A New Metadata-Aware Retrieval-Augmented Generation (RAG) Architecture for Trustworthy Legal Question Answering

Alexandra V. Jove-Ticona[1]*, Luis J. Duarte-Coaquera[2]*, Israel N. Chaparro-Cruz[3],
Silvana B. Cabana-Yupanqui[4], Americo Chaparro-Guerra[5]
School of Computer Science and Systems Engineering,
Universidad Nacional Jorge Basadre Grohmann, Tacna, Peru[1,2]
Department of Computer Science and Systems Engineering,
Universidad Nacional Jorge Basadre Grohmann, Tacna, Peru[3,4]
Department of Law, Universidad Nacional Jorge Basadre Grohmann, Tacna, Peru[5]

***Abstract*—Large Language Models (LLMs) offer strong capabilities for Natural Language Processing, yet their inherent uncertainty often produces hallucinations, confident but incorrect statements, which is critical in domains requiring precise knowledge representation. Retrieval-Augmented Generation (RAG) reduces this risk through information retrieval, but standard pipelines still suffer from fragmented context and weak alignment between queries and legal provisions, limiting trustworthy knowledge extraction. This study proposes a Metadata-Aware RAG architecture to improve grounding in large legal corpora. It integrates: 1) Sub-chunking with Legal Metadata Inheritance, which transforms unstructured legal PDFs into granular, metadata-rich fragments; and 2) an Adaptive Filter Creator, a pipeline that extracts structured constraints and compiles optimized hybrid retrieval queries. These components enhance semantic alignment, reduce uncertainty-driven hallucinations, and strengthen neural information retrieval. Using a curated Peruvian labor law corpus and 150 manually validated question–answer pairs, the system was evaluated across three LLMs (Llama-3.1-8B, GPT-OSS-20B, Gemma-3-27B). The proposed architecture achieves double-digit improvements over a Naive RAG baseline across all four RAGAS metrics—Context Precision, Context Recall, Factual Correctness, and Faithfulness—with gains ranging from 13.10% to 28.15%; notably, Faithfulness surpasses 0.90 for Gemma-3-27B. Statistical analysis confirms significance ($t(11) = 15.49$, $p = 4.06 \times 10^{-9}$) with an extremely large effect size (Cohen's $d = 4.47$). Regression results show minimal influence of model size (slope < 0.005), indicating that retrieval design has a stronger influence than parameter count in the evaluated setting.**

***Keywords*—RAG; Retrieval-Augmented Generation; LLMs; legal question answering**

I. Introduction

Workers' limited knowledge of their own rights remains a persistent barrier to fair and dignified employment. Informational asymmetries between employers and employees can translate into inconsistent treatment and tangible harm to workers [1]. In many contexts, employees are unaware of basic protections (e.g., those in the Employment Act 1955), which in turn fuels dissatisfaction and workplace disputes [2]. Evidence further indicates that workers with stronger knowledge of employment law are more inclined to advocate for change: seeking higher wages, better regulation, and improved conditions, and that union members, who often have superior access to information, are especially likely to act on this knowledge [3], [4]. Better-informed employees also tend to display deeper organizational commitment, suggesting that knowledge dissemination is not only a matter of rights but also of organizational health [5]. Yet the landscape is dynamic and complex: frequent legal updates, incomplete information, and perceived withholding of information pose substantial obstacles to workers becoming fully informed [4]. These factors call for stronger strategies by employers and policymakers to democratize access to labor-rights information, including comprehensive resources that integrate legal and economic perspectives to help workers navigate compliance complexities across jurisdictions [6].

Conversational agents (chatbots) offer a promising channel for narrowing these gaps at scale. Their always-on availability provides continuous support, enabling workers to explore policies, procedures, and rights without waiting for human agents [7], [8], [9], [10]. At the same time, responsible deployments must retain a human-in-the-loop for configuration, training, and escalation of complex or out-of-scope queries [7], [11]. Personalization can further increase impact by tailoring explanations to an individual's role, context, or language preferences, while automation of repetitive informational tasks ensures consistent, timely guidance [12]. Effectiveness also hinges on usable interaction: clear language, intuitive navigation, and supportive interface cues reduce cognitive load and help users find what they need [13]. Interactive feedback loops improve clarity and answer quality over time, and trust must be cultivated through consistency and transparency so that workers consider the chatbot's outputs reliable [14], [15].

A core obstacle to trust, however, is the phenomenon of *hallucinations* in large language model (LLM) systems: confidently stated but incorrect, illogical, or fabricated assertions [16], [17], [18], [19]. Hallucinations often arise from insufficient or biased training data and from the predictive nature of generative models, which optimize for fluency rather than grounded factuality [19], [17]. The legal domain's complexity and variability exacerbate these risks: nuanced statutory language, exceptions, temporal validity, and jurisdictional de-

*Corresponding author.

pendencies challenge purely generative approaches [20]. The consequences are non-trivial. Misinformation on labor rights can produce adverse legal and social effects, undermining user safety and institutional legitimacy [16], [17], [21]. Robust verification and validation workflows, augmented with contextual indicators and embedded markers, help constrain generation and surface provenance, thereby reducing the likelihood and impact of hallucinations [16], [18], [21].

Retrieval-Augmented Generation (RAG) has emerged as a practical mitigation strategy by grounding model outputs in authoritative, up-to-date sources at inference time [22], [23], [24]. In the legal domain, RAG improves accuracy and relevance by integrating external knowledge directly into the generation loop, a critical property when answering complex questions that demand precise citations and scope conditions [25], [26], [27], [28]. Compared with static fine-tuning, RAG-enabled assistants can draw from evolving corpora (statutes, regulations, jurisprudence, administrative guidance), thereby improving currency and reducing drift [27], [28], [29]. Evaluating such systems also requires adapted frameworks and metrics sensitive to retrieval-grounded reasoning [30]. When properly engineered, RAG-based legal chatbots can deliver expert-like answers while curbing hallucinations by privileging reliable documents during response synthesis [28], [31], [32].

Beyond model and prompt design, the quality of a RAG system hinges on how the corpus is partitioned ("chunking") and queried. Chunks that are too coarse dilute topical focus; chunks that are too fine lose essential legal context and citation anchors. Retrieval quality further depends on whether index structures preserve jurisdictional and versioned metadata and whether query-time filters can faithfully express the question's scope. Consequently, improvements in both storage (how documents are chunked and represented in the vector database) and search (how queries are mapped to structured constraints) have first-order effects on context Recall, context Precision, answer Faithfulness, and factual Correctness.

These limitations reveal a gap in current legal RAG approaches: existing pipelines do not systematically exploit the hierarchical structure and legal metadata of source documents to jointly guide document representation and query-time retrieval. To address this gap, this study makes three main scientific and technical contributions. First, it introduces a metadata-aware legal document representation based on hierarchical sub-chunking with legal metadata inheritance, preserving clause-level context and provenance during document segmentation and vector indexing. Second, it proposes an Adaptive Filter Creator that extracts structured constraints from user queries and combines them with hybrid similarity-based retrieval, progressive filter relaxation, and cross-encoder reranking. Third, it provides an empirical evaluation of the proposed retrieval architecture against a Naive RAG baseline across three contemporary open-weight LLMs of different scales, using Context Precision, Context Recall, Factual Correctness, and Faithfulness, complemented by statistical analysis of the observed improvements. Together, these contributions provide a retrieval architecture designed to improve the factual grounding, contextual relevance, and reliability of legal question-answering systems.

