Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 5 n.3 (2025)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Confidentiality-Preserving Retrieval-Augmented Generation over Educational Resources

Ludovico Boratto [1], Francesco Congiu [1], Gianni Fenu[1], Mirko Marras [1,*]

[1] Department of Mathematics and Computer Science, University of Cagliari
{ludovico.boratto, fenu, mirko.marras}@unica.it, f.congiu38@studenti.unica.it
* Correspondence: mirko.marras@unica.it;

Abstract: The evolution of Large Language Models (LLMs) has profoundly transformed access to educational resources, enabling interaction through natural language and fostering new learning modalities. However, the direct use of such models in educational settings raises critical issues related to response verification and the management of sensitive content. In this context, the Retrieval-Augmented Generation (RAG) paradigm represents a promising solution, as it combines the generative capabilities of LLMs with document retrieval from pre-processed knowledge bases, ensuring transparency and source traceability. This work presents a RAG framework designed for education, which integrates retrieval, fusion, and access-control modules to generate reliable responses that comply with users' authorization levels. Experimental results show that the system enhances the clarity and relevance of the retrieved resources while preserving both equity and confidentiality.

Keywords: educational content retrieval; learning management; information retrieval; large language models; retrieval-augmented generation; digital learning.

1. Introduction

In recent years, within educational platforms, the role of Information Retrieval (IR) has become crucial, as it allows students and teachers to access documents relevant to specific keywords entered in a search bar. Its main limitation, however, lies in the fact that it requires users to employ highly technical language to formulate effective queries, often making the process less intuitive for non-expert users. Conversely, the introduction of Large Language Models (LLMs) has profoundly changed the way students, teachers, and institutions interact with data. Their ability to understand and generate natural language has enhanced cognitive accessibility to educational resources, reducing the linguistic and conceptual barriers that traditionally characterized search interfaces (Luckin et al., 2016; Zawacki-Richter et al., 2019). These advances have opened new opportunities for adaptive learning and personalized support, enabling inclusive forms of interaction with digital knowledge (Holmes et al., 2019).

However, the direct use of LLMs in educational contexts still raises significant challenges. The generation of responses that are not always verifiable, the risk of hallucinations, and the potential exposure of sensitive or unauthorized content (such as student data or teaching materials intended for staff only) call for careful consideration of how such systems are deployed. As highlighted by Hwang et al. (2020) and

Holmes & Tuomi (2022), the educational use of AI must be guided by ethical principles, transparency, and the protection of personal data. Artificial intelligence should therefore not be regarded merely as an automated content generator, but rather as a technology of mediation, capable of supporting learning processes in a responsible and pedagogically meaningful way.

In this regard, the Retrieval-Augmented Generation (RAG) paradigm represents a key development, as it integrates the IR phase with that of LLM. By grounding model outputs on documentary evidence retrieved from external collections, it improves factual consistency and makes sources traceable (Lewis et al., 2020). Subsequent work has shown that coupling retrieval with pre-training further strengthens this effect (Guu et al., 2020). Related architectures such as Fusion-in-Decoder confirm that retrieved evidence can make generated answers more reliable and easier to inspect (Izacard & Grave, 2021). In the educational domain, this means that a student can explore a concept and receive an explanation supported by authentic references, transforming search into a learning process based on exploration and verification.

