

Functional Architecture of an Artificial Intelligence-Based Virtual Tax Assistant System

O. I. Ollaberganov, F. O. Yusupov, U. K. Ro'zimboyev

Abstract. *This paper presents the development of the functional architecture of an artificial intelligence-based virtual tax assistant system designed to operate on the basis of the national tax legislation. The proposed system is built upon a Large Language Model (LLM), the Retrieval-Augmented Generation (RAG) framework, and a vector database. To capture the semantic relationships among tax legislation documents, a knowledge graph was constructed, and a graph-based semantic chunking approach was employed. Furthermore, the document retrieval and context generation mechanisms were enhanced using the Graph-RAG approach. Experimental results demonstrate that the proposed method improves the accuracy of generating contextually relevant responses compared with the conventional RAG algorithm. The proposed architecture enhances the effectiveness of the virtual tax assistant and enables the provision of reliable advisory services based on the national tax legislation.*

Keywords: *Artificial Intelligence, Virtual Tax Assistant, Large Language Model (LLM), Retrieval-Augmented Generation (RAG), Graph-RAG, Knowledge Graph, Semantic Chunking, Qdrant.*

In recent years, the adoption of artificial intelligence (AI) technologies in the digital transformation of public services has accelerated significantly. According to the OECD Forum on Tax Administration (FTA), by 2024, 74.1% of tax administrations had deployed virtual assistants (chatbots), of which 97.5% were based on predefined rule-based mechanisms, while only 27.5% utilized AI technologies to provide personalized interactions with taxpayers. Furthermore, although 72% of tax administrations have incorporated AI into various business processes, its primary applications remain limited to tax risk assessment, fraud detection, and virtual assistant services [1].

Despite these advancements, research on AI-based virtual tax advisory systems capable of generating context-aware and explainable responses based on the tax legislation of the Republic of Uzbekistan remains limited. The frequent updates to tax regulations and the complex semantic relationships among legal documents make it challenging for conventional retrieval-based systems to provide accurate, reliable, and contextually relevant responses. Therefore, developing an intelligent virtual tax assistant that integrates Large Language Models (LLMs) with Graph-based Retrieval-Augmented Generation (Graph-RAG) represents a promising approach to improving the quality and reliability of tax advisory services.

The development of the proposed virtual tax assistant system was carried out in several stages. Initially, the Tax Code, laws, and other regulatory documents were collected, followed by data cleaning and preprocessing. In the next stage, a knowledge graph was constructed based on the semantic relationships among tax legislation documents, and a graph-based semantic chunking approach was applied. The resulting semantic chunks were converted into vector embeddings using the BGE-M3 embedding model [2] and stored in the Qdrant vector database [3]. To retrieve documents relevant to user queries, the Graph-RAG retrieval mechanism [4] was employed. The final responses were generated using the Qwen3 Large Language Model (LLM) [5], while the Low-Rank Adaptation (LoRA) technique [6] was utilized to adapt the model to the Uzbek tax legislation domain. The functional architecture of the proposed virtual tax assistant system is presented in Figure 1.

