# A Retrieval-Augmented Generation Approach for Enhancing Pedagogical Support in Education

**Ramkumar N**
Coimbatore Institute of Technology
ramkumar@cit.edu.in

**Rajendran S**
Coimbatore Institute of Technology
asrajendrayadav@gmail.com

## Abstract

The manual creation of high-quality study materials such as summaries, question–answer sets, and flashcards remains a time-consuming task for both educators and learners. While large language models can generate such content rapidly, they frequently suffer from hallucinations and lack grounding in the source material. This paper presents a Retrieval-Augmented Generation (RAG) framework designed to address these limitations by transforming educational documents into pedagogically structured learning resources. The system integrates document preprocessing, semantic chunking, embedding-based retrieval, and a locally deployed language model to produce concise summaries, targeted questions, and flashcards. A cross-sectional evaluation involving 48 undergraduate students and 12 educators was conducted to assess usability, accuracy, cognitive support, and overall pedagogical value. Findings indicate that the AI Study Agent is perceived as supportive of comprehension and revision activities, and objective evaluation supports factual consistency relative to non-retrieval generation. Challenges related to output accuracy, oversimplification, and user trust remain consistent with limitations identified in existing GenAI literature. The study concludes that while the system demonstrates strong potential to support teaching and learning workflows, its perceived effectiveness relies on responsible use, educator oversight, and continued refinement to ensure pedagogical alignment and reliability.

## 1 Introduction

Generative Artificial Intelligence has catalyzed profound transformations across education. Large Language Models now power intelligent tutoring systems, automated assessment, and adaptive learning platforms, offering unprecedented scalability in content creation (Dwivedi et al., 2023; Kasneci et al., 2023). Within this ecosystem, automated creation of structured study materials (summaries, question–answer sets, and flashcards) represents a particularly promising frontier. These resources serve as foundational scaffolds for active recall and self-assessment, strategies empirically validated for durable learning (Dunlosky et al., 2013).

Manual production of such materials, however, remains labor-intensive and scales poorly with increasing class sizes. Pure generative models offer a tempting shortcut, yet introduce significant liabilities: ungrounded outputs, factual hallucinations, and lack of domain-specific fidelity can undermine the comprehension they aim to support (Ji et al., 2023). Existing commercial tools frequently lack pedagogical alignment, transparency regarding source provenance, and local deployment options essential for institutional data sovereignty. Many operate as opaque black boxes with limited configurability, hindering trustworthy integration into formal learning workflows. This paper addresses these challenges through the design, implementation, and empirical evaluation of a Retrieval-Augmented Generation (RAG) framework for pedagogical support. The contributions are fourfold: (1) a fully articulated RAG pipeline encompassing document ingestion, semantic chunking, embedding-based retrieval, and local LLM inference; (2) generation of structured pedagogical outputs—summaries, question–answer sets, and flashcards—aligned with cognitive load reduction and active learning principles; (3) cross-sectional empirical evaluation with students and educators capturing both objective metrics and subjective perceptions; and (4) insights on residual hallucination, oversimplification, and trust calibration that inform responsible AI deployment in education.

## 2 Related Work

Generative Artificial Intelligence has transformed educational content creation, with Large Language



*Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 28–39
November 9, 2026 ©2026 Association for Computational Linguistics

Models now powering tutoring systems, assessment generation, and personalized feedback (Kasneci et al., 2023; Dwivedi et al., 2023). Critical scholarship documents significant risks: hallucinations, erosion of epistemic vigilance, and the potential for generative outputs to supplant rather than scaffold knowledge construction (Ji et al., 2023; Sullivan et al., 2023). The persuasive surface validity of model outputs can induce unwarranted confidence among learners, particularly when domain expertise is insufficient to detect subtle distortions.

Research on automated study materials spans summarization (Allahyari et al., 2017; Nenkova and McKeown, 2011), question generation (Kurdi et al., 2020; Alsubait, 2015), and flashcard creation (Dunlosky et al., 2013; Karpicke and Roediger, 2008). Existing tools frequently operate as opaque, proprietary services that prioritize lexical fluency over epistemic accuracy, generating plausible but potentially distorted content (Chiu et al., 2023; Bender et al., 2021). Many lack explicit pedagogical alignment, offering limited configurability for chunking strategies, retrieval parameters, or prompt engineering essential for fine-tuning outputs to specific educational contexts.

Retrieval-Augmented Generation addresses hallucination by grounding outputs in verifiable source documents (Lewis et al., 2020; Guu et al., 2020). Educational applications of RAG have shown potential for improving the factual consistency and contextual grounding of generated learning materials, while recent work has also explored large language models for educational question generation (Lamsiyah et al., 2024). Existing work has generally focused on individual educational artifacts or tutoring interactions. Comparatively less attention has been given to the *unified* generation of summaries, question–answer sets, and flashcards within a single, locally deployable hybrid RAG framework that is evaluated by both learners and educators. This gap motivates the present work: an integrated system that prioritizes institutional data sovereignty, pedagogical alignment, and transparent source grounding.

## 3 Theoretical Framework

The following three theories motivate both the system architecture and the evaluation framework.