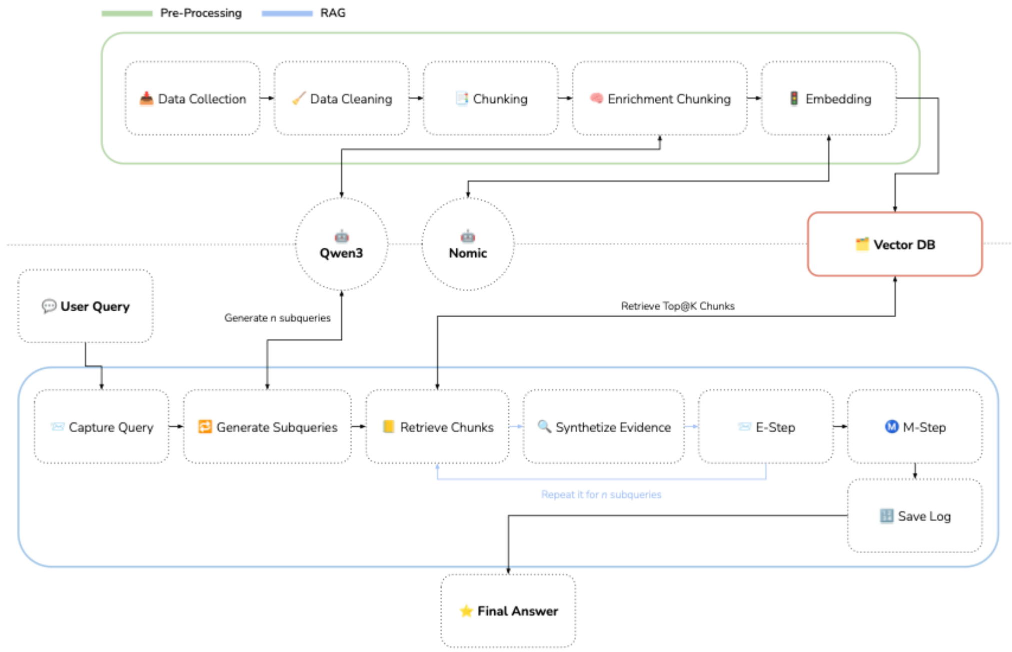
Despite the potential of this paradigm, current literature on RAG systems remains primarily focused on computational retrieval performance, overlooking the educational and ethical dimensions related to confidentiality and the reliability of responses. In a learning environment, not all documents can be made uniformly accessible: there are materials reserved for teaching staff, administrative or assessment data, and resources that require differentiated levels of authorization. The challenge, therefore, lies in designing confidentiality-aware RAG systems that combine informational accuracy with compliance to access policies, ensuring fairness, transparency, and the protection of sensitive data throughout teaching, learning, and administrative activities (Floridi & Cowls, 2021; European Commission, 2022).

In this article, we introduce a RAG-based framework tailored to the educational domain, integrating retrieval, fusion, and access-control strategies. The framework pursues a dual objective: (i.) enhancing the quality of responses through the combination of heterogeneous signals (retriever, reranker, and language models), and (ii.) ensuring that the retrieved content complies with the access levels assigned to each user role (student, teacher, tutor, or administrator). Through this work, we aim to explore the potential of this paradigm as a tool for delivering personalized and secure learning experiences, capable of ensuring and preserving the informational effectiveness of generated responses. Our broader vision aligns with that of scholars who view AI as a cultural and pedagogical mediator rather than a mere technological instrument (Maragliano, 2019; Selwyn, 2019), promoting learning experiences that are personalized, transparent, and ethically grounded.

2. Materials and Methods

The proposed framework was developed to explore the educational potential of the RAG paradigm, integrating retrieval, fusion, and access control techniques within a unified methodological pipeline. The primary objective was to analyze how generative systems grounded in document relevance can support personalized and reliable learning practices, while simultaneously ensuring confidentiality and protection against potential exposure to sensitive data. The framework consists of two operational modules: the first is related to data preparation and organisation, while the second concerns the functioning of the RAG system and the generation of responses.

Figure 1. The framework's architecture: high-level overview of the modules.



In the preliminary Pre-Processing phase (Figure 1), efforts were focused on adapting the reference corpus for educational use, ensuring semantic coherence and traceability of confidentiality levels. As a first step, corresponding to Data Collection, the FiQA (Financial Question Answering) dataset, part of the BEIR (Benchmark for Evaluation of Information Retrieval) benchmark (Thakur et al., 2021), was employed. It consists of 57.638 documents and 6.648 queries related to financial literacy education, chosen as a case study for its conceptual complexity and the presence of potentially sensitive content, similar to that found in real educational contexts. The data, already anonymized and publicly available, contain no personal information and therefore do not require ethical approval.

The second step, Data Cleaning, involved refining the documents and segmenting them into smaller, uniformly sized blocks with partial overlap to preserve contextual continuity. Next, in the Chunking and Enrichment phase, each block was augmented with textual metadata (e.g., title and summary) generated in zero-shot mode using a lightweight LLM (Qwen3-0.6B). During the Embedding stage, a confidentiality level was assigned to each document on a Likert scale (1–5), computed through an entailment-recognition classifier (BART-MNLI[1]). The resulting vector

[1] https://huggingface.co/facebook/bart-large-mnli

representations and metadata were then indexed in Qdrant[2], an open-source vector database enabling Approximate Nearest Neighbour (ANN) search. These operations complete the Pre-Processing phase illustrated in Figure 1.

