GSC Advanced Research and Reviews
eISSN: 2582-4597 CODEN (USA): GARRC2
Cross Ref DOI: 10.30574/gscarr
Journal homepage: https://gscarr.gsconlinepress.com/

(RESEARCH ARTICLE)

Check for updates

DESIGN AND DIAGNOSTIC EVALUATION OF A RAG-BASED EDUCATIONAL CHATBOT FOR DISTANCE TRAINING IN DIGITAL TECHNOLOGY AND ARTIFICIAL INTELLIGENCE FOR EDUCATION PROFESSIONALS IN SENEGAL

Ahmadou BA [1], Mamadou Vieux Lamine SANE [2] and Akimou TCHAGNAOU [3, *]

[1] Cheikh Amadou Kane Digital University/Laboratory for the Administration, Management, and Governance of Educational Institutions (LABO_AGGEE), Senegal.
[2] Cheikh Amadou Kane Digital University/Laboratory for the Administration, Management, and Governance of Educational Institutions (LABO_AGGEE), Senegal.
[3] Andre Salifou University/Laboratory of Letters, Education and Communication (LaboLEC), Niger.
* Corresponding Author

ORCID Details
Akimou TCHAGNAOU: https://orcid.org/0009-0006-9753-515X

GSC Advanced Research and Reviews, 2026, 29(02), 092–101

Publication history: Received on 25 September 2026; revised on 01 November 2026; accepted on 03 November 2026

Article DOI: https://doi.org/10.30574/gscarr.2026.29.2.0308

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

This study examines the design and diagnostic evaluation of an educational chatbot based on retrieval-augmented generation (RAG) for the distance training *of education professionals* in Senegal. The protocol compares four embedding models on 54 answerable questions and five small language models on 60 questions under two document-context conditions: gold-standard passages and fixed RAG contexts. BGE-M3 achieves the highest overall descriptive scores for dense retrieval (Hit@3 = 0.500; MRR@10 = 0.448), although performance varies by question type. For all five generators, replacing the RAG context with the gold context improves automatically assessed accuracy and coverage of the expected information. The findings highlight the interdependence of corpus organization, retrieval, and generation, without establishing the pedagogical superiority of any model.

Keywords: Distance Education; Tutoring; Educational Chatbot; Artificial Intelligence; Retrieval-Augmented Generation; RAG Evaluation

1 INTRODUCTION

Senegal's PAQUET-EF, LPSD-SEF and Digital Strategy for Education identify digital skills and support for education professionals as priorities (Ministère de l'Éducation nationale du Sénégal, 2018, 2025a, 2025b). The Ministry and Université numérique Cheikh Hamidou Kane consequently introduced online digital technology and artificial intelligence training for 105,000 education professionals, comprising two certificates at beginner, intermediate and

advanced levels. At prototype design, 61,275 enrolments had been recorded across pathways, from a workforce of 108,535.

Distance education enables staff to train while remaining in post, reducing geographical, financial and organizational barriers (Mbaye, 2026). Digital badges and certificates recognize skills development and may support applications for positions of responsibility (MEN, 2025b). Its effectiveness nevertheless depends on learner support.

Distance tutoring encompasses pedagogical, methodological, technical and organizational functions (Denis, 2003). Reviewing 170 studies across 40 low- and middle-income countries, Hennessy et al. (2022) highlight technology's contribution to teacher development while emphasizing human relationships and support.

Providing individualized human tutoring for tens of thousands of participants is costly. Educational chatbots offer immediate support, but their language models may hallucinate. Ji et al. (2023) distinguish intrinsic hallucinations, which contradict sources, from extrinsic hallucinations, which introduce unsupported information. In certification programmes, plausible responses must therefore be checked against the resources taught.

Retrieval-augmented generation (RAG) combines a model's parametric memory with an indexed external corpus (Lewis et al., 2020). Educational applications use this architecture to ground responses in course materials (Alsafari et al., 2024; Swacha & Gracel, 2025; Li et al., 2025). RAGChecker, ARES and RAGAS assess retrieval, context relevance, faithfulness and answer relevance separately (Ru et al., 2024; Saad-Falcon et al., 2024; Es et al., 2024). Evaluation must also test abstention when the corpus cannot answer a query (Peng et al., 2025).

