work is longitudinal stability: the question of whether the learner model maintains coherent estimates across periods of learner inactivity, or instead degrades in ways that undermine subsequent inference, a concern explicitly noted by Yamin et al. (2023).

Evaluation at the LLM Intelligence Layer must treat retrieval fidelity and generative quality as two analytically distinct dimensions, each requiring its own measurement strategy. Expert rubric-based rating of response correctness and pedagogical appropriateness addresses the generative dimension, with rubric dimensions capturing factual accuracy, scaffolding calibration to the learner's current state, and explanatory clarity. Retrieval fidelity is captured by RAG retrieval precision, the degree to which the retrieved learner state document accurately and completely reflects the knowledge state and misconception record relevant to the current interaction. This dimension must be assessed independently of generative output quality, because retrieval errors may propagate into responses that

appear pedagogically plausible while being substantively misleading (Ji et al., 2023). Two additional operational measures round out this layer's evaluation: grounding coverage, which quantifies the proportion of LLM output traceable to retrieved learner state content, and teacher review rate, which tracks the frequency with which automated tutoring outputs are flagged and corrected by human instructors, providing an ongoing signal of real-world validity (Lui et al., 2024; Makharia et al., 2024).

Within the Adaptive Learning Service Layer, the assessment and feedback sub-services each carry a distinct but intersecting evaluation burden. For automated assessment, the critical requirement prior to deployment is establishing rubric alignment and inter-rater reliability between LLM-generated grades and expert reference grades, quantified through Cohen's Kappa or its weighted variant applied to stratified response samples (Yamin et al., 2023). Aggregate performance metrics alone are insufficient: fairness analysis disaggregated by language background, gender, and prior achievement level is necessary to detect systematic performance disparities that would otherwise remain invisible (European Parliament and Council of the European Union, 2024; Khosravi et al., 2022). For feedback quality, Jürgensmeier and Skiera (2024) have demonstrated that perceived usefulness ratings and specificity scores are practically collectable at scale and provide a useful efficiency-oriented validity signal. Expert review of sampled outputs is additionally required to estimate the hallucination rate, defined as the proportion of feedback items containing factually incorrect or epistemically unsupported claims, because this dimension is not reliably detectable through automated methods alone. The most substantively meaningful validity signal at this layer is revision improvement: the measurable gain in assessed work quality attributable to LLM-generated feedback, evaluated against appropriate no-feedback or human-feedback control conditions (Lin & Crosthwaite, 2024; Ouyang et al., 2024).

Evaluation at the Trust, Ethics, and Governance Layer operates on a fundamentally different register from the layers above it: the primary questions are not about performance but about compliance, transparency, and accountability. Privacy compliance should be formally assessed through structured audit checklists referenced against the EU AI Act (European Parliament and Council of the European Union, 2024) and applicable institutional data protection obligations, covering data minimization, consent documentation, and access control practices. Bias audit reports must systematically document fairness testing outcomes across all adaptive services, with particular attention to the demographic subgroup disparities surfaced at the assessment layer. XAI log coverage measures what proportion of consequential system decisions including tutoring responses, feedback assignments, grading outcomes, and instructional intervention recommendations are accompanied by interpretable explanations available to both learners and their instructors, a requirement grounded in the framework developed by Khosravi et al. (Khosravi et al., 2022). The human-in-the-loop review log records the rate, category, and resolution outcome for cases flagged for human review, constituting a direct operational measure of governance mechanism effectiveness. Rounding out the governance evaluation picture, academic integrity safeguard metrics including the detection rate of potential misuse events and the false positive rate of automated alerts provide evidence relevant to concerns about learner misconduct facilitated by generative AI tools (Hadi Mogavi et al., 2024).

An important property of this evaluation protocol is its staged structure: the measures are organized as a sequential research agenda rather than a simultaneous validation battery to be applied at a single moment. In Horizon 1 (component validation), each layer is assessed independently using the measures described above; failures at this stage are diagnostic of component-level design issues that can be resolved before integration. In Horizon 2 (integrated architecture testing), inter-layer interactions become the primary object of study, with particular emphasis on the RAG grounding pathway connecting the CDT and LLM layers, where propagation effects are most likely. In Horizon 3 (longitudinal deployment), the research focus shifts to full-term deployment studies capable of capturing learning trajectory outcomes and sustained governance compliance over extended periods. This three-horizon structure is consistent with accepted practice in the evaluation of complex educational technology interventions (Alfredo et al., 2024; Gašević et al., 2015), and it guards against

the common failure mode in which integration problems are discovered only after large-scale deployment has already occurred.

