Large Language Models for Academic Internal Auditing

Houda CHAMMAA[1], Rachid ED-DAOUDI[2], Khadija BENAZZI[3]
Faculty of Economics-Law and Social Sciences, Cadi Ayyad University, Marrakech, Morocco[1]
LyRICA: Laboratory of Research in Computer Science, Data Sciences and Artificial Intelligence,
School of Information Sciences, B.P. 604, Rabat-Instituts, Rabat, Morocco[2]
Innovation-Responsibility and Sustainable Development Laboratory-INREED,
Cadi Ayyad University, Marrakech, Morocco[3]

***Abstract*—This research examines the application of Artificial Intelligence in internal auditing, focusing on document management and information retrieval in academic institutions. The study proposes using Large Language Models to streamline document processing during audit preparation, addressing inefficiencies in traditional document handling methods. Through experimental evaluation of three embedding models (BGE-M3, Nomic-embed-text-v1, and CamemBERT) on a dataset of 300 academic regulatory queries, the research demonstrates BGE-M3's superior performance with an nDCG3 score of 0.90 and top-1 accuracy of 82.5%. The methodology incorporates query expansion using GPT-4 and Llama 3.1, revealing robust performance across varied query formulations. While highlighting AI's potential to transform internal auditing practices, particularly in Morocco's academic sector, the study acknowledges implementation challenges including institutional constraints and resistance to technological change. The conducted experiments and result analysis provide useful criteria that can be applied to similar information retrieval challenges in other fields and real-world applications.**

Keywords—Large language models; internal auditing; information retrieval; embedding models; academic institutions

I. Introduction

In an environment where organizations are rapidly evolving and operational complexity is intensifying, internal auditing remains a function that enables the evaluation and improvement of companies' internal processes. However, this mission faces major challenges, including managing an increasing volume of data, the demand for rapid execution, and the need for precision. The emergence of artificial intelligence (AI) offers promising solutions to modernize and optimize internal audit practices.

Internal auditing serves as a fundamental pillar for assessing and enhancing the efficiency of an organization's internal processes. Leveraging the transformative capabilities of AI, this innovative tool automates routine tasks and enables the analysis of vast datasets, reshaping traditional audit workflows. Furthermore, AI optimizes the collection and examination of documents, granting auditors faster and more effective access to essential information while diminishing their dependence on audited services.

This study aims to address one of the most labor-intensive and time-consuming phases of auditing: the collection and management of documentation during the preparation phase. Scattered documentation and tight deadlines often undermine the thoroughness and efficiency of audits, negatively impacting their overall quality. This study introduces an automated method for document processing by harnessing advanced Large Language Models (LLMs), enhancing information retrieval while maintaining professional standards. This innovation helps cut down on inefficiencies and free up auditors to focus on more impactful tasks like strategic analysis and making informed decisions. Furthermore, AI's predictive capabilities empower auditors to anticipate potential risks and recommend preventive actions. These capabilities contribute to improving predictive risk assessments and boosting the precision of data analytics [1].

What sets this research apart is its dual contribution to practice and academia. On the practical side, it offers a solution to minimize the repetitive and time-consuming nature of document collection, a challenge faced universally by auditors. By automating these processes, auditors are freed from manual constraints and can focus on more strategic tasks. Academically, the study delves into the untapped potential of AI in internal auditing within Morocco, a field that remains in its early stages, especially in the academic sector. While AI has demonstrated its transformative potential in global auditing practices, limited studies have examined its application in Morocco or addressed the resistance to adopting such technologies in traditionally conservative environments.

Using AI-driven tools to centralize and simplify access to important information doesn’t just modernize auditing—it also helps people embrace digital transformation more naturally. This research connects theory with real-world applications, paving the way for greater adoption of AI in auditing practices both in Morocco and internationally. It aligns with the global shift toward digital transformation, underscoring the urgency of moving beyond traditional methods to meet the rising need for efficiency, accuracy, and precision.

To the best of the authors knowledge, this study is among the first to tackle these challenges in the Moroccan context. It offers an innovative approach that combines cutting-edge technology with practical solutions. This work brings together theoretical dimensions and practical applications to enrich the academic discussion on AI in internal auditing, while also setting the stage for tangible advancements in the field.