As public datasets do not provide user-related information, the framework introduces a simulation of educational personas, each associated with an institutional role (student, teacher, tutor, or administrator) and an access level L_u. The confidentiality policies follow the rule $L_u \geq A_d$, where A_d denotes the document's confidentiality level. This approach enables the system to reproduce realistic scenarios of differentiated access to digital resources, common in Learning Management Systems (LMSs).

Once the system's knowledge base had been established, the second phase, the RAG module, which constitutes the core operational layer of the framework, was initiated. The process begins with Capture Query, in which the user's input is collected and prepared for semantic expansion. Each query is then decomposed in the Generate Subqueries stage into three sub-queries produced by Qwen3-0.6B, with the aim of broadening semantic coverage and fostering exploratory search. This mechanism, inspired by the Expectation–Maximization (E-M) algorithm, enables the system to evaluate multiple informational trajectories, thereby capturing different semantic aspects of the same request. This corresponds to the E-Step in the workflow, where multiple candidate evidences are scored and compared, followed by an M-Step that selects the best trajectory.

During the Retrieve Chunks phase, document retrieval is performed using the Nomic text encoder, while Re-Ranking is managed by a Cross-Encoder Reranker (CE), which analyses query–document relevance with higher semantic precision (Nussbaum et al., 2024). The two scoring sources are then combined through two fusion strategies: Linear Fusion (LF), which integrates normalized scores with a balancing factor of α = 0.9, and Reciprocal Rank Fusion (RRF), used as a comparative baseline.

After scoring, the system applies a Confidentiality Filter to exclude documents that do not comply with the user's access profile, followed by the Backfill Safe-Aware mechanism that explores beyond the top-K retrieval threshold, reintegrating authorized documents that were not initially selected. This ensures a balance between precision and recall while maintaining adherence to access-control policies.

Finally, in the Final Answer stage, the selected evidences are summarized by a LLM into a coherent and verifiable explanatory text, representing the system's final output to the user. This design mitigates hallucination risk, promotes transparency, and ensures traceability of sources, key elements for the responsible application of artificial intelligence in educational contexts.

2 https://qdrant.tech/

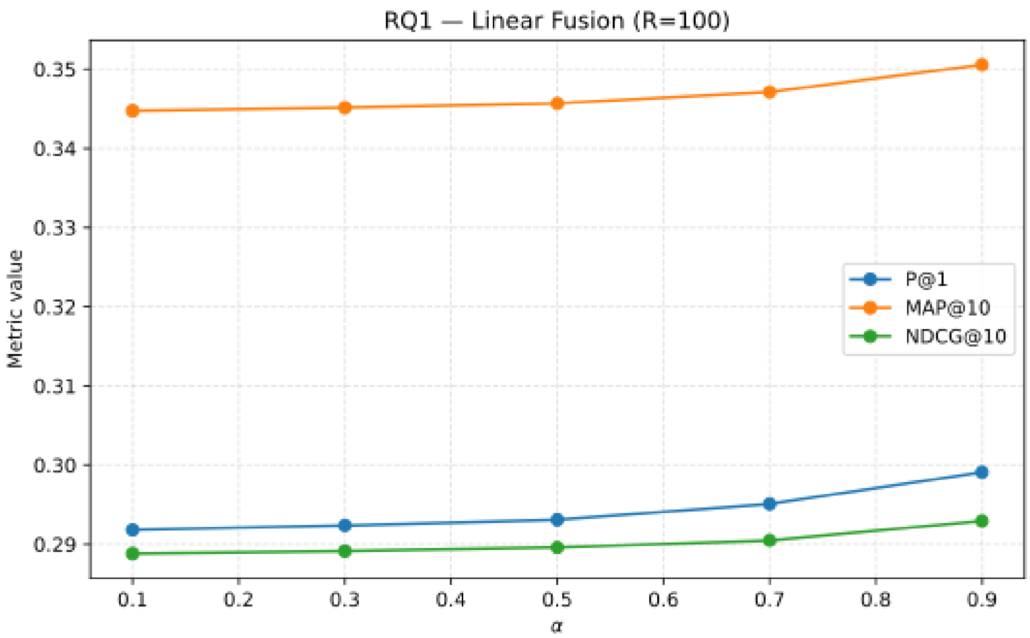
3. Results

The experimental validation of the framework aimed to assess the system ability to deliver informational content that was both educationally relevant, with respect to the documents stored in the knowledge base, and compliant with the selected data confidentiality policies (Lewis et al., 2020).

