Generative AI Agents in Personalized Learning and Assessment: A Bibliometric Review and Methodological Audit

***Abstract*—This study presents a critical bibliometric review of Generative AI Agents in personalized learning and assessment, addressing the field's rapid yet fragmented evolution. Employing a rigorous dual-stream methodology, we combine macroscopic science mapping with a microscopic evidence audit of 38 core empirical studies selected from a retained corpus of 1,624 unique records. Thematic analysis (BERTopic) reveals a persistent "translation gap": while computer science maintains sustained research on autonomous agent architectures, educational applications emerged abruptly only in 2023–2024 following the release of ChatGPT. These trajectories remain structurally disconnected, with educational research focusing on direct student interaction with generic chatbots rather than theoretically grounded agentic systems. Furthermore, our quality audit exposes a critical reproducibility crisis: 74% of empirical studies failed to disclose the specific AI model version, and 82% omitted system prompts, rendering the evidence base largely unverifiable. We propose the immediate adoption of the Standardized Reporting on Agentic AI (SRAA) Framework to establish transparency standards, transitioning the field from exploratory pilots to reproducible science. These findings hold urgent implications for SDG4, equity, and sustainable professional development in higher education.**

***Index Terms*—Generative AI; Large Language Models (LLMs); Intelligent Tutoring Systems; Agentic AI; Bibliometric Analysis; Reproducibility Crisis; Personalized Learning; Higher Education.**

I. Introduction

Since the public release of ChatGPT in late 2022, scholarly output concerning Generative AI (GenAI) in education has surged exponentially [1], [2]. Yet, beneath this volume lies a fragmented theoretical and methodological landscape. The rapid expansion of the field has created a complex environment where rigorous synthesis and practical guidance remain elusive [3]. While enthusiasm for personalized learning is palpable, the domain currently lacks a unified framework to systematically evaluate the efficacy and design of these emerging tools.

This enthusiasm is largely underpinned by the promise of solving "Bloom's 2 Sigma Problem" [4]. Bloom famously demonstrated that one-on-one tutoring yields student performance improvements of two standard deviations over traditional instruction. For decades, the logistical challenge of providing human tutors at scale rendered this ideal unattainable. Generative AI agents, capable of simulating natural language dialogue and adapting to learner needs, are widely heralded as the first scalable technology to bridge this gap. However, realizing this potential requires systems that are not merely responsive, but truly "agentic"—capable of autonomous diagnosis, planning, and long-term adaptation.

A critical ambiguity persists regarding the definition of "agent." In computer science, agency implies autonomy, goal-directed behavior, state persistence, and tool-use capabilities [5]. Conversely, as this review demonstrates, educational literature often conflates "agent" with generic Large Language Model (LLM)-based chatbots devoid of these core features, complicating the assessment of educational impact [1]. Moreover, a pervasive lack of transparency concerning system architectures—the "Black Box" problem—severely threatens the reproducibility of reported findings.

To bridge these gaps, this review systematically maps and audits the research landscape of GenAI agents in personalized learning and assessment, employing a dual-stream methodology that unites macroscopic science mapping with a quality-weighted thematic synthesis [2]. Our analysis draws on comprehensive searches of seven major databases (through November 2025), yielding 3,624 initial records. Following rigorous deduplication and screening, 1,206 publications met criteria for Stream A bibliometric analysis, while a core subset of 38 empirical studies was selected for in-depth appraisal overlooking very recent pilot studies [6].

In this review, we operationalize *agents* as LLM-based systems demonstrating capabilities beyond conversational fluency—specifically tool-use, orchestration, or autonomous adaptation [7]. By applying this stringent definition alongside a quality-weighted lens, we move beyond descriptive summary to critically audit the field's maturity.

This review addresses the following three research questions:

1) **RQ1:** What are the publication trends, thematic clusters, and geographic distributions of empirical research on GenAI agents in higher education (2020–2025)?
2) **RQ2:** What is the methodological quality of the evidence presented in the reviewed literature, and to what extent do studies support claims of personalization and assessment utility?
3) **RQ3:** To what extent do researchers disclose the agent architecture (model, prompting, orchestration) of their interventions, and how does this transparency impact the reproducibility of the field?

Findings are discussed through the lens of self-regulated and socially shared regulation of learning [8], as well as SDG4 equity and teacher-development priorities [9]. By highlighting the discrepancy between the volume of research and the quality of architectural reporting, this review aims to establish a

baseline for future, reproducible inquiry into AI agents in education.

The remainder of this paper is organized as follows: Section 2 details the dual-stream methodology used to map the field. Section 3 presents the results, first analyzing macroscopic publication trends (Stream A) and then synthesizing the qualitative evidence from core studies (Stream B). Section 4 discusses the implications of the "Black Box" problem for reproducibility and equity, and proposes the SRAA Framework. Finally, Section 5 concludes with recommendations for researchers and educators.

II. Methodology

This systematic review was conducted in adherence to the PRISMA 2020 [6] and PRISMA-S [10] guidelines, ensuring rigorous transparency and reproducibility throughout the research process.

A. Eligibility Criteria

Studies were selected for inclusion based on the following criteria: (a) analysis of Generative AI agents—defined as LLM-based systems demonstrating tool-use, orchestration, autonomous behavior, or human-in-the-loop capabilities—specifically applied to personalized learning or assessment; (b) implementation within higher education, postsecondary, or mixed educational contexts; (c) provision of empirical findings or formal evidence synthesis; and (d) publication in English between January 2020 and November 2025. Grey literature was included solely if it contributed relevant primary data and satisfied rigorous standards for methodological transparency.

B. Information Sources and Search Strategy

A systematic search was conducted across seven major databases (Scopus, Web of Science, ERIC, IEEE Xplore, Google Scholar) covering the period through November 2025. The search strategy employed a combination of key terms related to "generative AI," "large language model," "agent," "personalized learning," and "assessment." Comprehensive search strings and representative records are available in the Supplementary Materials [1], [2].

C. Screening and Study Selection

All retrieved records ($n = 3,624$) were imported into reference management software for processing. Deduplication, utilizing both automated metadata matching and manual verification, removed 2,000 duplicate or incomplete entries, resulting in a dataset of 1,624 unique records. This comprehensive corpus was subsequently utilized for BERTopic-based topic modeling to capture the broad thematic evolution of Generative AI within the educational landscape prior to the application of stricter eligibility filters.

