



# AI-Powered Voice Agent for Saudi Labor Law Inquiries

Prepared By

**Name: Hussam Almokayed**

**ID No.: 253001038**

**Name: Fatimah Aldeghaither**

**ID No.: 253000292**

Submitted in fulfillment of the requirements for the degree Master of  
Science in Artificial Intelligence

At

College of Informatics, Midocean University, UAE.

**Supervisor Name: Dr. Hager Saleh**

College of Informatics, Midocean University, Comoros

# Table of Contents

|                                                   |    |
|---------------------------------------------------|----|
| Abstract .....                                    | 3  |
| 1. Chapter 1: Introduction and Overview .....     | 3  |
| 1.1 Overview .....                                | 3  |
| 1.2 Problem Definition.....                       | 4  |
| 1.3 Thesis Objectives .....                       | 4  |
| 1.4 Thesis Contribution.....                      | 5  |
| 1.5 Thesis Organization .....                     | 6  |
| 2. Chapter 2: Background and Related Work .....   | 6  |
| 2.1 Introduction and Background .....             | 6  |
| 2.1.1 Artificial Intelligence (AI) .....          | 7  |
| 2.1.2 Machine Learning (ML) .....                 | 7  |
| 2.1.3 Deep Learning (DL).....                     | 7  |
| 2.1.4 Natural Language Process (NLP) .....        | 7  |
| 2.1.5 Large Language Models (LLMs).....           | 7  |
| 2.1.6 Retrieval-Augmented Generation (RAG) .....  | 8  |
| 2.1.7 Automatic Speech Recognition (ASR) .....    | 8  |
| 2.1.8 Text-to-Speech (TTS) .....                  | 8  |
| 2.1.9 Word Error Rate WER.....                    | 8  |
| 2.1.10 Role of AI in the Voice Agent System ..... | 8  |
| 2.2 Related Work .....                            | 9  |
| Review of Related Works .....                     | 9  |
| References .....                                  | 15 |

## Abstract

In recent years, conversational AI models have vastly improved their performance through the use of large language models (LLM) and retrieval-augmented generation (RAG) methods, allowing real-time processing, multimodal fusion, and answering complex and sensitive queries, such as legal and religious ones. In addition, real-time streaming Automatic Speech Recognition (ASR), Text-to-Speech (TTS) synthesis, and knowledge-retrieval-based (LLMs) can be combined to create autonomous voice assistants capable of performing with knowledgeable and fluent answers in a specialized legal area, including labor law, approaching human-level performance.

There are, however, still major challenges, particularly within the Arabic domain specific applications like dialectal speech recognition, factual correctness of legal responses and low-latency end-to-end voice processing systems.

This paper proposes an AI-based Arabic voice agent for Saudi Labor Law services to overcome these challenges. The suggested system combines (ASR), (RAG), (LLMs), and (TTS) modules into a single system to provide accurate and natural voice responses in Arabic.

There is no result on Abstract since the code has not finished yet, as recommended in the last semester

## 1. Chapter 1: Introduction and Overview

### 1.1 introduction

Artificial intelligence (AI) is an aspect that has gained significance in various contemporary systems and services in the past few years. Among its applications, conversational (AI) systems, which are set to engage the callers in a human-like way, hold great significance.

The fast evolution of the Automatic Speech Recognition (ASR), Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), and Text-to-Speech (TTS) systems has transformed the functionality of modern voice agents. The development opens a window of opportunity to develop voice agents in difficult areas, such as the legal sector, where users require conversational, accurate, and trustworthy help.

Nevertheless, there are some challenges when creating such an Arabic voice assistant. To begin with, Arabic cannot be easily identified because of the numerous dialects. Secondly, the information related to the laws should be accurate since even small mistakes can mislead the users and reduce their trust.

The purpose of this research project is to develop an Arabic voice assistant to answer questions about Saudi labor law. To do so, the proposed architecture incorporates a knowledge source, as well as ASR, RAG, LLMs, and TTS.

The proposed system aims to provide users with natural, real-time, and reliable legal assistance in Arabic.

## 1.2 Problem Definition

Government-run labor law contact centers in Saudi Arabia face several operational and technical difficulties. First, call volume is high, resulting in long wait times and a poor user experience. Moreover, the queries are repetitive, such as those on employee rights, end-of-service benefits, and contract termination procedures.

