Accepted and presented at the 2nd Workshop on Artificial Intelligence with and for Learning Sciences (WAILS '25). To appear in the post-proceedings of the Lecture Notes in Computer Science (LNCS) series, Springer.

Confidential Retrieval-Augmented Generation in Educational Contexts

Ludovico Boratto[1 [0000−0002−6053−3015]], Francesco Congiu[1,2 [0009−0000−7821−0663]], Gianni Fenu[1 [0000−0003−4668−2476]], Giacomo Medda[1 [0000−0002−1300−1876]], and Antonello Pau[1]

[1] University of Cagliari, Cagliari 09124, Italy
[2] University of Macerata, Macerata 62100, Italy
{ludovico.boratto,fenu,giacomo.medda,antonello.pau}@unica.it, f.congiu@unimc.it

Abstract. In recent years, the Retrieval-Augmented Generation (RAG) paradigm has become central to improving the reliability of systems based on Large Language Models (LLMs), as it grounds generation in evidence from external knowledge sources. However, much of the literature focuses almost exclusively on retrieval effectiveness, overlooking a crucial requirement in educational and professional settings: content confidentiality. The absence of mechanisms ensuring that only authorized documents are returned to the user risks limiting adoption in real scenarios. We present RetrievEM, a confidentiality-aware RAG framework validated on the BEIR/FiQA benchmark. Our approach pursues a dual objective: (i) improving retrieval by combining heterogeneous signals from different components, and (ii) ensuring that returned content complies with user-specific access constraints. RetrievEM integrates dense retrieval, reranking with cross-encoders, score-level fusion, and access-aware persona generation. Experimental results show that fusion yields substantial gains over individual components. Considering the limited accessible documents and RAG-related selection bias, we introduce Backfill, a post-processing algorithm that increases the search depth by exploring beyond the initial top-k results, preserving confidentiality without sacrificing retrieval utility. Overall, our RAG system can deliver pedagogically useful content while respecting access policies, demonstrating that effectiveness and confidentiality can coexist.

Keywords: Retrieval-Augmented Generation · Confidential Artificial Intelligence · Educational Technology

1 Introduction

Within modern information systems [3, 11, 6, 7, 2], Information Retrieval (IR) plays a key role in enabling users to access documents relevant to their queries. However, IR typically requires the formulation of precise, keyword-based queries, which can be challenging for non-technical users. Conversely, Large Language Models (LLMs) - a class of Transformer-based models capable of understanding and generating natural language - make information access more intuitive; yet their responses are non-deterministic and not always predictable or verifiable.

The combination of these two paradigms has given rise to Retrieval-Augmented Generation (RAG), which combines the retrieval precision with the generative fluency of LLMs. RAG systems enhance factual grounding and reduce hallucinations by incorporating evidence retrieved from external sources.

A major challenge in applying RAG to *educational contexts* concerns confidentiality: ensuring that information retrieval and generation processes comply with access-control and privacy requirements, preventing the exposure of sensitive materials to unauthorized users. For instance, a lecturer may use the system to retrieve access confidential assessment rubrics or internal teaching notes, whereas a student should only access publicly available learning materials. Similarly, academic administrators might retrieve aggregated analytics on student performance, while individual-level data should remain inaccessible. In financial literacy education (the domain of our experiments) these confidentiality issues become even more relevant, as educational resources may include proprietary datasets, students' financial scenarios, or graded assignments.

Developing a RAG system that is genuinely useful for education therefore requires careful *data governance*: managing access to different knowledge sources, protecting queries and responses in transit, and enforcing policies that prevent information leaks across roles (e.g., teacher, student, tutor). These safeguards ensure that the benefits of assistance do not come at the cost of data protection.

Despite the growing popularity of RAG in education, most studies have prioritized retrieval accuracy or generative fluency over privacy-aware design. Early works such as REALM [8] and RAG [10] established the paradigm of integrating retrieval with generation, while more recent methods like EXSEARCH [13] explored adaptive retrieval mechanisms. However, none of them natively address document confidentiality or access control, both crucial in learning environments.

