

A Retrieval-Augmented Generation Chatbot for Accessible, Hallucination-Resistant Access to the Constitution of Nepal

Sarina Nepal

Dept. of Software Engineering

Nepal College of Information Technology Nepal

Pokhara University

Kathmandu, Nepal

sarina.221741@ncit.edu.np

Sonu Giri

Dept. of Software Engineering

Nepal College of Information Technology Nepal

Pokhara University

Kathmandu, Nepal

sonu.221743@ncit.edu.np

Sneha Mahat

Dept. of Software Engineering

Nepal College of Information Technology Nepal

Pokhara University

Kathmandu, Nepal

sneha.221742@ncit.edu.np

Astha Thapa

Dept. of Software Engineering

Nepal College of Information Technology

Pokhara University

Kathmandu, Nepal

astha.221707@ncit.edu.np

Manil Vaidhya

Dept. of Software Engineering

Nepal College of Information Technology

Pokhara University

Kathmandu, Nepal

manil.baidhya@ncit.edu.np

Abstract—The constitution and laws of Nepal remain limited to static PDF documents, leaving users to interpret dense legal language largely unaided. This paper presents ArticleWise, a retrieval-augmented generation (RAG) system for question answering over the Constitution of Nepal (2015/2072) that is designed to be accessible and hallucination-resistant. The system combines a deterministic citation router that provides $O(1)$ lookup of statutes with a dense semantic search pipeline: the full set of 308 numbered constitutional provisions was extracted through layout-aware parsing, represented in a 1024-dimensional semantic space with multilingual-e5-large and asymmetric task prefixes, and indexed in PostgreSQL with pgvector using an exact flat nearest-neighbor scan. The top-ranked article is supplied to Google Gemini 3.5 Flash-Lite for a succinct, plain-language explanation generated under closed-domain prompting, while the statutory text itself is always returned verbatim from an immutable database record and never passed through the generative model. On a stratified benchmark of $N = 50$ questions spanning paraphrases, citizen phrasing, underspecified keywords, and multi-article workflow queries, the system achieves Top-1 retrieval accuracy of 62.0% (95% CI 48.2–74.1%) and Top-3 accuracy of 84.0% (95% CI 71.5–91.7%), with median retrieval-only latency of 72.2 ms and average end-to-end latency of 372.5 ms. A faithfulness evaluation shows a 0.0% hallucination rate relative to whatever article is retrieved, with unanimous inter-annotator agreement — a result that speaks to the fidelity of the explanation step but not to the correctness of article selection, a distinction we treat as a first-class limitation of the system rather than a footnote to the headline result.

Keywords—Retrieval-Augmented Generation, Legal Informatics, Natural Language Processing, Vector Search, Large Language Models, Constitution of Nepal, Accessible Legal Information

I. INTRODUCTION

A. Background and Motivation

The Constitution of Nepal, promulgated in 2072 BS (2015 CE), is the foundational legal document governing the rights, duties, and structure of the Nepali state. It comprises 35 parts, 308 articles, and 9 schedules, covering matters ranging from fundamental rights to the federal structure of government. For most citizens, engaging directly with this text means navigating dense legal language distributed across a 240-page PDF or printed volume, with no built-in way to ask a question and receive a direct, comprehensible answer. Retrieval-augmented generation has emerged as the standard architecture for building question-answering systems over a fixed, authoritative corpus precisely because it separates two functions a language model performs poorly when combined: recalling exact facts, and producing fluent natural-language explanation. Rather than relying on a model's internal parametric memory, which is known to hallucinate specific facts, citations, and numeric details, a RAG system retrieves the relevant source text at query time and supplies it directly to the model as context, restricting generation to only what is grounded in that retrieved evidence.

The motivation for this project is twofold. First, there is a genuine accessibility gap: constitutional literacy in Nepal is limited not by lack of interest but by the practical difficulty of searching and interpreting a dense legal document. Second, this project serves as a case study in responsible deployment of generative AI in a high-stakes information domain, demonstrating an architecture where the language model's role is deliberately constrained to explanation rather than fact

generation — a distinction this paper argues should be a design principle for any AI system operating over authoritative source documents, legal or otherwise.