Finally, human customer agents provide the service, but this does not guarantee the quality or accuracy of the responses. The automation currently used in contact centers is limited because it supports only a predefined menu of intents, which might not cover all possible caller queries. For instance, the caller may not find a matching intent in the menu, making it difficult to convey the purpose of the call. In addition, these services are usually available only during working hours, limiting their availability to callers. On the technical level, existing conversational AI services are unable to understand spoken Arabic queries in real time, especially when the domain of the queries is specialized, such as labor law. For example, accurately recognizing spoken Arabic dialects and ensuring the factual accuracy of the response are two challenging problems. Moreover, integrating speech processing technology (ASR) with a knowledge-grounded reasoning technology, retrieval-augmented generation (RAG), poses another technical challenge.

In this respect, a smart real-time voice-based service, which is capable of understanding spoken Arabic requests and giving consistent, accurate, and contextually relevant responses, is clearly needed. This paper meets this need by presenting an AI-powered voice-assistant, which uses ASR, RAG, LLM, and TTS to create an integrated system into the Saudi Labor Law service.

## 1.3 Thesis Objectives

The main aim of this study is to develop an Arabic Voice Agent that can respond to questions about Saudi labor law. This system will employ the use of artificial intelligence (AI) to comprehend and act on user queries.

The proposed system combines Automatic Speech Recognition (ASR), Retrieval-Augmented

Generation (RAG), Large Language Models (LLMs), and Text-to-Speech (TTS) modules, which are used to provide real-time human-system interaction, by using dialect-based ASR and knowledge-based generation capabilities, this work aims to:

1. To improve the precision of the legal responses produced by the system to go hand in hand with the fluency of the responses on the foundation of the Saudi Labor law articles.
2. To enhance the system performance in response time and the number of hallucinations of the model, which influences the accuracy of the legal responses.
3. To research the model configurations and the retrieval algorithms to identify the most suitable technique to use in Arabic applications in the legal field.

The goal of this thesis is to develop a real-time, scalable, and user-friendly voice-enabled system that provides around-the-clock access to Saudi legal services. The performance of the proposed model will be evaluated using the Word Error Rate (WER), response accuracy, and end-to-end latency.

## 1.4 Thesis Contribution

**Next semester required**

## 1.5 Thesis Organization

**Next semester required**

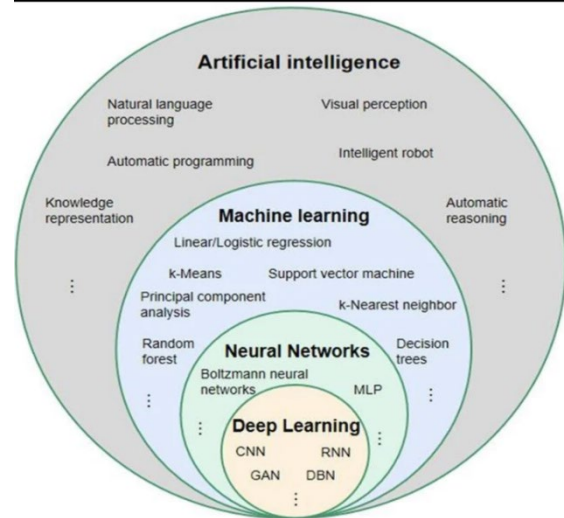
## 2. Chapter 2: Background and Related Work

### 2.1 Introduction and Background

Artificial Intelligence (AI) has played a key role in the development of conversational AI. Over the past few years, the incorporation of LLM and RAG has paved the way for the development of real-time, multi-modal, and domain-reasoning conversational AI systems over various domains and data types, such as legal and religious, which are highly sensitive, and enabled the model to go beyond response generation to contextual and knowledgeable response generation. For example, the incorporation of ASR, TTS, and knowledge-grounded LLMs has enabled the development of an end-to-end voice agent that can recognize spoken input and produce speech responses naturally and accurately. Such a voice agent is crucial for domain-specific applications where the domain requires precise and knowledgeable response generation. However, there are a few challenges facing the Arabic conversational AI system, such as accurate recognition of dialectal speech, factual correctness in response generation, and low end-to-end latency, all of which are critical in the legal domain. In this chapter, we will review the recent studies that focused on AI-based voice agents and their building blocks, such as speech processing, LLM, retrieval mechanisms, and optimization techniques, and the way in which these components can be integrated to build a scalable, efficient, and accurate voice agent for Saudi Labor Law applications.

### 2.1.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is a subdivision of computer science that deals with creating systems that can imitate human intelligence. These systems are also designed to do the tasks that usually involve human cognition, e.g., learning, reasoning, and problem solving, and enhance performance using data-driven methods. [1].



