Embedded Selforganizing Systems

Special Issue: "Learner Centered Learning (LCL 2026)" 5th International Workshop

Evaluating Retrieval, Grounding, and Refusal in a Citation-Grounded Local RAG Teaching Assistant

Battsetseg Erdenebat
Multimodal Cognition Laboratory
School of Information and Communication Technology Mongolian University of Science and Technology
Ulaanbaatar, Mongolia
https://orcid.org/0000-0002-0734-4063

Davaajargal Myagmarsuren
Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application
Harbin University of Science and Technology
Harbin, China
https://orcid.org/0009-0005-9156-3304

Uranchimeg Tudevdagva
Graduate school
Mongolian University of Science and Technology
Ulaanbaatar, Mongolia
https://orcid.org/0000-0001-9239-0760

Abstract*— Learner-centered education requires students to access course knowledge independently while retaining the ability to verify, question, and trace the information provided by AI-supported learning tools. Citation-grounded retrieval-augmented generation (RAG) can support this goal by providing on-demand explanations grounded in instructor-approved course materials and by directing learners to the original evidence rather than presenting generated responses as authoritative answers. To examine whether such support can be delivered reliably, this study evaluates retrieval, grounding, citation correctness, and refusal behavior in a locally deployed, citation-grounded RAG teaching assistant for university course materials. The implementation combines OCR-aware ingestion, document- and page-level provenance, sentence-aware chunking, local semantic retrieval, evidence-constrained generation, deterministic refusal normalization, and instructor-facing observability. Two full textbooks, Deep Learning and Power Systems Analysis, were indexed as 2,293 chunks. A source-grounded benchmark contains 60 questions: 52 answerable questions with validated multi-page evidence and eight unsupported questions. With Top-5 retrieval, validated Chunk Hit@5 and Page Hit@5 reached 71.15% and 75.00%, with chunk and page MRR of 0.5135 and 0.5375. Manual evidence-grounded review of end-to-end responses yielded 81.82% normalized correctness among generated answers, 86.36% groundedness, and 86.36% citation correctness. The system answered 42.31% of answerable questions; importantly, 19 of 52 questions (36.54%) were refused even though validated evidence was already present in the retrieved Top-5. Conversely, all eight unsupported questions were correctly rejected with no false answers or citation leakage. The results expose a measurable safety–coverage trade-off: conservative refusal improves unsupported-question safety, but evidence-sufficiency calibration is necessary to avoid suppressing answerable questions.

Keywords—retrieval-augmented generation, educational chatbot, citation grounding, semantic retrieval, grounded generation, refusal, local large language model.

I. INTRODUCTION

Generative artificial intelligence has expanded opportunities for conversational support in higher education. Large language models (LLMs) can explain concepts, summarize material, and respond to student questions, but their direct use also introduces risks related to hallucinated information, unverifiable claims, privacy, and over-reliance on fluent outputs [1-3]. These risks are particularly important when an LLM is positioned as a course teaching assistant, because the expected answer should be consistent with instructor-approved resources rather than with unrestricted model knowledge.

Retrieval-augmented generation (RAG) addresses part of this problem by retrieving external evidence before generation [4]. Dense retrieval [5], sentence-level embeddings [6], and efficient nearest-neighbor search [7] make it practical to ground answers in textbook pages, lecture notes, and other course documents. In education, however, citation presence is not sufficient: the retrieved evidence must actually support the answer, and the system should decline unsupported questions rather than attach citations to an answer generated from parametric memory.

From a learner-centered perspective, the role of the teaching assistant is not to supplant instructor guidance or to give cryptic answers, but to afford learners greater agency in when and how they can receive explanations about the course. Learners can formulate questions according to their individual learning needs, receive evidence-linked responses, inspect the cited textbook pages, and independently verify the generated explanation against instructor-approved materials. This interaction supports learner autonomy through self-directed questioning, transparency through explicit source attribution, and responsible AI use through evidence verification and refusal when the available course material is insufficient. In this sense, retrieval quality, grounding, citation correctness, and refusal behavior are not only technical reliability measures; they determine whether the system can function as a trustworthy learner-centered learning support tool.

Accordingly, this study investigates whether a locally deployed, citation-grounded RAG assistant can provide reliable learner-centered access to university course materials by evaluating evidence retrieval, grounded and verifiable

answer generation, and evidence-aware refusal of unsupported questions.

