

فاطمه الدغيثر

AI_Voice_ThesisHussam and Fatima v27.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:400629652

Submission Date

Jul 22, 2026, 1:40 PM GMT+3

Download Date

Aug 11, 2026, 12:47 PM GMT+3

File Name

AI_Voice_ThesisHussam and Fatima v27.docx

File Size

3.5 MB

68 Pages

19,141 Words

115,450 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **115 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
-  **24 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 4%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Hidden Text**
104 suspect characters on 2 pages
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 115 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
- 24 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 4% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	<1%
2	Internet	www.mdpi.com	<1%
3	Internet	blog.yueqianlin.com	<1%
4	Internet	argaamplus.s3.amazonaws.com	<1%
5	Internet	webthesis.biblio.polito.it	<1%
6	Publication	Abukhalaf, Seif. "Effective Prompt Augmentation Techniques for OCL Generation ...	<1%
7	Submitted works	Midocean University on 2025-09-05	<1%
8	Submitted works	Coventry University on 2026-04-19	<1%
9	Internet	www.fananews.com	<1%
10	Internet	artificialintelligence.oodles.io	<1%

11	Internet	ijireeice.com	<1%
12	Internet	www.the-saudi.net	<1%
13	Submitted works	City University of Hong Kong on 2026-07-20	<1%
14	Submitted works	Liverpool John Moores University on 2024-05-27	<1%
15	Submitted works	University of Hertfordshire on 2025-06-24	<1%
16	Internet	healthdocbox.com	<1%
17	Internet	researchgold.org	<1%
18	Submitted works	universititeknologimara on 2026-07-17	<1%
19	Submitted works	Liverpool John Moores University on 2026-03-15	<1%
20	Submitted works	University of Portsmouth on 2026-05-07	<1%
21	Internet	ir.uitm.edu.my	<1%
22	Submitted works	lahore garrison LGU on 2026-07-10	<1%
23	Internet	repository.aaup.edu	<1%
24	Submitted works	Brunel University on 2026-04-17	<1%

25	Internet	multiarticlesjournal.com	<1%
26	Internet	ri.itba.edu.ar	<1%
27	Internet	thesis.caltech.edu	<1%
28	Submitted works	German University of Technology in Oman on 2026-07-11	<1%
29	Internet	mail.ijra.net	<1%
30	Internet	www.shine.com	<1%
31	Submitted works	University of Bristol on 2022-10-11	<1%
32	Submitted works	University of New York Tirana on 2026-06-23	<1%
33	Submitted works	University of Technology, Sydney on 2026-05-18	<1%
34	Internet	dspace.lboro.ac.uk	<1%
35	Internet	ejournal.abcollab.id	<1%
36	Internet	umpir.ump.edu.my	<1%
37	Internet	www.medrxiv.org	<1%
38	Submitted works	King Mongkut's University of Technology Thonburi on 2026-05-22	<1%

39	Publication	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Re...	<1%
40	Publication	Wu, Ting-Wei. "How Machine Understands you and me: Naturally Learn and Gene...	<1%
41	Internet	dspace.library.uvic.ca	<1%
42	Internet	repository.najah.edu	<1%
43	Internet	5dok.net	<1%
44	Submitted works	The University of the West of Scotland on 2026-04-22	<1%
45	Submitted works	Thesis 2026 on 2026-06-23	<1%
46	Publication	Xingyu Feng, Dekang Guo, Mingshun Ye. "DMoE-LLM: A Dual-Branch Mixture-of-E...	<1%
47	Internet	developer.nvidia.com	<1%
48	Internet	www.saudi-expatriates.com	<1%
49	Publication	Cerveira, Diogo Dória Príncipe Botelho. "Offline Surgical Phase Timing Using Dee...	<1%
50	Submitted works	IT Department on 2025-09-08	<1%
51	Submitted works	Indian Institute of Technology, Madras on 2026-03-11	<1%
52	Submitted works	King Abdulaziz University on 2025-06-16	<1%

53	Submitted works	Loughborough University on 2026-05-28	<1%
54	Submitted works	Queen's University of Belfast on 2026-04-25	<1%
55	Submitted works	The Robert Gordon University on 2026-04-05	<1%
56	Submitted works	University of Bucharest on 2026-06-29	<1%
57	Internet	archives.univ-biskra.dz	<1%
58	Internet	www.uniphore.com	<1%
59	Publication	"Computational Science – ICCS 2025 Workshops", Springer Science and Business ...	<1%
60	Publication	Faizan Hamayat, Luqman Ejaz, Muhammad Danish, Ahsen Nazir, Pegah Ahadian, ...	<1%
61	Submitted works	Hanken School of Economics on 2023-09-19	<1%
62	Submitted works	Liverpool John Moores University on 2026-05-21	<1%
63	Submitted works	National Institute of Technology Warangal on 2026-07-20	<1%
64	Submitted works	University of Mauritius on 2026-05-17	<1%
65	Submitted works	University of Nottingham on 2025-09-09	<1%
66	Submitted works	University of Westminster on 2026-04-17	<1%

67	Submitted works	Vellore Institute of Technology on 2026-06-30	<1%
68	Submitted works	Vignana Bharathi Institute of Technology on 2026-05-15	<1%
69	Internet	digitallibrary.usc.edu	<1%
70	Internet	ijirt.org	<1%
71	Internet	mdpi-res.com	<1%
72	Internet	opus.bibliothek.uni-augsburg.de	<1%
73	Internet	www.asialeds.org	<1%
74	Internet	www.nigerianjournalonline.com	<1%
75	Publication	"Advances in Information Retrieval", Springer Science and Business Media LLC, 20...	<1%
76	Publication	"Cognitive Computing - ICC 2025", Springer Science and Business Media LLC, 2026	<1%
77	Publication	"General Audio Signal Processing with Deep Learning", Springer Science and Busi...	<1%
78	Submitted works	Charles University in Prague on 2026-07-16	<1%
79	Publication	Jakaria Habib, Monoarul Islam Srabon, Mst Lutfun Nahar, Mariam Lutfy et al. "Ge...	<1%
80	Submitted works	Liverpool John Moores University on 2025-07-21	<1%

81	Publication	Markonda Patnaikuni, Krishna Kanth. "Real-Time Detection of Regulatory Violatio...	<1%
82	Submitted works	UNICAF on 2024-12-21	<1%
83	Submitted works	University of Exeter on 2026-04-29	<1%
84	Submitted works	University of Glasgow on 2026-04-16	<1%
85	Submitted works	Vrije Universiteit Amsterdam on 2026-07-20	<1%
86	Internet	aclanthology.org	<1%
87	Internet	arabicnlp2025.sigarab.org	<1%
88	Internet	d-nb.info	<1%
89	Internet	dokumen.pub	<1%
90	Internet	download.bibis.ir	<1%
91	Internet	drum.lib.umd.edu	<1%
92	Internet	lrec-conf.org	<1%
93	Internet	odr.chalmers.se	<1%
94	Internet	www.ijcaonline.org	<1%

95	Internet	www.techrxiv.org	<1%
96	Publication	Bacelar, Patrícia Marquês. "Developing an AI Virtual Assistant for Higher Educatio...	<1%
97	Submitted works	Erasmus University Rotterdam on 2026-06-15	<1%
98	Publication	Gloria Corpas Pastor, Carlos Manuel Hidalgo-Ternero. "Perspectives on Technolog...	<1%
99	Submitted works	Indian Institute of Technology, Madras on 2026-03-11	<1%
100	Submitted works	Radboud Universiteit on 2026-06-15	<1%
101	Submitted works	University of Melbourne on 2026-05-31	<1%
102	Submitted works	Otto-von-Guericke-Universität Magdeburg on 2025-03-24	<1%
103	Submitted works	University of Birmingham on 2025-09-14	<1%
104	Internet	docslib.org	<1%



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

43

Table of Contents

Abstract	4
List of Abbreviations	5
Chapter 1: Introduction and Overview	7
1.1 Introduction	7
1.2 Problem Definition	7
1.3 Thesis Objectives	8
1.4 Thesis Contributions	8
1.5 Structure of the Thesis	9
Chapter 2: Background and Related Work	10
2.1 Introduction and Background	10
2.1.1 Artificial Intelligence (AI)	10
2.1.2 Machine Learning (ML)	10
2.1.3 Deep Learning (DL)	10
2.1.4 Natural Language Processing (NLP)	11
2.1.5 Large Language Models (LLMs)	11
2.1.6 Retrieval-Augmented Generation (RAG)	11
2.1.7 Automatic Speech Recognition (ASR)	11
2.1.8 Text-to-Speech (TTS)	11
2.1.9 Word Error Rate (WER)	11
2.1.10 Role of AI in the Voice Agent System	11
2.2 Related Work	12
Chapter 3: Methodology and Proposed Framework	17
3.1 Introduction	17
3.2 Dataset Description	19
3.3 Data Collection and Web Scraping	22
3.4 Data Preprocessing	23
3.4.1 Schema Standardization and Data Integration	23
3.4.3 Archiving and Exclusion of Repealed Articles	25
3.4.4 FAQ Cleaning and Leakage-Safe Splitting	25
3.5 Knowledge Base and Retrieval System	27
3.6 Hybrid RAG, Reranking, and Reliability Gate	30
3.7 Answer Generation Module	30

6

5

45

91

23

3.7.1 Intent and Reliability Gates.....31

The intent gate divides the query into one of several types: legal domain information request, numerical entitlement, conversational request, and out of scope. The legal information requests go through the RAG process; the numerical requests are addressed to the deterministic calculator; the conversational requests get the controlled responses; and out of scope requests get the safe refusal. 31

Prior to providing the final answer, the system performs routing and evidence checks. The reliability threshold regulates the legal generation on the basis of the retrieval; not the deterministic computation itself. A common post-processing layer then checks the answer format, cited evidence, Arabic readability, and the controlled refusal.31

3.7.2 Language Models and Prompting Strategy31

The generation layer includes two locally hosted instruction-tuned LLMs, ALLaM-7B and Qwen2.5-7B [37], [38] with identical retrieval and prompting strategies. They are run with 4-bit NF4 quantization [39]. Generation is limited to 160 tokens to control the latency and avoid long answers.31

The prompting strategy requires generating Arabic answer solely based on the retrieved Saudi Labor Law or HRSD FAQ evidence. The article number should be used if there is one and unsupported speculation is prohibited. The model is supposed to generate safe refusal if the evidence is not sufficient enough.3.7.3 Deterministic Legal Calculation Layer31

3.8 Conditions for triggering a deterministic calculation.....32

3.9 Calculator and RAG Evidence33

3.10 Reliability Threshold Calibration.....34

3.10.1 Selection Criterion34

3.10.2 Decision process for Reliability Gate34

3.11 Voice Processing Layer.....35

3.12 Conversational Memory Module36

3.12.1 Memory Design: Conversation Memory versus Retrieval Memory.....36

3.13 Evaluation Approach.....36

3.13.1 Retrieval Evaluation Metrics37

3.13.2 Answer Generation Metrics38

3.13.3 Voice and Latency Metrics39

3.13.4 Statistical Testing and Ablation Metrics40

3.15 Summary42

Chapter 4: Results and Discussion43

4.1 Introduction43

4.2 Experimental Setup43

4.2.1 Technical Environment43

4.2.2 Software Libraries and Versions.....	44
4.2.3 Data Split and RAG Parameters.....	45
4.3 Data Preparation Results	47
4.4 Retrieval Evaluation Results	48
4.4.1 Evaluation Sets and Sample Sizes.....	48
4.4.2 Comparison of Retrieval Methods	49
4.4.3 Final Legal Retrieval Test.....	50
4.4.4 Supplementary Robustness Check	50
4.5 Answer Generation Results.....	51
4.6 Statistical Evaluation.....	52
4.7 Voice Layer Results	53
4.8 No-RAG versus RAG Ablation	54
4.9 Latency Results	55
4.10 Deterministic Calculation Layer Implementation Checks	57
4.11 Overall Results Summary	57
4.12 Discussion	59
4.13 Limitations	60
4.14 Summary	60
Chapter 5: Conclusion and Future Work.....	61
5.1 Conclusion	61
5.2 Contributions.....	62
5.3 Limitations	62
5.4 Future Directions.....	62
5.5 Final Thoughts	62
References	64
Appendix A: Automated Evaluation Rubric and Judge Prompt	67
Appendix B: Data-Lineage and Deduplication Record	68

Abstract

The recent breakthroughs in LLMs and RAG techniques allowed designing QA systems capable of generating contextually relevant and evidence-based answers. The use of ASR and TTS technologies may allow deploying interactive voice services for such purposes in sensitive sectors like legal information. However, its application in the Arabic language faces additional problems of dialect/accents variation

recognition, ensuring the relevance of generated answers to authoritative sources, and maintaining low enough latency of the whole pipeline.

4 This thesis proposes and evaluates an Arabic evidence-based voice agent designed for answering questions about the Saudi Labor Law. The developed agent combines Whisper ASR, knowledge base based on Saudi Labor Law and labor-related materials provided by Ministry of Human Resources and Social Development, hybrid dense and BM25 retrieval, cross-encoder reranking, reliability gate trained on validation data, instruction-tuned language models, deterministic legal calculations, and TTS technology. The final snapshot of frozen knowledge base consists of 211 active records of legal articles and 278 indexed labor-related FAQs.

8 In the tests on manually crafted legal retrieval task with 24 questions, the final pipeline has demonstrated the following results: Recall@1 = 0.8750, Recall@5 = 0.9583, MRR = 0.9097, nDCG@5 = 0.9221. After applying the reliability gate, ALLaM-7B achieved 37 correct predictions out of 45 cases (82.22% vs. 71.11% for Qwen2.5-7B); an exact McNemar test showed that the difference is statistically insignificant (p-value = 0.1250). In the 24-question ablation study, the number of errors in legal citations raised from 1 out of 24 (RAG baseline) to 6 out of 24 (without retrieval). A controlled synthetic-speech test resulted in WER = 0.209 and CER = 0.0484. After optimization, the pipeline generated a full response in 2.96 seconds and answered in 0.82 seconds when using streaming technique.

Overall, it has been demonstrated that developing an evidence-first modular Arabic legal voice assistant is possible under controlled experimental settings. The developed pipeline is currently not ready for real deployment as legal service. Before doing so, it should be evaluated with larger and more diverse group of Saudi Arabic speakers, independently reviewed by legal experts, supplied with fully versioned legal corpus, and tested on bigger held-out test sets.

List of Abbreviations

ADI: Arabic Dialect Identification

AI: Artificial Intelligence

ASR: Automatic Speech Recognition

BM25: Best Matching 25 lexical retrieval algorithm

CER: Character Error Rate

CI: Confidence Interval

CTC: Connectionist Temporal Classification

DL: Deep Learning

FAQ: Frequently Asked Questions

51 FAISS: Facebook AI Similarity Search

GPU: Graphics Processing Unit

HRSD: Ministry of Human Resources and Social Development

32 LLM: Large Language Model

ML: Machine Learning

MRR: Mean Reciprocal Rank

nDCG: Normalized Discounted Cumulative Gain

NF4: NormalFloat 4-bit quantization

40 NLP: Natural Language Processing

NLU: Natural Language Understanding

OOS: Out of Scope

96 QA: Question Answering

RAG: Retrieval-Augmented Generation

RTF: Real-Time Factor

TTS: Text-to-Speech

TTFR: Time-to-First-Response

WER: Word Error Rate

Chapter 1: Introduction and Overview

1.1 Introduction

In this paper, we develop a case study of an Arabic voice agent which is asked to answer questions concerning Saudi Labor Law. The framework utilizes a modular pipeline comprising ASR, RAG, a local LLM, legal deterministic computations, and TTS. The answers are in Arabic and based on evidence collated from the official Saudi Labor Law, developed by the Ministry of Human Resources and Social Development (HRSD).

In this Study, the evaluation process involves various performance metrics, including retrieval effectiveness, answer correctness, citation accuracy, ASR errors, and end-to-end latency. In addition, we explore how retrieval grounding influences the occurrence of ungrounded citations.

1.2 Problem Definition

The Unified Call Center of the Ministry of Human Resources and Social Development (HRSD) receives thousands of calls daily on common issues related to salary, rights, end-of-service benefits, notice period, leave, etc., as per Saudi Labor Law. A traditional Interactive Voice Response (IVR) system exists and is built to be rule-based, with a large number of static menus that can't handle more complicated legal questions in natural language. That means all questions that are not standard or do not have a straightforward answer have to be handled by human agents.

The scarcity of human aid in this challenge is a huge contributing factor. Customer service operations and the ability to respond to legal inquiries are limited to official working hours and outside these hours and for public holidays, staff is not available and no customer service operations can be performed. This limitation in operation makes a significant break in the service continuity. Also, in humans, the quality and speed of their response may depend on the complexity of the case as well as on the completeness of the user's input. This underscores the necessity of an intelligent, 24×7 available conversational voice agent that can understand various natural language queries and provide precise, dependable, and evidence-based legal information without human support.

There are also additional technical issues for voice assistance. The regional variety, accent, dialect, and speech manner complicate Arabic speech recognition in real-time [10], [11]. The retrieved information should also be up-to-date and valid because otherwise the responses based on outdated or irrelevant regulations can mislead the user. Thus, the integration of ASR, hybrid RAG, reliability gate, deterministic legal computation, and TTS is a challenging system design task, but not just a collection of independent components.

Accordingly, the research problem is the creation of a real-time Arabic legal-information assistant which will be able to accept voiced or typed labor law queries, retrieve the relevant official information, provide the answers based on the information, and refuse to answer without enough information. This thesis proposes the modulated approach that uses ASR, hybrid RAG, local hosting of LLMs, validation-calibrated reliability gate, deterministic legal computation, and TTS.

1.3 Thesis Objectives

In this study, the goal is to create and test an Arabic voice agent, which can be used to answer questions about the Saudi Labor Law. The goals are to:

1. Develop an evidence-based that has been versioned using articles from the active Saudi Labor Law and Labor-related HRSD FAQ, tracing and cleaning and with explicit controls on leakage.
2. Evaluate the performance of dense, lexical, hybrid and reranked retrieval configurations on a held-out legal question set and report on quality, uncertainty and robustness of the ranking.
3. Compare two local Arabic- proficient LLMs based on the same grounded prompt and the same reliability gate (correct answer rate, correct refuses, citation, and paired statistical testing).
4. Evaluate the voice and deployment layers based on the results of the ASR error rate, latency, Streaming Time to First-Response and the determinant calculator for the deployed legal provisions.

1.4 Thesis Contributions

The thesis makes five main contributions:

1. End-to-end architecture for Arabic legal voice agent. Pipeline includes both spoken and typed queries about Saudi Labor Law and consists of modular components of automatic speech recognition (ASR), Arabic normalization, intent routing, hybrid dense and BM25 retrieval, cross-encoder reranking, reliability-controlled generation, deterministic legal computations, conversational memory, and Arabic speech synthesis.
2. Structured Arabic legal dataset created from official sources. Datasets were collected from official websites of the Ministry of Human Resources and Social Development and the FAQ section. Every data point is represented either as an article-level legal entry or as a question-answer pair. Source URL, document ID, article number (when applicable), and original text are kept in order to ensure reproducibility.
3. Traceable and leakage-aware knowledge base. Cleaning, normalization, and filtration were performed for the raw datasets to create the searchable knowledge base. Two representations were maintained for each data point: (i) official wording for output and citation and normalized Arabic wording for retrieval; and (ii) indexing, validation and held-out splits, where repealed entries are filtered out of the index in order to not leak evaluation questions into evidence base search.
4. Reliability-controlled answer or refuse mechanism with deterministic computation layer. Threshold for sufficient reliability, which is trained only on validation data, determines whether retrieved evidence can be considered reliable enough for answer generation. In case of insufficient evidence, lack of evidence or out-of-scope questions, the agent makes a controlled refusal. For certain numerical

labor rights, deterministic rule-based functions generate same answers for same inputs. Retrieval layer provides legal provision in this case to ensure that the calculated value can be explained and cited.