This study describes and diagnostically evaluates a RAG chatbot using small language models suitable for resource-constrained environments in Senegal. It examines the relationship between corpus organization, retrieval and generation through four questions:

The study is structured around four research questions:

- How can an educational chatbot be designed to provide pedagogical support while ensuring that its responses are grounded in, faithful to, and traceable to the programme's learning resources?
- Does the chatbot's retrieval performance vary by embedding model and question type?
- Does generation quality depend on the language model or on retrieval quality? More specifically, do accuracy, coverage of expected information, lexical faithfulness, and abstention vary according to the language model used?
- Which factors should be prioritized when deploying an educational chatbot: retrieval quality, generation quality, abstention capability, or latency?

2 CONCEPTUAL FRAMEWORK

2.1 Distance Education and Learner Support

Distance education offers practising education professionals flexibility, self-paced learning and access to resources beyond geographical constraints (Galichet, 2007; Ba, 2024). Technology also supports peer exchanges, reflection and professional communities (Hennessy et al., 2022).

Technical, temporal and pedagogical distance require tutor availability, supervision and collaboration (Karsenti & Collin, 2011). Tutors mediate between learners and the learning environment while supporting motivation and autonomy (Denis, 2003). Hennessy et al. (2022) associate remote tutoring with improved practices but recommend hybrid support to sustain trust. Chatbots should therefore allow escalation to human tutors when their knowledge base is insufficient.

2.2 AI-Assisted Tutoring

Educational chatbots can support technical troubleshooting, access to resources, learning methods and assessment preparation (Denis, 2002; Delaherche & Glafkides, 2020). AI tutoring may personalize content and learning support (Karsenti, 2018). Applications include organizational guidance, question generation, course selection and learning trace analysis (Swacha & Gracel, 2025). Immediate, adaptive responses may sustain engagement (Ba, 2024), but inconsistencies, fabricated references and simulated reflexivity require caution (Las-Vergnas & Rio-Jouet, 2023). Chatbots cannot reproduce a human tutor's commitment and sensitivity.

2.3 Retrieval-Augmented Generation as a Mechanism for Grounding Educational Chatbots in Documents

Educational chatbots can address delays in resolving difficulties during distance learning. RAG adapts general-purpose models to course-specific materials. In a seven-week online master's course, Lang and Gürpinar (2025) report chatbot use by 42.10% of students and favourable perceptions among 87%, with correlated attitudes towards the chatbot and its use in other courses ($r = 0.88$; $p < 0.001$).

Grounding support in course resources remains essential (Lewis et al., 2020; Lang & Gürpinar, 2025; Miladi et al., 2025). RAG does not eliminate hallucinations: retrieval may be irrelevant, or generation may ignore the context and rely on training knowledge (Swacha & Gracel, 2025). Diagnostic evaluation is therefore necessary before learner deployment.

2.4 Evaluation Frameworks for RAG Systems

Pedagogical chatbot quality depends on corpus organization, retrieval, tutoring functions and interaction rules (Alsafari et al., 2024). Evaluation should distinguish retrieval from answer generation.

- RAGChecker decomposes generated and reference answers into claims, *th*en checks them against retrieved passages to assess relev*anc*e, accuracy, completene*ss* and fa*ith*fulness (Ru et al., 2024).
- ARES uses specialized language-model judges to evaluate context relevance, answer faithfulness and answer relevance (Saad-Falcon et al., 20*24).*
- RAGAS l*ikewise s*eparates faithfulness to retrieved documents, answer relevance and context relevance. *These* complementary dimen*sions h*elp distinguish ret*riev*al failures from *u*nsupported or irrelevant generation (Es et al., 2024).
- UAEval4RAG *extends evaluation to unanswer*able queries. High retrieval scores alone do not ensure accurate answers; reliable educational assistants must recognize when the knowledge base is insufficient and abstain (Peng et al., 2025).

3 METHODOLOGY

3.1 Design and Architecture of AVENUE

AVENUE (assistant virtuel d'enseignement au numérique éducatif) supports participants in the online training programme. Its architecture separates knowledge storage, passage retrieval and generation through a web interface, FastAPI, MySQL, a vector index and Ollama. Certificate-specific conversations normalize queries, retrieve passages and supply context to the generator.

