computers

MDPI

Review

RAG in Self-Regulated Personalised Learning in Digital Education

Isaac Okiyai *, Lucy Mburu and Edwin Juma Omol

School of Technology, KCA University, Nairobi 00200, Kenya
* Correspondence: 2108557@students.kcau.ac.ke or okiyai52@gmail.com

Abstract

Large language models (LLMs) generate fluent text, but they hallucinate and work from knowledge that ages quickly. Retrieval-Augmented Generation (RAG) answers both challenges by grounding generation in external sources a reader can check. At issue is whether generative artificial intelligence (AI) can be used trustworthily in high-stakes domains such as education. Self-regulated learning (SRL) research has argued for decades that students who set goals, track their own progress, pick strategies deliberately and reflect on the results learn more deeply and retain more, and that these matter most in the loosely structured, learner-directed conditions typical of digital education platforms. This review examines where the two lines of work meet, and what evidence exists on how RAG systems are being designed, deployed and evaluated as support for self-regulated, personalised learning. We first trace RAG from classical information-retrieval models through neural dense retrieval to the current generation of large models. We then map its applications onto the phases of self-regulation: goal setting and planning, retrieval-grounded tutoring and conversational scaffolding, personalised content retrieval and adaptive sequencing, automated assessment and feedback, learning-analytics-based monitoring, curriculum-grounded content creation, and support for diverse learners. Thereafter we examine the challenges that qualify this promise, including retrieval-conditioned hallucination, data privacy and security risks specific to vector-based retrieval, algorithmic and retrieval fairness, the interpretability of retrieved evidence, infrastructural and implementation barriers, and disruptions to academic integrity. In this review conclusion, it is noted that RAG offers a feasible and increasingly evidenced pathway to a personalised, scalable and pedagogically grounded support for self-regulated learning, but that realising this potential responsibly requires deliberate attention to the quality of retrieval, along with rigour in terms of evaluation as well as learner agency rather than technical capability alone.

Keywords: retrieval-augmented generation; self-regulated learning; personalised learning; digital education; large language models; intelligent tutoring systems; learning analytics; educational technology

Check for updates

Academic Editor: Paolo Bellavista

Received: 2 September 2026
Revised: 17 September 2026
Accepted: 19 September 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Three terms get used interchangeably in discussions of AI in education, and they do not mean the same thing. Artificial intelligence (AI) is the broad field concerned with building systems that perform tasks associated with human intelligence. Machine learning (ML) is the subset of that field in which systems learn patterns from data rather than following rules written by hand. Generative AI is narrower again: large neural models that produce

new text, code or media rather than classifying inputs into fixed categories [1]. Retrieval-Augmented Generation (RAG) sits inside generative AI as one particular way of building such a system. Instead of drawing only on what the model absorbed into its parameters during pre-training, a RAG system retrieves passages from an external knowledge store at query time and writes its answer from what it finds there [2]. That store can be updated without retraining the model. The retrieve-then-generate design was introduced to fix two known weaknesses of purely parametric large language models (LLMs). They hallucinate, producing fluent content that is simply wrong. They also cannot reach anything published after their training cutoff, or anything held privately by an institution or a course [3].

Retrieval grounding and citation generation are related but not interchangeable. Retrieval grounding refers to the mechanism itself: constraining what the generator can say by conditioning it on passages actually retrieved from a defined corpus at query time. Citation generation refers to a further, optional step in which the system additionally surfaces which retrieved passage supports which part of its answer, typically as an inline reference or link. A RAG system can be grounded without generating visible citations, and a system can produce citation-like text without genuine retrieval grounding behind it, for instance, where a model fabricates a plausible-looking reference. The distinction matters pedagogically: grounding constrains what the model is likely to say, whereas citation generation is what lets a learner or instructor actually verify it, and the two should not be conflated when evaluating a system's trustworthiness [4,5].

The appeal of RAG for education follows directly from these properties. Digital education platforms increasingly rely on generative AI to power conversational tutors, automated feedback tools, and content-recommendation engines, yet the correctness and provenance of what such tools say to learners carries pedagogical and, in professional or credentialing contexts, ethical weight that generic consumer chatbot use does not [6]. By grounding generated responses in a curated corpus, such as a course's syllabus, textbook, lecture transcripts, or an institution's assessment rubrics, a RAG-based educational tool can in principle offer the fluency and interactivity of a large language model while remaining traceable to a specific, auditable source. That property matters for self-regulated learning (SRL), the research tradition concerned with how learners plan, monitor and adjust their own thinking and motivation while working towards a goal [7–9]. Learners in loosely structured, self-paced digital environments have to regulate themselves far more than learners in a tightly scaffolded classroom. Technology that ignores this, or undercuts it by handing over answers without engaging the learner's own monitoring and strategy use, can make outcomes worse [10,11]. Keeping effort and interest up when the work gets hard is its own part of SRL, and it is the part that gets overlooked. A support tool has to handle motivation as well as cognitive and metacognitive strategy [12]. Reviews of scaffolding in computer-based learning environments do find measurable gains in planning and monitoring, provided the prompts and tools are designed well [13]. A recent scoping review has started mapping SRL constructs onto technology-enhanced learning research in higher education, and RAG sits inside that map [14]. RAG systems can be tied to a specific curriculum and set up to scaffold rather than simply answer. This review asks how far that combination can carry SRL support at scale.

The bridge between RAG's technical affordances and SRL's pedagogical goals is worth making explicit, since the two literatures do not automatically synthesise just because they are placed side by side. RAG contributes three properties that matter for self-regulation specifically: groundedness, so that a learner can check a claim against its source rather than take it on faith; currency, so that content reflects an updated corpus rather than a frozen training snapshot; and traceability, so that the evidence behind a generated statement is, in principle, inspectable. Each property maps onto a specific SRL demand: groundedness

supports the monitoring sub-process, where learners judge whether their current understanding is correct; currency supports planning against material that has not gone stale; and traceability supports the kind of self-reflection that depends on being able to interrogate, not just receive, an answer. Section 4 organises the applications literature around this mapping, phase by phase, rather than treating RAG as a generic support layer bolted onto existing SRL scaffolds [2,3,7].

Adoption is moving quickly on both the commercial and the institutional side, which adds urgency. Industry analysts put the global AI-in-education market at about USD 8.3 billion in 2025, rising to USD 11.4 billion in 2026 and USD 57.2 billion by 2033, a compound annual growth rate near 25.9 percent [15]. The RAG market alone was estimated at USD 1.2 billion in 2024 and is forecast to grow at roughly 49.1 percent annually, reaching USD 11.0 billion by 2030 [16]. In a few years, RAG has gone from a research technique to deployed enterprise and educational infrastructure. That makes it worth asking what is actually known, technically and pedagogically, about RAG's role in supporting self-regulated personalised learning.

2. Materials and Methods

This paper adopts a narrative review of the literature on retrieval-augmented generation (RAG) in self-regulated and personalised learning. A narrative rather than systematic design was chosen because the field is young, methodologically heterogeneous, and dominated by system-description papers that do not share a common outcome measure, so meta-analytic synthesis is not yet feasible [17]. A narrative approach was therefore selected deliberately rather than by default, following the review-typology guidance that matches design to the maturity and heterogeneity of the underlying evidence base [18]. This review nonetheless follows a documented and reproducible search procedure, reported here so that readers can assess coverage and bias.

This review is best described as a structured narrative review: it combines a systematic, reproducible search and screening process with a non-meta-analytic, thematically organised synthesis appropriate to a young and methodologically heterogeneous evidence base. Of the PRISMA 2020 reporting items, this review adopts an explicit search strategy across named databases, defined eligibility criteria, a documented screening and inclusion count reported in Figure 1, and transparent reporting of information sources; it does not adopt PRISMA's requirements for risk-of-bias assessment of individual studies, a fixed outcome-level data-extraction template, or meta-analytic pooling, none of which suit a literature still dominated by system-description papers without a shared outcome measure. Where PRISMA conventions and narrative-review conventions diverge on this point, the sections that follow favour the interpretive, theme-based synthesis expected of a narrative review [18,19]. A completed PRISMA 2020 checklist, indicating where each applicable item is reported in this manuscript and identifying items not applicable to this narrative-review design, is provided as Supplementary Table S1 [19].

The literature was identified through structured searches of [Scopus, Web of Science, ACM Digital Library, IEEE Xplore, ERIC, and Google Scholar], covering publications from 2017 to August 2026. The lower bound was set at 2017 to capture the transformer architecture and the neural retrieval work that made RAG possible; foundational pre-2017 sources in information retrieval and self-regulated learning theory were added through backward citation chaining where they were required to establish conceptual lineage. Search strings combined three concept blocks with Boolean operators: (i) retrieval terms ("retrieval-augmented generation" OR "RAG" OR "retrieval-augmented" OR "grounded generation"); (ii) learning terms ("self-regulated learning" OR "personalised learning" OR "adaptive learning" OR "intelligent tutoring"); and (iii) context terms ("education" OR "digital learn-

ing" OR "learner"). Reference lists of included papers were screened for additional sources, consistent with evidence that no single database offers adequate coverage on its own and that multi-database searching with citation chaining improves retrieval completeness [20].

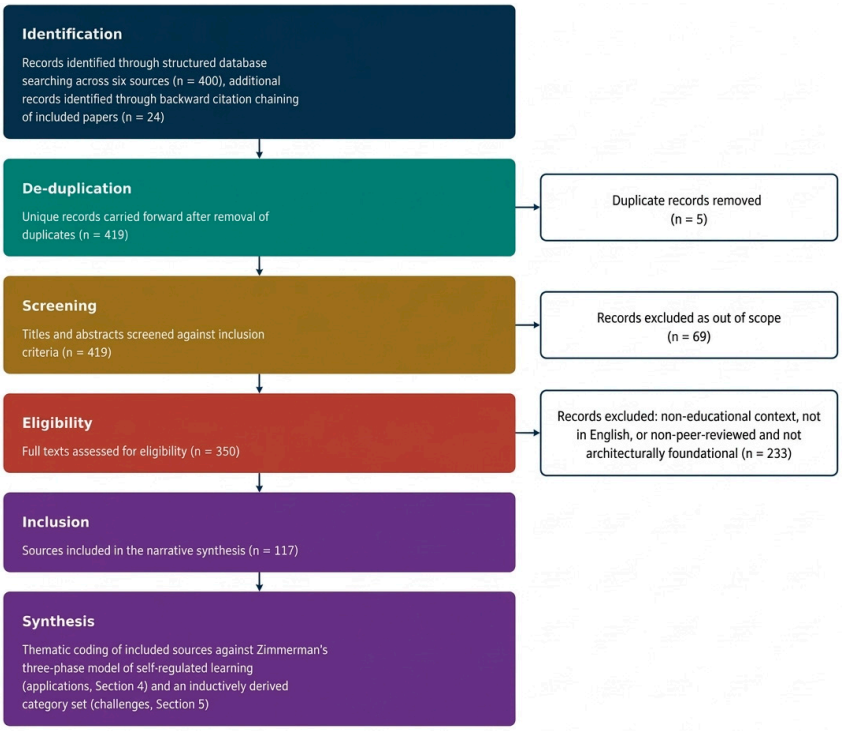


Figure 1. Literature identification, screening, and synthesis workflow for this review.