**Constructivist Learning Theory** positions AI-generated materials as scaffolds for active knowledge construction rather than inert knowledge repositories (Wood et al., 1976). The RAG pipeline operationalizes this through question–answer sets demanding active retrieval, flashcards cueing spaced rehearsal, and summaries preserving conceptual density for inferential reasoning. Source-grounded retrieval mitigates the risk of passive acceptance by positioning outputs as verifiable derivatives of authoritative texts.

**Cognitive Load Theory** distinguishes intrinsic, extraneous, and germane demands on working memory (Sweller, 1988). Structured summaries reduce extraneous load by eliminating redundancy and signaling hierarchical relationships; targeted questions channel resources toward germane processing. The chunking strategy and prompt engineering reflect this rationale, while the evaluation assesses whether perceived cognitive support aligns with theoretical predictions. Oversimplification and residual inaccuracies are treated as risks that may paradoxically increase load by forcing learners to reconstruct missing connections.

**Human–AI Interaction Theory** frames trust as a calibrated response shaped by reliability and transparency (Lee and See, 2004; Hoff and Bashir, 2015). Local deployment and source attribution support appropriate trust calibration, while the evaluation instruments measure perceived transparency and usability as distinct yet interdependent constructs. Educators are positioned as curators rather than passive consumers, retaining agency to validate and contextualize outputs.

## 4 System Design and Architecture

### 4.1 Overall Architecture

The system architecture implements a Retrieval-Augmented Generation pipeline that transforms raw educational documents into structured pedagogical artifacts through modular, sequential stages. Figure 1 illustrates the complete workflow. Document ingestion and preprocessing feed parallel retrieval streams: a dense semantic retriever employing sentence-transformer embeddings with cosine similarity, and a sparse lexical retriever employing BM25 scoring. The dual retrieval streams are fused via reciprocal rank fusion (RRF) to produce a consolidated context window of the top-$k$ most relevant chunks. This context, together with a task-specific prompt template, conditions the locally deployed large language model to generate one of three output types: narrative summaries, question–answer sets, or flashcards. Modularity ensures that

each functional component can be independently optimized, replaced, or audited without compromising broader system integrity. The pipeline operates entirely within local infrastructure, thereby eliminating external API dependencies and ensuring that sensitive educational materials remain within institutional boundaries.

### 4.2 Document Ingestion and Preprocessing

Document ingestion accommodates multiple formats including PDF, DOCX, and plain text, reflecting the heterogeneous document ecosystems typical of academic environments. Preprocessing routines perform structural normalization by removing headers, footers, and extraneous metadata while preserving semantic paragraph boundaries. Text chunking employs a hybrid strategy that combines recursive character splitting with semantic boundary detection; individual chunks are constrained to approximately 512 tokens with 64-token overlaps. This chunk size represents a deliberate compromise between retrieval granularity and contextual coherence. Smaller fragments improve precision yet risk severing conceptual relationships that span multiple sentences, whereas excessively large chunks dilute semantic specificity and introduce irrelevant context into generation prompts. Overlap mechanisms mitigate boundary effects by ensuring that transitional passages appear in adjacent chunks, thereby preserving thematic continuity across segment boundaries.

### 4.3 Embedding and Vector Storage

Chunk embeddings are computed using the `sentence-transformers/all-MiniLM-L6-v2` model, a 22.7M-parameter sentence-transformer that produces 384-dimensional dense vectors via mean-pooling over the final hidden states (Reimers and Gurevych, 2019). Input preprocessing truncates sequences to a maximum of 256 word-piece tokens and applies lowercase normalization and whitespace collapsing. No additional domain-specific fine-tuning is performed. Vector similarity is computed using cosine distance in the $L_2$-normalized embedding space.

Dense vector representations are stored in ChromaDB 0.5.0 with HNSW indexing (efConstruction = 128, M = 16), enabling approximate nearest-neighbor search with persistent local storage (Team, 2022). The selection of ChromaDB reflects a design priority for lightweight deployment without container orchestration overhead.

Table 1: Chunk Size and Overlap Ablation ($n = 30$ documents)

| Configuration | Precision@3 | Recall@5 | ROUGE-L F1 | Factual Consistency |
|---|---|---|---|---|
| $c = 256, o = 32$ | 0.71 | 0.64 | 0.62 | 3.78 |
| $c = 256, o = 64$ | 0.73 | 0.66 | 0.63 | 3.82 |
| $c = 256, o = 128$ | 0.72 | 0.65 | 0.61 | 3.75 |
| $c = 512, o = 32$ | 0.76 | 0.69 | 0.68 | 4.05 |
| $c = 512, o = 64$ | **0.78** | **0.71** | **0.71** | **4.12** |
| $c = 512, o = 128$ | 0.77 | 0.70 | 0.69 | 4.08 |
| $c = 1024, o = 64$ | 0.72 | 0.67 | 0.65 | 3.91 |
| $c = 1024, o = 128$ | 0.71 | 0.66 | 0.64 | 3.88 |
| $c = 1024, o = 256$ | 0.70 | 0.65 | 0.62 | 3.84 |

The choice of chunk size ($c = 512$ tokens) and overlap ($o = 64$ tokens) was validated through an ablation study rather than selected heuristically. Table 1 reports retrieval and generation performance across nine configurations on a held-out validation set of 30 documents. Retrieval quality is measured by Precision@3 and Recall@5 against manually annotated relevance judgments. Generation quality is measured by ROUGE-L F1 and expert-rated factual consistency (1–5 scale) for summarization outputs.

