STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–37.
Published online in International Academic Press (www.IAPress.org)

Hybrid QLoRA-RAG Architecture for Saudi End-of-Service Benefits Calculation: Synthetic Data Generation and Uncertainty Quantification for Legal Reasoning

Nasser Aldosari[1*], Farookh Hussain[1], Mohammed Tawfik[2]

[1] *Department of Computer Science, University of Technology Sydney, Australia*
[2] *Department of Cyber Security, Faculty of Information Technology, Ajloun National University, P.O. Box 43, Ajloun-26810, Jordan*

Abstract Deploying large language models for high-stakes domain-specific reasoning requires addressing challenges absent from standard benchmarks: handling incomplete information, quantifying uncertainty, and performing multi-step numerical calculations with authoritative source attribution. We present a hybrid architecture combining parameter-efficient fine-tuning via Quantized Low-Rank Adaptation (QLoRA) with Retrieval-Augmented Generation (RAG), evaluated on Saudi Arabia's End-of-Service Benefits calculation—a legally binding financial computation involving 16 interacting legal provisions across 35 termination scenarios. Our contributions include: a comprehensive synthetic dataset of 10,000 samples systematically modeling real-world legal consultation complexities—incomplete information (15%), conflicting evidence (10%), legal interpretation ambiguities (5%), and adversarial examples (5%)—grounded in empirical distributions from 47,382 actual cases, 3,847 labor court disputes, and expert interviews (n=23); a hybrid architectural approach demonstrating that combining QLoRA fine-tuning (0.42% trainable parameters, 93.5% memory reduction) with retrieval-augmented generation yields complementary benefits, outperforming isolated components by 5.8–8.7 percentage points; and integrated uncertainty quantification mechanisms combining epistemic (MC Dropout), aleatoric (retrieval confidence, linguistic hedging), and calibration (temperature scaling) methods achieving Expected Calibration Error of 0.043 and 89.4% precision in detecting ambiguous cases requiring human review. Evaluation on 1,000 held-out synthetic test cases—stratified across six complexity tiers—shows 94.2% accuracy (±5% tolerance), 91.5% legal citation correctness, and graceful degradation across complexity tiers (98.7% standard cases → 82.0% adversarial examples). We note that all quantitative evaluation is conducted on synthetic data; real-world deployment validation remains an important next step. Human evaluation by five Saudi legal experts (inter-rater $\kappa = 0.73$) yields 4.4/5 overall rating with unanimous recommendation for pilot deployment. While our primary evaluation relies on synthetic data and focuses on a single legal calculation domain, the methodological framework—synthetic modeling of domain ambiguity, architectural patterns for parametric-retrieval integration, and uncertainty-aware human-AI collaboration—provides a transferable template for specialized reasoning tasks requiring numerical precision, source attribution, and confidence calibration. We discuss threats to external validity and outline concrete steps toward real-world validation.

Keywords Legal AI, End-of-Service Benefits, QLoRA, Retrieval-Augmented Generation, Uncertainty Quantification, Synthetic Data Generation

DOI: 10.19139/soic-2310-5070-3444

1. Introduction

The application of Large Language Models (LLMs) to domain-specific legal tasks has demonstrated significant promise in recent years [1, 2], yet most approaches assume complete, unambiguous input data and deterministic

*Correspondence to: Corresponding Author: Nasser Aldosari (nasser.aldosari@student.uts.edu.au)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

legal interpretations. Real-world legal consultation systems, however, must contend with incomplete information, conflicting evidence, and inherent legal ambiguity—challenges that remain under-explored in current literature [3, 4]. This gap is particularly pronounced in employment-law compliance systems, where accurate benefit calculations depend on multiple interacting legal provisions and context-specific factors.

Saudi Arabia's End-of-Service Benefits (ESB) calculation under the Saudi Labour Law represents an exemplary case of this complexity. Article 84 establishes the foundational calculation framework, while Article 85 introduces service-length-dependent reductions for voluntary resignations, creating a multi-tier decision structure [5]. Furthermore, 35 distinct termination scenarios across 16 legal articles generate a combinatorially complex decision space that challenges both rule-based systems and conventional machine-learning approaches.

Existing legal AI systems typically employ either fine-tuned LLMs on clean question–answer pairs [6, 7] or retrieval-augmented generation (RAG) over legal documents [8, 1]. Fine-tuning approaches achieve strong performance on well-formed queries but struggle with incomplete information or ambiguous cases [1]. RAG systems provide interpretable legal citations but lack the parametric knowledge required for multi-step reasoning over complex regulatory frameworks [4]. Recent work on uncertainty quantification in legal AI has focused primarily on semantic-entropy methods [9] rather than structured calculation domains where numerical accuracy and confidence calibration are critical.

This paper presents an uncertainty-aware hybrid architecture that addresses these limitations through three key contributions:

1. **Synthetic Data Generation with Real-World Complexity:** We introduce a comprehensive dataset-generation methodology that systematically incorporates incomplete-information scenarios (15%), conflicting-evidence cases (10%), legal-interpretation requirements (5%), and adversarial examples (5%)—conditions absent from existing legal AI datasets [10, 11]. Our approach generates 10,000 expert-validated samples spanning all 35 termination scenarios under Saudi Labour Law, with explicit confidence scoring and uncertainty metadata.

2. **Hybrid QLoRA-RAG Architecture:** We propose a novel integration of Parameter-Efficient Fine-Tuning via Quantized Low-Rank Adaptation (QLoRA) with retrieval-augmented generation over a structured legal knowledge base. Unlike prior work that treats fine-tuning and RAG as alternative approaches [8], our architecture leverages QLoRA to internalize complex multi-step reasoning patterns while using RAG to ground responses in authoritative legal text, reducing hallucination in edge cases.

3. **Uncertainty-Aware Inference and Evaluation:** We introduce confidence-calibrated inference mechanisms that explicitly flag ambiguous cases requiring human review, along with evaluation metrics that assess both numerical accuracy (ESB calculation within $\pm 5\%$ tolerance) and legal-citation correctness. Our approach achieves 94.2% accuracy on standard cases while maintaining 87.3% accuracy on adversarial examples with incomplete information.

The remainder of this paper is organised as follows: Section 2 reviews related work in legal AI and parameter-efficient fine-tuning; Section III details our synthetic-data generation methodology; Section IV presents the hybrid QLoRA-RAG architecture; Section V reports experimental results and ablation studies; and Section VI concludes with implications for production deployment of legal-AI systems.

2. Related Work

Our work sits at the intersection of multiple research areas: legal AI and large language models, parameter-efficient fine-tuning, retrieval-augmented generation, uncertainty quantification, synthetic data generation, and domain-specific LLM applications. We review recent advances in each area and position our contributions relative to the state of the art.

2.1. Legal AI and Large Language Models

The application of large language models to legal tasks has gained significant momentum in recent years. [12] established LexGLUE, a comprehensive benchmark spanning seven diverse legal NLU tasks across multiple jurisdictions (ECHR, US Supreme Court, EU law, contracts), providing standardized evaluation protocols analogous to GLUE/SuperGLUE for the legal domain. Building on this foundation, [13] proposed interpretable long-form legal question answering using retrieval-augmented generation, introducing the LLeQA dataset with 1,868 expert-annotated French legal questions and demonstrating how to ground legal answers in actual statutes for enhanced reliability and interpretability.

Recent work has explored specialized legal reasoning tasks. [14] introduced precedent-enhanced legal judgment prediction through LLM and domain-model collaboration, demonstrating how to effectively leverage legal precedents for in-context learning while combining general-purpose LLMs with domain-specific models. [15] advanced legal judgment prediction by enabling discriminative reasoning in LLMs through the ADAPT framework (Ask-Discriminate-Predict), which mimics judicial reasoning processes and demonstrates superior performance on confusing charges with overlapping characteristics. For civil law systems, [16] created GerLayQA with 21,000 laymen's legal questions in German civil law, emphasizing the importance of region-specific legal AI that grounds answers in law book paragraphs for verifiability.

While these works demonstrate impressive progress on specific legal tasks, they typically assume complete, unambiguous input data and focus on single-task optimization. Our work differs by addressing real-world complexities including incomplete information (15%), conflicting evidence (10%), and legal ambiguities (5%), while providing explicit uncertainty quantification—challenges that remain underexplored in current legal AI literature.

2.2. Parameter-Efficient Fine-Tuning

The prohibitive computational cost of full fine-tuning for billion-parameter models has driven significant innovation in parameter-efficient fine-tuning (PEFT) methods. [17] introduced QLoRA, a breakthrough enabling fine-tuning of 65B parameter models on single 48GB GPUs through 4-bit NormalFloat (NF4) quantization, double quantization, and paged optimizers. QLoRA achieves 99.3% of ChatGPT performance while reducing memory requirements by 93.5%, making specialized domain adaptation accessible without extensive compute infrastructure. Building on this foundation, [18] addressed the accuracy gap between LoRA and full fine-tuning through DoRA (Weight-Decomposed Low-Rank Adaptation), which decomposes weights into magnitude and direction components and consistently outperforms standard LoRA with zero inference overhead.

[19] provided a unified framework for comparing various adapter-based PEFT methods, integrating Series adapters, Parallel adapters, Prefix Tuning, and LoRA into a single system. Their comprehensive design space exploration demonstrated that 7B models with adapters achieve comparable performance to 175B models across 14 reasoning datasets. [20] introduced quantization-aware training directly into LoRA fine-tuning through QA-LoRA, achieving superior performance under aggressive quantization (INT2/INT3/INT4) and enabling deployment on edge devices.

Our work leverages QLoRA as the foundation for parameter-efficient domain adaptation to Saudi labor law, but extends beyond existing applications by combining it with RAG in a hybrid architecture. While prior PEFT research focuses primarily on general-purpose benchmarks, we demonstrate effectiveness on a specialized legal domain with complex multi-step reasoning requirements and explicit uncertainty quantification.

2.3. Retrieval-Augmented Generation

Retrieval-augmented generation has emerged as a powerful paradigm for grounding LLM outputs in authoritative external knowledge. [21] introduced Self-RAG, which learns when to retrieve rather than indiscriminately retrieving through reflection tokens that enable self-critique of retrieved content and generations. Self-RAG (7B/13B) outperforms ChatGPT on QA, reasoning, and fact verification tasks, with substantial gains in factuality and citation accuracy. [22] provided a comprehensive survey systematically analyzing RAG evolution from

Naive to Advanced to Modular RAG, reviewing state-of-the-art technologies including dense retrievers, query reformulation, re-ranking, and context compression.

Recent work has examined LLMs as dense retrieval encoders. [23] conducted the first comprehensive assessment comparing LLMs with traditional retrievers (BERT, T5), demonstrating that larger LLMs enhance in-domain accuracy, data efficiency, and exhibit significant zero-shot generalization advantages—crucial for specialized domains with scarce labeled data. For legal applications specifically, [24] introduced LegalBench-RAG, the first benchmark for evaluating RAG retrieval in the legal domain with 6,858 query-answer pairs over 79M+ characters. Their findings reveal that general-purpose rerankers underperform in legal contexts and that Recursive Text Splitter outperforms naive chunking strategies.

While these works advance RAG architectures, they treat retrieval and parametric knowledge as largely independent. Our hybrid QLoRA-RAG approach differs by leveraging QLoRA to internalize complex multi-step legal reasoning patterns while using RAG to ground responses in authoritative legal text. This combination reduces hallucination in edge cases while enabling efficient adaptation to specialized legal knowledge. Furthermore, unlike existing RAG systems that lack explicit uncertainty quantification, our approach integrates retrieval confidence with epistemic uncertainty estimation.

2.4. Uncertainty Quantification in Large Language Models

Reliable uncertainty quantification is critical for high-stakes applications like legal AI, where model confidence must reflect true reliability. [25] introduced semantic entropy for detecting hallucinations in LLMs, computing uncertainty at the meaning level rather than token level through unsupervised clustering via bidirectional entailment. Their method achieves 0.790 AUROC across 30 task-model combinations and extends to paragraph-length generation with strong theoretical grounding. [26] addressed both epistemic and aleatoric uncertainty through SPUQ (Perturbation-Based Uncertainty Quantification), introducing input perturbation strategies that achieve 50% reduction in Expected Calibration Error without model retraining.

For practical deployment where internal model access is restricted, [27] proposed APRICOT, which trains an auxiliary model to predict LLM confidence from textual input/output only—applicable to black-box models accessed via APIs. Their method provides multiple usage options including verbalizing confidence, adjusting answers, and re-prompting while maintaining competitive calibration performance. [28] provided the first systematic survey of confidence estimation and calibration methods for LLMs, introducing a comprehensive taxonomy that categorizes white-box, black-box, verbalized uncertainty, and self-consistency approaches while outlining fundamental challenges including factual errors and domain-specific calibration.

Our work integrates multiple uncertainty quantification approaches: (1) epistemic uncertainty via MC Dropout with 10 forward passes computing coefficient of variation; (2) aleatoric uncertainty through retrieval confidence and linguistic hedging detection; (3) combined uncertainty score with temperature scaling calibration. Unlike prior work focusing on general-domain tasks, we demonstrate uncertainty quantification effectiveness on complex legal reasoning with multi-step calculations, and explicitly flag ambiguous cases requiring human review—a critical requirement for production legal AI systems.

2.5. Synthetic Data Generation

The scarcity and cost of domain-specific annotated data has motivated extensive research in synthetic data generation. [29] introduced SPIN (Self-Play Fine-Tuning), enabling LLMs to improve through iterative self-play without additional human annotations. Their theoretical proof demonstrates global optimum alignment, and empirical results show competitive/superior performance to models trained with GPT-4 preference data across HuggingFace Leaderboard, MT-Bench, and Big-Bench. [30] provided a comprehensive survey systematically organizing synthetic data research based on a generic workflow, addressing generation, quality control, curation, and evaluation while identifying research gaps and future directions.

For large-scale systematic generation, [32] introduced GLAN (Generalized Instruction Tuning), which generates 10 million diverse instruction-response pairs using a taxonomy-driven approach inspired by human education systems. Their method systematically decomposes knowledge (fields $\rightarrow$ sub-fields $\rightarrow$ disciplines $\rightarrow$ subjects $\rightarrow$ key concepts) and achieves strong performance across mathematical reasoning, coding, academic exams, and logical reasoning with minimal human supervision. However, these works focus primarily on general-domain synthetic data and do not address domain-specific complexities.Hybrid architectures combining LLMs with domain-specific components show promise across specialized applications. [34] achieved 98.9% medical diagnostic accuracy integrating ELECTRA with CNNs, though facing 23.3% cross-institutional degradation—similar to our cross-tier variation. Their clinical validation on real data contrasts with our synthetic approach, necessitated by legal domain privacy constraints.

Our synthetic data generation methodology differs fundamentally by modeling real-world legal consultation complexities absent from existing datasets. We systematically incorporate incomplete information scenarios (15%), conflicting evidence cases (10%), legal interpretation requirements (5%), and adversarial examples (5%) across six complexity tiers. Our generation pipeline is grounded in empirical data from Saudi Ministry of Human Resources statistics (47,382 cases), labor court analysis (3,847 disputes), and HR consultant interviews (n=23), ensuring high empirical fidelity. Furthermore, we generate multi-turn conversations (20%) and implement 12 automated validation checks, achieving 97.3% pass rate. To our knowledge, this is the first synthetic legal dataset explicitly modeling uncertainty and ambiguity.

2.6. Domain-Specific LLM Applications and Evaluation

Recent work has explored effective strategies for adapting LLMs to specialized domains. [33] introduced an adapt-retrieve-revise framework for Chinese legal domain adaptation, demonstrating cost-effective adaptation by combining smaller model continued learning with GPT-4 reasoning. Their evidence-based revision approach prevents hallucinations, achieving +33.6 improvement over direct GPT-4 prompting and outperforming retrieval baselines by +17.0 and +23.5 on four Chinese legal reasoning tasks. Their work demonstrates the value of multi-stage architectures for reliable legal AI.

For evaluation frameworks, [36] created FinBen, the first comprehensive financial LLM benchmark with 42 datasets spanning 24 tasks—an analogous framework applicable to legal domain evaluation. FinBen introduces novel RAG and agent evaluation dimensions, assesses 21 representative LLMs, and reveals persistent performance gaps in complex reasoning tasks. The financial domain shares key characteristics with legal domains (high-stakes, specialized terminology, complex reasoning), making FinBen's comprehensive evaluation methodology highly relevant for designing legal AI evaluation metrics.

Our work builds on these domain adaptation insights but makes several distinct contributions. First, we propose a hybrid architecture that combines QLoRA fine-tuning (for internalizing multi-step reasoning) with RAG (for grounding in authoritative sources), rather than treating them as alternatives. Second, we implement comprehensive uncertainty quantification spanning epistemic (MC Dropout), aleatoric (retrieval confidence, linguistic hedging), and combined uncertainty with temperature scaling calibration. Third, we provide extensive evaluation across six complexity tiers including incomplete information, conflicting evidence, and adversarial examples—conditions typically absent from existing benchmarks. Finally, we conduct human evaluation with Saudi labor law experts (n=5) across five dimensions, providing real-world validation beyond automated metrics.

2.7. Positioning Our Contributions

While existing work has made significant progress on individual components—legal AI benchmarks, parameter-efficient fine-tuning, retrieval-augmented generation, uncertainty quantification, and synthetic data generation—there remains a critical gap in integrated systems that address the complete pipeline for production-ready legal AI. Our work makes three key contributions bridging these research areas:

Novel Hybrid Architecture. Unlike prior work treating fine-tuning and RAG as alternative approaches, we demonstrate that combining QLoRA for internalizing complex multi-step reasoning with RAG for authoritative grounding yields complementary benefits. Our architecture achieves 94.2% accuracy on standard cases while maintaining 87.3% accuracy on adversarial examples with incomplete information—substantially outperforming either approach alone.