Recent research has begun to address these limitations by proposing privacy-preserving retrieval and secure RAG architectures. For example, Zeng et al. [18] mitigate privacy risks using fully synthetic data in RAG pipelines, showing that such corpora can preserve retrieval utility while preventing sensitive information leakage. Chakraborty et al. [4] investigate Federated RAG, demonstrating how decentralized retrieval allows collaboration across institutions without sharing private data. Similarly, Cheng et al. [5] introduce RemoteRAG, a privacy-preserving cloud service ensuring secure query handling and access-controlled retrieval. Nonetheless, these privacy-oriented strategies have seldom been applied to *educational systems*, leaving open the question of how to design RAG pipelines that are both effective and confidentiality-preserving.

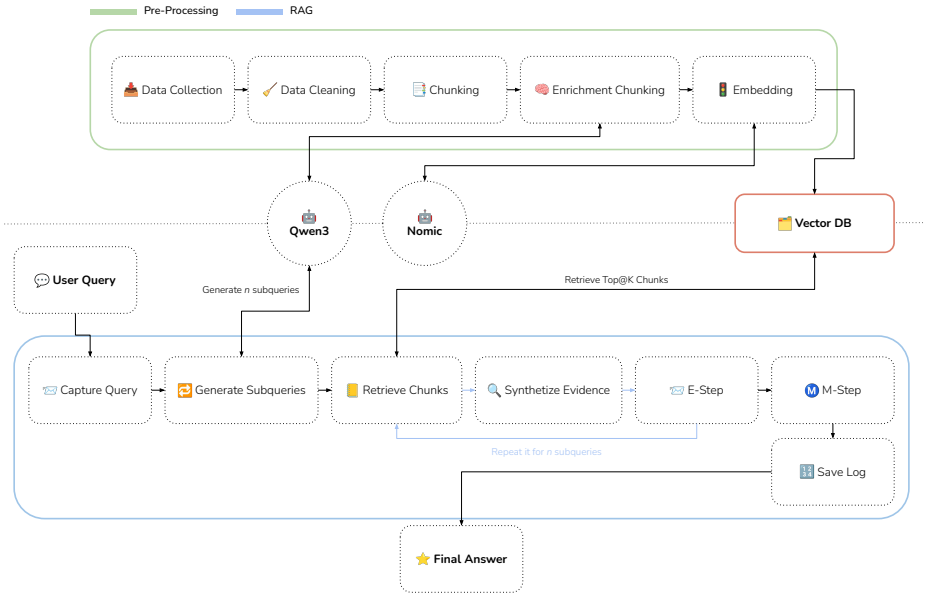

Fig. 1: RetrievEM architecture: high-level overview of core functional modules.

To explore this issue within a realistic domain, we focus on financial education—an area where both content and learner data are sensitive. Financial learning involves interpreting data-rich materials (e.g., market analyses, budget simulations) while ensuring that student queries, learning progress, and personal examples remain protected. As our reference dataset, we adopt the Financial Question Answering (FiQA) corpus from the BEIR framework [17], which enables controlled experimentation in a domain that mirrors authentic financial reasoning tasks.

In this work, we make three main contributions: (i) we examine the role of confidentiality in the design of RAG pipelines for educational support; (ii) we demonstrate how FiQA can serve as a secure yet realistic corpus for exploring confidentiality-aware RAG; and (iii) we outline learning scenarios and evaluation strategies showing how such systems can enhance financial information literacy while ensuring compliance with data protection requirements.

2 Methodology

In this section, we present the *Confidentiality-aware* architecture of RetrievEM and the design choices that guide the subsequent modules. Our goal is to maximize retrieval and generation effectiveness while managing the trade-offs and respecting access constraints, without resorting to fine-tuning base models.

2.1 Data Preparation

RAG systems aid the generation of reliable content through the retrieval of external knowledge. It is then crucial for such knowledge to be prepared before being fed into our architecture. Datasets for question answering typically consist of passages (documents to be retrieved), queries (user questions), and relevance judgments (qrels) indicating which documents are relevant. We follow common practices in the educational domain [15] and design a clear and reliable pipeline to ensure content is appropriate, coherent, and confidentiality-compliant:

- Cleaning: we apply a textual cleaning step by removing redundant headers, markup, extra spacing, and unwanted metadata.
- Chunking: we split passages into fixed-size segments with a 20% overlap between windows, ensuring better context coverage during retrieval.
- Chunk Enrichment: we augment each textual chunk through additional metadata to enable more accurate and less noisy retrieval. Specifically, metadata, such as title and summary, was generated with a compact LLM (*Qwen3-0.6B* [16]) in zero-shot mode to keep the process lightweight. Metadata also include the confidentiality levels associated with each document in the dataset.
- Embedding: we employ text encoders to embed chunks and their respective metadata into a latent representation. As the embedding process is crucial for the retrieval pipeline, we explore different solutions and assess their performance, focusing on state-of-the-art approaches.
- Ingestion: we use a vector database that supports large-scale approximate nearest neighbor search (ANN) to enable the document retrieval through a query. In particular, we relied on Qdrant[3], but RetrievEM can be easily adapted to any vector database. Embeddings are stored along with their respective confidentiality level to enable runtime filtering and ensure results respect access constraints and confidentiality policies.

2.2 Retrieval-Augmented Generation

RetrievEM builds on the steps of data preprocessing and ingestion in Qdrant, and it can operate in two main modes. The first is the *interactive mode*, in which the framework answers queries from students or lecturers with relevant content that complies with confidentiality constraints. In this scenario, the pipeline is designed to assign an access level to the user at runtime and manage the access policy dynamically. This component is currently under development and represents the natural evolution towards real-world system deployment.

The second is the *validation mode*, which enables systematic offline testing on benchmark queries. It is not common for corpora to provide confidentiality metadata at the query level. In light of this, RetrieveEM integrates an additional step of query enrichment that dynamically generates synthetic personas [14]. The process generates personas with an associated access level L_q linked to a query

[3] https://qdrant.tech/

q. Specifically, n_p queries from the corpus are sampled according to a uniform distribution and fed into an LLM (*Qwen3-0.6B* [16]) to generate n_p personas. The remaining queries are randomly assigned to these personas, simulating a set of queries performed by distinct users. This strategy enables confidentiality-aware analyses in setups lacking access-levels information.

Retrieval Stage Independently of the mode, RetrievEM integrates a query decomposition process to increase the semantic coverage of input queries. Each query q is converted into a set of sub-queries $\{q_i\}_{i=1}^{n_q}$, representing trajectories aimed at exploring sub-aspects of the search space [9]. We leverage the in-context learning ability of LLMs to perform this task by relying on the lightweight and fast *Qwen3-0.6*. Each generated sub-query q_i is then transferred to the vector database, which retrieves k candidate documents via approximate nearest neighbor (ANN) search. Although each retrieved document is associated with a similarity score, such a score is conditioned by the respective sub-query used for retrieval. Therefore, we employ reranking strategies to address the non-trivial task of selecting the k candidate documents that maximize the relevance for the initial query q.

Reranking Stage After dense retrieval, documents and queries are processed through a Cross-Encoder Reranker (CE) that reassess the relevance of retrieved documents. Unlike embedding-based similarity, the reranker jointly encodes the pair (q, d) (e.g., `[CLS] q [SEP] d [SEP]`), allowing it to capture more fine-grained semantic distitions. For example, two documents on compound interest may appear equivalent to the dense retriever, but the CE can recognize that one contains only a generic definition while the other provides a step-by-step explanation, resulting in more useful outputs in educational contexts. Although computationally more expensive, this step yields rankings where the most relevant and pedagogically rich documents receive more visibility and importance.

To integrate retriever and reranker signals, we adopt two strategies, namely Linear Fusion (LF) and Reciprocal Rank Fusion (RRF). LF combines normalized scores through a weighted average:

$$s_{\mathrm{LF}}(d \mid q;\, \alpha) \;=\; (1-\alpha)\, s_{\mathrm{dense}}(d \mid q) \;+\; \alpha\, s_{\mathrm{CE}}(d \mid q), \quad \alpha \in [0,1] \tag{1}$$

α is a factor that balances the retriever's and reranker's impact, with $\alpha = 0$ relying solely on the retriever and $\alpha = 1$ solely on the reranker. $\alpha \approx 0.5$ combines breadth (maximized by the retriever) and precision (maximized by CE).