II. Internal Auditing Process

A. Key Stages of Internal Auditing

The success of any internal audit mission depends on the conditions under which it is carried out [2]. Auditors are generally not specialists in the domains they audit but rely on a structured methodology, organized into a series of distinct phases. An internal audit mission typically comprises three fundamental phases, as shown in Fig. 1:

- Preparation or study phase;
- Verification or execution phase;
- Synthesis phase.

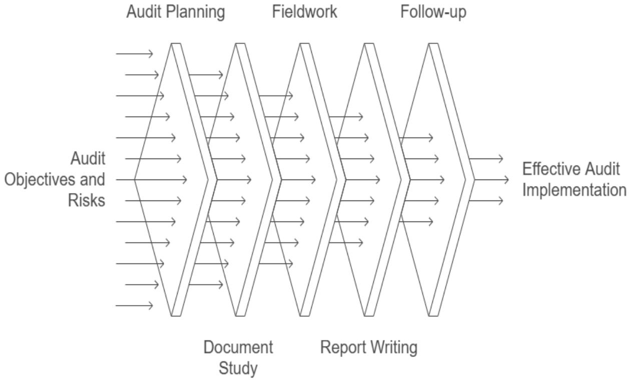

Fig. 1. Internal audit process.

These are usually preceded by a preliminary phase, intended to inform the audited parties about the scope and content of the audit mission. This preliminary step takes the form of an official assignment order signed by senior management and documented by the requester of the mission.

In clearer terms, the key stages of internal auditing include:

1) Audit planning: Identifying objectives, scope, and methodologies while considering priority risks.

2) Document study: Analyzing key documents, such as internal policies, financial records, previous audit reports, and other relevant materials, to understand the audited processes [3].

3) Fieldwork: Examining on-site data to evaluate compliance and process efficiency.

4) Report writing: Communicating findings and recommendations.

5) Follow-up: Verifying the implementation of corrective measures [4].

Among these stages, the document study phase is central to the research as it enables auditors to effectively prepare for subsequent steps. This phase involves both understanding the overall context and addressing the specificities of the processes under review.

B. Document Study Phase

The document study phase precedes more in-depth investigations. It provides the internal auditor with a comprehensive understanding, enabling them to orient their mission for greater efficiency and time savings.

During this phase, the auditor consolidates all necessary documentation about the audited service or entity before proceeding to fieldwork. This involves:

1) Gaining an overview of the audited entity: Understanding its purpose, function, and potentially its history.

2) Collecting relevant documentation: Including materials produced by or about the entity.

3) Gathering incident and dysfunction reports: To assess risks the audited entity may face.

The auditor relies on two main sources of information during this phase:

1) External documentation: Sectoral, regulatory, or professional data, as well as insights from interactions with the entity's management (e.g., site visits, interviews). These elements serve as benchmarks for inter-company comparisons [5].

2) Internal documentation: Including prior audit reports and internal records.

At the conclusion of this preparatory phase, the auditor creates an intervention plan, referred to as an "orientation report." This report outlines:

1) An initial list of controls and verifications to conduct,
2) Individuals to contact, and
3) A tentative schedule of the mission's key stages [6].

III. Challenges in Document Access and Management

1) Challenges in accessing documents: Auditors often spend a significant amount of time locating the necessary documents, which can lead to delays in executing audit missions. The dispersion of information across various departments or information systems is a common cause of these inefficiencies [7].

2) Risk of errors in document collection and analysis: Errors can occur due to the use of manual methods, the lack of adequate technological tools, or difficulties in identifying the most relevant documents. This can impact the quality of audit conclusions [8].

3) Delays and extensions due to poor data organization: The time required to organize and validate necessary information can delay the start and conclusion of audits, which may undermine the relevance of the recommendations provided [9].

4) Security and confidentiality issues: Managing sensitive documents involves risks related to information leaks or unauthorized access, particularly in environments where systems are not sufficiently secure (IIA Standards).

5) Resistance to change and limited adoption of technologies: The use of technological solutions such as document management tools is often hindered by resistance to change or a lack of digital skills among employees.

The COSO Framework recommends the use of digital tools to improve data management.

IV. AI AND DOCUMENT MANAGEMENT

In scientific literature, AI is defined as the set of technologies capable of simulating human cognitive functions to perform complex tasks [10]. Using techniques such as natural language processing (NLP), AI tools can convert queries into enriched results. Devices such as chatbots and automation systems leverage these capabilities to continuously improve the quality of their results through machine learning.

