Journal of Technology Research (JTR)
Volume 4, Special Issue, (2026), pp 856-864, ISSN 3005-639X
The 7th Conference of Engineering Sciences and Technology
(CEST-2025)

Building an AI-driven assistant for the Administrative and Legal Domain

Hajar Mansour [1] , Asma Elmangoush *[1] , Majdi Ashibani[2]

[1] College of Industrial Technology, Misurata, Libya,

[2]Libyan Academy for Telecom and Informatics, Misurata, Libya,

*Corresponding author email: asma_elmangoush@cit.edu.ly.

Received: 28-09-2025 | Accepted: 25-11-2025 | Available online: 03-01-2026 | DOI:10.26629/jtr.2026.**

ABSTRACT

The rapid growth of digital information has created significant challenges in managing, organizing, and retrieving knowledge, particularly within legal and administrative domains. Conventional keyword-based search tools often fail to capture the contextual meaning of complex documents, leading to inefficiencies in decision-making and legal research. To address this issue, this study presents the design and implementation of an AI-driven assistant that integrates vector databases with large language models (LLMs) to support administrators and legal professionals. The proposed system leverages semantic embeddings to transform unstructured legal and regulatory texts into high-dimensional vector representations. Experimental evaluation was conducted using a legal dataset from the Libyan Academy for Telecom and Informatics (LATI), comprising more than 200,000 words of official laws and regulations. The system was tested with legal queries and assessed both automatically and through expert review. Results demonstrated that the assistant retrieved accurate, contextually relevant passages, significantly reducing response times compared to manual search. Legal experts confirmed that most answers were precise and practically useful, although further refinement is required for handling ambiguous or nuanced cases. The system improves administrative and legal efficiency by enabling semantic search beyond keyword matching and providing actionable insights for decision-makers. Future work will focus on expanding the legal corpus, refining query handling, and exploring advanced indexing techniques to improve scalability, accuracy, and adaptability across different domains.

Keywords: Large Language Model, Vector Database, AI-driven Assistant.

بناء مساعد قائم على الذكاء الاصطناعي للمجال الإداري والقانوني

هاجر منصور[1]، أسماء المنقوش[1]، مجدي الشيباني[2]

[1]، قسم الهندسة الالكترونية ، كلية التقنية الصناعية ، مصراتة ، ليبيا

[2] الأكاديمية الليبية للاتصالات والمعلوماتية، مصراتة ، ليبيا

ملخص البحث

إن النمو السريع للمعلومات الرقمية قد أدى إلى تحديات كبيرة في إدارة وتنظيم واسترجاع المعرفة، خصوصًا في المجالات القانونية والإدارية. غالبًا ما تفشل أدوات البحث التقليدية المعتمدة على الكلمات المفتاحية في التقاط المعنى السياقي للنصوص المعقدة، مما يؤدي إلى قصور في كفاءة اتخاذ القرار والبحث القانوني. لمعالجة هذه المشكلة، تقدم هذه الدراسة تصميم وتنفيذ مساعد قائم على الذكاء

الاصطناعي يدمج قواعد البيانات المُتجهة **(Vector Databases)** مع النماذج اللغوية الكبيرة **(LLMs)** لدعم الإداريين والمهنيين القانونيين. يعتمد النظام المقترح على التمثيلات الدلالية **(Semantic Embeddings)** لتحويل النصوص القانونية والتنظيمية غير المهيكلة إلى متجهات عالية الأبعاد. تم إجراء تقييم تجريبي باستخدام مجموعة بيانات قانونية من **الأكاديمية الليبية للاتصالات والمعلوماتية(LATI)** ، تضم أكثر من 200,000 كلمة من القوانين واللوائح الرسمية. جرى اختبار النظام عبر استعلامات قانونية وتم تقييمه آليًا ومن خلال مراجعة خبراء. أظهرت النتائج أن المساعد استرجع مقاطع دقيقة وذات صلة سياقية، مع تقليل ملحوظ في زمن الاستجابة مقارنة بالبحث اليدوي. وأكد الخبراء القانونيون أن معظم الإجابات كانت دقيقة وذات فائدة عملية، رغم الحاجة إلى مزيد من التحسين لمعالجة الحالات الغامضة أو الدقيقة. يسهم النظام في تعزيز الكفاءة الإدارية والقانونية من خلال تمكين البحث الدلالي بما يتجاوز مطابقة الكلمات المفتاحية، وتوفير رؤى قابلة للتنفيذ لصنّاع القرار. وستركز الأعمال المستقبلية على توسيع قاعدة البيانات القانونية، وتحسين معالجة الاستعلامات، واستكشاف تقنيات فهرسة متقدمة لتعزيز قابلية التوسع والدقة والقدرة على التكيف عبر مجالات مختلفة.