The present work therefore treats reliability as a staged problem rather than a single end-to-end accuracy score. We evaluate whether validated evidence is retrieved, whether generated responses remain supported by that evidence, and whether the system refuses appropriately when evidence is absent or judged insufficient. The study reports the revised two-textbook implementation and a source-grounded benchmark designed specifically to separate these failure modes.

The contributions are threefold. First, we implement a fully local, citation-grounded RAG teaching assistant integrating optical character recognition (OCR)-aware document ingestion, document-level provenance, semantic retrieval, evidence-constrained generation, deterministic refusal control, and instructor-facing observability. Second, we construct and validate a 60-question benchmark comprising 52 answerable and eight unsupported questions across two technically distinct university textbooks, enabling separate evaluation of retrieval, grounded generation, citation correctness, and refusal behavior. Third, the evaluation reveals a measurable safety–coverage trade-off with direct implications for learner-centered AI support: generated responses achieve 81.82% manually reviewed normalized correctness and unsupported questions are rejected with 100% accuracy and zero citation leakage, yet 19 of 52 answerable questions are refused despite validated evidence being present in Top-5 retrieval.

II. Related Work

A. Educational Conversational Systems

Educational chatbots have been studied for tutoring, information support, learner interaction, and administrative assistance. Reviews report benefits in accessibility and responsiveness while emphasizing pedagogical design, evaluation, and instructor involvement [8-10]. Foundation models increase conversational capability but also increase the importance of evidence verification because fluent responses may be unsupported [1, 2].

B. Retrieval-Augmented Generation and Grounding

RAG combines retrieval with conditional generation so that a model can use external evidence at inference time [4]. Dense Passage Retrieval established neural passage retrieval for question answering [5], while Sentence-BERT enabled efficient semantic representations [6]. HNSW and GPU-based similarity search provide scalable retrieval mechanisms [7, 11]. For educational use, the principal advantage is that the knowledge collection can be restricted to instructor-selected material. The limitation is that generation quality remains bounded by ingestion and retrieval quality: a plausible answer with an irrelevant citation is not a grounded answer.

C. OCR and Document Provenance

University repositories often contain scanned textbooks, figures, equations, and inconsistent layouts. OCR is therefore necessary when native PDF text is unavailable; Tesseract provides a widely used open-source solution [12]. OCR errors can nevertheless propagate into headings, equations, metadata, and chunks. The revised system therefore treats provenance—including source file, document identity, page number, extraction path, and source hash—as part of the retrieval object rather than as presentation-only metadata.

III. System Design and Implementation

A. Architecture

Fig. 1 summarizes the current architecture. Instructor-provided PDFs are processed by a research-oriented ingestion pipeline. Native text is used when available; scanned pages use OCR fallback. The pipeline creates document-specific identifiers, preserves PDF-page provenance, stores source hashes, and forms sentence-aware overlapping chunks. The resulting local index is queried by semantic similarity, and the Top-K chunks are passed to an evidence-aware QA engine. Student interactions and retrieval diagnostics can be summarized in a teacher analytics dashboard.

B. Research-Oriented Ingestion and Indexing

The current ingestion design replaced manually fixed course labels and fixed-length character splitting. Document titles and identifiers are derived per source, and each chunk stores document type, document ID, document title, course label, PDF page, unique chunk ID, source filename, and SHA-256 hash. Sentence-aware chunking uses a target length of approximately 220 words, an upper bound of 300 words, and approximately 40 words of overlap. Rendered page images can also be preserved for later visual verification and multimodal extensions. The evaluated index contained 2,293 chunks from two full textbooks: Goodfellow, Bengio, and Courville's Deep Learning [13] and Bergen and Vittal's Power Systems Analysis, 2nd ed. [14]. The former is predominantly digitally extractable, whereas the latter contains scanned pages and therefore provides a substantially different ingestion condition.

C. Local Retrieval and Evidence-Aware Generation

The retriever uses the local nomic-embed-text embedding model and returns the Top-5 chunks. Generation is performed locally with Qwen2.5-7B through Ollama. Retrieved chunks expose their document and page metadata to the QA engine. The generator is instructed to use only retrieved course evidence and to provide citations only when evidence is sufficient. The response policy contains two safety stages. The corresponding decision logic is illustrated in Fig. 2.