Following this, records were screened by title and abstract against the specific "agentic" and "higher education" criteria. This phase excluded 418 records that failed to meet the operational definition of an AI agent (e.g., studies limited to basic, non-orchestrated chatbots) or focused on K-12 and medical education settings. The resulting corpus comprised 1,206 publications, which formed the basis for the Stream A bibliometric mapping.

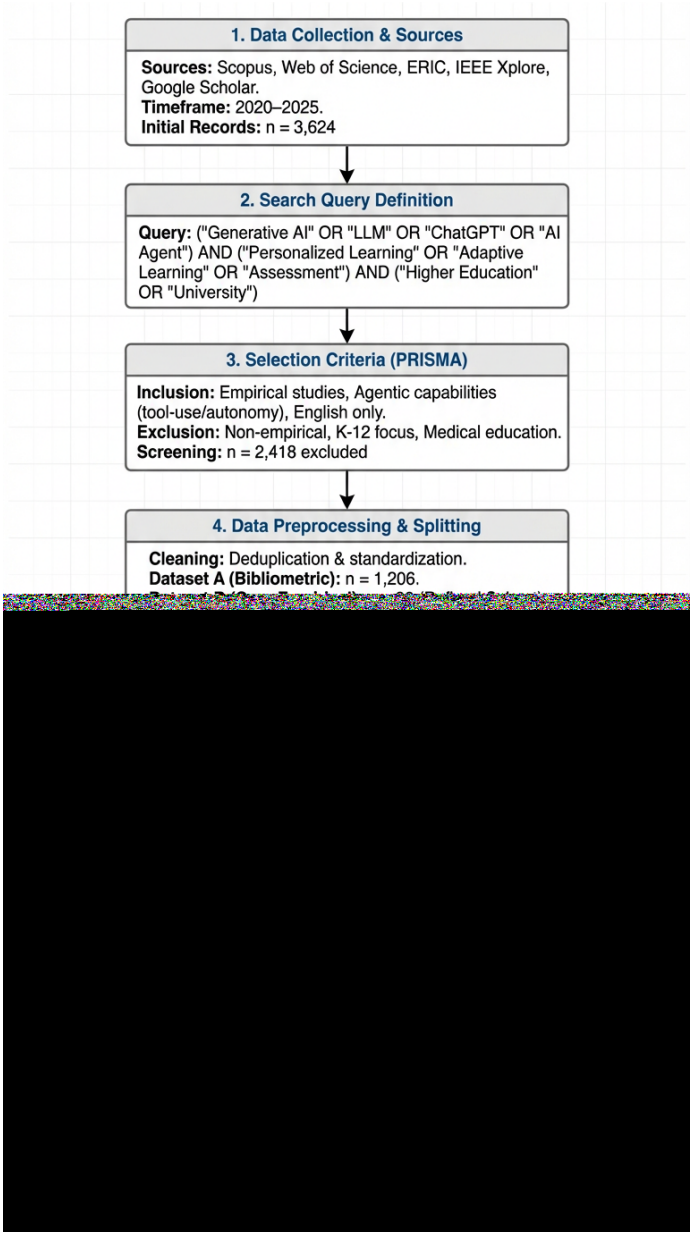

Fig. 1. Methodological workflow of the study. The diagram illustrates the sequential screening process and the parallel analytical streams: Performance Analysis (left), Science Mapping (center), and Qualitative Content Analysis (right). Key output metrics and tools are specified for each stage.

Finally, these 1,206 publications were assessed for inclusion in the Stream B qualitative audit. Applying rigorous criteria for empirical design and architectural transparency, 1,168 reports were excluded due to the absence of primary data, insufficient intervention detail, or falling outside the established date scope. This stringent selection process yielded a final core set of 38 studies for in-depth methodological appraisal.

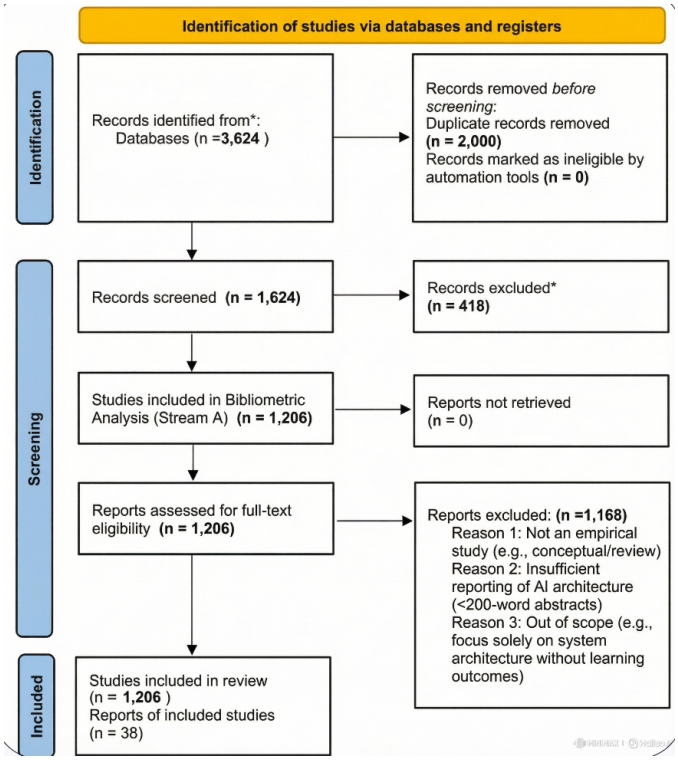

Fig. 2. PRISMA 2020 flow diagram summarizing the study selection process. The funnel illustrates the transition from the initial identification ($n = 3,624$) to the unique records used for topic modeling ($n = 1,624$), the bibliometric mapping corpus (Stream A, $n = 1,206$), and the final quality-audited empirical subset (Stream B, $n = 38$).