The configuration $c = 512$, $o = 64$ achieved the optimal balance: smaller fragments ($c = 256$) improved retrieval precision marginally but degraded generation coherence due to severed conceptual relationships, while larger chunks ($c = 1024$) diluted semantic specificity and introduced irrelevant context into generation prompts. The 64-token overlap was sufficient to preserve boundary continuity without excessive storage redundancy. This configuration was fixed for all subsequent experiments.

### 4.4 Retrieval and Generation

Dense retrieval identifies the top-$m = 10$ candidates via cosine similarity over the 384-dimensional embedding space, capturing paraphrastic and conceptual relationships that elude exact term matching. Sparse retrieval identifies the top-$m = 10$ candidates via BM25 with default parameters ($k_1 = 1.5$, $b = 0.75$), ensuring precise recovery of rare terminological and definitional content that dense representations may dilute. The two ranked lists are fused through reciprocal rank fusion with constant $k_{\text{RRF}} = 60$:

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k_{\text{RRF}} + \text{rank}_r(d)}$$

where $R$ denotes the set of retrievers and $\text{rank}_r(d)$ the rank of document $d$ in retriever $r$. The fused list is truncated to $k = 5$ chunks, which

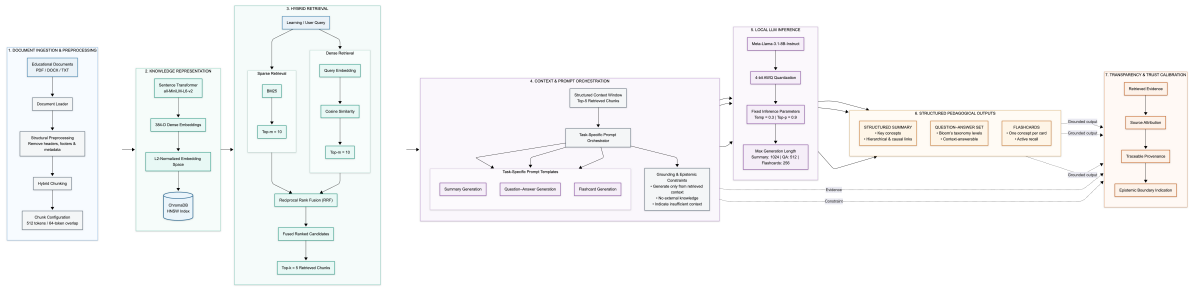


Figure 1: System Architecture. End-to-end pipeline from document ingestion to grounded pedagogical outputs (summaries, QA pairs, and flashcards), with hybrid retrieval, local LLM inference, and explicit transparency mechanisms.

are concatenated into a structured context window. This value of $k = 5$ was selected based on the context-window constraints of the generation model and the empirical observation that additional chunks beyond the fifth contributed diminishing relevance while increasing the risk of context dilution. The Full System significantly outperforms both unimodal retrieval baselines and the hybrid configuration without RRF, confirming that the fusion mechanism captures complementary signals rather than merely aggregating redundant ones.

Retrieved chunks are injected into task-specific prompt templates that constitute a secondary contribution. Three distinct templates govern output generation, each engineered to align with specific pedagogical objectives:

- *Summarization*: instructs condensation while preserving hierarchical concept relationships and key terminology, with explicit constraints against omission of causal or comparative relationships;
- *Question generation*: demands varied cognitive levels ranging from factual recall to analytical inference, with explicit constraints on answerability from the provided context and prohibitions against tautological or overly generic items;
- *Flashcard generation*: enforces concise question-answer pairs suitable for spaced repetition, with a maximum of 128 tokens per card face and requirements that each pair target a single, verifiable concept.

All templates include explicit grounding instructions that constrain the model to synthesize responses exclusively from retrieved context and to indicate when a request cannot be answered from the provided material. This design directly operationalizes the theoretical commitment to transparency: by mandating that the generator acknowledge epistemic boundaries, the system mitigates the risk of confabulation that persists even under retrieval augmentation.

Generation is performed by `meta-llama/Llama-3.1-8B-Instruct`, deployed locally via 4-bit AWQ quantization (group size 128, zero point quantization). Inference parameters are fixed across all tasks: temperature $= 0.3$, top-$p$ $= 0.9$, repetition penalty $= 1.1$, maximum generation length $= 1024$ tokens for summaries, $512$ tokens for question-answer sets, and $256$ tokens for flashcards. The low temperature prioritizes factual consistency and structural adherence over lexical diversity, an appropriate trade-off for educational content where precision supersedes creativity. Local deployment constitutes a core architectural decision motivated by three interlocking considerations: institutional data sovereignty, which precludes transmission of curricular materials to third-party APIs; elimination of network latency and external service dependencies, which ensures reliable classroom deployment; and mitigation of vendor lock-in, which preserves institutional autonomy over pedagogical infrastructure. These considerations reflect the theoretical mandate from Human-AI Interaction Theory that trust and transparency depend not merely on interface design but on underlying architectural control.