B. Existing Access Approaches and Their Gaps

Before describing ArticleWise’s design, it is useful to situate it against how citizens currently attempt to answer constitutional questions, since each existing approach fails in a different way that motivates a specific architectural choice later in this paper.

Manual search of the source PDF requires the reader to already know, or correctly guess, which part or article is relevant, and offers no natural-language entry point at all. Direct prompting of a general-purpose LLM removes that barrier but reintroduces a worse one: without retrieval grounding, such models cannot be trusted to reproduce article text or numbering correctly, since their training data may be incomplete, outdated, or simply insufficient to reproduce a specific legal document verbatim. Generic legal chatbots built on top of retrieval typically close part of this gap — they support natural-language queries and semantic search — but frequently blur the boundary between retrieved statutory text and generated explanation, so that a user cannot easily tell which words are the law’s and which are the model’s. Table I summarizes this comparison.

TABLE I
CONCEPTUAL COMPARISON OF LEGAL ACCESS APPROACHES

Approach	NL Query	Exact Text	Explanation	Semantic Search
Manual PDF Search	No	Yes	No	No
Direct LLM Prompting	Yes	No	Yes	Implicit
Generic Legal Chatbot	Yes	Partial	Yes	Yes
ArticleWise	Yes	Yes	Yes	Yes

The gap ArticleWise targets, then, is not natural-language access in isolation and not exact-text fidelity in isolation, but the combination of both guaranteed simultaneously and architecturally — not merely promised in a system prompt — for a formally numbered, hierarchically structured legal text extracted from an official but unstructured source PDF.

C. Problem Statement

Two related problems motivate this work. First, citizens and students seeking to understand a specific constitutional provision have no tool to query the constitution in natural language; they must know, or guess, which article addresses their question and then locate and interpret it manually. Second, general-purpose large language models, when asked directly about constitutional provisions without retrieval grounding, cannot be trusted to return accurate article text or numbering. There is a need for a system that combines the natural-language accessibility of a conversational AI system with the factual guarantees of a direct database lookup.

D. Objectives

- 1) **Structured Extraction and Normalization of the Corpus:** extract, clean, and structure the statutory English-language version of the Constitution of Nepal from PDF source documents into a standardized relational database of $N = 308$ Articles with hierarchy and ordering of clauses preserved.
- 2) **Dual-Path Retrieval System:** implement a hybrid retrieval architecture combining an $O(1)$ regular-expression router for explicit statutory citations with dense vector search over multilingual-e5-large embeddings in pgvector, so that untrained citizens can find relevant articles without knowing article numbers.
- 3) **Grounded, Hallucination-Resistant Legal Synthesis:** generate plain-English legal explanations through closed-domain prompting with Google Gemini 3.5 Flash-Lite based only on the retrieved statutory text, while keeping the generated explanation architecturally separate from the actual legal canon, which is always furnished verbatim from the database.
- 4) **Benchmarking for Empirical Faithfulness Evaluation:** construct a comprehensive evaluation suite ($N = 50$) spanning formal paraphrasing and citizen-constructed questions across multiple articles, and empirically characterize both retrieval accuracy and generation faithfulness against ungrounded baselines.

E. Significance and Contributions

At present, identifying the constitutional provision relevant to a particular legal question depends significantly on a person’s ability to comprehend and analyze complex legal language — a gradual, mistake-prone procedure largely unreachable for individuals lacking formal legal education. ArticleWise combines dense semantic retrieval, precise database queries, and constrained LLM generation to offer citizens, researchers, and legal professionals verified, easy-to-understand access to constitutional provisions. In contrast to earlier RAG demonstrations in classrooms or generic domains, this system operates on dense legal text with a formally numbered hierarchical structure (parts, articles, sub-clauses), using readily available pretrained embeddings without fine-tuning, and tackles parsing and retrieval challenges that are largely absent in less-structured text domains. Beyond its direct use, this work illustrates a design pattern — precise database retrieval paired with LLM-based explanation — applicable to any field where factual precision is critical for conversational coherence, including technical documentation, medical reference data, and regulatory compliance.