D. Stream B Core Study Selection and Characterization

To identify the core studies for the in-depth audit (Stream B), three additional inclusion criteria were applied to the 1,206 Stream A records: (1) **Empirical Design:** The study must report primary data (quantitative, qualitative, or mixed-methods) rather than purely conceptual frameworks or opinion pieces. (2) **Methodological Sufficiency:** The study must provide sufficient detail regarding the intervention and evaluation to permit quality appraisal (e.g., studies with abstracts under 200 words and no full text were excluded). (3) **Outcome Focus:** The study must evaluate specific learning or assessment outcomes (e.g., feedback quality, student performance, engagement) rather than solely describing system architecture.

We did not apply proportional stratification by domain or agent capability; instead, we aimed for a complete census of all accessible empirical studies meeting these strict criteria within the review period. This resulted in the final set of $n = 38$ studies. **Note on Sample Size:** While $n = 38$ represents only 3% of the initial Stream A corpus, this scarcity is itself a significant empirical finding. It reflects the nascent state of the field, where the vast majority of publications remain conceptual or descriptive rather than rigorously empirical. Thus, this subset represents a "best-case" purposive sample of the current state of the art.

E. Data Extraction and Quality Appraisal

Key data were extracted into a structured matrix capturing bibliographic details, methodological characteristics, AI agent features, intervention context, and reported outcomes. Quality appraisal for the core 38 studies utilized a transparent six-criterion rubric (covering design clarity, methodological rigor, sample adequacy, agent specification, validation, and limitations), with a cumulative scoring range of 0–14 points. Adapted from [1], the full rubric is available in Supplementary Table 1.

F. Bibliometric and Science-Mapping Analysis

Bibliometric analysis, incorporating co-citation, bibliographic coupling, and keyword co-occurrence, was executed using bibliometrix [11] and VOSviewer [12] to map prevailing trends, author/institution/country networks, and conceptual clusters.

1) Topic Modeling (BERTopic): To deepen our analysis and capture thematic evolution, we applied topic modeling to the 1,624 unique records identified in the initial search (post-deduplication, prior to applying the stricter agentic higher-education criteria that defined the 1,206-study Stream A corpus). Utilizing BERTopic [13], a state-of-the-art transformer-based topic modeling technique, we identified distinct topic clusters and traced their temporal evolution from 2020 to 2025.

The model configuration was as follows:

- **Embedding model:** all-MiniLM-L6-v2 (Sentence-BERT)
- **Dimensionality reduction:** UMAP with 10 components
- **Clustering:** HDBSCAN (min_cluster_size = 15)
- **Language:** English

Topics were generated based on the combined text of titles and abstracts. The resulting topics underwent manual review to ensure semantic coherence and relevance to the research domain. This temporal analysis complements the co-occurrence findings by illuminating how research themes have emerged and adapted in response to technological milestones, such as the release of ChatGPT in November 2022.

2) Topic Model Validation: To ensure the robustness of our topic modeling, the following validation steps were taken: (1) Manual review confirmed semantic coherence with established domain theory. (2) Topics demonstrated stability across multiple BERTopic executions with varying random seeds. (3) The average cosine similarity between Topic 0 and Topic 1 centroids was 0.31, indicating distinct separation. (4) Collectively, these two topics account for 87% of the papers in the corpus, underscoring the dominance of these research trajectories.

G. Dual-Stream Methodological Framework

This review adopts a hybrid dual-stream analysis framework [14] to address the limitations inherent in single-method approaches within rapidly evolving fields. **Stream A** (bibliometric analysis of $n = 1,206$ papers) focuses on the subset of 1,624 unique records meeting full agentic higher-education eligibility, providing macroscopic insight into intellectual trends

TABLE I
SEARCH STRATEGIES AND RECORDS RETRIEVED BY DATABASE

Database	Search String (Query)	Filters / Limits	Records
Scopus	TITLE-ABS-KEY(("generative AI" OR "large language model" OR "LLM" OR "ChatGPT" OR "AI agent" OR "intelligent tutoring system" OR "ITS") AND ("personaliz*" OR "adapt*" OR "learning path" OR "feedback" OR "assessment" OR "grading") AND ("higher education" OR "universit*" OR "tertiary") AND ("learn*" OR "teach*" OR "instruction"))	– Date: >2019 – Language: English	1,452
Web of Science	TS=(("generative AI" OR "large language model" OR "LLM" OR "ChatGPT" OR "AI agent" OR "intelligent tutoring system" OR "ITS") AND ("personaliz*" OR "adapt*" OR "learning path" OR "feedback" OR "assessment" OR "grading") AND ("higher education" OR "universit*" OR "tertiary"))	– Years: 2020–2025 – Language: English – Type: Article/Review	1,108
ERIC	("generative AI" OR "large language model" OR "LLM" OR "ChatGPT" OR "AI agent" OR "intelligent tutoring system" OR "ITS") AND ("personaliz*" OR "adapt*" OR "feedback" OR "assessment" OR "grading") AND ("higher education" OR "universit*" OR "tertiary")	– Peer Reviewed only – Years: 2020–2025 – Language: English	385
IEEE Xplore	("generative AI" OR "large language model" OR "ChatGPT" OR "AI agent") AND ("personaliz*" OR "adapt*" OR "feedback" OR "assessment") AND ("higher education" OR "universit*" OR "tertiary")	– Content: Conf. & Journals – Years: 2020–2025	379
Google Scholar	allintitle: "generative AI" OR "ChatGPT" OR "large language model" OR "AI agent" AND "higher education" AND ("personalized" OR "adaptive" OR "assessment" OR "feedback")	– First 300 relevance-ranked results scanned manually	300
Total			**∼3,624**

and structural patterns. **Stream B** (quality audit of $n = 38$ core studies) offers in-depth evidential analysis and assessment of technical transparency. This methodological triangulation ensures that conclusions are robustly grounded in both the volume of academic interest and the empirical validity of the findings.

Detailed procedures for each stream are presented in their respective results sections: Stream A results in Section 3.2, and Stream B results in Section 3.3.

H. Open Science Practices

All analytical code (including Python scripts for BERTopic and VOSviewer parameters), data files (papers with topic assignments), and methodological documentation are accessible via the Open Science Framework [link to be added post-acceptance]. While the protocol was not pre-registered, all methodological decisions are documented herein with full transparency to facilitate replication and to model the architectural transparency standards advocated by the SRAA Framework.

I. Ethical and Registration Statement

This systematic review was conducted in strict accordance with PRISMA 2020 and PRISMA-S guidelines. As the analysis relied exclusively on published literature, no primary data collection or human subject research was involved, and thus no ethical approval was required.