The remainder of this study is organized as follows. Section II presents the related work and research background. Section III describes the proposed Metadata-Aware RAG architecture and its main components. Section IV presents the experimental methodology, dataset, evaluation metrics, and implementation details. Section V reports and analyzes the experimental results, followed by a discussion of their implications and limitations. Finally, Section VI presents the conclusions and directions for future work.

II. Background

A. Large Language Models

Large Language Models (LLMs) are neural sequence models trained to predict the next token in text. Given a tokenized sequence $x_{1:T}$ over a vocabulary $\mathcal{V}$, an LLM learns a conditional distribution:

$$p_\theta(x_t \mid x_{<t}) = \mathrm{softmax}\big(W\, h_t + b\big) \tag{1}$$

where, h_t is the hidden state at position t and W, b project to $|\mathcal{V}|$ logits. Training maximizes the likelihood of observed text (equivalently minimizes cross-entropy):

$$\mathcal{L}_{\mathrm{LM}}(\theta) = -\sum_{(x_{1:T}\sim\mathcal{D})}\sum_{t=1}^{T}\log p_\theta(x_t \mid x_{<t}) \tag{2}$$

$$\mathcal{L}_{\mathrm{LM}}(\theta) = H(p_{\text{data}}) + \mathrm{KL}\big(p_{\text{data}} \parallel p_\theta\big) \tag{3}$$

so the model approximates the data distribution p_{data} by minimizing $\mathrm{KL}(p_{\text{data}}\|p_\theta)$ in expectation over the training corpus $\mathcal{D}$.

B. Hallucinations

Despite strong fluency, LLMs can generate confident but incorrect content. In our legal/labor rights context, the principal drivers are:

1) Objective mismatch: The learning target is *likelihood under* p_{data}, not *truth*. Because the expectation in $\mathcal{L}_{\mathrm{LM}}$ is over real data, assigning probability mass to plausible-but-false strings outside p_{data} is not directly penalized. This favors fluent, distributional plausibility over verified factuality.

2) Noisy / incomplete corpora: Pretraining data may contain errors, outdated laws, or conflicting jurisdictions; the model internalizes such patterns.

3) Distribution shift: Legal facts are temporal (effective dates, amendments). Queries about post-cutoff changes require knowledge the model never saw; it may extrapolate incorrectly.

C. Embeddings

An embedding is a vector representation of a discrete object (token, sentence, document) that places semantically related items close together in a continuous space. Formally, an encoder/embedding model f_ϕ maps text x to a point in $\mathbb{R}^d$:

$$\mathbf{z} = f_\phi(x) \in \mathbb{R}^d. \tag{4}$$

Such that semantic similarity in language corresponds to geometric proximity in the vector space. Concretely, an embedding model implements a function $f_\phi : \mathcal{X} \to \mathbb{R}^d$ that produces a fixed-length representation $\mathbf{z} = f_\phi(x)$ for input x, typically normalized to $\|\mathbf{z}\|_2 = 1$ to enable cosine or inner-product similarity. Unlike autoregressive LLMs that *generate* sequences, embedding models *encode* inputs into vectors that can be indexed and searched efficiently.

D. Retrieval-Augmented Generation (RAG)

A fundamental limitation of large language models is their bounded context window: they can only process a finite number of tokens in a single pass. This constraint precludes the possibility of simply feeding all relevant documents into the prompt, especially when working with large or multi-document corpora. To address this, a Retrieval-Augmented Generation (RAG) pipeline augments an autoregressive language model with a simple dense-retrieval step. Documents are preprocessed into fixed-size *chunks* and embedded; at inference, the query is embedded, the top-k nearest chunks are fetched, concatenated with the query, and passed to the LLM to generate an answer.

Let $\mathcal{C} = \{d_i\}$ be chunked documents with embeddings $\mathbf{z}_i = f_\phi(d_i)$. For a query x with embedding $\mathbf{q} = f_\phi(x)$, retrieval is:

$$C_k(x) = \underset{d \in \mathcal{C}}{\arg\,\text{top-k}}\ s(x, d) \tag{5}$$

$$s(x, d) = \mathbf{q}^\top \mathbf{z}_d = \frac{\langle f_\phi(x), f_\phi(d) \rangle}{\|f_\phi(x)\|\,\|f_\phi(d)\|} \tag{6}$$

The generator conditions on the query and retrieved context C_k:

$$p_\theta(y \mid x, C_k) = \prod_t p_\theta\big(y_t \mid x, C_k, y_{<t}\big) \tag{7}$$

and the answer $\hat{y}$ is obtained by decoding (e.g., greedy or top-p sampling).

1) Typical design in practice: A naive RAG baseline typically chunks documents into fixed-length windows (e.g., N tokens) with small overlaps and minimal cleaning; indexes each chunk in a vector database by storing its embedding–identifier pair $(\mathbf{z}_d, \text{doc-id})$ and enabling approximate nearest-neighbor (ANN) search; performs retrieval via pure dense similarity; assembles the prompt by *stuffing*, i.e., concatenating the top-k chunks into a single context block; and decodes with standard LLM settings (e.g., fixed temperature) without explicit verification or abstention logic. The main pipeline is illustrated in Fig. 1.

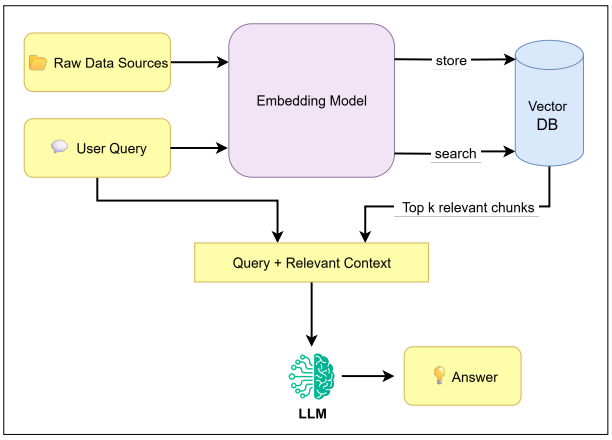

Fig. 1. Naive RAG.

However, this naive configuration introduces several well-known limitations. First, the use of fixed-size chunking with minimal preprocessing often leads to fragmented context, redundancy, or inclusion of irrelevant text, which degrades retrieval accuracy [33]. Second, relying solely on dense similarity for approximate nearest-neighbor (ANN) search increases the risk of retrieving semantically similar but contextually inappropriate passages [33]. Third, the practice of prompt "stuffing" by concatenating top-k chunks into a single context block frequently results in noisy or incoherent inputs to the LLM, especially when the retrieved passages overlap or conflict [33]. Moreover, decoding with fixed hyperparameters and without any verification or abstention mechanisms makes the system highly vulnerable to hallucinations and ungrounded responses [34]. Collectively, these limitations constrain the effectiveness and robustness of naive RAG pipelines, particularly in domain-specific or compliance-sensitive enterprise scenarios [34].

III. Proposed Metadata-Aware RAG Architecture

The proposed approach consists of a Retrieval-Augmented Generation architecture designed to address the limitations previously identified in conventional RAG pipelines, often referred to as naive RAG configurations. Fig. 3 shows the pipeline of our proposal. The approach is structured around two core components.