II. RELATED WORK

A. Retrieval-Augmented Generation

Lewis *et al.* [1] introduced retrieval-augmented generation as a method for combining a parametric language model with a non-parametric retrieval mechanism over an external corpus, allowing a model’s output to be grounded in retrieved documents rather than relying solely on knowledge encoded

during training. This approach has since become the dominant architecture for domain-specific question-answering systems, as it allows a system’s knowledge to be updated by changing the underlying document corpus rather than retraining the model, and, critically for legal applications, allows the exact source text to be surfaced alongside any generated explanation. Gao *et al.* [2] surveyed the rapid expansion of RAG architectures and noted that grounding generation in retrieved evidence remains the most effective mitigation for factual hallucination in knowledge-intensive tasks, though retrieval quality itself becomes the new bottleneck on overall system accuracy — an observation directly reflected in ArticleWise’s own results, where generation faithfulness is high but end-user correctness is bounded by Top-1 retrieval accuracy (Section IV, Section VI).

B. Legal Question-Answering Systems

Martinez-Gil [3] surveyed legal question-answering systems and identified two related but distinct tasks: statutory retrieval, which identifies the relevant law or provision for a given query, and legal text simplification, which rewrites dense legal language into more accessible prose. Systems that combine these tasks using generative models have consistently identified hallucination as the primary obstacle to reliable deployment, motivating architectures that separate retrieval from generation rather than relying on a language model’s internal knowledge of legal text.

Beyond Nepal, the Competition on Legal Information Extraction/Entailment (COLIEE) has for over a decade provided a standard benchmark for statute-law retrieval and entailment over the Japanese Civil Code and Canadian case law, requiring systems to retrieve the subset of articles that supports a yes/no bar-exam-style legal question [8]. COLIEE’s statute-retrieval task shares ArticleWise’s core structure — natural-language query against a fixed, numbered set of statutory articles — but differs in that a nontrivial share of its queries require multiple supporting articles rather than a single best match, a multi-article condition ArticleWise does not yet handle (Section VI). Closer to the present setting, Wagle *et al.* [9] built the first RAG pipeline for Nepali legal question answering, retrieving case law from the Nepal Kanun Patrika archive with BM25 and multilingual-e5-large and evaluating groundedness with both an automated judge model and human annotators; their reported top-1 precision (91% with BM25, 75% with e5-large) is substantially higher than ArticleWise’s 62%, a gap plausibly attributable to differences in corpus size, query difficulty, and domain (case law versus constitutional text) that is worth investigating in future comparative work rather than assumed away.

C. Multilingual Embedding Models and Nepali NLP Resources

Wang *et al.* [4] introduced the multilingual E5 family of text embedding models, trained across a large number of languages within a shared representation space, and demonstrated

strong transfer performance on multilingual retrieval benchmarks without requiring domain-specific fine-tuning. General-purpose multilingual embedding models of this kind transfer reasonably well to specialized domains such as legal text, though at some cost to retrieval precision compared to domain-adapted models — a trade-off discussed further in Section VI.

Because ArticleWise’s design implicitly relies on this multilingual transfer to eventually support Nepali-language (Devanagari) queries, it is important to note that Nepali remains a comparatively low-resource setting for retrieval evaluation specifically. Acharya and Bal [10] observed that Nepali still lacks a standardized IR benchmark and that reported cross-lingual transfer results for general-purpose multilingual encoders vary considerably by task and domain. ArticleWise’s own corpus and benchmark are English-only (Sections III and IV); we treat multilingual-e5-large’s Nepali retrieval capability for this specific legal corpus as an untested assumption rather than a validated claim, and revisit this explicitly as a limitation in Section VI.

D. Vector-Capable Relational Databases

As retrieval-augmented systems have moved from research prototypes to production systems, vector-capable extensions to conventional relational databases, such as pgvector [5] for PostgreSQL, have gained adoption as an alternative to dedicated vector database systems. This approach offers the benefit of combining structured metadata queries, such as exact lookup by article number, with approximate or exact nearest-neighbour vector search within a single, familiar database system, which is precisely the dual-retrieval requirement of this project.