الكلمات الدالة: بالنماذج اللغوية الكبيرة(LLM) ؛ قاعدة البيانات المُتجهة (Vector Database)؛ المساعد القائم على الذكاء الاصطناعي.

1. INTRODUCTION

The volume of digital information produced across platforms and applications continues to rise rapidly, creating new challenges for governments and organizations in managing and extracting value from the daily generated data. This data comes in various forms, including text, images, videos, and both structured and unstructured information, making its processing and analysis a significant challenge. As the volume of this data increases, it has become imperative to develop advanced technologies that can efficiently organize and extract value from it. [1]

Artificial intelligence (AI), particularly large language models (LLMs), now plays a central role in addressing these challenges, offering powerful methods for analyzing patterns and extracting knowledge beyond the capacity of conventional approaches [2]. AI technologies are utilized in various fields, including machine translation, text analysis, and intelligent predictions, revolutionizing how data is leveraged to inform decisions and boost productivity.

Despite these advancements, a persistent challenge lies in the efficient storage and retrieval of unstructured data especially in applications that depend on semantic search, such as recommendation engines, content analysis platforms, and intelligent virtual assistants. To address this issue, vector databases have gained prominence. These specialized systems manage unstructured data by encoding it into high-dimensional vector representations, thereby enabling semantic search based on similarity metrics rather than exact keyword matching. This paradigm shift significantly enhances the scalability and relevance of data retrieval in big data environments.

Given the growing demand for rapid and context-aware information retrieval, vector databases have become indispensable in applications such as image recognition, voice-based interfaces, and conversational AI. By providing a robust infrastructure for semantic analysis, these databases contribute to improved accuracy, responsiveness, and overall performance of AI systems in extracting actionable insights from complex datasets.

This paper presents a vector database-augmented generative question-answering assistant for administrators in organizations and companies. This assistant avoids LLM's model hallucinations by extracting content from a local professional database as supplementary knowledge. This system focuses on achieving the following objectives:

- Generating Analytical Insights: Assist managers and decision-makers by delivering accurate information extracted from large volumes of documents and data.
- Legal Comparisons and Recommendations: Facilitate comparisons between laws and provide recommendations for the most suitable legal options in various scenarios.
- AI-Powered Smart Search Tools: Enhance legal research with AI-based search tools that comprehend the legal context beyond traditional keyword searches.
- Improving Administrative and Legal Efficiency: Streamline and increase the efficiency of administrative and legal processes.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 outlines the methodology, Section 4 presents implementation and results, and Section 5 concludes with key findings and future directions

2. RELATED WORK

Recent advancements in artificial intelligence (AI) and large language models (LLMs) have intensified the demand for efficient systems capable of managing high-dimensional data. Vector databases (VDBs) have emerged as a critical component in this landscape, offering scalable solutions for storing, indexing, and retrieving dense embeddings that traditional database management systems struggle to handle.

In [3] a comprehensive survey of vector databases provides a detailed examination of storage and retrieval methodologies, with a particular focus on approximate nearest neighbor (ANN) search algorithms. The study compares leading VDB systems in terms of architecture, performance trade-offs, and application domains. It also highlights emerging research directions, such as novel indexing mechanisms and tighter integration with LLMs, aiming to guide future development and adoption of VDBs in AI ecosystems.