A. Sub-Chunking with Legal Metadata Inheritance

Our processing approach employs a multi-step method that transforms unstructured documents into enriched fragments, ready for indexing. The process is illustrated in Fig. 2, and comprises the following steps:

1) Document transformation: The process begins with unstructured documents. In an initial stage, these files are converted to an intermediate format, like Markdown (.md). Subsequently, a preprocessing phase is conducted to clean the text by removing headers, footers, and other irrelevant elements. This step is performed either algorithmically or manually, as required.

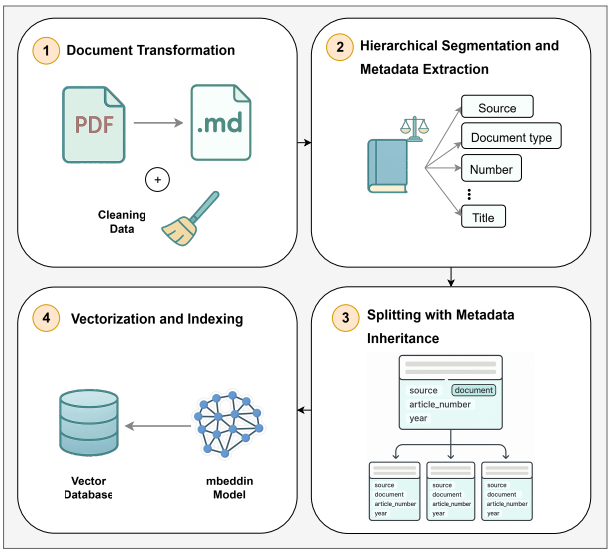

Fig. 2. Sub-chunking with legal metadata inheritance.

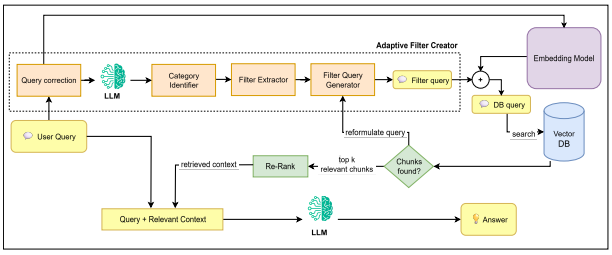

Fig. 3. Metadata-aware RAG (our proposal).

2) Hierarchical segmentation and metadata extraction: Next, a hierarchical splitting technique is applied to segment each document into coherent legal units. A key advantage of legal texts is that they are inherently structured and generally follow predictable patterns, which can be systematically exploited for segmentation. For example, a Compendium of Labor Regulations may include a heterogeneous set of instruments, such as supreme decrees, legislative decrees, ministerial resolutions, and laws, each identified by various attributes. This built-in regularity allows the segmentation process to align with the natural hierarchy of the documents rather than relying on arbitrary token-based splits.

This process not only divides the content but also extracts the metadata inherent to each section (e.g., the article number or the title of the law). The output is a semi-structured `JSON` document where the content and its metadata are formally linked.

3) Splitting with metadata inheritance: After the hierarchical segmentation and extraction of metadata, the resulting `JSON` documents are still relatively large. To facilitate effective semantic retrieval and to ensure that no relevant information is overlooked by the language model, these documents are further divided into smaller fragments.

Each fragment retains a full copy of the metadata from its parent document, ensuring that the structural context is preserved. This process allows the system to maintain precise references to the source of each fragment, which is critical for both accuracy and traceability in downstream tasks.

4) Vectorization and indexing: Finally, these enriched sub-fragments are processed through an embedding method to convert them into numerical vectors. These vectors are then indexed in a vector database, allowing fast and accurate semantic retrieval during queries.

B. Adaptive Filter Creator

This process involves a structured workflow of sequential stages designed to process the user query, optimally retrieve and organize information from the vector database, and extract the most relevant fragments. These fragments are then integrated into an enriched context, which forms the basis for generating the final response. Each step is illustrated in Fig. 4

Processing commences with the evaluation of the user's query.

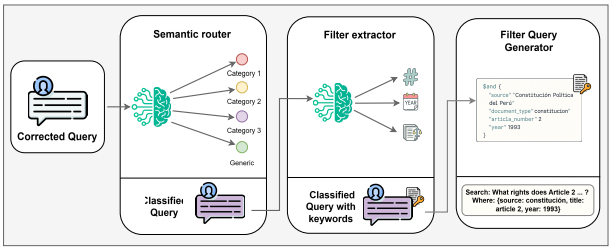

Fig. 4. Adaptive filter creator.

1) Query correction: The user's query is first corrected grammatically and cleaned. This step is important because it reduces ambiguities and ensures that subsequent retrieval and generation stages operate on a clear and well-structured input.

2) Semantic router: An LLM classifies the query intent into a set of predefined document categories, thereby narrowing the search to the most relevant data subset. If the LLM fails to assign a category with sufficient confidence, the system defaults to a comprehensive search across all documents to ensure information retrieval.

3) Filter extractor: The LLM analyzes the query to identify entities that correspond to explicitly mentioned metadata values (e.g., article number, year, document, etc.), which are then used to build the query for the database.

4) Filter query generator: This stage receives the outputs from the preceding modules to assemble the final optimized query. It follows a two-step strategy: first, metadata filtering is applied using the extracted entities (e.g., year, document source) to restrict the search space to a subset of documents that explicitly match those attributes; second, within this filtered subset, a semantic search is conducted with the user's embedded query vector to retrieve chunks that best align with the user's intent.

This component also includes a fallback mechanism: if a query is too restrictive and yields no results, filters are progressively removed to ensure that chunks are always retrieved.

The final output of this stage is an initial set of top-k relevant chunks from the vector database.

While the initial chunks are relevant, they are passed through a refinement stage to maximize precision. The top-k chunks are sent to a cross-encoder model, which performs a more thorough and computationally intensive analysis than the initial vector search. It directly compares the user's query against each chunk to calculate a highly precise relevance score. This process reorders the list, ensuring that the chunks with the highest semantic relevance are moved to the top.

Finally, the highest-scoring, re-ranked sub-chunks are concatenated to form a single, highly focused, and enriched context. This refined context is then supplied to the final generative LLM, along with the original user query, to synthesize the information and articulate a coherent final answer.

IV. Materials and Methods

A. Experimental Setup

The experiments were designed and executed on a local workstation equipped with an Intel(R) Core(TM) Ultra 9 285K Tetracosa-core (24 core) processor, 128 GB of RAM, and an NVIDIA GeForce RTX 5090 graphics processing unit (GPU) with 32 GB of VRAM for computational acceleration.

The software framework was implemented within a Python 3.10 environment. The LangChain library was employed for the orchestration of the RAG pipelines. Text embeddings were generated using the HuggingFace library, with the resulting vectors stored and managed in ChromaDB. The Ollama platform was utilized for local serving of language models, and quantitative evaluation was conducted using the Ragas framework.

B. Dataset

The corpus for this study was built to provide a representative knowledge base of the Peruvian legal framework, with a particular emphasis on labor law. The primary sources were the Political Constitution of Peru and three major legal compilations: the General Labor Compendium, the Private Labor Regime, and the Special Labor Regime. These compilations consist of a vast collection of legal instruments, including laws, supreme decrees, ministerial resolutions, and legislative decrees. Although they do not cover the entire body of Peruvian labor legislation, they represent a highly significant portion of it, with an aggregated volume exceeding 3,000 pages of legal text.