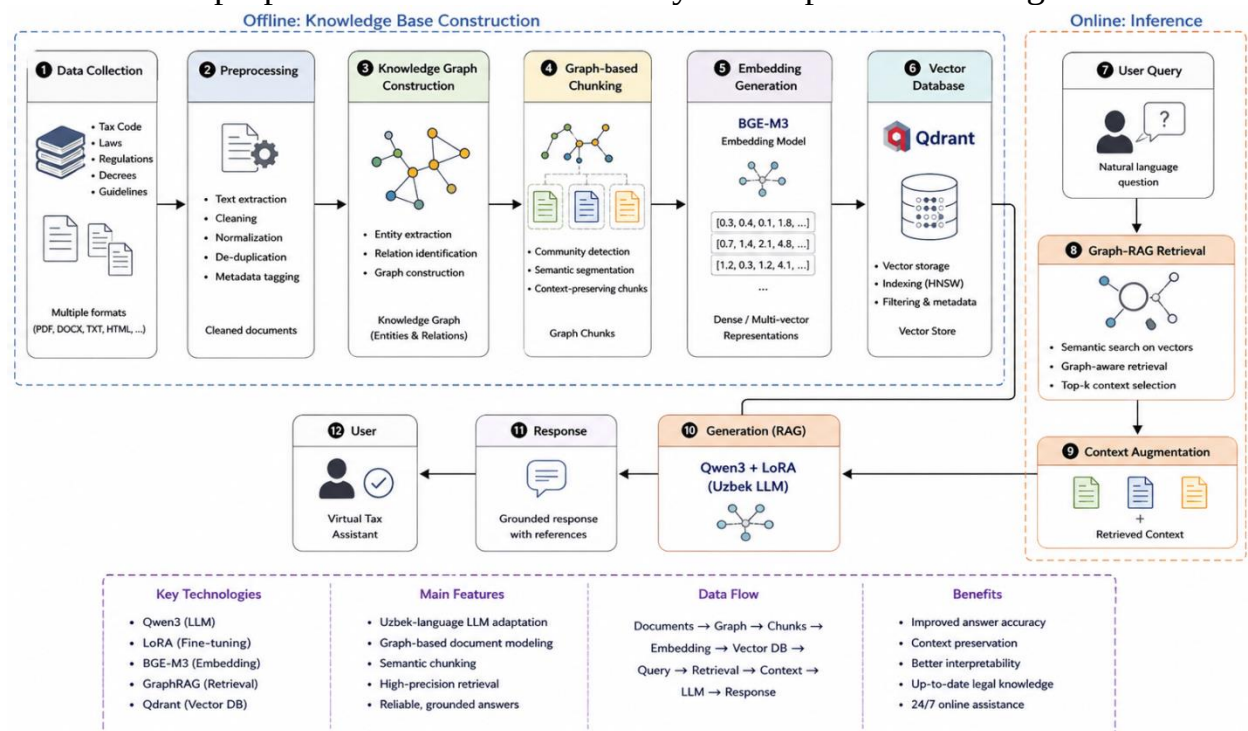


Figure 1. Functional architecture of the artificial intelligence-based virtual tax assistant system.

Initially, when the conventional Retrieval-Augmented Generation (RAG) algorithm was employed, the response generation accuracy of the virtual tax assistant was approximately 50%. To improve the system performance, the semantic relationships among the documents of the Tax Legislation of the Republic of Uzbekistan were analyzed, and a knowledge graph was constructed accordingly.

Based on the knowledge graph, the documents were reorganized using a graph-based semantic chunking approach, and the vector database was reconstructed.

To evaluate the effectiveness of the proposed approach, a benchmark dataset consisting of 270 question–answer pairs with expert-validated reference answers was prepared by tax domain specialists. Following each stage of system enhancement, the proposed architecture was re-evaluated using this benchmark dataset, and the response accuracy was measured based on the correspondence between the generated responses and the reference answers.

The main contribution of this work is the construction of a knowledge graph that captures the semantic relationships among the documents of the Tax Legislation of the Republic of Uzbekistan and the integration of a graph-based semantic chunking method into the virtual tax assistant system. Experimental results demonstrate that the proposed Graph-RAG approach improved the response generation accuracy from 50% to over 75%. Compared with the conventional RAG algorithm, the proposed method significantly enhanced both the accuracy and the contextual relevance of the responses provided to users.

Conclusion. This study presented the development of the functional architecture of an artificial intelligence-based virtual tax assistant system operating on the basis of the Tax Legislation of the Republic of Uzbekistan. By integrating a knowledge graph, graph-based semantic chunking, the Retrieval-Augmented Generation (RAG) framework, and a Large Language Model (LLM), the proposed system achieved more accurate retrieval of relevant documents and improved the accuracy of response generation for user queries. Experimental results demonstrated that the proposed approach provides higher performance than the conventional RAG algorithm. Future work will focus on extending the system with multimodal data processing, agentic artificial intelligence technologies, and automatic response evaluation mechanisms to further enhance the reliability and intelligence of the virtual tax assistant.

References

1. OECD. Tax Administration: Digitalisation and Digital Transformation Initiatives. Paris: OECD Publishing, 2025.
2. Chen J., Xiao S., Liu Z., et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Embeddings. arXiv, 2024.
3. Qdrant Team. Qdrant Vector Database Documentation. 2025.
4. Edge D., Trinh H., Cheng N., et al. From Local to Global: A GraphRAG Approach to Query-Focused Summarization. Microsoft Research, 2024
5. Yang A., Yang B., Hui B., et al. Qwen3 Technical Report. arXiv, 2025.

6. Hu E., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations (ICLR), 2022.

7. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), 2020.