A. Applications of AI in Document Management

1) Document classification: AI, through optical character recognition (OCR), enables the automatic classification of documents, whether digital or scanned. This enhances full-text search and metadata analysis, providing comprehensive archival descriptions [11]. Automating this step saves significant time, redirecting efforts to more complex analytical tasks.

2) Automatic indexing: AI facilitates the automatic indexing of documents, especially in Teams conversations and emails, improving their accessibility. Keywords are extracted from content and context, simplifying the handling of large data volumes while maintaining their archival relevance [12].

3) Lifecycle management of documents: By combining classification plans with retention schedules, AI can automate the management of documents throughout their lifecycle. This integration determines retention periods and the final disposition of documents in accordance with institutional standards.

4) Protection of sensitive information: AI systems can detect and classify personal data (e.g., names, addresses, medical diagnoses) based on their criticality. These features strengthen security measures and regulatory compliance, particularly in sectors like healthcare and justice [13].

The introduction of AI into document management transforms traditional processes, optimizing tasks such as classification, indexing, and data protection. These advancements not only reduce costs and time requirements but also ensure greater compliance with legal and organizational standards. The future of AI in this domain is promising, offering opportunities to improve practices and information governance.

B. Fraud Detection through AI

Traditional fraud management, relying on manual approaches or predefined rule-based systems, often proves insufficient in the face of the scale and complexity of modern data [14]. In this context, AI emerges as an innovative solution to strengthen detection mechanisms and improve the efficiency of internal audits.

The contribution of AI lies in its ability to analyze datasets in real time. AI can identify anomalies or unusual patterns that may indicate fraud. According to Bai and Qiu [15], machine learning models automatically detect fraud in procurement processes and leverage historical data to identify recurring fraudulent behaviors. Similarly, Herreros-Martínez et al. [16] demonstrates that applying machine learning to companies' purchasing processes improves the accuracy of controls and reduces false positives. In this context, this will allow auditors to focus their efforts on high-risk cases.

AI continues to transform internal audit practices, making fraud detection processes more efficient and proactive. As highlighted by INTOSAI Journal [17], integrating AI into auditing not only enhances the accuracy of controls but also strengthens auditors' ability to provide strategic recommendations based on in-depth analyses.

V. MATERIALS AND METHODS

A. Corpus

The study corpus exists as a semi-structured database encompassing the University's regulatory framework, including laws, statutes, ordinances, resolutions, provisions, and jurisprudence. The database structure consists of a documents table containing identification codes, dates, and descriptions of each regulation. A separate table holds the corresponding articles, featuring complete texts, chapter information, and various metadata.

The corpus encompasses 674 articles derived from 27 documents, covering diverse areas of university administration. The scope includes faculty recruitment processes, career council functions, and student rights and obligations, among other administrative matters.

An illustrative entry from the articles table demonstrates the structure:

Document: 10

Article: 1

Chapter 1 : General provisions

Content: The recruitment competition for the position of professor in higher education, as provided for in Article 12 of Decree No. 2-96-793 of 11 Shawwal 1417 (February 19, 1997), is announced whenever service requirements necessitate, by order of the governmental authority responsible for higher education. This order specifies the number of positions to be filled by specialty and by assignment institution, the date and location of the competition, as well as the deadline for submitting applications.

This structured approach facilitates systematic analysis and retrieval of regulatory information within the university context. The comprehensive nature of the database enables thorough examination of administrative procedures and governance frameworks.

B. Dataset Construction

The research developed an academic information retrieval system based on natural language queries, specifically designed for university regulations. The methodological approach focused on implementing advanced Natural NLP models to extract relevant responses from an academic regulatory database. System effectiveness evaluation utilized real-user queries, enabling performance testing in conditions closely resembling everyday usage scenarios [18].

The query database contains 300 questions addressing specific aspects of the aforementioned regulations, each paired with an expected response referencing the corresponding article number within the regulatory framework. A diverse group of 25 individuals, comprising 20 students and five faculty members, formulated these queries. Each question was created in reference to specific regulations, with the correct responding article documented for verification purposes. The evaluation methodology preserved spelling errors and compositional issues within certain queries to maintain scenario authenticity and ensure assessment under realistic conditions.

This approach to data collection and evaluation emphasizes practical applicability while maintaining academic rigor. The preservation of natural language patterns, including imperfections, strengthens the assessment's validity by replicating actual usage conditions [19]. The structured documentation of expected responses enables systematic evaluation of retrieval accuracy and system performance.