For reproducibility, the concept blocks above were translated into the following database-specific strings. Scopus (TITLE-ABS-KEY): ("retrieval-augmented generation" OR "RAG" OR "grounded generation") AND ("self-regulated learning" OR "personalised learning" OR "personalized learning" OR "adaptive learning" OR "intelligent tutoring") AND ("education" OR "digital learning" OR "learner"). Web of Science (Topic): analogous string using TS = in place of TITLE-ABS-KEY. IEEE Xplore and the ACM Digital Library: the same three concept blocks combined with AND, restricted to the Abstract field where the interface permits it. ERIC (ProQuest/EBSCO): the same Boolean string restricted to the Descriptor and Abstract fields. Google Scholar and arXiv, which do not support full field-restricted Boolean search, were searched with the shorter phrase pairs ("retrieval-augmented generation" AND "self-regulated learning") and ("RAG" AND "personalised learning"), sorted by relevance, with the first 200 results per query screened, since screening only the first 50–100 results is not itself an evidence-based cutoff and relevant records continue to appear well beyond that point [21].

Records were included where they reported the design, deployment, or evaluation of a retrieval-grounded generative system in an educational setting, or where they contributed theoretical or empirical grounding on self-regulated learning that the review draws on directly. Records were excluded where they addressed retrieval or generation outside an educational context without transferable implications, where they were not available in English, or where they were non-peer-reviewed material other than widely cited preprints that introduce architectures central to the field. Preprints were retained selectively on that last basis and are identified as such in the reference list.

Figure 1 summarises the identification, screening, and synthesis workflow, adapted for a narrative review from the PRISMA 2020 reporting conventions [19]. Searches returned 424 records (400 through structured database searching across the six sources and

24 through backward citation chaining of included papers). After the removal of 5 duplicate records, 419 unique records were screened by title and abstract, of which 69 were excluded as out of scope; 350 full texts were then assessed for eligibility, of which 233 were excluded (non-educational context, not in English, or non-peer-reviewed and not architecturally foundational), leaving 117 sources included in the narrative synthesis and cited throughout this review. Included sources were organised thematically rather than chronologically, using Zimmerman's three-phase model of self-regulated learning (forethought, performance, self-reflection) [7] as the organising frame for the applications discussed in Section 4, and an inductively derived category set for the challenges discussed in Section 5. Themes were developed by reading each source against both frames and revising the category boundaries iteratively until all included sources were accommodated, following the six-phase approach to the thematic analysis of qualitative and documentary material [22].

3. Historical Development of Retrieval-Augmented Generation

3.1. Early Foundations: Information Retrieval and Computer-Assisted Instruction

The retrieval component of RAG has intellectual roots that precede neural networks by decades. Salton, Wong, and Yang's vector space model, published in 1975, represented documents and queries as vectors in a shared term space and ranked documents by geometric similarity to a query, establishing the basic retrieve-by-similarity paradigm that all subsequent dense and sparse retrieval methods, including those used in modern RAG pipelines, still instantiate in some form [23]. Probabilistic ranking functions such as Okapi BM25, formalised by Robertson and Zaragoza, extended this paradigm with term-frequency and document-length normalisation and remain the dominant sparse retrieval baseline against which neural retrievers are still benchmarked [24]. On the generation and pedagogy side, computer-assisted instruction and early intelligent tutoring systems (ITSs) established the idea that a computer program could adapt its presentation of content to an individual learner's responses. Nwana's early overview of intelligent tutoring systems catalogued the architectural components, a domain model, a student model, and a pedagogical/tutoring model, that later RAG-based tutors would re-implement with a retrieval-augmented generative core in place of hand-authored production rules [25]. Meta-analytic evidence subsequently confirmed that well-designed ITSs could approach the effectiveness of human one-on-one tutoring, providing the empirical benchmark that generative-AI tutoring, including RAG-based tutoring, is still implicitly measured against [26,27]. Together, these two threads set the terms RAG would later merge: a retrieval paradigm proven at the level of document similarity, and a tutoring paradigm proven at the level of adaptive, individualised instruction, with the gap between them, connecting what is retrieved to what is taught, left for the neural and generative advances described next to close.

3.2. Emergence of Neural Information Retrieval and Neural Language Generation

The 2017–2020 period saw two previously separate research threads, neural information retrieval and neural language generation, converge into the technical substrate from which RAG emerged. Vaswani et al.'s transformer replaced recurrent and convolutional sequence models with self-attention, and it became the shared backbone for retrievers and generators alike [28]. Devlin et al.'s BERT then showed that bidirectional pre-training on large unlabelled corpora yields representations that transfer well to downstream tasks, passage ranking among them [29]. Reimers and Gurevych's Sentence-BERT adapted the same architecture to produce sentence-level embeddings usable for similarity search at scale [30]. Retrieval moved next. Karpukhin et al.'s Dense Passage Retrieval (DPR) [31] trained a dual-encoder to put questions and their answering passages near each other in embedding space, and it beat sparse lexical retrieval on open-domain question answering

by a wide margin. The original RAG architecture used DPR's retriever directly. Around the same time, Guu et al.'s REALM pre-trained a masked language model jointly with a learnable retriever, which put retrieval inside the pre-training objective instead of bolting it on afterwards [32]. Khattab and Zaharia's ColBERT offered late-interaction retrieval, more expressive than single-vector dense retrieval and more expensive to run [33]. None of this scales without fast approximate nearest-neighbour search over the resulting vector spaces, which FAISS [34] and hierarchical navigable small world (HNSW) graphs [35] made practical. Both sit under the vector databases used in production RAG systems today. On the generation side, Brown et al.'s GPT-3 [36] showed that a large enough autoregressive model can take on new tasks from a handful of in-context examples, with no task-specific fine-tuning. That few-shot capability is what RAG systems point at retrieved context.

3.3. The Retrieval-Augmented Generation Architecture and the Generative AI Revolution

Lewis et al. [2] themed the architecture in 2020: a pre-trained parametric sequence-to-sequence generator paired with a non-parametric dense vector index reached through a learned retriever. The hybrid beat purely parametric models on knowledge-intensive tasks, and its outputs were more specific, more varied and better grounded. Nearly every education-facing RAG system in Section 4 descends from it. As large language models multiplied and the public took up conversational AI from 2022, far more practitioners ran into the problem RAG seeks to mitigate: hallucination and insufficiently grounded responses, and answers that are neither current nor traceable to a source. The refinement of the architecture came quickly. Asai et al.'s Self-RAG introduced adaptive retrieval, in which the model itself learns when retrieval is necessary and generates explicit reflection tokens to critique the relevance and support of its own retrieved evidence [4]. Yan et al.'s Corrective RAG added a lightweight retrieval evaluator that can trigger corrective actions, including falling back to web search, when retrieved documents are judged irrelevant, improving robustness against poor retrieval [37]. Edge et al.'s GraphRAG combined automatically constructed knowledge graphs with retrieval to support holistic, corpus-level queries that single-passage retrieval struggles to answer [38], while Gao et al. formalised a "Modular RAG" framework, decomposing RAG systems into composable routing, retrieval, and generation modules [39]. Liu et al.'s empirical finding that LLMs use long contexts unevenly, performing worse when relevant information is buried in the middle of a long retrieved context, directly informs the context-ordering and reranking choices made by education-facing RAG systems [40], and Es et al.'s RAGAS framework supplied the reference-free evaluation metrics, faithfulness, answer relevance, and context precision and recall, that most subsequent RAG evaluation work, educational or otherwise, now reports against [5]. Comprehensive surveys by Gao et al. [3], Fan et al. [41], and, most recently, on agentic [42] and multimodal [43] RAG architectures document the field's rapid maturation from a single 2020 paper into a broad architectural family within only a few years, a maturation corroborated by a further systematic literature review that consolidates the fragmented terminology, evaluation metrics, and open technical challenges reported across this fast-growing body of work [44].

Stated formally, the dense retrieval of the kind used by DPR and its successors scores a candidate passage d against a query q with a dual-encoder inner product, $S(q, d) = E_Q(q) \cdot E_D(d)$, where E_Q and E_D are learned query and passage encoders and the top-scoring passages are returned as the retrieved set [31]. The RAG architecture itself marginalises the generator's output distribution over that retrieved set rather than conditioning on a single passage: $P(y|x) = \Sigma_z P(z|x)\, P(y|x, z)$, summed over the top-k retrieved documents z, so the final answer is shaped by all retrieved evidence jointly rather than by whichever single passage ranked first [2]. Table 1 compares these paradigms directly, contrasting

traditional information retrieval, neural retrieval, LLM-only generation, and RAG across the dimensions most relevant to educational use.

Table 1. Comparison of traditional information retrieval, neural retrieval, LLM-only generation, and RAG architectures.

Paradigm	Representative Method(s)	Basis for Output	Updatable Without Retraining?	Grounded/ Traceable Output?	Typical Educational Use	Key Limitation
Traditional (sparse) information retrieval	Vector space model; Okapi BM25 [23,24]	Term-overlap/geometric similarity between query and document	Yes (re-index)	Retrieval only—no generated answer	Library/catalogue search; early computer-assisted instruction content lookup	Returns ranked documents, not a natural-language answer
Neural (dense) retrieval	DPR; REALM; ColBERT [31–33]	Learned embedding similarity in a shared vector space	Yes (re-embed and re-index)	Retrieval only—no generated answer	Passage retrieval feeding a downstream QA or tutoring system	Needs re-embedding when the corpus changes materially; opaque similarity space
LLM-only (parametric) generation	GPT-3-class large language models [36]	Knowledge encoded in model parameters at training time	No (requires fine-tuning/re-training)	Fluent, but not inherently traceable to a source	Open-ended tutoring dialogue; content drafting	Hallucination; knowledge frozen at training cutoff
Retrieval-Augmented Generation (RAG)	Lewis et al. and successors, e.g., Self-RAG, Corrective RAG, GraphRAG [2,4,37,38]	Generation conditioned on passages retrieved from an external corpus at query time	Yes (corpus updatable independently of the model)	Traceable to retrieved passages, subject to the grounding/citation distinction above	Course-grounded tutoring; curriculum-aligned content and assessment generation	Residual hallucination; retrieval-quality dependence; corpus curation burden

3.4. Evolution of RAG Research Themes Towards Self-Regulated Personalised Learning

The application of RAG specifically to education, and more specifically still to self-regulated and personalised learning, is a recent and still-consolidating research theme. Early education-facing work such as Levonian et al.'s study of RAG for mathematics question answering, grounded in an OpenStax textbook, was already explicitly concerned not merely with factual accuracy but with the trade-off between groundedness and human (educator) preference, foreshadowing the pedagogical, rather than purely technical, evaluation criteria that would come to characterise the field [45]. By 2025, there were dedicated surveys of RAG chatbots in education [46] and of educational RAG applications generally [42,47], a sign the field had grown large enough to need synthesis. It covered in a few years what ML in education took decades to do. Part of that literature now ties RAG and generative AI directly to SRL theory. Instead of the generic "AI tutor" framing, these systems and their evaluations are organised around the phases of self-regulation, as reviewed in Section 4 [48–50].