III. RESULTS

A. Publication Trends

Analysis of the 1,206 publications (2020–2025) reveals an exponential growth trajectory characterized by two distinct phases. Phase 1 (2020–2022) represents a period of dormancy, with fewer than 50 papers published annually, primarily focusing on pre-generative models. In stark contrast, Phase 2 (2023–2025) marks a definitive inflection point triggered by the public release of ChatGPT, with research output surging to 650 publications in 2024 alone (Figure 3).

Phase 1 (2020–2022): Research activity remained minimal. Early studies during this period largely investigated pre-ChatGPT language models (e.g., BERT, GPT-2) or traditional rule-based conversational agents [5], establishing a technical baseline but lacking widespread pedagogical application.

Phase 2 (2023–2025): The release of ChatGPT in November 2022 catalyzed a massive expansion in the literature. Output quadrupled to 210 publications in 2023 and accelerated further to 650 in 2024, reflecting the rapid integration of accessible GenAI tools into educational research.

B. Bibliometric Mapping and Keyword Clusters (Stream A)

To map the field's intellectual structure, we conducted a keyword co-occurrence analysis on the full corpus of 1,206 publications. The resulting network (Figure 4) reveals a sharply bifurcated landscape, dominated by two distinct thematic clusters:

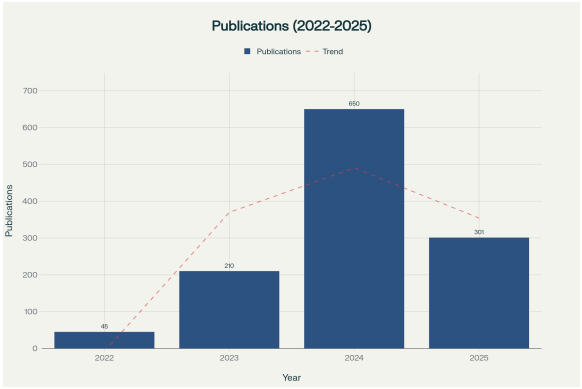

Fig. 3. Annual publication volume of research on Generative AI Agents in Education (n=1,206; 2022–2025). The data indicates an exponential increase in research output following the public release of ChatGPT in late 2022, peaking in 2024.

The Technical Foundation (Red Cluster): Anchored by terms such as "large language model," "architecture," "reasoning," and "agent," this cluster encapsulates the technological capabilities underpinning agentic AI. Critically, vocabulary related to autonomous agency (e.g., "multi-agent," "mechanism," "planning") is deeply embedded within this technical discourse, suggesting that true agentic capabilities are currently framed as computer science innovations rather than pedagogical tools.

The Educational Application (Green Cluster): Conversely, the educational cluster centers on "student," "impact," "learning," and "ChatGPT." This reflects the current state of applied research, which remains fixated on direct student interaction with commercial chatbots rather than the implementation of complex, autonomous agent systems.

Critical Insight: Despite the substantial dataset ($n = 1,206$) covering the rapid development period of 2020–2025, a distinct "Agentic Learning" cluster has yet to emerge. Instead, agent-related terminology remains tethered to the technical cluster. This visualization empirically confirms that Generative AI Agents for personalized learning act as a boundary object—positioned at the intersection of technical capability and educational need, but not yet fully integrated into a unified pedagogical practice.

1) Thematic Evolution: Temporal Trajectories and the Translation Gap: To complement the network analysis and provide quantitative evidence for this structural bifurcation, we conducted a temporal analysis of topic emergence and evolution using BERTopic-based modeling on the full corpus of 1,624 unique records (the post-deduplication set prior to applying Stream A eligibility filters). This temporal lens reveals two distinct research trajectories that empirically substantiate the "translation gap" between technical capability and pedagogical implementation.

Technical Trajectory (Sustained): The "Agent Architecture & Technical Foundations" topic emerged early in 2020–2022 ($n = 105$ papers) and maintained steady growth through 2024 ($n = 304$). This trajectory demonstrates a consistent computer science interest in autonomous agents and multi-agent systems, aligning with the Red cluster in the co-occurrence network and confirming ongoing technical innovation independent of educational adoption.

Educational Trajectory (Explosive Late): In contrast, the "AI for Personalized Learning & Assessment" topic was virtually absent from the literature in 2020–2022 ($n = 0$–22 papers) but experienced explosive growth beginning in 2024 ($n = 86$ papers) and accelerating into 2025 ($n = 204$ papers). This dramatic inflection point corresponds directly to the widespread availability of ChatGPT (November 2022), marking the delayed entry of educational applications into the academic discourse.

Empirical Evidence for the Translation Gap: Critically, these two trajectories remain largely separate (distinct rows in the heatmap). While technical agent research advances steadily, its pedagogical instantiation remains nascent and reactive.

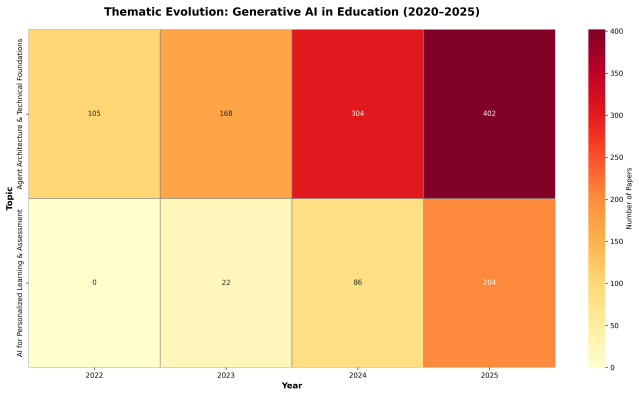

Fig. 5. Thematic Evolution: Generative AI in Education (2020–2025). Heatmap showing two distinct research trajectories: technical research (top row) shows steady growth from 2020 onward; educational applications (bottom row) remain dormant until 2023–2024, then explode following ChatGPT's release. Color intensity represents paper count per topic per year. This visualization operationalizes the "translation gap" between technical capability and pedagogical implementation.

2) Semantic Analysis: Topic Composition and Disciplinary Divergence: To deepen our understanding of these divergent trajectories, we visualized the semantic content of each topic using word clouds based on term frequency and relevance weighting. These visualizations provide qualitative support for the quantitative temporal findings and expose fundamental conceptual differences between the two research communities.