E. RAG Evaluation and Faithfulness Metrics

A separate line of work addresses how RAG outputs should be evaluated in the first place. Es *et al.* [7] introduced RAGAS, a reference-free evaluation framework that scores RAG pipelines along dimensions including context relevance, answer relevance, and faithfulness — the extent to which a generated answer is supported by its retrieved context — without requiring hand-labeled ground truth. RAGAS and related frameworks make an evaluative distinction this paper leans on directly: faithfulness (is the explanation consistent with what was retrieved?) and correctness (was the right thing retrieved in the first place?) are separate axes, and a system can score perfectly on the former while still failing the user on the latter. ArticleWise’s 0.0% hallucination rate (Table III) is a faithfulness result in this sense; Sections IV and VI make explicit why it does not, by itself, establish end-to-end correctness.

F. Research Gaps and Positioning of ArticleWise

Across the reviewed literature, prior RAG deployments have generally been demonstrated on either generic, loosely structured document corpora or on legal corpora from jurisdictions with existing structured legal databases and larger benchmark communities (COLIEE, statutory retrieval in French and other

languages). A gap remains for legal RAG systems applied to formally numbered, hierarchically structured legal text — such as a national constitution — extracted directly from an official but unstructured source PDF, where the parsing step itself introduces domain-specific challenges not addressed in the general RAG literature, and for such systems built for a specifically low-resource language setting where even the underlying embedding model’s effectiveness has not been locally validated. ArticleWise addresses the first gap directly; it does not yet close the second, and we say so plainly rather than implying multilingual readiness that has not been measured.

III. METHODOLOGY

A. System Architecture Overview

ArticleWise comprises five modules organized as a linear pipeline: (1) a Data Parsing Module that extracts structured article-level records (part number, part title, article number, article title, body text) from the official Constitution of Nepal PDF; (2) an Embedding Module that generates a dense vector representation for each article using the pretrained multilingual-e5-large sentence embedding model; (3) a Storage and Retrieval Layer, a PostgreSQL database with the pgvector extension supporting both exact lookup by article number and nearest-neighbour semantic search; (4) an Explanation Module that generates plain-English summaries using Google Gemini 3.5 Flash-Lite (*gemini-3.5-flash-lite*) via the Google GenAI SDK, receiving only the single top-ranked article ($k = 1$) and constrained to a 3–4 sentence explanation grounded strictly in that text (full prompt in Appendix ??); and (5) an API and UI Layer, a FastAPI backend linking the retrieval and explanation pipelines to a responsive React web chat interface.

B. Dataset Description

The primary text is the official English translation of the Constitution of Nepal (2015) from the Nepal Law Commission and Ministry of Law, Justice and Parliamentary Affairs [6]. The Constitution is structurally divided into a Preamble, 35 Parts, 308 Articles, and 9 Schedules; this work covers the Preamble and all 308 Articles. The 9 Schedules are excluded because they are presented in tabular format; their inclusion is discussed as future work in Section VI.

C. Data Extraction and Preprocessing

Text extraction was performed using layout-aware parsing, since standard linear extraction failed due to the hanging-indent formatting of the source PDF, where article numbers appear in the margin separate from clause bodies. The rule-based parser identified statutory boundaries by detecting left-aligned top-level article headings while preserving indented sub-clauses (e.g., (1), (2), and lettered sub-clauses). Running footers, headers, and pagination artifacts were filtered out prior to parsing. The resulting corpus consists of 309 structured records: the Preamble and all 308 Articles. Article numbers were confirmed to run sequentially from 1 to 308 without gaps

or duplicates, and text fidelity checks confirmed an exact word-count match with the raw extracted text, with only a 0.8% variance attributable to the removal of redundant structural headers and formatting artifacts.

D. Embedding Generation

The title and body content of each article were concatenated and encoded using the pretrained multilingual-e5-large sentence embedding model [4] to generate a dense 1024-dimensional vector representation for each constitutional provision. In accordance with the mandatory input format prescribed for asymmetric retrieval with E5 models, each stored article chunk was explicitly prepended with the "passage: " prefix prior to embedding, while incoming user queries were formatted with the "query: " prefix. The selection of multilingual-e5-large was guided by its strong zero-shot retrieval performance across benchmark datasets and its published ability to map many languages into a shared vector space without task-specific fine-tuning — a property this paper relies on as a design rationale but, as noted in Section II-C and revisited in Section VI, has not itself been empirically validated on this Nepali constitutional corpus.