### 4.5 Output Types

The system produces three structured output types through a unified retrieval-augmented pipeline, ensuring consistent source grounding while accommodating distinct pedagogical functions.

**Summaries** condense source documents into

narrative overviews that preserve core arguments and terminological precision. The summarization template instructs the model to maintain hierarchical concept relationships and causal links, avoiding the aggressive compression that strips contextual qualifiers necessary for advanced understanding.

**Question–answer sets** comprise paired items spanning multiple cognitive levels, from definitional recall to application and synthesis. The generation template demands varied cognitive demands aligned with a modified Bloom's taxonomy, with explicit constraints prohibiting tautological items or questions unanswerable from the provided context. These sets support both formative self-assessment and instructor-led review.

**Flashcards** adhere to a strict question-answer format optimized for active recall and spaced repetition protocols. Each card targets a single verifiable concept with a maximum of 128 tokens per face, ensuring concision suitable for rapid review cycles.

The tripartite design reflects the theoretical commitment to reducing extraneous cognitive load while supporting active knowledge construction.

# 5 Methodology

## 5.1 Participants and Procedure

A cross-sectional design was employed to capture user interactions and perceptions at a single temporal point. The sample comprised $N = 60$ participants purposively recruited from a single higher-education institution offering programmes in artificial intelligence and data science: $n = 48$ undergraduate students (24 male, 22 female, 2 non-binary; $M_{\text{age}} = 20.7$, $SD = 1.4$) distributed across second-year ($n = 16$), third-year ($n = 20$), and final-year ($n = 12$) study levels, and $n = 12$ educators (7 male, 5 female; $M_{\text{age}} = 38.4$, $SD = 6.2$) with 2 to 19 years of teaching experience ($M = 9.3$, $SD = 5.1$). All students were enrolled in or had recently completed a Fundamentals of Artificial Intelligence module, ensuring sufficient domain familiarity to evaluate output accuracy and pedagogical relevance. Inclusion required functional English proficiency and basic digital literacy; exclusion comprised prior exposure to the prototype or involvement in its development. Students reported moderate prior GenAI familiarity ($M = 3.1$, $SD = 1.0$ on a 5-point scale); educators reported lower familiarity ($M = 2.4$, $SD = 1.1$), with only 25% having integrated generative AI into formal instruction.

Participants completed a single 45–60 minute session with the prototype loaded with Fundamentals of AI materials. Students performed document upload, summary generation, question–answer set generation, flashcard creation, and output review. Educators performed identical interactions and additionally evaluated outputs against professional standards for pedagogical alignment, terminological precision, and classroom suitability. Task completion triggered administration of the structured questionnaire. All interactions and temporal markers were captured via automated logs.

## 5.2 Measures and Analysis

Data comprised interaction logs, a structured questionnaire, and expert artifact evaluations. Logs recorded uploads, generation requests, refinement iterations, and session duration. The questionnaire operationalized three theoretical constructs across four domains—Perceived Usefulness, Accuracy and Reliability, Perceived Cognitive Support, and Usability—with 16 five-point Likert items (1 = Strongly Disagree, 5 = Strongly Agree). Pilot testing ($n = 15$) confirmed internal consistency ($\alpha \geq .87$ for all domains). Complete items appear in Appendix A.

Objective artifact quality was evaluated by two domain experts. Summaries ($n = 40$) were rated on factual consistency (claim-level support against the source), completeness (retention of key concepts), relevance (alignment with inferred information need), and compression ratio (summary word count / source word count). Question–answer sets ($n = 120$) were assessed on correctness, answerability, relevance, and cognitive-level validity (match to intended Bloom's taxonomy level). Flashcards ($n = 80$) were assessed on factual correctness, pedagogical usefulness, and coverage (proportion of key terms represented). Inter-rater reliability was substantial ($\kappa \geq .82$).

Ablation analysis compared five system variants on a held-out corpus of 50 document sections: Vanilla LLM, Dense-only RAG, Sparse-only RAG, Hybrid without RRF, and the Full System. Each of the 50 document sections was processed independently by all five system variants, yielding repeated measurements for each document and enabling within-document comparisons. Two experts rated factuality on a 1–5 scale. Metrics included citation precision (verifiable claims against retrieved context; against full source for Vanilla LLM), hallucination rate (unsupported plus contra-

Table 2: Retrieval Performance Metrics Across System Configurations

| Metric | Dense-only | Sparse-only | Hybrid (no RRF) | Full System |
|---|---|---|---|---|
| Precision@3 | 0.62 | 0.58 | 0.71 | **0.78** |
| Recall@5 | 0.54 | 0.51 | 0.63 | **0.71** |
| MRR | 0.68 | 0.64 | 0.74 | **0.82** |
| nDCG@5 | 0.59 | 0.56 | 0.67 | **0.76** |

dicted claims), and ROUGE-L F1 against human-written reference summaries prepared by graduate teaching assistants instructed to condense source documents to approximately 30% length while preserving key concepts. ROUGE-L served as a lexical overlap metric, distinct from expert-rated factual consistency. Quantitative analysis employed descriptive statistics and independent-samples $t$-tests for subgroup comparisons. Effect sizes are reported as Cohen's $d$ with 95% confidence intervals. Optional open-text responses were analysed thematically to identify recurrent patterns (Braun and Clarke, 2006). Three evaluation subsets were used. A 30-document validation set was used for chunk-size/overlap selection, a manually annotated retrieval set was used to evaluate retrieval effectiveness, and a separate 50-document held-out corpus was used for the five-way system ablation.