### 2.1.2 Machine Learning (ML)

Machine Learning (ML) is an area of Artificial Intelligence (AI) focused on developing data-based models that learn and make predictions or act on data without being programmed. Today, machine learning-enhanced algorithms implement statistical processes and optimization algorithms in order to perform optimally, as additional data is collected [2].

### 2.1.3 Deep Learning (DL)

Deep learning (DL) is a subfield of AI that can be defined as deep neural network models composed of multiple layers that are trained to learn hierarchical representations from large-scale data. These models learn to extract features from raw input data to achieve state-of-the-art performance on tasks such as speech recognition, natural language processing, and computer vision [3].

### 2.1.4 Natural Language Process (NLP)

Natural Language Processing (NLP) refers to computational techniques that enable machines to process, understand, and generate human language by leveraging data-driven models and deep neural architectures. Recent advances in NLP have been driven by large language models, enabling significant improvements in tasks such as semantic understanding, dialogue systems, and information retrieval [4].

### 2.1.5 Large Language Models (LLMs)

Large Language Models (LLMs) are deep learning models built on transformer architectures and trained on large text datasets to comprehend, generate, and otherwise manipulate human language. By learning contextual patterns in data, they can achieve a variety of natural language capabilities such as text generation, question answering, and reasoning [5].

### 2.1.6 Retrieval-Augmented Generation (RAG)

RAG is a hybrid architecture that fuses parametric knowledge represented in LLMs with non-parametric external knowledge retrieved on-the-fly from external data stores. The approach facilitates the generation of context-dependent, up-to-date, and veracious responses, especially for domain-specific or real-time use cases [6].

### 2.1.7 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a machine's ability to recognize spoken words and transcribe them into text. ASR relies on deep learning models to learn patterns of sounds and language. State-of-the-art ASR is built with end-to-end neural networks trained on large datasets and performs well for real-time recognition and multilingual tasks [7].

### 2.1.8 Text-to-Speech (TTS)

Text-to-Speech (TTS) is the ability to convert text into a human-like voice. TTS relies on deep learning models to generate natural speech. State-of-the-art TTS produces a natural human-sounding voice with normal intonation [8].

### 2.1.9 Word Error Rate WER

Word Error Rate (WER) is a standard performance measurement for ASR systems. WER measures the difference between the actual transcript and the reference transcript. The measurement accounts for substitution, deletion, and insertion errors at the word level and returns a normalized value [9]. WER is especially relevant to Arabic ASR.

### 2.1.10 Role of AI in the Voice Agent System

AI is the main technology that helps the voice agent to understand the user's input and give a response. In the first step, the system uses ASR (Automatic Speech Recognition) to convert the speech into text. Then, NLP and NLU is used to find out what the user wants. After that, a language model generates the response. To provide a more accurate response, the system uses RAG to retrieve information from external knowledge sources. At last, TTS is used to convert the response into speech, so that the user can listen in a human way. AI is significant to the functional core of this system in order to provide an interface between the user intent and system action. It is not a simple task, but the voice agent passes through a complex pipeline whereby initially, the audio input is translated into text via ASR. Afterward, the text is processed by Natural Language Understanding (NLU) modules to extract its meaning. This is necessary when dealing with legal matters, as it is important to understand the context.

## 2.2 Related Work

### Review of Related Works

The authors [10] proposed an end-to-end voice processing pipeline designed to automate customer support in telecommunications and interactive voice response systems. They used a retrieval augmented generation (RAG) framework to retrieve information from RFC documents using Facebook AI Similarity Search (FAISS). The system incorporates a streaming automatic speech recognition (ASR) system based on a conformer transcribe-to-end (TTE) model to enable real-time voice processing. Additionally, the system incorporates a 4-bit-quantized large language model (Telecom Specific Language Model Mini-2B) and a real-time text-to-speech (TTS) system, dubbed Telecom Synthesis. They used a private dataset containing 500 human-recorded telecom utterances. Preprocessing involved denoising and sentence-level streaming. The result is a real-time voice response with an average latency of 0.94 seconds. They conclude that streaming workflows and quantization are critical to attaining knowledge-grounded voice responses.