The corpus was processed using the proposed sub-chunking approach, resulting in more than 6,000 sub-chunks.

The metadata associated with each sub-chunk and indexed in the vector database included: *source* (e.g., "General Labor Compendium"), *document_type* (e.g., "decree"), *title* (e.g., "Legislative Decree No. 728"), *topic* (e.g., "labor_law"), *year* (e.g., 1991), and the original source link.

C. Models

For embedding, the *BAAI/bge-m3* model was employed, chosen for its strong multilingual coverage and proven effectiveness in semantic similarity tasks. In the reranking stage, the *jina-reranker-v2-base-multilingual* model was applied to reorder retrieved passages, thereby enhancing contextual precision and mitigating irrelevant matches. Finally, for the generation stage, three Large Language Models (LLMs) of varying scales and characteristics were empirically compared: *llama-3.1* (8b), *gemma-3* (27b), and *gpt-oss* (20b).

This comparative evaluation enabled us to assess trade-offs between open-source and proprietary models, as well as between efficiency and generation quality. Table I presents the main specifications of the evaluated LLMs.

TABLE I. Main Specifications of the Evaluated LLMs

Specifications	Llama-3.1-8B	Gemma-3-27B	GPT-OSS-20B
Parameters	8B	27B	21B
Size	4.9 GB	17 GB	13 GB
Max Context	128K tokens	128K tokens	128K tokens

D. Metrics

The evaluation of Retrieval-Augmented Generation (RAG) systems requires metrics that go beyond conventional text similarity or lexical overlap measures. Traditional metrics such as BLEU or ROUGE primarily assess surface-level similarity between generated and reference texts, which fails to capture the degree of factual grounding or the contextual relevance of retrieved evidence. To address these limitations, the Retrieval-Augmented Generation Assessment Suite (RAGAS) [35] was employed in this study as a specialized framework for evaluating the quality, factual consistency, and contextual reliability of RAG-based systems.

RAGAS introduces a unified set of metrics designed to measure how faithfully and effectively a model integrates retrieved knowledge into its responses. Specifically, four key metrics were used in this research, Faithfulness, Factual Correctness, Context Recall, and Context Precision, to compare the performance of the Naive RAG and the proposed Metadata-Aware RAG. Together, these metrics quantify both the trustworthiness of generated answers and the relevance of the retrieved evidence, offering a comprehensive evaluation of retrieval and generation quality.

1) Faithfulness: Faithfulness measures the degree to which the generated answer is grounded in the retrieved context. It quantifies the proportion of factual statements in the generated answer that are supported by the retrieved passages. If S denotes the set of factual statements extracted from the answer and S_c the subset verifiable within the retrieved context, then:

$$\text{Faithfulness} = \frac{|S_c|}{|S|} \tag{8}$$

A high value indicates that the model relies primarily on retrieved knowledge rather than hallucinated or unsupported content. In practice, this metric is estimated using an entailment model to automatically determine whether each factual claim in the answer is logically entailed by the retrieved passages.

2) Factual correctness: Factual Correctness evaluates the alignment between the generated answer (A_{gen}) and a ground-truth reference answer (A_{ref}). It measures the proportion of statements in A_{gen} that are semantically consistent with A_{ref}, regardless of whether they appear in the retrieved documents. Formally:

$$\text{Factual Correctness} = \frac{|S_{match}(A_{gen}, A_{ref})|}{|S(A_{gen})|} \tag{9}$$

where, S_{match} represents the subset of factual statements in A_{gen} that are semantically equivalent or entailed by A_{ref}. This metric is computed using a natural language inference (NLI) model to evaluate textual entailment between answer segments. It reflects the objective accuracy of generated outputs, which is particularly crucial in domains such as legal question-answering, where factual errors may alter the legal meaning of the response.

3) Context recall: Context Recall quantifies how effectively the retriever retrieves all relevant information required to answer a query. Let R denote the set of all relevant information units in the knowledge base, and R_{ret} the subset retrieved by the system. Then:

$$\text{Context Recall} = \frac{|R_{ret} \cap R|}{|R|} \tag{10}$$

High recall implies that the retriever captures most of the necessary evidence from the knowledge base, thereby minimizing omission errors. This metric is calculated by comparing the semantic similarity between the retrieved passages and the reference evidence set, using embedding-based cosine similarity with a predefined relevance threshold τ.

4) Context precision: Context Precision evaluates the proportion of retrieved information that is truly relevant to the question, penalizing the inclusion of irrelevant or noisy passages. Let R_{ret} represent all retrieved passages and R_{rel} the subset deemed relevant:

$$\text{Context Precision} = \frac{|R_{rel}|}{|R_{ret}|} \tag{11}$$

This metric provides insight into the selectivity of the retriever: higher precision values indicate fewer irrelevant documents and a cleaner, more focused context for the generator. In this work, relevance was determined by computing the cosine similarity between the query embedding and each retrieved passage embedding, labeling a passage as relevant if $\text{sim}(q, d) \geq \tau$.

5) Implementation details: All metrics were computed automatically using the open-source `ragas` Python library, which implements these definitions through a pipeline combining:

- Semantic entailment models (`deberta-large-mnli`) for faithfulness and factual correctness,
- Sentence-transformer embeddings (`all-MiniLM-L6-v2`) for computing context recall and precision,
- Automatic parsing of model outputs into atomic factual statements via dependency-based segmentation.

Each score was averaged across the full evaluation dataset, consisting of a balanced set of legal question–answer pairs. All metrics were normalized to the range $[0, 1]$, where higher values indicate better performance.

6) Interpretation: Collectively, these four metrics provide a multidimensional view of RAG performance. *Faithfulness* and *Factual Correctness* assess the quality of generation, while *Context Recall* and *Context Precision* evaluate retrieval quality. A robust RAG system should exhibit both high recall (completeness) and high precision (relevance), along with faithful and factually correct generations. In this study, these metrics serve as the foundation for comparing the Naive RAG and the proposed Metadata-Aware RAG across different model sizes and configurations.

E. Experiments

The experimental procedure was designed to systematically evaluate the performance of three LLMs, comparing the traditional Naive RAG configuration with the proposed Metadata-Aware RAG approach.

The evaluation used a curated set of 150 question-answer pairs derived from the Peruvian legal corpus. The dataset was created in two stages. First, a generative language model was used to generate candidate questions and answers from randomly selected chunks of the corpus. Second, the generated pairs were manually reviewed and validated by two legal experts: one specializing in law and belonging to the research team, and one external expert specializing in labor law. Only those pairs that met the validation criteria were retained.

A question-answer pair was retained only if: 1) the question was clear and unambiguous; 2) the answer could be found in a single identifiable passage of the corpus; and 3) the answer appeared verbatim in the source text. The questions were also distributed across the main labor regimes and constitutional rights to provide coverage of the main legal topics in the corpus.

Regarding difficulty, the questions were designed to evaluate the retrieval of specific legal information from the corpus rather than complex multi-hop legal reasoning. They require the system to identify relevant provisions and retrieve the corresponding information from the appropriate legal passage, including details such as legal requirements, applicable dates, conditions, and article references. The requirement that the answer be explicitly stated in the source text excludes questions that depend on combining information from multiple documents or performing extensive legal interpretation. This design allows the evaluation to focus on retrieval accuracy and factual grounding, while reducing the influence of the reasoning capabilities and hallucination tendencies of the generation model.