## 6 Results

### 6.1 System Performance

Retrieval effectiveness was evaluated against 150 manually annotated query–chunk relevance judgments ($\kappa = .82$). The Full System achieved Precision@3 of 0.78, Recall@5 of 0.71, MRR of 0.82, and nDCG@5 of 0.76 (Table 2), outperforming both unimodal baselines and the hybrid configuration without RRF. The MRR gain of 0.08 attributable to RRF indicates that fusion elevates the most relevant chunks beyond the simple concatenation of retrieval streams.

Generated artifacts demonstrated strong overall quality across the three output types. Summaries ($n = 40$) achieved mean factual consistency of 4.12 ($SD = 0.68$), completeness of 3.85 ($SD = 0.74$), and relevance of 4.05 ($SD = 0.61$), with a mean compression ratio of 0.32 ($SD = 0.08$) (Table 3). The completeness score was marginally lower than consistency, reflecting occasional omission of secondary concepts during condensation.

Question–answer sets ($n = 120$) demonstrated strong relevance ($M = 4.21$, $SD = 0.58$) and answerability ($M = 4.08$, $SD = 0.66$), with lower correctness ($M = 3.94$, $SD = 0.79$) (Table 4). Discrepancies concentrated in items requiring inferential reasoning beyond explicit source statements. Cognitive-level validity reached 0.74 ($SD = 0.18$); misclassifications predominantly involved remember-level items generated when understand-level items were solicited.

Table 3: Objective Evaluation of Generated Summaries ($n = 40$)

| Criterion | Scale | $M$ | $SD$ |
|---|---|---|---|
| Factual Consistency | 1–5 | 4.12 | 0.68 |
| Completeness | 1–5 | 3.85 | 0.74 |
| Relevance | 1–5 | 4.05 | 0.61 |
| Compression Ratio | Proportion | 0.32 | 0.08 |

Table 4: Objective Evaluation of Generated Question–Answer Sets ($n = 120$)

| Criterion | Scale | $M$ | $SD$ |
|---|---|---|---|
| Correctness | 1–5 | 3.94 | 0.79 |
| Answerability | 1–5 | 4.08 | 0.66 |
| Relevance | 1–5 | 4.21 | 0.58 |
| Cognitive-Level Validity | Proportion correct | 0.74 | 0.18 |

Flashcards ($n = 80$) achieved factual correctness of 4.05 ($SD = 0.71$) and pedagogical usefulness of 3.88 ($SD = 0.82$), with coverage of 0.68 ($SD = 0.14$) (Table 5). The majority of key terms were represented, though specialized terminology in dense conceptual clusters was occasionally overlooked. Interaction logs recorded mean generation latency of 2.3 s for summaries, 1.1 s for question–answer sets, and 0.8 s for flashcards. Refinement requests constituted 12.4% of total interactions.

Participants reported an estimated 68.1% reduction in preparation time relative to *recalled* manual baselines ($M = 47.3$ minutes vs. $M = 15.1$ minutes, 95% CI $[65.7\%, 70.5\%]$). These figures reflect self-reported estimates rather than observed performance; no controlled manual baseline condition was administered. Interaction logs corroborated the AI-assisted duration ($M = 13.8$ minutes, $SD = 3.9$). We therefore interpret the reported time reduction as an indication of perceived efficiency gains contingent on successful workflow integration, not as causal evidence of productivity improvement.

Table 6 reports the distribution of individual reduction ratios.

### 6.2 User Perceptions

The final sample comprised $N = 60$ participants: $n = 48$ students and $n = 12$ educators. The instrument demonstrated strong internal consistency

Table 5: Objective Evaluation of Generated Flashcards ($n = 80$)

| Criterion | Scale | $M$ | $SD$ |
|---|---|---|---|
| Factual Correctness | 1–5 | 4.05 | 0.71 |
| Pedagogical Usefulness | 1–5 | 3.88 | 0.82 |
| Coverage | Proportion | 0.68 | 0.14 |

Table 6: Distribution of Self-Reported Time Reduction Ratios ($N = 60$)

| Statistic | Value |
|---|---|
| Minimum | 45.2% |
| 25th Percentile | 62.1% |
| Median | 69.3% |
| 75th Percentile | 74.5% |
| Maximum | 84.6% |
| Mean ($SD$) | 68.1% (9.4%) |

($\alpha \geq .87$ for all domains).