The authors [11] introduced SpeechGPT, a large model that supports speech perception and generation within a large-language-model architecture. LLaMA-7B was the primary architecture they used, and Hidden Unit Discovery was used to transform raw audio into discrete tokens that could be understood by the model. The Speech Instruct dataset, which consists of 800 hours of speech and text instructions that are carefully paired, was used to run the experiments to allow a better cross-modal alignment. They were trained in three phases: adaptation of modality and fine-tuning of text and speech instruction. The model scored highly in natural verbal interaction and reasoning, while retaining its original text-based reasoning. In general, they demonstrated that SpeechGPT provides low-latency, human-like speech responses to various conversational tasks without degrading on more standard language benchmarks.

The authors [12] proposed a group multi-agent Retrieval-Augmented Generation system (RAGentA) based on attributed question answering to render answers more reliable and grounded. They employed a full multi-agent RAG system in which four specialized agents (Predictor, Judge to filter, Final-Predictor adds in-line citations, and Reviser checks completeness and refines the query as needed) operate in a loop over the MAIN-RAG base. These models were Falcon-3-10B driving all agents, and a hybrid retrieval of BM25 sparse search and E5 dense embeddings with Pinecone. The dataset was a simulated benchmark of 500 pairs of different QA generated out of the FineWeb index. Top-20 hybrid retrieval, dynamic thresholds, adaptive document scoring and filtering, and query reformulation of missing parts were used as preprocessing. The result is fluent responses with explicit in-text references and second-stage polishing to complete coverage. Ultimately, they reasoned that this method increases faithfulness by more than 10% and provides more reliable responses than standard RAG, particularly with multi-source questions.

The researchers [13] have developed an easy, inexpensive method to address the lack of Arabic dialect speech samples for speech recognition. They optimized the OpenAI Whisper model across five major dialects: Gulf, Levantine, Iraqi, Egyptian, and Maghrebi. It was an unaugmented, pure Automatic Speech Recognition (ASR) study. They mainly used the smaller whisper-small model (244 million parameters) and compared it directly with the much bigger whisper-large-v3. The information was provided by Mozilla Common Voice of Modern Standard Arabic (approximately 40 hours) and the Massive Arabic Speech Corpus (MASC), which balanced each dialect to up to 20 hours. It was easy to preprocess: they resampled all to 16 kHz, removed log-Mel features, used the Whisper tokenizer with zero-padding, undersampled overrepresented dialects, and trimmed very long transcripts. The results were impressive enough that, even with a little polishing in Modern Standard Arabic, the small model could be compared to the giant when running much faster. The authors finally concluded that such a mixture of dialect data is surprisingly good and provides a more practical approach to producing working Arabic speech recognition systems at relatively low cost and does not rely on huge data sets.

The authors [14] created a system of the NADI 2025 competition that is aimed at two fundamental requirements: first, Arabic dialects Automatic Speech Recognition (ASR), second, Arabic dialects Dialect Identification (ADI) with voice-based inputs. They adopted a so-called transfer learning strategy, where they based their transcription on Whisper (large-v3), and specific encoders, including Wav2Vec 2.0 and HuBERT, to retrieve the required acoustic information to categorize the dialects. The dataset was the official corpus of the NADI 2025, and represents thousands of annotated audio clips, and covers a broad variety of regional languages, including Levantine, Gulf, Egyptian, and Maghrebi. The authors obtained quite impressive scores, particularly the great decrease of Word Error Rate (WER) during ASR, and the high F1-scores during dialect detection. Their findings indicate that self-supervised speech representations, along with dialect-specific fine-tuning are much more effective in Arabic than general-purpose speech models.

The authors [15] devised a special Retrieval-Augmented Generation (RAG) model to learn the nuances of the Moroccan Family Code (Moudawana). They ran the classy pipeline with sophisticated embedding models and applied a reranking mechanism to ensure the highest relevance of retrieved legal statutes. The team has conducted its research based on a specific set of Arabic legal texts and provisions on family law and the judicial process. Several text-segmentation (chunking) strategies were used to manage the formal and complex structure of legal Arabic, ensuring optimal performance. With this strategy, the system achieved substantial reduction in model hallucinations and high factual consistency and legal accuracy scores. The authors noted that domain-specific knowledge retrieval should be integrated to provide sound legal advice to citizens and ensure that the law is strictly adhered to in accordance with official documents.

The authors [16] proposed CTC-TTS to address lag issues in real-time AI voices by replacing heavy alignment tools with a slimmer CTC and bi-word interleaving model. In their approach,

they built two LLMs: one focused on high-quality audio and the other on feature stacking to achieve maximum speed. Training and testing of the model were based on the LibriTTS multi-speaker dataset. This involved certain preprocessing, such as bi-word interleaving. The performance was excellent, with the system recording a Real-Time Factor (RTF) of 0.05, approximately 30 times faster than the regular models, while maintaining very high speech similarity to humans. They concluded that CTC-based alignment is most appropriate for balancing high-quality generation with the rapid streaming required by interactive agents.