On the other hand, RRF merges rankings by positions rather than scores:

$$s_{\mathrm{RRF}}(d \mid q) = \sum_{m \,\in\, \text{Dense, CE}} \frac{1}{\gamma + r_m(d)} \tag{2}$$

where *Dense* denotes the ANN retrieval, $r_m(d)$ is the position of document d in method m's ranking, and γ is a smoothing constant (typically $\gamma = 60$

[1]) that prevents the top ranks from excessively dominating the result. This approach boosts documents ranked highly by both models and provides a robust compromise, even under different scoring scales. For instance, a document placed 2nd by the retriever and 3rd by the CE scores higher than one ranked 1st by one model but 50th by the other.

Post-processing Stage After dense retrieval and reranking, the system applies a post-processing stage to ensure that the final response is accurate, confidentiality-aware, and pedagogically useful. The process unfolds in three steps.

First, a confidentiality filter ensures that retrieved documents comply with user access policies. Each document has a confidentiality level A_d and each user/persona associated with the respective queries has an access level L_u, with the policy requiring $L_u \geq A_d$. Hence, any document that is not compliant with the confidentiality policy is filtered out and ignored in later steps. Second, the authorized evidence chunks are synthesized into a coherent summary through an LLM (*Qwen3-0.6*). For each sub-query q_i, evidence chunks are aggregated into a summarized evidence v_i. In other words, the system combines and reorganizes relevant fragments, avoiding redundancy and linking related concepts. From an educational standpoint, this is equivalent to taking notes from multiple sources and rephrasing them into a clear, linear explanation.

Third, the synthesized candidates undergo a re-scoring and selection process. Each trajectory is defined as a (sub-query, evidence) pair $\tau_i = [q_i; v_i]$, with $[\cdot;\cdot]$ the concatenation operator. This concatenated structure is used to compute a new score by comparison with the query q, and, in validation scenarios, also to a reference answer g. Formally, the score is defined as:

$$s(\tau_i \mid q, g;\, \lambda) \;=\; \lambda \cdot \cos\big(\mathbf{e}(\tau_i),\, \mathbf{e}(q)\big) \;+\; \big(1-\lambda\big) \cdot \cos\big(\mathbf{e}(\tau_i),\, \mathbf{e}(g)\big), \qquad (3)$$

where cos denotes cosine similarity, $\mathbf{e}$ the embedding operator, and λ the trade-off parameter. In evaluation and interactive settings without g, the score reduces to $s(\tau_i) = \cos(\mathbf{e}(\tau_i), \mathbf{e}(q))$. The trajectory with the highest score is selected as the basis for the answer, ensuring both relevance and didactic adequacy.

Finally, the selected evidence is passed to an LLM (*Qwen3-0.6*) in a structured prompt. At this stage, the model does not hallucinate content but generates a fluent explanation grounded in verified material. The resulting output integrates accuracy (through retrieval and scoring), confidentiality (through access control), and clarity (through synthesis and structured generation).

3 Experimental Results

In this section, we present and discuss the experimental results obtained with RetrievEM, with the aim of validating the architectural choices and analysing the extent to which the framework can improve retrieval quality while simultaneously enforcing confidentiality constraints. Accordingly, this section is structured as follows: we first describe the *Experimental Setup*, detailing the adopted dataset,

the procedure for enriching it with confidentiality metadata and personas, and the evaluation metrics. We then report the results of our experiments, discussing key findings and aiming to answer the following research questions:

RQ1 How effective are fusion strategies compared to standard retrieval?
RQ2 What is the impact of the backfill safe-Aware mechanism?

3.1 Experimental Setup

Dataset To evaluate the performance of RetrievEM, we adopted the `FiQA` dataset, which is part of the BEIR benchmark [17] and widely used in the literature for *Financial Question Answering*. `FiQA` is characterized by its highly specialized domain (finance and economics), and provides not only a large document corpus but also a set of queries and corresponding relevance judgments (*qrels*). In particular, the collection consists of 57,638 documents and 6,648 queries, making it well-suited for evaluating IR systems. However, `FiQA` does not include confidentiality levels for documents and users (i.e., queries). Therefore, we estimate a confidentiality level A_d in a zero-shot setting using a pre-trained entailment recognition classifier[4] to assess the logical relationship between premises and hypotheses. Specifically, the model processes the passage text and assigns an integer value from 1 to J representing the confidentiality level, we set $J = 5$ to reflect a five-point Likert scale. In educational contexts, this allows us to distinguish between public content intended for students (levels 1–2) and specialized or sensitive materials aimed at teachers or domain experts (levels 4–5).