Students reported significantly higher perceived usefulness, $t(58) = 3.02$, $p = .003$, $d = 0.97$ [0.32, 1.61], and perceived cognitive support, $t(58) = 2.62$, $p = .010$, $d = 0.84$ [0.20, 1.47], than educators. No significant group difference emerged for perceived accuracy and reliability, $t(58) = 1.15$, $p = .255$, $d = 0.37$ [−0.27, 1.00], or usability, $t(58) = 1.89$, $p = .064$, $d = 0.61$ [−0.03, 1.24]. Both groups identified summarization as the most beneficial feature.

### 6.3 Ablation Analysis

Table 7 reports performance across the five system variants. A one-way repeated-measures ANOVA documented significant condition effects for factuality, $F(4, 196) = 24.37$, $p < .001$, $\eta_p^2 = .33$; citation precision, $F(4, 196) = 41.52$, $p < .001$, $\eta_p^2 = .46$; hallucination rate, $F(4, 196) = 28.91$, $p < .001$, $\eta_p^2 = .37$; and ROUGE-L F1, $F(4, 196) = 19.84$, $p < .001$, $\eta_p^2 = .29$.

Bonferroni-corrected comparisons revealed that the Full System significantly outperformed the Vanilla LLM on all metrics (all $p < .001$, $d$ = 1.78–2.34). It also surpassed both Dense-only and Sparse-only configurations (all $p \leq .003$, $d$ = 0.82–1.21), confirming that hybrid retrieval captures complementary signals. Most critically, the Full System exceeded the Hybrid without RRF variant on factuality, $t(49) = 2.72$, $p = .009$, $d = 0.62$, and citation precision, $t(49) = 3.14$, $p = .003$, $d = 0.78$, demonstrating that reciprocal rank fusion contributes measurably beyond simple concatenation.

Table 7: Ablation Study Results Across System Variants ($N = 50$ documents)

| Variant | Factuality $M$ ($SD$) | Citation Precision $M$ ($SD$) | Hallucination Rate $M$ ($SD$) | ROUGE-L F1 $M$ ($SD$) |
|---|---|---|---|---|
| Vanilla LLM | 2.84 (0.91) | 0.12 (0.08) | 0.34 (0.11) | 0.42 (0.10) |
| Dense-only RAG | 3.62 (0.72) | 0.68 (0.12) | 0.18 (0.09) | 0.58 (0.09) |
| Sparse-only RAG | 3.44 (0.81) | 0.64 (0.14) | 0.21 (0.10) | 0.55 (0.11) |
| Hybrid without RRF | 3.88 (0.64) | 0.74 (0.10) | 0.14 (0.07) | 0.63 (0.08) |
| **Full System (Hybrid + RRF)** | **4.28 (0.52)** | **0.86 (0.08)** | **0.08 (0.06)** | **0.71 (0.07)** |

## 7 Discussion

### 7.1 Interpretation of Findings

The ablation study confirms that hybrid retrieval with reciprocal rank fusion yields measurable gains in factual consistency beyond both non-retrieval generation and unimodal retrieval baselines. The Full System achieved a hallucination rate of 0.08, representing a 76.5% reduction from the Vanilla LLM baseline, with disproportionate attenuation of unsupported claims. These improvements are attributable to the architecture itself. Dense retrieval excels at capturing paraphrastic and conceptual relationships, while sparse retrieval preserves precision for rare terminological and definitional content. Reciprocal rank fusion exploits the complementary strengths of both modalities, re-ranking candidates to elevate the most relevant chunks beyond what either stream achieves in isolation. The divergence between objective factual consistency and subjective trust ratings merits attention. The objective hallucination rate of 0.08 contrasts with perceived accuracy ratings of 3.45 on a five-point scale, indicating that participants maintained appropriate epistemic caution despite retrieval grounding. This pattern aligns with Human-AI Interaction Theory: transparency mechanisms, including source attribution and traceable provenance, enable users to calibrate trust without surrendering critical judgment.

### 7.2 Pedagogical Implications

Participant perceptions support the theoretical framing of the system as a scaffold rather than an autonomous instructional authority. Students reported significantly higher perceived usefulness and cognitive support than educators, a pattern consistent with Constructivist Learning Theory and its emphasis on learner agency. For students, structured sum-

maries and question sets functioned as temporary supports that reduced extraneous cognitive load and freed working memory resources for schema construction. Educators, by contrast, exhibited greater circumspection regarding accuracy and pedagogical alignment, reflecting their institutional responsibility for curriculum fidelity. This tension underscores that AI-generated materials facilitate active engagement only when deployed under human oversight. Self-reported time estimates indicated substantial perceived reduction in preparation time, though the absence of a controlled manual baseline precludes causal inference. The system should be understood as offering potential for reduced preparation effort contingent on workflow integration, not as empirically demonstrated productivity enhancement.

## 8 Limitations

The cross-sectional design captures perceptions at a single point in time and therefore cannot assess longitudinal effects such as sustained trust calibration or actual changes in knowledge retention. Claims regarding comprehension benefits and cognitive support consequently reflect participant perception rather than measured learning gains. The evaluation was restricted to a single higher-education institution and a single course domain. Performance may differ in disciplines with distinct epistemological structures or denser technical terminology. The sample size ($N = 60$) further limits generalizability. Despite retrieval grounding, residual hallucination and oversimplification of nuanced concepts persist. Educators consistently emphasized the need for human oversight. These limitations reinforce that the system is best deployed as a pedagogical scaffold under educator supervision rather than as an autonomous instructional authority.