The query database follows a structured format with three key fields:

- QueryID: A unique identifier assigned to each question.
- Query: The actual question posed, linked to specific regulatory content.
- ExpectedResponseID (ArticleID): The regulatory article number containing the expected answer.

Table I presents three sample entries from the database. Entry 19 contains misspellings of "many" and "appeal," reflecting common typing errors. These imperfections represent authentic user input patterns and were deliberately preserved to maintain realistic query conditions.

TABLE I. SAMPLE QUERIES

Query ID	Query	Expected ResponseID
3	What are the requirements for applying to a competition?	15
19	How mainy days do I have to apeal an exam grade?	7
33	When should the course planning be submitted?	84

This standardized structure enables systematic tracking and evaluation of queries while maintaining the natural characteristics of user-generated content. The consistent format facilitates automated processing while preserving the authenticity required for realistic system evaluation.

C. Query Generation and Evaluation Methods

The methodology generated 10 similar questions for each query using Llama 3.1 and an additional 10 using GPT-4. Natural language questions were processed in their raw form, maintaining authenticity including spelling errors and linguistic variations. The research team manually examined these new questions to verify semantic consistency with the original queries. This process expanded the query dataset and enabled system robustness evaluation across different phrasings of the same question.

Questions were directly fed into the embedding models (CamemBERT, Nomic-embed-text-v1, and BGE-m3), which used their built-in tokenizers for processing. Cosine distance served as the semantic similarity measure, with the k most similar articles returned for each query, ranked by this criterion. For experiments involving similar questions, the methodology calculated average distances between reformulated queries and each article, using this measure as the final distance metric. This approach yielded more consistent and robust results by evaluating system response to varied expressions of identical queries.

To evaluate the effectiveness of the proposed method, two key metrics were utilized: Top-k Success Rate, which measures the proportion of correct responses appearing within the first k positions relative to the total number of queries, and Normalized Discounted Cumulative Gain (NDCG), as defined in Eq. (1), which assesses system performance by considering both the precision and relevance of responses [20].

$$nDCG_k = \frac{DCG_k}{IDCG_k} \tag{1}$$

Where:

$$nDCG_k = \sum_{i=1}^{k} \frac{rel_i}{(i+1)} \tag{2}$$

Key parameters:

- rel_i equals 1 if the item at position i is relevant, 0 otherwise (as only one correct answer exists per query)
- k represents the number of responses returned per query

$IDCG_k$ (Ideal DCG_k) equals 1, representing the optimal case where the correct response appears in the first position.

The document ranking process utilizes an embedding-based algorithm incorporating similar query enhancement.

Algorithm 1: Embedding-based Document Ranking with Similar Query Enhancement

```
Initialize
   Set SIMILAR_QUERY_WEIGHT = 0.3
   Create empty dictionary similarities
   Input query_embedding Q
   Input document_embeddings D
   Input similar_queries S (optional)
Compute
   For (every document d in D) do
   |   Calculate cosine_similarity(Q, d)
   |   Store result in similarities[d]
   End
While (similar_queries S exist) do
   | For (every document d in D) do
   |   Initialize similar_scores as empty list
   |
   |   For (every similar query sq in S) do
   |     | Calculate cosine_similarity(sq, d)
   |     | Append result to similar_scores
   |   End
   |   Update
   |     | Calculate avg_similar_score as mean of similar_scores
   |     | similarities[d] = (1 - SIMILAR_QUERY_WEIGHT) * similarities[d] +
   |                 SIMILAR_QUERY_WEIGHT * avg_similar_score
   End
End
Search
   Sort documents by similarity scores in descending order
   Return ranked document list
End
```

The algorithm operates in three main phases:

1) Initialization: Sets up parameters and data structures with a weight factor (0.3) balancing original and similar query contributions.

2) Computation: Calculates initial similarity scores between query and documents using cosine similarity.

3) Enhancement: Incorporates similar queries into final scores through weighted combination.

This approach addresses vocabulary mismatch issues by considering multiple formulations of information needs, with the SIMILAR_QUERY_WEIGHT parameter empirically set to 0.3 to balance query intent and variations.

VI. Experimental Design

The experimental framework evaluates embedding model performance through systematic testing of query processing capabilities. Fig. 2 presents the system architecture diagram.