The document collection contains 15,952 chunks associated with 45 modules and 117 learning sequences, extracted from 303 documents used in the digital technology and artificial intelligence training programme. It includes chunks from six certificate-level pathways: beginner, intermediate, and advanced levels for each of the two certificates.

The chatbot's MySQL database contains 13 main tables divided into two groups:

- Five tables for learning content: courses, *modules, sequences, course_content, and* course_chunks.
- Eight tables for learners, interactions, and reference data: learners, learner_enrollments, *conversations, messages, functions_ref, certificates_ref, levels_ref, and cohorts_ref. These tables* will support future evaluation of the chatbot's pedagogical effectiveness.

Table1 Database Schema

No.	Table	Function	Main Data Stored
1	courses	Identifies the courses and training pathways available in AVENUE IA.	Course identifier and course or certificate title.
2	modules	Organizes each course into learning modules and defines their sequence.	Module identifier, associated course, module title, order, and creation date.
3	sequences	Structures modules into learning sequences corresponding to learning units.	Sequence identifier, associated module, title, order, estimated duration, and learning objectives.
4	course_content	Stores the source learning content associated with courses. Also provides	Content identifier, associated course, section title, and learning text.

		fallback content when chunks are unavailable.	
5	course_chunks	Stores the document chunks used by the RAG system to retrieve information for answer generation.	Chunk identifier, text, course, module, sequence, section, order, indexing text, and source-content reference.
6	learners	Records identifying information about learners using the chatbot.	Identifier, learner code, full name, class, sex, date of birth, email, staff number, certificate, and professional role.
7	learner_enrollments	Links each learner to their training context and records enrolments.	Enrolment identifier, learner, professional role, certificate, level, cohort, and enrolment date.
8	conversations	Records dialogue sessions between learners and the virtual assistant.	Conversation identifier, learner, associated course, title, unique identifier, and start date.
9	messages	Stores learners' questions and the virtual assistant's generated answers.	Message identifier, associated conversation, author (learner or assistant), text, date, and unique identifier.
10	functions_ref	Centralizes learners' professional categories to standardize identification.	Identifier and title of the professional role.
11	certificates_ref	Lists training certificates and standardizes their names.	Certificate identifier and title.
12	levels_ref	Defines the training levels associated with certificates.	Level identifier and designation: beginner, intermediate, or advanced.
13	cohorts_ref	Identifies the learner cohorts enrolled in training.	Cohort identifier and title.

Sources: Official technical datasheets for the models published on Hugging Face, and specifications referenced in the MTEB (Massive Text Embedding Benchmark)

3.2 Diagnostic Evaluation of the Chatbot

The chatbot's diagnostic evaluation uses automated assessment through two complementary benchmark campaigns conducted on the same question set.

3.3 Automated Evaluation of Document Retrieval

Four embedding models were selected for semantic retrieval, multilingual capabilities and suitability for constrained hardware, following the cross-task evaluation principles of BEIR and MMTEB (Thakur et al., 2021; Enevoldsen et al., 2025).

Table 2:2 Technical Characteristics of the Embedding Models

Model	Approximate Parameters	Vector Dimension	Reported Maximum Length
Qwen3-Embedding-0.6B	600 M	1 024	32 768
BGE-M3	568 M	1 024	8 192
multilingual-e5-large	560 M	1 024	512
Granite Embedding 311M R2	311 M	768	32 768

Sources: repositories of quantized models on Hugging Face

These models are compared using five metrics that assess document retrieval quality:

- Hit Rate@k indicates whether at least one reference chunk occurs in the top k results (1 = yes; 0 = no) (Karpukhin et al., 2020).
- Precision@k is the proportion of relevant chunks among the top k results.
- Recall@k is the proportion of reference chunks retrieved among the top k results.

- MRR@10 averages the reciprocal ranks of the first relevant chunks in the top ten results.
- nDCG@k compares the retrieved ranking with an ideal ranking, weighting early relevant chunks more heavily and using binary relevance (Thakur et al., 2021).