The authors [17] presented a more efficient text-to-speech (TTS) synthesis system named CM-TTS. The method of making TTS will accelerate the synthesis of regular diffusion models and optimize the sampling rate. The key point of the given method is that the Consistency Model (CM) will be included into the traditional synthesizing process of TTS so that it could produce the good one within only 1-2 iterations instead of dozens of them. The proposed technique also introduced a Weighted Sampler, which concentrates on complex time steps in order to not only improve the efficiency of the learning process but also make sure the vocal sounds in the learning process remain clear. To confirm the relevance of the proposed technique on the LJspeech dataset (consisting of 13100 short audio samples of a single speaker) and the LibriTTS dataset (consisting of multiple speakers of the experimental results) it was obvious that the proposed technique (when working with an NVIDIA A100 graphics card) provided an efficiency of 0.0028 Real-Time Factor (RTF).

The authors [18] reviewed the key transformer models that are applied in Arabic question answering. Their special attention was paid to AraBERT (the Arabic version of BERT) and MARBERT (the model that is specifically trained on dialectal and informal Arabic) which are effective in extractive question answering (QA) when the system only has to extract the answer right out of the text. To produce full, natural-sounding answers, what they refer to as abstractive QA, they talked about ArabicT5. They also tested the models of multilinguals like mBERT (multilingual BERT) and XLM-R. They took popular Arabic datasets such as Arabic-SQuAD, ARCD (Arabic Reading Comprehension Dataset), and TyDiQA-Arabic as their data. The authors also mentioned that they are working on their own specialized Fatwa based dataset on religious questions. Key techniques, such as morphological segmentation, lemmatization, and clitic separation, were pointed out during the preprocessing stage to deal with the complex morphology of Arabic and the low occurrence of diacritics.

The authors [19] used a particular standard that was aimed at evaluating the reasoning of law in the Saudi court setting. They evaluated the ability of the system to provide judgments and address legal arguments with a broad range of Large Language Models (LLMs), such as AraBERT. They analyzed a sample of 13,000 Saudi Arabian commercial court cases (real world cases). The study obtained high predictive performance in legal results using advanced reasoning schemes. They found that to capture the linguistic peculiarities of real Saudi legal texts, fine-tuning for specific domains is necessary.

One of the new frameworks to offer credible advice on Islamic and legal matters is NEBRAS that was proposed by the author [20]. To provide good Islamic and legal advice, they used retrieval-augmented generation (RAG) and Large Language Model (LLM) agents. Its retrieval method is a hybrid system that is backed by special agents that keenly check the authenticity of any fatwas or sources of law that it relies on. The author designed a pair of question-answer (QA) pairs, targeting legal and religious questions. The system achieved a BERTScore-F1 of 70.94 with such agentic layers of reasoning, which is better than GPT-4o in believing in the holiness and correctness of the law.

ChipChat is a sub-second on-device agent, that the authors [21] have created without cloud GPUs. They trained a pipeline of streaming a mixture of Experts (MoE) to do automatic speech recognition (ASR) and state action augmented Large Language Model LLM. It is powered by Apple Silicon via the Machine Learning experimental framework (MLX), which gives a smooth flow of ASR, dialogue, and text-to-speech (TTS) with full privacy. The information was founded on conventional ASR and LLM benchmarks. The system optimizes streaming and stacks features to enhance processing and speed. Therefore, it can execute in real time on a set of common consumer devices indicating that low-latency conversational agents are a realistic notion in reality.

To solve the sophisticated legal and mathematical problems of the Islamic inheritance law (Fara'id), [22] authors have suggested two-step optimization plan. It is based on the retrieval-augmented generation (RAG) and fine-tuning in particular to ensure that the model becomes more suitable to specialization in legal reasoning, and therefore alleviates the shortcomings of traditional language processing systems. Low-Rank Adaptation (LoRA) was used to fine-tune the 1-9B Fanar. This improvement therefore significantly added to the capabilities of the model to step-by-step, complex reasoning on the equitable allocation of inheritance as provided by Sharia law. The system was tested with the help of a well-selected selection of Islamic questions-answering data as a result of which the goal was to indicate the subtle concepts and mathematics behind the inheritance solution in Sharia. The plan, which united the two, was right 85.8%. The model proved to be more accurate as more and more explanations grew in length, resulting in increased confidence, which was considered to be empirical. This system was also particularly skillful at giving out coherent, legally valid interpretations that would not violate the prevailing legal principles.