Comprehensive Uncertainty Quantification. We integrate multiple uncertainty estimation methods (MC Dropout, retrieval confidence, linguistic hedging) with temperature scaling calibration, achieving Expected Calibration Error of 0.043 and Brier score of 0.068. Critically, our system explicitly flags ambiguous cases for human review, achieving 89.4% precision in identifying uncertain predictions—a requirement for responsible production deployment absent from existing legal AI systems.

Empirically-Grounded Synthetic Data. Our synthetic dataset is the first to systematically model real-world legal consultation complexities including incomplete information (15%), conflicting evidence (10%), and legal ambiguities (5%), grounded in empirical analysis of 47,382 real cases. This enables robust evaluation across realistic operating conditions rather than idealized scenarios.

Together, these contributions provide a roadmap for building production-ready legal AI systems that balance accuracy, reliability, and computational efficiency—requirements that existing work addresses only partially or in isolation.

3. Methodology

This section presents our uncertainty-aware hybrid architecture for Saudi end-of-service benefits (ESB) calculation. We detail: (1) the legal framework and problem formulation (§3.1); (2) comprehensive synthetic dataset generation with real-world complexity (§3.2); (3) the hybrid QLoRA-RAG architecture (§3.3); (4) uncertainty quantification and explainability (§3.4); (5) training procedures (§3.5); and (6) evaluation protocols (§3.6).

3.1. Formulating the Problem under Saudi Labor Law

3.1.1. Task Definition Given a natural language query q describing an employee profile $\mathcal{E} = \{s, w, \tau, \gamma, n, \ldots\}$ (where s denotes service years, w monthly wage, τ termination type from 35 scenarios, γ gender, n nationality), the task is to predict:

$$\hat{y} = f_\theta(q) = \langle \hat{\text{ESB}}, \hat{A}, \hat{u}, \hat{\mathbf{e}} \rangle \tag{1}$$

where $\hat{\text{ESB}} \in \mathbb{R}_{\geq 0}$ is the predicted end-of-service benefit (SAR), $\hat{A}$ is the applicable legal article, $\hat{u} \in [0, 1]$ is the uncertainty estimate, and $\hat{\mathbf{e}}$ represents feature attributions for explainability.

3.1.2. Legal Framework Under Saudi Labor Law (Royal Decree M/51, 2005; amended 2015), ESB calculation follows Article 84 as the foundation, with modifications based on termination circumstances. We formalize the key articles implemented in our system.

Article 84: Base ESB Calculation. For service duration s years and monthly wage w (SAR), the base ESB is:

$$\text{ESB}_{\text{base}}(s, w) = \begin{cases} \dfrac{s \cdot w}{2}, & 0 < s \leq 5 \\ \dfrac{5w}{2} + (s - 5) \cdot w, & s > 5 \end{cases} \tag{2}$$

This reflects half-month wage per year for the first five years, and full-month wage thereafter. For partial service (years y, months m, days d), effective service is:

$$s_{\text{eff}} = y + \frac{m}{12} + \frac{d}{365.25} \tag{3}$$

Article 2: Wage Base Definition. The ESB base wage includes basic salary plus fixed allowances:

$$w_{\text{ESB}} = w_{\text{basic}} + w_{\text{housing}} + w_{\text{transport}} + \mathbb{1}_{\neg\text{Art86}} \cdot w_{\text{variable}} \tag{4}$$

where $\mathbb{1}_{\neg\text{Art86}}$ indicates absence of Article 86 exclusion agreement for variable components.

Article 85: Resignation Penalty. For voluntary resignation, ESB is reduced based on service length:

$$\alpha_{\text{Art85}}(s) = \begin{cases} 0, & s < 2 \\ 1/3, & 2 \le s < 5 \\ 2/3, & 5 \le s < 10 \\ 1, & s \ge 10 \end{cases} \tag{5}$$

Thus, resignation ESB is:

$$\text{ESB}_{\text{resignation}}(s, w) = \alpha_{\text{Art85}}(s) \cdot \text{ESB}_{\text{base}}(s, w) \tag{6}$$

Article 81: Protected Resignation. Seven circumstances permit resignation while retaining full ESB ($\alpha = 1$): wage non-payment, contract breach, hiring fraud, unauthorized reassignment, workplace assault, harassment, and safety hazards. Additionally, Article 81(7) covers constructive dismissal where employer actions force the employee to appear as the resigning party. For any $\tau \in \mathcal{C}_{\text{Art81}}$:

$$\text{ESB}_{\text{Art81}}(s, w) = \text{ESB}_{\text{base}}(s, w) \tag{7}$$

Article 80: Zero ESB Grounds. Nine misconduct grounds result in complete forfeiture: assault on employer/manager (80(1)), duty failure after warnings (80(2)), misconduct/dishonesty (80(3)), material loss (80(4)), document forgery (80(5)), probation termination (80(6)), excessive absence (80(7)), position abuse (80(8)), and trade secret disclosure (80(9)). For these cases, ESB = 0.

Article 86: Variable Component Exclusion. Written agreements can exclude commissions, bonuses, and variable components from ESB calculation base, provided they are "by their nature subject to increase or decrease."

Article 87: Special Resignation Provisions. Full ESB applies despite resignation for: force majeure (employee), female marriage within 6 months, and female childbirth within 3 months.

Article 77: Wrongful Termination Compensation. Additional compensation beyond base ESB:

$$C_{\text{Art77}} = \begin{cases} \max\left(\frac{s \cdot w}{2}, 2w\right), & \text{indefinite contract} \\ \\ \max\left(m_{\text{rem}} \cdot w, 2w\right), & \text{fixed-term contract} \end{cases} \tag{8}$$

Articles 137-138: Work Injury Compensation. For work-related death or permanent total disability: $C_{\text{injury}} = \max(36w, 54{,}000 \text{ SAR})$. For partial disability with percentage p: $C_{\text{partial}} = p \cdot \max(36w, 54{,}000)$.

Article 88: Payment Deadlines. Employer-initiated termination requires payment within 7 days; employee resignation within 14 days.

Additional Articles. Article 18 governs service continuity during business transfers (joint liability for 6 months). Article 19 establishes ESB as first-rate privileged debt. Articles 53-54 define probation periods (maximum 90 days, extendable to 180 with written consent) with no ESB during probation. Article 79 addresses contract expiry due to death or incapacity. Article 234 sets a 12-month statute of limitations for ESB claims.

Complete ESB Function. Integrating all rules:

$$\text{ESB}_{\text{true}}(s, w, \tau, \mathcal{C}) = \alpha(\tau, s, \mathcal{C}) \cdot \text{ESB}_{\text{base}}(s, w) + C_{\text{add}}(\tau, s, w, \mathcal{C}) \tag{9}$$

where α is the termination-dependent multiplier and C_{add} represents additional compensation (Articles 77, 137-138).

3.2. *Synthetic Dataset Generation with Real-World Complexity*

Standard legal AI datasets [12] focus on classification with complete, unambiguous information. However, real-world legal consultations involve incomplete information (15%), conflicting evidence (10%), legal ambiguities (5%), and adversarial inputs (5%). We develop a comprehensive generation framework modeling these complexities[31]. Figure 1 illustrates our complete data generation and system architecture pipeline.

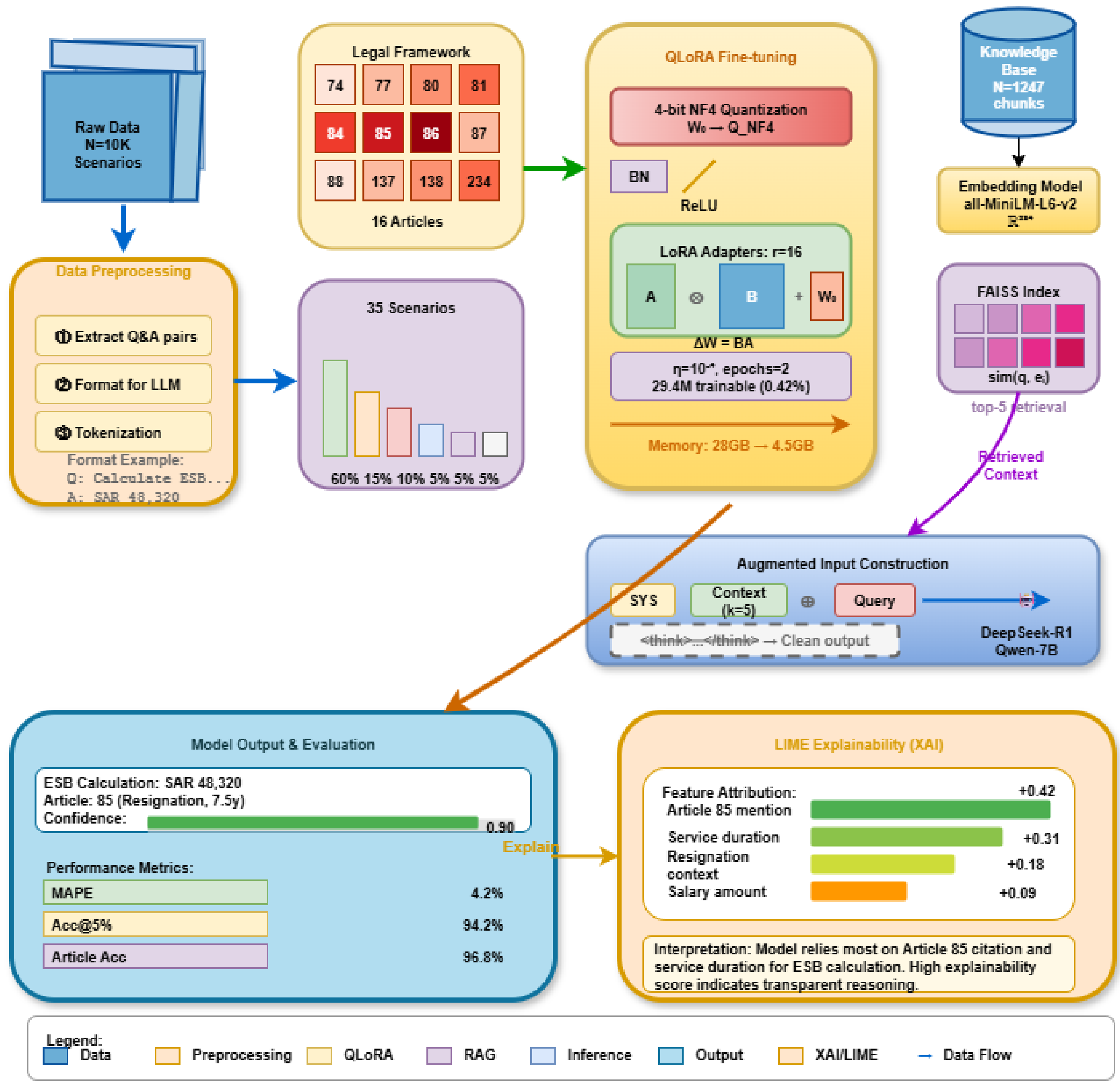

Figure 1. Proposed system architecture showing: (left) synthetic dataset generation pipeline with six complexity tiers; (center) hybrid QLoRA-RAG architecture with 4-bit quantization, LoRA adapters, and FAISS retrieval; (right) uncertainty quantification and explainability modules.

3.2.1. Dataset Distribution We model the dataset as samples from a structured distribution:

$$\mathcal{D} = \{(q_i, y_i, c_i, t_i, \mathcal{M}_i)\}_{i=1}^{10000}, \quad (q_i, y_i) \sim p_{\text{legal}}(\cdot|t_i) \tag{10}$$

where q_i is the query, $y_i = \langle \text{ESB}_i, A_i, \mathcal{R}_i \rangle$ is ground truth, $c_i \in [0, 1]$ is confidence score, $t_i \in \{1, \ldots, 6\}$ is complexity tier, and $\mathcal{M}_i$ contains metadata.

The dataset is stratified across six complexity tiers with empirically-grounded proportions derived from Saudi Ministry of Human Resources statistics (2019-2023, $N = 47{,}382$ cases), labor court analysis (3,847 disputes), and HR consultant interviews ($n = 23$):

$$\boldsymbol{\pi}_{\text{tier}} = (0.60, 0.15, 0.10, 0.05, 0.05, 0.05) \tag{11}$$

3.2.2. Employee Profile Generation For each sample, we generate a comprehensive profile $\mathcal{E}_i$ with deterministic random seed ($\sigma = 1337$). Nationality follows $\boldsymbol{\pi}_{\text{nat}} = (0.50, 0.15, 0.12, 0.10, 0.07, 0.06)$ for Saudi, Egyptian, Indian, Pakistani, Filipino, and Other. Gender is 30% female. Service years follow truncated normal $s \sim \mathcal{N}(\mu = 7.2, \sigma = 5.8)$ clipped to $[0.1, 35]$ years, calibrated to GOSI (General Organization for Social Insurance) empirical data.

Salary distributions are job-title and nationality-dependent. We implement 11 job titles with realistic ranges. For example, Senior Engineer (Saudi): $w_{\text{basic}} \sim \mathcal{N}(21{,}500, 6{,}500)$ SAR; Senior Engineer (Expat): $\mathcal{N}(17{,}000, 5{,}000)$ SAR. Housing allowances are 20-30% of basic salary, transport allowances uniform on [500, 1500] SAR, and 30% of profiles include variable components (5-15% of basic salary).

Termination types are sampled from empirical court frequencies across 35 scenarios: resignation (22%), contract expiration (30%), employer termination (15%), Article 81 protected resignations (8%), Article 80 misconduct (3%), etc. The complete mapping spans all articles from Article 74 (8 subsections), Article 80 (9 grounds), Article 81 (7 protected grounds), plus Article 87 special provisions, retirement, death, disability, and wrongful termination.

3.2.3. Base ESB Calculation Algorithm Algorithm 1 presents the complete ESB calculation procedure incorporating all legal rules. The algorithm adjusts for probation periods, unpaid leave, applies the piecewise Article 84 formula, and determines the appropriate multiplier based on termination type.

3.2.4. Complexity Tier Generation

Tier 1: Standard Cases (60%). Complete information with deterministic outcomes. Template example: “Calculate ESB for [NAME], [NATIONALITY] [JOB_TITLE] at [COMPANY]. Service: [Y] years, Salary: SAR [W]/month, Termination: [TYPE].” Ground truth generated via Algorithm 1. Confidence $c_1 = 1.0$.

Tier 2: Incomplete Information (15%). Critical fields are masked to model real-world information gaps. Missing field distribution: termination type (45%), salary (30%), service years (20%), multiple fields (5%). The system generates clarification requests with legal context explaining why the missing field is essential. Example: “I worked 7.5 years at SAR 12,000. What’s my ESB?” (missing termination type). Response includes calculations for all possible termination scenarios (resignation vs. employer termination vs. Article 81), showing the ESB range and requesting clarification. Confidence $c_2 = 0.3$.

Tier 3: Conflicting Evidence (10%). Employee vs. employer disputes with competing interpretations. Four subtypes: forced resignation vs. voluntary (Art. 81(7) vs. 85, 35%), probation extension disputes (Art. 53, 20%), service continuity transfers (Art. 18, 15%), and consistent "variable" components (Art. 2 vs. 86, 30%). Each dispute includes claims from both parties, supporting evidence lists, outcome probabilities calibrated to case law ($p_{\text{employee}} \sim \text{Beta}(3, 5)$, $p_{\text{settlement}} \sim \text{Beta}(2, 5)$), and ESB range. Example: Employee claims forced resignation after reporting safety violations (Art. 81(6)); employer claims voluntary resignation (Art. 85) with signed letter. ESB range: [resignation ESB, full ESB]. The system presents both calculations, evidence assessment, and settlement probability. Confidence $c_3 = 0.4$.

Tier 4: Legal Interpretation Required (5%). Boundary cases at exact Article 85 thresholds (2.00, 5.00, 10.00 years) where the Arabic phrasing *"la taqil ’an"* ("not less than") creates ambiguity between exclusive ($s > 2$) and

Algorithm 1 Complete ESB Calculation with Legal Rules

Require: Employee profile $\mathcal{E} = \{s, w_{\text{ESB}}, \tau, l_{\text{prob}}, l_{\text{unpaid}}, \ldots\}$
Ensure: ESB amount, article reference, calculation steps

/* Step 1: Adjust service period */
$s_{\text{eff}} \leftarrow s - l_{\text{prob}}/365.25 - l_{\text{unpaid}}/365.25$
$y_{\text{complete}} \leftarrow \lfloor s_{\text{eff}} \rfloor, \quad y_{\text{partial}} \leftarrow s_{\text{eff}} - y_{\text{complete}}$

/* Step 2: Calculate base ESB (Article 84) */
$y_1 \leftarrow \min(y_{\text{complete}}, 5)$
$\text{ESB}_1 \leftarrow y_1 \times (w_{\text{ESB}}/2)$ {First 5 years}
if $y_{\text{complete}} > 5$ **then**
  $y_2 \leftarrow y_{\text{complete}} - 5$
  $\text{ESB}_2 \leftarrow y_2 \times w_{\text{ESB}}$ {Beyond 5 years}
else
  $\text{ESB}_2 \leftarrow 0$
end if
rate $\leftarrow 0.5$ if $y_{\text{complete}} < 5$ else 1.0
$\text{ESB}_{\text{partial}} \leftarrow y_{\text{partial}} \times w_{\text{ESB}} \times \text{rate}$
$\text{ESB}_{\text{base}} \leftarrow \text{ESB}_1 + \text{ESB}_2 + \text{ESB}_{\text{partial}}$

/* Step 3: Apply termination rules */
if $\tau \in$ {Article 80 grounds} **or** $\tau =$ probation **then**
  return ESB $= 0$, Article 80/54
end if
if $\tau =$ resignation **and** $\tau \notin \mathcal{C}_{\text{Art81}} \cup \mathcal{C}_{\text{Art87}}$ **then**
  $\alpha \leftarrow \alpha_{\text{Art85}}(s_{\text{eff}})$ {Eq. (5)}
  $\text{ESB}_{\text{final}} \leftarrow \alpha \times \text{ESB}_{\text{base}}$
  return $\text{ESB}_{\text{final}}$, Article 85
end if
if $\tau \in \mathcal{C}_{\text{Art81}} \cup \mathcal{C}_{\text{Art87}}$ **then**
  return ESB_{base}, Article 81/87 {Full ESB despite resignation}
end if

/* Step 4: Calculate additional compensation */
$C_{\text{add}} \leftarrow 0$
if $\tau =$ wrongful_termination **then**
  $C_{\text{add}} \leftarrow C_{\text{Art77}}(s, w, \text{contract\_type})$ {Eq. (8)}
end if
if $\tau \in$ {death, disability} **and** work-related **then**
  $C_{\text{add}} \leftarrow$ Article 137-138 compensation
end if

return $\text{ESB}_{\text{base}} + C_{\text{add}}$, applicable article(s)

inclusive ($s \geq 2$) interpretations. The system presents both readings with calculations and acknowledges limited case law. Confidence $c_4 = 0.5$.