3.7. Implications and Threats to Validity

3.7.1. Implications

The implications of this review extend across multiple stakeholder groups. For researchers, the synthesis argues for a reorientation of contribution from isolated LLM tutoring or feedback prototypes toward longitudinally designed, learner-aware systems in which generative outputs are both grounded in and subsequently modify a maintained CDT record (Aboulsafa et al., 2023; Yamin et al., 2023). For developers and system architects, the five-layer framework carries a concrete design mandate: LLM deployment in educational settings should incorporate learner modeling components, RAG-mediated grounding pipelines, structured validation and review workflows, and auditable logging mechanisms from the outset (Abu-Rasheed et al., 2025; Madhubhashana et al., 2024). For institutional decision-makers, the evidence strongly supports maintaining human-in-the-loop review for all high-stakes applications of automated feedback, assessment, and intervention, regardless of the apparent quality of automated outputs (Alfredo et al., 2024). For policymakers and regulatory bodies, the review reinforces the position that privacy protection, algorithmic transparency, bias mitigation, and accountability mechanisms should be specified as non-negotiable system requirements rather than left to voluntary adoption (European Parliament and Council of the European Union, 2024).

3.7.2. Threats to Validity

Several limitations bear on the generalizability and completeness of this review. Database coverage was restricted to ScienceDirect and IEEE Xplore, studies indexed in ACM Digital Library, Scopus, Web of Science, or Google Scholar may meet inclusion criteria but fall outside the present synthesis. Search string construction inevitably introduces boundary effects, and the English-language filter excludes potentially relevant contributions in other languages. Single-reviewer screening and coding represent a further constraint. This was partially offset by four procedural safeguards—structured decision documentation, iterative consistency verification, rubric anchoring, and full workbook transparency—though future versions of this review should incorporate second-reviewer audit procedures and formally report Cohen's Kappa as a measure of inter-rater agreement (Kitchenham, 2007). Finally, readers should situate the findings within the temporal boundaries of the search: given the pace at which LLM research develops, the synthesis should be understood as an account of the evidence available up to the June 2026 screening closure, and material published subsequently is not represented (Brown et al., 2020).

4. CONCLUSION

This systematic review synthesized 111 Core and Supporting contributions spanning LLMs, GenAI, intelligent tutoring, adaptive assessment, automated feedback, learner modeling, Cognitive Digital Twins, and trustworthy AI in education. Across the five research questions, a consistent picture emerges. RQ1 establishes that LLM-based tutoring is actively developed but structurally session-bound. RQ2 and RQ3 together show that assessment and feedback automation have reached a degree of empirical maturity, yet neither has resolved the safety challenges posed by hallucination and output inconsistency. RQ4 identifies CDT and learner modeling as indispensable representational infrastructure whose integration with LLM services remains, in practical terms, unrealized. RQ5 documents ethics and governance concerns that are genuinely pervasive across the corpus but rarely addressed as embedded architectural properties of the systems in question.

The primary theoretical contribution advanced here is a five-layer LLM-driven CDT framework whose architecturally defining commitment is the RAG-mediated coupling of generative LLM capability to a persistent, updatable learner digital twin. This coupling moves the LLM from a context-free, session-bound responder to a learner-state-conditioned reasoning engine whose outputs can be

traced, personalized, and audited across the full arc of a learning trajectory rather than within isolated interactions. The framework is put forward as a conceptual architecture: it provides the design vocabulary and validation agenda for future empirical work, but should not be applied to high-stakes educational settings before component-level validation, integrated architecture testing, and governance evaluation have been satisfactorily completed.