5. Component-level evaluation of full system. Evaluation covered the following aspects: data preparation, retrieval, reliability gating, answer generation, paired statistical comparisons, speech recognition, latency, citation ablation, deterministic calculations. No-RAG vs RAG ablation allows for measuring retrieval grounding contribution.

36 1.5 Structure of the Thesis

This thesis is divided into five chapters covering different stages of creating and evaluating an Arabic voice agent powered by AI for providing information about Saudi Labor Law.

56 Chapter 1 states the problem, provides the motivation of creating an Arabic legal voice agent, specifies the aims and contributions, and presents the general structure of the thesis.

89 Chapter 2 discusses the background and previous works. This chapter presents the notions that should be reviewed for studying the issue and provides an overview of these notions such as artificial intelligence, machine learning, deep learning, natural language processing, large language models, retrieval-augmented generation (RAG), automatic speech recognition (ASR), text-to-speech (TTS), and related evaluation metrics. Chapter 2 analyzes the previous works devoted to Arabic NLP, legal QA, RAG system, and voice agent.

Chapter 3 discusses the methodology and the framework. The framework includes the stages of official data collection, cleaning and normalizing, leakage-proof data splitting, constructing a knowledge base, retrieving, reliability gating, answer generation, deterministic legal calculation, conversation history, voice processing, and metrics.

93 Chapter 4 presents and analyzes the experimental results on data preparation, retrieving, answer generation, statistical paired comparison, voice layer, latency, deterministic calculation, and No-RAG versus RAG ablation. This chapter discusses the implications and limitations of the results obtained during the experiment.

34 Chapter 5 discusses the key results and contributions of this thesis, restates the limitations of the framework and provides suggestions for the future work.

Chapter 2: Background and Related Work

2.1 Introduction and Background

The proposed voice agent involves five technology areas: speech processing, information retrieval, language generation, legal grounding, and speech synthesis. In this chapter, we present the concepts and literature behind understanding the design choices discussed in this thesis. We start with basic concepts of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). Then, we investigate technologies involved in this project: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and Word Error Rate (WER). Altogether, these concepts will shed light on how the Arabic input is handled, how legal evidence is retrieved, how the response is generated and evaluated, and how text and speech are evaluated.

The second part reviews the voice agents, Arabic speech recognition, Arabic question answering, legal RAG, and low-latency speech synthesis. We find an existing gap which is filled by our approach: a local pipeline of the voice-agent that retrieves legal evidence in official Arabic language, uses hybrid retrieval, reliability controls, deterministic computation, and voice interaction.

2.1.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) covers the computer science discipline concerned with developing machines that perform intelligent tasks such as learning, reasoning, problem-solving, and decision-making [1].

2.1.2 Machine Learning (ML)

Machine Learning (ML) falls under the AI umbrella, where the learned patterns are used to make decisions or predictions from data using models without task-specific programming. The approaches of ML involve statistical learning and optimization, and the higher the training dataset is, the better the performance of models usually is [2].

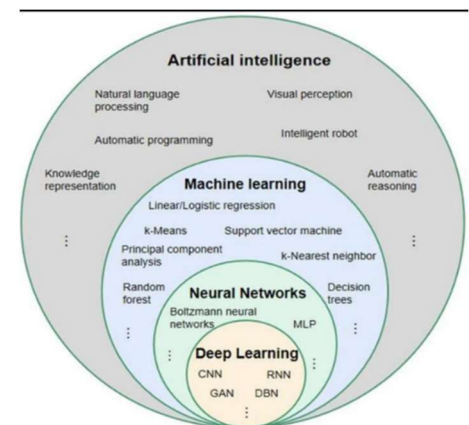


Figure 2.1. Relationship between artificial intelligence, machine learning, neural networks,

2.1.3 Deep Learning (DL)

DL is one of the subfields of AI that involves the use of deep neural network models with multiple layers that learn hierarchical representations from large-scale data. These models are trained to directly extract the features from raw input data, and they have achieved state-of-the-art performance in speech recognition, natural language processing, and computer vision applications [3].

2.1.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) is the name given to computational methods that allow computers to manipulate, understand, and create human language using data-driven models and architectures of deep neural networks. Recent advances in NLP have been driven largely by large language models, which have led to 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-scale text corpora to comprehend, generate, and manipulate human language. By learning contextual patterns from data, LLMs achieve a wide range of natural language capabilities, including text generation, question answering, and reasoning [5].

2.1.6 Retrieval-Augmented Generation (RAG)

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

2.1.7 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) refers to a machine's ability to recognize spoken words and transcribe them into text. ASR relies on deep learning models that learn acoustic and linguistic patterns. State-of-the-art ASR systems are built as end-to-end neural networks trained on large-scale datasets, and they perform well for real-time recognition and multilingual tasks [7].

2.1.8 Text-to-Speech (TTS)

Text-to-Speech (TTS) refers to the ability to convert text into a human-like voice. The nature-sounding speech synthesis is realized by deep learning models, and state-of-the-art TTS systems can synthesize nature-sounding voices with natural intonation [8].

2.1.9 Word Error Rate (WER)

WER, which is defined as the number of word errors per 100 words between the predicted and reference transcript, is a common performance metric of ASR systems. The measure takes into consideration the word level substitution, deletion and insertion errors and provides a normalized value [9]. WER has been used in particular for the evaluation of Arabic ASR [10].[11].

2.1.10 Role of AI in the Voice Agent System

The voice-agent pipeline is based on the computational facilities offered by artificial intelligence. ASR is used to convert the speech query to text. Language processing and intent detection next identify if retrieval, deterministic calculation, controlled conversational response or safe refusal is appropriate. When the user requests a voice response, the LLM generates an answer, validated by RAG, to the query for Arabic legal information seeking. In the legal domain, every transition is significant since it may affect

the meaning and trustworthiness of the overall response, depending on the correctness of the transcription, of the classification of the intent, of retrieval, or of generation.

2.2 Related Work

Review of Related Work

In this study, authors [10] fine-tuned OpenAI Whisper model for multi-dialectal Arabic ASR covering four major dialects: Gulf, levantine, Iraqi and Egyptian and maghribi. Experiments employed the Arabic Mozilla Common Voice data, supplemented with MASC dialectal speech, and were related to Whisper-small compared to other larger Whisper models. It was observed that fine-tuning of smaller models can result in a significant boost in performance for Arabic ASR and, in some cases, match that of non-fine-tuned larger models as well. This paper emphasizes that the evaluation needs to be dialect-specific for Arabic voice agents.

The authors [11] described the ELYADATA & LIA system for the NADI 2025 Arabic speech-processing shared task. The work addressed spoken Arabic dialect identification and multi-dialectal ASR. A lightly adapted Whisper-large-v3 encoder achieved 79.83% dialect-identification accuracy, while separately adapted SeamlessM4T-v2 Large models achieved an average WER of 38.54% and CER of 14.53%. The results demonstrate the value of large pretrained speech models and focused adaptation, but the study does not address legal retrieval or answer generation.

The authors [12] presented an end-to-end voice-processing pipeline for telecommunications customer support and interactive voice response (IVR). The system combined streaming ASR, FAISS-based RAG over RFC documents, a 4-bit quantized language model, and real-time TTS. Evaluation on 500 human-recorded telecommunications utterances reported an average end-to-end latency of approximately 0.934 s. The study demonstrates how sentence-level streaming, concurrent processing, retrieval grounding, and quantization can improve responsiveness, but it does not address Arabic legal question answering.

The authors [13] introduced SpeechGPT, a speech-language model that supports speech perception and generation within a large-language-model architecture. The model uses discrete speech representations and the SpeechInstruct speech-text instruction dataset and was trained through modality-adaptation pretraining, cross-modal instruction fine-tuning, and chain-of-modality instruction fine-tuning. The work demonstrates the potential of integrated speech-text interaction, but it does not address Arabic legal information, RAG-based grounding, or reliability-controlled legal generation.

The authors [14] presented a multi-agent RAG system for attributed question-answering named RAGentA. Prediction, document filtering, citation-aware answer generation and answer revision are carried out by four specialized agents. The system integrates BM25 sparse retrieval, E5 dense embeddings and uses a Pinecone vector store, and was tested over 500 synthetic question-answer pairs generated using FineWeb. Results demonstrate improvements for Recall@20 and faithfulness compared to a normal RAG baseline. While the study doesn't apply to Arabic legal assistance, it is applicable to attribution, source grounding, and answer faithfulness.

In an work by Hrimech et al. [15], researchers built a Arabic legal RAG system in the Moroccan Family Code. The authors created some 2500 documents of natural question answering instances in Arabic and tested several multilingual embedding model, namely BGE-M3, E5-large and E5-small, for legal information retrieval. The study also applied clause segmentation to maintain a legal provision's structure. Improved results of its outcomes prove the legitimacy of applying RAG and structured segmentation for Arabic legal text. But, it concerned only Arabic family law and not Arab speech interaction; also not the calculation of deterministic labor-entitlement..

3

Liu et al. [16] introduced CTC-TTS, a dual-streaming text-to-speech framework designed to reduce the latency of LLM-based speech synthesis. Instead of relying on a computationally expensive forced-alignment pipeline, the framework employs a CTC-based aligner and bi-word interleaving. Two variants were presented: CTC-TTS-L, which prioritizes synthesis quality, and CTC-TTS-F, which prioritizes lower latency. This work is relevant to the voice layer of the proposed system, but it does not address Arabic legal question answering or RAG-based legal grounding.

3

The authors [17] introduced a new text-to-speech system called consistency models and weighted samplers (CM-TTS), which reduced the number of steps involved in synthesis and made the TTS system more real-time-like, without using an adversarial training or the teacher model approach. The method was tested on the VCTK, LJSpeech, LibriSpeech and LibriTTS corpora. While it could be applicable to efficient speech synthesis it is not Arabic-specific and does not address the issue of legal-information interaction.

88

Aloqla et al. [18] conducted a review on the deep learning approaches of Arabic Question Answering systems based on extractive, generative and hybrid approaches. This review presents the difficulties encountered due to Arabic morphology problems, orthographic ambiguities, missing diacritics, dialectal variations and less domain-specific dataset.

52

In addition, the review covers the English and Arabic transformer models and datasets, such as Arabic-SQuAD, ARCD, TyDiQA-A, AraBERT, and mBERT. It highlights the importance of Arabic question answering specific to its domain and discusses about the possible use of retrieval augmented generation. To build an Arabic legal-argument-reasoning benchmark, named ALARB, Abu Shairah et al. [19] used over 13000 Saudi commercial court cases.

87

The ALARB benchmark covers case facts, reasoning, verdicts, and cited laws, and supports tasks such as verdict classification, legal-argument completion, and article identification. It is an important resource for Saudi Arabic legal reasoning, but it does not provide a voice-agent pipeline or a Saudi Labor Law question-answering service. Al-Huwari [20] presented NEBRAS, a RAG-based question-answering system for Islamic and legal guidance. The system applies agentic and hybrid retrieval with LLM agents and evaluates Islamic-fatwa and university help-desk domains using automatic metrics and human evaluation.

42

NEBRAS is relevant to retrieval grounding in sensitive Arabic guidance, but it addresses domains different from Saudi Labor Law voice assistance. The Seamless Communication team [21] introduced SeamlessM4T, a unified multilingual and multimodal model supporting automatic speech recognition,

1

speech-to-speech translation, speech-to-text translation, text-to-speech translation, and text-to-text translation. Although it is not a legal question-answering or RAG system, it demonstrates the potential of integrated multilingual speech and text processing.

The authors of [22] described QU-NLP, a system developed for the QIAS 2025 Islamic inheritance reasoning task. The study fine-tuned Fanar-1-9B using LoRA and integrated it with RAG to identify eligible heirs, apply fixed-share rules, and execute arithmetic calculations. The system achieved an overall accuracy of 85.8% on the final test data. The study is relevant because it combines retrieval grounding with rule-based reasoning and calculation; however, its scope is limited to Islamic inheritance, whereas the pre-sent work focuses on Saudi Labor Law.

The authors [23] designed LLMVoX, a lightweight 30-million-parameter, LLM-agnostic streaming autoregressive TTS model for low-latency speech synthesis. Its multi-queue streaming approach enables continuous synthesis, and an Arabic adaptation reported a CER of approximately 8%. The study supports separating LLM reasoning from the TTS layer, but it does not address Arabic legal retrieval or legal question answering.

The authors [24] conducted a systematic empirical analysis of Arabic RAG components across six datasets. The study compared chunking strategies, embedding models, rerankers, and language models using RAGAS metrics. Sentence-aware chunking performed strongly, while BGE-M3 and multilingual-E5-large were among the best embedding models; bge-reranker-v2-m3 also improved faithfulness on complex datasets. This work informed the component-level evaluation and model-selection approach adopted in this thesis.

Collectively, the reviewed studies address low-latency voice pipelines, Arabic dialect ASR, Arabic question answering, Arabic legal RAG, streaming TTS, and component-level RAG optimization. None of them combines official Saudi Labor Law and HRSD evidence, Arabic speech interaction, hybrid retrieval with reranking, a validation-calibrated reliability gate, deterministic legal calculation, and component-level evaluation within a single locally hosted Arabic legal voice-agent pipeline. This thesis addresses that integration and evaluation gap.

Table 2.1. Comparison of the reviewed studies.

Paper	Year	Main contribution	Limitation	Data	Model/method	Reported result
[10]	2025	Whisper fine-tuning for five major Arabic dialects	Dialectal resources remain limited	Common Voice MSA + MASC dialect data	Whisper-small and Whisper-large-v3	Fine-tuned smaller Whisper models approached larger-model performance in some settings
[11]	2025	Arabic dialect identification and multi-dialect ASR for NADI 2025	Shared-task setting; substantial ASR error remains	NADI 2025 Arabic speech data	Whisper-large-v3 encoder for ADI; SeamlessM4T-v2 Large for ASR	ADI accuracy 79.83%; ASR WER 38.54%, CER 14.53%
[12]	2025	Streaming telecom voice-agent pipeline with RAG, quantized LLM, and real-time TTS	Private telecommunications setting; not Arabic legal QA	500 human-recorded telecom utterances	T-Transcribe Engine + FAISS RAG + TSLAM-Mini-2B + T-SYNTH	Average end-to-end latency $\approx$ 0.934 s

[13]	2023	Speech perception and generation within a large-language-model architecture	Not designed for Arabic legal grounding or RAG-based legal QA	SpeechInstruct and related speech-text instruction data	LLaMA-based model + discrete speech units	Demonstrated cross-modal speech and text instruction-following abilities
[14]	2025	Multi-agent attributed QA with citation and revision agents	Synthetic benchmark; not Arabic legal or voice-based	500 FineWeb-derived synthetic QA pairs	Falcon-3-10B agents + BM25 + E5 + Pinecone	Hybrid retrieval improved Recall@20; faithfulness improved over standard RAG
[15]	2025	Arabic legal RAG for Moroccan Family Code with clause-based segmentation	Different jurisdiction; no Arabic speech interaction or labor-law calculation	Approximately 2,500 Arabic QA pairs	BGE-M3, E5-large, E5-small + LLM	Legal retrieval and generation evaluation for Moroccan family law
[16]	2026	Dual-streaming TTS using CTC alignment and bi-word interleaving	Preprint; TTS-focused and not Arabic legal QA	VoiceAssistant400 K / LibriSpeech-based evaluation	CTC-TTS-L and CTC-TTS-F	FPL-A = 159 ms (CTC-TTS-F) and 210 ms (CTC-TTS-L); WER/CER = 1.80/1.04 and 1.50/0.79, respectively
[17]	2024	Consistency-model TTS with weighted samplers for efficient synthesis	Not Arabic-specific and not legal-interaction focused	VCTK, LJSpeech, LibriSpeech, and LibriTTS (train-clean-100 for the zero-shot setting)	Consistency model + weighted sampler	RTF 0.02-0.07 depending on generation steps
[18]	2025	Review of deep-learning Arabic QA systems, models, datasets, and challenges	Survey paper; not a new experimental system	Arabic QA datasets surveyed	AraBERT, MARBERT, ArabicT5, mBERT, XLM-R, and other models	Taxonomy of Arabic QA methods and identified research gaps
[19]	2025	Saudi Arabic legal reasoning benchmark using authentic court cases	No voice-agent or RAG service pipeline	13,000+ Saudi commercial court cases	Legal reasoning benchmark tasks for LLMs	Verdict prediction, reasoning-chain completion, and article-identification tasks
[20]	2025	Agentic RAG system for Islamic and legal guidance	Master's thesis; different domain from Saudi Labor Law voice assistance	Custom Islamic fatwa and university help-desk datasets	Hybrid retrieval + LLM agents	BERTScore-F1 = 70.94% in Islamic fatwa evaluation
[21]	2023	Unified multilingual and multimodal model supporting ASR, speech translation, text translation, and TTS	Not a legal QA or RAG system	Large multilingual speech and text corpora including Seamless Align	SeamlessM4T / UnitY architecture with w2v-BERT 2.0	Strong multilingual speech and text translation performance across many languages
[22]	2025	Fine-tuning plus RAG for Islamic inheritance reasoning	Narrow inheritance domain; multiple-choice task	QIAS 2025 Islamic inheritance reasoning data	Fanar-1-9B + LoRA + RAG	Overall accuracy = 85.8%

[23]	2025	Lightweight LLM-agnostic streaming TTS	Not integrated with Arabic legal RAG	Streaming TTS benchmarks, including Arabic adaptation	LLMVoX, 30M parameters	Low-latency streaming TTS with Arabic adaptation
[24]	2025	Systematic comparison of Arabic RAG components	Not specific to Saudi legal documents	Six Arabic datasets	Chunking strategies, embedding models, reranker, and generators	Sentence-aware chunking, BGE-M3, multilingual-E5-large, and reranking performed strongly

Chapter 3: Methodology and Proposed Framework

3.1 Introduction

This chapter describes the methodology and suggested structure of an Arabic voice agent system to respond to queries concerning the Saudi Labour Law. It enables both verbal and written Arabic queries and includes automatic speech recognition (ASR), retrieval-augmented generation (RAG), large language models (LLMs), deterministic legal computing, and text-to-speech (TTS).

The system fetches appropriate evidence from a library with provisions of Saudi Labour Law and frequently asked questions published by the Ministry of Human Resources and Social Development (HRSD), to enhance the reliability and traceability of legal response. The retrieved passages with their informational sources are then used to produce grounded Arabic answers that cite the retrieved ones. This helps to minimise unsupported generation and enhances the verifiability of the answers.

The framework shall be developed based on the following three important requirements. Firstly, should be able to accept voice as well as text input in Arabic. Second, answers should be based on legal information retrieved and not just in the internal knowledge of the language model. Third, it shall be scalable, meaning that the important system components for evaluation shall be dissected into multiple modules that can be assessed separately in terms of speech-recognition accuracy, retrieval quality, correctness of answers, quality of citations, speech-synthesis quality and end-to-end latency.

The overall architecture to be proposed consists of input processing module, Arabic text normalisation module, intent classification, hybrid information retrieval module, evidence validation module, grounded answer generation module, safety processing module and speech synthesis (as an optional module). Questions with numbers are addressed as a separate deterministic computing path which follows certain rules derived from legal norms relating to the question. In the next sections, the knowledge base, the system architecture, their component implementation, and the evaluation methodology are discussed.