A lightweight, 30M, LLM-agnostic, and autoregressive streaming Text-to-Speech (TTS) system proposed by the author [23], called LLMVoX, can generate the speech of high quality and low-latency independent of the properties of the underlying LLM. It applies an autoregressive streaming model with only 30 million parameters, and has a multi-queue token streaming platform to support indefinitely long conversations, without attaching a text-to-speech (TTS) misalignment. The underlying LLM intelligence and language capabilities are maintained in real-time conversations by synthesizing speech in a modularized manner in the model. The system was critically tested on real time standards, and matched with the other built in models,

especially to reduce the latency experienced in difficult conversational environments. Consequently, the Word Error rate (WER) of LLMVOX was reduced drastically and the word and audio outputs were more harmonized. The researchers determined that such a decoupled and streaming model offers a more convenient-to-scale and efficient platform to create naturalistic and engaging voice agents, without the loss of the intrinsic intelligence of the model.

The authors [24] carried out a large-scale empirical investigation of the optimization of the Retrieval-Augmented Generation (RAG) aspects in Arabic through six datasets. Through the RAGAS framework they were in a position to test systematically and in comparison, the various chunking strategies, embedding models and rerankers to evaluate their impacts on the faithfulness and relevance of answers. The results revealed that sentence-aware chunking always was the most effective, semantically maintaining its coherence; additionally, BGE-M3 and Multilingual-E5-large were most effective Arabic embedding models. Additionally, reranked model was used to a great advantage to increase the accuracy of the results on a complex data set and the quality of generation was found to be better in Aya-8B model compared to others used as benchmark. The researchers concluded that the rational choice of component mixes, specific to the document structure is imperative in the construction of high-quality Arabic RAG systems. This research serves to advise developers on practical methods to resolve the most challenging morphological and syntactic challenges of Arabic to the current AI pipelines.

| Paper | Years | Strong                                                                                                       | Limitations                                                           | Datasets                                       | The Best models                                                  | Results                                                     |
|-------|-------|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|------------------------------------------------|------------------------------------------------------------------|-------------------------------------------------------------|
| [10]  | 2025  | Built a complete real-time voice pipeline with streaming ASR, 4-bit LLM, and instant TTS all grounded by RAG | Heavy computing is needed when scaling to large networks              | 500 real telecom calls recorded by humans      | Conformer TTE + Telecom-Specific Mini-2B LLM + Telecom-Synthesis | Average latency 0.94 seconds, RTF < 1.0                     |
| [11]  | 2025  | Integrates speech + LLM in one unified multimodal system                                                     | High training cost and not optimized for domain-specific Arabic tasks | SpeechInstruct (~800 hours paired speech-text) | LLaMA-7B + speech tokenization                                   | Strong conversational performance with human-like responses |
| [12]  | 2025  | Built four collaborative agents for filtering,                                                               | Standard RAG still hallucinates                                       | 500 Synthetic QA pairs from FineWeb            | Falcon-3-10B + BM25 + E5                                         | +10.7% faithfulness improvement                             |