The reliability and testing of vector database management systems (VecDBs) have become a pressing concern. A recent empirical study [4] investigates the unique challenges posed by the multidimensional nature of vector data, the probabilistic nature of ANN search, and the dynamic scaling requirements of LLM pipelines. The paper identifies critical gaps in software testing practices—particularly in test input generation, oracle definition, and evaluation metrics—and proposes a research roadmap for developing robust testing methodologies. These efforts are essential for ensuring the dependability of VDBMSs in data-intensive AI applications. These studies underscore the strategic importance of vector databases in enabling scalable, intelligent, and reliable AI systems.

Other researchers emphasize that vector databases can efficiently manage the high-dimensional embeddings generated by large language models, and that their integration can significantly enhance model accuracy and utility. A good systematic review in [5] explores the synergistic potential of LLMs and vector databases (VecDBs), focusing on how to address challenges such as hallucinations, outdated knowledge, and high application costs. The paper also discusses future developments and research and engineering challenges in the field, presenting diverse applications and prototypes, to encourage further research on improving the interaction between LLMs and VecDBs to enhance data processing and knowledge extraction capabilities.

The authors in [6] proposed an effective approach to accelerate the deployment of LLM models using INT4-precision automatic weight quantization with optimized runtime on CPUs, improving inference efficiency. The approach has been successfully applied to popular models such as Llama2 and GPT-NeoX, offering superior performance compared to open-source solutions.

Another study explores the integration of generative AI, such as Llama3 8B, Mistral 7B, and Phi-3 Mini 3.8B, into clinical trial design in the field of pharmacogenomics [7]. A comprehensive comparison of models was conducted in terms of accuracy, relevance, and operational efficiency using a local environment equipped with an RTX 4080 graphics card and an Intel Core i9 processor. The results showed superior accuracy and relevance, particularly for Llama3 8B and Phi-3 Mini 3.8B, with variations in efficiency and scalability. The study suggests that generative AI can significantly enhance clinical trial design by improving patient stratification and data management, although challenges such as bias remain and further validation and development are needed.

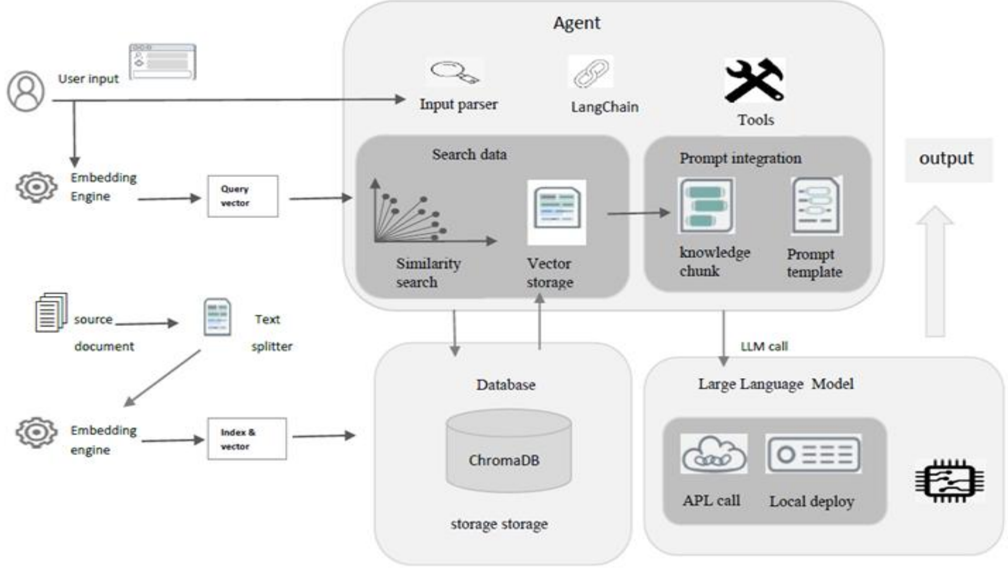


Fig 1. Information retrieval system using a vector database and a large language model

In [8] a multi-pronged strategy for developing an intelligent academic chatbot based on the Retrieval Augmented Generation (RAG) model is presented. The work aims to improve textual and image information retrieval and support high-quality content production. The system combines large-scale language models such as LLaMA and Gemini, vector search techniques via FAISS, and web scraping using Beautiful Soup to gather information about courses, faculty, and admissions. The integration of Groq AI enables reduced response time, while the Streamlit interface provides an interactive experience for uploading documents and asking queries. Results demonstrate the system's effectiveness in enhancing academic engagement and automating institutional communication, while improving the contextual accuracy of responses. Future work includes multilingual support, advanced image analysis, fine-tuning of query domains, integration with knowledge graphs, and cloud deployment to ensure scalability and accessibility.