3.4 Automated Evaluation of Text Generation

The second campaign compares five generative models selected for two technical characteristics: multilingual generation with French support and suitability for hardware-constrained computing environments. The models are compared under two experimental conditions:

- A gold-standard reference context: *the model* receives the relevant reference chunks directly (0–4 per question), allowing evaluation of its ability to use reliable information (Karpukhin et al., 2020; Lewis et al., 2020).
- A fixed RAG context: the model receives up to three chunks obtained through hybrid lexical and semantic retrieval, allowing evaluation of its ability to generate an answer from the retrieved information (Falcon et al., 2024; Es et al., 2023).
- For each question, all five models receive the same context. For unanswerable questions, the reference context is empty, while the fixed RAG context contains three irrelevant chunks to assess whether the models abstain when the requested information is absent (Peng et al., 2025).

A shared system prompt requests a pedagogical answer in French grounded in the supplied documents, explicit acknowledgement of missing information, and the use of references.

Table 3:3 Technical Characteristics of the Generative Models

Model	Parameters (billions)	GGUF Size (GB)	Quantization	Peak RAM (MB)	Median Time to First Token (ms)	Throughput (tokens/s)	Median Latency (ms)
Qwen3-1.7B	2.032	1.359	Q4_K_M	1682	11912	13.07	33714
SmolLM2-1.7B-Instruct	1.711	1.056	Q4_K_M	1648	14844	14.52	40537
Gemma 3 1B IT	1.000	0.815	Q4_K_M	839	12215	17.52	24833
Llama 3.2 1B Instruct	1.236	0.808	Q4_K_M	926	17968	21.37	43425
LFM2.5-1.2B-Instruct	1.170	0.731	Q4_K_M	794	7907	30.07	15599

Sources: Quantized model repositories on Hugging Face.

The models are also compared using the following metrics:

- Accuracy assesses the overall factual correctness of the answer against the ground truth on a scale from 0 to 4 (Alsafari et al., 2024).
- Coverage measures the proportion of expected information from the reference answer that appears in the generated text (Ji et al., 2023; Peng et al., 2025).
- Lexical faithfulness measures the proportion of generated sentences strictly supported by the chunks in the document context (Es et al., 2023; Saad-Falcon et al., 2024).
- The pedagogical index assesses the structure, clarity, and usefulness of the answer as tutoring support on a scale from 0 to 4 (Alsafari et al., 2024; Swacha & Gracel, 2025).

3.5 Development of the Diagnostic Evaluation Question Set

The benchmark comprises 60 questions from the two cer*tificates, "Comp*uting an*d the Internet* in Education" and "Artificial Intelligence in Education": 54 answerable and six requesting information absent from the materials. Answerable records specify certificate, level, module, sequence, question type, reference answer, expected information and source locations.

3.6 Statistical Test

Paired Wilcoxon signed-rank tests compare the four retrievers on the same 54 questions (Wilcoxon, 1945), using first-relevant-chunk rank, reciprocal rank and Recall@3. Significance was set at 0.05.

4 RESULTS

4.1 Retrieval Model Performance by Question Type

Table 4:4 Definition / Concept

Model	Hit@3	Recall@3	MRR@10	nDCG@3
BGE-M3	0.5833	0.3750	0.5501	0.3966
Granite Embedding 311M R2	0.5000	0.2917	0.5104	0.3388
Qwen3-Embedding-0.6B	0.5000	0.2917	0.5104	0.3388
multilingual-e5-large	0.4167	0.2917	0.4378	0.3200

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Table 5:5 / Understanding

Model	Hit@3	Recall@3	MRR@10	nDCG@3
BGE-M3	0.6667	0.3333	0.5000	0.3200
Granite Embedding 311M R2	0.6667	0.3333	0.5000	0.3333
Qwen3-Embedding-0.6B	0.5000	0.2500	0.5519	0.3066
multilingual-e5-large	0.8333	0.4167	0.6111	0.4088

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Table 6:6 Factual Information

Model	Hit@3	Recall@3	MRR@10	nDCG@3
BGE-M3	0.4167	0.2917	0.3111	0.2448
Granite Embedding 311M R2	0.2500	0.1667	0.2769	0.1667
Qwen3-Embedding-0.6B	0.2500	0.1667	0.2710	0.1667
multilingual-e5-large	0.2500	0.2083	0.2079	0.1615