E. Storage and Retrieval Architecture

Articles and their associated dense embeddings are indexed in a PostgreSQL database using the pgvector extension [5], which provides a native 1024-dimensional vector data type and optimized vector distance operators. Because the constitutional corpus is strictly bounded ($N = 308$), vector retrieval is executed via an exact flat nearest-neighbor scan rather than approximate nearest-neighbor indexing (e.g., IVFFlat or HNSW). At this corpus scale, exact sequential vector scanning completes with negligible latency (< 25 ms) and avoiding the indexing distortions or recall degradation inherent in approximate clustering methods.

For an incoming user query q , the query vector $\mathbf{e}_q$ is computed with the identical encoder, and semantic relevance to article i is quantified via cosine similarity:

$$\text{sim}(q, i) = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{\|\mathbf{e}_q\| \|\mathbf{e}_i\|} \quad (1)$$

The semantic retrieval module returns the top- k highest-ranking statutory articles:

$$D_k(q) = \arg \text{top-}k \text{ sim}(q, i)_{i \in \{1, \dots, 308\}} \quad (2)$$

ArticleWise supports two complementary retrieval pathways. *Deterministic exact-match lookup* uses a regular-expression router that intercepts explicit statutory citations (e.g., “Article 20”, “Art. 76”) and executes an $O(1)$ indexed SQL query against the primary key, bypassing both the embedding model and the language model and returning verbatim statutory text with sub-10 ms latency. *Dense semantic search* is used for conceptual, scenario-based, or natural-language queries, computing $D_k(q)$ over the exact vector space and returning ranked constitutional provisions by semantic similarity.

F. Grounded Generation

During the generation phase, the title and full text of the top-ranked retrieved article $D_1(q)$ are injected into the prompt context for the language model, which is constrained via prompt instructions to explain the provision in plain English based solely on the retrieved text, without incorporating external or ungrounded assertions. Formally, rather than sampling an explanation y from the unconditioned distribution $P(y | q)$, generation is conditioned directly on the retrieved context:

$$y \sim P(y | q, D_1(q)) \quad (3)$$

where $D_1(q)$ denotes the highest-ranking article for query q . This constraint is what the faithfulness result in Section IVD measures: the model is never tasked with recalling or generating legal facts from parametric memory, but is restricted to paraphrasing the provided statutory passage. The verbatim constitutional text returned to the user is never passed through the language model; it is retrieved and served directly from the database record as $D_1(q)$.

G. Evaluation Protocol

Retrieval effectiveness was evaluated on a stratified test set of 50 natural-language questions covering fundamental rights, the governmental system, constitutional bodies, and procedural provisions, with each query mapped to its true article number by the authors during dataset construction. For each question, the top-3 closest articles were retrieved via dense vector search, and performance was measured by whether the correct article appeared as the first result (Top-1) or within the top three (Top-3). A parallel faithfulness evaluation compared ArticleWise’s grounded explanations against an ungrounded direct-LLM baseline on the same 50 queries, with each generated explanation rated for factual consistency against its retrieved (or, for the baseline, assumed) source article.

IV. RESULTS AND FINDINGS

A. Parsing Accuracy

The parsing pipeline retrieved all 308 articles without missing or duplicating any article numbers, verified programmatically by checking that article numbers fell within the expected sequential range. Manual checks on randomly selected articles — including the first and last articles and articles from different parts — confirmed accurate boundary detection.

B. Retrieval Evaluation

Table II reports Top-1 and Top-3 retrieval accuracy on the $N = 50$ benchmark, together with 95% confidence intervals computed using the Wilson score interval for a binomial proportion ($z = 1.96$), applied to the reported hit counts (31/50 for Top-1, 42/50 for Top-3).

For example, the natural-language query “what does it say about right to live with dignity” — which contains no article numbers or formal legal terms — correctly returned Article 16 (“Right to live with dignity”) as the top result with a cosine similarity score of 0.892; Article 37 (“Right to housing”)

TABLE II
RETRIEVAL EVALUATION RESULTS ($N = 50$)

Metric	Value	95% CI (Wilson)
Top-1 accuracy	62.0% (31/50)	48.2%–74.1%
Top-3 accuracy	84.0% (42/50)	71.5%–91.7%
Median retrieval-only latency	72.2 ms	—
Avg. end-to-end latency (incl. gen.)	372.5 ms	—

and Article 30 (“Right relating to clean environment”) were returned second and third. This demonstrates that semantic search successfully identifies the relevant provision based on meaning rather than direct keyword matching, though the wide confidence interval on Top-1 accuracy (48.2–74.1%) is itself informative: at $N = 50$, the point estimate of 62.0% carries substantial uncertainty, and the true population Top-1 accuracy could plausibly be anywhere in roughly that range.