Figure 2 sets this out as a timeline: classical information retrieval and early intelligent tutoring systems (pre-2017), the neural retrieval and transformer era (2017–2020), the arrival of RAG and its rapid branching into variants (2020–2024), and the current education-specific, SRL-oriented work on application and evaluation (2023–2026).

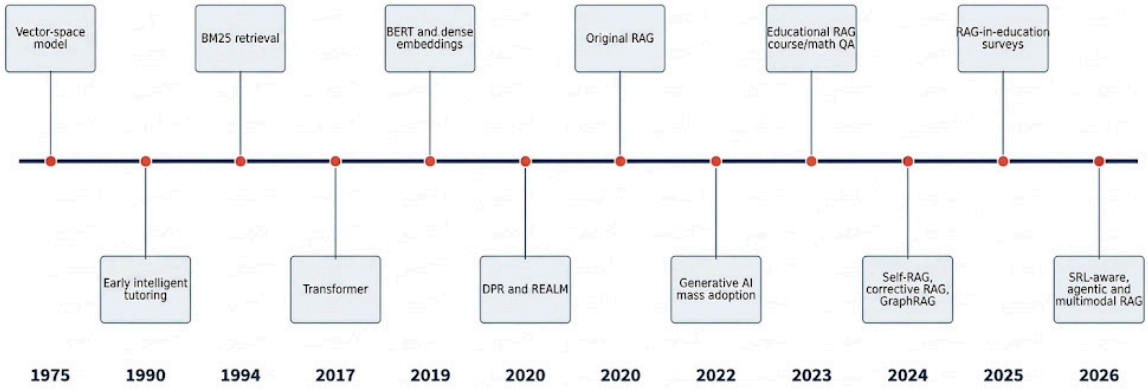


Figure 2. Timeline of key milestones in the development of Retrieval-Augmented Generation and its convergence with self-regulated learning research, 1975–2026.

4. Applications of RAG in Self-Regulated Personalised Learning

Organising the applications literature by the phase of self-regulation it primarily supports, rather than purely by technical function, makes explicit the theoretical link between RAG's technical affordances and SRL's pedagogical goals. Figure 3 presents this mapping as a categorised overview of the RAG-in-SRL application space; the following subsections review the evidence behind each category. Table 2 sets out the same mapping in tabular form, listing the representative systems reviewed under each SRL phase.

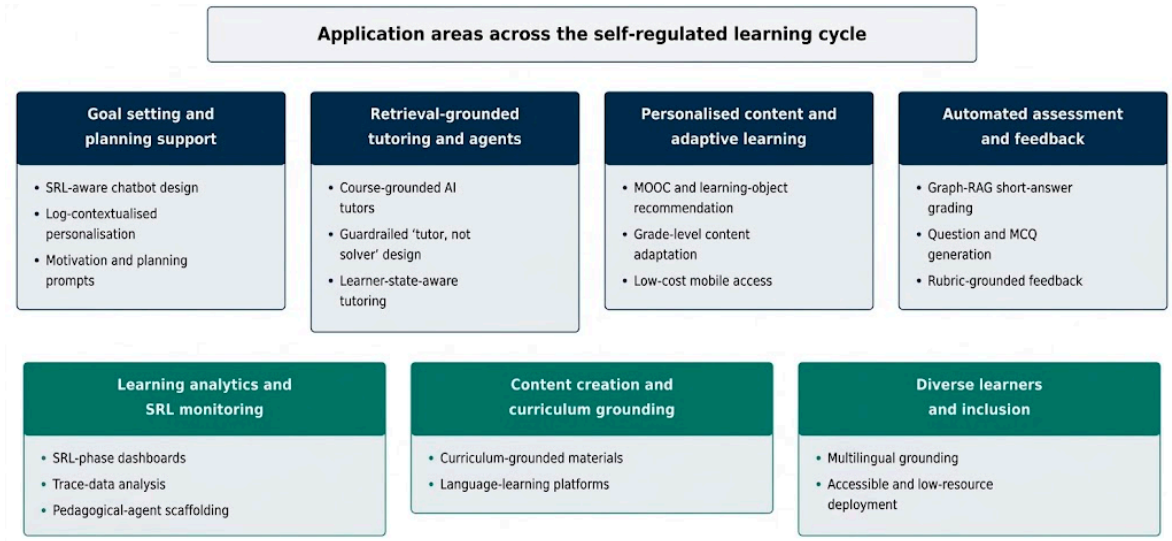


Figure 3. A categorised overview of RAG applications in self-regulated personalised learning, organised by SRL phase: goal setting and planning; tutoring and conversational scaffolding; personalised content retrieval and adaptive learning; automated assessment and feedback; learning analytics and monitoring; content creation and curriculum grounding; and support for diverse learners.

It is also useful to distinguish systems by who the retrieval and generation are ultimately for. Most of the tutoring, content-retrieval, and assessment-feedback systems reviewed in Sections 4.1–4.4 are student-facing: the learner interacts with the RAG system directly, and self-regulation is exercised, or bypassed, in that interaction. Others are teacher- or institution-facing: the learning-analytics dashboards discussed in Section 4.5 and the curriculum-grounded content-generation tools in Section 4.6 are typically operated or consumed by instructors and administrators, with any effect on the learner's self-regulation mediated by how that person acts on the system's output. The distinction matters for evaluation, since a student-facing system's effect on SRL can be measured through learner behaviour and

outcomes, while a teacher-facing system's effect on SRL is necessarily indirect and harder to attribute.

Table 2. Taxonomy of RAG-in-education applications mapped to self-regulated learning (SRL) phase.

Application Category (Section)	Primary SRL Phase	SRL Sub-Process Supported	Representative Systems/Studies
Goal setting and planning (Section 4.1)	Forethought	Goal setting; strategic planning	Log-contextualized RAG [51]; SRL-oriented chatbot design work [49,52]
Retrieval-grounded tutoring and conversational agents (Section 4.2)	Performance	Self-monitoring; strategy application	PeteChat [53]; MDKAG [54]; MCP/ACP protocol-level ITS [55]
Personalised content retrieval and adaptive learning (Section 4.3)	Forethought/Performance	Strategy selection; adaptive sequencing	RAMO [56]; CLARE [57]; SP-TeachLLM [58]
Automated assessment and feedback (Section 4.4)	Self-reflection	Outcome evaluation; feedback uptake	Xu and Corter GraphRAG grading [59]; Maity et al. [60]
Learning analytics and monitoring (Section 4.5)	Performance/Self-reflection	Self-monitoring; trace-based reflection	Jivet et al.'s dashboard review [61]; MetaTutor [62]
Content creation and curriculum grounding (Section 4.6)	Forethought	Resource selection supporting planning	College English platform [63]
Supporting diverse learners and inclusive education (Section 4.7)	Cross-phase	Equitable access across all three phases	College English platform [63]; mobile-first RAG [64]

4.1. Supporting Goal Setting and Planning

The forethought phase of SRL, in which learners set goals, activate prior knowledge, and plan a strategy, is the least directly addressed by current RAG systems, most of which are reactive, responding to a learner's query rather than proactively scaffolding goal formation, but a growing body of work treats this gap explicitly. Conceptual and empirical work on adapting SRL theory to generative-AI chatbot contexts argues that chatbots must be deliberately designed to prompt goal articulation and planning rather than simply answering whatever is asked, or risk short-circuiting the forethought phase entirely [49]. Reviews of chatbot design for SRL make the same point, among them one built on Habermas's theory of communicative, strategic and emancipatory interests. A chatbot that supports SRL has to be built to draw learners into planning dialogue rather than answer questions as they come [52]. Cohn et al. [51] show one way to do that. Their log-contextualized RAG retrieves from a learner's own interaction history to personalize what the agent says, so the system is grounded in course content and in the learner's stated goals and past work, which makes the planning support specific to that learner.

4.2. Retrieval-Grounded Intelligent Tutoring and Conversational Agents

The performance phase of SRL, where learners monitor their understanding and apply strategies as they work, gets the most attention in the current literature, mostly through retrieval-grounded conversational tutors. Course-grounded RAG tutors have been evaluated in technical education [65], introductory psychology [66], chemistry (where the chatbots were used to support and assess student reasoning rather than deliver answers) [67], and language learning [63]. One classroom study found that grounding LLM tools in course-specific retrieval improved both the relevance and instructor-rated usefulness of generated responses compared with ungrounded generation [68]. Deployments run from

locally hosted, domain-specific tutoring systems [69] to tutors wired straight into learning management systems like Moodle [70], and general-purpose tools such as NotebookLM repurposed as a Socratic physics tutor [71]. One worry comes up repeatedly: an agent optimised for helpfulness alone becomes a shortcut that dispenses answers and erodes self-regulation. The PeteChat case study frames its assistant as a guardrailed "tutor, not solver," constraining retrieval and generation so the system scaffolds reasoning instead of handing over final answers [53]. Comparative and evaluative studies reinforce this concern from the other direction: an expert-rated comparison of a prompt-only LLM tutor against a learner-state-aware RAG tutor found the retrieval-grounded, state-aware system rated as more pedagogically appropriate [72], while a multimodal knowledge-graph-grounded intelligent tutoring system demonstrated that structured retrieval sources can support more coherent step-by-step tutoring than flat passage retrieval alone [73]. A closely related architecture, MDKAG, extracts entities from textbooks, slides, and lecture video into a multimodal disciplinary knowledge graph and reports a measurable reduction in hallucination rate alongside improved answer accuracy relative to text-only retrieval, reinforcing that the structure of the retrieval source, not only its content, shapes tutoring quality [54]. Protocol-level architecture is a further emerging design axis: a hybrid intelligent-tutoring system combining RAG with a Model Context Protocol for session- and course-level state and an Agent Communication Protocol for coordinating specialised agents was evaluated against 240 annotated question–answer pairs and human raters, and reduced unsupported claims relative to a baseline RAG pipeline, suggesting that standardised inter-agent protocols are becoming as consequential to tutoring reliability as the underlying retriever or generator [55]. Algorithm-tracing and problem-solving tutors in computer science education [74], scalability-focused deployments of RAG chatbots for higher-education student support [75,76], and domain-specific applications such as a RAG chatbot for obstetric medical education [77] collectively indicate that retrieval-grounded conversational tutoring is now the most mature and empirically tested RAG-in-education use case.