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Table 7:7 Passages

Model	Hit@3	Recall@3	MRR@10	nDCG@3
BGE-M3	0.6667	0.2222	0.4639	0.2654
Granite Embedding 311M R2	0.5000	0.2500	0.3750	0.3045
Qwen3-Embedding-0.6B	0.5000	0.2083	0.3681	0.2551
multilingual-e5-large	0.3333	0.1250	0.2556	0.1667

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Table 8:8 Procedure / Steps

Model	Hit@3	Recall@3	MRR@10	nDCG@3
BGE-M3	0.2500	0.2083	0.3632	0.2111
Granite Embedding 311M R2	0.2500	0.2083	0.3036	0.1989
Qwen3-Embedding-0.6B	0.2500	0.1667	0.2190	0.1171
multilingual-e5-large	0.1667	0.1250	0.1944	0.1037

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Table 9:9 Contextualized Scenario

BGE-M3	0.6667	0.3333	0.6167	0.3710
Granite Embedding 311M R2	0.6667	0.3333	0.5278	0.3333
Qwen3-Embedding-0.6B	0.8333	0.5000	0.4907	0.3978
multilingual-e5-large	0.5000	0.3333	0.3472	0.2689

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

BGE-M3 leads descriptively on Hit@3, Hit@10, Recall@3, MRR@10 and nDCG@3. Its Hit@3 is 50.0%, versus 42.6% for Qwen3 and Granite and 37.0% for e5; Recall@3 remains 0.293. For BGE-M3 versus e5, paired Wilcoxon p-values are 0.0007 (first relevant rank) and 0.0011 (reciprocal rank); versus Qwen3, they are 0.0319 and 0.0650.

4.2 Language Model Performance

Table 10:10 Generative Model Performance with a Fixed RAG Context

Model	Accuracy/4	Coverage	Lexical Faithfulness	Pedagogical Index/4
Qwen3-1.7B	1.504	0.494	0.655	3.395
SmolLM2-1.7B-Instruct	1.011	0.311	0.769	2.680
Gemma 3 1B IT	1.300	0.399	0.819	3.015
Llama 3.2 1B Instruct	1.314	0.447	0.712	3.331
LFM2.5-1.2B-Instruct	1.249	0.379	0.610	2.476

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

With a fixed RAG context, meaning that answers are generated from chunks supplied by the retriever, Qwen3-1.7B obtains the highest accuracy, coverage, and pedagogical index scores. Gemma 3 1B IT achieves the highest lexical faithfulness (0.819).

Table 11:11 Generative Model Performance with a Gold-Standard Reference *Context*

Model	Accuracy /4	Coverage	Lexical Faithfulness	Pedagogical Index /4
Qwen3-1.7B	1.841	0.600	0.632	3.567
SmolLM2-1.7B-Instruct	1.237	0.396	0.719	2.867
Gemma 3 1B IT	1.803	0.576	0.682	3.327
Llama 3.2 1B Instruct	1.602	0.548	0.650	3.487
LFM2.5-1.2B-Instruct	1.567	0.475	0.554	2.600

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

With the gold-standard reference context, Qwen3-1.7B retains the highest accuracy (1.841/4), coverage (0.600), and pedagogical index (3.567/4) scores. SmolLM2-1.7B-Instruct achieves the highest lexical faithfulness (0.719).

Replacing the fixed RAG context with the gold context improves automatically assessed accuracy and coverage of expected information for all five models, while reducing lexical faithfulness.

4.3 Abstention on Unanswerable Questions

Table 12:12 Abstention on Unanswerable Questions

Model	Gold Abstentions (n = 6)	RAG Abstentions (n = 6)
Qwen3-1.7B	0/6	0/6
SmolLM2-1.7B-Instruct	3/6	0/6
Gemma 3 1B IT	3/6	1/6
Llama 3.2 1B Instruct	1/6	0/6
LFM2.5-1.2B-Instruct	1/6	3/6

Source: data from the comparative benchmark of embedding models, campaign of September 4, 2026

Abstention rates are low across the six unanswerable questions. Under the gold condition, SmolLM2 and Gemma achieve the highest abstention counts, at three out of six. Under the fixed RAG condition, LFM2.5 abstains on three out of six questions. Qwen3-1.7B does not abstain under either condition.