Two main research dimensions were analyzed. The first research question (RQ1) investigates the extent to which the combination of multiple information sources, such as dense retrievers, cross-encoders, and fusion mechanisms, can enhance the semantic relevance and expository clarity of the resources provided to students, thereby fostering a deeper understanding of the concepts being learned. On the other hand, the second research question (RQ2) explores the effectiveness of control and substitution mechanisms for unauthorized content in ensuring equitable access and the protection of sensitive data, guaranteeing that each user can benefit from complete and secure learning experiences consistent with their level of authorization and without any risk of exposure to inaccessible materials.

3.1. RQ1 – Information relevance and clarity of the retrieved resources

Figure 2. Trend of the Linear Fusion (LF) strategy across different values of the weighting parameter α on the considered dataset.



The first research question assessed the impact of integrating heterogeneous retrieval signals on the quality and clarity of the materials presented to users. As shown in Table 1, the LF strategy, which combines the scores produced by the dense retriever and the cross-encoder through a weighted scheme, consistently outperforms both single models (Dense and CE) and the simpler RRF. In particular, LF achieves the highest values for P@1 (0.2991) and MAP@10 (0.3506), indicating that the weighted combination improves ranking precision and overall relevance. While CE attains the best HIT@10 (0.5878), LF offers a more balanced trade-off between precision and coverage, generating clearer and more pedagogically useful responses.

These findings are further detailed in Table 2, which reports an ablation study on the fusion parameter α. The progressive increase of α (from 0.10 to 0.90) corresponds to a monotonic improvement across all metrics, confirming that greater reliance on the reranker enhances semantic coherence and content clarity. This trend is visually summarized in Figure 2, where the curve illustrates the positive slope of LF performance as α approaches 0.9. From a practical perspective, this approach reduces

Table 1. RQ1: performance comparison among Dense, CE, LF, and RRF. Best results are highlighted in **bold**.

Metric	Dense	CE	LF	RRF
P@1	0.2169	0.2884	**0.2991**	0.2169
MAP@10	0.2730	0.3397	**0.3506**	0.2730
NDCG@10	0.2250	0.2808	**0.2929**	0.2250
HIT@10	0.4498	**0.5878**	0.5247	0.4498

Table 2. RQ1: ablation study on Linear Fusion across different values. Best results are highlighted in **bold**.

Metric	α = 0.10	α = 0.30	α = 0.50	α = 0.70	α = 0.90
P@1	0.2918	0.2924	0.2931	0.2951	**0.2991**
MAP@10	0.3448	0.3452	0.3452	0.3471	**0.3506**
NDCG@10	0.2888	0.2891	0.2891	0.2905	**0.2929**
HIT@10	0.5200	0.5202	0.5202	0.5213	**0.5247**

redundancy among the retrieved documents, generating responses that are more readable and comprehensible, particularly for students dealing with complex or specialized topics. In contrast, RRF, while offering a simpler and more computationally efficient fusion approach, exhibits lower performance: the ranking based solely on document positions tends to dilute the contribution of the reranker, prioritizing coverage over precision.