RAG Setup A key design choice is the selection of the textual encoder for the retrieval stage. Specifically, the selection process involved testing several state-of-the-art encoders, and *Nomic* [12] emerged as the most effective model for the financial domain. We set $n_p = 10$ for persona generation, $n_q = 3$ for query decomposition (each input query is expanded into three sub-queries generated by *Qwen3*), and $\gamma = 60$ for RRF smoothing. Unless otherwise indicated, we set the number of retrieved documents to 100.

Metrics For the performance analysis, we adopted a set of classical IR and more recent dimensions related to confidentiality. The former include Precision@k (P@k), which measures the proportion of relevant documents among the top k results; MAP@k, which computes the mean of the cumulative precision values at the ranks where relevant documents occur; and nDCG@k, which assesses the overall ranking quality by penalizing relevant documents retrieved at lower ranks; Hit@k, which represents the probability of finding at least one relevant document within the top k positions.

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

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

Metric	$\alpha = 0.10$	$\alpha = 0.30$	$\alpha = 0.50$	$\alpha = 0.70$	$\alpha = 0.90$
P@1	0.2918	0.2924	0.2931	0.2951	0.2991
MAP@10	0.3448	0.3452	0.3457	0.3471	0.3506
NDCG@10	0.2888	0.2891	0.2896	0.2905	0.2929
HIT@10	0.5200	0.5202	0.5205	0.5213	0.5247

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

3.2 RQ1 - Fusion Strategies Effectiveness

The first Research Question investigates the impact of different fusion strategies within the RetrievEM framework. The underlying intuition behind this analysis is that heterogeneous signals, such as the scores produced by the dense retriever and the cross-encoder, may be *complementary*: the former provides coverage, while the latter offers higher ranking precision. The question, therefore, is whether a weighted combination of these signals can overcome the limitations of the individual models, and to what extent simpler methods (e.g., RRF) can compete with more sophisticated approaches.

The analysis begins by comparing the proposed technique against the main baselines. As shown in Table 1, *LF* outperforms both *Dense* and *CE* on *P@1*, *MAP@10*, and *nDCG@10*. In particular, *P@1* increases from 0.2884 (CE) to 0.2991 and *MAP@10* from 0.3397 to 0.3506, highlighting the benefits of a weighted combination: the retriever contributes coverage, while the re-ranker improves precision in the top ranks. For *HIT@10*, CE achieves the highest value (0.5878), while fusion remains competitive (0.5247), suggesting that the method prioritizes quality in the top positions over broad coverage.

A relevant comparison is with *RRF*, a simple but widely adopted method for combining heterogeneous models. In our setting, *RRF* exhibits clearly inferior performance compared to *LF* across all metrics, essentially matching the dense retriever. This behavior suggests that the reranker's contribution is diluted in the *RRF* combination, likely because the dense retriever dominates the ranking. These findings are consistent with previous observations that *RRF* tends to be competitive only when sources provide partially complementary signals.

The performance of *LF* in Table 1 refers to the configuration with the fusion weight yielding best results, according to an ablation study conducted on a validation set by varying α. Table 2 shows the effect of this study on the test set, with a clear monotonic improvement across all metrics as the contribution of the re-ranker increases and $\alpha = 0.90$ consistently emerging as the best configuration.

3.3 RQ2 - Confidentiality Preservation

The second RQ examines the effectiveness of the Backfill mechanism in mitigating key structural limitations that arise in retrieval scenarios subject to confidentiality constraints, without excessively compromising performance.

Two main factors motivate this analysis: (i) Reduction of the accessible catalog — when strict access policies are enforced, the set of documents actually available to each user is drastically reduced. This increases the risk of failing to retrieve relevant content simply because it is not authorized. (ii) Selection bias introduced by top-k — in RAG systems, the choice of the parameter k (i.e., the number of documents retrieved in the initial stage) determines which candidates are passed to the reranker. A fixed k value introduces a structural bias: if relevant documents do not appear in the initial top-k, they will never be considered, even if potentially important.

The Backfill mechanism directly addresses these two issues. After applying *confidentiality filtering* (which ensures full compliance with access policies $L \geq A$), Backfill increases the search depth by exploring beyond the initial top-k results and re-ranking these additional candidates using the reranker. This deeper exploration reaches documents that may be difficult to surface directly through the user's query, but which are both relevant and authorized as well.