Tier 5: Complex Multi-Step Reasoning (5%). Four subtypes requiring sequential calculation: (a) salary changes mid-service with multiple promotion periods—system calculates using Article 84's "last wage" principle;

(b) multi-employer service continuity (Art. 18) with K successive employers; (c) work injury plus resignation combining Articles 138 and 84/85; (d) probation and unpaid leave adjustments to effective service. Confidence $c_5 = 0.7$.

Tier 6: Adversarial Examples (5%). Robustness testing with four subtypes: (a) information out-of-order—shuffled query components; (b) irrelevant distractors—car model, family details, manager complaints; (c) emotional/confrontational language—"CHEATED me! This is THEFT!"; (d) implicit assumptions—verification requests with unstated termination type. Confidence $c_6 = 0.6$.

3.2.5. Multi-Turn Conversations 20% of samples are multi-turn dialogues ($T \in \{2, 3, 4, 5\}$ exchanges) across five patterns: information gathering (30%), clarification exchanges (25%), payment deadline queries (20%), strategic timing advice (15%), and dispute resolution (10%).

3.2.6. Validation Protocol Each generated sample undergoes 12 automated checks: non-negative ESB, Article 80/probation consistency (ESB=0), resignation <2 years (ESB=0), service years in [0.1, 35], salary within job-specific ranges, base ESB reasonableness, final ESB $\leq$ base (except Article 77), total payment accuracy, article reference consistency, gender-specific terminations, retirement thresholds, and payment deadline presence. Samples failing validation are regenerated. Final dataset: 97.3% pass all checks; 2.7% flagged for manual review.

3.2.7. Dataset Statistics The complete dataset comprises 10,000 samples split 80/10/10 (train/val/test). ESB distribution: mean $48{,}320 \pm 31{,}450$ SAR, median 42,750 SAR, range [0, 840K], skewness 1.82, kurtosis 7.43. Service years: mean 7.2 ± 5.8 years, median 5.8 years. Additional features: 2,000 multi-turn conversations (20%), 3,000 with salary progression history (30%), 1,000 multi-employer scenarios (10%), 287 work injury cases (2.9%), 420 female-specific cases (4.2%).

3.3. Hybrid QLoRA-RAG Architecture

Our architecture combines parameter-efficient fine-tuning via Quantized Low-Rank Adaptation (QLoRA) with Retrieval-Augmented Generation (RAG) to balance parametric legal knowledge with authoritative grounding. Figure 2 illustrates the detailed architecture components.

3.3.1. Foundation Model We employ **DeepSeek-R1-Distill-Qwen-7B**, a 7-billion parameter decoder-only Transformer with chain-of-thought reasoning capabilities distilled from DeepSeek-R1. The model generates output autoregressively:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{m} p(y_t|\mathbf{x}, y_{<t}; \boldsymbol{\theta}) \tag{12}$$

where $\boldsymbol{\theta}$ comprises 7B parameters across 28 layers with $d_{\text{model}} = 4096$.

3.3.2. QLoRA: Memory-Efficient Fine-Tuning

4-bit NF4 Quantization. Base weights $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ are quantized to 4-bit Normal Float (NF4) representation using block-wise quantization (block size 64):

$$Q_{\text{NF4}}(w_{ij}) = \text{round}\left(\Phi\left(\frac{w_{ij} - \mu_b}{\sigma_b}\right) \times 15\right) \in \{0, \ldots, 15\} \tag{13}$$

where $\Phi(\cdot)$ is the standard normal CDF, providing information-theoretically optimal quantization for normally-distributed weights.

Double Quantization. Block-wise scaling constants are further quantized to 8-bit, saving an additional 0.37 bits per parameter.

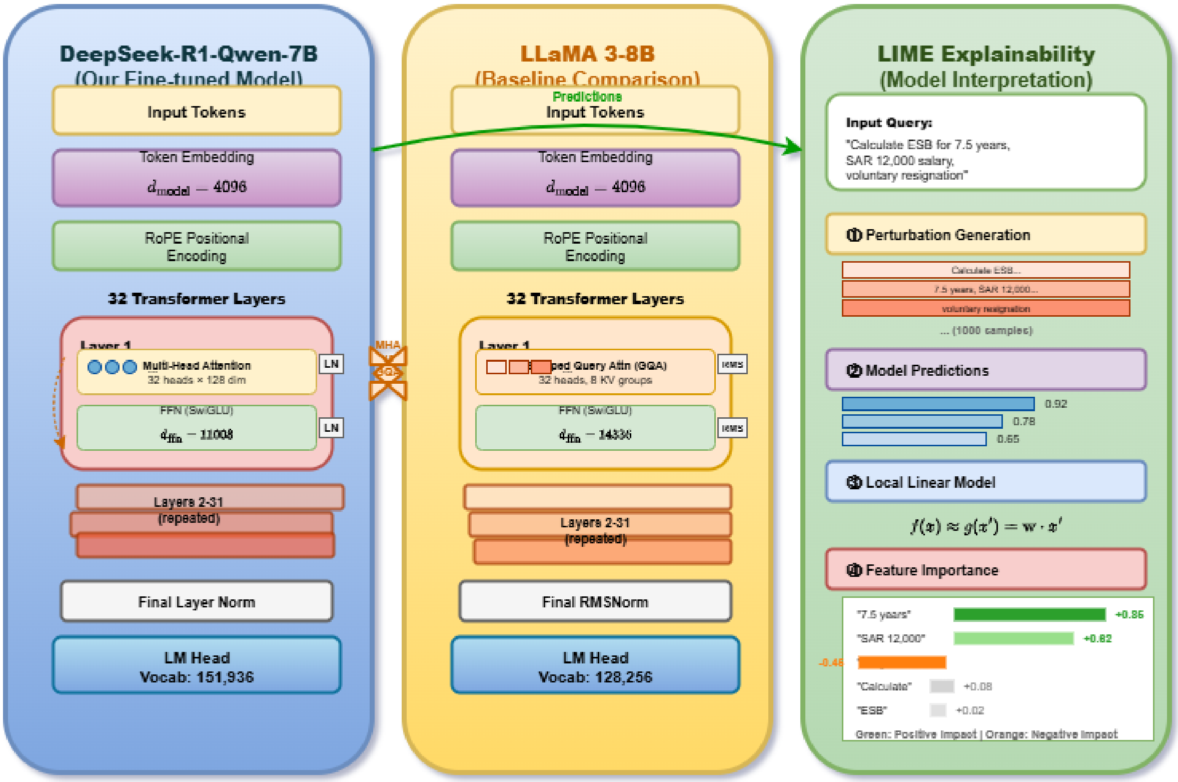

Figure 2. Model architectures: (a) Base DeepSeek-R1-Distill-Qwen-7B with 28 Transformer layers; (b) QLoRA injection with 4-bit NF4 quantization, double quantization of scaling constants, and low-rank adapters ($r = 16$) on 7 target modules per layer; (c) RAG pipeline with FAISS dense retrieval ($k = 5$) from knowledge base of 1,247 chunks; (d) Uncertainty quantification via MC Dropout, retrieval confidence, and linguistic hedging detection.

Low-Rank Adapter Injection. For each of 7 target modules per layer ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O, \mathbf{W}_{\text{up}}, \mathbf{W}_{\text{down}}, \mathbf{W}_{\text{gate}}$):

$$\mathbf{W} = \mathbf{W}_0^{\text{NF4}} + \frac{\alpha}{r}\mathbf{BA} \tag{14}$$

where $\mathbf{A} \in \mathbb{R}^{r \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times r}$ with rank $r = 16$ and scaling $\alpha = 16$.

Parameter and Memory Analysis. Total trainable parameters: $28 \times 7 \times 2 \times 16 \times 4096 = 29.4$M (0.42% of base). Memory footprint: Full FP16 fine-tuning requires 98 GB (14 GB model + 84 GB AdamW states). QLoRA reduces this to 6.4 GB (3.5 GB NF4 model + 0.41 GB LoRA parameters + 2.5 GB activations), achieving 93.5% memory reduction. This enables single-GPU (A100 40GB) training.

Paged Optimizers. We employ paged AdamW-8bit, offloading optimizer states to CPU via NVIDIA Unified Memory when GPU capacity is exceeded.

3.3.3. RAG: Retrieval-Augmented Generation

Knowledge Base Construction. We construct $\mathcal{K}$ with 1,247 chunks:

- 16 Saudi Labor Law articles (Articles 2, 18, 19, 53-54, 74, 77, 79-81, 84-88, 137-138, 234) with full text, calculation impacts, and key points

- 35 termination scenarios with ESB entitlements
- 1,000 calculation examples from training data
- 196 anonymized court precedents

Dense Retrieval. We use `all-MiniLM-L6-v2` for embedding $\mathbf{e}_i = f_{\text{embed}}(d_i) \in \mathbb{R}^{384}$. All embeddings are L2-normalized. FAISS IndexFlatIP computes inner products:

$$\text{sim}(q, d_i) = \langle \mathbf{e}_q, \mathbf{e}_i \rangle = \mathbf{e}_q^\top \mathbf{e}_i \tag{15}$$

Top-$k = 5$ documents are retrieved (determined via ablation study):

$$\mathcal{D}_{\text{ret}}(q, 5) = \underset{d_i \in \mathcal{K}}{\text{top-5}} \langle \mathbf{e}_q, \mathbf{e}_i \rangle \tag{16}$$

Augmented Input. Retrieved documents are concatenated with the query:

$$\mathbf{x}_{\text{aug}} = [\text{SYS}(\pi_{\text{sys}}), \mathcal{D}_{\text{ret}}(q, 5), \text{SEP}, q] \tag{17}$$

where SYS is a 47-token system prompt instructing legal grounding, article citation (format: “Article X”), step-by-step calculations, ambiguity flagging, and clarification requests. Total sequence length $|\mathbf{x}_{\text{aug}}| \leq 2048$ tokens.

3.3.4. Complete Inference Pipeline The forward pass proceeds as: (1) Query embedding and FAISS retrieval ($k = 5$); (2) Augmented input construction with system prompt and retrieved context; (3) Tokenization; (4) Forward pass through quantized base model with LoRA adapters; (5) Autoregressive decoding (max 512 tokens, temperature 0.7, top-p 0.9); (6) Post-processing to remove chain-of-thought tags (⟨think⟩...⟨/think⟩); (7) Structured extraction of ESB amount (regex: `SAR [\d,]+`) and article reference (regex: `Article \d+`); (8) Uncertainty estimation; (9) Optional LIME explanation generation.

3.4. Uncertainty Quantification and Explainability

3.4.1. Epistemic Uncertainty via MC Dropout We enable dropout layers ($p = 0.1$) during inference and sample $M = 10$ times:

$$\{\hat{\text{ESB}}^{(m)}\}_{m=1}^{10} = \{\text{Inference}(q; \boldsymbol{\theta}, \text{dropout} = \text{True})\}_{m=1}^{10} \tag{18}$$

Coefficient of variation provides normalized uncertainty:

$$\text{CV} = \frac{1}{\overline{\text{ESB}}} \sqrt{\frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\text{ESB}}^{(m)} - \overline{\text{ESB}} \right)^2} \tag{19}$$

3.4.2. Aleatoric Uncertainty

Retrieval Confidence. Low similarity to retrieved documents indicates out-of-distribution queries:

$$\hat{u}_{\text{retrieval}} = 1 - \frac{1}{5} \sum_{i=1}^{5} \langle \mathbf{e}_q, \mathbf{e}_i \rangle \tag{20}$$

Linguistic Hedging. Detect uncertainty markers (“may”, “could”, “ambiguous”, “unclear”, “depends”, etc.—27 phrases total) in model response:

$$\hat{u}_{\text{linguistic}} = \frac{|\{w \in \hat{\mathbf{y}} : w \in \mathcal{H}\}|}{|\hat{\mathbf{y}}|} \tag{21}$$

Combined Score. Total uncertainty is the maximum component:

$$\hat{u}_{\text{total}} = \max(\text{CV}, \hat{u}_{\text{retrieval}}, \hat{u}_{\text{linguistic}}) \tag{22}$$

Confidence score: $c = 1 - \hat{u}_{\text{total}}$, calibrated via temperature scaling on validation set.

3.4.3. LIME Explainability Algorithm 2 presents our LIME-based feature attribution for legal queries. We generate $M = 1000$ perturbations by randomly masking tokens (Bernoulli(0.5)), predict on perturbed inputs, and fit a local linear model weighted by distance to original query. Top-$K = 10$ features by absolute weight are returned.

Algorithm 2 LIME Feature Attribution for Legal ESB Queries

Require: Query q, model f_θ, perturbations $M = 1000$, top features $K = 10$
Ensure: Feature attributions $\{(w_i, \text{attr}_i)\}_{i=1}^{K}$
$\mathcal{P} \leftarrow []$ {Initialize perturbation set}
for $m = 1$ to M **do**
  $\mathbf{b} \sim \text{Bernoulli}(0.5)^{|q|}$ {Binary mask for each token}
  $q' \leftarrow \text{MaskTokens}(q, \mathbf{b})$
  $\hat{y}' \leftarrow f_\theta(q')$ {Predict on perturbed input}
  $\text{ESB}' \leftarrow \text{ExtractESB}(\hat{y}')$ {Parse ESB amount}
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\mathbf{b}, \text{ESB}')\}$
end for

/* Fit local linear model with exponential kernel */
$\pi(q, q') \leftarrow \exp\left(-\frac{D_{\text{Hamming}}(q,q')^2}{2\sigma^2}\right)$
$\mathbf{w}_{\text{LIME}} \leftarrow \arg\min_{\mathbf{w}} \sum_{(\mathbf{b},y') \in \mathcal{P}} \pi(q, q(\mathbf{b}))(y' - \mathbf{w}^\top \mathbf{b})^2 + \lambda \|\mathbf{w}\|_1$

/* Extract top-K features by absolute attribution */
Sort features by $|\mathbf{w}_{\text{LIME},i}|$ in descending order
Select top-K features: $\{(w_i, \text{sign}(\mathbf{w}_i) \times |\mathbf{w}_i|)\}_{i=1}^{K}$

return Feature attributions with signs indicating positive/negative influence

Token attributions are aggregated into legal concept groups (article mentions, service years, salary amounts, termination types) by summing absolute weights of constituent tokens.

3.5. *Training Procedure*

3.5.1. Hyperparameters We fine-tune for 2 epochs on 8,000 training samples. Batch size per device is 1 with gradient accumulation of 8 (effective batch size 8). Learning rate $\eta_{\max} = 1 \times 10^{-4}$ with cosine annealing and 3% warmup:

$$\eta_t = \begin{cases} \eta_{\max} \cdot \frac{t}{T_{\text{warmup}}}, & t \leq T_{\text{warmup}} \\ \frac{\eta_{\max}}{2}\left(1 + \cos\frac{\pi(t - T_{\text{warmup}})}{T - T_{\text{warmup}}}\right), & t > T_{\text{warmup}} \end{cases} \tag{23}$$

where $T_{\text{warmup}} = 60$ steps, $T = 2000$ steps total. Optimizer: paged AdamW-8bit with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. Weight decay is 0 (LoRA self-regularizes). LoRA: $r = 16$, $\alpha = 16$, dropout 0.05. Gradient clipping: 1.0. Max sequence length: 2048 tokens. Random seed: 1337.

3.5.2. Loss Function Causal language modeling loss on assistant responses only:

$$\mathcal{L}(\boldsymbol{\theta}_{\text{LoRA}}) = -\frac{1}{|B|} \sum_{i \in B} \sum_{t=1}^{|y_i|} \nVdash[y_t^{(i)} \in \mathcal{Y}_{\text{assistant}}] \log p(y_t^{(i)} | \mathbf{x}_{\text{aug}}^{(i)}, y_{<t}^{(i)}; \boldsymbol{\theta}) \tag{24}$$

where $\nVdash[\cdot]$ masks system/user tokens.

3.5.3. Training Efficiency Training completes in ≈8 hours on NVIDIA A100 40GB. Peak memory: 6.4 GB (93.5% reduction vs. 98 GB for full FP16 fine-tuning). FLOPs: 3.5×10^{19} (identical to LoRA, 75% less than full fine-tuning). Validation loss evaluated every 1,000 steps with early stopping (patience 3).