Proposed architecture of the evidence-grounded Arabic voice agent for Saudi Labor Law inquiries

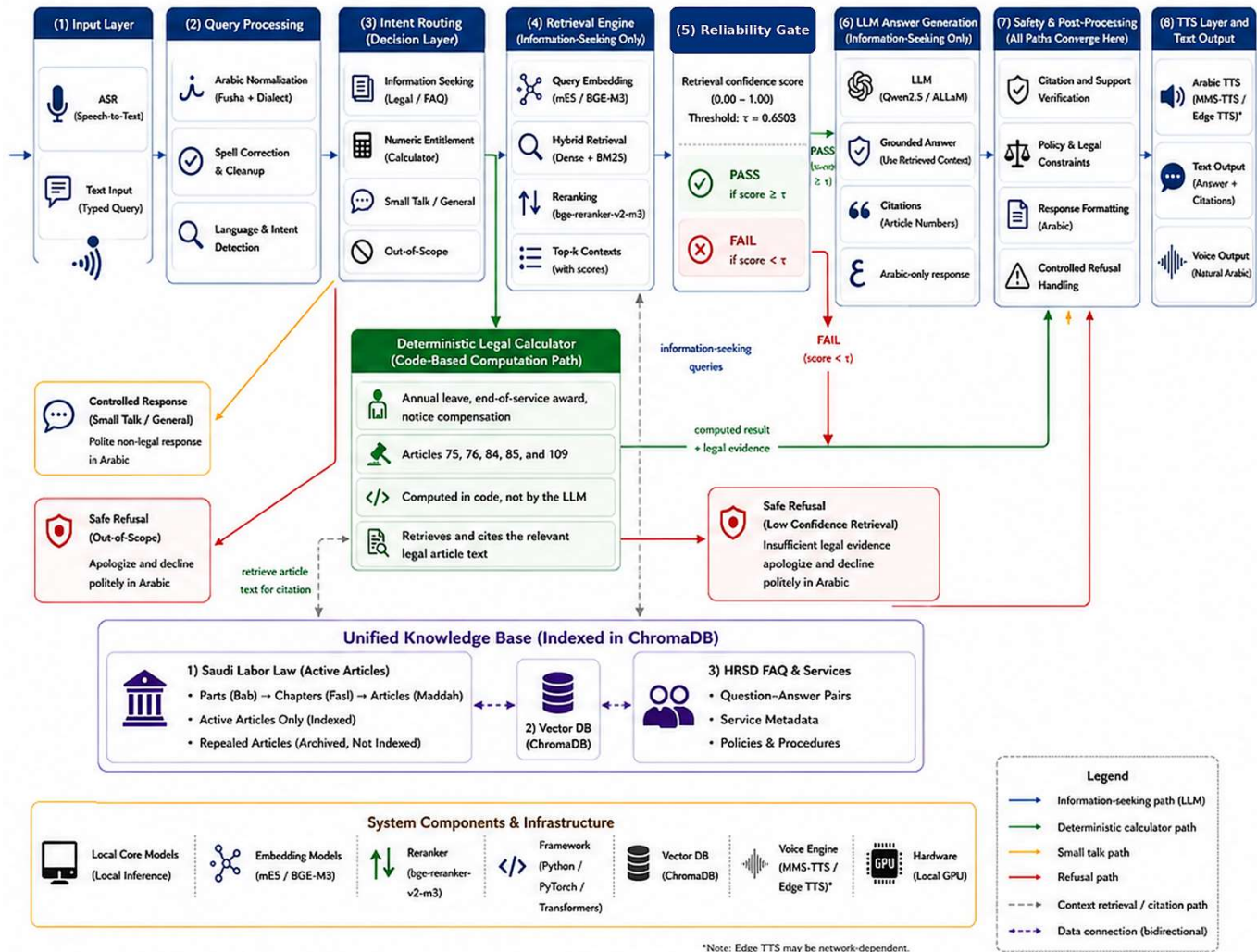


Figure 3.1. Runtime architecture and intent-based routing paths of the proposed evidence-grounded Arabic voice agent.

Figure 3.1 summarizes the runtime architecture and its intent-based routing. After query processing, legal information-seeking questions proceed to retrieval and grounded generation; numerical-entitlement requests are directed to the deterministic calculation branch; and small-talk or out-of-scope inputs receive controlled responses or safe refusals. All paths converge at the safety and post-processing layer before the final text or voice response is returned.

Step 1: User Input Handling. The system accepts Arabic queries as text or speech. Typed queries pass directly to query processing; spoken queries are first transcribed by the ASR module using Whisper-large-v3 or the lower-latency Whisper-small.

Step 2: Query Processing. After being cleaned and normalized, the Arabic text is made ready for routing by detecting its language and intent.

Step 3: Intent Routing. The query is sorted into one of the four routes previously mentioned: Retrieval engine, deterministic calculator, controlled response, or safe refusal.

75 Step 4: Retrieval and Evidence Selection. For information-seeking queries, the retrieval engine runs dense semantic retrieval and BM25 lexical retrieval over the legal articles and FAQ records in the ChromaDB knowledge base [25]. The process works as follows. Normalized dense and lexical scores are merged into the hybrid score, which is further used for re-ranking candidate documents by the cross-encoder. The top-scoring articles or FAQ entries become the evidentiary base.

76 78 Step 5: Reliability Gate. The highest confidence score is checked against the calibrated threshold $\tau = 0.6503$. If the score is not lower than the threshold, the supporting evidence is forwarded to the Large Language Model (LLM). Otherwise, the system rejects the request with a safe error message instead of producing an invalid legal answer.

81 Step 6: Answer Generation with groundings. The output of the LLM is an answer in Arabic, supported only by selected evidence from Step 4, if available, followed by the number of the article or source reference. The prompt tells the model not to mention any ulterior facts that aren't provided or not to use its own knowledge.

20 Step 7: Post-processing and safety. This is the step in which the processing paths converge together. The layer verifies correctness of the answer in the light of the evidence obtained, verifies the correctness of the reference or quotation, the legal or policy limits, formats the Arabic answer and manages controlled refusal.

Step 8: Text and Voice Output. The response, validated in Step 7, is presented as Arabic text. If the mode of operation involves voice, the response is additionally forwarded to the text-to-speech (TTS) pipeline and presented as Arabic speech.

Legal calculations run in parallel with the retrieval-LLM path rather than being an additional numbered step. Whenever the query asks for a numerical entitlement, which can be calculated by rule-based functions from the values in the Articles 75, 76, 84, 85, and 109 of the Saudi Labor Law [26], deterministic legal computation is performed. No calculations are done by the LLM. The relevant article is retrieved and cited to substantiate the calculated value, and the output goes through the same safety and post-processing layer.

57 3.2 Dataset Description

4 The dataset in this study was collected from two officially published sources in Saudi Arabia – the Saudi Labor Law articles pages and the Ministry of Human Resources and Social Development (HRSD) FAQ portal [26], [27]. Legal documents were collected at the level of an article, whereas HRSD data included service-oriented question-answer content. Every single record in the dataset contained information on the source URL, source type, original text, normalized retrieved text, and document identifier.

Legal documents collection included 249 article records. Status screening left 211 active articles to be indexed, 38 repealed articles were put aside out of the searchable collection to make sure that outdated

provisions will not be retrieved. Experiments used the fixed legal collection generated by the data-preparation pipeline, but not the dynamic legal source itself for evaluation purposes.

HRSD extraction provided 711 question-answer pairs. Basic cleaning resulted in 703 valid question-answer pairs, domain filtering led to the collection of 402 labor-relevant question-answer pairs. Leakage and near-duplicates screening resulted in 400 question-answer pairs eligible for the identifier-based split. It corresponded to 399 unique question texts, as there was one question represented in two separate source records.

70/10/20 identifier-based split yielded 278 indexing question-answer pairs, 41 validation pairs, and 81 held-out pairs. Only indexing question-answer pairs were added to the FAQ knowledge base. Exact-question text deduplication in the held-out partition showed one duplicate of the question in `faq_id` 248 and `faq_id` 94. Thus, the final manifest included 80 unique held-out HRSD questions: `faq_id` 248 remained as the only valid one, whereas `faq_id` 94 was deleted before scoring. The decision about deduplication was made based on data quality and not because of the possible performance issues.

Evaluation resources were intentionally divided into sets according to their purpose. The final set of legal retrieval test included 24 manually created questions. Another set of 201 automatically generated legal questions was intended exclusively as a robustness check; it was not used to define the reliability threshold or to serve as the main benchmark of the legal retrieval system.

For the paired LLM comparison, 45 independently collected stratified questions were used, including 14 HRSD questions, 17 legal questions, 12 out-of-scope questions, and 2 small talk inputs instead of 80 held-out HRSD questions. All 45 questions passed through the same fixed reliability-gated pipeline, and both grounded answers and appropriate safe refusals were scored against the task-specific rubric.

Table 3.1. Dataset construction and cleaning summary.

Source / Stage	Processing step	Records after step	Explanation
Saudi Labor Law	Scraped from official HRSD legal pages	249	Article-level legal records collected.
	Active articles retained	211	Active articles indexed in the searchable knowledge base.
	Repealed articles archived	38	Repealed articles excluded from retrieval to avoid obsolete citations.
HRSD FAQ	Scraped from official HRSD FAQ pages	711	Initial question-answer records collected.
	Basic cleaning	703	Empty, malformed, or unusable records removed.
	Domain filtering	402	Labor-related FAQ records retained.
	Leakage / near-duplicate screening	400 rows / 399 unique questions	Records prepared for leakage-safe splitting after screening and exact-question review.

Source / Stage	Processing step	Records after step	Explanation
	Indexing subset	278	FAQ records used in the searchable RAG knowledge base.
	Validation subset	41	FAQ records reserved for validation and threshold calibration.
	Held-out FAQ subset	80	Unique FAQ questions reserved outside the index for held-out evaluation.
Legal evaluation	Final legal retrieval test	24	Manually curated legal questions used as the final retrieval test.
	Supplementary robustness set	201	Auto-generated legal questions used only as a robustness check.
Generation evaluation	Final mixed-type reliability-gated set	45	Stratified mixed-type questions used for paired LLM comparison under the fixed reliability-gated pipeline.
Threshold calibration	Validation examples	56	41 validation records combined with 15 out-of-scope examples for Youden's J calibration.

Table 3.1 describes the process that leads from the original data to the knowledge base search engine and the evaluation data. Starting with the legal data source, 249 articles were gathered, out of which 211 active articles were indexed, while 38 repealed articles were archived without retrieval. In the HRSD data source, 711 FAQ rows were acquired. After processing for noise removal and domain filtering, 402 labor-related rows were kept. After leakage and de-duplication filtering, 400 rows, which correspond to 399 unique questions, were filtered into 278 indexing rows, 41 validation rows, and 81 held-out rows. These 81 rows correspond to 80 unique held-out questions after exact de-duplication. Legal data, robustness data, FAQ data, generation data, and calibration data sets were all different because of the different experimental tasks they served.

bab_label	bab_title	fasl	article_title	article_number_label	article_status	article_text
0	الباب الأول / التعريفات / الأحكام العامة	الفصل الأول	المادة الأولى:	1	active	يسمى هذا النظام نظام العمل
1	الباب الأول / التعريفات / الأحكام العامة	الفصل الأول	: المادة الثانية	2	active	يقصد بالألفاظ والعبارات الآتية - أينما وردت في هذا النظام - المعاني المبينة أمامها ما لم يقتض السياق خلاف ذلك: -الوزارة: وزارة الموارد البشرية والتنمية الاجتماعية. -الوزير: وزير
2	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	: المادة الثالثة	3	active	العمل حق للمواطن ، لا يجوز لغيره ممارسته إلا بعد توافر الشروط المنصوص عليها في هذا النظام ، والمواطنون متساوون في حق العمل دون أي تمييز على أساس الجنس أو الإعاقة أو ... السن أو أي
3	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	: المادة الرابعة	4	active	يجب على صاحب العمل والعامل عند تطبيق أحكام هذا النظام الالتزام بمقتضيات أحكام . الشريعة الإسلامية
4	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	المادة الخامسة:	5	active	تسري أحكام هذا النظام على الآتي: كل عقد عمل يلتزم بمقتضاه أي شخص بالعمل لمصلحة صاحب عمل وتحت إدارته أو إشرافه ؛ مقابل أجر. عمال الحكومة والهيئات والمؤسسات العامة ...، بمن فيهم الذ
5	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	المادة السادسة:	6	active	تسري على العامل العرضي والموسمي والمؤقت الأحكام الخاصة بالواجبات وقواعد التأديب ، والحد الأقصى لساعات العمل ، وفترات الراحة اليومية والراحة الأسبوعية ، والتشغيل الإضافي ... والعط
6	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	المادة السابعة :	7	active	يستثنى من تطبيق أحكام هذا النظام كل من: -أفراد أسرة صاحب العمل، وهم زوجه 1- وأصوله وفروعه الذين يعملون في المنشأة التي لا تضم سواهم. ب- لاعبو الأندية والاتحادات ...الرياضية وم
7	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	: المادة الثامنة	8	active	يبطل كل شرط يخالف أحكام هذا النظام ، ويبطل كل إبراء ، أو مصالحة عن الحقوق الناشئة للعامل بموجب هذا النظام ، أثناء سريان عقد العمل ، ما لم يكن أكثر فائدة للعامل
8	الباب الأول / التعريفات / الأحكام العامة	الفصل الثاني	المادة التاسعة:	9	active	اللغة العربية هي الواجبة الإستعمال في البيانات والسجلات والملفات وعقود العمل وغيرها مما هو منصوص عليه في هذا النظام ، أو في أي قرار صادر تطبيقاً لأحكامه ، وكذلك ... التعليمات التي

question	answer	beneficiaries	sector	subsite
هل تستخدم البوابة والانظمة التابعة لها eSignature؟	تقدم الوزارة خدمة التوقيع الالكتروني في اطار ادارة وتوثيق عقود العمل عبر منصة "قوى" يمكن للمنشآت انشاء وتوثيق وانهاء عقود العمل للموظفين بشكل الكتروني بالكامل.	أصحاب عمل أفراد العمالة المنزلية كبار السن	قطاع العمل	مركز الرياض للسياسات السلوكية
ماذا أفعل إذا تم إرجاع الطلب للمراجعة؟	الدخول إلى "مهامي" - اختبار مهمة طلب مراجعة - تعديل البيانات - إعادة الإرسال	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
ماذا يحدث بعد إرسال الطلب؟	يتم تحويل الطلب إلى اللجنة للمراجعة ويتم إشعار المتقدم بحالة الطلب	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
كيف أقدم طلب تعديل فئة مهنية؟	من خلال تسجيل الدخول - الخدمات - اختيار "طلب تعديل الفئة المهنية" - تعبئة النموذج - إرسال الطلب	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
ما متطلبات التقديم؟	أن يكون المتقدم حاصل على ترخيص سابق	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
من يمكنه التقديم على خدمة تعديل الفئة المهنية؟	الممارس الحاصل على ترخيص من خلال خدمة التسجيل والتصنيف المهني	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
هل يمكن للمتقدم أن يحصل على أكثر من تصنيف وبالتالي أكثر من رخصة؟	نعم، يمكن للمتقدم الحصول على أكثر من رخصة إذا كان يحمل مؤهلات أو دورات أو خبرات تؤهله لذلك	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
إذا رفض طلب التسجيل والتصنيف هل يحق لي استرداد رسوم الخدمة؟	نعم، يحق لك استرداد رسوم خدمة التسجيل والتصنيف إذا رفض الطلب	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
ماذا يحدث في حال رفض الطلب؟	يمكنك الاطلاع على سبب الرفض من خلال تفاصيل الطلب، ثم التقديم بطلب جديد بعد معالجة الأسباب	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية
هل يمكن إلغاء الفاتورة؟	نعم، يمكن إلغاء الفاتورة قبل السداد، مع ضرورة إدخال سبب الإلغاء	التخصصات الاجتماعية العمالة المنزلية كبار السن	قطاع التنمية الاجتماعية	مركز الرياض للسياسات السلوكية

Figure 3.2. Actual screenshots of the structured legal and HRSD FAQ datasets.

Figure 3.2 illustrates the two structured record types. A legal-article record includes an article identifier, its position in the legal hierarchy, its status, and the official text. An FAQ record contains the official question and answer together with beneficiary and service metadata. These fields preserve source traceability and allow every retrieved passage to be linked back to its origin.

3.3 Data Collection and Web Scraping

The raw data were extracted directly from the website of the Ministry of Human Resources and Social Development (HRSD) of the Kingdom of Saudi Arabia. The corpus included two kinds of content: articles from Saudi Labor Law and HRSD FAQ pages with questions and answers. By confining the data extraction to only these sources, we could be sure that our knowledge base would consist only of the legal and informational resources.

Data collection was done in Python by using Requests library together with BeautifulSoup. Pages with Saudi Labor Law were processed in article-wise manner while maintaining hierarchy levels of parts, chapters, and article numbers. Each HRSD FAQ question-answer pair was stored as an individual page with its metadata.

Arabic titles of the articles, such as "المادة الحادية والخمسون بعد المائة"، were turned into numeric representations (e.g., 151) with preservation of "مكرر" tag if necessary; e.g., "المادة الحادية والثلاثون بعد المائة مكرر" was converted to "131مكرر". Such numeric tags were used for citations, retrieval experiments, and gold standard alignments.

Table 3.2. Examples of Arabic article-number normalization.

Raw Article Title (as scraped)	Normalized article number label	Note
المادة الحادية عشرة	11	Two-digit ordinal (10–19 range)
المادة التاسعة والسبعون	79	Units + tens combination (below 100)
المادة الحادية والثلاثون بعد المائة مكرر	131 مكرر	Adds 100; repeated-article marker preserved
المادة التاسعة والعشرون بعد المائتين	229	Adds 200

3.4 Data Preprocessing

This approach of data cleansing ensured that the raw data scrapped was ready to be retrieved. This process involved schema normalization and normalization of Arabic text, removal of redundant articles, filtering for domain-specific FAQs, leakage-free data split, near-duplicate identification and text segmentation. All the intermediate products were stored and verified before next processing, to ensure reparability and to determine processes mistakes.

3.4.1 Schema Standardization and Data Integration

The two raw sources have been mapped to the common schema with the identifiers, sources, text blocks, and metadata columns. Since the articles of the Saudi Labor Law and HRSD FAQ entries have different original schemas, the merged table includes the `source_type` column. Figure 3.3 shows an example of the resulting knowledge base.

Schema standardization and integration: two raw sources with different structures (13 and 10 columns) → one unified 23-column schema distinguished by `source_type`

Stage	Schema / columns
BEFORE • raw legal source (scraped)	(13 columns) bab_order, bab_label, bab_title, fasl, article_title, article_number, article_number_label, article_number_int, article_text, source_url, searchable_text, text_len, article_status
BEFORE • raw HRSD FAQ source (scraped)	(10 columns) faq_id, question, answer, beneficiaries, sector, subsite, page_number, page_url, source, searchable_text
AFTER • unified knowledge base (source_type)	(23 columns) record_id, source_type, source_name, legal_category, bab_label, bab_title, fasl, article_title, article_number, article_number_label, article_number_int, article_status, rag_usage, question, answer, beneficiaries, sector, subsite, original_text, source_url, document_unit_id, text_for_indexing, normalized_text

This schema was chosen for use due to four reasons:

- Firstly, a standardized schema makes it possible to clean, filter, chunk, embed, index and perform Retrieval through a single common process.
- Secondly, the retrieval engine is capable of making a query on both formal law articles and the FAQ entries without the necessity of keeping separate query paths.
- Thirdly, the FAQ entries contain the actual phrasing that is used by the user to ask the question regarding service information, whereas the law articles provide the legal foundation for the answer.
- Lastly, common fields include record_id, source_type, source_url, original_text, normalized_text and document_unit_id.