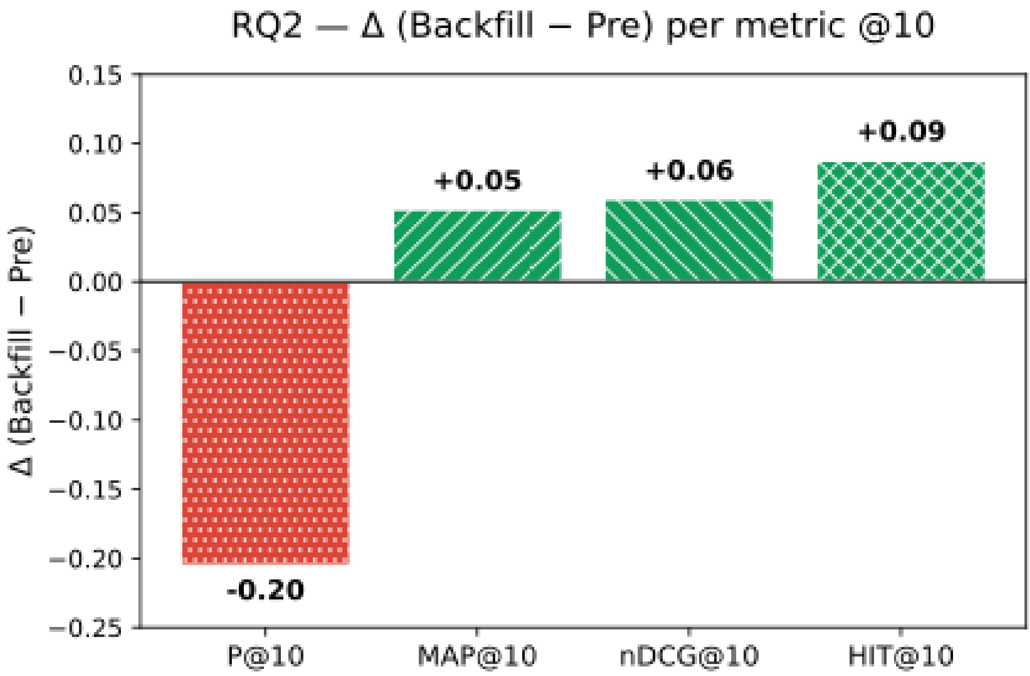
3.2. RQ2 – Equity of access and protection of confidential data

The second research question evaluates the ability of the Backfill Safe-Aware mechanism to preserve equitable access under confidentiality constraints. In educational settings, restrictive access policies can significantly narrow the set of materials available to certain users, reducing learning opportunities for those with lower authorization levels. The proposed approach addresses this limitation by automatically replacing unauthorized documents with relevant, authorized alternatives, thereby maintaining the overall informational coherence of the retrieved set. Specifically, the

Backfill process extends the search depth beyond the predefined top-K threshold, exploring additional candidates that satisfy the user's access level ($L_u \geq A_d$).

As shown in Figure 3, this strategy produces a measurable change in retrieval performance. After applying Backfill, HIT@10 and nDCG@10 increase, meaning that a larger number of queries retrieve at least one relevant and authorized document, and that these relevant items appear in higher ranking positions. Conversely, a moderate decrease in P@10 reflects a trade-off: deeper exploration introduces a few less relevant documents, reducing precision while improving coverage and inclusiveness.

Figure 3. Change in performance after applying the Backfill mechanism with respect to the Pre-filtering step corresponding to LF with α=0.9.



Overall, these results confirm that the Backfill mechanism maintains the pedagogical quality of retrieved resources while ensuring compliance with confidentiality constraints. This behavior reflects real-world educational dynamics, where the balance between information accessibility and data protection is essential for responsible and equitable AI-supported learning.

4. Discussion

The results highlight the possibility of integrating classical retrieval techniques and LLMs within an architecture tailored to educational needs, capable of balancing informational accuracy and data confidentiality. Overall, the empirical evidence confirms that the adoption of controlled fusion strategies and mechanisms for substituting unauthorised content makes it possible to overcome two of the main challenges associated with the use of artificial intelligence in educational contexts: the lack of transparency in generative processes and the management of sensitive data.

From a pedagogical perspective, the use of retrieval fusion strategies had a direct impact on the cognitive readability and expository clarity of the retrieved resources. The system's ability to integrate multiple information sources enabled it to provide responses that were more consistent with students' learning intentions (and thus with their queries), facilitating the understanding of complex concepts and reducing informational fragmentation. This evidence suggests that a properly designed RAG system can act as a cognitive scaffolding tool, capable of supporting knowledge construction processes and promoting more active and reflective learning.