The study in [9] addresses the recurring problem of ambiguity in contract terms in the construction sector, and the resulting legal disputes and project delays. The authors proposed a solution by employing LLMs tailored to industry regulations. A specialized chatbot was developed using ChatGPT, powered by a building regulations database, to interpret contracts and automate document management tasks. The development process included architecture design, data preparation, vector embedding, and model integration. Evaluation showed that the specialized bot achieved an accuracy of 88% compared to 36% for standard ChatGPT, with the potential to save 70% time and achieve an accuracy of up to 85% in analysis, extraction, and interpretation tasks. The research confirms that the integration of vector databases and embedding techniques enables advanced semantic search beyond keyword search, improving the efficiency of information extraction from unstructured data.

3. METHODOLOGY

The methodology employed in this research centers on the development of a legal information retrieval system powered by vector-based semantic search and LLM. In this section, the main stages of the system and design considerations are described. Figure 1 illustrates the AI-Assistance system's main components.

3.1 Dataset Extraction and Processing

The dataset utilized in this study was sourced from the official digital repository of the Libyan Academy for Telecom and Informatics (LATI). It comprises a curated collection of legal documents, including national laws and regulatory frameworks, obtained directly from verified governmental sources. These documents were digitized and systematically organized within LATI's internal filing infrastructure to facilitate structured processing and analysis. The legal corpus was integrated into the study's designated legal software environment, enabling semantic and quantitative evaluation. In total, approximately 212,880 words were uploaded and processed, encompassing a diverse range of legislative and regulatory texts.

The legal texts analyzed in this study were ingested using the fitz library (PyMuPDF) [10], which facilitated the extraction of content from digitally formatted documents. Each page was programmatically parsed and converted into an individual document unit, enabling granular control over text segmentation. This preprocessing step was essential for optimizing downstream tasks such as tokenization, semantic indexing, and contextual analysis. The structured page-level decomposition significantly enhanced the efficiency and accuracy of legal text processing within the adopted analytical framework.

3.2 Text Segmentation Strategy

Following the initial upload, the legal texts were segmented into overlapping blocks using the RecursiveCharacterTextSplitter algorithm. Each block was configured to contain 1,000 characters, with a 100-character overlap between consecutive segments. This segmentation strategy was deliberately selected to preserve the semantic continuity and contextual integrity of legal discourse, while minimizing the risk of generating ambiguous or fragmented embeddings during downstream processing. The overlapping design ensures that critical legal references and dependencies—often spanning across sentence or paragraph boundaries—are retained, thereby enhancing the fidelity of semantic analysis and improving the performance of language model-based interpretation. Table 1 shows the characteristics of dataset segmentation and processing.

Table 1. Dataset segmentation and processing characteristics

Characteristic	Description
Text Block Size	Approximately 1000 words
Overlap Words	100 words
Text Source	PDFDocuments
Transformation Function	fitz (PyMuPDF)
Library Used	LangChain
Trade-off Considerations	Context vs. Embedding Quality
Database Population Efficiency	High

3.3 Vector Database Configuration

During the embedding phase, the study employed the intfloat/multilingual-e5-small model, a compact yet high-performing member of the E5 family of multilingual encoders [11]. This model is specifically designed to generate semantically rich vector representations across diverse languages, including Arabic, Persian, and English, thereby supporting cross-lingual semantic search and retrieval tasks. Each input segment was transformed into a 384-dimensional embedding, offering a favorable balance between computational efficiency and semantic fidelity. The model's lightweight architecture makes it particularly well-suited for applications requiring rapid processing and low resource consumption, without compromising the accuracy of meaning representation in multilingual legal texts.