Sample of the unified knowledge base — legal articles + FAQ merged into one schema (source_type)

record_id	source_type	article_no	article_title	question	text / answer	beneficiaries
3	legal_article	3	المادة الثالثة:	—	العمل حق للمواطن ، لا يجوز لغيره ممارسته إلا بعد توافر الشروط المنصوص عليها في هذا النظام ، والمواطنون متساوون في حق العمل دون أي تمييز على أساس الجنس أو الإعاقة أو السن أو أي شكل من أشكال التمييز الأخرى، سواء أثناء أداء العمل أو عند التوظيف أو الإعلان عنه.	—
6	legal_article	6	المادة السادسة:	—	تسري علي العامل العرضي والموسمي والمؤقت الأحكام الخاصة بالواجبات وقواعد التأديب ، والحد الأقصى لساعات العمل ، وفترات الراحة اليومية والراحة الأسبوعية ، والتشغيل الإضافي ، والعطلات الرسمية ، وقواعد السلامة ، والصحة المهنية ، وإصابات العمل والتعويض عنها ، وما يقرره الوزير .	—
8	legal_article	8	المادة الثامنة:	—	يبطل كل شرط يخالف أحكام هذا النظام ، ويبطل كل إبراء ، أو مصالحنة عن الحقوق الناشئة للعامل بموجب هذا النظام ، أثناء سريان عقد العمل ، ما لم يكن أكثر فائدة للعامل.	—
213	faq	—	—	هل يطبق قرار توظيف المهن بالتوازي مع نطاقات؟	نعم، يُطبق قرار توظيف المهن على المهن المستهدفة بالقرار داخل المنشأة، وتطبق العقوبات المنصوص عليها نطاقًا بغض النظر عن نطاق المنشأة في نطاقات، نطاق المنشأة لا يؤثر في عملية حساب نسبة التوظيف لمهنة مشترف مسكن.	العمالة المنزلية كبار السن
214	faq	—	—	هل يحق صاحب العمل في الاعتراض في عقد العمل الموثق سنديًا تنفيذًا؟	يحق لصاحب العمل الاعتراض على طلب التنفيذ عن طريق منصة (ناجز)، حيث يتم إشعار صاحب العمل بطلب التنفيذ ويتم إعطائه (5) أيام للاعتراض أو السداد.	أصحاب عمل أفراد العمالة المنزلية كبار السن
215	faq	—	—	كيف يتم التحقق من استلام العامل للأجر الذي سيتم التنفيذ عليه في عقد العمل الموثق سنديًا تنفيذًا؟	يتم التحقق إلكترونيا وفق بيانات برنامج حماية الأجور (منصة مدد) والذي يوثق دفع أجور العاملين بالمنشآت حسب تاريخ الاستحقاق ومقدار الأجر المتفق عليه.	أصحاب عمل أفراد العمالة المنزلية كبار السن

Figure 3.3. Sample of the unified knowledge base showing Saudi Labor Law articles and HRSD FAQ records integrated into a single schema and distinguished by the source_type field.

3.4.2 Normalization of Arabic Text

Normalization and cleaning of the Arabic text was done before indexing to minimize orthographic differences. All diacritics and tatweel symbols were stripped away; common Alef, Yeh, and Ta-marbuta characters were normalized; and all scrapings and extra whitespace was stripped away. The normalized version of the text was used for search purposes, whereas the actual wording was kept separately.

Having both forms is crucial in the legal sphere. On one hand, normalization improves the matching between user query and indexed document, but on the other hand, the text that is shown to the user should be an exact copy of the original text from the official source.

Arabic text before and after normalization (remove tashkeel & tatweel, unify Alef / Yeh / Ta-marbuta; the original wording is preserved separately)

#	Before — (محفوظ كما نُشر)	After — (للمهسة والبحث)
1	يجب على صاحب العمل والعامل عند تطبيق أحكام هذا النظام الالتزام بمقتضيات أحكام الشريعة الإسلامية .	يجب على صاحب العمل والعامل عند تطبيق احكام هذا النظام الالتزام بمقتضيات احكام الشريعة الاسلاميه .
2	تحتسب جميع المدد والمواعيد المنصوص عليها في هذا النظام بالتقويم الهجري ، مالم ينص في عقد العمل أو لائحة تنظيم العمل على خلاف ذلك .	تحتسب جميع المدد والمواعيد المنصوص عليها في هذا النظام بالتقويم الهجري ، مالم ينص في عقد العمل او لايحه تنظيم العمل على خلاف ذلك .
3	على الوزير — في سبيل تنفيذ أحكام هذا النظام — التنسيق مع الجهات ذات العلاقة كلما اقتضى الأمر ذلك.	على الوزير في سبيل تنفيذ احكام هذا النظام التنسيق مع الجهات ذات العلاقة كلما اقتضى الامر ذلك.

Figure 3.4. Example of Arabic legal text before and after normalization, including removal of diacritics and tatweel and unification of Alef, Yeh, and Ta-marbuta letter forms. The original wording is preserved separately for citation and traceability.

3.4.3 Archiving and Exclusion of Repealed Articles

Each legal article was assigned an article_status value indicating whether it was active, repealed, or amended. Only active articles were indexed in the searchable knowledge base. Repealed and inactive provisions were retained in an archive for auditability but excluded from retrieval because citing an obsolete article could mislead the user and undermine the reliability of the service.

عدد المواد الملغاة (مستبعدة من قاعدة البحث): 38
Repealed labor-law articles — archived for auditability and excluded from the searchable knowledge base

article_no	article_title	article_status	article_text	in knowledge base?
14	المادة الرابعة عشرة:	repealed	(ملغاة)	Excluded (archived)
149	المادة التاسعة والأربعون بعد المائة:	repealed	(ملغاة)	Excluded (archived)
150	المادة الخمسون بعد المائة:	repealed	(ملغاة)	Excluded (archived)
152	المادة الثانية والخمسون بعد المائة:	repealed	(ملغاة)	Excluded (archived)
156	المادة السادسة والخمسون بعد المائة:	repealed	(ملغاة)	Excluded (archived)
195	المادة الخامسة والتسعون بعد المائة:	repealed	(ملغاة)	Excluded (archived)

Figure 3.5. Sample of repealed (cancelled) labor-law articles that are archived for auditability and excluded from the searchable knowledge base.

3.4.4 FAQ Cleaning and Leakage-Safe Splitting

The FAQ records were cleaned to remove incomplete, duplicate, noisy, and out-of-domain entries. Only records related to labor rights, employment contracts, termination, workplace regulation, and employment services were retained. Figure 3.6 contrasts retained labor-related examples with excluded out-of-domain records.

BEFORE - classified FAQ: 703

AFTER - kept (labor-related): 402 | excluded: 301

FAQ domain cleaning — kept labor-related questions vs. excluded out-of-domain questions (703 classified → 402 kept / 301 excluded)

question	decision	reason	indexed (RAG)
هل يطبق قرار توظيف المهن بالتوازي مع نطاقات؟	KEEP (labor)	Strong labor-law or labor-service evidence found in question	Yes
هل يحق صاحب العمل في الاعتراض في عقد العمل الموثق سندياً تنفيذياً؟	KEEP (labor)	Strong labor-law or labor-service evidence found in question	Yes
كيف يتم التحقق من استلام العامل للأجر الذي سيتم التنفيذ عليه في عقد العمل الموثق سندياً تنفيذياً؟	KEEP (labor)	Strong labor-law or labor-service evidence found in question	Yes
ماذا أفعل إذا تم إرجاع الطلب للمراجعة؟	EXCLUDE (non-labor)	Social-development FAQ without strong labor-law evidence.	No
ماذا يحدث بعد إرسال الطلب؟	EXCLUDE (non-labor)	Social-development FAQ without strong labor-law evidence.	No
ما متطلبات التقديم؟	EXCLUDE (non-labor)	Social-development FAQ without strong labor-law evidence.	No

Figure 3.6. FAQ domain cleaning: examples of kept labor-related questions versus excluded out-of-domain questions (703 classified FAQ entries, of which 402 are retained and 301 excluded).

The retained FAQ records were then divided through an identifier-based 70/10/20 split. The 70% subset was indexed in the searchable RAG knowledge base, the 10% subset was reserved for validation and threshold calibration, and the remaining 20% was kept as held-out test data. Validation data were used only before final evaluation, and the held-out partition remained unseen during threshold selection. The Stage 04 auto-generated legal set was handled separately as a supplementary robustness resource and was not used for tuning.

26

Formally, leakage prevention required the three subsets to be mutually disjoint: $D_{train} \cap D_{val} = D_{train} \cap D_{test} = D_{val} \cap D_{test} = \emptyset$. Here, D_{train} denotes the 70% records indexed in the searchable RAG knowledge base, D_{val} denotes the 10% validation questions used for configuration selection, and D_{test} denotes the 20% held-out questions and their gold references used for evaluation.

The duplicate retained for evaluation was faq_id 248. The removed instance was faq_id 94, with document_unit_id b0ebcab351b4739f0566a90080daa23d875b475bac36659a0088539697b8a266. Although the two source records differed, their question text was identical. Keeping one canonical instance prevented the same user question from being counted twice in the held-out analysis.

3.4.5 Detection of Near-Duplicates

In this research, the technique involving character n-gram of the TF-IDF together with cosine similarity was used to detect possible near duplicates [28]. Character n-grams have been used since they can be useful in uncovering substantial text overlap in cases where there are no word-for-word similarities. Scores that are higher than 0.90 and between 0.85 to 0.90 were considered as review heuristics in this research and were manually reviewed. This was done to avoid having very similar documents in both

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \log \frac{N}{1 + \text{df}(t)}$$

index and evaluation sets.

$$\text{cosine_similarity}(x_i, x_j) = (x_i \cdot x_j) / (\|x_i\| \|x_j\|) \quad (3.1)$$

2

3.4.6 Text Segmentation and Source Traceability

After data cleaning, the source text was segmented into retrievable units while keeping the metadata necessary for source traceability. The segmentation procedure occurred before indexing since the retrieval system processes evidence units and not whole source pages. Two segmentation techniques were considered:

1) Structural Segmentation

Under this technique, each article of the Saudi Labor Law was considered as one legal unit, and each FAQ question-answer combination of the HRSD as one service-information unit. The advantage of this approach is that it keeps intact the structure of the original record and connects it to the article number, source type, and other metadata.

2) Fixed-Size Segmentation

Under the fixed-size baseline scenario, the text was split into 500-word pieces with 80-word overlap to prevent information loss around the boundaries of the chunks. Due to the highly inflected nature of the Arabic language, fixed size word-based segmentation may break up legal concepts, numbers, and references to articles, resulting in reduced retrieval quality.

Both approaches were tested experimentally. Structural segmentation was chosen as the basic approach since it allows to keep complete legal and FAQ records, while the fixed size segmentation provided the baseline. Every segment contained the metadata such as source type, article number, document type, section reference, parent document identifier, and segment identifier.

Preprocessing Output	Purpose
Cleaned legal articles	Provide standardized active legal records for the knowledge base.
Cleaned HRSD FAQ records	Provide service-oriented question-answer knowledge.
Held-out evaluation set	Support leakage-safe evaluation.
Structural chunks	Preserve article and FAQ boundaries.
Fixed-size chunks	Provide a baseline chunking strategy.
Data-quality and leakage audits	Document preprocessing checks and potential risks.

Table 3.3. Main preprocessing outputs.

3.5 Knowledge Base and Retrieval System

The knowledge base contains the legal articles and FAQ records supplied to the answer-generation module as evidence. Before any response is generated, dense semantic retrieval and BM25 lexical retrieval search the index, the two score streams are fused, and a cross-encoder reranks the resulting candidates.

3.5.1 Embedding and Indexing

In the dense retrieval stage, the articles and the FAQs were embedded with the help of multilingual embeddings that consist of multilingual-e5-large and bge-m3. The embeddings persist in ChromaDB, which is responsible for similarity search [25]. A BM25 lexical index has been built alongside the dense

index because not all legal terms, article number, and other domain-specific expressions may be identified based on the semantic similarity alone.

3.5.2 Retrieval Algorithms

Retrieval module includes the following algorithms and settings:

- In the dense retrieval process, the query and the indexed item are encoded separately with the multilingual embedding and ranked according to the cosine similarity metric. For the experiments, the two multilingual-e5-large and bge-m3 were tested [29], [30], relying on the general sentence embeddings framework [31]. Dense retrieval is useful when a user's wording differs from the wording of the legal source.

$$\text{sim}(q, d) = \frac{q \cdot d}{\|q\| \|d\|}$$

$$\cos(q, d) = (q \cdot d) / (\|q\| \|d\|) \quad (3.2)$$

- BM25 is a lexical ranking function that weights query-term frequency by inverse document frequency and normalizes for document length, as defined in Eq. (3.3) [32]. The experiments used $k_1 = 1.5$ and $b = 0.75$. Lexical retrieval is particularly useful for exact legal terms, article numbers, and expressions that may be weakened by semantic normalization.

$$\text{BM25}(q, d) = \sum_{t \in q} \text{idf}(t) \frac{f(t, d) (k_1 + 1)}{f(t, d) + k_1 (1 - b + b |d| / \text{avgdl})}$$

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot ((f(t, d)(k_1 + 1)) / (f(t, d) + k_1 (1 - b + b |d| / \text{avgdl}))) \quad (3.3)$$

- Hybrid retrieval combines the normalized dense and BM25 scores in one ranking. A weighted sum controlled by α assigns the relative contribution of semantic and lexical retrieval, allowing the two signals to complement one another.
- Cross-encoder reranking re-scores the top hybrid candidates using bge-reranker-v2-m3 [33], [34]. Unlike independent query and passage embeddings, a cross-encoder jointly processes each query-passage pair and can provide a more discriminative relevance score [35]. The reranker therefore changes the ordering of the candidates rather than expanding the candidate pool.

The final configuration retrieved an initial pool of 50 passages and passed the five highest-ranked evidence passages to the answer-generation module. The dense weight was evaluated over $\{0.65, 0.80, 0.90\}$, and $\alpha = 0.80$ was selected for the final configuration. The reliability gate used the separately calibrated threshold $\tau = 0.6503$. Chapter 4 reports the numerical comparison of these settings.

$$\text{HybridScore}(q, d) = \alpha \cdot \text{DenseScore_norm}(q, d) + (1 - \alpha) \cdot \text{BM25_norm}(q, d) \quad (3.4)$$

RAG Retrieval and Reliability Gate Pipeline

AI-Powered Arabic Voice Agent for Saudi Labor Law Inquiries

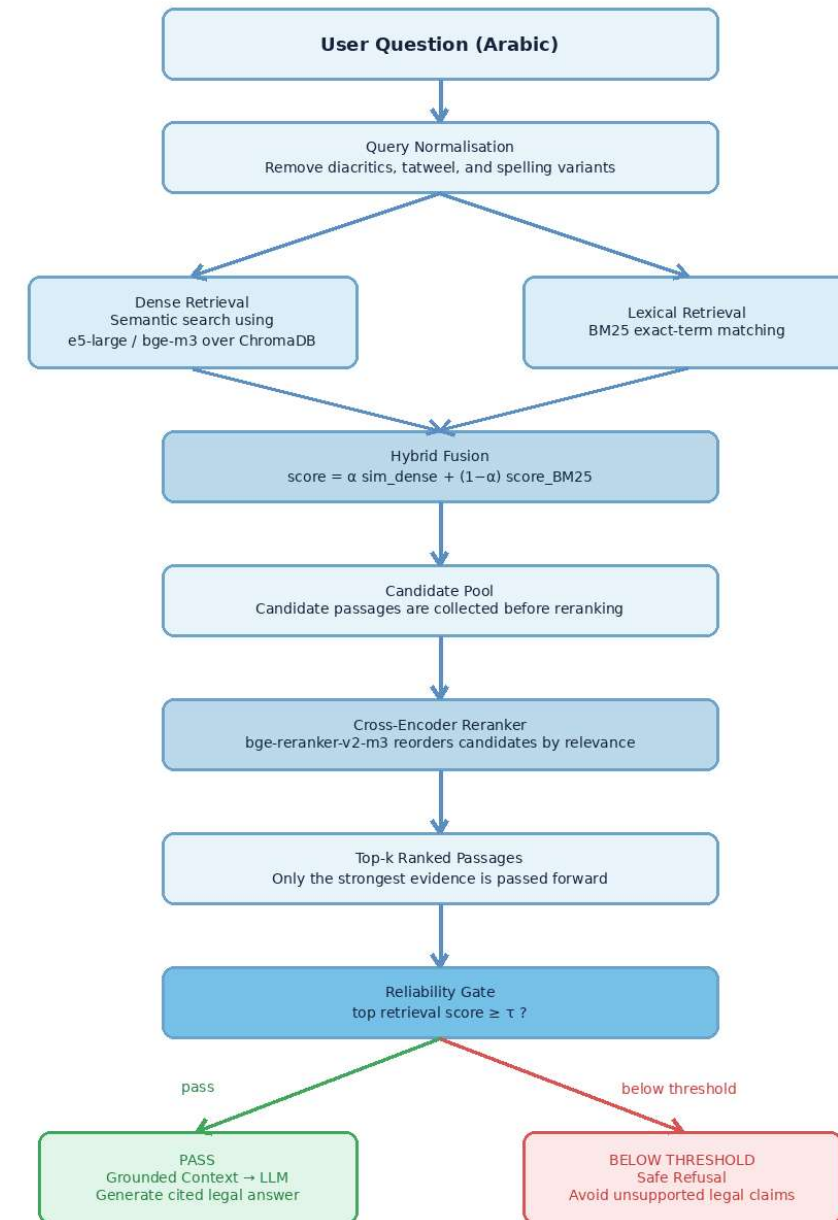


Figure note: Dense retrieval captures semantic meaning, BM25 captures exact legal terminology, hybrid fusion combines both signals, and the reranker improves the final ranking. The reliability gate prevents unsupported answer generation when evidence is weak.

Figure 3.7. Retrieval, reranking, and reliability-gate pipeline used before grounded answer generation.

Figure 3.7 shows the retrieval sequence. The Arabic query is normalized, then submitted to **dense semantic retrieval** and **BM25 lexical retrieval**. **Hybrid fusion** combines the two score streams into a single candidate ranking, and a cross-encoder reorders the candidates by joint query-passage relevance. The **top-k passages** are then selected as evidence. Before generation, the reliability gate determines whether the strongest retrieved evidence is sufficient; otherwise, the system issues a safe refusal.

Retrieval configurations were compared on a frozen 155-question validation-ablation manifest before the final legal test was reported. This manifest was used only for configuration comparison and was not reused as the 24-question final legal test, the 201-question robustness set, the FAQ held-out resource, or the 45-question generation set. The selected configuration combined structural segmentation, multilingual-e5-large embeddings, hybrid retrieval ($\alpha = 0.80$), and cross-encoder reranking, following the component-wise evaluation approach recommended for Arabic RAG pipelines [24].

3.6 Hybrid RAG, Reranking, and Reliability Gate

The last retrieval step is the use of BM25 lexical matching and dense semantic retrieval. BM25 keeps the exact terms and article references, while dense retrieval maintains the semantic similarity between various Arabic expressions. A weighted fusion method, which considers contributing from both signals, is used in this circumstance instead of relying on either method alone [24].

The hybrid pool of candidates is reordered by the cross-encoder before the top 5 passages are considered to generate. The purpose of the reliability gate is different – it checks if the top confidence in retrieval score passes a defined threshold, and if there is some evidence available that supports an answer.

66

If the retrieval-confidence is too low, the system issues a controlled default denial, instead of a "generation from weak context". By doing this the design would benefit evidence prioritisation and minimises the possibility of getting an apparently fluent legal answer from irrelevant content.

3.7 Answer Generation Module

The answer-generation module provides Arabic responses about legal information based on selected evidentiary passages. The LLM doesn't conduct direct search on the knowledge base and isn't asked to produce responses based on unsupported internal memory. It uses exclusively the evidentiary passages approved during the retrieval and reliability phases thus decreasing the probability of producing unsupported legal information.

55

The prompt describes the role of the model as the Arabic Saudi Labor Law information assistant. The model is forced to generate answers only based on the provided context of the Saudi Labor Law or HRSD FAQ, to use the corresponding article number or the source when possible and to generate safe refusal when the evidence is not sufficient enough. This design implements the evidence-grounding principle of the retrieval-augmented generation (RAG) method and is consistent with the reliability criteria described in the literature on RAG [6], [36].