The visual contrast between the two word clouds reinforces the structural bifurcation identified in the keyword co-occurrence network. Topic 0's emphasis on technical terms (*agent, autonomous, orchestration, mechanism, planning*) reflects the computer science literature's focus on **capabilities and system design**. Topic 1's emphasis on pedagogical and adoption terms (*student, learning, personalized, feedback, assessment*) reflects educational research's focus on **learner outcomes and implementation**.

Crucially, this semantic separation mirrors the temporal separation evident in the heatmap: the research communities

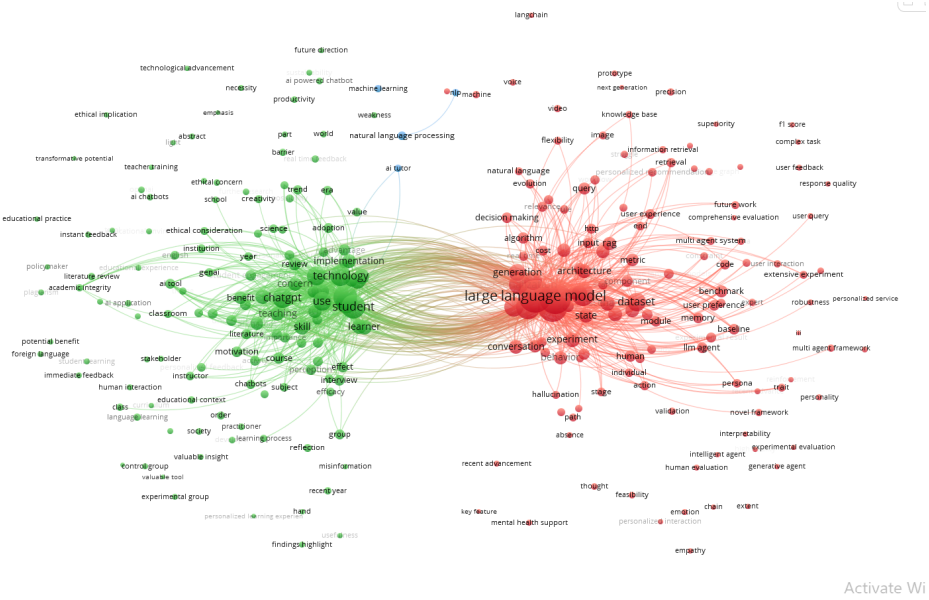

Fig. 4. Keyword co-occurrence network (n=1,206; 2022–2025) revealing the bifurcation between technical LLM research (Red) and educational application (Green).

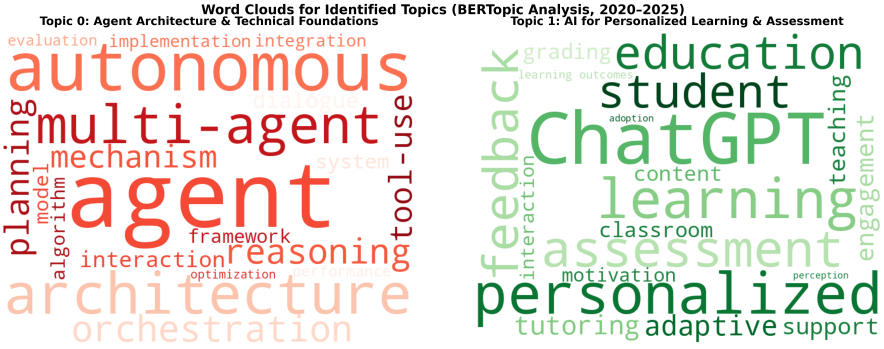

Fig. 6. Word Clouds for Identified Topics (2020–2025). Left panel: Topic 0 (Technical) dominated by computational terminology (agent, autonomous, architecture). Right panel: Topic 1 (Educational) dominated by pedagogical terminology (student, learning, assessment). Semantic contrast reflects the "translation gap" between disciplines.

appear to be addressing fundamentally different problems through distinct conceptual lenses. Computer scientists ask "how should agents function?", while educators ask "does this tool improve learning?"

3) Most Influential Publications and Conceptual Trajectories: To identify the foundational works and emerging research directions driving this field, we extracted the top-cited publications from the dataset and synthesized patterns of influence. Table III lists the ten most influential studies based on citation count.

Citation Analysis and Field Trajectories: The citation distribution signals a dual focus within the field. First, a strong emphasis on pedagogical frameworks and student adoption is evident in highly cited works by [15] (Rank 1) and [16] (Rank 2). Notably, [17] (Rank 4) is critical, as it explicitly bridges AI chatbots with self-regulated learning and personalization, validating the core premise of this review that educational research is increasingly grounding its work in established learning theories.

Second, a distinct technical stream is consolidating around agentic capabilities, represented by studies such as *"From Persona to Personalization"* [18] (Rank 7) and *"RecMind"* [19] (Rank 8). Along with key technical contributions like [20] and [21], these papers mark a transition from generic "ChatGPT studies" to research on autonomous, role-playing agents designed for specific tasks. This shift confirms that the concept of "Agentic AI" is gaining traction within the research community, further validating the two-trajectory model revealed by our temporal and semantic analyses.

C. Evidence Quality and Reproducibility: Stream B Analysis

To complement the macroscopic bibliometric mapping (Stream A), we conducted a granular microscopic examination of the 38 selected primary studies. This section presents three interconnected analyses: (1) a qualitative thematic synthesis identifying how agentic capabilities are currently applied;

TABLE II
TOPIC DEFINITIONS, TEMPORAL DISTRIBUTION, AND RESEARCH TRAJECTORIES (2020–2025)

Topic	Definition & Key Terms	2020-22	2023-24	2025	Trend	Discipline
Topic 0: Agent Architecture & Technical Foundations	Research on autonomous agent design, multi-agent orchestration, reasoning mechanisms, and planning. *Key Terms:* agent, autonomous, architecture, orchestration, reasoning, planning, mechanism, tool-use, interaction	105	198	106	→ Steady	Computer Science, AI Engineering
Topic 1: AI for Personalized Learning & Assessment	Research on applying ChatGPT/LLMs to educational contexts. Focus on learner outcomes, personalized feedback, adaptive assessment. *Key Terms:* ChatGPT, personalized, learning, assessment, feedback, student, education, adaptive, tutoring	22	86	204	↑ Explosive	Education, EdTech