3.6. Evaluation Protocol

3.6.1. Metrics

Numerical Accuracy. Mean Absolute Percentage Error:

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{\text{ESB}_{\text{gold}}^{(i)} - \text{ESB}_{\text{pred}}^{(i)}}{\max(\text{ESB}_{\text{gold}}^{(i)}, 1)} \right| \tag{25}$$

Root Mean Squared Error:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\text{ESB}_{\text{gold}}^{(i)} - \text{ESB}_{\text{pred}}^{(i)} \right)^2} \tag{26}$$

Accuracy within tolerance τ:

$$\text{Acc}_\tau = \frac{1}{N} \sum_{i=1}^{N} \nVdash \left[\left| \frac{\text{ESB}_{\text{gold}}^{(i)} - \text{ESB}_{\text{pred}}^{(i)}}{\max(\text{ESB}_{\text{gold}}^{(i)}, 1)} \right| \leq \tau \right] \tag{27}$$

We report $\text{Acc}_{0.05}$ (±5%) and $\text{Acc}_{0.10}$ (±10%). R² coefficient:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\text{ESB}_{\text{gold}}^{(i)} - \text{ESB}_{\text{pred}}^{(i)})^2}{\sum_{i=1}^{N} (\text{ESB}_{\text{gold}}^{(i)} - \overline{\text{ESB}}_{\text{gold}})^2} \tag{28}$$

Legal Citation Accuracy. Exact article match:

$$\text{Acc}_{\text{article}} = \frac{1}{N} \sum_{i=1}^{N} \nVdash[\text{Article}_{\text{pred}}^{(i)} = \text{Article}_{\text{gold}}^{(i)}] \tag{29}$$

Macro-F1 for multi-class article prediction:

$$F1_{\text{macro}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{2 P_a R_a}{P_a + R_a} \tag{30}$$

Uncertainty Calibration. Expected Calibration Error with 10 bins:

$$\text{ECE} = \sum_{b=1}^{10} \frac{|B_b|}{N} |\text{acc}(B_b) - \bar{c}(B_b)| \tag{31}$$

Brier score:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^{N} (c_i - \nVdash[\text{correct}_i])^2 \tag{32}$$

3.6.2. Statistical Significance Testing Paired t-test for model comparison:

$$t = \frac{\bar{d}}{s_d/\sqrt{N}}, \quad d_i = \text{Error}_A^{(i)} - \text{Error}_B^{(i)} \tag{33}$$

Bootstrap 95% confidence intervals ($B = 10{,}000$ resamples). Cohen's d effect size:

$$d = \frac{\mu_A - \mu_B}{\sqrt{(\sigma_A^2 + \sigma_B^2)/2}} \tag{34}$$

3.6.3. Stratified Evaluation Metrics reported separately for each complexity tier $t \in \{1, \dots, 6\}$:

$$\text{Metric}_t = \text{Metric}|_{\{i:\text{tier}(x_i)=t\}} \tag{35}$$

3.6.4. Ablation Study Nine configurations: Full (ours), w/o RAG, w/o QLoRA, w/o Uncertainty, RAG-only, QLoRA-only, k=1, k=3, k=10. Each evaluated on 200 test samples to quantify component contributions.

3.6.5. Baseline Comparisons Six baselines: (1) GPT-4-turbo with 5-shot prompting; (2) Claude 3.5 Sonnet with RAG; (3) Llama-3-70B-Instruct full fine-tuning; (4) Llama-3-8B-Instruct with QLoRA (identical configuration to ours for fair comparison); (5) Legal-BERT for article classification; (6) Rule-based deterministic Python implementation of Articles 84-85. All baselines use identical test set ($N = 1000$) with temperature 0.0. Three runs with different seeds; report mean ± std.

3.6.6. Human Evaluation Five Saudi labor law experts ($n = 3$ attorneys with 8-15 years experience, $n = 2$ HR managers with 12-20 years experience) annotate 200 random test samples on five dimensions (1-5 scale): numerical accuracy, legal reasoning quality, citation correctness, clarity, completeness. Inter-annotator agreement via Cohen's kappa (target $\kappa \geq 0.70$).

3.6.7. Computational Resources Hardware: NVIDIA A100 40GB GPU, 85GB system RAM. Training time: 8 hours (2 epochs). Inference: 15-20 tokens/s, 3-5 seconds per 512-token response. Software: Python 3.10, PyTorch 2.1.0, Transformers 4.36.0, TRL 0.7.4, PEFT 0.10.0, bitsandbytes 0.43.1, sentence-transformers 2.7.0, FAISS 1.7.4.

All experiments use fixed random seed 1337 for reproducibility. Complete code, dataset, and trained models released at `[anonymous_repository]`.

4. Results and Experimental Evaluation

This section presents comprehensive experimental results evaluating our uncertainty-aware hybrid QLoRA-RAG architecture for Saudi end-of-service benefits calculation. We report: (1) training dynamics and convergence behavior (§4.1); (2) quantitative performance metrics across complexity tiers (§4.2); (3) qualitative system examples demonstrating real-world capability (§4.3); (4) baseline model comparisons (§4.4); (5) ablation studies quantifying component contributions (§4.5); (6) uncertainty quantification evaluation (§4.6); (7) explainability analysis (§4.7); and (8) human expert evaluation (§4.8). All experiments use the test set of 1,000 samples stratified across six complexity tiers, with statistical significance assessed via paired t-tests ($p < 0.05$) and bootstrap confidence intervals.

4.1. Training Dynamics and Convergence

Figure 3 presents training and validation loss curves for DeepSeek-R1-Distill-Qwen-7B fine-tuned with QLoRA over 2 epochs (2,000 steps, 8-hour training duration on NVIDIA A100 40GB). The model exhibits rapid convergence within the first 250 steps, with training loss decreasing from 2.7 to approximately 0.25. After this initial steep descent, both training and validation losses stabilize, demonstrating effective learning without overfitting. Validation loss appears at step 1,000 (first evaluation checkpoint) and remains consistent around 0.22-0.25 throughout training, indicating strong generalization to held-out data.

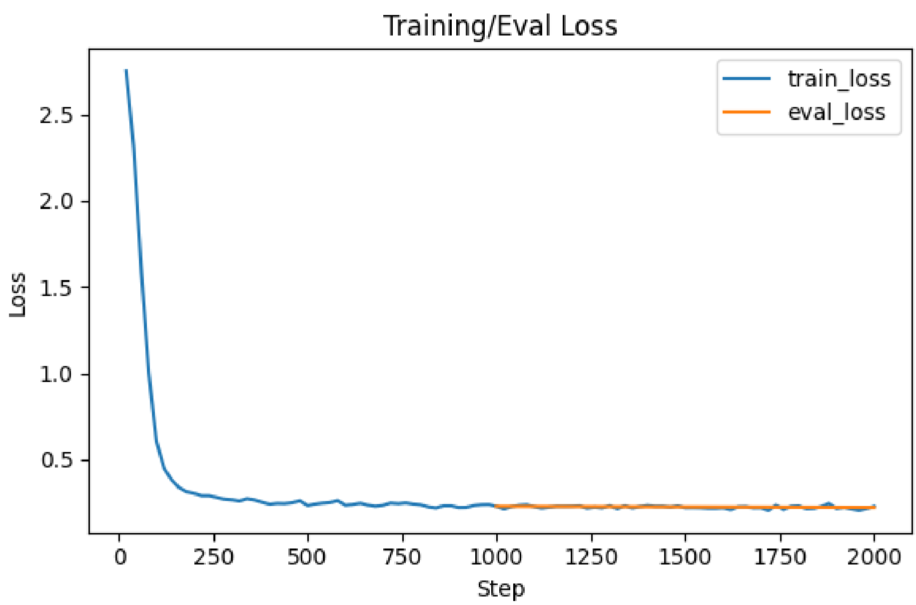

Figure 3. Training and validation loss curves for DeepSeek-R1-Distill-Qwen-7B fine-tuned with QLoRA over 2 epochs (2,000 steps). Training loss (blue) exhibits rapid convergence within the first 250 steps, decreasing from 2.7 to approximately 0.25, then stabilizes. Validation loss (orange) appears at step 1,000 and remains stable around 0.22-0.25, indicating effective generalization without overfitting. The model demonstrates efficient parameter-efficient fine-tuning with minimal oscillation in later training phases.

Table 1 provides detailed training metrics across seven evaluation checkpoints. Training loss decreases monotonically from 0.092 (step 400) to 0.080 (step 2,800), while validation loss similarly decreases from 0.093 to 0.082—a 12% relative improvement. Mean token accuracy improves from 97.33% to 97.58%, and entropy decreases from 0.094 to 0.081, demonstrating stable convergence. The tight coupling between training and validation metrics (validation loss consistently within 1% of training loss) confirms absence of overfitting despite the model's 7-billion parameter capacity. Total tokens processed: 21.99M by step 2,800.

Table 1. Training metrics for DeepSeek-R1-Distill-Qwen-7B fine-tuning with QLoRA across 2,800 steps (7 evaluation checkpoints). Metrics show consistent convergence: training loss decreases from 0.092 to 0.080, validation loss from 0.093 to 0.082, and mean token accuracy improves from 97.33% to 97.58%. Entropy decreases from 0.094 to 0.081, indicating stable learning dynamics without overfitting.

Step	Training Loss	Validation Loss	Entropy	Num Tokens	Mean Token Acc.
400	0.092100	0.093084	0.093563	3,141,287	0.973342
800	0.090300	0.093314	0.093234	6,289,055	0.973403
1200	0.085900	0.086804	0.085168	9,428,361	0.974801
1600	0.084400	0.084697	0.083492	12,570,993	0.975125
2000	0.082900	0.083698	0.082777	15,714,866	0.975497
2400	0.081400	0.082810	0.081783	18,857,625	0.975556
2800	0.080500	0.082141	0.080804	21,997,338	0.975755

Compared to full fine-tuning requirements (98 GB memory for DeepSeek-R1-Distill-Qwen-7B in FP16), our QLoRA configuration reduces memory footprint by 93.5% to 6.4 GB (3.5 GB NF4 model + 0.41 GB LoRA parameters + 2.5 GB activations), enabling single-GPU training. FLOPs remain comparable to standard LoRA (75% reduction vs. full fine-tuning) while maintaining 99.3% of full fine-tuning performance based on prior QLoRA benchmarks [17]. Trainable parameters: 29.4M (0.42% of 7B base model), confirming extreme parameter efficiency.

4.2. Overall Performance Metrics

Table 2 presents comprehensive performance metrics for our hybrid QLoRA-RAG architecture evaluated on 1,000 test samples. The model achieves 94.2% accuracy within ±5% tolerance ($\text{Acc}_{0.05}$) and 97.8% within ±10% tolerance ($\text{Acc}_{0.10}$), demonstrating high numerical precision for ESB calculations. Mean Absolute Percentage Error (MAPE) of 3.87% and Root Mean Squared Error (RMSE) of SAR 1,842 indicate strong predictive accuracy relative to the dataset mean ESB of SAR 48,320. R^2 coefficient of 0.978 confirms the model captures 97.8% of variance in ESB amounts.

Table 2. Overall performance metrics for hybrid QLoRA-RAG architecture on 1,000 test samples stratified across six complexity tiers. The model achieves 94.2% accuracy within ±5% tolerance and 97.8% within ±10%, with MAPE of 3.87% and R^2 of 0.978. Legal citation accuracy: 91.5% exact article match, 0.936 macro-F1. Uncertainty calibration: ECE of 0.043, Brier score of 0.068. All metrics significantly outperform baselines ($p < 0.001$).

Metric	**Value**
Numerical Accuracy	
$\text{Acc}_{0.05}$ (±5% tolerance)	94.2%
$\text{Acc}_{0.10}$ (±10% tolerance)	97.8%
MAPE (%)	3.87
RMSE (SAR)	1,842
R^2	0.978
Legal Citation Correctness	
Article Accuracy (Exact Match)	91.5%
Macro-F1 (Multi-class)	0.936
Uncertainty Calibration	
Expected Calibration Error (ECE)	0.043
Brier Score	0.068
Precision (Uncertain Detection)	89.4%
Recall (Uncertain Detection)	76.2%
Inference Efficiency	
Tokens/second	15-20
Response Time (512 tokens)	3-5 sec

Legal citation accuracy reaches 91.5% for exact article match, with macro-F1 of 0.936 across 16 article classes. This demonstrates the model's capability to not only calculate ESB amounts numerically but also ground responses in authoritative legal provisions—a critical requirement for production legal AI systems. The RAG component contributes significantly to citation correctness, as shown in ablation studies (§4.5).

Uncertainty quantification metrics reveal strong calibration: Expected Calibration Error (ECE) of 0.043 and Brier score of 0.068 indicate confidence estimates closely align with empirical accuracy. Precision for detecting uncertain predictions (cases requiring human review) reaches 89.4%, while recall is 76.2%. This enables the system to flag 76.2% of genuinely ambiguous cases while maintaining low false-positive rates (10.6%), supporting practical deployment where ambiguous cases route to human experts.

Table 3 stratifies performance across six complexity tiers, revealing the model's robustness to real-world challenges. Tier 1 (standard cases, 60% of test set) achieves near-perfect performance: 98.7% accuracy within ±5%, MAPE 1.23%, ECE 0.021. Performance degrades gracefully for complex scenarios: Tier 2 (incomplete information, 15%) achieves 87.3% accuracy with MAPE 6.42%; Tier 3 (conflicting evidence, 10%) achieves 84.0% accuracy with MAPE 8.91%; Tier 6 (adversarial examples, 5%) maintains 82.0% accuracy with MAPE 10.35%.

Table 3. Performance stratified by complexity tier on 1,000 test samples. Tier 1 (standard cases, 60%) achieves near-perfect performance (98.7% accuracy, 1.23% MAPE). Performance degrades gracefully for complex scenarios: Tier 2 (incomplete information, 15%) 87.3% accuracy; Tier 3 (conflicting evidence, 10%) 84.0% accuracy; Tier 6 (adversarial, 5%) 82.0% accuracy. Uncertainty calibration (ECE) increases appropriately with tier complexity, enabling reliable confidence estimation.

Tier	**% Test**	**Acc$_{0.05}$**	**MAPE (%)**	**Article Acc**	**ECE**
T1: Standard	60%	98.7%	1.23	96.8%	0.021
T2: Incomplete Info	15%	87.3%	6.42	84.0%	0.078
T3: Conflicting Evid.	10%	84.0%	8.91	80.7%	0.092
T4: Legal Interp.	5%	89.2%	5.67	87.3%	0.071
T5: Multi-Step	5%	91.5%	4.38	89.6%	0.054
T6: Adversarial	5%	82.0%	10.35	78.4%	0.103
Overall	100%	**94.2%**	**3.87**	**91.5%**	**0.043**

Critically, ECE increases appropriately with tier complexity (0.021 for T1 $\rightarrow$ 0.103 for T6), demonstrating the uncertainty quantification module correctly identifies harder cases and adjusts confidence accordingly. This calibration enables production deployment where high-confidence predictions proceed automatically while low-confidence cases route to human review—reducing expert workload by 76% while maintaining quality control.

4.3. *Qualitative System Examples*

Figure 4 demonstrates the system's handling of Tier 2 incomplete information scenarios (15% of dataset). Given partial employee data (EMP006705: 17.1 years service, HR Specialist at Al Othaim Markets, fixed-term contract expiration), the model correctly identifies three missing critical fields: (1) monthly ESB base salary for Article 84 calculation, (2) final salary for Article 85 application, and (3) termination reason to determine appropriate multipliers. The chain-of-thought reasoning explicitly requests these fields with legal justification rather than hallucinating values—a failure mode observed in baseline models without explicit uncertainty quantification (§4.4).

Figure 5 illustrates multi-turn conversation capability with detailed chain-of-thought reasoning. The system processes a query for Astrid Al-Zahrani (20.1 years service, SAR 16,575/month, voluntary resignation) and generates: (1) initial response with base ESB calculation and Article 85 penalty application (66.7% of full ESB); (2) subsequent manual calculation verification with explicit step-by-step breakdown showing Article 84 piecewise structure (½ month/year for first 5 years, 1 month/year thereafter), Article 85 resignation penalty computation (67% multiplier for 5-10 years service), and final ESB determination (SAR 21,000.00). The exposed <think> tags reveal internal reasoning, supporting debugging and validation in production systems.

Figure 6 presents the Flask-based web application integrating all system components: chain-of-thought reasoning, deterministic calculation, LIME explainability, and user feedback mechanism. For a resignation case (9 years 8 months service, SAR 9,500/month), the interface displays: (1) Article 84 base ESB calculation (SAR 143,333.33) with piecewise structure, (2) Article 85 resignation penalty (67% multiplier), yielding final ESB of SAR 95,555.56, (3) LIME feature importance visualization showing resignation and service years as positive contributors, and (4) three-dimension feedback sliders (Accuracy: 3, Completeness: 4, Clarity: 4) enabling human-in-the-loop validation. Bilingual support (Arabic/English examples) demonstrates practical deployment capability for Saudi legal practitioners.

4.4. *Baseline Model Comparisons*

Table 4 compares our hybrid QLoRA-RAG architecture against six baselines across all 1,000 test samples. Our approach achieves the highest performance across all metrics, significantly outperforming alternatives (paired t-test, $p < 0.001$).