4.3. Personalised Content Retrieval and Adaptive Learning

Beyond conversational tutoring, RAG has been applied directly to the problem of retrieving or recommending the right content for an individual learner at the right time, the core technical problem in adaptive and personalised learning. RAMO applies RAG to improve MOOC recommendation, retrieving semantically relevant course content rather than relying on collaborative-filtering signals alone [56], while the CLARE framework extends this idea with category-aware retrieval for recommending discrete learning objects in both educational and professional-training contexts [57]. Nguyen et al.'s context-aware RAG system explicitly adapts retrieved and generated content to a learner's grade level, operationalising personalisation as a retrieval-and-generation constraint rather than a separate recommendation layer [78], and the RAG-PRISM framework frames adaptive retrieval-augmented tutoring around rapid, personalised skill mastery [79]. In programming education specifically, the SP-TeachLLM framework combines curriculum decomposition, reinforcement-learning-based strategy selection, and reflective assessment to personalize instruction, and reports consistent gains in code-generation accuracy across benchmark tasks, with the largest gains for smaller, more deployable student-facing models, a result directly relevant to the cost and scalability constraints discussed in Section 5.5 [58]. This personalisation ambition is not unique to RAG; a systematic review of self- and shared regulation within adaptive digital learning environments more broadly finds that adaptivity alone does not guarantee support for self-regulation unless the system is explicitly designed to make its adaptive decisions visible and actionable to the learner, a design requirement equally applicable to RAG-based personalisation [80]. Equitable access has driven interest

in lightweight deployment. One example is a scalable, low-cost mobile RAG architecture built to bring AI-augmented learning to institutions and personal devices with limited computing resources [64]. Retrieval reliability is a related focus, since inaccurate retrieved sources feed straight into the personalised output a learner sees [81].

The applications reviewed in this section personalise along two different bases that are worth keeping separate. Some systems personalise from what a learner explicitly states a stated goal, a declared skill level, a selected topic, treating the learner as the authority on their own needs. Others personalise from behavioural analytics inferred from interaction data response times, error patterns, click paths treating the system as the authority on what the learner needs, whether or not the learner agrees [82]. The RAG-PRISM and Nguyen et al. [78] systems described above lean towards the first basis, adapting retrieval to stated grade level or mastery targets, while log-contextualized approaches such as Cohn et al. [51] lean towards the second, inferring what to retrieve from accumulated interaction history. The distinction is not cosmetic: explicit-preference personalisation preserves learner agency by construction, while inferred personalisation risks narrowing what a learner is shown based on a behavioural pattern the learner never consented to or may not recognise as accurate, a tension neither category of system reviewed here resolves.

4.4. Automated Assessment and Feedback

Self-reflection in SRL depends on feedback that reaches learners quickly and holds up on the substance. That is close to what RAG does well: it ties generated text back to source material a learner or instructor can check. Xu and Corter's GraphRAG-based short-answer grading system explicitly bridges pedagogical feedback needs with technical grading capability by grounding both the score and the accompanying explanation in retrieved rubric and reference content [59], while a modular RAG-based scoring system similarly emphasises generating feedback, not only a score, from retrieved reference answers [83]. On the content-generation side of assessment, Maity et al. compared in-context learning, RAG, and hybrid approaches for automatic question generation grounded in course material [60], and subsequent work has applied RAG specifically to multiple-choice question generation [84], including in the resource-constrained setting of generating exam-quality questions with smaller, more deployable models for Brazil's national computing examination, POSCOMP [85]. An LLM-powered, RAG-based assessment system aimed specifically at higher education further illustrates the trend towards retrieval-grounded assessment tooling as a distinct application category rather than a byproduct of tutoring systems [86].

4.5. Learning Analytics and Monitoring of Self-Regulation

RAG's relevance to the monitoring phase of SRL is more indirect than its relevance to tutoring or assessment, operating primarily through its capacity to ground the generation of natural-language explanations and recommendations on top of existing learning-analytics infrastructure. Jivet et al.'s systematic review of learning-analytics dashboards from an SRL perspective found that most existing dashboards visualise activity data without adequately supporting the monitoring and strategy-adjustment functions SRL theory identifies as central, a gap that retrieval-grounded natural-language explanation could plausibly help close [61], and subsequent dashboard-design work has begun evaluating tools built explicitly around SRL-phase support [87]. Winne and Hadwin's [8] methodological argument for using fine-grained trace data, rather than retrospective self-report, to study SRL as it actually unfolds is directly relevant to RAG-based tutoring systems, which by their nature log detailed interaction traces that could in principle serve as both a monitoring signal for the learner and a retrieval context for the system itself, as in Cohn et al.'s log-contextualized

approach [51]. The MetaTutor project, an intelligent tutoring system built explicitly to scaffold SRL using multichannel learner data, provides a substantial evidence base for what phase-aware, monitoring-integrated tutoring can look like even where the underlying generative component is not itself RAG-based [62], and related work on the pedagogical-agent scaffolding of SRL within ITSs offers a design template that RAG-based conversational agents are beginning to adopt [88]. A recent meta-analysis quantifying the effects of intelligent tutoring systems on SRL outcomes provides the most direct existing evidence for how much of a difference phase-aware tutoring support can make, forming a benchmark against which RAG-specific systems can eventually be compared once a comparable body of controlled studies accumulates [89]. This benchmark should be read with caution: it is drawn from conventional, non-RAG intelligent tutoring systems, and because comparable controlled studies of RAG-based tutoring have not yet accumulated, any comparison between the two remains provisional rather than a like-for-like evaluation of RAG's own effectiveness [47,48].

4.6. Content Creation and Curriculum Grounding

RAG is also used to generate lesson materials, explanations, and curricular content drawn from an approved source corpus instead of whatever the model already holds in its weights. This overlaps with assessment-item generation but counts as a separate use case. A 2025 systematic review of technology-enabled personalised learning and SRL research argues that an adaptive content-generation system works only if the content itself is pedagogically sound; personalisation machinery does not make up for weak material. The point applies to RAG-based generation too [90]. The College English online teaching-resource platform illustrates this application concretely, using RAG to ground automatically assembled or generated language-learning materials in a curated instructional corpus rather than open-domain generation [63], and the broader survey literature on RAG chatbots and RAG applications in education situates curriculum-grounded content generation as one of several converging use cases alongside tutoring and assessment [46–48].

The quality of curriculum grounding depends on how the underlying corpus is assembled and maintained, not only on the retrieval architecture layered on top of it. A corpus built from an institution's own syllabi, approved textbooks, and assessment rubrics gives retrieval a curriculum-aligned anchor that a general-purpose or open-domain corpus cannot match [91], but it also concentrates the maintenance burden: outdated syllabi, unindexed revisions, or gaps between what a course actually covers and what the corpus contains surface directly in what the system retrieves and, downstream, in what a learner is shown. Domain-specific educational corpora therefore function as a curriculum-alignment mechanism as much as a technical retrieval resource, and their governance who updates them, on what schedule, against what curriculum-change process, is a design question the applications reviewed above have generally left implicit.

4.7. Supporting Diverse Learners and Inclusive Education

RAG's capacity to adapt both the content retrieved and the language in which it is presented gives it particular relevance to inclusive and accessible education, though this remains the least developed of the seven application areas reviewed here. Language-learning platforms that ground content generation in curated instructional material, such as the College English RAG platform, illustrate one pathway towards supporting learners studying in a second or additional language [63], while low-cost, mobile-first RAG architectures directly target learners whose access to conventional computing infrastructure is limited [64]. This potential must, however, be read alongside the documented risk that AI systems, including retrieval-grounded ones, can produce systematically different quality of support for dif-

ferent learner populations if the underlying retrieval corpus or generation model embeds cultural or linguistic assumptions that do not generalise; Stanford HAI's applied analysis of AI bias harms specifically for diverse student populations makes this caution explicit, and it applies with equal force whether the underlying generative system is retrieval-augmented or not [92].

Retrieval quality itself is not language-neutral. Dense retrievers and the embedding models beneath them are typically trained on English-dominant or high-resource-language corpora, and recent cross-lingual retrieval work finds that retrieval-based gains are asymmetric by direction: substantial when retrieving from a low-resource language into English, much weaker the other way round, so retrieval precision, and therefore grounding quality, cannot be assumed to transfer evenly across languages with less training data or a smaller indexed corpus [93]. A constraint that falls hardest on exactly the multilingual and low-resource educational settings where equitable access is most at stake. None of the applications reviewed in this section report evaluation in a genuinely low-resource language setting, which leaves an open question about whether the personalisation and grounding gains documented elsewhere in this review transfer to those contexts at all.

Table 3 is also intended to serve as the requested cross-system comparison: its System (Model & Retriever) column reports the generation model and retrieval/indexing approach used by each study, its SRL Phase and Comparator/Design columns report the target domain and evaluation design, and its Key Metric(s) & Result column reports the evaluation metrics and findings, so a separate table was not added to avoid duplicating the same studies twice.

Table 3. Study characteristics and evidence summary for key empirical and case-study systems discussed in Section 4 (NR = not reported in the source text).

Study/Year	System (Model + Retriever)	Corpus & Educational Context	Comparator/Study Design	SRL Phase	Key Metric(s) & Result	Limitation
Cohn et al. [51]	Log-contextualized RAG	Programming problem-solving discourse logs; secondary/high-school education (grades 9–10)	Discourse-only baseline (no environment logs); controlled comparison, 1080 LLM-as-judge comparisons across 5 embedding models	Forethought	Win rate vs. baseline (21.8–62.0% across models, LC-RAG ahead in 4 of 5); inter-rater QWK = 0.49; focus-group participants reported the agent supported problem-solving	No learning-outcome measures (process/discourse-quality metrics only); small sample (24 and 16 students across the two sub-studies)
Slade, Hyk and Gurung [66]	RAG tutor	Course material; introductory psychology	Absence-of-grounding comparison; pilot evaluation	Performance	Posttest quiz score: control M = 9.22, RAG tutor M = 10.81, GPT-4 Turbo M = 11.31; RAG vs. control significant ($p = 0.036$), GPT-4 vs. control significant ($p = 0.003$), RAG vs. GPT-4 not significant ($p = 0.283$)	Pilot scale (109 undergraduates); RAG tutor not distinguishable from an ungrounded GPT-4 Turbo comparator