C. Generation Quality

Review of the explanations produced for the test queries showed that they were consistent with the retrieved source article and did not make unsupported statements, adhering to the grounding constraint in the prompt. For the Article 16 example above, the model accurately elaborated on both the right to dignity and the prohibition of the death penalty in the article’s second clause.

D. Faithfulness and Hallucination Evaluation

Table III compares ArticleWise’s grounded generation against a direct-LLM (no-RAG) baseline on the same 50 queries.

TABLE III
EMPIRICAL FAITHFULNESS AND HALLUCINATION EVALUATION

Approach	Faithful	Hallucinations	Agreement	κ
Direct LLM (No RAG)	30.0%	35/50 (70.0%)	90.0%	0.780
ArticleWise (Proposed)	100.0%	0/50 (0.0%)	100.0%	N/A*

* Undefined: both annotators assigned the same label to all 50 items, so observed and chance agreement are equal.

It is important to state explicitly what this table does and does not show. “Faithful” here means the generated explanation is consistent with whatever article was retrieved and presented to the language model — it is a measure of the generation step alone. It says nothing about whether the retrieved article was the correct one for the user’s underlying question. Because Top-1 retrieval accuracy is 62.0% (Section IVB), roughly 38% of queries in this benchmark would, in practice, receive a fluent, well-grounded, zero-hallucination explanation of the *wrong* constitutional article. That is a

user-facing correctness failure that a 0.0% hallucination rate cannot, by construction, detect, and we consider it the single most important qualification on this paper’s headline results (elaborated further in Section VI).

E. System Latency

Latency in the retrieval-and-generation pipeline is dominated by the generation step, as expected for an API call to a large language model; exact article lookup, which requires neither embedding nor generation, has negligible latency. The reported end-to-end average of 372.5 ms substantially exceeds the median retrieval-only latency of 72.2 ms, consistent with generation dominating overall latency.

F. Comparative Model Evaluation

Table IV situates ArticleWise’s empirical retrieval performance against a sparse lexical baseline (BM25) and an ungrounded direct-LLM approach.

TABLE IV
EMPIRICAL COMPARISON OF RETRIEVAL APPROACHES ON
CONSTITUTIONAL BENCHMARK

Approach			Hit@1	MRR@3	Latency
Direct LLM (No RAG)	(No	–	–	–	~800.0 ms
BM25 (Sparse Base-line)			60.0%	0.703	0.6 ms
ArticleWise (Hybrid)			62.0%	0.717	72.2 ms

G. Discussion

Taken together, these results support the paper’s core design principle: separating what the constitution says (database query precision) from what it means (LLM generation) lets the system remain conversationally accessible without compromising the factuality of the underlying legal text. The Article 16 case study shows this separation does not come at the cost of query precision for well-matched examples. At the same time, the aggregate Top-1 accuracy and its confidence interval (Section IVB) show that this precision is not yet uniform across the benchmark, and Section IVD’s discussion of faithfulness-versus-correctness should be read as this paper’s central empirical caveat rather than an incidental one.

V. CONCLUSION

This paper introduced ArticleWise, a domain-specific retrieval-augmented generation architecture for accessible, hallucination-resistant public interaction with the Constitution of Nepal (2015/2072). By converting a layout-rich legal PDF into a structured corpus of 308 canonical articles, the system combines a deterministic regular-expression router for precise citations with dense vector retrieval using multilingual-e5-large and exact flat scanning in pgvector, while plain-language descriptions are produced by Google Gemini 3.5 Flash-Lite under constrained closed-domain prompting and statutory text is always served verbatim from an immutable database. On

a stratified $N = 50$ benchmark, the system achieves 62.0% Top-1 (95% CI 48.2–74.1%) and 84.0% Top-3 (95% CI 71.5–91.7%) retrieval accuracy, a median retrieval-only latency of 72.2 ms, and an average end-to-end latency of 372.5 ms including generation. A faithfulness evaluation shows a 0.0% hallucination rate relative to retrieved context, with unanimous annotator agreement, against a 70.0% hallucination rate for an ungrounded baseline.