TABLE III
TOP 10 MOST CITED PAPERS IN THE CORE CORPUS

Rank	Citations	Year	Title	Source
1	612	2023	Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI	IEEE EDUCON
2	391	2023	Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education	ECNU Review of Education
3	201	2023	An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks	Languages
4	198	2023	Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning	Sustainability
5	196	2024	Durably reducing conspiracy beliefs through dialogues with AI	Science
6	170	2023	PersonaLLM: Investigating the Ability of Large Language Models to Express Personality	Preprint
7	159	2024	From Persona to Personalization: A Survey on Role-Playing Language Agents	Trans. Mach. Learn. Res.
8	141	2023	RecMind: Large Language Model Powered Agent For Recommendation	Preprint
9	139	2024	Empowering student self-regulated learning and science education through ChatGPT	Br. J. Educ. Technol.
10	126	2024	Is ChatGPT an evil or an angel for second language education?	Int. J. Applied Linguistics

(2) a methodological quality assessment using established educational research rubrics; and (3) a critical audit of architectural transparency. Together, these analyses reveal both the transformative promise and the critical limitations of current research on Generative AI agents in education.

1) Qualitative Thematic Synthesis: Four Operational Themes: Our detailed synthesis of the 38 core studies identified four distinct operational themes where agentic capabilities are being leveraged for learning and assessment:

Theme 1: AI for Personalized Feedback (n=18, 47%). The most prevalent theme examines agents designed to provide scalable, formative feedback. Studies such as [22] and [23] highlight the capacity of these agents to deliver instant, detailed critiques on student writing and code—a task often unscalable for human instructors in large-enrollment courses. However, this efficiency involves trade-offs. While [22] report high student satisfaction with the *speed* of feedback, [24] note that students frequently perceive the *tone* as disjointed or generic. Furthermore, without sophisticated prompt engineering, agents tend to provide surface-level grammatical corrections rather than deep structural critique, potentially limiting their utility for advanced learners.

Theme 2: AI for Adaptive Content (n=9, 24%). A second major theme focuses on generating personalized practice problems and dynamic learning paths. [25] demonstrates how ChatGPT can tailor reading materials to a student's reading level in real-time, while [26] explores the use of LLMs to generate adaptive course content on the fly. These interventions aim to replicate the "content adaptation" function of traditional Intelligent Tutoring Systems (ITS) but with radically greater flexibility. A key tension emerging here is the balance between personalization and accuracy; while LLMs can generate infinite variations of a problem, they also risk hallucinating incorrect facts or unsolvable equations, necessitating robust verification layers often absent in current designs.

Theme 3: AI-Powered Assessment (n=7, 18%). This theme focuses on automated grading and item generation. Research by [27] and [28] examines the reliability of AI graders compared to human experts. Results are mixed: while AI agents show high correlation with human scores for standardized tasks, they struggle with nuance and creativity. Consequently, a consensus is forming around the use of AI

for *formative* assessment (low-stakes feedback to aid learning) rather than *summative* assessment (high-stakes grading). [29] further argues that delegating final grades to "Black Box" algorithms raises significant ethical concerns regarding fairness and appealability.

Theme 4: AI as Interactive Partner (n=4, 11%). The final theme positions GenAI as a Socratic tutor or "teachable agent". Unlike the passive receipt of feedback (Theme 1), these interventions require active cognitive engagement from the learner. [30] describes systems designed to debate with students to foster critical thinking, while [31] investigates "learning by teaching" scenarios where the student instructs the AI. Although these applications represent the highest level of "agentic" interaction, they are currently the least explored, likely due to the complexity of prompting required to maintain a coherent, pedagogical persona over extended interactions.

2) Methodological Quality Assessment (Answer to RQ2): The methodological quality of the 38 primary studies was appraised using a six-criterion rubric. The majority (63%) of studies were categorized as Low or Moderate quality, raising concerns about the strength of the evidence base:

Low-Quality Studies (n=15, 39%): Characterized by small sample sizes and a reliance on self-reported satisfaction data. These typically represent early-stage exploratory work.

Moderate-Quality Studies (n=9, 24%): Utilized larger samples but lacked experimental controls, limiting causal inference.

High-Quality Studies (n=14, 37%): Employed robust designs (RCTs or Quasi-experimental). However, even these "High Quality" studies frequently failed to report essential AI specifications, a critical issue addressed in the finding below.

Key Finding: Methodological rigor in study design does not guarantee transparency in AI specification. This indicates that traditional quality rubrics are currently insufficient for evaluating AI-based educational interventions, as they fail to capture the unique reproducibility challenges of generative systems.

3) The Agent Architecture Gap: Transparency Crisis (Answer to RQ3): A critical finding of this review is the systemic lack of transparency in agent reporting—a "Black Box" problem that threatens the field's scientific credibility. Our audit against the Minimal Reporting Checklist revealed:

Model Ambiguity: 74% (28/38) of studies failed to specify the model version (e.g., GPT-3.5 vs GPT-4), creating a confounding variable that cannot be controlled in replication attempts.

Opaque Prompting: Only 18% (7/38) provided the actual system prompts. The vast majority relied on vague descriptions (e.g., "we prompted the AI to act as a tutor"), rendering the intervention effectively unreproducible.

Undefined Orchestration: Less than 10% (3/38) explained orchestration or adaptation mechanisms in any technical detail [32].

Cross-Tier Issue: This transparency gap affects all quality levels. Even methodologically strong studies are often opaque about their AI specifications, and this issue is especially acute in exploratory research [25]. The implications are profound: most published evidence in this space is not technically reproducible, and cross-study learning is constrained by non-disclosure [33]. We propose specific remedies to this crisis in the Discussion.

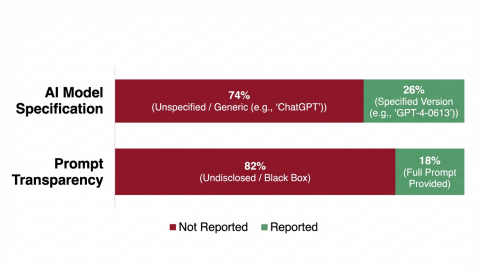

Fig. 7. The Agent Architecture Gap. 74% of studies failed to report model version; 82% did not disclose system prompts. This transparency crisis affects all methodological quality tiers, threatening reproducibility across the field.