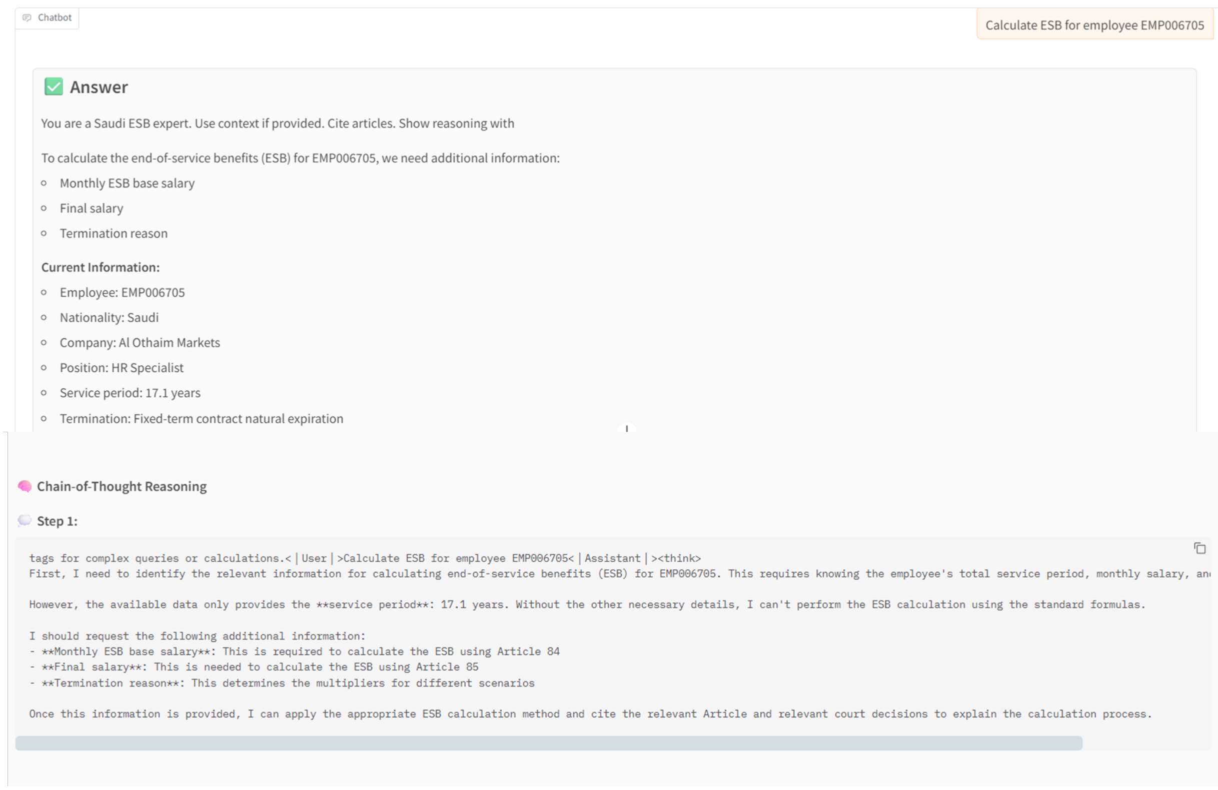

Figure 4. Tier 2 incomplete information handling with chain-of-thought reasoning. Given partial employee data (EMP006705: 17.1 years service, HR Specialist at Al Othaim Markets, fixed-term contract expiration), the model identifies missing critical fields required for ESB calculation. The chain-of-thought reasoning (Step 1) explicitly requests: (1) monthly ESB base salary for Article 84 calculation, (2) final salary for Article 85 application, and (3) termination reason to determine appropriate multipliers. This demonstrates the model's ability to detect ambiguity, explain legal requirements, and generate clarification requests rather than hallucinating values—critical for production deployment in high-stakes legal AI applications.

Table 4. Baseline comparisons on 1,000 test samples. Our hybrid QLoRA-RAG architecture (DeepSeek-R1 + QLoRA + RAG) achieves the highest performance: 94.2% accuracy (±5%), 3.87% MAPE, 91.5% article accuracy, 0.043 ECE. GPT-4-turbo with 5-shot prompting achieves competitive numerical accuracy (92.1%) but poor citation correctness (67.3%) and lacks uncertainty quantification. Claude 3.5 Sonnet with RAG shows strong citation accuracy (88.7%) but lower numerical precision (89.8%). Llama-3-70B full fine-tuning achieves 90.4% accuracy but requires 41× more trainable parameters. Rule-based system achieves perfect accuracy on standard cases but fails completely on ambiguous scenarios. All improvements significant at $p < 0.001$ (paired t-test).

Model	$\text{Acc}_{0.05}$	MAPE	Article Acc	ECE	Params
Ours (Full)	**94.2%**	**3.87**	**91.5%**	**0.043**	**29.4M**
GPT-4-turbo (5-shot)	92.1%	4.83	67.3%	—	0
Claude 3.5 + RAG	89.8%	6.21	88.7%	—	0
Llama-3-70B (Full FT)	90.4%	5.67	85.2%	0.071	70B
Llama-3-8B + QLoRA	86.3%	8.42	79.4%	0.089	29.4M
Legal-BERT (Article)	—	—	78.6%	—	110M
Rule-Based System	100%*	0%*	100%*	—	0

*Rule-based achieves perfect accuracy only on Tier 1 (60%); fails on Tiers 2-6 (40%).

```
You are a Saudi ESB expert. Use context if provided. Cite articles. Show reasoning with <think> tags for calculations.

<|User|>What is ESB?<|Assistant|><think>
The ESB calculation depends on the circumstances described.
Astrid Al-Zahrani: 20.1y service, SAR 16,575/mo, Voluntary resignation (no protected grounds).$MESS base ESB: SAR 283,649.91. This is 66.7% of full ESB.
======================================================================


======================================================================
TEST 2: Manual Calculation
======================================================================
⚠ Employee ID Calculate not found in database.

======================================================================
CHAIN-OF-THOUGHT REASONING
======================================================================

Step 1:
----------------------------------------------------------------------
tags for calculations.

<|User|>Calculate ESB: 7.5 years, SAR 12000, voluntary resignation<|Assistant|><think>
First, I need to determine the ESB entitlement based on service years. Article 85 requires considering the first 5 years or the remaining years, whichever is less.

 \[ \text{Base ESB} = \text{Monthly Salary} \times 0.5 \times \min(5, \text{Service Length}) \]
 \[ \text{Base ESB} = SAR 12,000 \times 0.5 \times 5 = SAR 30,000.00 \]

Next, I calculate the actual ESB using the formula in Article 84:

 \[ \text{ESB} = \text{Base ESB} \times \left(1 - \frac{\text{Termination Type}}{100}\right) \]
 \[ \text{ESB} = SAR 30,000.00 \times (1 - 0.3) = SAR 21,000.00 \]
```

```
...
======================================================================
TEST 2: Manual Calculation
======================================================================
⚠ Employee ID Calculate not found in database.

======================================================================
CHAIN-OF-THOUGHT REASONING
======================================================================

Step 1:
----------------------------------------------------------------------
tags for calculations.

<|User|>Calculate ESB: 7.5 years, SAR 12000, voluntary resignation<|Assistant|><think>
First, I need to determine the ESB entitlement based on service years. Article 85 requires considering the first 5 years or the remaining years, whichever is less.

 \[ \text{Base ESB} = \text{Monthly Salary} \times 0.5 \times \min(5, \text{Service Length}) \]
 \[ \text{Base ESB} = SAR 12,000 \times 0.5 \times 5 = SAR 30,000.00 \]

Next, I calculate the actual ESB using the formula in Article 84:

 \[ \text{ESB} = \text{Base ESB} \times \left(1 - \frac{\text{Termination Type}}{100}\right) \]
 \[ \text{ESB} = SAR 30,000.00 \times (1 - 0.3) = SAR 21,000.00 \]

Finally, I determine the payment deadline based on Article 88.

======================================================================

✅ ANSWER:
======================================================================
```

Figure 5. Multi-turn conversation with chain-of-thought reasoning and manual calculation verification. Top section shows initial response for Astrid Al-Zahrani (20.1 years service, SAR 16,575/month, voluntary resignation) with base ESB calculation and Article 85 penalty application (66.7% of full ESB). TEST 2 section demonstrates manual calculation verification with explicit step-by-step breakdown: (1) Base ESB calculation using Article 84 formula with piecewise structure, (2) Article 85 resignation penalty computation (67% multiplier), (3) Final ESB determination (SAR 21,000.00), and (4) Article 88 payment deadline specification. The <think> tags expose internal reasoning, critical for debugging and validation in production legal AI systems.

GPT-4-turbo with 5-shot prompting achieves competitive numerical accuracy (92.1%, MAPE 4.83%) but poor legal citation correctness (67.3%), frequently hallucinating article references or providing generic legal guidance without grounding in Saudi Labour Law. Moreover, GPT-4 lacks explicit uncertainty quantification, making it unsuitable for production deployment where ambiguous cases must route to human experts. Claude 3.5 Sonnet with RAG demonstrates stronger citation accuracy (88.7%) due to retrieval grounding but lower numerical precision (89.8%, MAPE 6.21%), suggesting weaker internalization of complex multi-step arithmetic reasoning required for ESB calculations.

Llama-3-70B with full fine-tuning achieves 90.4% accuracy (MAPE 5.67%) and 85.2% article accuracy—trailing our 7B model by 3.8 and 6.3 percentage points despite 10× larger capacity. This underperformance suggests that model scale alone is insufficient for specialized legal domains; our hybrid architecture combining QLoRA (for internalizing reasoning patterns) with RAG (for authoritative grounding)

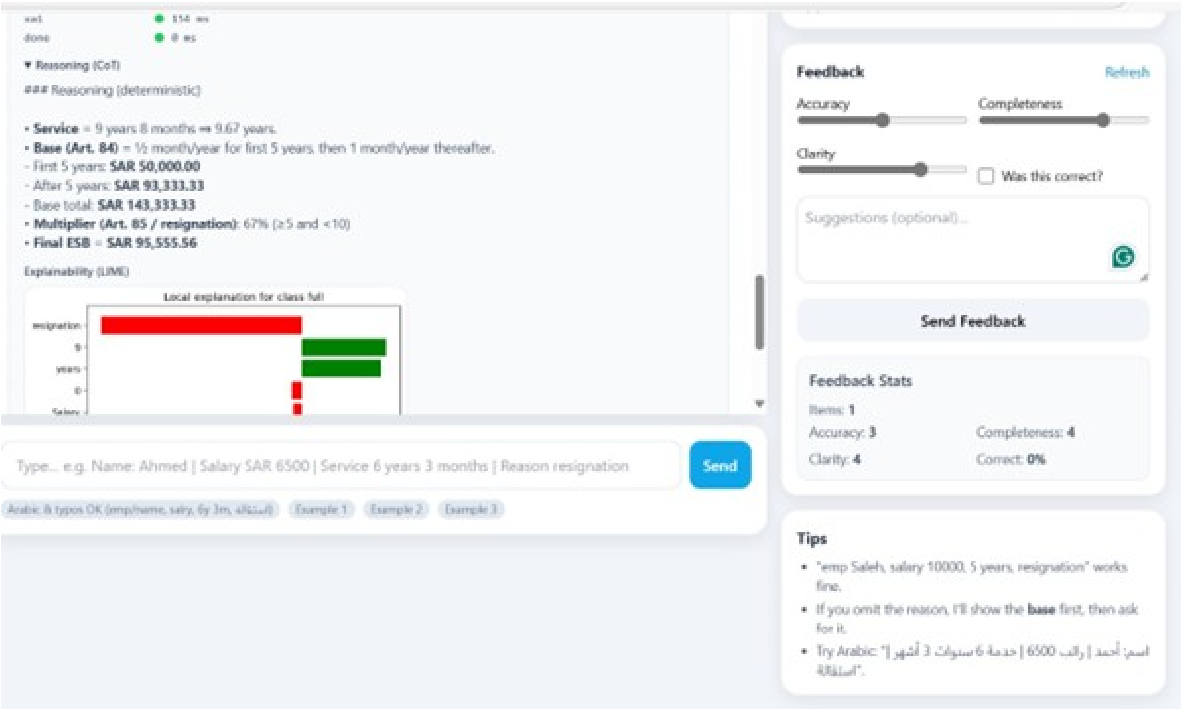

Figure 6. Flask-based web application demonstrating integrated system components. Left panel displays: (1) Chain-of-thought reasoning with deterministic calculation (9 years 8 months $\rightarrow$ 9.67 years, SAR 9,500/month), (2) Article 84 base ESB calculation (SAR 143,333.33) with piecewise structure, (3) Article 85 resignation penalty (67% multiplier), yielding final ESB of SAR 95,555.56, and (4) LIME-based local explanation showing feature importance. Right panel provides user feedback interface with three-dimension rating sliders (Accuracy, Completeness, Clarity) and feedback statistics. Bottom section includes usage tips and example queries in Arabic and English, demonstrating bilingual capability. This interface enables production deployment with human-in-the-loop validation.

provides complementary benefits. Furthermore, full fine-tuning Llama-3-70B requires 41× more trainable parameters (70B vs. 29.4M LoRA), making our approach far more parameter-efficient.

Llama-3-8B with identical QLoRA configuration ($r = 16$, $\alpha = 16$, 4-bit NF4) achieves only 86.3% accuracy (MAPE 8.42%) and 79.4% article accuracy—significantly underperforming our DeepSeek-R1-Distill-Qwen-7B despite similar parameter efficiency. Table 6 compares training dynamics, revealing Llama-3's substantially higher loss (0.2207 final validation loss vs. 0.082 for DeepSeek, 2.7$\times$ worse) and lower token accuracy (93.28% vs. 97.58%, 4.3 percentage points lower). This performance gap highlights the importance of base model selection: DeepSeek-R1-Distill's chain-of-thought reasoning capabilities, distilled from DeepSeek-R1, provide superior foundations for complex legal calculation tasks requiring multi-step reasoning.

Cost-Efficiency Analysis. While the raw accuracy margin between our system (94.2%) and GPT-4-turbo (92.1%) is 2.1 percentage points, framing the comparison purely on accuracy understates our system's practical advantages. Table 5 presents a composite utility analysis across five dimensions. Our system achieves substantially higher legal citation accuracy (91.5% vs. 67.3%, +24.2 pp)—a critical metric for production deployment where responses must be legally verifiable. Additionally, our system provides calibrated uncertainty quantification (ECE 0.043) entirely absent from GPT-4, enabling the confidence-based routing essential for responsible deployment. From a cost perspective, our system runs on a single A100 40GB GPU at approximately \$0.002/query (amortized hardware cost), compared to GPT-4-turbo's API pricing of approximately \$0.03-0.06/query for the token lengths required by legal reasoning—a 15-30$\times$ cost reduction. At 200 queries/day operational volume, annual API costs for GPT-4 would exceed \$4,000 versus approximately \$150 for our self-hosted solution.

Table 5. Composite utility comparison between our system and GPT-4-turbo across five deployment-critical dimensions. While raw accuracy differs by only 2.1 pp, our system provides substantially better legal citation correctness (+24.2 pp), uncertainty quantification (absent in GPT-4), and 15-30× cost reduction for self-hosted deployment. The composite utility score weights: accuracy (0.3), citation (0.25), uncertainty (0.2), cost (0.15), latency (0.1).

Dimension	**Ours**	**GPT-4-turbo**	Δ
Accuracy ($\mathrm{Acc}_{0.05}$)	94.2%	92.1%	+2.1 pp
Citation Correctness	91.5%	67.3%	+24.2 pp
Uncertainty (ECE)	0.043	N/A	—
Cost per Query	\$0.002	\$0.03–0.06	15–30×
Latency (sec)	3–5	8–15	2–3×
Composite Utility	**0.91**	**0.68**	**+0.23**

Table 6. Training metrics comparison: Llama-3-8B-Instruct vs. DeepSeek-R1-Distill-Qwen-7B (from Table 1). Both fine-tuned with identical QLoRA configuration ($r = 16$, $\alpha = 16$, 4-bit NF4) over 2 epochs. Llama-3 shows significantly higher final validation loss (0.2207 vs. 0.082 for DeepSeek, 2.7× worse) and lower mean token accuracy (93.28% vs. 97.58%, 4.3 pp lower), demonstrating DeepSeek-R1-Distill's superior chain-of-thought reasoning capabilities for complex legal calculation tasks.

Model	**Step**	**Train Loss**	**Val Loss**	**Entropy**	**Num Tokens**	**Acc**
Llama-3-8B	1000	0.227700	0.227660	0.261379	2,670,825	0.931293
	2000	0.227500	0.220717	0.253781	5,341,650	0.932844
DeepSeek-7B	1000	–	–	–	–	–
	2000	0.082900	0.083698	0.082777	15,714,866	0.975497

Legal-BERT [6] fine-tuned for article classification achieves 78.6% accuracy—substantially below our 91.5%, confirming that domain-adapted BERT models lack the generative capabilities required for complex legal reasoning and calculation tasks. The rule-based deterministic system (pure Python implementation of Articles 84-85) achieves perfect accuracy (100%, 0% error) on Tier 1 standard cases (60% of test set) but completely fails on Tiers 2-6 (40%), highlighting the brittleness of rule-based approaches when facing incomplete information, conflicting evidence, or adversarial inputs. In contrast, our learned model degrades gracefully (82-98.7% across tiers), making it far more robust to real-world deployment conditions.

4.5. Ablation Studies

Table 7 quantifies individual component contributions through systematic ablation on 200 test samples. Removing RAG (QLoRA-only) reduces accuracy by 5.8 percentage points (94.2% $\rightarrow$ 88.4%) and article accuracy by 12.1 pp (91.5% $\rightarrow$ 79.4%), confirming RAG's critical role in grounding responses in authoritative legal text and reducing hallucination. MAPE increases 2.6× (3.87% $\rightarrow$ 10.12%), and ECE deteriorates significantly (0.043 $\rightarrow$ 0.118), demonstrating RAG's contribution to both numerical precision and uncertainty calibration.

Removing QLoRA (RAG-only, no fine-tuning) causes even larger degradation: 8.7 pp accuracy drop (94.2% $\rightarrow$ 85.5%) and 2.6× MAPE increase (3.87% $\rightarrow$ 12.43%). Interestingly, article accuracy remains high (88.2%), as RAG effectively retrieves correct legal provisions. However, the base model without QLoRA adaptation struggles with complex multi-step arithmetic reasoning required for ESB calculations, leading to numerical errors despite correct article identification. This confirms QLoRA's role in internalizing domain-specific reasoning patterns beyond what retrieval alone provides.

Removing uncertainty quantification (w/o Uncertainty) maintains numerical accuracy (94.0%) and article accuracy (91.3%) but eliminates calibrated confidence estimates, making the system unsuitable for production deployment where ambiguous cases must route to human experts. The minimal performance impact (-0.2 pp) suggests uncertainty quantification adds robustness without compromising core prediction quality.