For both the Naive RAG baseline and the proposed Metadata-Aware RAG, retrieval depth was fixed at $k = 5$,

and generation was performed with a temperature of 0.1 to balance determinism and lexical variation. Both pipelines were evaluated under identical conditions, ensuring that the only difference between them lay in the retrieval architecture. Each was tested with three open-weight LLMs: Llama-3.1-8B, GPT-OSS-20B, and Gemma-3-27B. These models were selected based on three complementary considerations. First, they represent recent high-performing open-weight LLMs with different parameter scales, enabling the evaluation of the proposed retrieval architecture across increasing model capacities. Second, these model families have broad practical availability through major LLM inference services, supporting their relevance to contemporary real-world LLM applications. Third, the selected range allows the influence of model size on the proposed architecture to be assessed under controlled conditions. Finally, the 8B to 27B parameter range reflects the maximum capacity feasible for local inference, as the 27B model reaches the 32 GB VRAM physical limit of the RTX 5090 workstation utilized in this study.

The following objectives were considered in order to validate the proposed architecture:

- To evaluate the performance of Naive RAG and the proposed Metadata-Aware RAG using three language models of different sizes.
- To measure the improvement achieved by the proposed RAG over the Naive RAG across all evaluation metrics.
- To determine the influence of model size on the performance of the proposed RAG.
- To statistically validate the effectiveness of the proposed RAG compared to the Naive baseline.

F. Implementation Details

To ensure experimental reproducibility, the main hyperparameters, retrieval settings, and model configurations are specified as follows. The document chunking pipeline utilizes a primary structural split based on legal boundaries (e.g., full decrees, laws, or articles), followed by recursive sub-chunking with a maximum window of 250 tokens and a 30-token overlap to preserve boundary context. Semantic vector embeddings are generated using BAAI/bge-m3. During query execution, the Adaptive Filter Creator performs an initial hybrid vector search retrieving $k_{initial} = 20$ candidate passages. If metadata filters yield no results, progressive filter relaxation is applied. Candidates are then re-ranked using jina-reranker-v2-base-multilingual, with the top $k = 5$ passages selected for prompt synthesis. A cosine similarity threshold of $\tau = 0.65$ is enforced during context evaluation. All generative inference is performed with a temperature of $T = 0.1$ and top-$p = 0.9$ sampling.

The source code and implementation materials are publicly available at https://github.com/avjovet/Metadata-AwareRAG.

V. Results

A. Evaluation of Naive and the Oroposed RAG Across Different Language Models

The first experimental objective was to evaluate the performance of the *Naive RAG* and the *proposed Metadata-Aware RAG* pipelines using three large language models (LLMs) of different sizes: *Llama 3.1 8B*, *GPT-OSS 20B*, and *Gemma 3 27B*. Each model was assessed under identical retrieval and generation conditions to ensure comparability. The evaluation employed the four RAGAS metrics previously described: Context Precision, Context Recall, Factual Correctness, and Faithfulness, to capture both retrieval quality and generation trustworthiness.

TABLE II. Comparison of Naive RAG and Proposed Metadata-Aware RAG Across Three Language Models and Four RAGAS Metrics. Bold Values Indicate the Best Performance for Each Metric

RAG Pipeline	Metric	llama-3.1-8b	gpt-oss-20b	gemma-3-27b
Naive RAG	Context Precision	0.730	0.732	**0.734**
	Context Recall	0.710	**0.714**	**0.714**
	Factual Correctness	0.597	**0.620**	0.615
	Faithfulness	0.725	**0.762**	0.755
Our RAG	Context Precision	0.917	0.922	**0.929**
	Context Recall	0.905	0.912	**0.915**
	Factual Correctness	0.745	0.758	**0.772**
	Faithfulness	0.820	0.868	**0.913**

As observed in Table II, the proposed Metadata-Aware RAG consistently outperformed the Naive RAG across all evaluated metrics and models. The improvement is particularly pronounced in the *Faithfulness* and *Context Precision* dimensions, which measure the degree of grounding and relevance of retrieved evidence. For instance, when using *Gemma 3 27B*, faithfulness increased from 0.755 to 0.913, and context precision from 0.734 to 0.929, demonstrating a significant reduction in unsupported or noisy content generation.

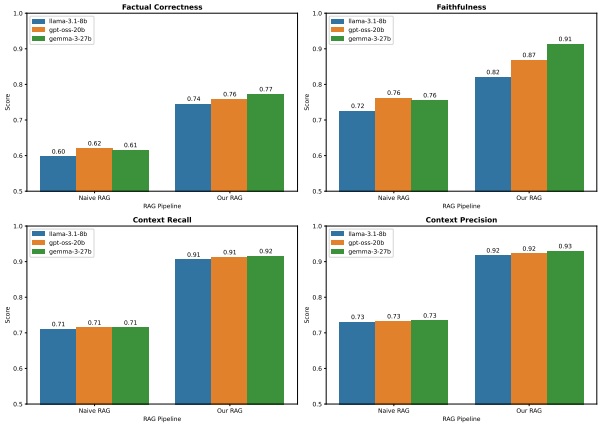

Fig. 5. Performance comparison of Naive RAG and proposed Metadata-Aware RAG across three language models and four RAGAS metrics. The proposed RAG shows consistent improvements across all models and evaluation dimensions.

Fig. 5 visually reinforces the numerical trends observed in Table II. The proposed RAG demonstrates clear and consistent gains in all four RAGAS metrics, with particularly strong effects in *Faithfulness* and *Context Precision*. The uniformity of the improvements across the three models indicates that the proposed retrieval enhancements generalize effectively across architectures of varying capacities.

Overall, these results confirm that the proposed RAG achieves a robust balance between retrieval coverage (high recall) and relevance (high precision) while improving factual

grounding and semantic accuracy. This experiment establishes a quantitative baseline for subsequent analyses, which further measure relative improvements, assess the influence of model size, and statistically validate the observed differences.

B. Improvement Achieved by the Proposed RAG Over the Naive RAG

The second objective of this study was to quantify the improvement achieved by the proposed Metadata-Aware RAG pipeline relative to the Naive RAG baseline across all evaluation metrics and language models. The aim of this analysis is to determine the magnitude of gains attributable to the introduction of metadata-guided retrieval and structured context filtering mechanisms.

As illustrated in Fig. 6, the proposed Metadata-Aware RAG exhibits consistent and substantial performance improvements across all metrics: *Factual Correctness*, *Faithfulness*, *Context Recall*, and *Context Precision*. The visual trends reveal that for each language model, the proposed RAG achieves higher scores, with particularly pronounced gains in *Context Recall* and *Context Precision*. These results confirm that metadata-aware retrieval contributes both to broader information coverage and to more relevant evidence selection, directly enhancing the generation phase by improving factual and contextual grounding.

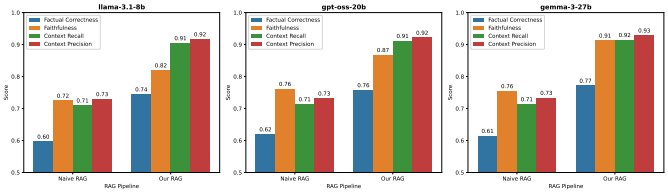

Fig. 6. Comparison of Naive RAG and proposed Metadata-Aware RAG across three language models and four RAGAS metrics. The proposed RAG consistently demonstrates higher scores, reflecting significant improvements in retrieval precision, contextual grounding, and factual reliability.