|      |      |                                                                                                            |                                                       |                                                             |                                       |                                                            |
|------|------|------------------------------------------------------------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------------|---------------------------------------|------------------------------------------------------------|
|      |      | citing, and refining answers                                                                               | on complex queries                                    |                                                             |                                       |                                                            |
| [13] | 2025 | Fine-tuned a small Whisper on mixed Arabic dialects and matched large-model performance at lower cost      | Dialect data remains scarce and unbalanced            | Common Voice (MSA ~40 hours) + MASC (~20 hours per dialect) | Whisper-small & large-v3              | The small model achieved similar accuracy to the large one |
| [14] | 2025 | Fine-tuned Whisper-large-v3 for multi-dialect ASR and developed wav2vec2 classifier for ADI                | Noisy real-world audio requires further improvement   | NADI 2025 (~30.72 hours across 8 dialects)                  | Whisper-large-v3 + Wav2Vec 2.0/HuBERT | WER 38.5% for ASR, CER 14.5% for ADI (top rank)            |
| [15] | 2025 | Implemented RAG to reduce hallucinations in Arabic legal QA                                                | Limited to family code domain                         | ~2,500 custom QA pairs from Arabic Family Code              | Multilingual embeddings + LLM         | BERTScore-F1 70%, 25% fewer hallucinations                 |
| [16] | 2026 | Replaced heavy alignment with CTC and bi-word interleaving for true dual-streaming TTS                     | Alignment still requires careful tuning               | Public dataset LibriTTS multi-speaker English speech        | LLM variants + CTC aligner            | Much lower latency and high human-like similarity          |
| [17] | 2024 | Generated high-quality speech in 1-2 steps using consistency models, skipping teacher-student distillation | Weighting needs precise tuning                        | LJSpeech (single-speaker) + LibriTTS (multi-speaker)        | Consistency Models + Weighted Sampler | Vastly superior RTF with excellent audio quality           |
| [18] | 2025 | Provided a comprehensive taxonomy of Arabic QA models and highlighted the need for RAG                     | Very few domain-specific Arabic QA datasets available | Arabic-SQuAD, ARCD, TyDiQA-Arabic + upcoming Fatwa dataset  | AraBERT, MARBERT, ArabicT5            | Clear roadmap and strong emphasis on RAG                   |
| [19] | 2025 | Created the first Saudi legal reasoning benchmark using real court cases                                   | Lacks voice integration or full RAG                   | 13,000+ real Saudi commercial court cases                   | Various LLMs (e.g., AraBERT)          | strong verdict prediction and reasoning performance        |
| [20] | 2025 | Developed RAG with LLM agents for Arabic legal/fatwa QA                                                    | Restricted to fatwa/Islamic domain                    | Custom fatwa QA pairs                                       | LLM agents + hybrid retrieval         | BERTScore-F1 70.94%, outperformed GPT-4o                   |

|      |      |                                                                                    |                                             |                                     |                                   |                                                         |
|------|------|------------------------------------------------------------------------------------|---------------------------------------------|-------------------------------------|-----------------------------------|---------------------------------------------------------|
| [21] | 2025 | Achieved sub-second responses fully on device with no cloud GPUs                   | Limited to Apple Silicon hardware           | Standard ASR and LLM benchmarks     | MoE ASR + Mixtral LLM via MLX     | Real-time performance with full privacy                 |
| [22] | 2025 | Proposed two-phase RAG with LoRA fine-tuning for Arabic legal reasoning            | Domain-limited to inheritance               | Custom Islamic QA pairs             | Fanar-1-9B + LoRA RAG             | 85.8% accuracy                                          |
| [23] | 2025 | Introduced lightweight streaming TTS compatible with any LLM and supporting Arabic | Not yet integrated with domain-specific RAG | Multilingual Arabic speech datasets | LLMVoX (30M parameters)           | CER ~8% with very low latency                           |
| [24] | 2025 | Systematically evaluated RAG components for Arabic text                            | Not specific focus on the legal domain      | Arabic corpora for RAG testing      | various embeddings (e.g., BGE-m3) | Improved recall@k and better Arabic morphology handling |

## References

- [1] N. J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, 2025.
- [2] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*, MIT Press, 2022.
- [3] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, MIT Press, 2023.
- [4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Draft, 2023.
- [5] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, MIT Press, 2023.
- [6] S. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2024.
- [7] A. Baevski et al., "A Survey on Modern Speech Recognition Systems and Architectures," *arXiv preprint*, 2025.
- [8] J. Shen et al., "Advances in Neural Text-to-Speech Systems: A Survey," *arXiv preprint*, 2025.
- [9] J. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint*, 2023.
- [10] V. Ethiraj, A. David, S. Menon, and D. Vijay, "Toward Low-Latency End-to-End Voice Agents for Telecommunications Using Streaming ASR, Quantized LLMs, and Real-Time TTS," *arXiv preprint arXiv:2508.04721*, 2025.
- [11] D. Zhang et al., "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 15757–15773.
- [12] I. Besrour, J. He, T. Schreieder, and M. Färber, "Ragenta: Multi-agent retrieval-augmented generation for attributed question answering," *arXiv preprint arXiv:2506.16988*, 2025.