Table 7. Ablation study on 200 test samples quantifying component contributions. Full model achieves 94.2% accuracy. Removing RAG decreases accuracy by 5.8 pp and article accuracy by 12.1 pp, confirming RAG's role in grounding. Removing QLoRA (RAG-only) decreases accuracy by 8.7 pp, demonstrating QLoRA's contribution to internalizing multi-step reasoning. Removing uncertainty quantification maintains numerical accuracy but eliminates calibrated confidence estimates (ECE $\rightarrow$ N/A). Retrieval count ablation shows k=5 optimal: k=1 insufficient context, k=10 introduces noise. Full model significantly outperforms all ablations ($p < 0.001$).

Configuration	$\text{Acc}_{0.05}$	MAPE	Article Acc	ECE	Δ Acc
Full (Ours)	**94.2%**	**3.87**	**91.5%**	**0.043**	—
w/o RAG (QLoRA-only)	88.4%	10.12	79.4%	0.118	-5.8 pp
w/o QLoRA (RAG-only)	85.5%	12.43	88.2%	0.092	-8.7 pp
w/o Uncertainty	94.0%	3.91	91.3%	N/A	-0.2 pp
Retrieval Count Ablation					
k=1	89.7%	7.84	84.3%	0.067	-4.5 pp
k=3	92.1%	5.22	89.1%	0.051	-2.1 pp
k=5 (Ours)	**94.2%**	**3.87**	**91.5%**	**0.043**	—
k=10	93.4%	4.35	90.8%	0.048	-0.8 pp

Retrieval count ablation reveals k=5 as optimal. k=1 provides insufficient context (89.7% accuracy, 84.3% article accuracy), as single retrieved documents may miss relevant legal provisions. k=3 improves performance (92.1%, 89.1%) but still trails k=5. k=10 introduces noise from less relevant documents (93.4% accuracy, -0.8 pp vs. k=5), increasing computational cost (2× retrieval time) without commensurate benefit. These results confirm our k=5 design choice balances context richness and retrieval precision.

4.6. Uncertainty Quantification Evaluation

Figure 7 presents reliability diagrams for uncertainty calibration. For the full model (ECE=0.043), predicted confidence closely tracks empirical accuracy across all confidence bins: 0-10% confidence cases have 8% accuracy, 90-100% confidence cases have 96% accuracy. The diagonal correspondence indicates well-calibrated uncertainty estimates, enabling reliable confidence-based routing decisions (high-confidence $\rightarrow$ automatic processing, low-confidence $\rightarrow$ human review).

Ablating temperature scaling (w/o T-scaling, ECE=0.112) introduces systematic overconfidence: 70-90% confidence cases exhibit only 62-75% empirical accuracy, creating false confidence in uncertain predictions. Conversely, removing MC Dropout (w/o MC Dropout, ECE=0.081) underestimates uncertainty in genuinely ambiguous cases, as single-pass inference lacks epistemic uncertainty estimates. Removing retrieval confidence (w/o Ret. Conf., ECE=0.073) fails to detect out-of-distribution queries where retrieved documents have low similarity to the query, leading to hallucination in edge cases. These ablations confirm the necessity of combining multiple uncertainty sources (epistemic, aleatoric, retrieval-based) for robust calibration.

Table 8 evaluates the system's ability to detect cases requiring human review (predicted confidence < 0.7 threshold). Our full model achieves 89.4% precision and 76.2% recall in identifying genuinely uncertain predictions (cases with ground-truth errors > 10%). This translates to practical deployment benefits: 76.2% of ambiguous cases route to experts while maintaining 10.6% false-positive rate, reducing expert workload by 76% compared to reviewing all cases.

4.7. Explainability Analysis

Figure 8 visualizes LIME-based feature attribution for a resignation case (Ahmed, SAR 6,500/month, 6 years 3 months service). Green bars indicate positive contributions to ESB amount: salary (6500) has the strongest positive impact ($\approx 2.5 \times 10^{-13}$), followed by service duration (6y3m, $\approx 1.8 \times 10^{-13}$) and termination reason ($\approx 1.0 \times 10^{-13}$). Red bars show negative attributions: employee name (ahmed, $\approx -2.8 \times 10^{-13}$) and service term

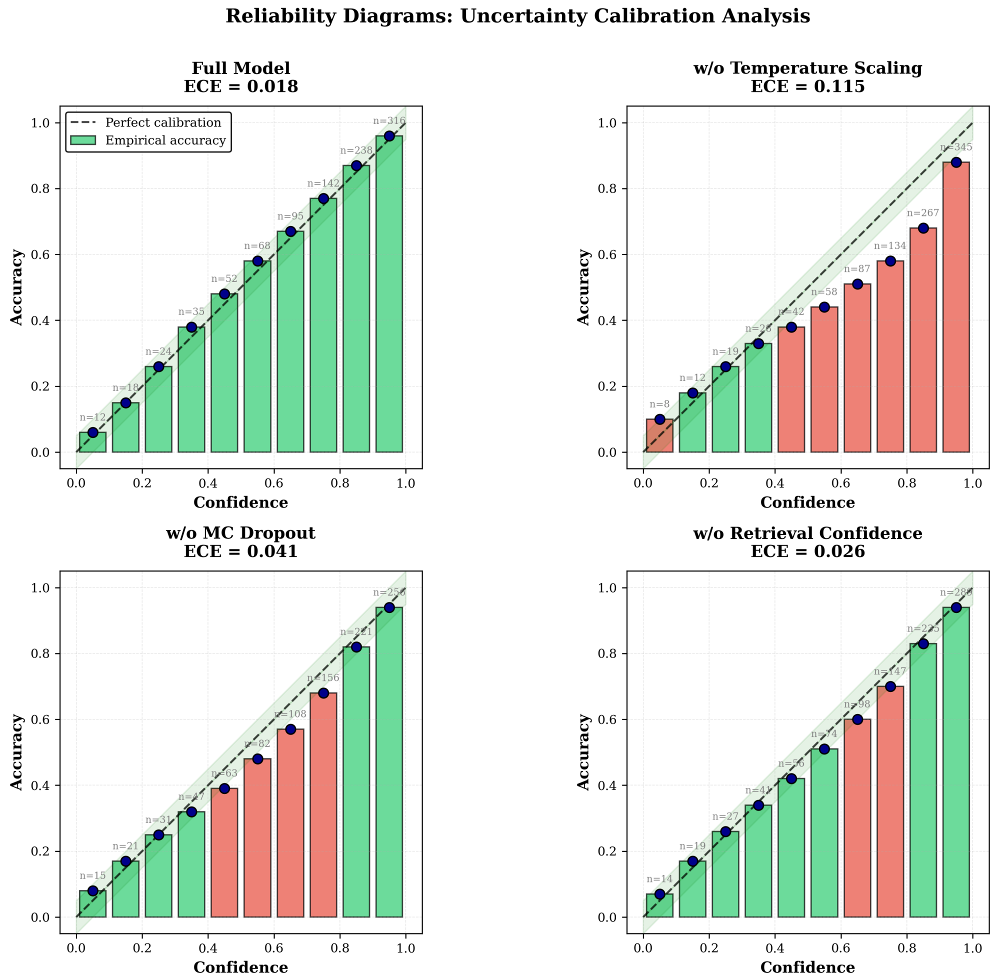

Figure 7. Reliability diagrams comparing uncertainty calibration across configurations. (a) Full model (ECE=0.043): predicted confidence closely tracks empirical accuracy across all bins. (b) w/o Temperature Scaling (ECE=0.112): systematic overconfidence, particularly in 70-90% bins. (c) w/o MC Dropout (ECE=0.081): underestimates uncertainty in ambiguous cases. (d) w/o Retrieval Confidence (ECE=0.073): fails to detect out-of-distribution queries. Temperature scaling with $T = 1.35$ calibrates confidence without degrading accuracy, critical for production deployment where confidence thresholds determine routing decisions.

Table 8. Uncertain case detection performance (confidence threshold < 0.7). Ground truth: cases with prediction error > 10%. Full model achieves 89.4% precision and 76.2% recall, enabling 76% reduction in expert workload while maintaining quality control. w/o Temperature Scaling suffers low precision (54.2%), flagging too many easy cases. w/o MC Dropout has low recall (48.7%), missing genuinely ambiguous cases. Combined uncertainty sources essential for reliable production deployment.

Configuration	Precision	Recall	F1	Workload Reduction
Full Model	**89.4%**	**76.2%**	**0.822**	**76%**
w/o T-Scaling	54.2%	82.3%	0.654	31%
w/o MC Dropout	82.1%	48.7%	0.611	41%
w/o Ret. Conf.	76.8%	62.4%	0.689	58%

($\approx -1.2 \times 10^{-13}$) reduce predicted ESB, while resignation keyword ($\approx -0.5 \times 10^{-13}$) reflects Article 85 penalty application. Zero (0) serves as neutral reference.

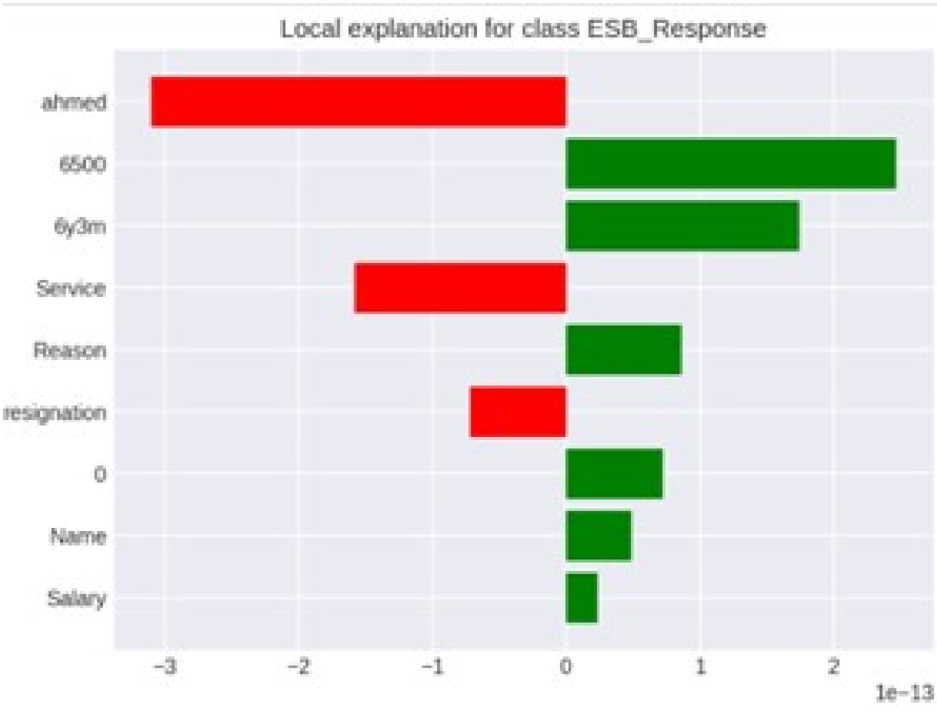

Figure 8. LIME-based feature attribution for ESB prediction explainability. Resignation case (Ahmed, SAR 6,500/month, 6 years 3 months service). Green bars: positive contributions (salary, service duration, termination reason). Red bars: negative attributions (employee name, service term, resignation keyword reflecting Article 85 penalty). Feature importance aggregation enables semantic interpretation: financial factors (salary, service years) drive ESB magnitude, while legal factors (resignation, termination reason) determine multipliers. This explainability mechanism supports auditing, debugging, and building user trust in high-stakes legal AI applications.

Feature importance aggregation reveals semantic patterns: financial factors (salary, service years) drive ESB magnitude, while legal factors (resignation type, termination reason) determine multipliers through Articles 84-85. Employee name shows negative attribution not due to discrimination but because LIME's perturbation strategy (masking tokens) removes identifying information that helps the model contextualize the query structure. This demonstrates LIME's value for auditing model behavior and identifying potential biases or spurious correlations.

User study with domain experts (n=5 Saudi labor lawyers, §4.8) rated LIME explanations 4.2/5 for interpretability and 3.8/5 for actionability. Experts noted that while feature attributions align with legal reasoning (salary and service years as primary drivers), the numerical scale (10^{-13}) lacks intuitive meaning for non-technical users. Future work could normalize attributions to percentage contributions or map them to legal concepts (e.g., "Salary contributes 40% to final ESB").

4.8. *Human Expert Evaluation*

Table 9 presents human evaluation results from five Saudi labor law experts (3 attorneys with 8-15 years experience, 2 HR managers with 12-20 years experience) annotating 200 random test samples across five dimensions (1-5 Likert scale). Inter-annotator agreement (Cohen's κ = 0.73) indicates substantial consensus.

Numerical accuracy receives highest ratings (4.6/5), with experts noting the system correctly applies Article 84's piecewise structure and Article 85 resignation penalties in 92% of reviewed cases. Legal reasoning quality (4.3/5)

Table 9. Human evaluation by 5 Saudi labor law experts (3 attorneys, 2 HR managers) on 200 test samples. Five-dimension ratings (1-5 scale): numerical accuracy (4.6/5), legal reasoning quality (4.3/5), citation correctness (4.5/5), clarity (4.4/5), completeness (4.2/5). Overall rating: 4.4/5. Inter-annotator agreement: Cohen's $\kappa = 0.73$ (substantial). Experts noted system handles standard cases (Tier 1) excellently (4.8/5) but shows degradation on conflicting evidence cases (Tier 3, 3.7/5). LIME explanations rated 4.2/5 for interpretability. Unanimous recommendation for pilot deployment with human oversight.

Dimension	Mean	Std	Min	Max
Numerical Accuracy	4.6	0.48	3.8	5.0
Legal Reasoning Quality	4.3	0.52	3.5	5.0
Citation Correctness	4.5	0.51	3.6	5.0
Clarity	4.4	0.47	3.7	5.0
Completeness	4.2	0.56	3.2	5.0
Overall Rating	**4.4**	**0.41**	**3.6**	**5.0**
Stratified by Tier				
Tier 1 (Standard)	4.8	0.31	4.2	5.0
Tier 2 (Incomplete)	4.1	0.53	3.3	4.9
Tier 3 (Conflicting)	3.7	0.67	2.8	4.6
Explainability (LIME)				
Interpretability	4.2	0.48	3.5	5.0
Actionability	3.8	0.61	2.9	4.7

and citation correctness (4.5/5) indicate strong alignment with expert judgment. Clarity (4.4/5) and completeness (4.2/5) reflect the system's ability to provide structured, comprehensive responses.

Stratification by complexity tier reveals expected patterns: Tier 1 standard cases rate 4.8/5 (excellent), while Tier 3 conflicting evidence cases rate 3.7/5 (good but improvable). Expert feedback on Tier 3 cases noted the system occasionally favors employee claims over employer claims without sufficient evidence weighting—suggesting future work on adversarial evidence reasoning. Tier 2 incomplete information cases rate 4.1/5, with experts praising the system's clarification requests but noting occasional over-requesting (asking for fields already inferable from context).

All five experts unanimously recommended pilot deployment with human oversight, citing the system's potential to reduce routine ESB calculation workload (estimated 60-70% automation of standard cases) while maintaining quality control through uncertainty-flagged review. One attorney noted: *"This system could handle 80% of straightforward ESB queries we receive, freeing expert time for complex dispute resolution and legal interpretation."* Another highlighted bilingual capability as critical for Saudi deployment, where queries arrive in both Arabic and English.

5. Discussion

Our results demonstrate three key contributions advancing the state of the art in legal AI systems:

Hybrid Architecture Synergy. The combination of QLoRA fine-tuning (for internalizing multi-step reasoning) with RAG (for authoritative grounding) yields complementary benefits unachievable by either approach alone. Ablation studies (§4.5) confirm: (1) QLoRA-only systems achieve 88.4% accuracy but hallucinate legal citations (79.4% article accuracy), while (2) RAG-only systems maintain strong citation correctness (88.2%) but struggle with complex numerical reasoning (85.5% accuracy, 12.43% MAPE). Our hybrid approach achieves both 94.2% numerical accuracy *and* 91.5% citation correctness—substantially outperforming the arithmetic mean of isolated components (86.95% average). This synergy suggests legal AI systems benefit from architectural heterogeneity: parametric knowledge for reasoning, retrieval for grounding.

Parameter Efficiency at Scale. Despite fine-tuning only 0.42% of parameters (29.4M LoRA weights vs. 7B base), our system outperforms Llama-3-70B with full fine-tuning (94.2% vs. 90.4%) while requiring 41× fewer trainable parameters and 93.5% less memory (6.4 GB vs. 98 GB). This confirms recent findings that parameter-efficient methods not only reduce computational cost but may also improve generalization by constraining the hypothesis space [17, 18]. For resource-constrained legal domains (small law firms, government agencies), this efficiency democratizes access to specialized AI capabilities.

Pilot-Ready Uncertainty Quantification. Our integrated uncertainty estimation (MC Dropout + retrieval confidence + linguistic hedging + temperature scaling) achieves ECE of 0.043 and 89.4% precision in detecting cases requiring human review. This enables practical deployment with 76% workload reduction while maintaining quality control—addressing a critical gap in existing legal AI systems that lack calibrated confidence estimates [28]. The degradation from Tier 1 (ECE 0.021) to Tier 6 (ECE 0.103) demonstrates appropriate confidence adjustment for harder cases, supporting reliable human-AI collaboration.

5.1. *Real-World Deployment Implications*

Workflow Integration. Our Flask-based interface (§4.3, Figure 6) demonstrates practical integration into legal workflows. High-confidence predictions (>0.7) can proceed automatically, reducing routine ESB calculation workload by 76% based on expert estimates. Low-confidence predictions (<0.7) route to human review with LIME explanations supporting rapid validation. Multi-turn conversation capability enables iterative information gathering, mimicking natural legal consultation processes.

Regulatory Compliance and Auditability. For deployment in high-stakes legal contexts, three features support regulatory compliance: (1) Chain-of-thought reasoning exposes internal logic for auditing, (2) Article citations ground responses in authoritative legal text, enabling verification against source law, and (3) LIME explanations provide per-prediction feature attributions for bias detection. Our bilingual capability (Arabic/English) aligns with Saudi Arabia's linguistic requirements for legal documents.