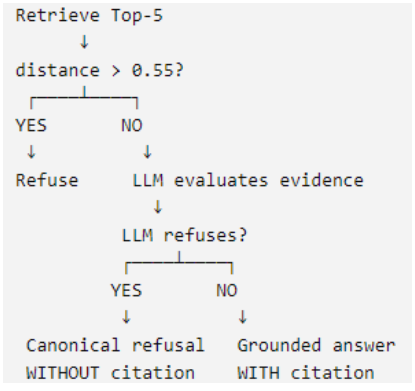


Fig. 2. Revised local citation-grounded educational RAG architecture.

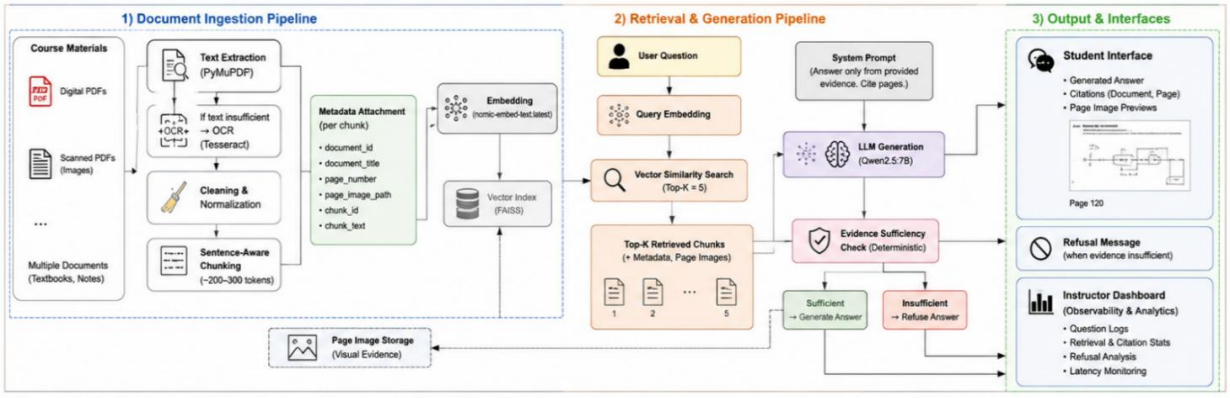


Fig. 1. Revised local citation-grounded educational RAG architecture.

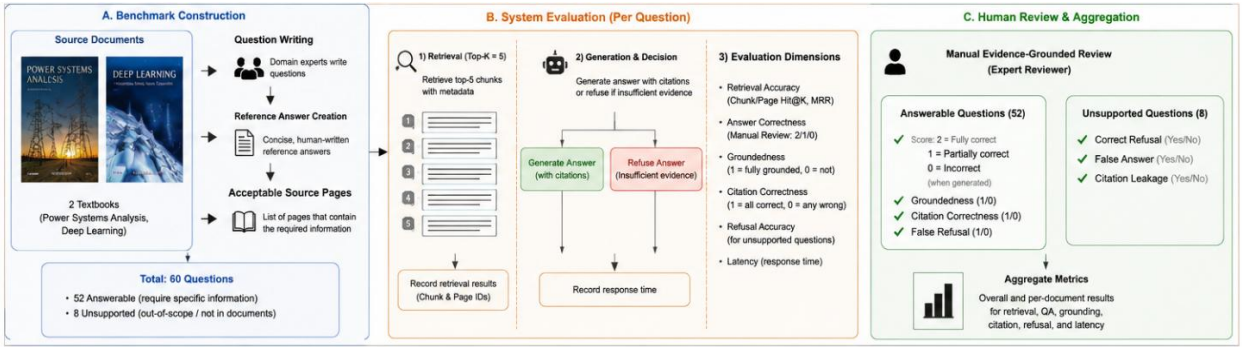


Fig. 3. Evaluation protocol separating retrieval/QA quality from unsupported-question safety.

A retrieval-distance rule can refuse low-relevance results, while the generator can also determine that the retrieved evidence is insufficient. Any generation-level refusal is normalized to a canonical citation-free response. This deterministic post-processing was introduced after preliminary testing showed that an LLM could correctly refuse yet still append irrelevant citations because of output-format instructions.

D. *Instructor-Facing Observability*

The Streamlit-based instructor interface records anonymous interaction-level signals such as question text, language, retrieved evidence, retrieval distance, latency, citation behavior, and refusal status. These signals are summarized to support inspection of system behavior and recurring retrieval/refusal patterns. The dashboard is treated as an observability component rather than as an evaluated learning-analytics contribution; no claim is made that these signals directly measure student learning outcomes.

