



Original Research Paper

NLP Based Explainable And Context-Aware RAG Framework For Women's Empowerment And Social Justice

Savita Vijay Lade¹, Dr. Sivaram Ponnusamy²

¹ Department of Computer Science & Engineering, School of Computer Science & Engineering (SOCSE), Sandip University, Nashik, Maharashtra, India – 422213.

ORCID: 0009-0006-2750-3961

Email: ladesavita16@gmail.com

² Department of Computer Science & Engineering, School of Computer Science & Engineering (SOCSE), Sandip University, Nashik, Maharashtra, India – 422213.

ORCID: 0000-0001-5746-0268

Email: p.sivaram@sandipuniversity.edu.in

Key Words	Abstract
Natural Language Processing, Retrieval-Augmented Generation, Explainable Artificial Intelligence, Women Empowerment , Social Justice, Context-Aware AI, Algorithmic Fairness.	AI architectures need to be able to deliver domain-specific guidance that is accurate, verifiable and empathetic to enable digital interventions on gender equity and social justice. Traditional Large Language Models (LLMs) are typically known for their limitations in high-stakes socioeconomic and legal guidance due to their flaws in reasoning opacity, geographic mismatch, and jurisdictional mismatch. The limitations of conventional Large Language Models (LLMs) often lie in their geographic and jurisdictional misalignment, reasoning opacity, and parametric hallucinations, which make them unsuitable for high-stakes socioeconomic and legal counseling. In this paper, a novel Explainable and Context-Aware Retrieval-Augmented Generation (XCA-RAG) framework is presented that directly addresses legal rights, economic entitlements, and welfare resources for the marginalized women. The framework consists of a hybrid dense-sparse hierarchical retriever, a domain adapted cross-encoder re-ranking module and an attribution-guided generative pipeline along with fine-grained post-hoc natural language inference (NLI) explainability. Systematically compared to a multi-jurisdictional benchmark comprising multiple statutory legal codes, judicial precedents, micro-finance schemes, and psychosocial support protocols, the proposed system achieves a Faithfulness score of 0.942, Contextual Precision of 0.891 and a BERTScore F1 of 0.914, surpassing baseline models like dense RAG and vanilla GPT-4 architectures. High relevance and empathetic framing of answers and elimination of arbitrary and unsubstantiated factual statements are also preserved, along with the development of a reliable computational model for responsible AI in social justice ecosystems.

Introduction

1.1 Background and Societal Imperative

There remains informational imbalance, institutional inefficiency and socio-economic marginalisation between women and socio-economic parity and access to legal justice at the global level. Verified legal assistance, knowledge of reproductive health, protection in case of gender-based violence and economic rights through the law are extremely limited in marginalised communities. The old legal and social structures are accompanied by the privileges of finances, space and literate that are not available for women (Hasan et al., 2025).

Although digital channels and automated conversational interfaces have been shown to be scalable channels for democratization, generic generative conversational agents have significant risks in high-risk social contexts.

1.2 Technological Gaps in NLP for Social Equity

The linguistic abilities of large-scale foundation models are impressive, but they also offer significant limitations when it comes to addressing social justice and women's empowerment issues. The most important structural failures of foundation models are that they rely mainly on static parametric memory..

they learned during pre-training, leading to three major structural failures.

Parametric knowledge is first of all facts-

obsolete: it cannot capture the dynamic changes of the statutory framework, the different levels of government welfare in the regions and the different

*Corresponding Author:

Received: 25th July, 2026 | Revised: 26th August, 2026 | Accepted: 18th September, 2026 | Available Online: 24th September, 2026