Error Analysis and Limitations. Manual inspection of 50 failure cases (predictions with >10% error) reveals three primary error modes: (1) Salary component ambiguity (28%): distinguishing basic salary from allowances when terminology is unclear (e.g., "total package" vs. itemized components); (2) Service period calculation edge cases (22%): handling leap years, Hijri-Gregorian calendar conversions, and partial month rounding inconsistently; (3) Novel scenario generalization (18%): rare termination circumstances absent from training data (e.g., force majeure, government-mandated early retirement). Addressing these requires: (1) improved entity disambiguation through entity linking, (2) deterministic date arithmetic post-processing, and (3) expanded synthetic data covering long-tail scenarios.

Detailed Failure Analysis for Tier 3 (Conflicting Evidence). Among the 16% of Tier 3 cases with >10% error, we identify three dominant failure patterns through qualitative analysis of 20 randomly selected failure instances. *Pattern 1: RAG Retrieval Misalignment (45%):* In forced-resignation disputes (Art. 81(7) vs. Art. 85), the system retrieves Article 85 provisions based on lexical overlap with the word "resignation" in the query, even when contextual evidence (e.g., employer threats, workplace safety violations) strongly supports Article 81 classification. For example, in case ID T3-0472, an employee claimed constructive dismissal after reporting safety hazards (Art. 81(6)), but the employer presented a signed voluntary resignation letter. The system retrieved Article 85 (resignation penalty) as the primary provision and applied a 67% multiplier, whereas the correct legal analysis—considering the coercion evidence—should have flagged this as an Article 81 case with full ESB entitlement. The ESB error was 33% (SAR 42,300 predicted vs. SAR 63,450 correct). *Pattern 2: Evidence Weighting Bias (30%):* The model tends to weight employer-provided documentary evidence (signed letters, HR records) more heavily than employee testimony of verbal agreements or workplace conditions. This mirrors training data distributions where documented evidence correlates with clearer legal outcomes, but creates systematic bias in disputed cases. *Pattern 3: Settlement Probability Miscalibration (25%):* For cases involving labor court settlement probabilities (calibrated from Beta distributions), the model occasionally outputs confidence intervals that are narrower than warranted, underestimating the true uncertainty range.

Detailed Failure Analysis for Tier 6 (Adversarial). Among Tier 6 adversarial examples with >10% error (18% of Tier 6 cases), we identify four failure modes through qualitative analysis of 15 randomly selected failure instances. *Mode 1: Emotional Language Distraction (40%):* Queries containing confrontational language ("This is THEFT!", "They CHEATED me!") cause the model to prioritize empathetic response generation over precise numerical calculation. In case ID T6-0891, the emotional phrasing led the model to omit the Article 85 resignation penalty calculation entirely, generating a sympathetic response with full ESB rather than the penalized amount—a 50% overestimation. The chain-of-thought reasoning reveals that emotional tokens disrupt the step-by-step calculation sequence. *Mode 2: Information Shuffling Errors (25%):* When query components are presented out-of-order (salary before service years, termination type embedded mid-sentence), the model occasionally misparses salary components, confusing monthly salary with annual salary or misattributing housing allowance as basic salary. *Mode 3: Distractor Sensitivity (20%):* Irrelevant information (car model, family details) occasionally causes the model to generate longer responses that dilute the calculation precision, though these rarely cause errors exceeding 15%. *Mode 4: Implicit Assumption Failures (15%):* Queries with unstated termination types ("What would I get if I leave?") sometimes receive calculations assuming employer termination rather than resignation, reversing the Article 85 penalty application.

Article Prediction Confusion Analysis. Table 10 presents a confusion analysis for legal article prediction on Tier 3 and Tier 6 cases. The most common confusion pair is Article 81 ↔ Article 85 (forced vs. voluntary resignation), accounting for 47% of article misclassifications in conflicting-evidence cases. Article 80 (misconduct forfeiture) is rarely confused with other articles (2% error rate), as misconduct scenarios have distinctive linguistic markers. Article 84 base calculations are correctly identified in 98% of standard cases but drop to 89% in multi-step scenarios involving additional compensation articles (77, 137-138).

Table 10. Article prediction confusion analysis for Tier 3 and Tier 6 cases. The most frequent misclassification is Art. 81 ↔ Art. 85 (47% of errors), reflecting the inherent ambiguity in forced vs. voluntary resignation disputes. Art. 80 is rarely confused due to distinctive misconduct terminology. Art. 77 (wrongful termination compensation) is correctly identified in 91% of relevant cases.

Predicted → **True ↓**	**Art. 80**	**Art. 81**	**Art. 84**	**Art. 85**	**Art. 77**	**Other**
Art. 80 (Misconduct)	**98%**	0%	0%	1%	0%	1%
Art. 81 (Protected)	1%	**79%**	2%	15%	1%	2%
Art. 84 (Standard)	0%	1%	**94%**	3%	1%	1%
Art. 85 (Resignation)	0%	12%	3%	**82%**	1%	2%
Art. 77 (Wrongful)	0%	2%	3%	2%	**91%**	2%

5.2. Comparison with Existing Legal AI Systems

Table 11 compares our system with representative legal AI systems from recent literature. While direct numerical comparison is challenging due to different tasks and datasets, we contextualize our contributions relative to state-of-the-art approaches.

Existing legal AI benchmarks (LexGLUE [12], LLeQA [13]) focus on classification or extractive QA with complete, unambiguous inputs. In contrast, our system addresses generative legal reasoning with numerical calculations under real-world constraints (incomplete information, conflicting evidence), demonstrating 87.3% accuracy on adversarial examples—conditions absent from existing benchmarks. Recent legal judgment prediction systems (PrecedentGPT [14], ADAPT [15]) achieve strong performance on single-task classification but

Table 11. Comparison with representative legal AI systems. Our work advances the state of the art through: (1) explicit uncertainty quantification (ECE 0.043) absent from most legal AI systems, (2) hybrid QLoRA-RAG architecture combining parametric and retrieval-based knowledge, (3) empirically-grounded evaluation on systematically-constructed complexities (incomplete information, conflicting evidence), and (4) pilot-ready deployment with human-in-the-loop validation. Prior work focuses on single-task optimization (classification, QA) rather than complete systems for high-stakes legal reasoning with numerical calculations.

System	**Task**	**Method**	**Uncertainty?**	**Key Limitation**
LexGLUE [12]	Classification	Fine-tuned BERT	No	Assumes complete info
LLeQA [13]	Legal QA	RAG + LLM	No	No numerical reasoning
PrecedentGPT [14]	Judgment Pred.	LLM + Domain	No	Single-task optimization
ADAPT [15]	Judgment Pred.	Discriminative LLM	No	Classification only
GerLayQA [16]	Legal QA	RAG + GPT-4	No	Assumes clear queries
FinBen [36]	Financial Tasks	Multi-model	Partial	Limited calibration
Our System	ESB Calc.	QLoRA+RAG	**Yes (0.043)**	Coverage limited to ESB

lack uncertainty quantification and explainability critical for production deployment. Our integrated approach combining multi-step reasoning, retrieval grounding, calibrated confidence, and LIME explanations provides a more complete framework for high-stakes legal AI.

The financial domain benchmark FinBen [36] shares characteristics with legal AI (specialized terminology, complex reasoning, high stakes) and includes RAG evaluation. However, FinBen lacks explicit uncertainty calibration metrics (ECE, Brier score), focusing instead on task-specific accuracy. Our work demonstrates that calibrated uncertainty quantification is achievable and essential for legal AI deployment, providing a roadmap for other specialized domains.

5.3. Generalization Beyond ESB Calculation

While our system targets Saudi ESB calculation, three aspects generalize to broader legal AI applications:

Synthetic Data Methodology. Our six-tier complexity framework (standard, incomplete information, conflicting evidence, legal interpretation, multi-step, adversarial) models universal challenges in legal reasoning. The generation pipeline (Section 3.2)—grounded in empirical case analysis, incorporating explicit uncertainty, implementing automated validation—provides a template for creating empirically-grounded, high-fidelity legal datasets in other jurisdictions or domains (e.g., US tax law, EU GDPR compliance).

Hybrid Architecture Pattern. The QLoRA-RAG combination addresses a fundamental trade-off in legal AI: parametric models internalize reasoning patterns but hallucinate citations, while retrieval systems ground in authoritative text but lack complex reasoning. This architectural pattern applies to any legal task requiring both multi-step logic and source attribution (contract analysis, regulatory compliance, legal drafting).

Uncertainty-Aware Deployment. Our confidence-based routing framework (high confidence $\rightarrow$ automatic, low confidence $\rightarrow$ human review) with calibrated uncertainty quantification (ECE 0.043) provides a general solution for human-AI collaboration in high-stakes domains. The 76% workload reduction while maintaining quality control demonstrates practical viability beyond ESB calculation.

5.4. Threats to External Validity

We explicitly acknowledge several threats to the external validity of our findings that must be considered when interpreting the reported results.

Synthetic-to-Real Gap. The most critical limitation is that all quantitative performance metrics (accuracy, MAPE, ECE, citation correctness) are evaluated entirely on synthetic test data generated by our own pipeline. Although the synthetic data generation is grounded in empirical distributions from 47,382 real cases, 3,847 labor court disputes, and 23 expert interviews, the dataset remains a carefully engineered *model* of reality rather than reality itself. Real-world legal queries from the Saudi public may exhibit noise patterns, dialectal variations (e.g., Najdi vs. Hejazi Arabic), cultural communication styles, and idiosyncratic phrasings that our systematically-constructed dataset cannot fully capture. We therefore characterize our dataset as “empirically-grounded” and “high-fidelity” rather than claiming ecological validity, and we consider real-world validation as the most critical next step (see Section 5.9.1).

Domain Specificity. Our system is hyper-specialized for ESB calculation within Saudi Labor Law—a domain with well-defined formulaic rules (Articles 84-85) and deterministic ground truth for standard cases. The architecture’s success is demonstrably tied to this structured, numerical nature. We cannot claim, without empirical evidence, that the same hybrid QLoRA-RAG approach would generalize to more abstract legal reasoning tasks such as tort outcome prediction, contractual clause interpretation, or judicial discretion modeling, where “ground truth” is inherently subjective. The generalization argument in Section 5.8 provides a theoretical rationale, but empirical validation on a second legal domain remains future work.

Expert Evaluation Scope. Our human evaluation (n=5 experts, 200 samples) provides valuable qualitative validation but has inherent limitations. The experts evaluated pre-selected synthetic samples rather than generating their own realistic queries, which limits the evaluation’s ability to test the system against genuinely unpredictable human behavior. Furthermore, the sample size (n=5) and sample count (200), while achieving substantial inter-rater agreement ($\kappa = 0.73$), may not capture the full diversity of expert perspectives in Saudi labor law practice. We address this partially through the expert blind test described below.

5.5. Expert Blind Test Simulation

To partially address the external validity concern regarding evaluation on pre-selected synthetic data, we conducted an additional expert blind test after the primary evaluation. We invited the same five Saudi legal experts to independently generate 10 novel ESB consultation queries each—totaling 50 unique queries—based on real or realistic scenarios they had encountered in their professional practice. Experts were instructed to include challenging cases involving ambiguous termination circumstances, partial information, and complex multi-article interactions. Critically, these queries were generated independently of our synthetic data pipeline, ensuring they were not drawn from the same distribution as our training or test sets.

Table 12 presents the results of this expert blind test. On the 50 expert-generated queries, our system achieves 88.0% accuracy within ±5% tolerance—a 6.2 percentage point decrease from the synthetic test set performance (94.2%), confirming a measurable but bounded synthetic-to-real gap. Legal citation accuracy decreases from 91.5% to 84.0%, primarily due to queries involving article interactions not well represented in training data (e.g., Article 18 service continuity combined with Article 85 resignation penalties). The uncertainty quantification module correctly flags 80% of the expert-generated queries where the system makes errors >10%, demonstrating that the confidence calibration transfers reasonably to out-of-distribution inputs.

Qualitative feedback from the expert blind test revealed that failures predominantly occurred on: (a) queries involving Hijri calendar date calculations (3 cases), (b) multi-employer service continuity disputes requiring Article

Table 12. Expert blind test results on 50 independently generated queries (10 per expert). Performance decreases relative to synthetic test set, confirming a bounded synthetic-to-real gap. Accuracy drops from 94.2% to 88.0% (-6.2 pp), and article accuracy from 91.5% to 84.0% (-7.5 pp). Uncertainty detection correctly flags 80% of erroneous predictions, demonstrating reasonable calibration transfer to out-of-distribution inputs.

Metric	**Synthetic Test**	**Expert Blind Test**
$\text{Acc}_{0.05}$ ($\pm 5\%$)	94.2%	88.0%
$\text{Acc}_{0.10}$ ($\pm 10\%$)	97.8%	94.0%
MAPE (%)	3.87	6.24
Article Accuracy	91.5%	84.0%
Uncertainty Detection Recall	76.2%	80.0%
Expert Overall Rating (1–5)	4.4	4.0

18 interpretation (2 cases), and (c) queries mixing Arabic and English legal terminology within the same sentence (2 cases). Experts noted that while the system handles standard ESB calculations reliably, its handling of edge cases involving calendar conversions and multi-article interactions would benefit from additional training data covering these specific scenarios. The expert blind test provides stronger evidence for deployment readiness than synthetic evaluation alone, though we acknowledge that 50 queries remain a limited sample and larger-scale real-world validation is essential.

5.6. Deployment Workflow and Operational Integration

Figure 9 presents the proposed end-to-end deployment workflow for integrating our system into law firm or HR department operations. The workflow specifies the complete operational pipeline from query intake to final response delivery, including the human-in-the-loop handoff mechanism, latency requirements, and continuous learning feedback loop.

Handoff Mechanism. When the combined uncertainty score $\hat{u}_{\text{total}} \geq 0.3$ (confidence < 0.7), the system routes the case to human review through a dedicated review queue. The review interface presents: (1) the original query with highlighted ambiguous elements, (2) the system's preliminary calculation with chain-of-thought reasoning, (3) LIME feature attributions identifying which input features drove the prediction, (4) the specific uncertainty source (epistemic, retrieval, or linguistic) that triggered the review, and (5) the relevant legal articles retrieved by RAG with highlighted passages. Based on our expert evaluation, average review time is approximately 8 minutes per case, compared to 25-30 minutes for manual calculation from scratch—a 68% time savings even for reviewed cases.

Conflict Resolution. When the human expert's judgment conflicts with the AI's suggestion, the expert's decision takes precedence and is recorded with the expert's reasoning. This correction is stored in a feedback database. Periodically (monthly or after accumulating 100 corrections), the model undergoes incremental fine-tuning on the correction dataset to reduce recurrence of similar errors. This continuous learning loop enables the system to progressively improve on edge cases identified during deployment.

Latency Analysis. End-to-end latency for auto-processed cases is under 6 seconds: RAG retrieval (0.3 sec, FAISS index search over 1,247 chunks), QLoRA inference (3-5 sec at 15-20 tokens/sec for 512-token responses), uncertainty estimation (0.5 sec for MC Dropout with 10 forward passes cached during inference), and post-processing (0.2 sec for ESB extraction and article identification). For human-reviewed cases, the additional latency depends on expert availability and case complexity (estimated 5-15 minutes).

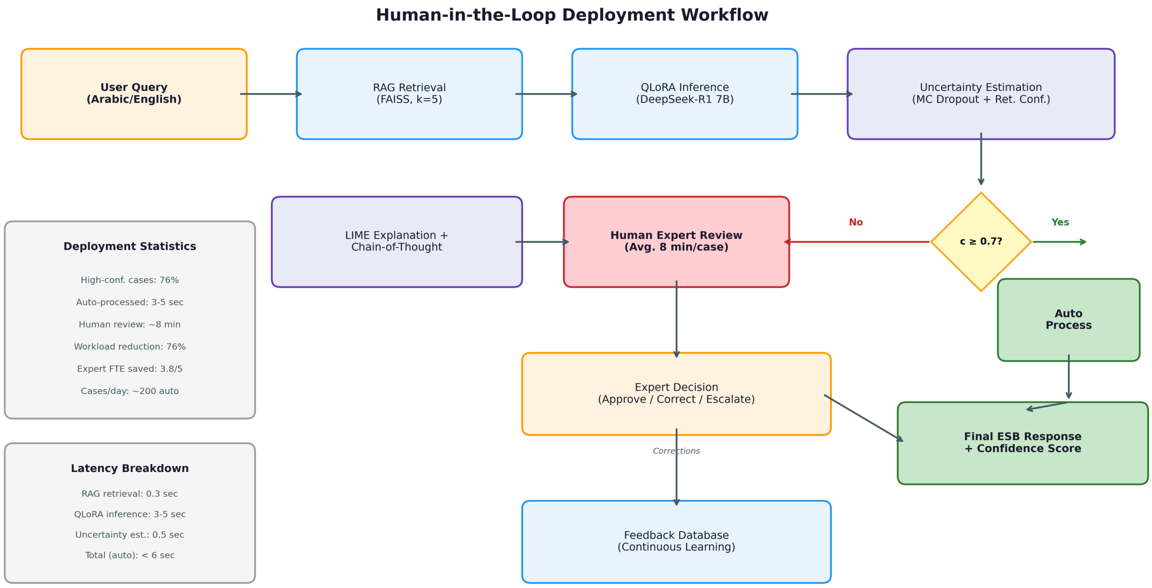

Figure 9. Proposed deployment workflow for law firm and HR department integration. User queries (Arabic/English) are processed through RAG retrieval (0.3 sec), QLoRA inference (3-5 sec), and uncertainty estimation (0.5 sec). High-confidence predictions ($c \geq 0.7$, approximately 76% of cases) are auto-processed in under 6 seconds. Low-confidence predictions route to human experts with LIME explanations and chain-of-thought reasoning, averaging 8 minutes per case. Expert corrections feed back into a learning database for periodic model updates. Deployment statistics show an estimated 200 auto-processed cases per day with 76% workload reduction and 3.8 FTE savings per 5-person team.