For vector storage and retrieval, the study employed ChromaDB [12], selected for its lightweight architecture, seamless integration with large language models, and suitability for local deployment without reliance on complex infrastructure. ChromaDB supports advanced contextual querying through high-dimensional text embeddings, enabling efficient semantic search and retrieval across multilingual legal corpora. Its modular design and compatibility with embedding models such as intfloat/multilingual-e5-small made it an ideal choice for scalable experimentation and rapid prototyping within the study's legal informatics framework.

3.4 Embedding Dimensionality Configuration

The dimensionality of the vector embeddings stored in the database was configured to 384 dimensions to ensure full compatibility with the output of the intfloat/multilingual-e5-small model. This model, optimized for multilingual semantic representation—including support for Arabic, Persian, and English—produces compact yet semantically rich embeddings that balance interpretability with computational efficiency. The 384-dimensional vector space enables high-speed processing and effective semantic search, making it well-suited for legal text analysis in multilingual environments.

3.5 Indexing Method:

For internal indexing within ChromaDB, the study employs the Hierarchical Navigable Small World (HNSW) algorithm. HNSW is a graph-based approximate nearest neighbor (ANN) search technique known for its high recall rates and low latency, making it particularly suitable for large-scale vector retrieval tasks. Its hierarchical structure enables efficient functional matching of high-dimensional embeddings by navigating through layered proximity graphs. This configuration allows ChromaDB to perform rapid and resource-efficient semantic searches, thereby enhancing the responsiveness and scalability of the legal text retrieval system.

3.6 Similarity Metric

To evaluate semantic similarity between embedded text segments, Cosine Distance was employed as the primary metric. This approach is particularly effective for contextual comparison of high-dimensional vectors, as it considers the angular relationship between vectors rather than their absolute Euclidean distance. Such a formulation is well-suited for SBERT-style embeddings, where semantic proximity is encoded in vector orientation.

Mathematically, the cosine distance between two vectors A and B is defined as:

$$\frac{\mathrm{A} \cdot \mathrm{B}}{\| \mathrm{A} \| \times \| \mathrm{B} \|} - 1 = \text{Cosine Distance}(\mathrm{A}, \mathrm{B}) \tag{1}$$

This formulation ensures that vectors with similar semantic content yield lower distance values, thereby enhancing the precision of retrieval tasks in multilingual legal corpora. Upon receiving an input query vector, the system initiates a similarity search within the vector database to identify the most semantically aligned text segments. Leveraging the configured indexing mechanism and cosine distance metric, the database efficiently computes proximity scores between the input vector and stored embeddings. The output is a ranked list of the top-matching paragraphs, selected from the entire corpus of legal documents, thereby enabling context-aware retrieval and supporting downstream tasks such as legal interpretation, summarization, or decision support.

Algorithm 1 Vector Database Configuration and Query Execution

```
Input: Input Query Q, Number of Top Matches to Retrieve n
Initialize Chroma Vector Database:
  Set Embedding Model to SBERT (paraphrase-multilingual-MiniLM-L12-v2)
  Set Dimensionality to 384
  Set Distance Metric to Cosine Similarity
function EMBEDDOCUMENTSINTOVECTORS
  Load PDF documents using PyPDFLoader
  Split documents into chunks using RecursiveCharacterTextSplitter
  Embed each chunk using HuggingFaceEmbeddings
  Store embedded vectors into Chroma vector store
end function
function PERFORMQUERY(Q, n)
  Embed input query Q using same SBERT model
  Search Chroma vector store for top n similar vectors using cosine similarity
  Sort and rank results based on similarity score
  Retrieve associated document chunks (paragraphs)
  Output top n results
end function
Execute PERFORMQUERY(Q, n)
```

4. IMPLEMENTATION AND RESULTS

4.1 Model Configuration

To generate semantic embeddings for legal texts, the multilingual model intfloat/multilingual-e5-small was selected. This model supports a wide range of languages—including Arabic—and offers an optimal balance between semantic accuracy and computational efficiency. As a pre-trained encoder, it requires no additional fine-tuning and is used solely to transform textual input into high-quality vector representations for retrieval tasks.