Table 3. *Cont.*

Study/Year	System (Model + Retriever)	Corpus & Educational Context	Comparator/Study Design	SRL Phase	Key Metric(s) & Result	Limitation
PeteChat [53]	Guardrailed RAG tutor	Course syllabi, lecture notes, exams, discussion boards (ECE 20875, Purdue); undergraduate electrical/computer engineering	NR; case study	Performance	No learning-outcome metric reported (design case study); baseline message-coding reliability Cohen's kappa = 0.82 (284 messages, 31 conversations); 4-person expert evaluation	No controlled comparison; student deployment data not yet reported (later phases ongoing at time of writing)
Learner-state-aware RAG tutor [72]	State-aware RAG vs. prompt-only LLM	Curated algebra-tutoring corpus (six task types, e.g., simplifying expressions, linear equations, error identification); no real classroom deployment (24 simulated two-turn episodes)	Prompt-only LLM tutor baseline; 24 episodes rated by 8 expert reviewers (of 10 invited) across 5 dimensions on a 1–5 scale	Performance	Pedagogical-quality composite: 4.33 (baseline) vs. 4.65 (RAG), $p = 0.00024$, rank-biserial r = 0.86; but unsupported-content flags more frequent in the RAG condition (13/192 vs. 1/192 rating cells, $p = 0.0015$)	Scripted learners, no student-outcome data; confounded intervention (retrieval, learner-state tracking and prompting changed together); small 8-reviewer panel with ceiling effects and limited inter-rater reliability
MDKAG [54]	Multimodal disciplinary knowledge-graph RAG	Textbooks, slides, lecture recordings from three university courses (Digital Television, Costume History, Calligraphy Appreciation); upper-level university coursework across humanities and technical disciplines	Text-only RAG and knowledge-augmented generation (KAG) baselines; comparative evaluation with a 5-point answer-quality scoring system	Performance	Hallucination rate reduced by up to 23%; answer accuracy improved by 11% vs. text-only RAG/KAG baselines; entity-extraction F1 = 0.848 (22.4% improvement over a GPT-3.5 baseline)	Modest sample (300 samples across three domains); no participant demographics or significance testing reported; performance varies by domain
MCP/ACP protocol ITS [55]	RAG + Model Context Protocol + Agent Communication Protocol	Course materials from two undergraduate/graduate STEM courses (Virtual Instrumentation; Software for Telecommunications), Transilvania University; 842 and 1068 indexed text chunks respectively	Baseline dense retrieval and heuristic re-ranking pipelines; 240 annotated QA pairs plus human ratings from 36 students and 12 instructors, and a real deployment with 27 and 21 students	Performance	Citation fidelity improved from 0.68 to 0.86; unsupported numerical claims reduced from 12% to 3%; human factuality/pedagogical-appropriateness ratings ≈ 4.5–4.8/5	Single institution, two course corpora; visual retrieval relies on a caption/OCR fallback; production deployment would need stronger authentication and access controls

Table 3. *Cont.*

Study/Year	System (Model + Retriever)	Corpus & Educational Context	Comparator/Study Design	SRL Phase	Key Metric(s) & Result	Limitation
SP-TeachLLM [58]	Curriculum decomposition + RL strategy selection + reflective assessment	HumanEval (164 tasks) and MBPP (974 tasks) programming benchmarks; model fine-tuned on 100,000 programming QA pairs; programming education from basic syntax to advanced concepts	Five LLM tutors (e.g., GPT-4o, GPT-4, LLaMA-3-70B, Claude 3 Opus) paired with four simulated student models; ablation across five framework modules	Forethought/Performance	Code-generation accuracy gains for smaller student models up to 19.1% (MBPP); largest single-module contribution from strategy-selection removal (~6.1–6.2% drop)	Simulated LLM-based students only, no real classroom validation; the accuracy metric captures task outcomes, not the reasoning process behind them
Xu and Corter [59]	GraphRAG-based grading	Short Answer Feedback (SAF) dataset, 31 topics; automated short-answer grading with feedback (ASAG-F) for large-scale deployment	Controlled comparison against a vector-based RAG baseline and a fine-tuned LLM baseline, across multiple LLM sizes and prompting configurations	Self-reflection	Grading accuracy comparable to, and generally exceeding, the vector-based RAG and fine-tuned LLM baselines; more transparent source attribution to instructor-curated facts	Reported limitations are at the abstract level only (no enumerated methodological caveats in the source); cost/accuracy trade-offs vary by model size
POSCOMP question generation [85]	RAG-based MCQ generation, smaller models	Dataset of 1340 past POSCOMP questions (POSCOMP = Brazilian national computing-competency exam); Brazilian higher education	Few-shot baseline across five LLMs; 120 generated MCQs evaluated by human experts and an LLM-as-judge using a detailed rubric	Self-reflection	RAG approach improved question quality on up to half of the evaluated rubric criteria, including for smaller models	Per-model and per-criterion results not detailed in the publicly available abstract; full-text check recommended before final submission
MetaTutor [62]—benchmark, not RAG-based	Multichannel ITS (not RAG)	Hypermedia-based ITS teaching the human circulatory system to college students; over 10 years of research synthesized (log files, eye-tracking, facial-expression and think-aloud data)	Adaptive scaffolding vs. no scaffolding of self-regulatory behaviour across dozens of studies (individual study samples e.g., 117 or 174 undergraduates); mixed quantitative/qualitative methods	Performance/Self-reflection	Collaboratively set subgoals associated with higher proportional learning gains; students often over-confident and inaccurate in metacognitive judgments; boredom and persistent confusion predicted less learning	Predominantly Caucasian female college-student samples limit generalizability; inconsistent findings across studies due to differing methods/system versions; not RAG-based, so no directly comparable controlled RAG study yet

5. Challenges of RAG in Self-Regulated Personalised Learning

The literature reviewed here converges on six recurring problem areas, set out in Figure 4: ethical and fairness risks; data privacy and security; pedagogical appropriateness and effectiveness; hallucination, retrieval quality, and interpretability; infrastructure and implementation barriers; and academic integrity. Each is treated in turn below.

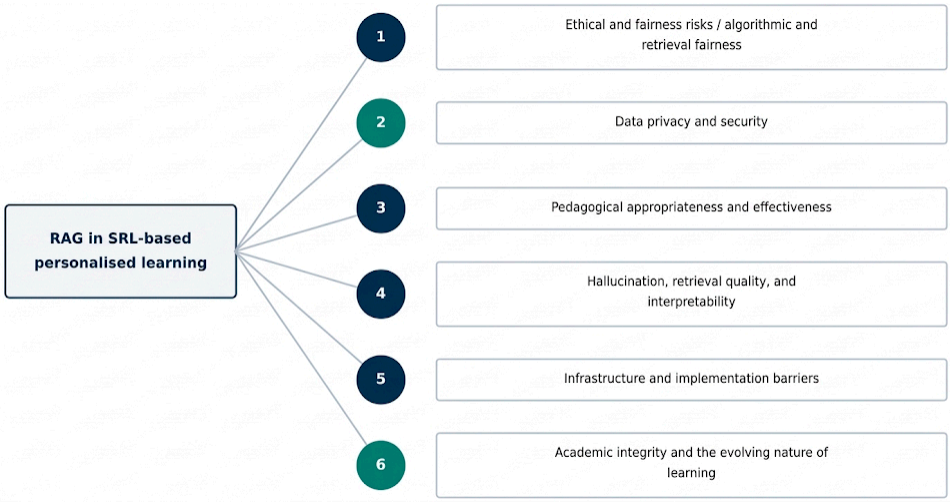


Figure 4. Principal challenges of RAG in self-regulated personalised learning: ethical and fairness risks; data privacy and security; pedagogical appropriateness and effectiveness; hallucination, retrieval quality, and interpretability; infrastructure and implementation barriers; and academic integrity.

5.1. Ethical Challenges and Algorithmic and Retrieval Fairness

RAG systems inherit the general fairness concerns documented for large language models, which Gallegos et al. catalogue across bias sources from training data through decoding strategy [6]. Retrieval adds a problem of its own: the corpus itself can be unrepresentative or out of date, even when the generative model is well aligned. Debiased LLMs have been shown to retain biased associations at an implicit level, so grounding generation in retrieved evidence does not settle the fairness question on its own; it moves the question of whose knowledge and whose framing reaches the learner rather than answering it [94]. In education, the cost of that bias falls hardest on students who are already underserved, which is what gives the problem its pedagogical weight [92].

5.2. Data Privacy and Security in Retrieval Systems

Because RAG systems retrieve from an external corpus at inference time, they introduce privacy and security risks distinct from those of purely parametric models. Bodea et al.'s [95] systematisation of privacy risks in RAG systems documents threats including membership inference against the retrieval corpus and embedding inversion, by which an attacker can partially reconstruct sensitive source text from its vector representation, both directly relevant where the retrieval corpus contains student work, assessment data, or other personally identifiable educational records. The OWASP Gen AI Security Project's treatment of vector and embedding weaknesses provides a complementary, practitioner-oriented taxonomy of the same attack surface [96]. These technical risks sit alongside long-standing legal obligations, such as the United States' Family Educational Rights and Privacy Act (FERPA), whose application to AI tools that process student data is still being actively clarified by institutions [97]. Privacy-preserving RAG design in education therefore depends on institutional governance as much as on the technical controls themselves. On the technical-countermeasure side, recent work on educational vector-database management combines per-record privacy-budget scheduling, which caps how much a given student's or document's embedding can be re-released or re-queried over time. This is paired with access-pattern-aware query routing that adds controlled randomness to index probing so that which records were retrieved is harder to infer from the access trace alone. On one educational benchmark, this reduced representation-leakage and access-pattern-leakage

rates by roughly a third relative to an unprotected baseline, while retrieval quality and latency both held up [98].

5.3. Pedagogical Appropriateness and Effectiveness

A retrieval-grounded answer is not automatically a pedagogically appropriate one, and a growing strand of the literature evaluates RAG-based educational tools specifically on this dimension rather than on retrieval accuracy alone. Slade, Hyk, and Gurung's pilot evaluation of a RAG tutor for introductory psychology directly tested whether grounding actually improved learning-relevant outcomes rather than assuming it would [66], and the expert-rated comparison of prompt-only versus learner-state-aware RAG tutoring found that pedagogical quality depended on more than the presence of retrieval alone, specifically on whether the system was also aware of and responsive to the learner's current state [72]. Systematic surveys of RAG in education converge on the observation that the evidence base for learning-outcome effectiveness remains thinner than the rapidly growing body of system-description papers, a methodological gap that mirrors the broader machine-learning-in-education literature's own evidence deficit [47,48].

5.4. Hallucination, Retrieval Quality, and Interpretability

RAG is frequently, and only partially accurately, described as "solving" hallucination. Huang et al.'s comprehensive taxonomy of LLM hallucination shows that retrieval augmentation mitigates but does not eliminate the problem, since a generator can still ignore, misread, or contradict correctly retrieved evidence [99], and benchmarks such as HaluEval provide the large-scale evaluation infrastructure needed to measure this residual risk empirically [100]. A focused review of hallucination mitigation specifically for retrieval-augmented LLMs catalogues the main technical countermeasures, including retrieval-evaluator gating, citation-grounded decoding, and post hoc verification [101], and applied work in structured-output settings demonstrates that narrowly scoped, high-quality retrieval corpora can substantially reduce hallucination rates in practice [102]. Retrieval quality itself is a second, closely related failure mode: Liu et al.'s finding that models underuse information placed in the middle of a long retrieved context [40] and the broader RAG evaluation literature's emphasis on faithfulness and context precision as first-class metrics [5,103,104] both indicate that a technically successful retrieval step does not guarantee a faithful generated answer. Interpretability compounds this concern: the dense embedding models underlying RAG retrieval are themselves largely opaque, and recent work on interpretable text embeddings [105] and on explainable knowledge-graph RAG that traces which retrieved facts actually drove a generated answer [106] represents an active but still immature research response to the resulting "why did the system say this" problem, which is especially consequential when a RAG tutor's explanation is itself the thing a self-regulating learner is meant to evaluate and act on. Entity linking has been proposed as a complementary, more lightweight remedy for a closely related failure mode, ambiguous or underspecified retrieval queries in educational platforms, where grounding query terms to disambiguated entities before retrieval has been shown to improve retrieval precision without the computational overhead of a full knowledge-graph pipeline [107].