We emphasize, as the central conclusion of this paper rather than a caveat appended to it, that these two headline results answer different questions. Zero-hallucination generation demonstrates that ArticleWise does not invent statutory content once an article has been retrieved; it does not demonstrate that the correct article is retrieved in the first place, which happens for roughly six in ten queries at Top-1. A hallucination-resistant system that is confidently wrong about which article applies is still, from a citizen’s perspective, an unreliable one. The architecture presented here is a defensible foundation for closing that gap — because retrieval and generation are already cleanly separated, retrieval accuracy can be improved (through re-ranking, a domain-adapted embedding model, or multi-article fusion; Section VI) without touching the generation guarantees this paper establishes.

VI. LIMITATIONS AND FUTURE RECOMMENDATIONS

A. Current Limitations

1) *English-Only Corpus and Excluded Schedules*: The system currently includes only the English text of the 308 articles. It excludes the 9 Schedules, which detail the division of legislative powers among federal, provincial, and local governments, because they are presented in tabular form and require specialized table extraction.

2) *General-Purpose Embedding Model*: ArticleWise uses a general multilingual embedding model (multilingual-e5-large) rather than a domain-adapted legal model. This can make it harder for the search to distinguish between semantically similar but legally distinct provisions, and is a plausible contributor to the 62.0% Top-1 accuracy ceiling.

3) *Single-Article Retrieval Grounds a Correctness Ceiling on the Whole System*: This is the limitation we consider most consequential, and we state it without softening: the generation module conditions on only the single top-ranked article ($k = 1$). At 62% Top-1 retrieval accuracy, this means that for roughly 38% of queries in the evaluated benchmark, the system produces a fluent, non-hallucinated explanation of the wrong constitutional article — a user-facing correctness failure that the 0.0% hallucination metric in Table III cannot see, because that metric only measures faithfulness to whichever article was retrieved, never whether that article was the right one. A citizen using this system therefore cannot distinguish, from the output alone, a correct grounded answer from a confidently wrong one. Any deployment of this architecture beyond a research prototype should treat retrieval accuracy, not hallucination rate, as the primary safety metric, and should consider surfacing retrieval confidence (e.g., the cosine similarity score, or the margin between the top result

and the runner-up) to the user rather than presenting every answer with uniform confidence.

4) *Evaluation Methodology and Sample Size*: The $N = 50$ benchmark, while stratified across fundamental rights, governmental structure, constitutional bodies, and procedural provisions, is small enough that the Top-1 and Top-3 accuracy estimates carry wide confidence intervals (Section IVB). The annotator methodology behind the faithfulness results in Table III — annotator count, blinding, and rubric — is not yet documented in this paper (flagged in Section IIIG) and should be reported before the 100% faithfulness figure is treated as conclusive. Larger benchmarks incorporating real citizen queries, adversarial phrasing, and harder multi-article cases are needed before real-world deployment.

B. Future Recommendations

1) *Incorporation of Tabular Schedules*: The nine constitutional schedules, currently excluded due to their tabular structure (e.g., exclusive and concurrent legislative jurisdiction lists across federal, provincial, and local tiers), will be incorporated using specialized table-parsing and schema-aware chunking techniques, expanding the corpus beyond the 308 articles covered here.

2) *Cross-Lingual (Devanagari) Retrieval*: This paper’s design rationale for multilingual-e5-large (Sections II-C and IIIE) rests on its published multilingual training, but ArticleWise’s benchmark and corpus are English-only; Nepali-language (Devanagari) retrieval has not been empirically evaluated in this work. We therefore retract, rather than repeat, any implication that the system is currently capable of native Nepali-language querying, and treat this as an open empirical question rather than an assumed capability. The concrete next step is a small probe evaluation — even 10 to 15 hand-translated Devanagari versions of benchmark queries, scored for Top-1/Top-3 accuracy against the same 308-article corpus — to establish whether multilingual-e5-large’s cross-lingual transfer holds for this specific legal domain and script before any Nepali-facing claim is made in a future version of this paper.