A set of experimental queries was developed based on legal documents in PDF format. These documents were processed using the PyMuPDF library to extract text content, followed by segmentation via LangChain's RecursiveCharacterTextSplitter to preserve contextual coherence. The resulting text chunks were embedded using the SentenceTransformer framework and stored in the ChromaDB vector database.

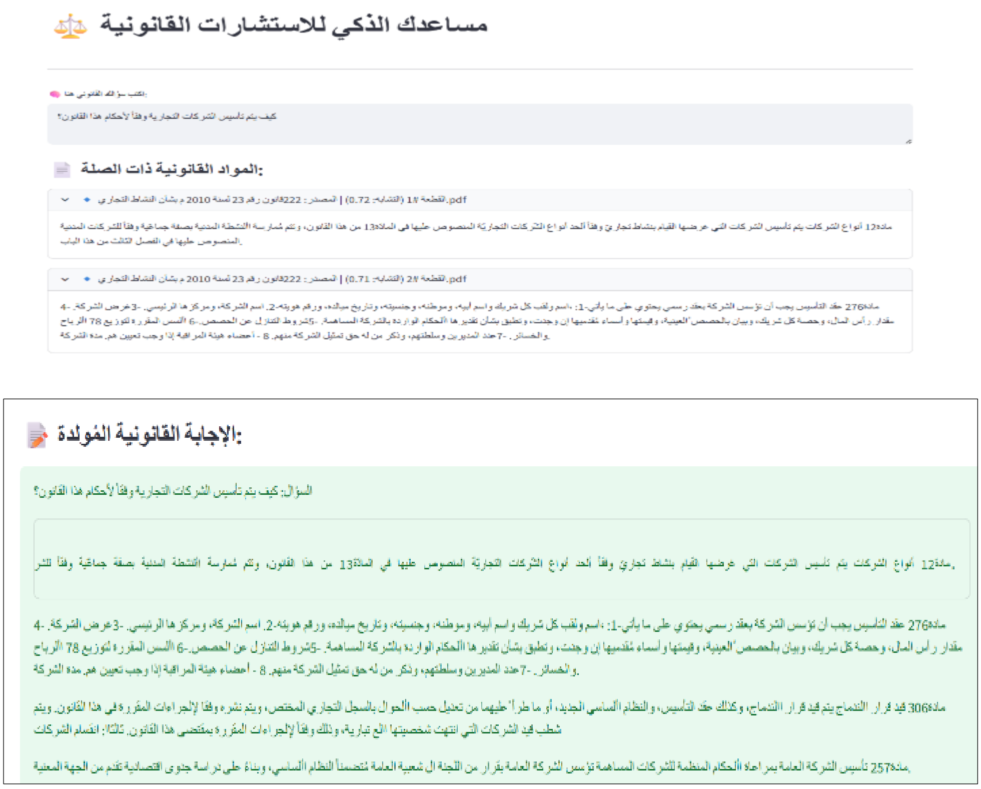
During inference, queries are matched against stored embeddings using cosine similarity. The top-ranked results are retrieved from ChromaDB and passed through a Streamlit interface to the language model for optional summarization or direct response generation. This architecture ensures low-latency interaction and efficient semantic search. The system is implemented and tested entirely in a local environment. However, the system could be implemented on the cloud with scalability options for larger datasets and multi-user access.

4.2 Experimental testing

The developed system was tested by posing a set of legal queries to a local model built using HuggingFace Transformers and linked to the vector database. The results were recorded and analyzed, with each experiment including the question posed, the resulting answer, the response time, and the case identification code.

In addition, the results were presented to a legal expert, who gave their opinion on the accuracy of the answers and their relevance to the legal context. They considered the system to provide correct answers in most cases, while noting that some contextual details could be improved to be more comprehensive.

The results in Figure 3 reveal distinct variation in response times across 18 queries. The initial queries exhibited relatively high latencies of 14 to 16 seconds, which can be attributed to system initialization and model-loading overhead. A peak was observed at q3, followed by a period of relative stability with average response times of approximately 15 seconds up to q11, indicating consistent performance under continuous processing load. A notable inflection occurred at q12, where response time sharply decreased to 8.5 seconds. This reduction reflects the completion of initialization processes and the effective utilization of cache memory, resulting in improved efficiency.