5.5. Infrastructure and Implementation Barriers

Deploying RAG at institutional scale requires vector database infrastructure, embedding computation, and ongoing corpus curation that many educational institutions, particularly under-resourced ones, do not currently possess, a concern documented in recent evidence mapping of the generative-AI digital divide in higher education [108] and in policy-oriented analysis of AI's potential to widen rather than narrow existing educational

access gaps [109]. Lightweight and mobile-first RAG architectures answer the hardware side of this barrier [64]. The organisational side is harder: someone has to keep the corpus current and keep staff trained, and the institution has to fund both long after rollout. That is what decides whether a deployment stays pedagogically useful.

Cost and latency compound the hardware barrier rather than following automatically from it: embedding an institution's full corpus, running retrieval at query time, and re-embedding whenever the corpus changes all add computational overhead and, at API-billed scale, direct per-query cost on top of whatever the underlying language model already costs to run, and each additional retrieval hop adds latency a learner experiences directly as lag in a tutoring dialogue. Model maintenance is a related, ongoing cost rather than a one-time deployment expense: the generator, the retriever, and the corpus can each drift out of alignment with the others as any one of them is updated, so a RAG system that worked well at launch can degrade silently if maintenance is not budgeted as a recurring line item alongside initial deployment [110].

5.6. Academic Integrity and the Evolving Nature of Learning

Generative AI has already disrupted conventional academic-integrity practice, and RAG-based tools, precisely because they can produce fluent, source-grounded, and therefore more convincing output, intensify rather than resolve this disruption. Cotton, Cotton, and Shipway's early and widely cited analysis of academic integrity in the era of ChatGPT established the terms of this debate [111], and subsequent empirical work confirms that generative AI, including retrieval-grounded tools, continues to undermine assumptions built into authentic-assessment design [112]. Evidence on the unreliability of automated AI-text detection tools indicates that institutions cannot rely on detection as a primary response [113], a conclusion reinforced by systematic reviews calling instead for redesigned assessment and explicit institutional policy [114]. Broader systematic reviews of the ethical and regulatory landscape [115] and of responsible-AI risk mitigation in education [116] situate academic integrity as one instance of a wider question this review's other challenge sections also raise: whether the profession is designing RAG-based tools around genuine pedagogical goals, including the goal of preserving and strengthening learners' capacity for self-regulation, or around technical capability alone. That disruption extends well beyond plagiarism in the narrow sense of submitting text a student did not write. It includes over-reliance, where a learner defers to a RAG tutor's answer instead of attempting the reasoning step themselves, even when no policy is technically violated; unauthorised assistance during assessments that were designed assuming independent work; a broader erosion of assessment validity, where a grade increasingly measures a student's skill at prompting and evaluating a RAG system rather than the underlying competency the assessment was built to certify [117]; and, over time, an erosion of learners' capacity for independent work if scaffolded retrieval substitutes for practice rather than supporting it. These are distinct risks with distinct mitigations, and treating academic integrity as solely a plagiarism-detection problem understates what is actually at stake.

6. Discussion and Conclusions

Retrieval-Augmented Generation offers a technically coherent answer to a problem that predates it: how to make generative AI trustworthy enough, in the specific sense of being traceable to a verifiable source, for use in a domain where factual accuracy and pedagogical appropriateness both matter. The work reviewed here tests that promise across every phase of self-regulated learning, from goal setting to self-reflection. Most of the evidence sits in retrieval-grounded tutoring and assessment. Goal setting and monitoring remain thin, and in those phases, RAG works in the background rather than facing the

learner directly. The challenges in Section 5 bear directly on whether these systems end up supporting self-regulation or undermining it. A tutor that answers too readily, retrieves from a narrow corpus, cannot say which evidence it used, or logs interaction data without proper safeguards fails at more than a technical benchmark. It reproduces, in more fluent and more convincing form, the unscaffolded support that SRL theory predicts will fail the learners who need help most. That trade-off has a concrete operational form: the retrieval and re-ranking steps that make a RAG tutor's answers verifiable also add latency on top of generation itself. In real-time tutoring or conversational scaffolding, a system that is slow enough to disrupt the learner's flow of thought can undermine the same self-regulated engagement it is meant to support. Therefore, compute overhead and response latency are properly designed constraints on pedagogical responsiveness, not just infrastructure costs to be minimised separately.

Three implications follow for researchers and system designers working at this intersection. First, the evaluation of RAG-based educational tools should be organised explicitly around SRL-relevant outcomes, planning quality, monitoring accuracy, and strategy adaptation, not only around retrieval-faithfulness metrics borrowed from the general NLP literature, since a system can score well on the latter while still failing to support the former. Second, retrieval corpus design and curation deserve the same design attention currently given to prompt engineering and model selection, since the corpus, not only the model, determines what a learner is actually shown, and both the fairness and pedagogical-appropriateness challenges reviewed above trace back substantially to corpus composition. Third, given the demonstrated unreliability of both hallucination and of AI-text detection, institutions adopting RAG-based tools should treat retrieval-groundedness as a risk-reduction measure rather than a guarantee, maintaining human oversight and authentic-assessment redesign as complementary, not alternative, safeguards. Figure 5 draws these implications together as a deployment pipeline with governance checkpoints, spanning corpus curation and documentation, retrieval and generation configuration, learner-facing scaffolding design, the monitoring of SRL-relevant outcomes, and continuous review.

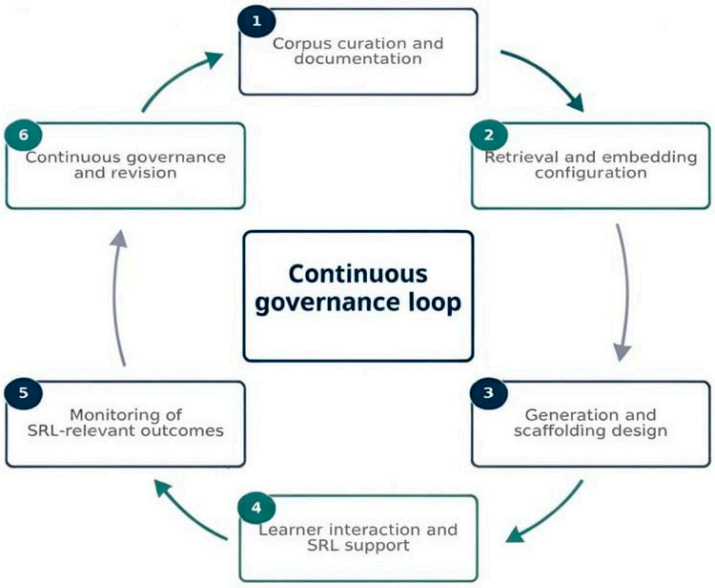


Figure 5. A RAG-in-SRL deployment pipeline with governance checkpoints, spanning corpus curation and documentation, retrieval and generation configuration, learner-facing scaffolding design, monitoring of SRL-relevant outcomes, and continuous review.

To be clear about what this review has and has not established, its limitations follow. This review has limitations characteristic of a narrative synthesis of a fast-moving field: a sub-

stantial share of the applications literature consists of preprints and conference papers not yet subject to full peer review, reflecting the genuine pace of the field rather than a deliberate lowering of evidentiary standards, and the concentration of controlled, outcome-focused evaluation studies remains small relative to the number of system-description papers.

The points above describe constraints on the evidence already reviewed; the following are proposed directions for work still to be done, and should not be read as additional findings. Future research should prioritise randomised or quasi-experimental comparisons of RAG-based and non-RAG-based educational tools on SRL-specific outcome measures, the longitudinal study of whether retrieval-grounded scaffolding builds or erodes learners' independent self-regulatory capacity over time, and the continued development of interpretability methods that make a RAG system's retrieved evidence, not only its final answer, visible and actionable to the learner. RAG is not a solved problem for education, but the trajectory traced in this review, from classical information retrieval, through neural dense retrieval, to increasingly SRL-aware application design, indicates a field converging on the recognition that grounding a generative system in evidence is necessary but not sufficient; grounding it in sound pedagogy is the harder and still-unfinished task.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/computers15110792/s1, Table S1: PRISMA 2020 Checklist.

Author Contributions: Conceptualization, I.O., L.M. and E.J.O.; methodology, I.O.; investigation, I.O.; writing—original draft preparation, I.O.; writing—review and editing, L.M. and E.J.O.; supervision, L.M. and E.J.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Haenlein, M.; Kaplan, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *Calif. Manag. Rev.* **2019**, *61*, 5–14. [CrossRef]
2. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
3. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997.
4. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the ICLR 2024, Vienna, Austria, 7–11 May 2024. Available online: https://arxiv.org/abs/2310.11511 (accessed on 18 September 2026).
5. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the EACL 2024 (System Demonstrations), St. Julians, Malta, 17–22 March 2024. [CrossRef]
6. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist.* **2024**, *50*, 1097–1179. [CrossRef]
7. Zimmerman, B.J. Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of Self-Regulation*; Boekaerts, M., Pintrich, P.R., Zeidner, M., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 13–39. [CrossRef]
8. Winne, P.H.; Hadwin, A.F. Studying as Self-Regulated Learning. In *Metacognition in Educational Theory and Practice*; Hacker, D.J., Dunlosky, J., Graesser, A.C., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 1998; pp. 277–304. Available online: https://www.researchgate.net/publication/247664651_Studying_as_Self-Regulated_Learning (accessed on 18 September 2026).
9. Pintrich, P.R. A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educ. Psychol. Rev.* **2004**, *16*, 385–407. [CrossRef]
10. Broadbent, J.; Poon, W.L. Self-Regulated Learning Strategies & Academic Achievement in Online Higher Education Learning Environments: A Systematic Review. *Internet High. Educ.* **2015**, *27*, 1–13. [CrossRef]