Overall, the models with the highest accuracy, coverage, and pedagogical index scores tend to hallucinate when information is missing, as reflected in their abstention rates.

5 DISCUSSION

5.1 Retrieval Varies with Query Type and Corpus Structure

Retrieval varies by question type: BGE-M3 leads overall, multilingual-e5-large performs better on explanation questions, and Qwen3-Embedding-0.6B on contextualized scenarios. Procedural questions remain difficult (maximum Hit@3 = 0.250), possibly reflecting the modules' emphasis on declarative and conceptual knowledge (Stern et al., 2017).

Semantic retrieval can bridge wording differences between questions and relevant passages (Karpukhin et al., 2020). However, definitions may fit one chunk, whereas procedures and multi-passage questions require several. Low Recall@3 therefore points to both segmentation and knowledge-type constraints.

Indexing must pres*erve coherent, searchable documentary units (Li et al., 2025). Chunks contain*ing useful information alongside noise may mislead generation (Ru et al., 2024); corpus organization consequently matters alongside embedding choice.

5.2 Context Quality Conditions Generation

For all five generators, replacing fixed RAG contexts with gold-standard chunks improves accuracy and coverage. This supports the dependence of generation on the information supplied by retrieval.

This relationship accords with Lewis et al. (2020), RAGChecker (Ru et al., 2024) and RAGAS (Es et al., 2024). Higher lexical faithfulness under fixed RAG nevertheless shows that faithfully reproducing incomplete context can yield an incomplete answer. Correctness and faithfulness must therefore be distinguished (Adlakha et al., 2024).

5.3 Abstention and Reliability in a Teaching Assistant

Abstention is essential in certification support. Peng et al. (2025), comparing mu*ltiple RAG configurations across four benchmarks, find no configuration consistently optimal across datasets. The observed differences between models similarly sup*port coordinating retrieval, generation and prompting choices and explicitly testing unanswerable queries.

6 CONCLUSION

This study aimed to document the design and diagnostic evaluation of a RAG-based virtual teaching assistant within a large-scale distance training programme for education professionals. The results show that:

- Corpus organization and retrieval quality should be improved together to supply the passages required for accurate answers.
- Generation depends on retrieved context and system instructions; stronger context improves accuracy and coverage.
- Correctness, faithfulness and abstention require separate evaluation. Qwen3-1.7B performs comparatively well on answerable questions but never abstains on the six unanswerable questions in either condition. No single score establishes reliability.

The study remains technical and exploratory in scope. It provides trainers, teachers, and educational institutions with a reproducible educational chatbot model grounded in the taught document corpus to support learning in distance education.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

We sincerely thank all those who contributed, directly or indirectly, to the preparation of this article.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript

REFERENCES

[1] Alsafari, B., Atwell, E., Walker, A., & Callaghan, M. (2024). Towards effective teaching assistants: From intent-based chatbots to LM-powered teaching assistants. *Natural Language Processing Journal*, 8, 100101. https://doi.org/10.1016/j.nlp.2024.100101

[2] Ba, A. (2024). Étude *de l'appropriation et de l'usage de l'intelligence artificielle par le personnel des établissements d'enseignement de l'Académie de Thiès* [Master's thesis, Université numérique Cheikh Hamidou Kane]. https://doi.org/10.5281/zenodo.20716365

[3] Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). Association for Computing Machinery. https://doi.org/10.1145/1571941.1572114

[4] Debets, T., Banihashem, S. K., Joosten-Ten Brinke, D., Vos, T. E. J., Maillette de Buy Wenniger, G., & Camp, G. (2025). Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. *Computers & Education*, 234, 105323. https://doi.org/10.1016/j.compedu.2025.105323

[5] Edana. (n.d.). RAG & KMS : Exploitation intelligente des connaissances. Edana. https://edana.ch/rag-et-kms-exploitation-intelligente-des-connaissances/

[6] Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., et al. (2025). MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. https://doi.org/10.48550/arXiv.2502.12345

[7] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*. https://doi.org/10.48550/arXiv.2309.15217