In parallel, the management of confidentiality and access rights represents a key contribution to the development of responsible practices in the educational domain. The Backfill Safe-Aware mechanism showed that it is possible to ensure differentiated access to information without compromising educational equity. The system's ability to automatically replace unauthorised content with coherent and relevant alternatives preserved the overall pedagogical quality, fostering a balance between inclusion, transparency, and security. This result aligns with privacy-by-designprinciples and supports the ethical adoption of artificial intelligence in education.

From a methodological view, the proposed framework should not be interpreted merely as a technical framework, but as an educational mediation device capable of making AI decision-making processes visible and understandable. In this regard, the system contributes to strengthening artificial intelligence literacy, that is, the critical awareness of students and teachers on functioning, limitations, and implications of generative technologies. Understanding how the responses are constructed from authorised documents and filtering mechanisms enhances the ability to evaluate the credibility of sources and to use intelligent tools in a responsible manner.

It is, however, important to acknowledge certain limitations of this study. The experiments were conducted in a simulated environment, using public datasets and synthetic personas, and do not yet involve the direct participation of real users. Although this configuration allows for a controlled analysis of the effects of individual framework components, further research will be necessary to validate its effectiveness in authentic educational scenarios (for instance, through offline evaluations with real users). In the future, the integration of the proposed framework into real LMSs could provide richer data on user interactions, paving the way for qualitative assessments and analyses of the framework's impact on teaching and learning practices.

5. Conclusions

The study presented an innovative approach to the integration of IR and LLMs within the educational domain, with the aim of combining informational effectiveness, inclusiveness, and data confidentiality. Indeed, the proposed framework demonstrated its ability to improve the clarity and relevance of the responses provided to users, offering content consistent with learning objectives and compliant with the access policies defined by user roles. The evidence collected confirms the feasibility of a RAG paradigm that not only optimizes the quality of retrieved information but also incorporates principles of security and ethical responsibility in data management.

From an educational perspective, the proposed framework opens new avenues for the design of learning support systems capable of adapting to students' needs and ensuring a transparent and secure learning experience. Its modular architecture allows for scalability across different disciplinary domains (from *economics* to *linguistics and the sciences*) promoting a flexible and transferable approach. In the future, the framework could be integrated into existing Learning Management Systems (LMSs) to provide adaptive and personalised support during study activities, incorporating components of explainability and adaptive feedback.

The next stages of development will include the introduction of specific educational metrics capable of evaluating not only informational accuracy but also the communicative clarity and educational value of the generated responses. In parallel, experimentation with real samples of students and teachers will make it possible to measure the system's actual impact on the level of trust towards artificial intelligence in educational contexts. Finally, the use of larger and more diverse datasets may further refine the framework's performance and strengthen its generalizability across domains.

In conclusion, it can be stated that the proposed framework represents a concrete step towards an ideal model of AI for Education, founded on transparency, equity, and responsibility, in which technology does not replace the educational process but becomes its conscious, ethical, and cognitively empowering ally.

References

European Commission. (2022). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. *Harvard Data Science Review, 3*(1). https://doi.org/10.1162/99608f92.8cd550d1

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval-augmented language model pre-training. *In Proceedings of the 37th International Conference on Machine Learning (ICML 2020)* (pp. 3929–3938). PMLR. http://proceedings.mlr.press/v119/guu20a.html

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://curriculumredesign.org/our-work/artificial-intelligence-in-education/

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence, 1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the ACL (EACL 2021)* (pp. 874–880). https://aclanthology.org/2021.eacl-main.74

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. *In Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson. https://discovery.ucl.ac.uk/1475756/

Maragliano, R. (2019). *Educazione digitale e intelligenza artificiale*. FrancoAngeli. https://www.francoangeli.it/Libro/Educazione-degli-adulti,-sviluppo-di-comunità,-pedagogica-critica-Angela-Zucconi-e-il-Progetto-Pilota-Abruzzo?Id=25378

Nussbaum, Z., Morris, J. X., Duderstadt, B., & Mulyar, A. (2024). *Nomic Embed: Training a reproducible long-context text embedder*. arXiv. https://arxiv.org/abs/2402.01613

Selwyn, N. (2019). *Should robots replace teachers? AI and the Future of Education*. (1st ed.) Polity Press.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks Track (Round 2).* https://openreview.net/forum?id=wCu6T5xFjeJ

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education, 16*, 39. https://doi.org/10.1186/s41239-019-0171-0