## 9 Conclusion

This study presented a locally deployable Retrieval-Augmented Generation framework for generating structured educational materials, including summaries, question–answer sets, and flashcards. The architecture integrates hybrid dense–sparse retrieval with reciprocal rank fusion, local quantized inference, and task-specific prompt engineering. The 76.5% reduction in hallucination rate, driven primarily by attenuation of unsupported claims, provides direct empirical support for the central architectural motivation. A cross-sectional evaluation with 60 participants yielded converging evidence from objective system metrics and subjective user perceptions. Students reported significantly higher perceived usefulness and cognitive support than educators, while both cohorts shared comparable levels of perceived accuracy and trust calibration. Residual challenges, including occasional inaccuracies, oversimplification of complex topics, and persistent trust calibration concerns, reinforce that the system functions most effectively as a pedagogical scaffold under educator oversight rather than as an autonomous instructional authority. Future work should prioritize longitudinal experimental designs with pre/post knowledge assessments and delayed retention measures to establish whether perceived benefits translate into measurable learning outcomes. Multi-domain evaluation is essential to determine generalizability beyond artificial intelligence education. Technical priorities include improved hallucination mitigation through chain-of-verification generation, confidence indicators, and expanded personalization features that accommodate heterogeneous learning levels. Hybrid human–AI curation workflows, wherein educators actively validate and contextualize generated materials, represent the most promising direction for ensuring pedagogical alignment and ethical soundness in AI-supported learning environments.

## AI Usage Statement

The authors used generative AI only for auxiliary writing support, including grammar correction, and wording refinement. The technical contributions were developed and validated by the authors. The authors take full responsibility for the content of the paper.

## Ethics Statement

This study involved human participants (students and educators) who interacted with the system and completed questionnaires. All participants provided informed consent prior to participation. No personally identifiable information was collected, and all interaction logs were anonymized. The educational materials used in the evaluation were institutionally approved course content. The system is designed as a pedagogical support tool and is not intended to replace human educators. We emphasize that generated outputs should always be reviewed by instructors before classroom use.

## References


Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Tahani Alsubait. 2015. A review of ontology-based question generation techniques. *Journal of King Saud University-Computer and Information Sciences*, 27(2):145–153.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Thomas K. F. Chiu, Qing Xia, Xiaoqin Zhou, Ching Sing Chai, and Miao Cheng. 2023. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4:100118.

John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1):4–58.

Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, Emma L. Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71:102642.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Kevin A. Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jeffrey D. Karpicke and Henry L. Roediger. 2008. The critical importance of retrieval for learning. *Science*, 319(5865):966–968.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Darya Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Gihan Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Salima Lamsiyah, Abdelkader El Mahdaouy, Aria Nourbakhsh, and Christoph Schommer. 2024. Fine-tuning a large language model with reinforcement learning for educational question generation. In *Artificial Intelligence in Education*, pages 424–438. Springer Nature Switzerland.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Michael Sullivan, Ann Kelly, and Peter McLaughlan. 2023. Chatgpt in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, 6(1):31–37.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Chroma Team. 2022. Chroma: The ai-native open-source embedding database. `https://github.com/chroma-core/chroma`. Accessed: 2024.

David Wood, Jerome S. Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

# A Questionnaire Items

Table 8 presents the 16 questionnaire items. All items employ a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

# B Local Deployment Configuration

Table 9 reports hardware specifications and inference performance.

# C Prompt Templates

The following task-specific prompt templates were used for all generation tasks. Each template includes explicit grounding instructions and epistemic boundary constraints.

## C.1 Summary Generation Prompt

```
You are an educational summarization assistant. Your task is to produce a concise summary of the provided educational text.
```

**Instructions:**

- Condense the text to approximately 30% of its original length.
- Preserve all key concepts, technical terminology, and hierarchical relationships.
- Maintain causal, comparative, and definitional links.
- Do not introduce information not present in the context.
- If the context lacks sufficient detail for a complete summary, indicate which aspects are omitted.

**Output format:** A single coherent paragraph with section headings where appropriate.

**Context:** {retrieved_chunks}

**Summary:**

## C.2 Question–Answer Generation Prompt

```
You are an educational assessment designer. Your task is to generate question-answer pairs from the provided educational text.
```

**Instructions:**

- Generate exactly {n} question–answer pairs.
- Vary cognitive levels across Bloom's taxonomy: remember, understand, apply, analyze.
- Each question must be unambiguously answerable from the provided context.
- Do not create tautological questions (e.g., "What is X?" followed by "X is X").
- Do not generate questions requiring external knowledge.
- Provide a concise, accurate answer for each question.