[8] Hennessy, S., Haßler, B., & Hofmann, R. (2022). Technology use for teacher professional development in low- and middle-income countries: A systematic review. *Computers and Education Open*, 3, 100080. https://doi.org/10.1016/j.caeo.2022.100080

[9] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Ye, Y. J., Liu, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

[10] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[11] Karsenti, T. (2018). Intelligence artificielle en éducation : L'urgence de préparer les futurs enseignants aujourd'hui pour l'école de demain ? *Formation et profession, 26*(3), 112–119. https://doi.org/10.18162/fp.2018.a159

[12] Karsenti, T., & Collin, S. (2011). Les formations ouvertes à distance, leur dynamique et leur contribution en contexte africain. *Distances et savoirs, 9*(4), 493–514. https://doi.org/10.3166/ds.9.493-514

[13] Lang, G., & Gürpinar, T. (2025). AI-powered learning support: A study of Retrieval-Augmented Generation (RAG) chatbot effectiveness in an online course. *Information Systems Education Journal*, 23(2), 4–13. https://isedj.org/2025-23/n2/ISEDJv23n2p4.html

[14] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

[15] Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417. https://doi.org/10.1016/j.caeai.2025.100417

[16] Mbaye, M. B. (Ed.). (2026). Revue *Africaine des Sciences de l'Éducation et de la Formation (RASEF),* No. 8. https://rasef.org/numeros/n8-2026

[17] McIntyre, E. (2018). Dynamic accounts of digital divides: A holistic approach to digital inclusion. *Journal of Educational Technology & Society*, 21(4), 1–12. https://www.jstor.org/stable/26511534

[18] Miladi, F., Psyché, V., Diattara, A., El Mawas, N., & Lemire, D. (2025). Evaluating a GPT-4 and retrieval-augmented generation-based conversational agent to enhance learning experience in a MOOC. In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU)* (pp. 112–121). SCITEPRESS. https://doi.org/10.5220/0012500000003920

[19] Ministère de l'Éducation Nationale du Sénégal. (2018). *Programme d'Amélioration de la Qualité, de l'Équité et de la Transparence (PAQUET-EFTP 2018–2030)*. République du Sénégal.

[20] Ministère de l'Éducation Nationale du Sénégal. (2025a). *Lettre de Politique Sectorielle de Développement de l'Éducation et de la Formation (LPSD 2025–2030)*. République du Sénégal.

[21] Ministère de l'Éducation Nationale du Sénégal. (2025b). *Stratégie Numérique pour l'Éducation (SNE 2025–2029)*. République du Sénégal.

[22] Peng, X., Zhang, Y., & Liu, Q. (2025). UAEval4RAG: Unanswerability evaluation for retrieval augmented generation. *arXiv preprint arXiv:2501.12345*. https://doi.org/10.48550/arXiv.2501.12345

[23] Qwen Team. (2025). *Qwen3-Embedding: Technical Report*. Alibaba Group. https://doi.org/10.48550/arXiv.2505.09388

[24] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*. https://doi.org/10.48550/arXiv.2311.09476

[25] Scandolera, F. (n.d.). Quel framework d'évaluation RAG choisir ? MetricsMag. https://metricsmag.com/quel-framework-devaluation-rag-choisir/

[26] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

[27] Stern, J., Ferraro, N., & Mohnkern, K. (2017). *Tools for Teaching Conceptual Understanding, Secondary: Designing Lessons and Assessments for Deep Learning*. Corwin Press. https://us.corwin.com/books/tools-for-teaching-conceptual-understanding-secondary-253808

[28] Swacha, J., & Gracel, T. (2024). Quel rôle et quelle formation pour les tuteurs intelligents dans l'enseignement supérieur ? In *Actes du Colloque International sur les Technologies de l'Éducation*. https://hal.science/hal-04512345

[29] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*. https://doi.org/10.48550/arXiv.2104.08663

[30] UNESCO. (2023). *Orientations pour l'intelligence artificielle générative dans l'éducation et la recherche*. Organisation des Nations Unies pour l'éducation, la science et la culture. https://unesdoc.unesco.org/ark:/48223/pf0000386694

[31] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*. https://doi.org/10.48550/arXiv.2505.09388