- [13] Ö. T. Özyilmaz, M. Coler, and M. Valdenegro-Toro, "Overcoming data scarcity in multi-dialectal arabic asr via whisper fine-tuning," *arXiv preprint arXiv:2506.02627*, 2025.
- [14] H. Elleuch, Y. Saidi, S. Mdhaflar, Y. Estève, and F. Bougares, "ELYADATA & LIA at NADI 2025: ASR and ADI Subtasks," in *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, 2025, pp. 762–766.
- [15] J. Hrimech, M. Mghari, and Y. Zaz, "Retrieval-augmented generation for Arabic legal information: the family code case study," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 6, pp. 1495–1505, 2025.
- [16] H. Liu, S. Yusuyin, H. Huang, and Z. Ou, "CTC-TTS: LLM-based dual-streaming text-to-speech with CTC alignment," *arXiv preprint arXiv:2602.19574*, 2026.
- [17] X. Li *et al.*, "Cm-tts: Enhancing real time text-to-speech synthesis efficiency through weighted samplers and consistency models," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 3777–3794.
- [18] A. Aloqla, R. Khalifa, and W. Alghamdi, "A Review of Question-Answering Systems Using Deep Learning in the Arabic Language," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29214–29228, 2025.
- [19] H. A. Shairah *et al.*, "ALARB: An Arabic Legal Argument Reasoning Benchmark," in *Proceedings of The Third Arabic Natural Language Processing Conference*, 2025, pp. 389–406.
- [20] S. Al-Huwari, "NEBRAS: A RAG-BASED QUESTION ANSWERING SYSTEM FOR ISLAMIC AND LEGAL GUIDANCE," Faculty of Graduate Studies NEBRAS: A RAG-BASED QUESTION ANSWERING SYSTEM ..., 2025.
- [21] T. Likhomanenko *et al.*, "Chipchat: Low-latency cascaded conversational agent in mlx," *arXiv preprint arXiv:2509.00078*, 2025.
- [22] M. AL-Smadi, "QU-NLP at QIAS 2025 Shared Task: A Two-Phase LLM Fine-Tuning and Retrieval-Augmented Generation Approach for Islamic Inheritance Reasoning," *arXiv preprint arXiv:2508.15854*, 2025.
- [23] S. Shikhar *et al.*, "Llmvox: Autoregressive streaming text-to-speech model for any llm," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 20481–20493.
- [24] J. Alsubhi *et al.*, "Optimizing rag pipelines for arabic: A systematic analysis of core components," *arXiv preprint arXiv:2506.06339*, 2025.

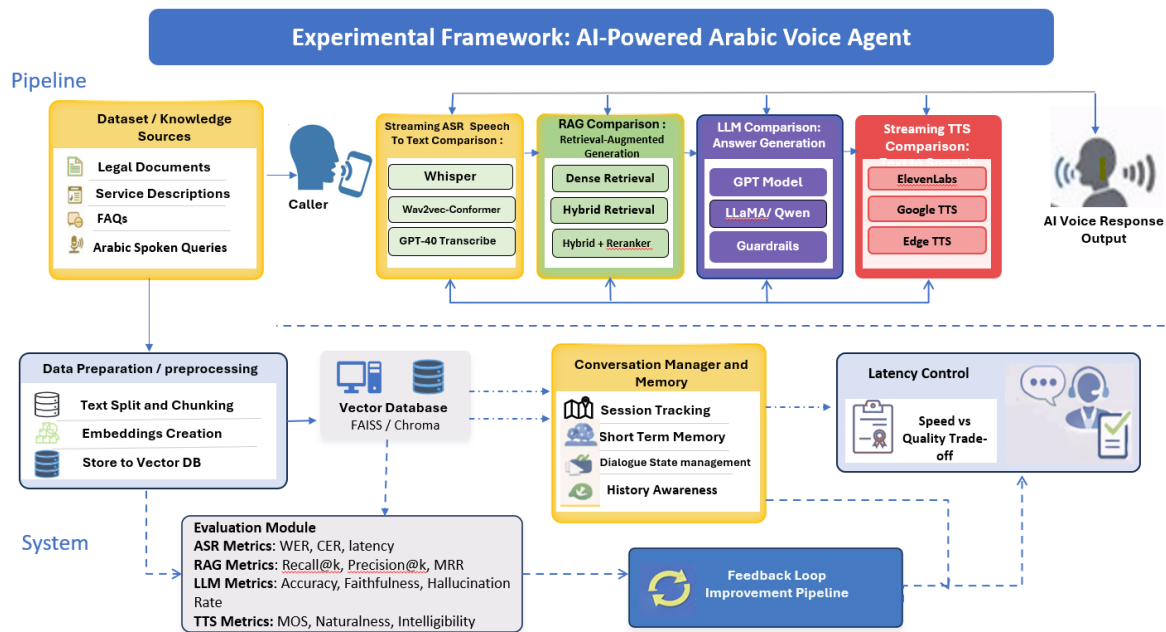


Figure 1: Proposed Experimental Framework for an AI-Powered Arabic Voice Agent