80

3.7.1 Intent and Reliability Gates

The intent gate divides the query into one of several types: legal domain information request, numerical entitlement, conversational request, and out of scope. The legal information requests go through the RAG process; the numerical requests are addressed to the deterministic calculator; the conversational requests get the controlled responses; and out of scope requests get the safe refusal.

Prior to providing the final answer, the system performs routing and evidence checks. The reliability threshold regulates the legal generation on the basis of the retrieval; not the deterministic computation itself. A common post-processing layer then checks the answer format, cited evidence, Arabic readability, and the controlled refusal.

3.7.2 Language Models and Prompting Strategy

The generation layer includes two locally hosted instruction-tuned LLMs, ALLaM-7B and Qwen2.5-7B [37], [38] with identical retrieval and prompting strategies. They are run with 4-bit NF4 quantization [39]. Generation is limited to 160 tokens to control the latency and avoid long answers.

The prompting strategy requires generating Arabic answer solely based on the retrieved Saudi Labor Law or HRSD FAQ evidence. The article number should be used if there is one and unsupported speculation is prohibited. The model is supposed to generate safe refusal if the evidence is not sufficient enough.

3.7.3 Deterministic Legal Calculation Layer

Some labor-law questions require numerical computation, including annual-leave entitlement, end-of-service awards, resignation entitlements, and selected notice-compensation cases. These values are calculated deterministically rather than delegated to the LLM.

With a successful numerical-entitlement request identified by the intent router, the calculator gathers the inputs needed by the rule implemented. These could involve salary, term of employment, whether or not the contract should be terminated, who's terminating the contract, how often wages were paid, and whether notice was already given. Instead of making up a value, the system prompts a clarification question if it does not get an input. The calculation result is calculated in a code program, then it is called out and explained by referring to the related official article [26].

Computation	Article	Codified Rule
Annual leave	Article 109	Annual leave entitlement is computed from service duration and prorated when required.
End-of-service award (termination)	Article 84	The award is computed according to the termination entitlement formula stated in the law.

Computation	Article	Codified Rule
End-of-service award (resignation)	Article 85	The resignation case applies the legally defined reduction levels according to service duration.
Notice period	Articles 75-76	Notice requirements and compensation depend on contract type, terminating party, wage periodicity, and the notice actually provided under the applicable amended provisions.

Table 3.4. Legal provisions implemented in the deterministic calculation layer.

The numerical value is calculated in the code, but the result is evidence typed because it manages to return and display the appropriate legal article. Questions that still need to be ‘interpreted’ instead of computed are still processed through the standard retrieval and generation process [6]. Below, Figure 3.8 shows how the routing decisions between deterministic calculator and RAG-LLM pipeline take place.

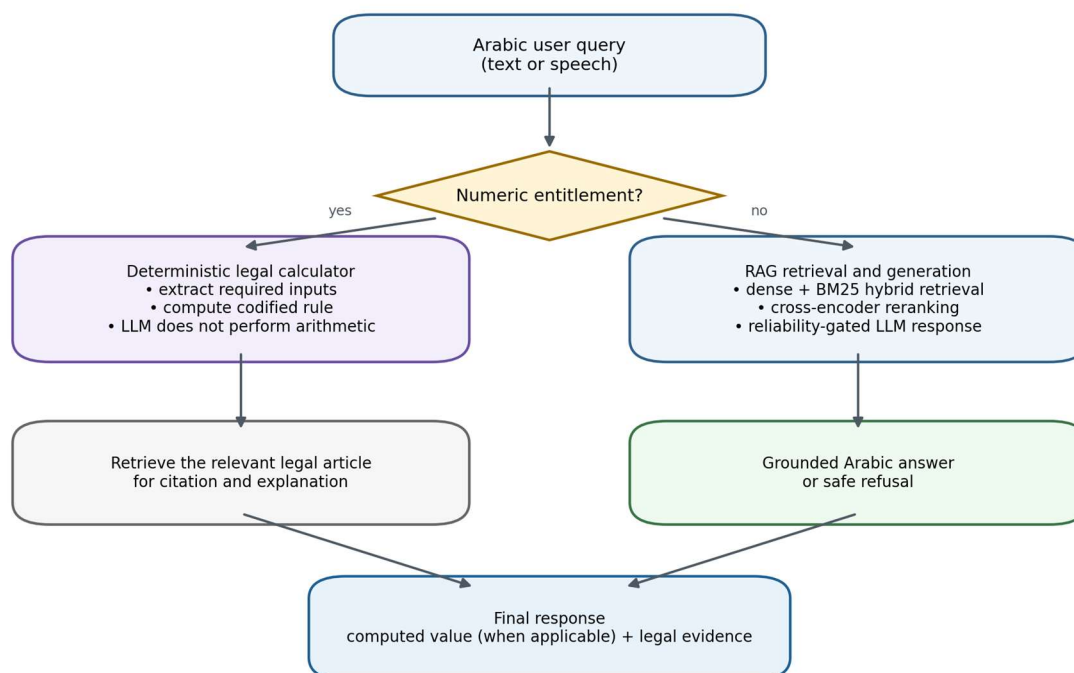


Figure 3.8. Hybrid answer generation: deterministic calculator versus RAG-LLM routing.

3.8 Conditions for triggering a deterministic calculation

The deterministic calculation layer is only turned on for the user when they are seeking a numerical answer. Including a reference to a legal right (such as an annual leave or end of service benefit) is not enough: the intent router must differentiate between a call for a calculation and a call for any general legal explanation.

A query is sent to the calculator only when three conditions are met. First, the user requests a numerical result. Second, the entitlement is covered by the implemented rule set, such as annual leave, an end-of-service award, resignation entitlement, or a supported notice-compensation case. Third, all inputs required by the rule are available. Notice-related calculations additionally require the contract type, terminating

party, wage periodicity, and the notice actually provided, because the amended Article 75 does not define one universal notice period for every case [26].

Questions that seek a general legal explanation, or that do not contain all required numerical inputs, remain on the RAG-and-LLM path. When a calculation intent is detected but information is missing, the system asks the user to provide it [26]. This prevents the LLM from filling in unspecified values and confines numerical outputs to fully specified cases.

Query type	Example	System action
General legal explanation	What is the end-of-service award?	RAG + LLM explanation
Complete numerical calculation	Calculate end-of-service for 7 years and 6,000 SAR salary.	Calculator + RAG evidence
Missing required inputs	How much is my end-of-service award?	Ask for salary, service duration, and termination type
Out-of-scope question	What is the price of a house in Riyadh?	Safe refusal
Follow-up calculation	What if it was resignation?	Use conversation memory + calculator

3.9 Calculator and RAG Evidence

The deterministic calculator and the RAG have different roles to play in the system. Numbers are generated using functions in Python, while RAG searches for the relevant articles required for justification. In particular, Article 84 is the basis for the end-of-service formula, whereas Article 85 is used whenever a resignation ratio needs to be calculated, while Articles 75 and 76 are considered only after assuming the relevant contract and notice conditions [26]. The language model outputs the correct answer with evidence in Arabic, but it does not do the calculations.

The separation of the calculation and generation process helps in avoiding arithmetic hallucinations.

Calculation type	Legal article	Implemented rule
Annual leave entitlement	Article 109	21 days annually; 30 days annually after five or more years of service.
End-of-service award	Article 84	Half-month wage for each of the first five years and one-month wage for each following year.

Calculation type	Legal article	Implemented rule
Resignation entitlement	Article 85	Applies the resignation fraction according to service duration.
Notice-related calculation	Article 75	Apply the amended notice rule only after contract type, terminating party, wage periodicity, and actual notice are established.
Selected compensation cases	Article 76	Compensation in lieu is computed only for the selected, fully specified cases implemented in code.

3.10 Reliability Threshold Calibration

The reliability threshold was calibrated exclusively on the validation split. Neither the 24-question final legal test nor the 201-question robustness set contributed to threshold selection. Once calibration was complete, the threshold was fixed before the final generation experiments to reduce test leakage.

16

Youden's J statistic, $J(\tau) = \text{sensitivity}(\tau) + \text{specificity}(\tau) - 1$, was used to select the threshold [40]. Because unsupported legal generation was considered more harmful than a conservative refusal, the selected operating point favored perfect specificity on the available validation negatives. Across 56 validation examples, $\tau = 0.6503$ produced $J = 0.7805$, sensitivity = 0.7805, and specificity = 1.0.

3.10.1 Selection Criterion

37

A confusion matrix was generated for each of the candidate thresholds, and the sensitivity and specificity values were determined by the numbers of true-positive, false-negative, true-negative, and false-positive cases. Youden's J was calculated for each operating site and an operating site was chosen at which J was maximum. There were 41 in-descriptive and 15 out-of-scope negatives in the validation set.

At $\tau = 0.6503$, TP = 32, FN = 9, FP = 0, and TN = 15, giving sensitivity = 0.7805, specificity = 1.0, and $J = 0.7805$. All of the 15 out of scope retrieval questions were refused as the selected operating point is above the highest observed out of scope retrieval score.

16

$J(\tau) = \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1$

$\tau^* = \arg \max_{\tau} J(\tau)$ τ^* was determined by calibration and kept the same for the last answer generation stage. It remained the same without making adjustments to account for the reported test results, retaining the separation between calibration and evaluation.

3.10.2 Decision process for Reliability Gate

The reliability gate is used after retrieval and reranking and before answer generation. It aims to ensure that, when the retrieved context is weak, the LLM does not generate a legal response. The retrieval-confidence score is given to the top evidence candidate: if this is $\geq \tau = 0.6503$, the query and the selected evidence are sent to the LLM; if this is $< \tau = 0.6503$, the LLM is fed with a safe refusal.

Yields to LLM if $\text{score}(q) \geq \tau$; otherwise, Safe Refusal.

Gate is based on retrieval confidence, and is not an assurance of legal accuracy. If $\tau > S$, it is refused; if $S \geq \tau$, then it can be grounded. Each mixed-type in the separate mixed-type generation set was judged a full pipeline event, whereby generated responses as well as responses of refusal were judged.

3.11 Voice Processing Layer

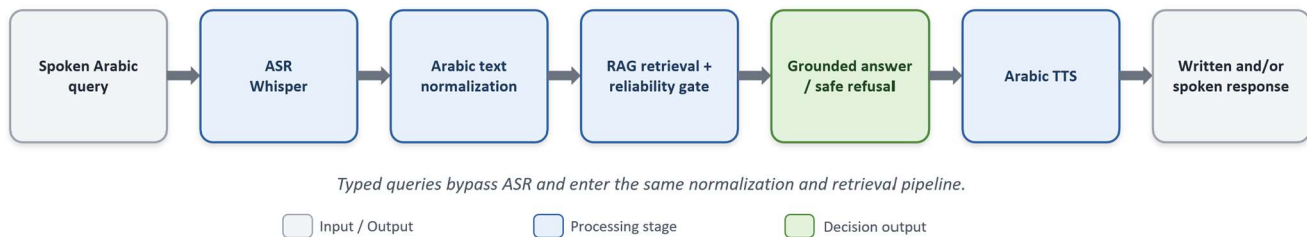
1

The voice-processing layer supports spoken interaction by converting the user's speech into text, passing the transcript through the same retrieval and generation pipeline used for typed input, and converting the validated Arabic response back into speech when requested.

Whisper-large-v3 was used as the main ASR model, with Whisper-small as a lower-latency alternative [7], [41]. MMS-TTS-Arabia, from the Massively Multilingual Speech family, provided the local synthetic-speech proxy [42], while Edge TTS [43] was measured in the runtime latency setting. These conditions are reported separately because synthesis quality, network-dependent runtime, and ASR retranscription measure different aspects of the voice layer.

End-to-End Voice-Processing Pipeline

Arabic Voice Agent for Saudi Labor Law Inquiries



14

Figure 3.9. End-to-end voice-processing pipeline.

Figure 3.9 presents the voice-processing sequence:

Step 1: The user submits a spoken Arabic query.

92

Step 2: The ASR module transcribes the speech into text.

Step 3: The transcribed text is normalized and routed according to the detected intent.

Step 4: For information-seeking queries, the RAG layer retrieves relevant legal or FAQ evidence and the reliability gate evaluates the top retrieval-confidence score.

Step 5: If the reliability threshold is met, the LLM generates a grounded Arabic answer from the retrieved evidence; otherwise, the system returns a safe refusal.

Step 6: When voice mode is enabled, the TTS engine synthesizes the validated Arabic response.

Step 7: The system returns the final response as written text and, when requested, as spoken Arabic.

3.12 Conversational Memory Module

The conversational-memory module retains short-term session context so that the agent can handle multi-turn exchanges. Its primary role is query reformulation: an ambiguous follow-up is rewritten as a self-contained legal query before retrieval.

The reformulation process can be expressed formally as follows:

$$q^*_t = \text{Reformulate}(q_t, H_{(t-1)}) \quad (3.5)$$

Here, q_t denotes the user's current query, q^*_t is the self-contained reformulated query sent to retrieval, and $H_{(t-1)}$ is the preceding conversation history.

For example, a follow-up such as "Can it be extended?" is expanded using the previous turn into a complete legal question. Reformulation reduces ambiguity while keeping the final answer anchored in official evidence.

3.12.1 Memory Design: Conversation Memory versus Retrieval Memory

There is a distinction of short-term conversation memory and retrieval memory. Conversation memory holds all the context information from the current session – so follow up questions can be rephrased, and input provided to the calculation (such as salary or length of service, or termination type) can be reused in the conversation the next time the calculation is used.

Retrieval memory is implemented through ChromaDB. It is not a user-specific profile; it is the persistent legal knowledge base containing normalized legal articles and FAQ records stored as embeddings with metadata. The system does not maintain long-term personal user profiles.

3.13 Evaluation Approach

The evaluation was carried out per-component and per-experiment role. For retrieval, the metrics that were used were Recall@k, MRR, and nDCG@5. For the comparison of the hybrid generation approach, task-specific automated correctness tests, an unsupported claim test, incorrect citations, and secondary LLM-based judgments were performed. The voice component evaluation was done using word error rate (WER) and character error rate (CER). No human or legal-expert adjudicator participated in the reported answer evaluation.

The final legal retrieval test contained 24 manually curated questions. A frozen 155-question validation-ablation manifest was used only to compare retrieval configurations and remained separate from all final-test resources. The 201 automatically generated legal questions formed a supplementary robustness set, with metrics computed over all 201 items. The paired generation comparison used a separate 45-question mixed-type set and should not be conflated with the 80 unique held-out FAQ questions.

3.13.1 Retrieval Evaluation Metrics

- **Recall@k**

Definition: Recall@k measures the proportion of evaluation questions for which the gold legal article appears within the top-k retrieved evidence passages.

Formula:

2

$$\text{Recall@k} = (1/N) \sum_{i=1..N} 1[\text{gold_i} \in \text{TopK}(q_i)]$$

Use: This metric evaluates whether the retrieval layer retrieves the correct legal article within the evidence window.

104

- **Mean Reciprocal Rank (MRR)**

Definition: MRR is the average reciprocal rank of the first correct retrieved passage across the evaluation questions.

Formula:

2

$$\text{MRR} = (1/N) \sum_{i=1..N} 1 / \text{rank_i}$$

101

Use: This metric evaluates whether the correct legal article is ranked near the top of the retrieved list.

- **nDCG@k**

Definition: nDCG@k measures normalized ranking quality by rewarding relevant evidence placed at higher ranks [44].

Formula:

44

$$\text{DCG@k} = \sum_{j=1..k} \text{rel_j} / \log_2(j+1); \quad \text{nDCG@k} = \text{DCG@k} / \text{IDCG@k}$$

85

Use: This metric evaluates the ranking quality of the retrieved evidence passages.

- **Reliability Threshold**

Definition: the reliability threshold is the lowest top retrieval-confidence score that will let the LLM generate an answer.

Formula:

Generate answer if $s_top1 \geq \tau$; otherwise return a safe refusal.

Use: This rule prevents weakly supported evidence from being passed to the LLM for legal answer generation.

Since every legal evaluation question has a single designated gold article, Recall@k and MRR are especially informative in this single-gold retrieval setting.

3.13.2 Answer Generation Metrics

- Answer Accuracy

Definition: Answer accuracy is defined as the share of all pipeline outcomes that fall into the “correct” category according to the automated rubric tailored for the task. In this regard, an appropriately grounded legal or FAQ answer, an appropriate controlled small-talk answer, and a refusal are considered correct if the latter was expected.

Formula:

Answer Accuracy = Number of correct pipeline outcomes / Number of evaluated questions

Usage: The metric allows comparing two generation models against the same set of 45 questions with identical gating decisions. It functions as an automated correctness proxy but not as a replacement of the legal expert evaluation.

- Faithfulness

Definition: Faithfulness measures to what extent the answer is grounded on the retrieved context. A judge provides a score on a 1-5 scale, which gets normalized to $[0,1]$ as $(s - 1) / 4$.

Formula:

Normalized score = $(s_{1-5} - 1) / 4$; mean faithfulness = $(1/N) \sum_{i=1..N} \text{normalized_faithfulness}_i$

Usage: The mean normalized faithfulness is used as a second grounding proxy. Since Qwen2.5-7B-Instruct has been one of the examined generators, the score could be affected by the self-preference bias.

- Answer Relevance

20

Definition: Answer relevance is a 1-5 rating given by the judge of how directly the answer addresses the user’s question. As in the previous example, $(s - 1) / 4$ transformation is applied to the ordinal score to normalize it to $[0,1]$.

Usage: The reported answer-relevance value equals the mean of normalized item-level scores and, similarly to the previous measure, could be affected by the self-preference bias.

- Automated Evaluation Protocol

The task-specific rubric varied depending on the type of the input: legal questions entailed a grounded answer and the expected number of articles; FAQ questions entailed supporting the answer with the retrieved context; out-of-scope questions entailed a safe refusal; and small-talk questions entailed providing

an appropriate controlled answer. For Qwen2.5-7B-Instruct, faithfulness and answer relevance were measured through greedy decoding (do_sample = False) with 160 new tokens and 4-bit NF4 loading. The full judge prompt is in Appendix A.

- **Hallucination Rate**

Definition: Hallucination rate is the fraction of generated answers that contain unsupported information or fabricated legal claims.

Formula:

22

Hallucination Rate = Number of answers with unsupported claims / Number of evaluated answers

Use: This metric quantifies the risk of unsupported legal generation.

- **Wrong Citation Rate**

Definition: The wrong-citation rate is the fraction of evaluated answers that cite an incorrect or unsupported legal article. Correct citations, incorrect citations, omitted citations, and refusal outcomes are retained as separate categories.

Formula:

Wrong Citation Rate = Number of answers with incorrect or unsupported citations / Number of evaluated answers

Use: This metric evaluates citation safety under grounded and baseline generation settings without treating citation omission as equivalent to an incorrect article citation.

64

3.13.3 Voice and Latency Metrics

- **Word Error Rate (WER)**

Definition: WER measures word-level transcription error in the ASR output.

Formula:

$WER = (S + D + I) / N_{\text{words}}$

62

Use: This metric evaluates the quality of Arabic ASR transcription.

- **Character Error Rate (CER)**

Definition: CER measures character-level transcription error in the ASR output.

Formula:

1

$CER = (S_c + D_c + I_c) / N_{\text{characters}}$

Use: This metric captures Arabic spelling and character-level transcription errors.

- **TTS Intelligibility Proxy**

Definition: The TTS intelligibility proxy synthesizes text into speech, retranscribes the audio, and compares the transcription with the original text.

Formula:

TTS Intelligibility Proxy = WER/CER(original text, ASR retranscription of synthesized speech)

Use: This metric estimates whether the synthesized Arabic speech remains intelligible.

- **End-to-End Latency**

Definition: End-to-end latency is the measured wall-clock time from user input to the final system response.