11. Panadero, E. A Review of Self-Regulated Learning: Six Models and Four Directions for Research. *Front. Psychol.* **2017**, *8*, 422. [CrossRef] [PubMed]
12. Wolters, C.A. Regulation of Motivation: Evaluating an Underemphasized Aspect of Self-Regulated Learning. *Educ. Psychol.* **2003**, *38*, 189–205. [CrossRef] [PubMed]
13. Devolder, A.; van Braak, J.; Tondeur, J. Supporting Self-Regulated Learning in Computer-Based Learning Environments: Systematic Review of Effects of Scaffolding in the Domain of Science Education. *J. Comput. Assist. Learn.* **2012**, *28*, 557–573. [CrossRef]
14. Araka, E.; Maina, E.; Gitonga, R.; Oboko, R. Self-Regulated Learning and Technology-Enhanced Learning Environments in Higher Education: A Scoping Review. *Sustainability* **2021**, *13*, 7281. [CrossRef]
15. Grand View Research. *AI in Education Market Size, Share and Growth Report, 2033*; Grand View Research: San Francisco, CA, USA, 2026. Available online: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-education-market-report (accessed on 18 September 2026).
16. Grand View Research. *Retrieval Augmented Generation Market Report, 2025–2030*; Grand View Research: San Francisco, CA, USA, 2025. Available online: https://www.grandviewresearch.com/industry-analysis/retrieval-augmented-generation-rag-market-report (accessed on 18 September 2026).
17. Baumeister, R.F.; Leary, M.R. Writing Narrative Literature Reviews. *Rev. Gen. Psychol.* **1997**, *1*, 311–320. [CrossRef]
18. Grant, M.J.; Booth, A. A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies. *Health Info. Libr. J.* **2009**, *26*, 91–108. [CrossRef] [PubMed]
19. 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.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
20. Bramer, W.M.; Rethlefsen, M.L.; Kleijnen, J.; Franco, O.H. Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study. *Syst. Rev.* **2017**, *6*, 245. [CrossRef] [PubMed]
21. Haddaway, N.R.; Collins, A.M.; Coughlin, D.; Kirk, S. The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching. *PLoS ONE* **2015**, *10*, e0138237. [CrossRef] [PubMed]
22. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
23. Salton, G.; Wong, A.; Yang, C.S. A Vector Space Model for Automatic Indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
24. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
25. Nwana, H.S. Intelligent Tutoring Systems: An Overview. *Artif. Intell. Rev.* **1990**, *4*, 251–277. [CrossRef]
26. VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
27. Alkhatlan, A.; Kalita, J.K. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *arXiv* **2018**, arXiv:1812.09628.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
29. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
30. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019. [CrossRef]
31. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the EMNLP 2020, Online, 16–20 November 2020. [CrossRef]
32. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. REALM: Retrieval-Augmented Language Model Pre-Training. In Proceedings of the ICML 2020, Online, 13–18 July 2020. Available online: https://arxiv.org/abs/2002.08909 (accessed on 18 September 2026).
33. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the SIGIR 2020, Online, 25–30 July 2020. [CrossRef]
34. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [CrossRef]
35. Malkov, Y.A.; Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 824–836. [CrossRef] [PubMed]
36. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
37. Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884.
38. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
39. Gao, Y.; Xiong, Y.; Wang, M.; Wang, H. Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks. *arXiv* **2024**, arXiv:2407.21059.

40. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
41. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024), Barcelona, Spain, 25–29 August 2024. [CrossRef]
42. Singh, A.; Ehtesham, A.; Kumar, S.; Khoei, T.T.; Vasilakos, A.V. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv* **2025**, arXiv:2501.09136.
43. Abootorabi, M.M.; Zobeiri, A.; Dehghani, M.; Mohammadkhani, M.; Mohammadi, B.; Ghahroodi, O.; Baghshah, M.S.; Asgari, E. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. In Proceedings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025. [CrossRef]
44. Brown, A.; Roman, M.; Devereux, B. A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *Big Data Cogn. Comput.* **2025**, *9*, 320. [CrossRef]
45. Levonian, Z.; Li, C.; Zhu, W.; Gade, A.; Henkel, O.; Postle, M.-E.; Xing, W. Retrieval-Augmented Generation to Improve Math Question-Answering: Trade-Offs between Groundedness and Human Preference. *arXiv* **2023**, arXiv:2310.03184.
46. Swacha, J.; Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Appl. Sci.* **2025**, *15*, 4234. [CrossRef]
47. Li, Z.X.; Wang, Z.J.; Wang, W.M.; Hung, K.; Xie, H.R.; Wang, F.L. Retrieval-Augmented Generation for Educational Application: A Systematic Survey. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100417. [CrossRef]
48. Arthana, R.I.K.; Gunantara, N.; Sudarma, M.; Sukarsa, M. A Systematic Literature Review of Retrieval-Augmented Generation Implementation for Enhancing Large Language Models in Education. *J. Nas. Pendidik. Tek. Inform. (JANAPATI)* **2025**, *15*, 91–109. [CrossRef]
49. Lai, J.W. Adapting Self-Regulated Learning in an Age of Generative Artificial Intelligence Chatbots. *Future Internet* **2024**, *16*, 218. [CrossRef]
50. Ng, D.T.K.; Tan, C.W.; Leung, J.K.L. Empowering Student Self-Regulated Learning and Science Education through ChatGPT: A Pioneering Pilot Study. *Br. J. Educ. Technol.* **2024**, *55*, 1328–1353. [CrossRef]
51. Cohn, C.; Rayala, S.; Snyder, C.; Fonteles, J.; Jain, S.; Mohammed, N.; Timalsina, U.; Burriss, S.K.; Ashwin, T.S.; Srivastava, N.; et al. Personalizing Student-Agent Interactions Using Log-Contextualized Retrieval-Augmented Generation (RAG). *arXiv* **2025**, arXiv:2505.17238.
52. Wu, X.Y.; Radloff, J.D.; Yeter, I.H.; Wang, L.; Chiu, T.K.F. Designing Artificial Intelligence Chatbots for Self-Regulated Learning from a Systematic Review Based on Habermas's Three Interests. *Interact. Learn. Environ.* **2025**, *34*, 3141–3164. [CrossRef]
53. Li, B.; Tan, L.; Zakharov, W.; Qiu, Q.; Acton, C.B. Tutor, Not Solver: Designing a Guardrailed AI Assistant for Learning in Higher Education—A Design Case of PeteChat. *arXiv* **2026**, arXiv:2606.09845.
54. Zhao, X.; Wang, G.; Lu, Y. MDKAG: Retrieval-Augmented Educational QA Powered by a Multimodal Disciplinary Knowledge Graph. *Appl. Sci.* **2025**, *15*, 9095. [CrossRef]
55. Modran, H.A. Leveraging RAG with ACP & MCP for Adaptive Intelligent Tutoring. *Appl. Sci.* **2025**, *15*, 11443. [CrossRef]
56. Rao, J.; Lin, J. RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations. *arXiv* **2024**, arXiv:2407.04925.
57. Ennaceur, I.; Elghazel, H.; Aussem, A.; Lefebvre, G.; Sonnati, M. CLARE: A Category-Aware RAG-Based Framework for Recommending Learning Objects in Education and Professional Training. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2026. [CrossRef]
58. Huang, S.; Sun, Y.; Yu, X. SP-TeachLLM: An LLM-Driven Framework for Personalized and Adaptive Programming Education. *Information* **2025**, *16*, 1045. [CrossRef]
59. Xu, G.; Corter, J. Graph RAG for Automated Short Answer Grading with Feedback: Bridging Pedagogical Needs and Technical Capabilities. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 40916–40924. [CrossRef]
60. Maity, S.; Deroy, A.; Sarkar, S. Leveraging In-Context Learning and Retrieval-Augmented Generation for Automatic Question Generation in Educational Domains. In Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE 2024), Gandhinagar, India, 12–15 December 2024. [CrossRef]
61. Matcha, W.; Ahmad Uzir, N.; Gašević, D.; Pardo, A. A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Trans. Learn. Technol.* **2019**, *13*, 226–245. [CrossRef]
62. Azevedo, R.; Bouchet, F.; Duffy, M.; Harley, J.; Taub, M.; Trevors, G.; Cloude, E.; Dever, D.; Wiedbusch, M.; Wortha, F.; et al. Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning with an Intelligent Tutoring System. *Front. Psychol.* **2022**, *13*, 813632. [CrossRef] [PubMed]
63. Dou, J.H.; Zhao, X.H. Design and Application of Online Teaching Resource Platform for College English Based on Retrieval-Augmented Generation. In Proceedings of the 2nd International Conference on Educational Knowledge and Informatization (ICEKI 2024), Shanghai, China, 19–21 July 2024. [CrossRef]