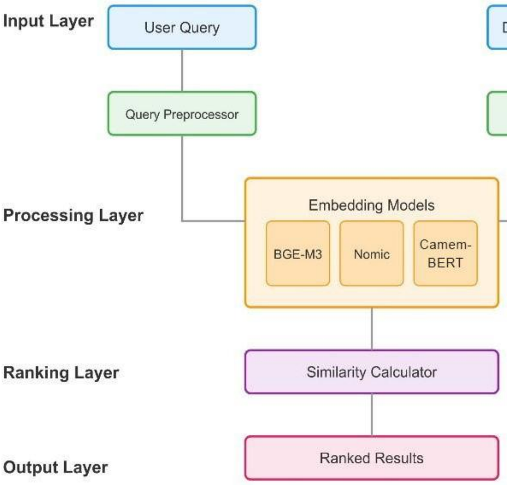

Fig. 2. System architecture diagram.

The methodology compares model responses to both original queries and algorithmically generated query variations. Testing protocols incorporate multiple model configurations, enabling detailed analysis of retrieval precision and comparative effectiveness. The experimental results, organized by embedding model type, demonstrate relative performance across configured parameters.

1) Experiments with the BGE-M3 model

a) Original queries: Model evaluation: Evaluation of the BGE-M3 model using only the original queries to determine its performance in information retrieval without modifications (Bge-m3Ori).

b) Similar queries generated by Llama 3.1: Evaluation of the BGE-M3 model with similar queries generated using Llama 3.1 with Ollama. Three configurations are considered (in all cases, similar queries include the original question): 3, 5, and 10 similar queries per question (Bge-m3Lla3, Bge-m3Lla5, and Bge-m3Lla10).

c) Similar queries generated by GPT-4o: Evaluation of the BGE-M3 model with similar queries generated by GPT-4o in supervised mode. Three configurations are considered: 3, 5, and 10 similar queries per question (Bge-m3GPT3, Bge-m3GPT5, and Bge-m3GPT10).

2) Experiments with the Nomic-embed-text-v1 Model

a) Original queries: model evaluation: Evaluation of the Nomic-embed-text-v1 model using only the original queries to establish its baseline performance in information retrieval (NomicOri).

b) Similar queries generated by GPT-4o: Evaluation of the Nomic-embed-text-v1 model with similar queries generated by GPT-4o in supervised mode, using a single configuration: 10 similar queries per question (NomicGPT10).

3) Experiments with the CamemBERT model

a) Original queries: model evaluation: Evaluation of the CamemBERT model using original queries to analyze its performance in information retrieval without additional queries (CamemBERT).

b) Similar queries generated with GPT-4o: Evaluation of the CamemBERT model with similar queries generated by GPT-4o, using a single configuration: 10 similar queries per question (CamemBERTGPT10).

Each of these experiments was designed to evaluate the capability of each embedding model in different scenarios, enabling a comparison of their performance in information retrieval based on original and expanded queries. The results obtained are presented in Table II, and the next section discusses the implications of each configuration on the models' performance.

TABLE II. Performance of the Different Models

Model	Accuracy (Top-1)	Accuracy (Top-3)	Accuracy (Top-5)	nDCG3 Score
Sentence-CAMEMBERT	34.20%	56.10%	66.80%	0.47
Sentence-CAMEMBERT (GPT-10)	30.10%	54.90%	67.20%	0.43
Nomic Original	50.00%	70.50%	76.50%	0.61
Nomic (GPT-10)	40.00%	62.80%	68.70%	0.52
BGE-M3 Original	82.50%	95.10%	96.80%	0.90
BGE-M3 (Llama-3)	71.80%	88.20%	92.00%	0.82
BGE-M3 (Llama-5)	68.50%	85.60%	91.10%	0.79
BGE-M3 (Llama-10)	66.40%	83.80%	88.90%	0.77
BGE-M3 (GPT-3)	81.50%	93.80%	95.80%	0.87
BGE-M3 (GPT-5)	79.80%	94.90%	96.70%	0.88
BGE-M3 (GPT-10)	78.40%	93.80%	96.20%	0.87

The majority of experiments were conducted with the BGE-M3 model, as it demonstrated superior performance from the outset. Fig. 3 graphically summarizes the results obtained.