Formula:

End-to-End Latency = $t_{\text{final response}} - t_{\text{user input}}$

Use: This metric evaluates practical usability under real-time interaction.

- **Time-to-First-Response (TTFR)**

Definition: TTFR is the time until the first streamed response segment reaches the user.

Formula:

TTFR = $t_{\text{first streamed output}} - t_{\text{user input}}$

Use: This metric evaluates perceived latency when streaming is enabled.

3.13.4 Statistical Testing and Ablation Metrics

- **Bootstrap 95% Confidence Interval**

Definition: The reported 95% intervals were estimated with 10,000 percentile-bootstrap resamples of the item-level score vector, sampled with replacement. The lower and upper limits are the 2.5th and 97.5th percentiles of the bootstrap distribution. No BCa, bias-correction, or acceleration adjustment was applied [45].

Formula:

95% CI = [P2.5(bootstrap estimates), P97.5(bootstrap estimates)]

Use: This interval quantifies sampling uncertainty around the reported point estimate under the observed evaluation set and resampling procedure.

- **McNemar Exact Test**

Definition: McNemar's exact test compares two systems evaluated on the same questions using the discordant outcome counts b and c [46]. In this study, $b = 6$ represented ALLaM-correct/Qwen-incorrect cases and $c = 1$ represented Qwen-correct/ALLaM-incorrect cases.

Formula:

$$p = 2 \times \min\{P(X \leq \min(b,c)), P(X \geq \max(b,c))\}, \text{ where } X \sim \text{Binomial}(b+c, 0.5)$$

Use: This test examines whether the difference between two paired systems is statistically significant.

• Ablation Difference

Definition: Ablation difference measures the change in performance after removing or modifying a system component.

Formula:

$$\Delta_{\text{metric}} = \text{metric}(\text{full system}) - \text{metric}(\text{ablated system})$$

Use: This metric estimates the contribution of components such as RAG, reranking, or latency control.

Module	Metrics Defined
RAG Retrieval	Recall@k, MRR, nDCG@k, reliability decision rule
Answer Generation	Automated rule-based correctness, normalized judge faithfulness and relevance, hallucination proxy, wrong-citation rate
Voice Layer	WER, CER, TTS intelligibility proxy
Latency	End-to-end latency, time-to-first-response
Statistical and Ablation Analysis	10,000-resample percentile bootstrap, paired exact McNemar test, ablation difference

Table 3.5. Metrics defined for the final experiments. Numerical results are discussed in Chapter 4.

3.14 Workflow Implementation and Reproducibility

The workflow was developed as a multi-step pipeline in Python, in which each step saved its output, allowing further steps to be executed without running the whole pipeline again. ChromaDB [25] was used as the persisting vector database, while configuration files contained the main parameters for retrievals and generations.

All embeddings, rerankings, automatic speech recognition (ASR), and answer generation were performed using the locally available models. The used embeddings were multilingual-e5-large and bge-m3, bge-reranker-v2-m3 [33] was used as the reranking model, Whisper-large-v3 [41] was the main ASR model, and ALLaM-7B and Qwen2.5-7B were used as answer generators. Their comparative performance is analyzed in Chapter 4.

Component	Specification
Development platform	Anaconda (conda environment), Jupyter Notebook
Operating system	Windows 11 Pro
Programming language	Python 3.13.13

46

Component	Specification
Processor (CPU)	AMD Ryzen 9 9950X3D (16-core)
Memory (RAM)	128 GB
GPU	NVIDIA RTX 5090, 32 GiB VRAM (approximately 34.2 GB decimal)
CUDA toolkit	12.8
Deep-learning framework	PyTorch 2.12.0.dev20260408+cu128
Random seed	42 (fixed across all stochastic components)

Table 3.6. Hardware and software environment used in the experiments.

35

Library	Version
pandas	3.0.3
NumPy	2.4.6
scikit-learn	1.9.0
BeautifulSoup4	4.14.3
Sentence Transformers	5.5.1
Transformers	5.12.1
ChromaDB	1.5.9
PyTorch	2.12.0.dev20260408+cu128

Table 3.7. Principal Python libraries and their versions.

3.15 Summary

This chapter described the complete methodology: official-data collection, Arabic preprocessing, leakage-safe splitting, ChromaDB indexing, hybrid retrieval with reranking, reliability-controlled answer generation, deterministic legal calculation, voice processing, conversational memory, and the evaluation metrics. Experimental findings are presented and interpreted in Chapter 4.

18

Chapter 4: Results and Discussion

4.1 Introduction

This chapter presents the experimental results for the Arabic voice agent developed for Saudi Labor Law inquiries. The analysis is organized by system component: data preparation, retrieval, reliability gating, answer generation, paired statistical evaluation, voice processing, citation ablation, latency, and deterministic legal calculation.

4.2 Experimental Setup

This section provides information about the hardware used, software versions, data splits and the error matrix (RAG) used in experiments.

This is useful since the model loading speed, and retrieval, generation, voice processing and latency, are all dependent on the environment and the configuration in which they were run.

4.2.1 Technical Environment

All experiments were conducted in a local Anaconda environment using a workstation running Windows 11 Pro version and Python 3.13.13, along with an NVIDIA RTX 5090 GPU. The stochastic components have been seeded with 42, to promote reproducibility.

Table 4.1. Technical environment.

Component	Specification
Development platform	Anaconda (conda environment)
Programming language	Python 3.13.13
Operating system	Windows 11 Pro
GPU	NVIDIA RTX 5090, 32 GiB VRAM (CUDA 12.8)
Random seed	42

Technical environment and GPU verification

```
In [1]: import sys, torch, platform
print("Python :", sys.version.split()[0])
print("PyTorch:", torch.__version__)
print("CUDA available:", torch.cuda.is_available())
print("CUDA version :", torch.version.cuda)
print("GPU :", torch.cuda.get_device_name(0))
p = torch.cuda.get_device_properties(0)
print("VRAM:", round(p.total_memory/1e9,1), "GB")
```

```
Out: Python : 3.13.13
PyTorch: 2.12.0.dev20260408+cu128
CUDA available: True
CUDA version : 12.8
GPU : NVIDIA GeForce RTX 5090
VRAM: 34.2 GB
```

Figure 4.1. Code notebook screenshot: technical environment and GPU verification (Stage 03).

Figure 4.1 shows the notebook output used to verify the execution environment and GPU availability. The recorded output confirms that the retrieval and generation experiments were run under the hardware setting reported in Table 4.1.

53

4.2.2 Software Libraries and Versions

Table 4.2 summarizes the principal software libraries used in the experiments. PyTorch, Transformers, Sentence-Transformers, and ChromaDB [25] supported the retrieval and generation components. The bitsandbytes library enabled 4-bit model loading, while jiwer was used to compute WER and CER.

Table 4.2. Core libraries and their versions.

Library	Version	Role
PyTorch	2.12.0.dev20260408 + cu128	Deep-learning backend
Transformers	5.12.1	LLM and ASR loading
Sentence-Transformers	5.5.1	Embeddings and reranker
ChromaDB [25]	1.5.9	Persistent vector database
rank-bm25	0.2.2	Lexical (BM25) retrieval
bitsandbytes	0.49.2	4-bit model quantization
accelerate	1.14.0	Model loading and placement
Gradio	6.19.0	Web interface
jiwer	4.0.0	WER / CER computation
scikit-learn	1.9.0	TF-IDF and evaluation utilities

Library	Version	Role
pandas / numpy	3.0.3 / 2.4.6	Data processing
edge-tts [43]	7.2.8	Arabic speech synthesis

```

import torch, transformers, sentence_transformers
import chromadb, rank_bm25, bitsandbytes, accelerate
import gradio, jiwer, sklearn, pandas, numpy, edge_tts

libs = {
    'torch': torch.__version__,
    'transformers': transformers.__version__,
    'sentence-transformers': sentence_transformers.__version__,
    'chromadb': chromadb.__version__,
    'rank-bm25': rank_bm25.__version__,
    'bitsandbytes': bitsandbytes.__version__,
    'accelerate': accelerate.__version__,
    'gradio': gradio.__version__,
    'jiwer': jiwer.__version__,
    'scikit-learn': sklearn.__version__,
    'pandas': pandas.__version__,
    'numpy': numpy.__version__,
    'edge-tts': edge_tts.__version__,
}

for k, v in libs.items(): print(f'{k:<25} {v}')

```

```

torch                2.12.0.dev20260408+cu128
transformers          5.12.1
sentence-transformers 5.5.1
chromadb              1.5.9
rank-bm25             0.2.2
bitsandbytes          0.49.2
accelerate            1.14.0
gradio               6.19.0
jiwer                 4.0.0
scikit-learn          1.9.0
pandas                3.0.3
numpy                 2.4.6
edge-tts              7.2.8

```

Figure 4.2. Code notebook screenshot: library version check output (Stage 03).

4.2.3 Data Split and RAG Parameters

The FAQ data were split as follows: there are 278 indexing rows, 41 validation rows and 81 raw held-out rows. There was one question that came up twice in two held out source records; Exact deduplication kept one instance and created 80 unique held out question texts. Figure 4.3 defects the data as a split and thus represents number of rows and not the number of unique questions.

Table 4.3. Data split and RAG configuration parameters.

Parameter	Value
Data split (indexing/validation/held-out)	70% / 10% / 20% (leakage-safe, ID-based)
Chunk size (fixed-size baseline)	500 words
Chunk overlap (fixed-size baseline)	80 words
Chunking (best configuration)	Structural (article / FAQ boundary)

Number of retrieved documents (top_k)	5
Candidate pool (candidate_k)	50
Hybrid weight (α)	0.80
Embedding models	multilingual-e5-large [29], bge-m3 [30]
Reranker	bge-reranker-v2-m3 [33]
Language models	ALLaM-7B-Instruct [37], Qwen2.5-7B-Instruct [38], loaded using 4-bit NF4 quantization [39]
ASR model	Whisper-large-v3 [41] (main); Whisper-small (latency-oriented alternative)
Reliability threshold	0.6503 (validation-only Youden's J)
TTS engines	MMS-TTS-Arabia [42] (local synthetic-speech proxy); Edge-TTS [43] (runtime latency)

Figure 4.3 shows the raw leakage-safe split from Stage 02: 278 indexing rows, 41 validation rows, and 81 held-out rows. Exact question-text deduplication then reduced the held-out FAQ resource to 80 unique questions. The figure should therefore not be interpreted as an 81-question unique evaluation set.

config.py - Best RAG Configuration Parameters

```

In [1]: # Central configuration - single source of truth
BEST_CONFIG = {
    'chunking_strategy' : 'structural',
    'embedding_model'   : 'e5_large',
    'retriever'         : 'hybrid_reranker',
    'alpha'             : 0.80,
    'top_k'             : 5,   'candidate_k' : 50,
}
CHUNK_SIZE = 500;  CHUNK_OVERLAP = 80
RELIABILITY_THRESHOLD = 0.6503  # validation-calibrated (Youden J)
RANDOM_SEED = 42

```

```

Out[1]: Configuration loaded.
Reliability threshold = 0.6503 (validation-calibrated, Youden's J).
Threshold selected on the validation split only;
final test and robustness sets were not used.

```

Figure 4.4. Code notebook screenshot: RAG configuration parameters from config.py.

Figure 4.4 lists the parameters used by the final RAG pipeline, including the chunking strategy, retrieval depth, candidate-pool size, hybrid weight, embedding models, reranker, language models, ASR model, and reliability threshold.

The final configuration was chosen to balance evidence coverage, preservation of legal context, computational cost, and latency. Structural segmentation keeps complete legal articles and FAQ pairs together. An initial candidate pool of 50 provides broad coverage before the five strongest passages are passed to generation. The hybrid weight combines semantic and lexical evidence, which is useful for Arabic legal text because both paraphrased meaning and exact terminology or article numbers can determine relevance. Cross-encoder reranking then places the most relevant candidates near the top of that fixed pool.

4.3 Data Preparation Results

Data preparation yielded an official HRSD labour law and FAQ data source and a clean knowledge base, implementing leakage awareness. Legal articles were prepared for inclusion, articles marked as repealed were moved out of retrievability, FAQ articles were filtered by domain, and articles that were retained were chunked into retrievable units. The resulting composition is summarised in Table 4.4.

Table 4.4. Knowledge base composition after preparation. Table 4.4. Knowledge base composition after preparation.

Component	Count	Role
Active legal articles	211	Indexed official legal evidence
Repealed articles (archived)	38	Excluded from active retrieval
Indexed FAQ records	278 rows	Service-oriented evidence in the searchable knowledge base
Raw FAQ held-out partition	81 rows	Raw 20% split output
Unique FAQ held-out resource	80 questions	One exact duplicate question removed before scoring
Final legal retrieval test	24 questions	Manually curated final retrieval benchmark
Supplementary robustness set	201 questions	Supplementary automatically generated legal retrieval resource
Mixed-type generation comparison	45 questions	Independent mixed-type set for paired LLM comparison

The legal searchable knowledge base had 211 live legal articles and 278 indexed FAQ records. During the deduction of the held-out partition, the raw dataset contained 81 rows of questions in the FAQ, which was reduced to 80 unique questions with exact question-text deduplication. A distinct mixed type 45 question set of legal, FAQ, out of scope and small talk inputs was used for the paired generation comparison.

The evaluation sources were separated during the experiments. The primary retrieval test was the 24 question manually curated legal test, with the 201 question automatically generated legal test being a robustness check. A total of 81 raw FAQ rows of unique questions were not added to the searchable FAQ index. The 45 question paired LLM comparison set was independent mixed-type vs. not a sample from the FAQ holdout.

Retrieval evaluation, robustness testing, FAQ holdout, threshold calibration, and generation comparison were separated to ensure that results were for specific component. This design diminishes data spillage, and increases the transparency of the strengths and limitations of each stage.

4.4 Retrieval Evaluation Results

The retrieval evaluation focused on the capability of the RAG pipeline to retrieve relevant legal evidence from the Saudi Labor Law and HRSD knowledge base. Grounded generation cannot be separated from the quality of the retrieved context, hence, retrieval was assessed without considering the behavior of the language model.

4.4.1 Evaluation Sets and Sample Sizes

Various data sets were allotted different tasks in the experiments. The manually curated set of 24 questions was the last legal retrieval test. Validation was done on the same 155 questions on the frozen manifest and the full manifest consisted of 201 questions and was automatically generated and was saved as a supplementary robustness check. There were 81 raw FAQ holdout rows with 80 distinct question texts, and a mixed-type set of 45 questions was employed for the paired LLM comparison.

Table 4.5. Evaluation sets and their roles in the experimental analysis.

Evaluation set	Size	Role
Final legal retrieval test	24 questions	Manually curated held-out legal set; principal final retrieval result
Retrieval validation ablation subset	155 questions	Frozen configuration-comparison manifest; separate from the final legal test and other evaluation resources
Supplementary robustness set	201 questions	Automatically generated stress test; all 201 included, with 194 top-five successes
FAQ held-out resource	81 raw rows / 80 unique questions	Held outside the FAQ index; one exact duplicate question removed before scoring
Mixed-type paired generation set	45 questions	14 FAQ, 17 legal, 12 out-of-scope, 2 small-talk; 29 generated answers and 16 safe refusals

The robustness metrics in Table 4.8 were calculated over all 201 automatically generated legal questions. The gold article appeared within the top five for 194 questions, producing $\text{Recall}@5 = 194/201 = 0.9652$. A combined score based on the 24 manual questions and only the 194 successful robustness cases would omit seven failures and would not constitute an independent benchmark; no such score is therefore reported.

4.4.2 Comparison of Retrieval Methods

Several retrieval configurations were evaluated before the final setting was fixed. Table 4.6 compares hybrid retrieval with the same candidate pool followed by cross-encoder reranking on the frozen 155-question validation-ablation manifest. This experiment supported configuration selection and remained separate from the final legal test.

Table 4.6. Retrieval-method validation ablation: hybrid vs hybrid + reranker (n = 155).

Method	Recall@1	Recall@3	Recall@5	MRR
Hybrid	0.8645	0.9097	0.9226	0.8846
Hybrid + Reranker	0.9161	0.9871	0.9935	0.9508

Reranking improved every reported retrieval metric in Table 4.6. Recall@1 increased from 0.8645 to 0.9161, and MRR increased from 0.8846 to 0.9508. Because the reranker worked on the same candidate pool, it did not introduce new passages; it improved the ordering of candidates already retrieved. The largest practical effect was therefore at the earliest ranks, particularly Recall@1 and MRR.

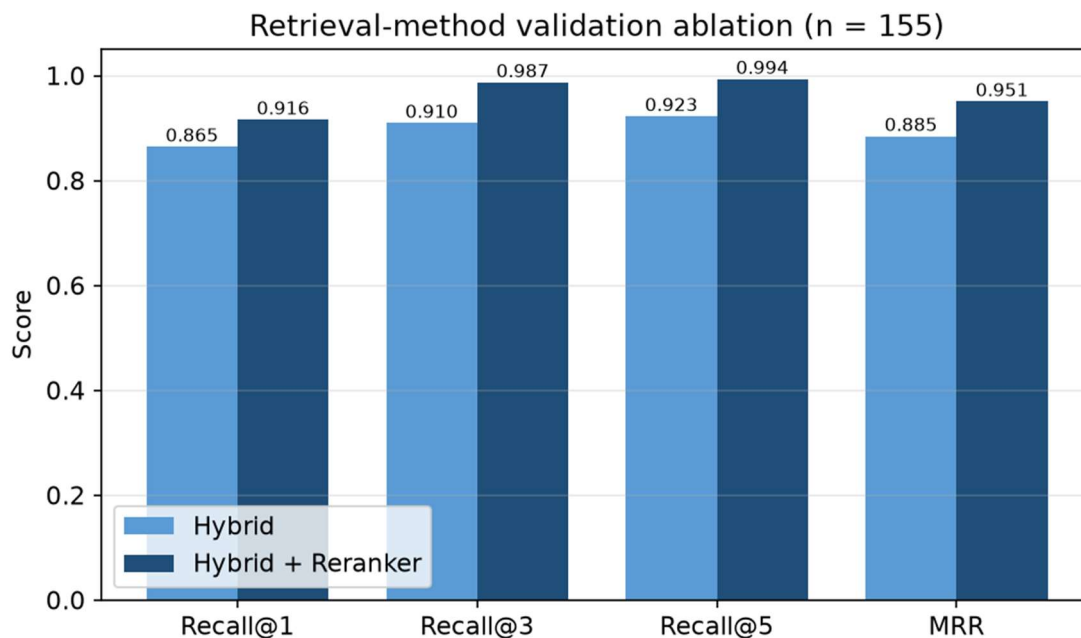


Figure 4.5. Retrieval method comparison on the legal-article task (higher is better).

Figure 4.5 visualizes the same pattern: adding the cross-encoder reranker improved early-rank performance, with the clearest gains in Recall@1 and MRR.

4.4.3 Final Legal Retrieval Test

After selecting the final retrieval configuration, the system was evaluated on the manually curated, held-out legal test set. The final configuration used structural segmentation, multilingual-e5-large embeddings, hybrid retrieval, and cross-encoder reranking. Table 4.7 presents the final legal retrieval result.

Table 4.7. Final legal retrieval test on the manually curated held-out legal set ($n = 24$).

Metric	Score
Recall@1	0.8750
Recall@3	0.9583
Recall@5	0.9583
MRR	0.9097
nDCG@5	0.9221
Sample size (n)	24 (manually curated legal test)

According to Table 4.7, the system retrieved the correct legal article at the first rank for 87.50% of the final legal questions, and the correct article appeared within the top five results for 95.83% of the questions. The MRR value of 0.9097 further demonstrates that the correct article was consistently ranked near the top. These results indicate that the selected retrieval configuration provides strong evidence coverage for the generator. Since the final legal test comprises only 24 manually curated questions, this result is reported cautiously and is supported by the supplementary robustness check.