IV. EXPERIMENTAL PROTOCOL

A. *Benchmark Construction*

The final benchmark contains 60 questions and is separated into answerable and unsupported subsets (Fig. 3). Fifty-two answerable questions are evenly divided between Deep Learning and Power Systems Analysis. Each answerable item contains a reference answer and validated acceptable evidence pages; Each answerable item contains a reference answer and validated acceptable evidence pages; subsequent evidence auditing and review were used to identify alternative supporting pages and construct the final multi-evidence reference set. Eight additional questions intentionally fall outside the two-course corpus and are used only for refusal/out-of-scope evaluation.

B. *Retrieval Metrics*

Retrieval is evaluated independently from generation. Chunk Hit@K records whether a validated supporting chunk appears within the first K results, while Page Hit@K checks whether any retrieved result originates from an acceptable evidence page. MRR measures the reciprocal rank of the first validated hit. This separation avoids attributing retrieval failures to the generator or, conversely, crediting a fluent answer when the correct evidence is not retrieved.

C. *End-to-End QA and Manual Evidence-Grounded Review*

For each of the 52 answerable questions, the system retrieves Top-5 chunks and either generates or refuses an answer. An independent local Llama3.1-8B judge was used only to flag questionable cases; its scores were not treated as final because some semantically correct responses were penalized for wording differences. All 48 flagged cases were then manually reviewed against the benchmark reference answer, validated pages, retrieved context, and generated response using predefined rubric. Answer scores were 2 for fully correct, 1 for partially correct, and 0 for incorrect or unanswered. Normalized correctness is the obtained score divided by the maximum possible score. Groundedness and citation correctness are reported for generated answers. Because this was a single-reviewer evidence-grounded assessment rather than a multi-rater annotation study, the results are described as manually reviewed rather than as inter-rater-validated ground truth.

D. Refusal Evaluation

The eight unsupported questions test whether the system avoids answering from outside knowledge. Correct refusal rate, false-answer rate, citation leakage, and response latency are recorded. This evaluation uses the same Top-5 retrieval and Qwen2.5-7B generation configuration as the answerable evaluation.

E. Implementation Configuration and Reproducibility

All evaluated components run locally. Ollama provides the embedding and generative-model services, with nomic-embed-text: latest used for embedding and Qwen2.5-7B used for answer generation. The retrieval index is rebuilt from the revised processed_v2 outputs rather than from the earlier prototype database, ensuring that the evaluation uses the document-specific metadata produced by the final ingestion pipeline. Top-K is fixed at 5 for all reported retrieval and QA experiments. The benchmark, validated evidence mappings, detailed retrieval outputs, refusal logs, and human-review tables are stored as separate experiment artifacts so that retrieval, generation, and review decisions can be reproduced independently.

The evaluation was intentionally staged. First, retrieval was measured without involving the generator. Second, the unsupported subset was used to test refusal behavior and citation leakage. Third, end-to-end QA was executed on the answerable subset. Finally, cases flagged by automated evaluation were reviewed manually. This sequence reduces the risk of diagnosing a retrieval miss as a language-model error and makes it possible to identify where a failure enters the pipeline.

F. Evidence Validation Procedure

The initial benchmark associated each question with a nominal gold page. During evidence audit, retrieved chunks were compared with the question and reference answer. Because textbook explanations can span neighboring pages or repeat a concept in more than one location, a single gold page was not always sufficient. Candidate alternative evidence was therefore exposed for review, and only pages judged to directly support the reference answer were added to the acceptable evidence set. The final retrieval evaluation consequently uses validated multi-evidence references rather than assuming that one page is the only correct source.

This distinction materially affects evaluation quality. A retrieval result from a neighboring page should not be counted as a miss when that page independently contains the required evidence. Conversely, semantic similarity alone is not enough to declare a hit: the retrieved page must support the answer required by the benchmark. The evidence-audit stage therefore provides the ground truth used by both Page Hit@K and the subsequent human QA review.

V. Results

A. Validated Retrieval Performance

Table I shows that the revised full-textbook index retrieved an acceptable evidence page within the Top-5 for 75.00% of answerable questions and a validated chunk for 71.15%. Performance was reasonably balanced across the two books at K=5, although Deep Learning produced higher mean reciprocal rank (MRR), indicating that its first relevant result tended to appear earlier.