To complement the visual analysis, Table III reports the percentage improvement of the proposed RAG over the Naive RAG for each metric and model. The proposed RAG yields double-digit gains in every setting, with improvements ranging from 13.10% to 28.15%. The highest relative improvement was observed for *Context Recall* with *Gemma-3-27B* (28.15%), followed closely by *Context Precision* (26.57%) and *Factual Correctness* (25.53%) using the same model. These results confirm that larger models benefit even more strongly from metadata-aware retrieval, amplifying the improvements over the baseline pipeline.

The consistent improvement across all metrics and models provides strong evidence that the proposed RAG enhances the retrieval and generation process in a robust and scalable manner. The proportional increases in Context Precision and Context Recall confirm that metadata-guided retrieval reduces noise while increasing the availability of relevant information. The accompanying rise in Faithfulness and Factual Correctness demonstrates improved grounding and factual reliability, reinforcing the claim that retrieval quality directly influences answer trustworthiness.

TABLE III. PERFORMANCE IMPROVEMENT (%) OF THE PROPOSED METADATA-AWARE RAG OVER THE NAIVE RAG ACROSS METRICS AND MODELS

Metric	Model	Naive RAG	Our RAG	Improvement
Factual Correctness	llama-3.1-8b	0.597	**0.745**	24.79%
	gpt-oss-20b	0.620	**0.758**	22.26%
	gemma-3-27b	0.615	**0.772**	**25.53%**
Faithfulness	llama-3.1-8b	0.725	**0.820**	13.10%
	gpt-oss-20b	0.762	**0.868**	13.91%
	gemma-3-27b	0.755	**0.913**	**20.93%**
Context Recall	llama-3.1-8b	0.710	**0.905**	27.46%
	gpt-oss-20b	0.714	**0.912**	27.73%
	gemma-3-27b	0.714	**0.915**	**28.15%**
Context Precision	llama-3.1-8b	0.730	**0.917**	25.62%
	gpt-oss-20b	0.732	**0.922**	25.96%
	gemma-3-27b	0.734	**0.929**	**26.57%**

These results strongly support the conclusion that metadata-aware retrieval mechanisms significantly strengthen RAG pipelines, providing systematic benefits regardless of model size and suggesting that such strategies may be essential components of future trustworthy LLM-based information systems.

C. Influence of Model Size on the Performance of the proposed RAG

To assess whether larger language models yield additional benefits for the proposed Metadata-Aware RAG, we evaluated the evolution of each RAGAS metric across three model sizes: 8B, 20B, and 27B parameters. Fig. 7 visualizes the performance tendencies for both pipelines. The dashed lines correspond to the Naive RAG, while the solid lines show the behavior of the proposed RAG.

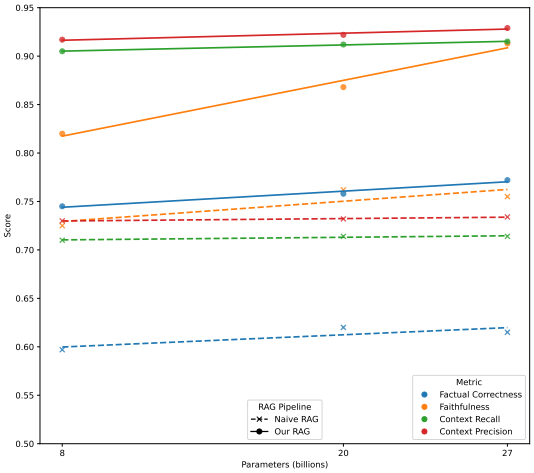

Fig. 7. Regression of model size (in billions of parameters) against performance scores across the four RAGAS metrics. Solid lines represent the proposed Metadata-Aware RAG, while dashed lines represent the Naive RAG.

The visual trends reveal two consistent patterns.

First, at every model size and for every metric, the proposed RAG scores are *always* higher than the Naive RAG scores. This performance gap remains stable across all parameter scales, confirming that the metadata-aware retrieval mechanism provides benefits that are independent of model capacity.

Second, the slopes of the regression lines are extremely small for most metrics, indicating that increases in parameter count, from 8B to 27B, lead to only marginal improvements. For *Context Precision*, *Context Recall*, and *Factual Correctness*, the curves grow slightly but remain nearly flat, suggesting that model size exerts minimal influence on these dimensions. In practical terms, each additional billion parameters yields only a very small increase in score, and the effect is nearly negligible.

TABLE IV. LINEAR REGRESSION BETWEEN MODEL SIZE (BILLIONS OF PARAMETERS) AND METRIC SCORES FOR NAIVE AND THE PROPOSED RAG. LARGER SLOPES INDICATE GREATER SENSITIVITY TO MODEL SCALING

Metric	RAG Pipeline	slope	intercept	r value	p value	std error
Context Precision	Naive RAG	0.00021	0.72823	**0.98865**	**0.09599**	**0.00003**
	Proposed RAG	**0.00061**	**0.91151**	0.96972	0.15706	0.00015
Context Recall	Naive RAG	0.00022	0.70856	0.93130	0.23734	0.00009
	Proposed RAG	**0.00053**	**0.90090**	**0.99710**	**0.04850***	**0.00004**
Factual Correctness	Naive RAG	0.00105	0.59137	0.83592	0.36987	0.00069
	Proposed RAG	**0.00138**	**0.73295**	**0.98522**	**0.10960**	**0.00024**
Faithfulness	Naive RAG	0.00174	0.71540	0.85156	0.35131	0.00107
	Proposed RAG	**0.00480**	**0.77904**	**0.99128**	**0.08414**	**0.00064**

To formalize these observations, Table IV reports the regression coefficients and statistical indicators. Across all metrics, the proposed RAG consistently exhibits slightly higher slopes than the Naive RAG, reflecting its better capacity to extract value from larger models. However, even these slopes remain small in absolute magnitude, again confirming that model size is not a decisive factor for performance gains.

The regression results confirm the visual findings: Model size has very limited influence on the performance of either RAG pipeline. The extremely small slopes indicate that the improvement per billion parameters is minimal and practically negligible for most metrics.

The only partial exception is Faithfulness in the proposed RAG, which shows the steepest slope (0.00480) and the strongest correlation ($r = 0.99128$). Although still small, this slope is meaningfully larger than those of the other metrics, suggesting that larger models are better at maintaining contextual grounding when combined with metadata-aware retrieval.

Overall, the results demonstrate that the proposed RAG consistently outperforms the Naive RAG across all models, regardless of size. Model size itself contributes little additional performance**, except for a modest gain in Faithfulness. The architecture of the retrieval pipeline matters far more than the parameter count, emphasizing the importance of metadata-aware design in RAG systems.

D. Effectiveness of the Proposed RAG Compared to the Naive Baseline

Before conducting the statistical validation, a qualitative visual analysis was performed to compare the overall behavior of both RAG pipelines. Fig. 8 illustrates the distribution of performance scores for the Naive RAG and the proposed RAG across all models and evaluation metrics.

The first panel aggregates all scores by pipeline, showing a clear upward shift in the distribution for the proposed RAG. The second and third panels further decompose the results by model and by pipeline, respectively, consistently revealing higher median values and reduced variability for the proposed approach. Together, these visual patterns suggest a systematic improvement in performance that motivates the subsequent statistical analyses.