4.4.4 Supplementary Robustness Check

The chosen configuration was tested on the 201 question auto-generated legal set to check for stability on a larger sample. The questions were formulated based on articles of law and were awaiting expert examination.

Therefore, this experiment is not used as a substitute to the manually curated final test, but instead, as a supplementary robustness check.

Table 4.8. Supplementary robustness check on the auto-generated legal set ($n = 201$, pending expert review).

Metric	Result
Recall@1	0.8358
Recall@3	0.9403
Recall@5	0.9652

Metric	Result
MRR	0.8902
nDCG@5	0.9092

The system obtained high recall with $\text{Recall}@1 = 0.8358$ and $\text{Recall}@5 = 0.9652$ for all 201 robustness questions. The top five retrievals are 194 and the failures are 7. The questions and their gold assignments have not been developed and adjudicated independently by legal experts, and it is therefore to be understood as a consistency test for retrieval behavior, but not as legal validation.

Prior to generation comparison, the reliability threshold was calibrated with only the validation split. This step set the threshold for a retrieval-confidence score to advance to LLM to a legal answer.

Youden's J was used for 56 validation examples to perform the calibration. The selected threshold, $\tau = 0.6503$, produced $J = 0.7805$, sensitivity = 0.7805, and specificity = 1.0. The final legal test (24 questions) and robustness set (201 questions) and the generation set (45 questions) were not utilized during calibration.

This split minimizes threshold-selection leakage, but is somewhat uncertain due to the small size of the validation sample

4.5 Answer Generation Results

Both ALLaM-7B and Qwen2.5-7B were compared using the same loading/retrieval configuration (4-bit NF4), prompt, and reliability gate. There were a total of 45 questions in this stratified evaluation set, 14 of which were FAQ, 17 legal, 12 out of scope, and 2 small-talk. The gate was able to generate 29 questions and returned safe refusals for 16 questions.

The accuracy was based on the total of 45 pipeline outcomes, a refusal was correct if refusal was expected.

Table 4.9. Generation accuracy of the two language models.

Model	Correct outcomes	Accuracy	95% percentile-bootstrap CI	Unsupported-claim proxy
ALLaM-7B	37 / 45	0.8222	[0.7111, 0.9333]	0.0667 (3/45)
Qwen2.5-7B	32 / 45	0.7111	[0.5778, 0.8444]	0.0667 (3/45)

ALLaM-7B produced 37 correct pipeline outcomes out of 45 (0.8222), while Qwen2.5-7B produced 32 out of 45 (0.7111). The corresponding 95% percentile-bootstrap intervals were [0.7111, 0.9333] and [0.5778, 0.8444]. These totals include correct grounded answers, appropriate controlled responses, and correct safe refusals under the automated rubric.

Since the two models were tested using the same 45 questions, this is a paired comparison. Thirty-one were correct for both models, six were correct just for ALLaM-7B, one was correct just for Qwen2.5-7B, and seven were incorrect for both models. Nevertheless, the observed accuracy was higher for ALLaM-7B, and the exact McNemar test did not show a statistically significant difference.

The difference between the seven discordant pairs was not significant at the 0.05 level, and may be due to paired sampling variation. Therefore, automated correctness should not be considered alone when selecting the model, and also the adherence of evidence, relevance of answer, latency, resource requirements, quality of Arabic answer, and limitations of the evaluation.

In this setting, ALLaM-7B was a viable generator, although the experiment does not prove its betterment in general. The comparison includes one prompt, one retrieval setting, one quantization and 45 mixed-type questions.

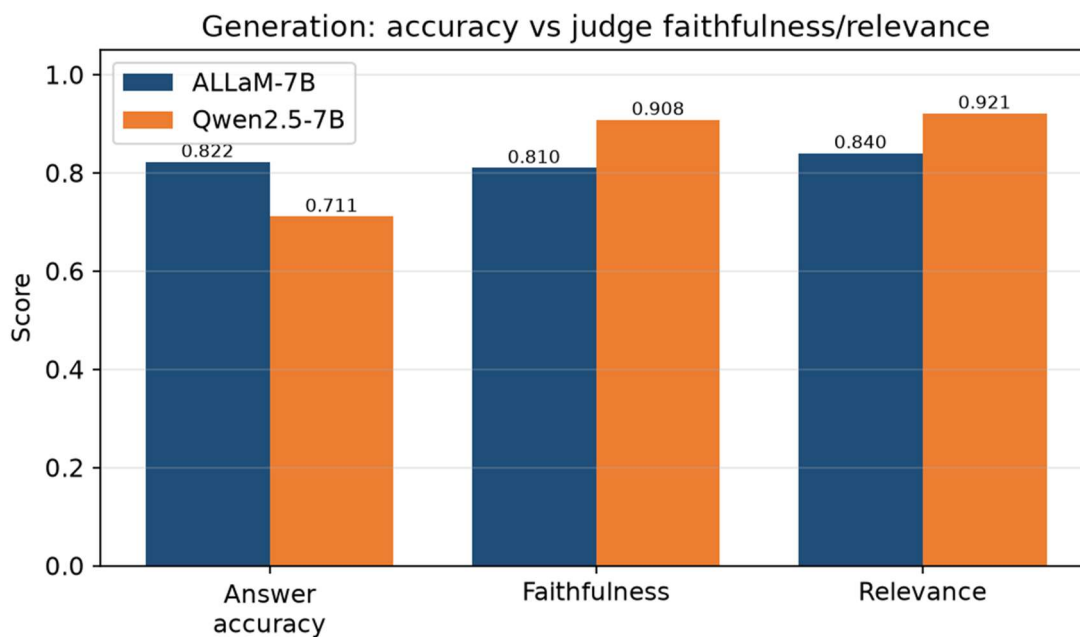


Figure 4.6. Answer accuracy and faithfulness of ALLaM-7B and Qwen2.5-7B.

Figure 4.6 compares automated answer accuracy with the judge-based faithfulness and relevance proxies. ALLaM-7B achieved the higher observed accuracy, whereas Qwen2.5-7B received higher proxy scores from the Qwen-based judge.

4.6 Statistical Evaluation

A paired exact McNemar test was appropriate because ALLaM-7B and Qwen2.5-7B were evaluated on the same 45 questions. The discordant counts were $b = 6$ and $c = 1$, giving seven discordant pairs. Accuracy uncertainty was estimated with 10,000 percentile-bootstrap resamples. Qwen2.5-7B-Instruct also served as the automated judge for faithfulness and answer relevance, so those measures are treated as secondary proxies rather than independent evaluation.

Table 4.10. Statistical comparison and faithfulness results.

Measure	ALLaM-7B	Qwen2.5-7B
Answer accuracy	0.8222 (37/45)	0.7111 (32/45)
Faithfulness proxy	0.810	0.908
Answer relevance proxy	0.840	0.921
Both correct / both incorrect	31 / 7	31 / 7
Discordant outcomes	b = 6 (ALLaM only correct)	c = 1 (Qwen only correct)
Exact two-sided McNemar p-value	0.1250 (not significant)	0.1250 (not significant)
95% percentile-bootstrap CI	[0.7111, 0.9333]	[0.5778, 0.8444]

The exact two-sided McNemar test yielded $p = 0.1250$. ALLaM-7B had the higher observed accuracy, but the experiment did not establish a statistically significant difference between the two models. The paired counts and p-value are reported explicitly because overlapping confidence intervals are not, by themselves, a test of paired significance.

The judge-based measures showed the opposite ordering. Qwen2.5-7B obtained faithfulness = 0.908 and answer relevance = 0.921, compared with 0.810 and 0.840 for ALLaM-7B. The judge assigned ordinal scores from 1 to 5, which were normalized with $(s - 1) / 4$ and averaged. Because Qwen2.5-7B-Instruct was both the judge and one of the evaluated generators, the scores may contain self-preference bias. No human or legal-expert adjudication was performed, so these measures cannot be interpreted as independent legal validation.

4.7 Voice Layer Results

The voice layer was evaluated with Word Error Rate (WER) and Character Error Rate (CER). Stage 07 computed both measures for the controlled Arabic synthetic-speech proxy, and Table 4.11 summarizes the results.

Table 4.11. Voice-layer error rates (Stage 07).

Metric	Result	Interpretation
ASR WER	0.209; 95% CI [0.162, 0.264]	Approximately one in five reference words was affected; controlled synthetic-speech proxy.
ASR CER	0.0484; 95% CI [0.0357, 0.0631]	Character-level error remained below 5% in the controlled proxy.

The WER and CER for the synthetic-speech proxy were equal to 0.209 and 0.0484 respectively. The error rate at the lower character level indicates that at least some of the word-level differences were not due to completely different contents, but rather to short orthographic and/or segmentation changes. The results of this test were limited to non-natural Saudi-dialect speakers with no microphone variation or background noise.

4.8 No-RAG versus RAG Ablation

The effect of retrieval grounding was isolated using a No-RAG-versus-RAG ablation. Language model answered from its parametric knowledge, retrieval component not included in the baseline condition. The RAG condition provided evidence from Saudi Labor Law and HRSD knowledge base prior to generation. The model, prompt family, evaluation questions and scoring criteria were the same for both conditions, and the manipulated factor was availability of retrieved evidence.

RAG can be used to anchor knowledge-intensive generation with external evidence [6] and parts of surveys identify unsupported claims and source attribution as significant evaluation topics [36]. In this experiment, comparisons allow an estimate of the contribution of retrieval under one fixed configuration, but not to prove causal relationships outside the tested sample and controls.

The main indicator of citation-safety was the percentage of answers that were assessed and contained an illegal article citation or a citation with no support. The results were counted and differentiated between correct citations, incorrect citations, omitted citations, and refusals. The latter distinction is important as it may still lead the user astray if it quotes the wrong legal passage in the answer.

The results showed that in the sample of 24 questions, the wrong-citation rate was reduced from 0.250 to 0.042 with the use of retrieval grounding. But neither the rate of correct citation increased nor was the paired correctness comparison statistically significant. Thus, under the tested conditions (as opposed to claim of "improved answer completeness" or "improved legal correctness") the ablation provides a limited citation-safety benefit.

Table 4.12. Effect of retrieval on wrong legal citation rate.

Metric (n = 24)	Rate
No-RAG correct-citation rate	0.292
RAG correct-citation rate	0.208
No-RAG wrong-citation rate	0.250
RAG wrong-citation rate	0.042

The rate of wrong citation dropped from 0.250 without RAG to 0.042 with RAG (6 out of 24 answers from without RAG and 1 out of 24 with RAG). The rate of correct citation for the No-RAG condition was 0.292 and the rate for the RAG condition was 0.208. The result therefore shows a lower number of harmful citation

errors not necessarily that there is a higher frequency of correct citations, and the paired correctness test was not significant ($p = 0.7266$).

The ablation is consequently interpreted as an observed citation-safety improvement within the tested sample. Retrieval grounding reduced incorrect or unsupported article references, but it did not produce a statistically significant increase in overall correctness. A system can reduce wrong citations by refusing or omitting a citation, so the complete distribution of response outcomes should remain part of the experiment archive.

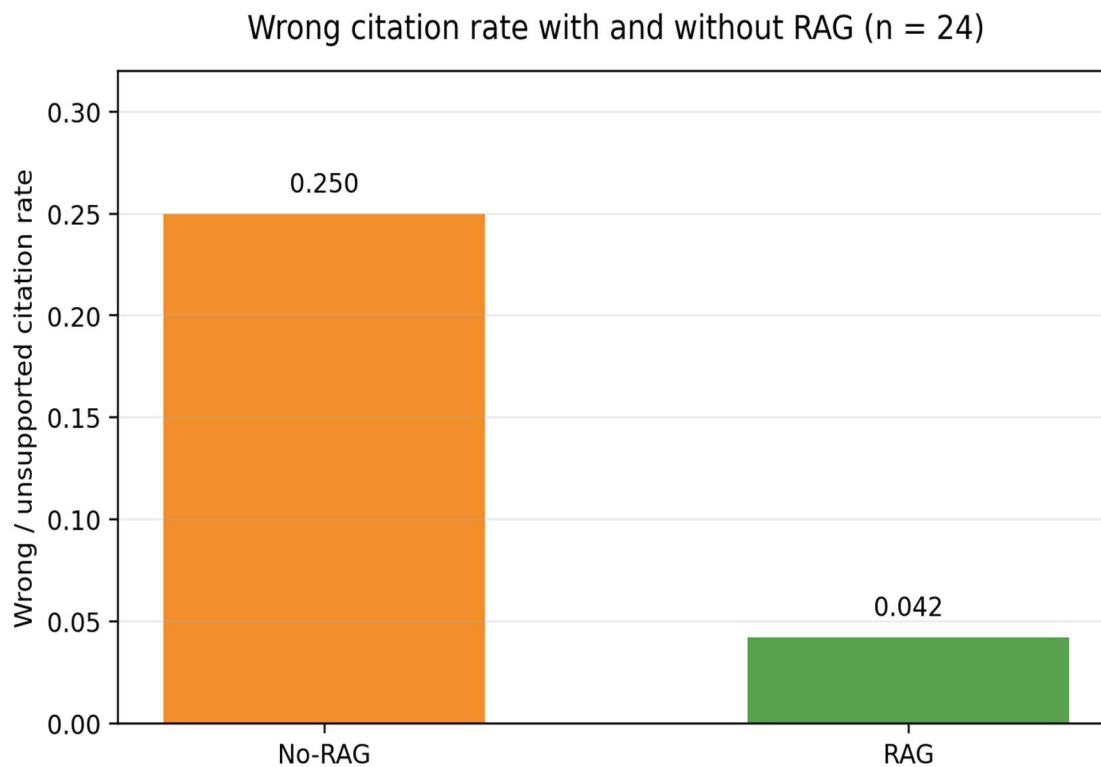


Figure 4.7. Effect of retrieval grounding on the wrong legal-citation rate.

Figure 4.7 shows the decline in the wrong legal-citation rate after retrieval grounding was introduced, from 0.250 to 0.042 on the 24-question ablation sample.

4.9 Latency Results

Latency was measured to determine whether the full pipeline could support interactive voice use. Stage 09 recorded ASR, retrieval, LLM generation, TTS, complete end-to-end latency, optimized latency, and streamed time-to-first-response. Table 4.13 reports the breakdown.

Table 4.13. Component latency, complete-response latency, and streamed time-to-first-response.

Component/setting	Time
ASR (Whisper large-v3)	0.68 s
ASR (Whisper small)	0.31 s

Component/setting	Time
Retrieval	0.25 s
LLM generation (full)	1.88 s
LLM generation (capped)	1.23 s
TTS (edge-tts)	1.18 s
Full pipeline (large ASR, full answer)	3.99 s
Optimized (small ASR, capped tokens)	2.96 s
Perceived (streaming, time-to-first-response)	0.82 s

The full pipeline returned a complete response in 3.99 seconds. Using Whisper-small and capped generation reduced complete-response latency to 2.96 seconds, while streaming reduced the time-to-first-response to 0.82 seconds. Retrieval required 0.25 seconds and was not the dominant delay in the tested configuration. These rounded component values should not be interpreted as an exact arithmetic decomposition of the reported end-to-end measurements.

The reported values are wall-clock measurements from the most recent verified execution in the stated environment. They should therefore be interpreted as environment-specific measurements rather than population-level latency estimates. GPU load, device temperature, and network- and service-dependent Edge-TTS behavior may introduce variation across runs.

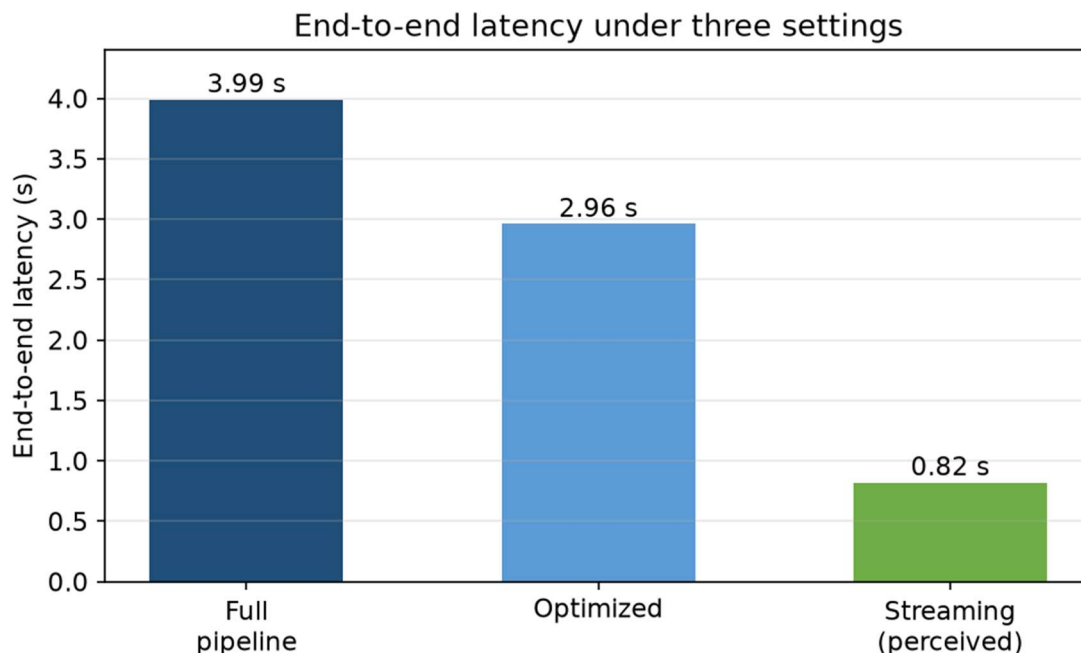


Figure 4.8. End-to-end latency under the three settings.

Figure 4.8 compares the full pipeline, the optimized configuration, and streamed time-to-first-response. Streaming improves perceived responsiveness even though the complete answer continues to be generated in the background.

4.10 Deterministic Calculation Layer Implementation Checks

The representative numerical test cases were presented to the deterministic calculation layer. These implementation checks ensured that the Python functions gave the values prescribed by the codified rules, and that the corresponding legal parts could be retrieved for citing. But neither were they independent, competent legal validation. The cases and the assumptions underlying each case are presented in Table 4.14.

Table 4.14. Representative implementation test cases for the deterministic calculation layer.

Query / assumptions	Article(s)	Computed result
Annual leave for 1.5 years of service assuming no leave was taken	109	31.5 days
End-of-service award: 7 years, employer termination, 6,000 SAR monthly wage	84	27,000 SAR
End-of-service award: 7 years, resignation, 6,000 SAR monthly wage	84, 85	18,000 SAR (two-thirds of the Article 84 award)
Employer terminates an indefinite-term, monthly paid contract after 7 years without the required 60-day notice; wage = 6,000 SAR	84, 75, 76	39,000 SAR (27,000 award + 12,000 notice compensation; unused leave excluded)

The implemented functions returned the expected values for the representative cases in Table 4.14. The 39,000 SAR result applies only to the fully specified scenario in which an employer terminates an indefinite-term, monthly paid contract after seven years without providing the required 60-day notice: 27,000 SAR under Article 84 plus 12,000 SAR in notice compensation under Articles 75 and 76. It is not a general dismissal value because the applicable rule depends on contract type, terminating party, wage periodicity, and the notice actually provided [26].

4.11 Overall Results Summary

Table 4.15 consolidates the principal system results. It provides an overview and should be read together with the detailed analyses in the preceding sections.

Table 4.15. Overall system performance summary.