TABLE I. Validated retrieval results on 52 answerable questions

Metric	Overall	Deep Learning	Power Systems
Chunk Hit@1	40.38%	46.15%	34.62%
Chunk Hit@3	65.38%	65.38%	65.38%
Chunk Hit@5	71.15%	73.08%	69.23%
Page Hit@1	40.38%	46.15%	34.62%
Page Hit@3	71.15%	65.38%	76.92%
Page Hit@5	75.00%	73.08%	76.92%
Chunk MRR	0.5135	0.5558	0.4712
Page MRR	0.5375	0.5558	0.5192

B. Manually Reviewed End-to-End QA

The system generated substantive answers for 22 of 52 answerable questions, corresponding to 42.31% answer coverage. Sixteen responses were fully correct, four were partially correct, and two generated answers were incorrect. Using the 2/1/0 rubric, normalized correctness among generated responses was 81.82%. When all 52 answerable questions are included and refusals receive zero score, normalized end-to-end correctness was 34.62%. Groundedness and citation correctness among generated responses were both 86.36%.

The 30 answerable-question refusals should not all be interpreted as false refusals. Manual failure decomposition showed that 19 cases had an acceptable validated evidence page already present in the retrieved Top-5, but the evidence-aware generation stage still refused. We define these as evidence-available false refusals, representing 36.54% of the 52 answerable questions. In the remaining 11 cases (21.15%), no validated evidence page was present in Top-5, so refusal was reasonable given the evidence supplied to the generator. Thus, the overall answerable-question refusal rate of 57.69% reflects two distinct causes: retrieval insufficiency and over-conservative evidence-sufficiency control in Table II.

TABLE II. Manually reviewed end-to-end QA results

Metric	Overall	Deep learning	Power system
Answer coverage	42.31%	57.69%	26.92%
Answerable-question refusal rate	57.69%	42.31%	73.08%
Fully correct/ all questions	30.77%	38.46%	23.08%
Partial/ all questions	7.69%	11.54%	3.85%
Normalized correctness/ questions	34.62%	44.23%	25.00%
Normalized correctness/ answered	81.82%	76.67%	92.86%
Groundedness/ answered	86.36%	100.00%	57.14%
Citation correctness/ answered	86.36%	86.67%	85.71%
Average response time	1.135 s	1.227 s	1.044 s

C. Unsupported-Question Safety

All eight unsupported questions were rejected without false answers or citation leakage in Table III. Importantly, the retrieval-distance threshold alone did not explain these refusals; semantically related chunks could still be retrieved for out-of-scope questions. The evidence-aware generation

stage and canonical refusal normalization were therefore necessary to prevent unsupported output.

TABLE III. REFUSAL / OUT-OF-SCOPE EVALUATION

Metric	Overall
Unsupported questions	8
Correct refusal rate	100.00%
False answer rate	0.00%
Citation leakage rate	0.00%
Average response time	0.924 s

D. Manual Failure Decomposition

Table IV separates the 57.69% answerable-question refusal rate into two actionable causes. Nineteen questions (36.54%) reached the generator with validated evidence already present, yet the generator still judged the evidence insufficient; these evidence-available false refusals are the clearest targets for calibration of the evidence-sufficiency policy. Eleven questions (21.15%) lacked a validated evidence page in Top-5; in those cases, refusal protects the system from answering without benchmark-supported evidence. This decomposition distinguishes retrieval improvement from refusal-policy improvement and avoids labeling every refusal on an answerable benchmark item as a generation error. The larger evidence-available group (19 versus 11 cases) indicates that post-retrieval evidence-sufficiency control, rather than retrieval failure alone, is a major source of reduced answer coverage. From a learner-centered perspective, such evidence-available refusals can unnecessarily restrict learners' self-directed access to verifiable course knowledge.

The 22 generated answers also show why correctness, grounding, and citation quality should be measured separately. Two Deep Learning responses were judged incorrect despite being grounded in retrieved text because they did not answer the requested concept. Four additional responses were partially correct, typically because they omitted an important qualifier or replaced the benchmark formulation with a less directly supported explanation. In a small number of otherwise correct responses, extra explanatory claims reduced groundedness even though the core answer was correct. These cases motivate future answer-generation prompts that favor concise evidence-bounded explanations over elaboration.