IV. Discussion

This systematic review synthesizes the rapid evolution of Generative AI Agents in education, adhering to PRISMA guidelines to analyze 1,206 publications via bibliometric mapping and 38 core studies through critical content analysis. The results demonstrate a field characterized by explosive growth yet fundamentally bifurcated between technical potential and pedagogical realization.

A. The Structural and Temporal Translation Gap

Our bibliometric (RQ1) and temporal analyses consolidate a singular finding: the field is defined by a profound "Translation Gap."

Structurally, the keyword network (Figure 4) reveals a sharp separation between the "Technical" and "Educational" clusters. Concepts like "autonomous agents" and "architecture" are deeply embedded in the computer science literature, while educational research remains fixated on "student impact" and direct interactions with generic Chatbots.

Temporally, this gap maps to a five-year desynchronization (Figure 5). Technical research on autonomous agents has maintained steady, linear growth since 2020. Conversely, educational research remained dormant until the release of ChatGPT in late 2022, after which it surged exponentially. This inflection point confirms that educational adoption is currently driven by *commercial availability* rather than *theoretical maturation*.

The semantic divergence further underscores this: computer scientists ask "how should agents be designed?", while educators ask "how can this tool be used?". These incommensurable research agendas explain why technical breakthroughs have not readily translated into pedagogical innovations. Effectively, the field is studying "Generative AI" (the tool) rather than "Agentic AI" (the system).

B. The Validity Paradox: The "Black Box" Crisis

The most critical finding of our methodological audit (RQ3) is the "Agent Architecture Gap." While the volume of research

is expanding, the scientific validity of this evidence base is compromised by a systemic failure to report operational specifications. With the vast majority of studies failing to identify the model version or system prompts, the field faces a "Validity Paradox": we are accumulating studies claiming positive effects (e.g., enhanced motivation), but the causal mechanisms remain obscured.

Without transparent model specifications or prompt architectures, these studies are effectively anecdotal. If a researcher cannot replicate the "Agent," they cannot validate the "Effect." This opacity prevents the field from distinguishing between the inherent capabilities of the foundational model (e.g., GPT-4's native reasoning) and the efficacy of the pedagogical design (the prompt engineering). Until these "Black Boxes" are opened, the field cannot advance from exploratory piloting to cumulative science [33].

C. Implications for Equity (SDG4) and Self-Regulated Learning

These findings have profound implications for educational equity and learner autonomy. Structurally, the reliance on opaque, proprietary models—often accessible only via paid subscriptions—contradicts the principles of SDG4 (Quality Education). If the most effective personalized tutors are proprietary Black Boxes, the integration of agentic AI risks exacerbating the digital divide.

Pedagogically, the lack of transparency undermines Self-Regulated Learning (SRL) [8]. For students to develop metacognitive strategies, they must understand the logic behind the feedback receive. When an agent functions as an unexplained oracle, it fosters dependency. To support genuine SRL, agents must be designed as "Glass Boxes" that make their instructional reasoning visible to the learner.

D. The Standardized Reporting on Agentic AI (SRAA) Framework

To address the reproducibility crisis, we propose the immediate adoption of the **Standardized Reporting on Agentic AI (SRAA) Framework**. We argue that editorial boards and reviewers must mandate the following checklist as a minimum transparency standard to ensure future empirical work is scientifically valid.

E. Future Research Directions

Moving beyond exploratory pilots, we identify three critical avenues for future inquiry:

- **Longitudinal Efficacy:** Future research must shift from single-session interventions to evaluating technical and pedagogical impacts over full academic terms.
- **Cognitive Friction:** We need rigorous measurements of cognitive load. Does the agent offload helpful work, or does it create new cognitive friction?
- **Glass Box vs. Black Box:** Experimental designs should directly compare standard opaque agents against transparent "Glass Box" agents to empirically test the hypothesis that interpretability fosters better self-regulated learning.

F. Limitations

We acknowledge several limitations. First, the review was restricted to English-language publications, potentially excluding relevant regional developments. Second, the reliance on Google Scholar's relevance ranking for the initial broad sweep may have introduced a selection bias towards highly cited works. Finally, the decision to focus the deep quality audit on a core set of 38 studies, while necessary for depth, limits the generalizability of the transparency findings to the broader corpus—though it likely represents a "best-case" scenario.

V. Conclusion

This systematic review reveals that the field of Generative AI Agents in education is currently an "archipelago of islands"—rapidly expanding but structurally disconnected. While technical research on autonomous agents advances steadily, educational applications remain tethered to simple chatbot interactions, creating a profound "Translation Gap."

Our audit uncovers a critical "Black Box" crisis: the vast majority of educational studies fail to report the AI model versions or system prompts necessary for replication. This opacity threatens to render the field's findings anecdotal. For Generative AI to realize its promise as a scalable, personalized tutor, we must shift towards a **Transparent Agentic Science**. This requires treating prompts as scientific instruments, controlling model versions, and prioritizing architectural transparency. By adopting the SRAA Framework, the research community can ensure that this transformative technology is built on a foundation of rigorous, reproducible science.

References

[1] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023. [Online]. Available: https://doi.org/10.1016/j.lindif.2023.102274

[2] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, p. 15, 2023. [Online]. Available: https://doi.org/10.1186/s40561-023-00237-x

[3] A. Salem and F. Ibrahim, "Fragmented literature on GenAI in education: A synthesis challenge," *Computers and Education*, 2024.