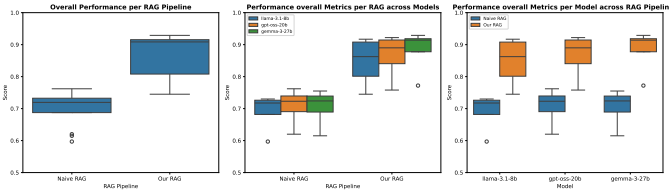

Fig. 8. Qualitative comparison of overall performance between the Naive RAG and the proposed RAG. Left: aggregated scores per pipeline. Center: performance across models. Right: performance per model across pipelines.

To statistically validate whether the proposed Retrieval-Augmented Generation (Our RAG) provides superior performance compared to the Naive RAG baseline, a series of inferential tests were conducted. These include 1) a normality assessment using the Shapiro–Wilk test, 2) a paired t-test for mean comparison, and 3) the computation of Cohen's d to quantify the magnitude of the effect.

1) Normality test: Prior to performing the parametric analysis, the assumption of normality was verified on the paired differences between the two systems. The Shapiro–Wilk test was selected due to its sensitivity in small samples ($n = 12$). The results are summarized in Table V.

TABLE V. SHAPIRO–WILK NORMALITY TEST ON PAIRED DIFFERENCES BETWEEN THE PROPOSED RAG AND NAIVE RAG

Statistic	p-value	Decision (at $\alpha = 0.05$)
0.869	0.064	Fail to reject normality

As the p-value exceeds 0.05 ($p = 0.064$), the null hypothesis of normality cannot be rejected, indicating that the paired differences are approximately normally distributed. Consequently, a paired t-test was deemed appropriate for subsequent analysis.

2) Paired t-Test: The paired t-test was applied to determine whether the mean performance of the proposed RAG was significantly higher than that of the Naive RAG across the 12 paired observations. The results are presented in Table VI.

TABLE VI. PAIRED t-TEST RESULTS COMPARING THE PROPOSED RAG AND NAIVE RAG

t-statistic	df	p-value (one-tailed)	Mean diff.	95% CI	Decision
15.49	11	4.06×10^{-9}	0.164	[0.141, 0.187]	Reject H_0

The test revealed a statistically significant improvement in performance for the proposed RAG system ($t(11) = 15.49$, $p < 0.001$). The mean difference in scores was 0.164, with a 95% confidence interval ranging from 0.141 to 0.187, confirming that the proposed approach consistently outperformed the baseline.

3) Cohen's Effect Size Analysis: To quantify the magnitude of this improvement, Cohen's d was computed based on the mean and standard deviation of the paired differences:

$$d = \frac{\bar{X}_{\text{diff}}}{s_{\text{diff}}} = \frac{0.164}{0.0367} = 4.47$$

This value ($d = 4.47$) corresponds to an *extremely large effect size*, indicating that the difference in performance is not only statistically significant but also practically meaningful. In practical terms, the proposed RAG consistently produced higher metric scores across all evaluation criteria.

The combination of a very low p-value ($< 10^{-8}$) and a large effect size ($d > 4$) provides strong evidence that the improvements achieved by the proposed RAG are not attributable to random variation. This confirms the effectiveness of the architectural modifications and the metadata-aware retrieval strategy integrated into the proposed RAG. These results validate the hypothesis that the proposed pipeline significantly enhances factuality, faithfulness, and contextual grounding compared to the Naive baseline.

VI. Discussion

The results obtained in this study provide evidence that the proposed Metadata-Aware RAG architecture substantially improves performance relative to the Naive RAG baseline under the evaluated legal question-answering setting. This controlled comparison demonstrates the benefit of incorporating hierarchical metadata-aware retrieval into a conventional RAG pipeline, while broader comparisons with alternative advanced RAG architectures remain an important direction for future evaluation.

A first point of discussion concerns the mechanisms underlying these gains. The substantial increases in Context Precision and Context Recall suggest that metadata-aware retrieval improves both the coverage and the relevance of retrieved evidence. By explicitly encoding jurisdictional markers, structural identifiers, and article-level metadata, the system reduces the incidence of semantically close but contextually irrelevant matches, while simultaneously enhancing the likelihood that all necessary fragments are retrieved. These retrieval gains translate directly into improvements in Factual Correctness and Faithfulness, confirming that more accurate and legally grounded evidence reduces opportunities for hallucinations during generation. The cross-encoder reranking module further reinforces this effect by prioritizing fragments with high semantic alignment to the query.

The regression analysis provides additional insight into the role of model size in the proposed architecture. Although larger models typically yield improved generation capabilities, the slopes of the regression lines across metrics were extremely small, revealing that parameter count only marginally influences performance. This finding underscores a central implication of our work: architectural improvements in retrieval are far more impactful than scaling model size for legal RAG. The only notable exception was Faithfulness, where larger models exhibited a modest but consistent advantage in grounding answers to retrieved evidence. Even in this case, however, the effect was minor relative to the improvements brought by metadata-aware retrieval.

Another relevant aspect concerns the practical significance of the proposed architecture for trustworthy legal assistants. The results suggest that investments in corpus engineering, including cleaning, hierarchical segmentation, metadata extraction, and structured retrieval, can provide substantial improvements in retrieval relevance, factual correctness, and grounding without requiring larger language models. In practical deployments, preserving legal metadata and provenance can also facilitate the identification and verification of the source provisions supporting an answer, which is particularly relevant in legal settings where traceability is essential. The results further suggest that improving the retrieval layer can be a more practical strategy than relying solely on model scaling, since the proposed architecture consistently improved performance across models of different sizes. These findings support the use of structured retrieval mechanisms, robust query filtering, and fine-grained provenance tracking when designing reliable legal information systems.

Several limitations should be acknowledged. The evaluation focuses on Peruvian labor law; however, the proposed retrieval mechanisms can be adapted to other legal domains by adjusting their domain-specific metadata and document structures. The dataset consists of synthetic questions with manual verification but does not include real-world user queries or adversarial prompts. Additionally, all evaluations were conducted offline; user studies and longitudinal deployment analyses were outside the scope of this work. Future research may extend the proposed architecture to multi-jurisdictional settings, incorporate temporal validity constraints, evaluate robustness to ambiguous or adversarial queries, and explore integration with active learning strategies for corpus refinement.

VII. Conclusions

This work introduced a Metadata-Aware Retrieval-Augmented Generation architecture designed to improve trustworthiness, factual consistency, and contextual grounding in legal question–answering systems. The proposed approach integrates hierarchical sub-chunking with legal metadata inheritance and an Adaptive Filter Creator capable of generating structured, metadata-guided retrieval queries. Evaluated across three language models of different scales, the system demonstrated substantial and consistent improvements over a Naive RAG baseline.

Across all RAGAS metrics (context precision, context recall, factual correctness, and faithfulness), the proposed RAG yielded double-digit gains for every model. These improvements were statistically significant, with a large effect size, confirming that metadata-aware retrieval and reranking materially enhance the quality of both retrieved evidence and generated answers. Importantly, the architecture consistently outperformed the baseline regardless of model size, indicating that retrieval design has a stronger impact on performance than increasing parameter count. Only Faithfulness showed a modest, incremental benefit from larger models.

The findings highlight the importance of corpus engineering, metadata extraction, and structured retrieval pipelines for deploying reliable, legally grounded conversational agents. By improving the relevance and coverage of retrieved fragments, the proposed method reduces hallucination risks and strengthens factual alignment, offering a practical and scalable strategy for legal information systems.