مساعدك الذكي للاستشارات القانونية ⚖️

اكتب سؤالك القانوني هنا:

كيف يتم تأسيس الشركات التجارية وفقاً لأحكام هذا القانون؟

المواد القانونية ذات الصلة:

القطعة #1 (التشابه: 0.72) | المصدر: 222قانون رقم 23 لسنة 2010 م بشأن النشاط التجاري.pdf

مادة12 أنواع الشركات يتم تأسيس الشركات التي غرضها القيام بنشاط تجاري وفقاً لأحد أنواع الشركات التجارية المنصوص عليها في المادة13 من هذا القانون، وتتم ممارسة الأنشطة المدنية بصفة جماعية وفقاً للشركات المدنية المنصوص عليها في الفصل الثالث من هذا الباب.

القطعة #2 (التشابه: 0.71) | المصدر: 222قانون رقم 23 لسنة 2010 م بشأن النشاط التجاري.pdf

مادة276 عقد التأسيس يجب أن تؤسس الشركة بعقد رسمي يحتوي على ما يأتي:-1. اسم ولقب كل شريك واسم أبيه، وموطنه، وجنسيته، وتاريخ ميلاده، ورقم هويته.-2 اسم الشركة، ومركزها الرئيسي. -3 غرض الشركة. -4 مقدار رأس المال، وحصة كل شريك، وبيان بالحصص العينية، وقيمتها وأسماء مقدميها إن وجدت، وتطبق بشأن تقديرها الأحكام الواردة بالشركة المساهمة. -5 شروط التنازل عن الحصص. -6 النسب المقررة لتوزيع الأرباح والخسائر. -7 عدد المديرين وسلطتهم، وذكر من له حق تمثيل الشركة منهم. 8 - أعضاء هيئة المراقبة إذا وجب تعيينهم. مدة الشركة

الإجابة القانونية المُولدة:

السؤال: كيف يتم تأسيس الشركات التجارية وفقاً لأحكام هذا القانون؟

مادة12 أنواع الشركات يتم تأسيس الشركات التي غرضها القيام بنشاط تجاري وفقاً لأحد أنواع الشركات التجارية المنصوص عليها في المادة13 من هذا القانون، وتتم ممارسة الأنشطة المدنية بصفة جماعية وفقاً للشر

مادة276 عقد التأسيس يجب أن تؤسس الشركة بعقد رسمي يحتوي على ما يأتي:-1. اسم ولقب كل شريك واسم أبيه، وموطنه، وجنسيته، وتاريخ ميلاده، ورقم هويته.-2 اسم الشركة، ومركزها الرئيسي. -3 غرض الشركة. -4 مقدار رأس المال، وحصة كل شريك، وبيان بالحصص العينية، وقيمتها وأسماء مقدميها إن وجدت، وتطبق بشأن تقديرها الأحكام الواردة بالشركة المساهمة. -5 شروط التنازل عن الحصص. -6 النسب المقررة لتوزيع الأرباح والخسائر. -7 عدد المديرين وسلطتهم، وذكر من له حق تمثيل الشركة منهم. 8 - أعضاء هيئة المراقبة إذا وجب تعيينهم. مدة الشركة

مادة306 قيد قرار الاندماج يتم قيد قرار الاندماج، وكذلك عقد التأسيس، والنظام الأساسي الجديد، أو ما طرأ عليهما من تعديل حسب الأحوال بالسجل التجاري المختص، ويتم نشره وفقاً للإجراءات المقررة في هذا القانون. ويتم شطب قيد الشركات التي انتهت شخصيتها الاعتبارية، وذلك وفقاً للإجراءات المقررة بمقتضى هذا القانون. ثالثاً: انقسام الشركات

مادة257 تأسيس الشركة العامة مع مراعاة الأحكام المنظمة للشركات المساهمة تؤسس الشركة العامة بقرار من اللجنة الشعبية العامة متضمناً النظام الأساسي، وبناءً على دراسة جدوى اقتصادية تقدم من الجهة المعنية

Fig2. Implementation results of the AI-powered legal assistant system

5. CONCLUSIONS

This study presents the design and implementation of an intelligent legal chatbot system that leverages vector databases for the storage, analysis, and retrieval of statutory and regulatory texts. By integrating ChromaDB with multilingual language models from the Hugging Face Transformers library, the system demonstrates high efficiency in processing legal documents and delivering contextually accurate responses to user queries. The architecture enables rapid semantic search and retrieval, thereby supporting legal research workflows and institutional decision-making.