**Output format:**

```
Q1: [Question]
A1: [Answer]
...
Q{n}: [Question]
A{n}: [Answer]
```

**Context:** {retrieved_chunks}

## C.3 Flashcard Generation Prompt

```
You are a flashcard designer for spaced repetition study. Your task is to generate flashcards from the provided educational text.
```

**Instructions:**

- Generate flashcards targeting discrete, verifiable concepts.
- Each card must contain exactly one question and one answer.
- Maximum length: 128 tokens per card face (question or answer).
- Optimize for active recall: questions should prompt retrieval of a specific fact, definition, or relationship.
- Do not include information not present in the context.
- If a concept cannot be adequately captured in 128 tokens, split it across multiple cards.

**Output format:**

```
Q: [Question]
A: [Answer]
```

**Context:** {retrieved_chunks}

# D Evaluation Rubrics

Two domain experts independently rated generated artifacts using the rubrics below. Discrepancies were resolved through discussion.

## D.1 Summary Evaluation

The evaluation criteria and rating scales for summaries are listed in Table 10.

## D.2 Question–Answer Set Evaluation

The evaluation criteria and rating scales for question–answer sets are listed in Table 11.

## D.3 Flashcard Evaluation

The evaluation criteria and rating scales for flashcards are listed in Table 12.

Table 8: Questionnaire Structure: Theoretical Constructs, Domains, and Items

| **Theoretical Foundation** | **Domain** | **Item** |
|---|---|---|
| Constructivist Learning Theory | Perceived Usefulness | 1. The generated materials helped me identify key concepts I might have overlooked.<br>2. The system encouraged me to think more deeply about the source content.<br>3. I felt actively involved in constructing my understanding while using the materials.<br>4. The outputs served as a useful starting point for my own further exploration of the topic. |
| Human–AI Interaction Theory | Accuracy and Reliability | 5. The generated content was factually consistent with the uploaded documents.<br>6. I could trace most claims back to specific parts of the source material.<br>7. I felt confident that the outputs did not contain fabricated information.<br>8. The system provided sufficient transparency for me to judge the reliability of its outputs. |
| Cognitive Load Theory | Perceived Cognitive Support | 9. The summaries helped me grasp the main ideas without re-reading the entire document.<br>10. The structured format (summaries, questions, flashcards) reduced the mental effort needed to study the material.<br>11. The system helped me focus on important concepts rather than getting lost in details.<br>12. The generated questions prompted me to think about relationships between concepts rather than memorize isolated facts. |
| Human–AI Interaction Theory | Usability | 13. The interface was intuitive and easy to navigate.<br>14. The generation process was predictable and transparent.<br>15. I understood how to refine or regenerate outputs when needed.<br>16. The system responded in a timeframe that supported my workflow. |

Table 9: Local Deployment Configuration and Inference Performance

| **Category** | **Metric** | **Value** |
|---|---|---|
| Hardware | GPU | NVIDIA RTX 4090 (24 GB) |
| | System RAM | 64 GB |
| | CPU | AMD Ryzen 9 7950X |
| Model | Model | Llama 3.1 8B Instruct |
| | Quantization | 4-bit AWQ |
| Memory | Model memory | $\approx$4.7 GB VRAM |
| | Peak inference memory | 6.2 GB VRAM |
| Latency | Summaries | 2.3 s ($SD = 0.4$) |
| | Question–answer sets | 1.1 s ($SD = 0.2$) |
| | Flashcards | 0.8 s ($SD = 0.2$) |
| Throughput | Sequential load | 18 requests/min |

Table 10: Summary Evaluation Rubric

| **Criterion** | **Definition and Scale** |
|---|---|
| Factual Consistency | Proportion of claims fully supported by the source document. Rated 1–5: 1 = majority unsupported/contradicted; 3 = mixed support; 5 = all claims fully supported. |
| Completeness | Degree to which key concepts from the source are retained. Rated 1–5: 1 = majority omitted; 3 = most present with minor omissions; 5 = all critical concepts retained. |
| Relevance | Alignment between summary content and the inferred information need of a learner studying this material. Rated 1–5: 1 = minimal alignment; 3 = adequate; 5 = precise alignment. |
| Compression Ratio | Objective measure: summary word count divided by source word count. |

Table 11: Question–Answer Set Evaluation Rubric

| **Criterion** | **Definition and Scale** |
|---|---|
| Correctness | Factual accuracy of the provided answer. Rated 1–5: 1 = incorrect; 3 = partially correct; 5 = fully accurate. |
| Answerability | Whether the question can be answered unambiguously from the source context. Rated 1–5: 1 = unanswerable; 3 = answerable with inference; 5 = directly and unambiguously answerable. |
| Relevance | Alignment with the source topic and pedagogical utility. Rated 1–5. |
| Cognitive-Level Validity | Proportion of items matching their intended Bloom's taxonomy classification (remember, understand, apply, analyze). |

Table 12: Flashcard Evaluation Rubric

| **Criterion** | **Definition and Scale** |
|---|---|
| Factual Correctness | Accuracy of the answer component. Rated 1–5. |
| Pedagogical Usefulness | Suitability for spaced repetition and active recall. Rated 1–5: 1 = unsuitable; 3 = adequate; 5 = optimally structured. |
| Coverage | Proportion of key terms from the source represented across the full set of generated flashcards. |