Component	Main result	Interpretation
Final legal retrieval	Recall@1 = 0.8750; Recall@5 = 0.9583 (n = 24)	Strong observed top-five coverage; manual test remains small

Component	Main result	Interpretation
Robustness retrieval	Recall@5 = 194/201 = 0.9652	All 201 automatically generated questions included
Generation correctness	ALLaM = 0.8222; Qwen = 0.7111 (n = 45)	Higher observed ALLaM score; no significant paired difference
Reliability gate outcomes	29 generated answers; 16 safe refusals	Outcomes reported over the independent 45-question mixed-type set
Paired statistical test	b = 6, c = 1; exact p = 0.1250	No statistically significant model superiority established
Judge-based proxies	Faithfulness 0.810 / 0.908; relevance 0.840 / 0.921	Qwen judge favored Qwen; self-preference limitation
Voice proxy	WER = 0.209; CER = 0.0484	Synthetic-speech evaluation; real-speaker testing required
RAG ablation	Wrong citations 0.250 → 0.042 (n = 24)	Observed citation-safety improvement; no significant correctness gain
Latency	2.96 s complete; 0.82 s streamed TTFR	Environment-specific local-workstation measurement
Calculation layer	Deterministic for selected fully specified rules	Reduces arithmetic hallucination but is not complete legal coverage

The final legal retrieval test produced Recall@1 = 0.8750 and Recall@5 = 0.9583. This coverage is important because the answer-generation module relies on the retrieved article or FAQ record as its source context.

Both ALLaM-7B and Qwen2.5-7B produced viable outputs under the fixed RAG protocol. According to the results of the automated-correctness score from the automated judge, ALLaM-7B outperforms Qwen2.5-7B, while the latter achieved higher faithfulness and relevance scores when evaluated by the Qwen-based judge. The exact McNemar test was not significant and therefore the thesis does not argue that any of the two models is superior.

The rate of wrong citations decreased from 0.250 without RAG to 0.042 with RAG in the ablation. Although there was no significant improvement in overall correctness, the reduction in the number of incorrect legal citations is a practically relevant safety result in the context of a legal-information assistant and would merit further investigation in the form of independent human adjudication.

The voice and latency times indicate that interactive use is technically possible on the tested workstation, especially in streaming mode. The voice assessment was based on synthetic speech, and latency measurements were made in a single setting. Both results are preliminary and need to be validated to larger sample of real speakers and deployment situations.

4.12 Discussion

Before the final retrieval, robustness and generation experiments, the answer-or-refuse threshold was calibrated on validation data. No direct tuning of parameters was done based on the reported test results. Although this process decreases the likelihood of test leakage, it is not a replacement for an independent adjudication by a legal expert and does not eliminate the uncertainty of validation and test samples that are small in size.

The retrieval results answer the first of the empirical questions. The $\text{Recall@5} = 0.9583$ on the final legal test with 24 questions, and $\text{Recall@5} = 0.9652$ on the robustness set with 201 questions, demonstrate that the selected hybrid-plus-reranker setup, in general, spotted the right legal article at the top of the evidence list. The selected evidence window is supported by these findings, however the small manual test and automatically generated robustness questions limit the ability to make broader conclusions.

The generation comparison produced mixed evidence. Automated correctness was higher for ALLaM-7B, whereas the Qwen-based judge assigned higher faithfulness and relevance scores to Qwen2.5-7B. The exact McNemar test did not establish a significant correctness difference, and the judge was not independent of one of the evaluated models. The observed pattern is therefore reported without selecting a statistically superior generator.

The No-RAG comparison provides evidence for a limited citation-safety role of retrieval grounding. Incorrect citations decreased from six to one across the 24 legal questions, but the number of correct citations did not increase and the paired correctness comparison was not significant. A larger sample reviewed by qualified legal experts is needed to determine whether the same pattern extends to legal accuracy and user utility.

The latency experiment showed that retrieval was not the dominant source of delay on the tested workstation. The optimized complete response required 2.96 seconds, and streaming reduced time-to-first-response to 0.82 seconds. These measurements are environment specific, and the voice evaluation used

synthetic speech. Real-speaker, noisy-channel, concurrent-user, and deployment-network tests are required before any claim of operational readiness.

4.13 Limitations

1. Absence of human legal-expert adjudication. The reported answer outcomes were assessed with task-specific rules and an LLM-as-judge proxy. Because a qualified Saudi Labor Law specialist did not review the complete answer set, the metrics should be interpreted as experimental proxies rather than legal certification.
2. Automatically generated robustness questions. The supplementary set contained 201 automatically generated legal questions. The gold article appeared within the top five for 194 questions, and all 201 items were included in the reported metrics. Their wording, ambiguity, and gold-article assignments have not been independently validated by qualified legal experts.
3. Synthetic-speech evaluation. WER and CER were measured with a controlled synthetic-speech proxy rather than recordings from natural Saudi-dialect speakers. The evaluation therefore does not capture spontaneous speech, speaker variation, microphone differences, or background noise.
4. Non-independent LLM judge. Qwen2.5-7B-Instruct was both an evaluated generator and the automated judge for faithfulness and answer relevance. The resulting proxy scores may therefore be affected by self-preference bias.
5. Limited and version-sensitive calculation rules. The calculator implements selected provisions under explicit assumptions and depends on the version of the legal source. It is not a substitute for case-specific legal analysis and must be updated when the law or its implementing regulations change.
6. Limited scope and deployment context. The 24-question final legal test, 45-question generation comparison, synthetic voice evaluation, and single-workstation latency measurements support exploratory conclusions only. They do not establish population-level performance or production readiness.

4.14 Summary

This chapter reported the retrieval, generation, statistical, voice, citation-ablation, latency, and deterministic-calculation results. The selected retrieval setting achieved $\text{Recall}@5 = 0.9583$ on the 24-question final legal test and 0.9652 on the 201-question robustness set. ALLaM-7B obtained the higher

63

observed automated-correctness score, but the paired model difference was not statistically significant. RAG reduced the observed wrong-citation rate from 0.250 to 0.042. The automated, synthetic, and modest-scale nature of the evaluation limits any claim beyond controlled experimental feasibility.

27

Chapter 5: Conclusion and Future Work

5.1 Conclusion

This thesis proposes the development and experimental analysis of the Arabic voice agent for Saudi Labor Law facts. The system combines automatic speech recognition (ASR), versioned evidence base, hybrid dense-lexical retrieval, cross-encoder reranking, fixed reliability gate, local large language model (LLM) generation, deterministic legal calculation, and text-to-speech (TTS). Decoupling retrieval, probabilistic generation, legal arithmetic, and voice processing allows conducting independent evaluation of each subsystem.

Legal corpus includes 249 article entries: 211 active articles have been indexed into searchable knowledge base and 38 obsolete articles have been stored in unindexed archive. Searchable FAQ corpus has 278 records related to labor law. Data split has resulted in 41 rows for validation and 81 rows for raw held-out split, amounting to 80 unique held-out question texts after exact deduplication. Legal test set, robustness set, FAQ resource, and generation set have been kept different to serve their purposes in experiments.

Selected retrieval setup demonstrated $\text{Recall}@1 = 0.8750$, $\text{Recall}@5 = 0.9583$, $\text{MRR} = 0.9097$, and $\text{nDCG}@5 = 0.9221$ on hand-curated 24 legal questions test. Among 201 automatically generated robustness questions, $\text{Recall}@5$ was 0.9652, meaning 194 top-5 retrievals have been successfully performed. For a separate 45 question mixed-type comparison, ALLaM-7B scored 0.8222 accuracy and Qwen2.5-7B scored 0.7111; a paired McNemar test returned $p = 0.1250$.

Ablation analysis showed that retrieval-augmented generation (RAG) helped decrease the number of incorrect citations per question from 0.250 to 0.042 among 24 legal questions test; however, the frequency of correct citations and correctness did not improve noticeably. Under considered conditions, the optimized pipeline took 2.96 seconds to generate a complete response and 0.82 seconds to produce first response in streaming mode. These results confirm the technical viability of local evidence-based system; they do not constitute legal certification or production readiness evidence.

5.2 Contributions

Firstly, they explain the full pipeline for the purposes of labor law queries in spoken and written Arabic, including retrieval-based generation, controlled refusal, deterministic calculation, and optional voice generation. Second, it brings a validation based calibration procedure that ensures that a threshold of reliability is established before the final testing, avoiding gate tuning based on test results. Third, it separates the deterministic numerical calculation from the LLM generation, and takes the corresponding legal article as a support to explain and justify each calculation. Fourth, it constructs a transparent data pipeline to know the leakage, that is, distinguishing between the source rows and the unique queries, evidence indexing, validation samples, samples of FAQ out-of-sample, legal tests, robustness query, and test set of mixed type. Fifth, it offers component-specific metrics, alongside paired statistical testing, percentile bootstrap uncertainty, citation ablation and clear statement of automatic evaluation limitations.

5.3 Limitations

24

The main limitations are that there is no full human or human expert evaluation, that the generator and the LLM evaluator are the same (Qwen2.5-7B-Instruct), and that the size of the evaluation sets is small. There is no natural Saudi Arabic voice, and robustness queries are generated automatically. Only certain (version specific) rules are covered by calculator, and latency was measured only on one machine. The system's purpose is to give general information on the law, and it has not been tested as a means for giving personalized legal advice or for production deployment.

5.4 Future Directions

There are multiple critical areas that future research needs to focus on. To begin with, the system outputs need to be evaluated anonymously by Saudi Labor Law experts according to some predefined criteria, and inter-rater reliability should be reported. The voice component has to be tested on a wide variety of Saudi speakers with different dialects, microphones, and acoustic environments. In addition, the corpus and the set of rules have to be version controlled and protected through regression testing in case of legislative changes. More comparisons should be performed with other local models, prompts, retrieval settings, and reliability techniques. Finally, there should be an experiment on deployment.

5.5 Final Thoughts

This research shows that it is possible to develop a modular voice-agent for Arabic language that can use the official Saudi Labor Law to get evidence, provide grounded answers or refuse due to lack of evidence

and make numeric entitlement determinations with interactional latency times on local computers. The most reliable outcome from the conducted experiment is that the RAG condition resulted in significantly fewer mistakes in legal references than other conditions. Any conclusions about the legal correctness and safety of the agent operation require further verification.

References

- [1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.
- [2] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2022.
- [3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [4] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., online manuscript, Jan. 6, 2026. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>. Accessed: Jul. 18, 2026.
- [5] A. Vaswani et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems* 30, 2017, pp. 5998–6008.
- [6] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems* 33, 2020, pp. 9459–9474.
- [7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proc. ICML*, vol. 202, 2023, pp. 28492–28518.
- [8] J. Shen et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in *Proc. IEEE ICASSP*, 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.
- [9] A. C. Morris, V. Maier, and P. Green, “From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition,” in *Proc. Interspeech*, 2004, pp. 2765–2768, doi: 10.21437/Interspeech.2004-668.
- [10] Ö. T. Özyilmaz, M. Coler, and M. Valdenegro-Toro, “Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning,” in *Proc. Interspeech* 2025, pp. 1158–1162, doi: 10.21437/Interspeech.2025-2260.
- [11] H. Elleuch, Y. Saidi, S. Mdhaftar, Y. Estève, and F. Bougares, “ELYADATA & LIA at NADI 2025: ASR and ADI Subtasks,” in *Proc. Third Arabic Natural Language Processing Conference: Shared Tasks*, 2025, pp. 762–766, doi: 10.18653/v1/2025.arabicnlp-sharedtasks.105.
- [12] 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:2508.04721*, 2025.
- [13] 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, pp. 15757–15773, doi: 10.18653/v1/2023.findings-emnlp.1055.
- [14] I. Besrou, J. He, T. Schreieder, and M. Färber, “RAGentA: Multi-Agent Retrieval-Augmented Generation for Attributed Question Answering,” in *Proc. SIGIR 2025 LiveRAG Challenge*, Padua, Italy, 2025, *arXiv:2506.16988*.
- [15] J. Hrimech, M. Mghari, and Y. Zaz, “Retrieval-Augmented Generation for Arabic Legal Information: The Family Code Case Study,” *TELKOMNIKA*, vol. 23, no. 6, pp. 1495–1505, 2025, doi: 10.12928/TELKOMNIKA.v23i6.27400.
- [16] H. Liu, S. Yusuyin, H. Huang, and Z. Ou, “CTC-TTS: LLM-Based Dual-Streaming Text-to-Speech with CTC Alignment,” *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*, pp. 3777–3794, doi: 10.18653/v1/2024.findings-naacl.240.
- [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, doi: 10.48084/etasr.14229.
- [19] H. Abu Shairah et al., “ALARB: An Arabic Legal Argument Reasoning Benchmark,” in *Proc. Third Arabic Natural Language Processing Conference*, 2025, pp. 389–406, doi: 10.18653/v1/2025.arabicnlp-main.32.
- [20] S. N. Al-Huwari, “NEBRAS: A RAG-Based Question Answering System for Islamic and Legal Guidance,” master’s thesis, An-Najah National University, Nablus, Palestine, 2025. [Online]. Available: <https://hdl.handle.net/20.500.11888/19994>. Accessed: Jul. 18, 2026.
- [21] Seamless Communication et al., “SeamlessM4T: Massively Multilingual & Multimodal Machine Translation,” arXiv:2308.11596, 2023.
- [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,” in *Proc. Third Arabic Natural Language Processing Conference: Shared Tasks*, 2025, pp. 892–898, doi: 10.18653/v1/2025.arabicnlp-sharedtasks.123.
- [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*, pp. 20481–20493, doi: 10.18653/v1/2025.findings-acl.1051.
- [24] J. Alsubhi et al., “Optimizing RAG Pipelines for Arabic: A Systematic Analysis of Core Components,” arXiv:2506.06339, 2025.
- [25] Chroma, “Introduction,” Chroma Documentation. [Online]. Available: <https://docs.trychroma.com/docs/overview/introduction>. Accessed: Jul. 18, 2026.
- [26] Ministry of Human Resources and Social Development, Kingdom of Saudi Arabia, “Labor Law in the Kingdom of Saudi Arabia,” Royal Decree No. M/51, as amended including Royal Decree No. M/44 dated 8 Safar 1446 AH (Aug. 12, 2024); frozen experimental snapshot. [Online]. Available: <https://www.hrsd.gov.sa/knowledge-centre/decisions-and-regulations/regulation-and-procedures/%D9%86%D8%B8%D8%A7%D9%85-%D8%A7%D9%84%D8%B9%D9%85%D9%84; consolidated text: https://www.hrsd.gov.sa/sites/default/files/2025-11/labor-law.pdf>. Accessed: Jul. 12, 2026.
- [27] Ministry of Human Resources and Social Development, Kingdom of Saudi Arabia, “Frequently Asked Questions,” official service-information portal, including labor-system and employment-rights categories. [Online]. Available: <https://www.hrsd.gov.sa/en/contact-us/faq>. Accessed: Jul. 12, 2026.
- [28] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
- [29] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, “Multilingual E5 Text Embeddings: A Technical Report,” arXiv:2402.05672, 2024.
- [30] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” in

- Findings of the Association for Computational Linguistics: ACL 2024, pp. 2318–2335, 2024, doi: 10.18653/v1/2024.findings-acl.137.
- [31] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
 - [32] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/15000000019.
 - [33] Beijing Academy of Artificial Intelligence (BAAI), “BAAI/bge-reranker-v2-m3,” Hugging Face model card. [Online]. Available: <https://huggingface.co/BAAI/bge-reranker-v2-m3>. Accessed: Jul. 18, 2026.
 - [34] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, “C-Pack: Packed Resources For General Chinese Embeddings,” in *Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 641–649, 2024, doi: 10.1145/3626772.3657878.
 - [35] R. Nogueira and K. Cho, “Passage Re-Ranking with BERT,” arXiv:1901.04085, 2019.
 - [36] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv:2312.10997, v5, 2024.
 - [37] M. S. Bari et al., “ALLaM: Large Language Models for Arabic and English,” in *Proc. International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: <https://openreview.net/forum?id=MscdsFVZrN>. Accessed: Jul. 18, 2026.
 - [38] Qwen Team, “Qwen2.5 Technical Report,” arXiv:2412.15115, v2, 2025.
 - [39] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs,” in *Advances in Neural Information Processing Systems 36*, 2023, pp. 10088–10115.
 - [40] W. J. Youden, “Index for Rating Diagnostic Tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.
 - [41] OpenAI, “openai/whisper-large-v3,” Hugging Face model card. [Online]. Available: <https://huggingface.co/openai/whisper-large-v3>. Accessed: Jul. 18, 2026.
 - [42] V. Pratap et al., “Scaling Speech Technology to 1,000+ Languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
 - [43] rany, “edge-tts,” ver. 7.2.8, Python Package Index (PyPI), Mar. 22, 2026. [Online]. Available: <https://pypi.org/project/edge-tts/>. Accessed: Jul. 18, 2026.
 - [44] K. Järvelin and J. Kekäläinen, “Cumulated Gain-Based Evaluation of IR Techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.
 - [45] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993, ISBN: 978-0-412-04231-7.
 - [46] Q. McNemar, “Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, doi: 10.1007/BF02295996.
 - [47] S. Li, Y.-Q. Deng, Z.-L. Zhu, H.-L. Hua, and Z.-Z. Tao, “A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging,” *Diagnostics*, vol. 11, no. 9, Art. no. 1523, 2021, doi: 10.3390/diagnostics11091523.

Appendix A: Automated Evaluation Rubric and Judge Prompt

A.1 Rule-Based Correctness Rubric. Legal questions required a grounded answer and the expected article number; FAQ questions required support from the retrieved context; out-of-scope questions required a safe refusal; and small-talk inputs required an appropriate controlled response. No human or legal-expert adjudicator participated in the reported evaluation.

A.2 LLM Judge Configuration. Faithfulness and answer relevance were evaluated using Qwen2.5-7B-Instruct loaded in 4-bit NF4 [39]. Decoding was deterministic (`do_sample = False`) with `max_new_tokens = 160`. Raw scores from 1 to 5 were normalized with $(s - 1) / 4$, so the mapped values were 0.00, 0.25, 0.50, 0.75, and 1.00.

A.3 Judge Prompt. The judge was instructed to evaluate the answer only against the retrieved context and to return valid JSON with three fields: `faithfulness`, `answer_relevance`, and `reason`. Faithfulness was scored from 1 to 5 according to whether the answer was supported by the retrieved context. Answer relevance was scored from 1 to 5 according to whether the response directly answered the user question. The `reason` field contained a short explanation of the assigned scores.

A.4 Judge Limitation. Qwen2.5-7B-Instruct was both an evaluated generator and the judge. Faithfulness and relevance are therefore secondary proxy measures that may be affected by self-preference bias and must not be presented as independent human or legal-expert validation.

A.5 Statistical Procedure. Accuracy intervals used 10,000 percentile-bootstrap resamples with limits at the 2.5th and 97.5th percentiles. No BCa, bias-correction, or acceleration adjustment was applied. The paired exact McNemar table was: both correct = 31, ALLaM only correct = 6, Qwen only correct = 1, both incorrect = 7; exact two-sided $p = 0.1250$.

Appendix B: Data-Lineage and Deduplication Record

B.1 FAQ Split. The identifier-based split produced 278 indexing rows, 41 validation rows, and 81 raw held-out rows. The raw split therefore contained 400 rows.

B.2 Exact Duplicate. The question “ما هي أهداف برنامج حماية الأجور؟” occurred in `faq_id` 248 and `faq_id` 94. `faq_id` 248 was retained as the canonical evaluation record. `faq_id` 94 was removed before scoring; its `document_unit_id` was `b0ebcab351b4739f0566a90080daa23d875b475bac36659a0088539697b8a266`. The operation was exact question-text deduplication, not a quality exclusion.

B.3 Final FAQ Resource. The 81 raw held-out rows represented 80 unique question texts after deduplication.

B.4 Robustness Set. The supplementary legal robustness set contained 201 automatically generated questions. The gold article appeared within the top five for 194 questions; all metrics were computed over all 201 questions.

B.5 Generation Comparison. The paired generation set was independent of the FAQ held-out resource and contained 45 questions: 14 FAQ, 17 legal, 12 out-of-scope, and 2 small-talk. The fixed gate produced 29 generated answers and 16 safe refusals. Accuracy counted both correct answers and correct refusals.

unted both correct answers and correct refusals.