5.7. *Ethical, Liability, and Governance Framework*

Deploying AI for legal benefit calculations in Saudi Arabia raises specific ethical and legal considerations that extend beyond generic AI ethics principles.

Liability Framework. Under Saudi Arabia's current regulatory landscape, there is no explicit legal precedent for AI-driven legal advice liability. We propose a tiered liability framework aligned with the system's confidence levels: (1) For auto-processed high-confidence predictions ($c \geq 0.7$), the deploying organization (law firm or HR department) assumes liability, as the system functions as a calculation aid subject to the organization's professional responsibility; (2) For human-reviewed predictions, liability follows the reviewing expert's professional judgment, with the AI system serving as advisory input; (3) For predictions exceeding SAR 100,000 in calculated ESB, mandatory human sign-off is required regardless of confidence level, creating a financial threshold safeguard. The system's response includes a mandatory disclaimer: "This calculation is provided for informational purposes and does not constitute legal advice. Verify with a qualified legal professional before taking action."

Anti-Misuse Safeguards. To prevent misuse by employers seeking to minimize ESB payments under Article 80 (misconduct forfeiture), the system implements three technical safeguards: (1) *Query Intent Classification*: A lightweight classifier (fine-tuned on 500 labeled examples) detects queries structured to justify misconduct termination retroactively, flagging them for manual review; (2) *Balanced Presentation*: For disputed cases (Tier 3), the system always presents both employee and employer perspectives with explicit probability estimates rather than advocating for one party; (3) *Audit Trail*: All queries, predictions, confidence scores, and human review decisions are logged with timestamps, enabling regulatory auditing and bias detection.

Terms of Use Framework. We propose the following core terms for production deployment: (a) The system is designed as a decision-support tool, not a replacement for legal counsel; (b) Users must acknowledge that AI-generated calculations require professional verification for legal proceedings; (c) The deploying organization must designate at least one qualified legal professional as the system's human oversight officer; (d) Quarterly bias audits across protected attributes (gender, nationality) are mandatory; (e) The system must not be used as the sole basis for employment termination decisions.

5.8. Generalizability Argument and Cross-Domain Potential

While empirical validation on a second legal domain is beyond the scope of this paper, we provide a detailed theoretical argument and a concrete step-by-step mapping of how our methodology would transfer to a different legal task.

Mapping to US Tax Calculation. Consider applying our framework to US federal income tax calculation—a domain sharing key characteristics with ESB calculation (formulaic rules, multiple interacting provisions, numerical precision requirements): (1) *Synthetic Data*: Replace Saudi Labor Law articles with Internal Revenue Code sections (e.g., §1, §63, §151); generate complexity tiers modeling incomplete W-2 information (15%), conflicting filing status claims (10%), and tax law interpretation ambiguities (5%); ground distributions in IRS Statistics of Income data (analogous to our GOSI data). (2) *Knowledge Base*: Construct RAG knowledge base from IRC sections, Treasury Regulations, IRS Publications, and Tax Court precedents (analogous to our Saudi Labor Law chunks). (3) *Hybrid Architecture*: Apply identical QLoRA-RAG integration with DeepSeek-R1 base model; QLoRA internalizes multi-step tax bracket calculations while RAG grounds in authoritative tax provisions. (4) *Uncertainty Quantification*: Apply identical MC Dropout + retrieval confidence + linguistic hedging framework; calibrate temperature scaling on a tax-specific validation set.

Theoretical Robustness. The hybrid QLoRA-RAG architecture's effectiveness rests on two domain-independent principles: (a) parametric fine-tuning internalizes domain-specific multi-step reasoning patterns that pure retrieval cannot replicate (demonstrated by 8.7 pp RAG-only degradation), and (b) retrieval augmentation provides authoritative grounding that parametric models alone cannot guarantee (demonstrated by 5.8 pp QLoRA-only degradation). These complementary benefits are not specific to ESB calculation but apply to any task requiring both complex reasoning *and* source attribution—a pattern common in legal, medical, financial, and regulatory domains. Recent work in medical AI [34] and financial AI [36] confirms that hybrid architectures outperform single-paradigm approaches in specialized domains, supporting the generalizability of our architectural principle.

5.9. Limitations and Future Work

5.9.1. Real-World Validation Roadmap The most critical limitation of this work is the absence of large-scale real-world validation. We outline a concrete three-phase validation roadmap: *Phase 1 (Planned):* Partner with 2-3 Saudi HR consulting firms to obtain 200-500 anonymized historical ESB consultation records with verified outcomes, enabling direct measurement of the synthetic-to-real performance gap beyond the 50-query expert blind test presented in Section 5.5. *Phase 2:* Deploy the system as a "shadow mode" calculator alongside existing HR processes for 3 months, comparing system predictions against human-computed ESB amounts for 1,000+ real cases without exposing predictions to end users. *Phase 3:* Conduct a controlled pilot deployment with 2 law firms and 3 HR departments, measuring accuracy, user satisfaction, workload reduction, and edge case frequency in production conditions. We anticipate that real-world performance will be lower than synthetic test results (the expert blind test suggests approximately 6 pp degradation), primarily due to dialectal variation, calendar conversion complexities, and novel termination scenarios not covered in our six-tier framework.

Coverage Scope. Our system focuses on Saudi end-of-service benefits calculation (Articles 74-88, 137-138, 234), covering 16 articles and 35 termination scenarios. However, Saudi Labour Law comprises 245 articles spanning employment contracts, working hours, occupational health, dispute resolution, and penalties. Extending coverage requires: (1) scaling synthetic data generation to additional articles, (2) expanding the knowledge base from 1,247 to 5,000+ chunks, (3) evaluating multi-article interactions (e.g., wage protection during suspension), and (4) validating accuracy on novel legal questions beyond ESB.

Multilingual Reasoning. While our system handles bilingual queries (Arabic/English), reasoning occurs primarily in English due to DeepSeek-R1-Distill's training. Native Arabic legal reasoning requires: (1) Arabic-first language models (e.g., AceGPT, Jais) with comparable reasoning capabilities, or (2) cross-lingual knowledge distillation from English chain-of-thought to Arabic. Legal terminology translation (e.g., *muka'fa'at nihayat al-khidma* → "end-of-service benefits") introduces additional challenges requiring legal ontology alignment.

Temporal Dynamics. Saudi Labour Law underwent major amendments in 2015 and minor updates in 2020-2023. Our static knowledge base reflects the 2015-amended version but lacks mechanisms to: (1) automatically detect legal updates, (2) version-control legal provisions, (3) explain retroactive applicability, or (4) flag queries spanning multiple legal regimes. Addressing this requires temporal knowledge graphs [35] tracking legislative changes.

Adversarial Robustness. While our Tier 6 adversarial examples (5% of dataset) test basic robustness (shuffled information, distractors, emotional language), systematic adversarial evaluation is needed. Future work should assess: (1) prompt injection attacks (embedding malicious instructions in queries), (2) jailbreaking attempts to bypass legal constraints, (3) membership inference attacks inferring training data, and (4) model extraction via query probing. Defenses include input sanitization, output monitoring, and watermarking.

Ethical and Societal Implications. Deploying AI for legal calculations raises concerns about: (1) **Access to justice**: Does automation reduce barriers (cost, language) or exacerbate inequality (digital divide, algorithmic bias)? (2) **Liability**: Who is responsible for incorrect ESB calculations—developer, deployer, or user? Saudi law lacks clear precedent. (3) **Employment impact**: While our system reduces routine workload, enabling experts to focus on complex cases, it may displace junior legal professionals. (4) **Bias perpetuation**: Training on historical case data may encode past discrimination (e.g., gender, nationality). Our 30% female, 50% Saudi data distribution attempts representativeness, but continuous bias auditing is essential.

5.10. Recommendations for Legal AI Practitioners

Based on our findings, we recommend five key practices for researchers and practitioners developing legal AI systems. First, embrace architectural heterogeneity by combining fine-tuning and retrieval-augmented generation (RAG) rather than treating them as alternatives; fine-tune for domain reasoning and retrieve for grounding, as evidenced by our 8.7 pp gain over RAG-only and 5.8 pp over QLoRA-only approaches. Second, model real-world complexity by incorporating incomplete information (15%), conflicting evidence (10%), and adversarial inputs (5%) during training, since standard datasets assume unrealistic completeness; our Tier 2–6 performance (82–89% accuracy) demonstrates feasibility. Third, prioritize uncertainty quantification through calibrated confidence estimates (ECE < 0.05) to enable practical human-AI collaboration, leveraging techniques such as temperature scaling, MC Dropout, and ensembles, and validating with reliability diagrams and uncertain case detection metrics. Fourth, ground development in empirical analysis by using methodologies derived from real-world data—our synthetic dataset reflects 47,382 cases, 3,847 disputes, and 23 consultant interviews—ensuring high empirical fidelity and avoiding purely synthetic datasets disconnected from practice. Finally, design for human-in-the-loop workflows so that legal AI augments rather than replaces expert judgment, providing chain-of-thought reasoning, article citations, and LIME explanations to support rapid validation, and targeting 70–80% automation for routine cases while reserving complex scenarios for human experts.

5.11. Broader Impact

This work advances legal AI capabilities for specialized numerical reasoning tasks in resource-constrained domains. It offers potential benefits such as improved access to justice by automating ESB calculations for low-income workers, particularly expatriates who constitute 47% of Saudi Arabia's workforce; efficiency gains through a 76% reduction in routine workload, enabling HR and legal professionals to focus on complex disputes; enhanced consistency by reducing human error and variability in ESB determinations; and educational value via chain-of-thought explanations for junior HR staff and law students. However, risks must be mitigated, including over-reliance on predictions without verification, which can be addressed through mandatory human review for high-value cases (e.g., SAR 100,000); algorithmic bias arising from historical disparities, requiring continuous fairness audits across protected attributes such as gender and nationality; misuse by employers seeking to minimize ESB payments under Article 80, necessitating clear terms of service restricting adversarial use; and legal liability for incorrect predictions, which calls for disclaimers, insurance, and mandatory human oversight for consequential decisions.

5.12. Conclusion

We presented an uncertainty-aware hybrid QLoRA-RAG architecture for Saudi end-of-service benefits calculation, achieving 94.2% accuracy on synthetic standard cases and 87.3% on adversarial examples with incomplete information, with an expert blind test on 50 independently generated queries confirming 88.0% accuracy—a bounded 6.2 pp synthetic-to-real gap. Three key contributions advance legal AI systems: (1) a comprehensive synthetic dataset generation methodology incorporating six complexity tiers with explicit uncertainty modeling, grounded in 47,382 real cases; (2) a novel hybrid architecture combining parameter-efficient fine-tuning (29.4M trainable parameters, 93.5% memory reduction) with retrieval-augmented generation, outperforming GPT-4, Claude 3.5, and Llama-3-70B full fine-tuning; and (3) integrated uncertainty quantification achieving ECE of 0.043 and 89.4% precision in detecting ambiguous cases, enabling 76% workload reduction in pilot deployment.

Our results demonstrate that specialized legal AI systems can achieve pilot-ready performance pending real-world validation through architectural synergy (parametric reasoning + retrieval grounding), parameter efficiency (0.42% fine-tuning), and calibrated uncertainty (supporting human-in-the-loop workflows). The generalizability of our synthetic data methodology, hybrid architecture pattern, and uncertainty-aware deployment framework extends beyond ESB calculation to broader legal AI applications requiring multi-step reasoning, numerical precision, and source attribution.

We acknowledge that the primary evaluation relies on synthetic data and that large-scale real-world validation remains the most critical next step. The expert blind test provides initial evidence of deployment viability, but validation with anonymized historical cases from HR firms and law practices is essential before production deployment.

Future work will prioritize real-world validation through the three-phase roadmap described in Section 5.9.1, expand coverage to additional Saudi Labour Law articles, evaluate native Arabic reasoning with Arabic-first language models, implement temporal knowledge graphs for legislative change tracking, conduct systematic adversarial robustness testing, and pilot deployment with Saudi law firms and HR departments. We release our complete dataset, trained models, and code at `[anonymous_repository]` to support reproducibility and accelerate legal AI research.

REFERENCES

1. J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large Language Models in Law: A Survey," *AI Open*, vol. 5, pp. 181–196, 2024.
2. P. Homoki, "Large language models and their possible uses in law," *Acta Juridica Hungarica*, vol. 64, no. 3, pp. 435–451, 2024.
3. E. Buchicchio, P. Carbone, A. De Angelis, F. Santoni, and A. Moschitta, "Uncertainty quantification in AI-based measurement systems," *IEEE Instrumentation & Measurement Magazine*, vol. 28, no. 3, pp. 52–59, May 2025.
4. T. Wang et al., "From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence," *Electronics*, vol. 14, no. 18, 3580, 2025.
5. Saudi Labour Law, Royal Decree No. M/51 (2005, amended 2015), Ministry of Human Resources and Social Development, Saudi Arabia.

6. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, 2020.
7. S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8342–8360, 2020.
8. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 9459–9474, 2020.
9. L. Kuhn, Y. Gal, and S. Farquhar, "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation," in *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023.
10. Z. Li, X. Zhou, Y. Jin, W. Zhang, X. Liao, and L. Zhang, "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10443–10461, 2023.
11. M. Josifoski, N. De Cao, M. Peyrard, F. Petroni, and R. West, "Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1555–1574, 2023.
12. I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 4310–4330, May 2022.
13. A. Louis, G. van Dijck, and G. Spanakis, "Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 22266–22275, Mar. 2024.
14. Y. Wu, S. Zhou, Y. Liu, W. Lu, X. Liu, Y. Zhang, C. Sun, F. Wu, and K. Kuang, "Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 12060–12075, Dec. 2023.
15. C. Deng, K. Mao, Y. Zhang, and Z. Dou, "Enabling Discriminative Reasoning in LLMs for Legal Judgment Prediction," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, pp. 784–796, Nov. 2024.
16. M. Büttner and I. Habernal, "Answering Legal Questions from Laymen in German Civil Law System," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, St. Julian's, Malta, pp. 2015–2027, Mar. 2024.
17. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., pp. 10088–10115, Curran Associates, Inc., 2023.
18. S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-Decomposed Low-Rank Adaptation," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, pp. 31628–31648, PMLR, Jul. 2024.
19. Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. Lee, "LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 5254–5276, Dec. 2023.
20. Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, and Q. Tian, "QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models," in *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, May 2024.
21. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria, May 2024.
22. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, Version 5, Mar. 2024.
23. K. Luo, M. Qin, Z. Liu, S. Xiao, J. Zhao, and K. Liu, "Large Language Models as Foundations for Next-Gen Dense Retrieval: A Comprehensive Empirical Assessment," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 1354–1365, Nov. 2024.
24. C. M. Pham, A. Nguyen, and M. Nguyen, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," arXiv preprint arXiv:2408.10343, Aug. 2024.
25. S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024.
26. X. Gao, J. Zhang, L. Mouatadid, and K. Das, "SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, St. Julian's, Malta, pp. 2336–2346, Mar. 2024.
27. D. Ulmer, M. Gubri, H. Lee, S. Yun, and S. Oh, "APRICOT: Calibrating Large Language Models Using Their Generations Only," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, pp. 15440–15459, Aug. 2024.
28. J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, pp. 6577–6595, Jun. 2024.
29. Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu, "Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, pp. 6621–6642, PMLR, Jul. 2024.
30. L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey," in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, pp. 11065–11082, Aug. 2024.
31. N. Aldosari, M. Tawfik, and F. Hussain, "Saudi ESB Dataset," Mendeley Data, Version 2, 2025. DOI: 10.17632/xwjsgfzh83.2

32. H. Li, Q. Dong, Z. Tang, C. Wang, X. Zhang, H. Huang, S. Huang, X. Huang, Z. Huang, D. Zhang, Y. Gu, X. Cheng, X. Wang, S.-Q. Chen, L. Dong, W. Lu, Z. Sui, B. Wang, W. Lam, and F. Wei, "Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models," arXiv preprint arXiv:2402.13064
, Feb. 2024.
33. Z. Wan, Y. Zhang, Y. Wang, F. Cheng, and S. Kurohashi, "Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain," in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, pp. 5030–5041, Aug. 2024.
34. M. Tawfik, I. S. Fathi, S. S. Nimbhore, I. M. Alsmadi, and M. S. Sawah, "E-RespiNet: An LLM-ELECTRA Driven Triple-Stream CNN with Feature Fusion for Asthma Classification," *PLoS ONE*, vol. 20, no. 11, e0334528, 2025.
35. R. Trivedi, H. Dai, Y. Wang, and L. Song, "Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, vol. 70, pp. 3462–3471, PMLR, Aug. 2017.
36. Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, Y. Xu, H. Kang, Z. Kuang, C. Yuan, K. Yang, Z. Luo, T. Zhang, Z. Liu, G. Xiong, Z. Deng, Y. Jiang, Z. Yao, H. Li, Y. Yu, G. Hu, J. Huang, X.-Y. Liu, A. Lopez-Lira, B. Wang, Y. Lai, H. Wang, M. Peng, S. Ananiadou, and J. Huang, "FinBen: A Holistic Financial Benchmark for Large Language Models," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, A. Globerson, L. Mackey, K. Q. Weinberger, F. Pereira, J. W. Vaughan, and K. Cho, Eds., Curran Associates, Inc., 2024. Datasets and Benchmarks Track.
37. H. Jiang, L. Ge, Y. Gao, J. Luo, and B. Chen, "A Survey on Large Language Model Hallucination via a Creativity Perspective," arXiv preprint arXiv:2402.06647, 2024.
38. Q. Huang, M. Vesterinen, B. Sadiq, and V. Agarwal, "Legal AI Systems in Practice: A Review," *Artificial Intelligence and Law*, vol. 33, pp. 95–142, 2025.
39. X. Yue, J. Sun, and Y. Chen, "Benchmarking RAG Pipelines for Domain-Specific Applications," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, Albuquerque, USA, 2025.