Figure 2 clearly illustrates the impact of this strategy on retrieval metrics. The use of Backfill leads to an increase in HIT@10, indicating that a larger number of queries are able to retrieve at least one relevant document. At the same time, an improvement in NDCG@10 can be observed, as relevant documents are promoted to higher and therefore more visible positions in the final ranking. Conversely, P@10 tends to decrease, since deeper exploration also introduces fewer relevant documents, reducing the average number of relevant results per query. Overall, these findings show that Backfill introduces a controlled trade-off between precision and coverage: by extending the search depth, the system can satisfy more queries and improve the visibility of relevant documents, at the cost of a moderate reduction in precision. This behavior reflects realistic educational scenarios in which confidentiality constraints significantly restrict the space of accessible documents, and deeper search becomes essential to preserve the pedagogical utility of the system.

4 Conclusions and Future Work

This work introduced RetrievEM, a *Confidentiality-Aware* RAG framework designed for educational contexts and validated on the `BEIR/FiQA` benchmark. The

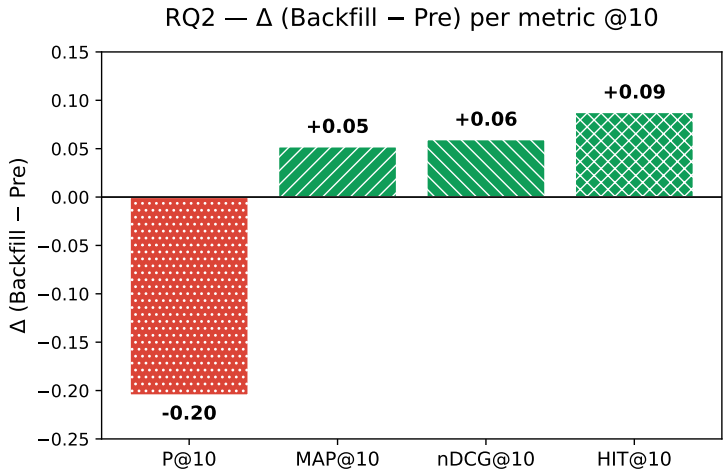

Fig. 2: RQ2. Change in performance after applying the *Backfill* mechanism with respect to the *Pre*-filtering step (corresponding to LF with $\alpha = 0.9$).

system demonstrates that it is possible to combine retrieval effectiveness with strict access-control constraints, providing accurate and pedagogically useful responses while preserving confidentiality.

In terms of effectiveness, fusion strategies proved superior to individual components: Linear Fusion with $\alpha = 0.9$ achieved the best performance, outperforming both dense retriever and cross-encoder taken individually, and surpassing that achieved by reciprocal rank fusion. On the confidentiality side, enforcing access policies inevitably reduces the pool of accessible documents while amplifying the bias introduced by top-k retrieval. To mitigate this, we proposed the Backfill strategy, which extends the search depth beyond the initial top-k results and re-ranks additional candidates. This approach improves *Hit* and *NDCG*, ensuring higher coverage and better visibility of relevant documents, but it reduces P@10 due to broader inclusion. In short, Backfill trades precision for coverage, providing a practical balance between strict access enforcement and utility of the retrieved results.

The study also highlights several limitations. First, performance depends heavily on the embedding space, with results and rankings sensitive to the choice of model. Moreover, gains from reranking and query decomposition are not uniform across all queries, exposing variability across queries. Computational costs also increase substantially when incorporating multiple strategies after the retrieval stage, such as fusion-based scoring and the Backfill mechanism.

Looking ahead, future work will focus on adaptive policies that dynamically select retrieval, fusion, and backfill depth based on query characteristics, developing cost-aware mechanisms that optimize efficiency. We also plan to extend

the evaluation to educational datasets with real confidentiality constraints and to conduct user studies assessing the framework's pedagogical value and usability.