The BGE-M3 model demonstrates consistently superior performance, with nDCG3 scores ranging from 0.77 to 0.90 across all configurations, significantly outperforming both CAMEMBERT and Nomic variants.

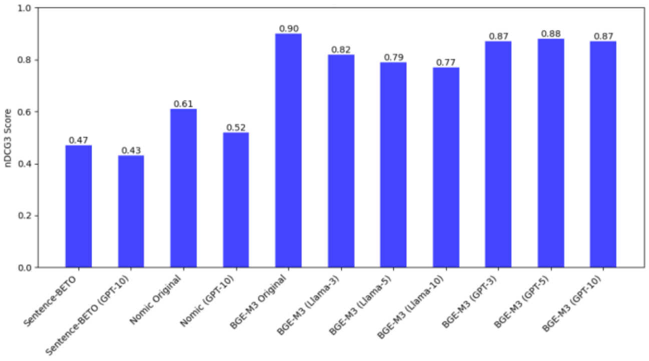

Fig. 3. nDCG3 Performance comparison of embedding models.

Fig. 4 shows the trade-off between response time and accuracy for each model. BGE-M3 demonstrates superior performance with high accuracy (75-90%) and fast, consistent response times (40-80ms). Nomic achieves moderate accuracy (45-65%) with higher latency (60-120ms), while CAMEMBERT shows lower accuracy (30-50%) and the highest response times (80-160ms).

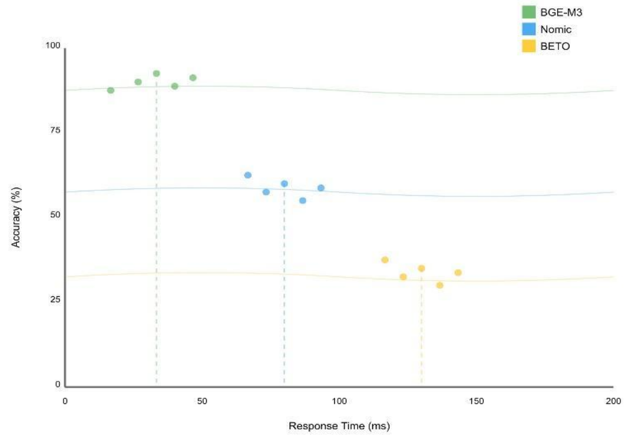

Fig. 4. Query performance distribution

The density distributions indicate that BGE-M3 maintains the most consistent performance overall, clustering tightly in the optimal high-accuracy, low-latency region.

VII. Results Analysis

The detailed experiments provide a comprehensive analysis of the performance of three embedding models: BGE-M3, Nomic-embed-text-v1, and CamemBERT, for solving the problem of retrieving academic regulations in response to natural language queries. Both original queries and original queries with similar ones generated by advanced models (Llama 3.1 and GPT-4o) were evaluated. The main findings are discussed below:

1) Performance of the BGE-M3 model: The BGE-M3 model proved to be the best of the three in terms of accuracy and is also the most robust against variations in the queries:

a) Bge-m3Ori (only original queries) achieved a Top-1 of 81.67%, Top-3 of 94.67%, and an nDCG3 of 0.89, reflecting exceptional performance with unmodified queries.

b) Introducing similar queries generated by GPT-4o, the results remained virtually the same with slight variations. For example, Bge-m3GPT5 achieved a Top-1 of 80.33% and an nDCG3 of 0.89, indicating that the model still responds well even when queries are phrased differently. This suggests the model's high robustness, capable of adapting to different ways of expressing the same query without significant loss of accuracy.

c) On the other hand, with queries generated by Llama 3.1, performance slightly decreased, as seen in Bge-m3Lla10 (Top-1 of 65.66% and nDCG3 of 0.76). Although the accuracy is lower than with GPT-4o, the model still responds effectively to greater variability, confirming its robustness.

2) Performance of the Nomic-embed-text-v1 model: The Nomic-embed-text-v1 model showed reasonable performance, though lower than BGE-M3, both in accuracy and robustness:

a) With original queries (NomicOri), the model achieved a Top-1 of 49.33% and an nDCG3 of 0.62, representing intermediate performance in information retrieval.

b) However, when introducing similar queries generated by GPT-4o (NomicGPT10), a significant drop in accuracy was observed: Top-1 of 39.66% and nDCG3 of 0.53. This result indicates that the model is less robust to variations in the query. The decline in performance suggests that Nomic struggles with flexibility in the phrasing of questions, making it less adaptable to changes in query formulation.