64. Bojorque, R.; Plaza, A.; Morquecho, P.; Moscoso, F. A Scalable and Low-Cost Mobile RAG Architecture for AI-Augmented Learning in Higher Education. *Appl. Sci.* **2026**, *16*, 963. [CrossRef]
65. Vu-Minh, A.; Nguyen-The, Q.; Nguyen, N.; Pham, X.-L.; Nguyen, A. Designing a Course-Grounded AI Tutor with Retrieval-Augmented Generation: A DSR Approach to Technical Education. 2025. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5607971 (accessed on 18 September 2026).
66. Slade, J.J.; Hyk, A.; Gurung, R.A.R. Transforming Learning: Assessing the Efficacy of a Retrieval-Augmented Generation System as a Tutor for Introductory Psychology. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2024**, *68*, 1827–1830. [CrossRef]
67. Cooper, M.M.; Klymkowsky, M.W. Let Us Not Squander the Affordances of LLMs for the Sake of Expedience: Using Retrieval Augmented Generative AI Chatbots to Support and Evaluate Student Reasoning. *J. Chem. Educ.* **2024**, *101*, 4847–4856. [CrossRef]
68. Mullins, E.; Portillo, A.; Ruiz Rohena, K.; Piplai, A. Enhancing Classroom Teaching with LLMs and RAG. In Proceedings of the 25th Annual Conference on Information Technology Education (SIGITE 2024), El Paso, TX, USA, 9–12 October 2024. [CrossRef]
69. Gala, A.; Marcondes, F.S.; Machado, J.; Novais, P. Designing a Local RAG-Based Intelligent Tutoring System for Domain-Specific Education. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2026. [CrossRef]
70. Ostrowska, A.; Kukla, M.; Majstrak, G.; Opala, J.; Pergała, S.; Skwarek, J.; Wróblewska, A. From Surface Learning to Deep Understanding: A Grounded AI Tutoring System for Moodle. *arXiv* **2026**, arXiv:2605.06963.
71. Tufino, E. NotebookLM as a Socratic Physics Tutor: Design and Preliminary Observations of a RAG-Based Tool. *arXiv* **2025**, arXiv:2504.09720.
72. Adenuga, J.A. Retrieval-Augmented Generation for Pedagogically Aware Educational AI: An Expert-Rated Comparison of a Prompt-Only LLM Tutor and an Integrated, Learner-State-Aware RAG Tutor. *Front. Educ.* **2026**, *11*, 1896839. [CrossRef]
73. Deng, C.; Yuan, B. Research on an Intelligent Tutoring System Based on Automatic Construction of Multimodal Knowledge Graphs and Retrieval-Augmented Generation. *Front. Comput. Sci.* **2026**, *8*, 1777749. [CrossRef]
74. Jain, M.; Bhatt, T.; Pitts, G.; Pandya, A.; Brusilovsky, P.; Norouzi, N.; Hellas, A.; Leinonen, J.; Akram, B. Retrieval-Augmented Tutoring for Algorithm Tracing and Problem-Solving in AI Education. *arXiv* **2026**, arXiv:2605.12988.
75. Retrieval-Augmented Chatbots for Scalable Educational Support in Higher Education. CEUR Workshop Proc. 2025, 3994, paper 3. Available online: https://ceur-ws.org/Vol-3994/paper3.pdf (accessed on 18 September 2026).
76. Lang, G.; Gürpinar, T. AI-Powered Learning Support: A Study of Retrieval-Augmented Generation (RAG) Chatbot Effectiveness in an Online Course. *Inf. Syst. Educ. J.* **2025**, *23*, 4–13. [CrossRef]
77. Optimization of a RAG Model Chatbot for Obstetric Medical Education. 2025. Available online: https://scholarworks.indianapolis.iu.edu/items/015f038b-b0df-49c9-b431-3bfd8ac508d4 (accessed on 18 September 2026).
78. Nguyen, V.D.; Tran, N.H.; Dao, K.A.; Nguyen, P.V. Towards Personalized AI Education: Context-Aware Retrieval-Augmented Generation with Grade-Level LLM Adaptation. *Comput. Appl. Eng. Educ.* **2026**, *34*, e70153. [CrossRef]
79. Raul, G.; Lin, Y.Z.; Patel, K.; Shih, P.J.; Redondo, M.W.; Saber Latibari, B. RAG-PRISM: A Personalized, Rapid, and Immersive Skill Mastery Framework with Adaptive Retrieval-Augmented Tutoring. *arXiv* **2025**, arXiv:2509.00646.
80. Sharma, K.; Nguyen, A.; Hong, Y. Self-Regulation and Shared Regulation in Collaborative Learning in Adaptive Digital Learning Environments: A Systematic Review of Empirical Studies. *Br. J. Educ. Technol.* **2024**, *55*, 1398–1436. [CrossRef]
81. Lian, Y. Machine Assistant with Reliable Knowledge: Enhancing Student Learning via RAG-Based Retrieval. *arXiv* **2025**, arXiv:2506.23026.
82. Khor, E.T.; K, M. A Systematic Review of the Role of Learning Analytics in Supporting Personalized Learning. *Educ. Sci.* **2024**, *14*, 51. [CrossRef]
83. Fateen, M.; Wang, B.; Mine, T. Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring with Feedback. *arXiv* **2024**, arXiv:2409.20042.
84. Pradeesh, N.; Remya, T.; Thushara, M.G.; Krishna, K.A.; Pranav, V. Retrieval-Augmented Generation for Multiple-Choice Questions and Answers Generation. *Procedia Comput. Sci.* **2025**, *259*, 504–511. [CrossRef]
85. da Silva Araujo Junior, J.R.; Marinho, L.B.; Campos, L.S.; Lima, K.N.d.S.; Pereira, D.E.; Cavalcanti, H.B.; Ramos, A.L.C.; de Araújo, E.C. Smarter Questions, Smaller Models: RAG-Enhanced Multiple-Choice Question Generation for POSCOMP. In Anais do Simósio Brasileiro de Informática na Educação (SBIE 2025). Available online: https://sol.sbc.org.br/index.php/sbie/article/view/38499 (accessed on 18 September 2026).
86. Vatankhah Barenji, R.; Salimi, N.; Khoshgoftar, S. An LLM-Powered Assessment Retrieval-Augmented Generation (RAG) for Higher Education. *arXiv* **2026**, arXiv:2601.06141.
87. de Vreugd, L.; van Leeuwen, A.; Jansen, R.; van der Schaaf, M. Learning Analytics Dashboard Design and Evaluation to Support Student Self-Regulation of Study Behaviour. *J. Learn. Anal.* **2024**, *11*, 249–262. [CrossRef]
88. Dever, D.A.; Sonnenfeld, N.A.; Wiedbusch, M.D.; Schmorrow, S.G.; Amon, M.J.; Azevedo, R. A Complex Systems Approach to Analyzing Pedagogical Agents' Scaffolding of Self-Regulated Learning within an Intelligent Tutoring System. *Metacognition Learn.* **2023**, *18*, 659–691. [CrossRef]

89. Zhang, K.; Zhou, Y.; Zhang, C.; Chen, J. The Effects of Intelligent Tutoring Systems on Self-Regulated Learning: A Meta-Analysis. *Educ. Psychol. Rev.* **2026**, *38*, 103. [CrossRef]
90. Halder Adhya, D.; Panda, S. An Integrative Systematic Review Analysis of Research on Technology-Enabled Personalized Learning and Self-Regulated Learning. *Smart Learn. Environ.* **2026**, *13*, 3. [CrossRef]
91. Fu, S.; Liu, F.; Wu, H.; Zhou, J.; Pan, Z. TeachPlanAlign: Dual-Profile Personalized and Curriculum-Grounded Lesson Plan Generation Schema via Retrieval-Augmented Fine-Tuning. *Mathematics* **2026**, *14*, 1492. [CrossRef]
92. Stanford Institute for Human-Centered Artificial Intelligence. *How Harmful Are AI's Biases on Diverse Student Populations?* Stanford HAI: Stanford, CA, USA, 2024. Available online: https://hai.stanford.edu/news/how-harmful-are-ais-biases-on-diverse-student-populations (accessed on 18 September 2026).
93. Park, G.; Park, J.; Lee, H. Cross-Lingual Summarization for Low-Resource Languages Using Multilingual Retrieval-Based In-Context Learning. *Appl. Sci.* **2025**, *15*, 7800. [CrossRef]
94. Bai, X.; Wang, A.; Sucholutsky, I.; Griffiths, T.L. Explicitly Unbiased Large Language Models Still Form Biased Associations. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2416228122. [CrossRef] [PubMed]
95. Bodea, A.-E.; Meisenbacher, S.; Klymenko, A.; Matthes, F. SoK: Privacy Risks and Mitigations in Retrieval-Augmented Generation Systems. *arXiv* **2026**, arXiv:2601.03979.
96. *OWASP Gen AI Security Project. LLM08:2025—Vector and Embedding Weaknesses*; OWASP Foundation: Wilmington, NC, USA, 2025. Available online: https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses/ (accessed on 18 September 2026).
97. University of Wisconsin–Madison Office of the Registrar. *FERPA and Artificial Intelligence (AI)*; University of Wisconsin–Madison: Madison, WI, USA, 2026. Available online: https://registrar.wisc.edu/ferpa-and-artificial-intelligence-ai/ (accessed on 18 September 2026).
98. Fu, S.; Liu, F.; Zhou, J.; Cai, J.; Pan, Z. PP-EDUVec: Privacy-Preserving Intelligent Management Algorithms for Educational-Corpus Vector Databases Under Retrieval-Augmented Learning. *Electronics* **2026**, *15*, 943. [CrossRef]
99. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
100. Li, J.; Cheng, X.; Zhao, W.X.; Nie, J.-Y.; Wen, J.-R. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Proceedings of the EMNLP 2023, Singapore, 6–10 December 2023. Available online: https://aclanthology.org/2023.emnlp-main.397/ (accessed on 18 September 2026).
101. Zhang, W.; Zhang, J. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics* **2025**, *13*, 856. [CrossRef]
102. Béchard, P.; Marquez Ayala, O. Reducing Hallucination in Structured Outputs via Retrieval-Augmented Generation. In Proceedings of the NAACL 2024 (Industry Track), Mexico City, Mexico, 16–21 June 2024. Available online: https://aclanthology.org/2024.naacl-industry.19/ (accessed on 18 September 2026).
103. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv* **2024**, arXiv:2405.07437.
104. Wu, S.; Xiong, Y.; Cui, Y.; Wu, H.; Chen, C.; Yuan, Y.; Huang, L.; Liu, X.; Kuo, T.W.; Guan, N.; et al. Retrieval-Augmented Generation for Natural Language Processing: A Survey. *arXiv* **2024**, arXiv:2407.13193.
105. Opitz, J.; Moeller, L.; Michail, A.; Padó, S.; Clematide, S. Interpretable Text Embeddings and Text Similarity Explanation: A Survey. *arXiv* **2025**, arXiv:2502.14862.
106. Moghaddam, Z.Z.S.; Dehghani, Z.; Rani, M.; Aslansefat, K.; Mishra, B.K.; Kureshi, R.R.; Thakker, D. Explainable Knowledge Graph Retrieval-Augmented Generation (KG-RAG) with KG-SMILE. *arXiv* **2025**, arXiv:2509.03626.
107. Granata, F.; Poggi, F.; Mongiovì, M. Enhancing Retrieval-Augmented Generation with Entity Linking for Educational Platforms. *Big Data Cogn. Comput.* **2026**, *10*, 120. [CrossRef]
108. Jamal Eddine, R.; Gide, E.; Al-Sabbagh, A. Systematised Evidence Mapping of Generative Artificial Intelligence (GenAI) and Digital Divide Phenomena in Higher Education. *Discov. Comput.* **2026**, *29*, 157. [CrossRef]
109. Brookings Institution. *AI and the Next Digital Divide in Education*; Brookings: Washington, DC, USA, 2024. Available online: https://www.brookings.edu/articles/ai-and-the-next-digital-divide-in-education/ (accessed on 18 September 2026).
110. Mishra, A.; Brahmanapally, N. A Comparative Performance Analysis of Locally Deployed Large Language Models Through a Retrieval-Augmented Generation Educational Assistant Application for Textual Data Extraction. *AI* **2025**, *6*, 119. [CrossRef]
111. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innov. Educ. Teach. Int.* **2024**, *61*, 228–239. [CrossRef]
112. Kofinas, A.K.; Tsay, C.H.-H.; Pike, D. The Impact of Generative AI on Academic Integrity of Authentic Assessments within a Higher Education Context. *Br. J. Educ. Technol.* **2025**, *56*, 2522–2549. [CrossRef]
113. Van Vlasselaer, M.; Van Droogenbroeck, F.; Spruyt, B. Who Wrote This? Evaluating the Reliability of AI Detection Tools in Higher Education. *Int. J. Educ. Integr.* **2026**, *22*, 16. [CrossRef]

114. Bittle, K.; El-Gayar, O. Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information* **2025**, *16*, 296. [CrossRef]
115. García-López, I.M.; Trujillo-Liñán, L. Ethical and Regulatory Challenges of Generative AI in Education: A Systematic Review. *Front. Educ.* **2025**, *10*, 1565938. [CrossRef]
116. Zhu, H.; Sun, Y.; Yang, J. Towards Responsible Artificial Intelligence in Education: A Systematic Review on Identifying and Mitigating Ethical Risks. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 1111. [CrossRef]
117. Brattli, H.; Utne, A.; Lynch, M. Assessment Validity in the Age of Generative AI: A Natural Experiment. *Informatics* **2026**, *13*, 56. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.