This study is limited by its focus on a single jurisdiction, an offline evaluation setting, and a curated dataset of synthetic questions. Future work should include ablation studies and benchmarking against enhanced and legal-domain RAG variants to further assess the individual contributions and state-of-the-art positioning of the proposed architecture.

In summary, the proposed Metadata-Aware RAG demonstrates that retrieval architecture, rather than model size, is the dominant factor in achieving accurate, faithful, and contextually grounded responses in legal RAG systems.

REFERENCES

[1] MM Sulphey. Development and standardization of a tool to measure knowledge of labour laws among employees. 2021.

[2] Om Prakash Gupta. *Commitment to work of industrial workers*. Concept Publishing Company, 1982.

[3] Mary Nell Trautner, Erin Hatton, and Kelly E Smith. What workers want depends: Legal knowledge and the desire for workplace change among day laborers. *Law & Policy*, 35(4):319–340, 2013.

[4] Dijana Šobota and Sonja Špiranec. Information literacy in the context of workers' rights; [informacijska pismenost u kontekstu radniČkih prava]. *Vjesnik Bibliotekara Hrvatske*, 65(3):1 – 36, 2022.

[5] Wan Faridatul Akmal binti Wan Yaacob. Employees' knowledge in employment act 1955 and level of commitment towards organization. volume 1, page 483 – 490, 2013.

[6] Galina Ognqnova Yolova, Andriyana Andreeva, Darina Nedelcheva Dimitrova, Hristina Vilhelm Blagoycheva, Plamena Nedyalkova, and Hristosko Bogdanov. *Legal and economic aspects of state control over compliance with Labor legislation*. IGI Global, 2023.

[7] Tushar Tanmay, Akanksha Bhardwaj, and Shilpi Sharma. E-health bot to change the face of medicare. In *2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH)*, pages 49–54. IEEE, 2020.

[8] Hriakumar Pallathadka, Domenic T Sanchez, Larry B Peconcillo Jr, Malik Jawarneh, Julie Anne T Godinez, and John V De Vera. Conversational chatbot-based security threats for business and educational platforms and their counter measures. *Conversational Artificial Intelligence*, pages 107–126, 2024.

[9] Rashmi P Karchi, Sanjeevakumar M Hatture, TS Tushar, and BN Prathibha. Ai-enabled sustainable development: An intelligent interactive quotes chatbot system utilizing iot and ml. In *International Conference on Sustainable Development through Machine Learning, AI and IoT*, pages 195–203. Springer, 2023.

[10] N Rakesh, Nikita Ravi, Omkar N Daivajna, and Shashank Ramesh. A proposed academic chatbot system using nlp techniques. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1300–1304. IEEE, 2022.

[11] Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. Human-aided bots. *IEEE Internet Computing*, 22(6):36–43, 2018.

[12] Xiaolin Lin, Bin Shao, and Xuequn Wang. Employees' perceptions of chatbots in b2b marketing: Affordances vs. disaffordances. *Industrial Marketing Management*, 101:45–56, 2022.

[13] Shimmy Francis and Sangeetha Rangasamy. Unveiling the transformative role of chatbots: An insight from industry. In *Marketing Intelligence, Part B: AI, Trust, and Innovation in the Modern Business Landscape*, pages 79–96. Emerald Publishing Limited, 2025.

[14] ST Nikhil, Rodrigues Sharlene, KR Shivani, C Shyni Supritha, and JA Shravani. Niles: The next-gen employee chatbot. In *AIP Conference Proceedings*, volume 2742, page 020040. AIP Publishing LLC, 2024.

[15] Quoc Nguyen Phan, Chin-Chin Tseng, Thu Thi Hoai Le, and Thi Bich Ngoc Nguyen. The application of chatbot on vietnamese misgrant workers' right protection in the implementation of new generation free trade agreements (ftas). *AI & SOCIETY*, 38(4):1771–1783, 2023.

[16] Varun Saxena, Aneesh Sathe, and S Sandosh. Mitigating hallucinations in large language models: A comprehensive survey on detection and reduction strategies. In *International Conference on Sustainable Computing and Intelligent Systems*, pages 39–52. Springer, 2024.

[17] Timothy R Hannigan, Ian P McCarthy, and André Spicer. Beware of botshit: How to manage the epistemic risks of generative chatbots. *Business Horizons*, 67(5):471–486, 2024.

[18] Cem Uluoglakci and Tugba Taskaya Temizel. Hypotermqa: Hypothetical terms dataset for benchmarking hallucination tendency of llms. *arXiv preprint arXiv:2402.16211*, 2024.

[19] SM Nazmuz Sakib. Bane and boon of hallucinations in the context of generative ai. In *Cases on AI Ethics in Business*, pages 276–299. IGI Global Scientific Publishing, 2024.

[20] Frances Flanagan and Michael Walker. How can unions use artificial intelligence to build power? the use of ai chatbots for labour organising in the us and australia. *New Technology, Work and Employment*, 36(2):159–176, 2021.

[21] Alexei A Birkun and Adhish Gautam. Large language model (llm)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehospital and Disaster Medicine*, 38(6):757–763, 2023.

[22] Surendra Yadav. Aeroquery rag and llm for aerospace query in designs, development, standards, certifications. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE, 2024.

[23] Mahimai Raja, E Yuvaraajan, et al. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1128–1133. IEEE, 2024.

[24] Kohei Arai. Design of on-premises version of rag with ai agent for framework selection together with dify and dsl as well as ollama for llm. *International Journal of Advanced Computer Science and Applications*, 15(12), 2024.

[25] Ha-Thanh Nguyen and Ken Satoh. Consrag: Minimize llm hallucinations in the legal domain.

[26] Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy. Enhancing the precision and interpretability of retrieval-augmented generation (rag) in legal technology: A survey. *IEEE Access*, 2025.

[27] Samir Amri, Saida Bani, and Rkia Bani. Moroccan legal assistant enhanced by retrieval-augmented generation technology. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security*, pages 1–5, 2024.

[28] Youssra Amazou, Faouzi Tayalati, Houssam Mensouri, Abdellah Azmani, and Monir Azmani. Accurate ai assistance in contract law using retrieval-augmented generation to advance legal technology. *International Journal of Advanced Computer Science & Applications*, 16(2), 2025.

[29] Jian Zhu, Huajun Zhang, Jianpeng Da, Hanbing Huang, Chongxin Luo, and Xu Peng. Knowledge graph path-enhanced rag for intelligent residency q&a. *International Journal of Advanced Computer Science & Applications*, 16(2), 2025.

[30] Gineke Wiggers and Christian Hartz. Lererag: Measuring legal relevance in retrieval augmented generation applications. In *Legal Knowledge and Information Systems*, pages 404–406. IOS Press, 2024.

[31] K Karthick, T Pooja, VG Oviya, J Damodharan, and S Senthamizh Selvi. Interactive legal assistance system using large language models. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 931–937. IEEE, 2024.

[32] Kurnia Muludi, Kaira Milani Fitria, Joko Triloka, et al. Retrieval-augmented generation approach: Document question answering using large language model. *International Journal of Advanced Computer Science & Applications*, 15(3), 2024.

[33] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199, 2024.

[34] Tilmann Bruckhaus. Rag does not work for enterprises. *arXiv preprint arXiv:2406.04369*, 2024.

[35] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.