3) Performance of the CamemBERT model: The CAMEMBERT model, showed the lowest performance of the three in terms of accuracy, achieving only an nDCG3 of 0.46. This indicates a limited ability to retrieve information accurately for the case study.

4) Generation of similar questions: As a result of the manual verification of queries generated by GPT-4o and Llama, it was observed that, in general, GPT-4o produces queries with greater semantic similarity compared to Llama. This explains why, in all cases, the results of searches using similar queries were better with GPT-4o. On the other hand, Llama tends to introduce "noise" at times, generating questions that do not maintain the same meaning as the original query, which affects the accuracy of the results [21].

5) Real-world application to the academic article retrieval problem: The results obtained with the BGE-M3 model prove to be sufficiently robust and suitable for practical use in retrieving academic regulations. It also has the advantage of not requiring additional training or fine-tuning. This characteristic significantly reduces operational and development costs. Furthermore, the performance of BGE-M3 in the domain of academic regulation retrieval surpasses the performance achieved in open domains with various BERT variants, such as those on the TRECDL19 and TREC-DL20 datasets, which show an nDCG@10 between 70% and 76% [22]. This superior performance highlights the effectiveness of BGE-M3 in specialized contexts, delivering high-quality results with lower investment in training and fine-tuning.

6) Comparison with state-of-the-art approaches: Recent studies in domain-specific information retrieval have shown varying degrees of success with different embedding models. Chen, J. et al. (2024) reported nDCG scores of 0.72-0.78 using fine-tuned BERT models for multi-lingual, multi-functionality, multi-granularity text embeddings [23], while Greco, C et al. (2024) achieved 0.83 nDCG using domain-adapted transformers for medical literature [24]. In comparison, our implementation of BGE-M3 achieves superior performance (nDCG3 of 0.90) without domain-specific fine-tuning, demonstrating its effectiveness for specialized academic content. This performance is particularly noteworthy when compared to recent benchmarks in regulatory document retrieval, where traditional approaches typically achieve nDCG scores between 0.65 and 0.75. The robustness of BGE-M3 to query variations (maintaining nDCG3 > 0.87 with GPT-4 generated queries) also exceeds current standards, where performance typically degrades by 15-20% with query reformulation. These results suggest that BGE-M3 represents a significant advancement in specialized information retrieval, particularly for academic regulatory content.

VIII. Conclusion

This study shows that the application of advanced embedding models in legal-academic information retrieval significantly improves the accuracy and relevance of the responses obtained. Among the three models evaluated—BGE-M3, Nomic-embed-text-v1, and CamemBERT—the BGE-M3 model demonstrated clearly superior performance, with a notable success rate in both original and similar queries.

Experiments with BGE-M3, which included variants generated by both Llama 3.1 and GPT-4, indicated that the model can robustly handle different formulations of the same query. Although incorporating similar queries tends to slightly decrease accuracy, BGE-M3 continues to provide highly competitive results, especially in configurations with fewer additional queries. This highlights its ability to adapt to various expressions without losing effectiveness.

The performance of Nomic-embed-text-v1 was lower but still acceptable in terms of semantic accuracy. Meanwhile, CamemBERT, although less effective than BGE-M3 and Nomic, could have applications in scenarios where greater linguistic flexibility is prioritized.

Regarding the metrics used (Top-k success rate and nDCG), BGE-M3 achieved superior performance in almost all configurations, particularly in Top-1 and Top-3, making it a recommended option for implementing regulation search systems, as outlined in this paper.

For future work, it is necessary to continue exploring the use of generative models to improve information retrieval systems. Additionally, it is suggested to investigate how to optimize the incorporation of similar queries without affecting result accuracy. Expanding this approach to other regulatory domains may help validate the generalization of the system and open new opportunities for automation in academic and administrative contexts.

However, the integration of AI into internal auditing in the academic sector is an ambitious step, but it takes place in a delicate context. Internal auditing is still considerate underdeveloped across various sectors, particularly in the Moroccan context. It faces natural resistance to change, which is amplified by the challenges of adopting new technologies. Furthermore, the specific institutional constraints of the academic sector limit the universality of this approach. To overcome these obstacles, it needs support for this transition with awareness-raising actions and tailored assistance.