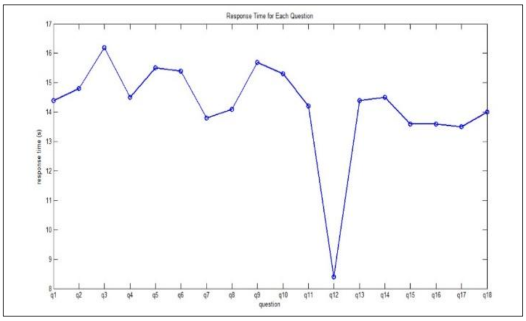


Fig. 1. Response time for different 18 questions

Empirical evaluation, including expert review of system outputs, confirms the system's practical utility in the domain of legal

informatics. The chatbot consistently produced accurate and concise responses, grounded in relevant legal references, with minimal latency. Nonetheless, certain limitations were observed, particularly in handling nuanced or edge-case queries, which could be mitigated by expanding the legal corpus and refining the underlying natural language processing mechanisms.

Overall, the findings underscore the value of vector database integration in managing high-dimensional legal data and enhancing semantic retrieval performance. Future work will focus on broadening the scope of legal content, optimizing language model architectures for domain-specific tasks, and incorporating advanced indexing algorithms to further improve the scalability, precision, and reliability of legal information systems

6. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Libyan Academy for Telecom and Informatics for providing access to the legal dataset used in this study. We also thank our colleagues at the College of Industrial Technology, Misurata, for their valuable feedback and technical support.

REFERENCES

[1] Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. 2022. doi:10.3390/info1010000.

[2] Fernandez RC, Elmore AJ, Franklin MJ, Krishnan S, Tan C. How large language models will disrupt data management. Proc VLDB Endow. 2023;16(11):3302–9. doi:10.14778/3611479.3611527.

[3] Ma L, et al. A comprehensive survey on vector database: storage and retrieval technique, challenge. 2025. Available from: http://arxiv.org/abs/2310.11703

[4] Wang S, et al. Towards reliable vector database management systems: a software testing roadmap for 2030. 2025. doi:10.48550/arXiv.2502.20812.

[5] Jing Z, et al. When large language models meet vector databases: a survey. 2025. Available from: https://ann-benchmarks.com/

[6] Shen H, Chang H, Dong B, Luo Y, Meng H. Efficient LLM inference on CPUs. 2025. p. 33–46. doi:10.1007/978-3-031-85747-8_3.

[7] Fatharani A, Alsayegh A. Pharmacogenomics meets generative AI: transforming clinical trial design with large language models. *J Pharmacol Pharmacother*. 2025. doi:10.1177/0976500X251321885.

[8] Tahseen A, Sumathi J, Reddy S, Shravani D. RAG-based query engine using LLM and vector DB for college details. J Front Multidiscip Res. 2025. doi:10.54660/IJFMR.2025.6.1.86-91

[9] Saparamadu PVIN, et al. Optimising contract interpretations with large language models: a comparative evaluation of a vector database-powered chatbot vs. ChatGPT. Buildings. 2025;15(7):1144.

[10] PyMuPDF documentation. 2025 Sep 6. Available from: https://pymupdf.readthedocs.io/en/latest/

[11] Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual E5 text embeddings: a technical report. 2024. Accessed 2025 Sep 6. Available from: https://huggingface.co/intfloat/multilingual-e5-base

[12] GitHub – chroma-core/chroma: Open-source search and retrieval database for AI applications. 2025 Sep 6. Available from: https://github.com/chroma-core/chroma