REFERENCES

Aboulsafa, E. I., El Khayat, G. A., & Elmorsy, S. A. (2023). An educational human digital twin proposed model for personalized e-learning. In *2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI)* (pp. 1–8). https://doi.org/10.1109/AMCAI59331.2023.10431503

Abu-Rasheed, H., Jumbo, C., Al Amin, R., Weber, C., Wiese, V., Obermaisser, R., & Fathi, M. (2025). LLM-assisted knowledge graph completion for curriculum and domain modelling in personalized higher education recommendations. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–5). https://doi.org/10.1109/EDUCON62633.2025.11016377

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence, 6*, 100215. https://doi.org/10.1016/j.caeai.2024.100215

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*.

Cagliero, L., Farinetti, L., Fior, J., & Manenti, A. I. (2024). ChatGPT, be my teaching assistant! Automatic correction of SQL exercises. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 81–87).

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278. https://doi.org/10.1007/BF01099821

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., et al. (2024). Extended TAM based acceptance of AI-powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon, 10*(8), e29317. https://doi.org/10.1016/j.heliyon.2024.e29317

Du, H., Rashid, P., Jia, Q., & Gehringer, E. (2024). Interactive rubric generator for instructor's assessment using prompt engineering and large language models. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9).

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Frank, L., Herth, F., Stuwe, P., Klaiber, M., Gerschner, F., & Theissler, A. (2024). Leveraging GenAI for an intelligent tutoring system for R: A quantitative evaluation of large language models. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–9).

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gopi, S., Sreekanth, D., & Dehbozorgi, N. (2024). Enhancing engineering education through LLM-driven adaptive quiz generation: A RAG-based approach. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1–8). https://doi.org/10.1109/FIE61694.2024.10893146

Hadi Mogavi, R., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujar, S., Nacke, L. E., & Gao, H. (2024). ChatGPT in education: A blessing or a

curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans, 2*(1), 100027. https://doi.org/10.1016/j.chbah.2023.100027

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12). https://doi.org/10.1145/3571730

Jürgensmeier, L., & Skiera, B. (2024). Generative AI for scalable feedback to multimodal exercises. *International Journal of Research in Marketing, 41*(3), 468–488. https://doi.org/10.1016/j.ijresmar.2024.05.005

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074.

Kinsner, W. (2021). Digital twins for personalized education and lifelong learning. In *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 1–6).

Kitchenham, B. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Version 2.3, EBSE Technical Report).

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*.

Lui, R. W. C., Bai, H., Zhang, A. W. Y., & Chu, E. T. H. (2024). GPTutor: A generative AI-powered intelligent tutoring system to support interactive learning with knowledge-grounded question answering. In *2024 International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)* (pp. 702–707).

Madhubhashana, S., Ranhinda, V., Jayakody, D., Shavinda, L., Nawarathne, M., & Haddela, P. (2024). Adapting pedagogical frameworks to online learning: Simulating the tutor role with LLMs and the QUILT model. In *2024 6th International Conference on Advancements in Computing (ICAC)* (pp. 581–586).

Makharia, R., Kim, Y. C., Jo, S. B., Kim, M. A., Jain, A., Agarwal, P., et al. (2024). AI tutor enhanced with prompt engineering and deep knowledge tracing. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (pp. 1–6). https://doi.org/10.1109/IATMSI60426.2024.10503187

Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the effects of instructor manual feedback and ChatGPT intelligent feedback on collaborative programming in China's higher education. *IEEE Transactions on Learning Technologies, 17*, 2227–2239. https://doi.org/10.1109/TLT.2024.3486749

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Prasetyo, M. M., & Muslaini, F. (2025). Utilization of digital twin technology in education ecosystem: A systematic literature review. *4*(2), 13–26.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare, 11*(6).

Sinha, A., Goyal, S., Sy, Z., Kuperus, R., Dickey, E., & Bejarano, A. (2024). BoilerTAI: A platform for

enhancing instruction using generative AI in educational forums. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1–8).

Susnjak, T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education, 34*(2), 452–482.

Vargas-Murillo, A. R., Pari-Bedoya, I. N. M. D. L. A., & Guevara-Soto, F. D. J. (2023). The ethics of AI assisted learning: A systematic literature review on the impacts of ChatGPT usage in education. *ACM International Conference Proceedings Series, 22*(7), 8–13.

Yamin, M. M., Imran, A. S., & Katt, B. (2023). Towards a digital twin for lifelong learning. In *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1–6).