TABLE IV. MANUAL FAILURE DECOMPOSITION FOR 52 ANSWERABLE QUESTIONS

Failure	N	Interpretation
Evidence retrieved; refusal	19	Over-conservative sufficiency
No valid evidence; refusal	11	Retrieval insufficiency; reasonable
Fully correct answer	16	Successful response
Partially correct answer	4	Incomplete/ mildly imprecise
Incorrect answer	2	Generation/ selection error

VI. DISCUSSION

The staged evaluation is more informative than a single end-to-end score because it isolates where reliability is gained or lost. Page Hit@5 of 75.00% shows that validated evidence is frequently available to the generator, while 81.82% normalized correctness among generated responses indicates that local generation can be effective when the system elects to answer. Groundedness and citation correctness of 86.36% further show that answer correctness, evidential support, and source attribution are related but distinct properties.

At the same time, the results reveal a clear safety–coverage trade-off. The refusal mechanism eliminated false answers and citation leakage on the eight unsupported questions, but answer coverage on the answerable benchmark was only 42.31%. More specifically, 19 of 52 answerable questions (36.54%) were refused despite validated evidence being present in Top-5. Therefore, simply strengthening refusal rules is not an adequate route to educational reliability; evidence-sufficiency detection must be calibrated jointly with retrieval quality and answer coverage. From a learner-centered perspective, educational RAG should balance safe refusal with learner autonomy and access to verifiable course knowledge.

The difference between groundedness and correctness is also instructive. A response may be supported by retrieved text yet fail to answer the benchmark question completely, while another response may contain a correct core answer but add unsupported explanatory material. Human review was necessary to distinguish these cases because the independent LLM judge occasionally penalized acceptable paraphrases or rewarded conservative responses. For a small research benchmark, human validation therefore provides a more defensible final assessment.

The scanned Power Systems Analysis textbook exhibited a more difficult generation condition despite slightly higher Page Hit@5 than the digitally extracted Deep Learning textbook (76.92% vs. 73.08%). Although retrieval performance was comparable, answer coverage was substantially lower for the OCR-derived material (26.92% vs. 57.69%). This contrast indicates that page-level retrieval success does not necessarily translate into sufficient evidence by the generation; however, document format is confounded with textbook domain and content, so the observed difference cannot be attributed causally to OCR. Preserved page images provide a practical basis for future figure-, equation-, and table-aware verification, but those multimodal capabilities are not evaluated in the present study.

Several limitations should be noted. The benchmark contains 60 questions, including only eight unsupported items, which limits generalization; therefore, the 100% refusal result should be interpreted strictly within this test set. The manual evidence-grounded review was performed on flagged cases by a single reviewer rather than through a multi-rater study, and inter-rater agreement is therefore unavailable. The evaluation uses two textbooks and one local model configuration; model, embedding, chunking, threshold, and Top-K sensitivity remain to be studied. Finally, the instructor dashboard provides system observability, but its relationship to student learning outcomes has not yet been evaluated.

REFERENCES

[1] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and individual differences,* vol. 103, p. 102274, 2023.

[2] W. Holmes, M. Bialik, and C. Fadel, *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign, 2019.

[3] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in

higher education–where are the educators?," *International journal of educational technology in higher education,* vol. 16, no. 1, p. 39, 2019.

[4] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems,* vol. 33, pp. 9459-9474, 2020.

[5] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 6769-6781.

[6] N. Reimers and I. Gurevych, "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 (EMNLP-IJCNLP)*, 2019, pp. 3982-3992.

[7] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE transactions on pattern analysis and machine intelligence,* vol. 42, no. 4, pp. 824-836, 2018.

[8] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence,* vol. 2, p. 100033, 2021.

[9] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet?-a systematic literature review on chatbots in education," *Frontiers in artificial intelligence,* vol. 4, p. 654924, 2021.

[10] A. Kerlyl, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models," in *International conference on innovative techniques and applications of artificial intelligence*, 2006: Springer, pp. 179-192.

[11] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE transactions on big data,* vol. 7, no. 3, pp. 535-547, 2019.

[12] R. Smith, "An overview of the Tesseract OCR engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, 2007, vol. 2: IEEE, pp. 629-633.

[13] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* (no. 2). MIT press Cambridge, 2016.

[14] A. R. Bergen and V. Vittal, *Power Systems Analysis*. Prentice Hall, 2000.