3) *Multi-Article Context Aggregation and Fusion*: Complex constitutional inquiries frequently span multiple interrelated provisions — for example, presidential discretionary powers intersecting with prime ministerial appointment procedures, or questions about constitutional amendment that implicate several articles at once. Because generation currently conditions on only the single top-ranked article ($k = 1$; Section IIIF), such questions cannot be fully answered in one step even when all the relevant articles are individually retrievable. Addressing this requires two related extensions: (i) raising k and applying a cross-encoder or LLM-based re-ranker over the resulting candidate set to select a coherent subset of articles rather than a single best match, and (ii) a context-fusion step in the generation prompt that explicitly synthesizes an explanation spanning multiple retrieved articles while preserving per-article attribution, so the user can still tell which explanation text traces back to which article. This also compounds with the correctness-ceiling limitation above: multi-article fusion done

carelessly could mask retrieval errors behind a longer, more plausible-sounding explanation, so any re-ranking and fusion mechanism should be evaluated for faithfulness per-article, not just as an aggregate response.

4) *Empirical Evaluation by Users*: Structured usability and trust evaluations with law students, legal professionals, and non-expert citizens are needed to assess explanation readability, cognitive load, user trust calibration (whether users correctly judge when to double-check an answer), and end-to-end task completion in realistic deployment settings — particularly given the correctness-ceiling limitation discussed above.

C. Privacy, Ethics, and Responsible Deployment

Using AI for legal information requires strong ethical safeguards, since incorrect or fabricated answers can mislead users. ArticleWise follows several key principles: explanations are generated strictly from retrieved legal text, with the model never permitted to add outside information, and the original legal text is always fetched directly from the database without AI modification; the system is positioned as an educational tool, not a substitute for legal advice, with a disclaimer directing users with real legal issues to a qualified lawyer; every response cites the exact Part, Article number, and Article title so users can cross-check against the official constitution; the system does not collect, record, or store personal user information or private chat history, and any anonymized logs kept for performance testing contain no personal data; and because constitutional text can be amended, the current system reflects the text available at the time of data collection, with an automated update pipeline needed for long-term public use to keep it synchronized with the Nepal Law Commission’s official text.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, 2023.
- [3] J. Martinez-Gil, “A survey on legal question answering systems,” *Computer Science Review*, vol. 48, p. 100552, 2023.
- [4] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, “Multilingual E5 text embeddings: A technical report,” *arXiv preprint arXiv:2402.05672*, 2024.
- [5] A. Kane, “pgvector: Open-source vector similarity search for Postgres,” 2024. [Online]. Available: <https://github.com/pgvector/pgvector>
- [6] Government of Nepal, *The Constitution of Nepal*, Ministry of Law, Justice and Parliamentary Affairs, 2015. [Online]. Available: <https://www.lawcommission.gov.np>
- [7] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAs: Automated evaluation of retrieval augmented generation,” *arXiv preprint arXiv:2309.15217*, 2023.
- [8] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka, “Overview of benchmark datasets and methods for the legal information extraction/entailment competition (COLIEE) 2024,” in *New Frontiers in Artificial Intelligence (JSAL-isAI 2024)*, Lecture Notes in Computer Science, vol. 14741. Singapore: Springer, 2024, pp. 109–124.

- [9] S. Wagle, A. Adhikari, R. Khanal, B. Bhandari, P. Manandhar, P. Acharya, and B. K. Bal, "Retrieval augmented generation framework for the Nepali legal domain question answering," in *Proc. 8th Int. Conf. Natural Language Processing (ICNLP)*, 2026, pp. 583–592.
- [10] P. Acharya and B. K. Bal, "Towards building a standard benchmark for low-resource Nepali information retrieval," in *Proc. 49th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2026, pp. 5256–5258.
- [11] S. Ru, M. Wang, C. Xu, L. Wang, and J. Zhao, "RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation," *arXiv preprint arXiv:2408.08067*, 2024.
- [12] Y. Li, J. Zhang, and H. Chen, "A reasoning-focused legal retrieval benchmark," *arXiv preprint arXiv:2505.03970*, 2025.