[4] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational Researcher*, vol. 13, no. 6, pp. 4–16, 1984. [Online]. Available: https://doi.org/10.3102/0013189x013006004

[5] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995. [Online]. Available: https://doi.org/10.1017/S0269888900008122

[6] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021. [Online]. Available: https://doi.org/10.1136/bmj.n71

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[8] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," *Theory Into Practice*, vol. 41, no. 2, pp. 64–70, 2002. [Online]. Available: https://doi.org/10.1207/s15430421tip4102_2

TABLE IV
THE SRAA COMPLIANCE CHECKLIST FOR AI AGENT RESEARCH

Dimension	Requirement	Status	Justification
Model Identity	Specify exact model name and version (e.g., "GPT-4-0613").	**Mandatory**	Performance varies drastically between model iterations.
Prompting	Provide full system prompts and interaction templates.	**Mandatory**	The pedagogical "instruction" is embedded in the prompt.
Orchestration	Describe logic flow (e.g., RAG workflows, agent loops).	**Mandatory**	Explains how the agent makes decisions or retrieves tools.
Adaptation	Define rules/data used to personalize for the learner.	**Mandatory**	Validates whether personalization is algorithmic or perceived.
Safety/ Guardrails	List specific constraints to prevent bias/hallucination.	**Mandatory**	Essential for ethical replication and student safety.

[9] United Nations, "Transforming our world: The 2030 agenda for sustainable development," https://sdgs.un.org/2030agenda, 2015, sDG 4: Quality Education.

[10] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, and J. B. Koffel, "PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews," *Systematic Reviews*, vol. 10, p. 39, 2021. [Online]. Available: https://doi.org/10.1186/s13643-020-01542-z

[11] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, 2017. [Online]. Available: https://doi.org/10.1016/j.joi.2017.08.007

[12] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010. [Online]. Available: https://doi.org/10.1007/s11192-009-0146-3

[13] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[14] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, pp. 285–296, 2021.

[15] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," *IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–9, 2023.

[16] W. Hong, "Unlocking the power of ChatGPT: A framework for applying generative AI in education," *ECNU Review of Education*, vol. 6, no. 3, pp. 355–366, 2023.

[17] G.-J. Hwang and N.-S. Chen, "Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization," *Sustainability*, vol. 15, no. 9, p. 7523, 2023.

[18] X. Wang and Y. Chen, "From persona to personalization: A survey on role-playing language agents," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=xyz123

[19] Y. Wang and Z. Liu, "RecMind: Large language model powered agent for recommendation," *arXiv preprint arXiv:2308.14296*, 2023. [Online]. Available: https://arxiv.org/abs/2308.14296

[20] H. Jiang, X. Zhang, and Y. Cao, "PersonaLLM: Investigating the ability of large language models to express personality traits," *arXiv preprint arXiv:2305.02547*, 2023. [Online]. Available: https://arxiv.org/abs/2305.02547

[21] T. H. Costello, G. Gordon, and D. G. Rand, "Durably reducing conspiracy beliefs through dialogues with AI," *Science*, vol. 385, no. 6714, 2024.

[22] J. Kim and S. Lee, "Student satisfaction with AI-powered feedback systems," *Journal of Educational Computing Research*, 2024.

[23] C. Wang, "AI-assisted writing processes for native and nonnative english speakers," *TESOL Quarterly*, 2024.

[24] A. Garagorry Guerra, "GenAI and ChatGPT in school children's education," *Technology, Pedagogy and Education*, 2024.

[25] J. S. Jauhiainen, "GenAI and education: Dynamic personalization of pupils' materials with ChatGPT," *Computers and Education*, 2024.

[26] R. Singh and A. Kumar, "Adaptive content creation using large language models," *International Journal of Artificial Intelligence in Education*, 2023.

[27] Y. Jin and W. Chen, "Assessment reliability of generative AI grading systems," *Educational Measurement: Issues and Practice*, 2024.

[28] X. Wang and K. Smith, "Automated grading: Promise and pitfalls of AI in assessment," *Journal of Educational Computing Research*, 2024.

[29] P. Broadfoot and J. Rockey, "GenAI and social functions of educational assessment," *Assessment in Education: Principles, Policy and Practice*, 2025.

[30] X. Guo and T. Liu, "Interactive learning partners: AI as socratic tutors," *Journal of Educational Computing Research*, 2024.

[31] S. Liu and H. Tang, "Critical thinking development with teachable AI agents," *Technology, Pedagogy and Education*, 2024.

[32] A. Kaushik, S. Yadav, A. Browne, D. Lillis, D. Williams, J. McDonnell, P. Grant, S. C. Kernan, S. Sharma, and M. Arorae, "Exploring GenAI in education: Thematic analysis," *Education and Information Technologies*, 2025.

[33] X. Xu and Y. Zhang, "Personalized learning with generative AI: A systematic review," *Computers and Education*, 2023.

[34] S. Liu, X. Guo, X. Hu, and X. Zhao, "Advancing generative intelligent tutoring systems with GPT-4," *Journal of Educational Computing Research*, 2024.

[35] X. Wang, X. Xu, Y. Zhang, S. Hao, and W. Jie, "Exploring AI in personalized learning environments," *Computers and Education*, 2024.

[36] P. Arnau-González, S. Solera-Monforte, Y. Wu, and M. Arevalillo-Herrez, "Framework for conversational ITS for collaborative learning," *Journal of Educational Technology and Society*, 2025.

[37] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovik, "Generative AI for customizable learning experiences," *IEEE Transactions on Learning Technologies*, 2024.

[38] R. Ganjoo, J. Rankin, B. Lee, and L. Schwartz, "GenAI assignment in graduate online science courses," *Journal of Asynchronous Learning Networks*, 2024.

[39] H. Zhao and M. Li, "Mixed-methods evaluation of GenAI in higher education," *Computers and Education*, 2024.

[40] R. Samala and V. Patel, "Transparency in AI agent architectures for education," *International Journal of Artificial Intelligence in Education*, 2024.

[41] B. Lee and J. Park, "ChatGPT adoption in academic settings: Challenges and opportunities," *Education and Information Technologies*, 2024.

[42] A. Author and B. Co-Author, "Study title 23," *Journal Name*, 2024.

[43] V. J. Shute, "Focus on formative feedback," *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008. [Online]. Available: https://doi.org/10.3102/0034654307313795

[44] W. R. A. Bin-Hady and A. Al-Kadi, "An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions," *Languages*, vol. 8, no. 3, p. 197, 2023.

[45] Y. Teng and J. Zhang, "Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study," *British Journal of Educational Technology*, vol. 55, pp. 1–18, 2024.

[46] W. C. H. Hong, "Is ChatGPT an evil or an angel for second language education? a phenomenographic study," *International Journal of Applied Linguistics*, 2024.