Acknowledgments. We acknowledge financial support from the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.1 - Call for tender No. 3277, published on December 30, 2021, by the Italian Ministry of University and Research (MUR), funded by the European Union – Next Generation EU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – Grant Assignment Decree No. 1056 adopted on June 23, 2022, by the MUR (CUP F53C22000430001)

References

1. Benham, R., Culpepper, J.S.: Risk-reward trade-offs in rank fusion. In: Proceedings of the 22nd Australasian Document Computing Symposium. ADCS '17, Association for Computing Machinery, New York, NY, USA (2017), https://doi.org/10.1145/3166072.3166084
2. Boratto, L., Fabbri, F., Fenu, G., Marras, M., Medda, G.: Counterfactual graph augmentation for consumer unfairness mitigation in recommender systems. In: Frommholz, I., Hopfgartner, F., Lee, M., Oakes, M., Lalmas, M., Zhang, M., Santos, R.L.T. (eds.) Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023. pp. 3753–3757. ACM (2023). https://doi.org/10.1145/3583780.3615165, https://doi.org/10.1145/3583780.3615165
3. Boratto, L., Fenu, G., Marras, M., Medda, G.: Practical perspectives of consumer fairness in recommendation. Inf. Process. Manag. **60**(2), 103208 (2023). https://doi.org/10.1016/J.IPM.2022.103208, https://doi.org/10.1016/j.ipm.2022.103208
4. Chakraborty, A., Dahal, S., Gupta, H.: Federated retrieval-augmented generation: A systematic mapping study. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 4521–4535 (2025), https://aclanthology.org/2025.findings-emnlp.388
5. Cheng, Z., Zhang, Z., Wang, Y., Yuan, C., Yao, J.: Remoterag: A privacy-preserving llm cloud rag service. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 2341–2355 (2025), https://aclanthology.org/2025.findings-acl.197
6. Dessí, D., Dragoni, M., Fenu, G., Marras, M., Reforgiato Recupero, D.: Deep learning adaptation with word embeddings for sentiment analysis on online course reviews. In: Deep learning-based approaches for sentiment analysis, pp. 57–83. Springer (2020)
7. Fenu, G., Galici, R., Marras, M., Recupero, D.R.: Exploring student interactions with AI in programming training. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024, Cagliari, Italy, July 1-4, 2024. ACM (2024). https://doi.org/10.1145/3631700.3665227, https://doi.org/10.1145/3631700.3665227
8. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: Proceedings of the 37th International Conference

on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 3929–3938. PMLR (2020), http://proceedings.mlr.press/v119/guu20a.html
9. Huang, J., Wang, M., Cui, Y., Liu, J., Chen, L., Wang, T., Li, H., Wu, J.: Layered query retrieval: An adaptive framework for retrieval-augmented generation in complex question answering for large language models. Applied Sciences **14**(23) (2024), https://www.mdpi.com/2076-3417/14/23/11014
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
11. Medda, G., Fabbri, F., Marras, M., Boratto, L., Fenu, G.: GNNUERS: fairness explanation in gnns for recommendation via counterfactual reasoning. ACM Trans. Intell. Syst. Technol. **16**(1), 6:1–6:26 (2025). https://doi.org/10.1145/3655631, https://doi.org/10.1145/3655631
12. Nussbaum, Z., Morris, J.X., Duderstadt, B., Mulyar, A.: Nomic embed: Training a reproducible long context text embedder (2024)
13. Shi, Z., Yan, L., Yin, D., Verberne, S., de Rijke, M., Ren, Z.: Iterative self-incentivization empowers large language models as agentic searchers. CoRR **abs/2505.20128** (2025), https://doi.org/10.48550/arXiv.2505.20128
14. Shin, J., Hedderich, M.A., Rey, B.J., Lucero, A., Oulasvirta, A.: Understanding human-ai workflows for generating personas. In: Proceedings of the 2024 ACM Designing Interactive Systems Conference. p. 757–781. DIS '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3643834.3660729
15. Takagi, S., Yamauchi, R., Kumagai, W.: Towards autonomous hypothesis verification via language models with minimal guidance (2023), https://arxiv.org/abs/2311.09706
16. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
17. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021), https://openreview.net/forum?id=wCu6T5xFjeJ
18. Zeng, J., Zhang, Z., Li, Y., Li, W., Zhang, J.: Mitigating the privacy issues in retrieval-augmented generation (rag) via pure synthetic data. arXiv preprint arXiv:2406.14773 (2024), https://arxiv.org/abs/2406.14773