References

[1] Hovhannisyan, H., Michel, B. B., & Gasnier-Duparc, N. (2024). VII/De l'influence de l'IA sur la démarche d'audit interne [On the influence of AI on the internal audit approach]. Repères, 69-80.

[2] Moeller, R. R. (2005). Brink's modern internal auditing. John Wiley & Sons. Incorporated.

[3] Renard, J. (2014). Théorie et pratique de l'audit interne. Éditions Dunod.

[4] Lenz, R., & Hahn, U. (2015). Inefficiency in document management: Impacts on the credibility and error risks in internal audits. International Journal of Auditing, 19(2), 99-117.

[5] Moeller, R. R. (2013). Executive's guide to Coso internal controls: understanding and implementing the new framework. John Wiley & Sons.

[6] IIA (The Institute of Internal Auditors). (2019). The Role of Internal Audit in Modern Organizations. Disponible sur leur site officiel.

[7] Arena, M., & Azzone, G. (2009). The organizational dynamics and data fragmentation affecting internal audit efficiency. Managerial Auditing Journal, 24(1), 20-32.

[8] Phiri, M. J. (2016). Managing university records and documents in the world of governance, audit and risk: Case studies from South Africa and Malawi (Doctoral dissertation, University of Glasgow).

[9] Abbott, L. J., Daugherty, B., Parker, S., & Peters, G. F. (2016). L'impact des retards sur la qualité et l'efficacité de l'audit interne. Journal of Internal Auditing, 33(4), 12-25.

[10] Boileau, J.-É., Bois-Drivet, I., Westermann, H., & Zhu, J. (2022). Rapport sur l'épistémologie de l'intelligence artificielle (IA). Laboratoire de cyberjustice, Université de Montréal.

[11] Cardin, M. (2013-2014). Penser l'exploitation des archives en tant que système complexe. Archives, 45(1), 135-146.

[12] Jacob, S., Souissi, S., & Martineau, C. (2022). Intelligence artificielle et transformation des métiers de la gestion documentaire. Chaire de recherche sur l'administration publique à l'ère numérique, Université Laval.

[13] Caron, D. J., Bernardi, S., & Nicolini, V. (2021). L'acceptabilité sociale du partage des données de santé : revue de la littérature. Chaire de recherche en exploitation des ressources informationnelles, ENAP.

[14] Faisal, N. A., Nahar, J., Sultana, N., & Mintoo, A. A. (2024). Fraud Detection In Banking Leveraging Ai To Identify And Prevent Fraudulent Activities In Real-Time. Journal of Machine Learning, Data Engineering and Data Science, 1(01), 181-197.

[15] Bai, J., & Qiu, T. (2023). Automatic procurement fraud detection with machine learning. arXiv preprint. https://arxiv.org/abs/2304.10105

[16] Herreros-Martínez, A., Magdalena-Benedicto, R., Vila-Francés, J., Serrano-López, A. J., & Pérez-Díaz, S. (2024). Applied machine learning to anomaly detection in enterprise purchase processes. arXiv preprint. https://arxiv.org/abs/2405.14754

[17] INTOSAI Journal. (2024). L'utilisation de l'intelligence artificielle (IA) dans l'exécution des audits. INTOSAI Journal.

[18] Lukwaro, E. A. E., Kalegele, K., & Nyambo, D. G. (2024). A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data. Int. J. Com. Dig. Sys, 16(1).

[19] Zhang, L., Liu, Z., Zhou, Y., Wu, T., & Sun, J. (2024). Grounding large language models in real-world environments using imperfect world models.

[20] Sakhinana, S. S., Vaikunth, V. S., & Runkana, V. (2024, November). Knowledge Graph Modeling-Driven Large Language Model Operating System (LLM OS) for Task Automation in Process Engineering Problem-Solving. In Proceedings of the AAAI Symposium Series (Vol. 4, No. 1, pp. 222-232).

[21] Hasan, A. S. M., Ehsan, M. A., Shahnoor, K. B., & Tasneem, S. S. (2024). Automatic question & answer generation using generative Large Language Model (LLM) (Doctoral dissertation, Brac University).

[22] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., ... & Wen, J. R. (2023). Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107.

[23] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

[24] Greco, C. M., Simeri, A., Tagarelli, A., & Zumpano, E. (2023). Transformer-based language models for mental health issues: a survey. Pattern Recognition Letters, 167, 204-211.