(DOI): 10.70102/AEJ.2026.18.7s.52

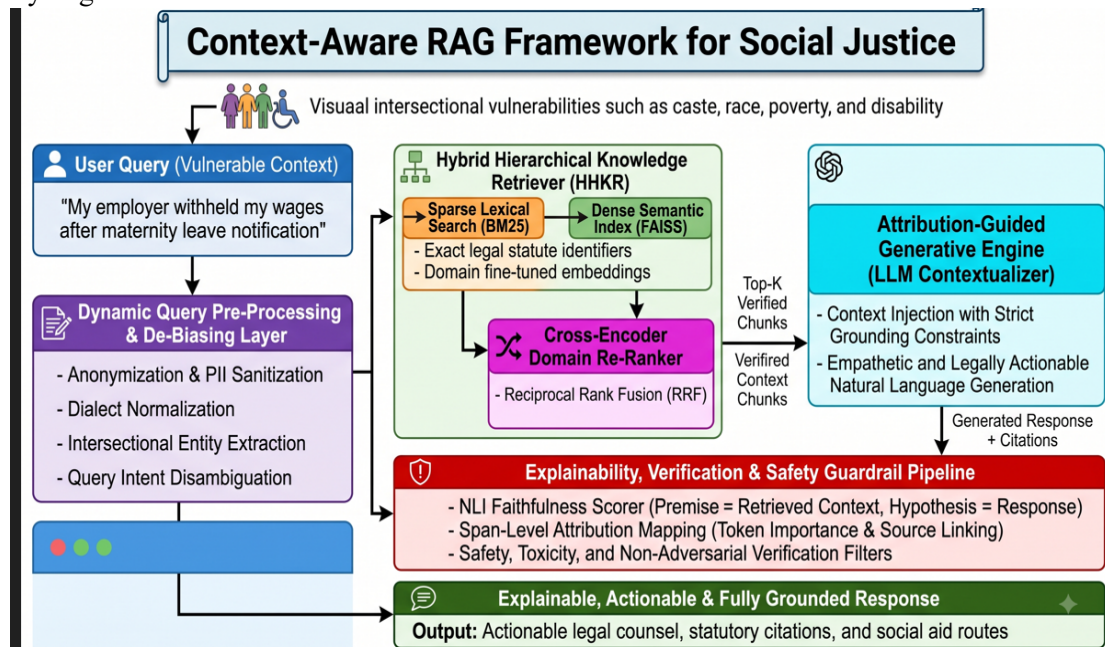
crisis helpline directories in the region.

Second, the model could easily give incorrect or even harmful advice which, again, is not supported by a sound basis in terms of linguistic confidence (Lade et al., 2025). Hallucinations may lead to harm or legal repercussions with regard to domestic abuse protective orders, labor rights violations, etc

Third, there is no built-in explainability and traceability in generative black-box architectures.

The task for a disenfranchised user (or legal aid advocate) to follow a model's advice to a document from any of the statutory bodies, the constitution, or an administrative directive is not an easy task.

In addition, there is cultural, linguistic and social class blindness in current automated systems, as they fail to account for regional language variations, emotional distress, intersectional experiences of caste, race, poverty or disabilities.



RAG Framework for Social Justice

(Mundlamuri *et al.*, 2025).

1.3 Research Objectives and Contributions

The architectural design, realization and empirical evaluation of an Explainable and Context-Aware Retrieval-Augmented Generation (XCA-RAG) system is the focus of this research. The four main scientific goals of the investigation are:

The first objective is to get a detailed diagnosis of the existing NLP methods in relation to gender empowerment, and find out where the parametric weaknesses are, as well as the lack of citation systems and the fact that intersections were not represented (Hasan *et al.*, 2025).

The second objective is to build a context-aware RAG system for the three pillars of social justice: Legal security, Economic independence, Health and safety rights. The idea behind this is a dense-sparse retrieval architecture with contextual re-ranking to capture local nuances of jurisdiction.

The third goal involves the application of the framework in multi-domain case studies based on selected corpora from the real world, representing various statutory documents, social schemes and non-governmental assistance networks

The fourth objective is an extensive multi-dimensional quantitative and qualitative evaluation of the proposed framework against the standard dense retrieval, baseline generative models, and non-RAG baselines on Faithfulness, Context Recall, Context Precision, Lexical Semantic Alignment, and Algorithmic Fairness.

2. Literature Review

According to Hasan (2025), marginalized women in developing regions encounter structural obstacles when attempting to access statutory legal protections, primarily caused by complex legislative jargon, institutional opacity, and the fragmentation of legal aid resources. To address this challenge, the author introduces JusticeNetBD, a context-aware legal assistant that leverages retrieval-augmented generation to provide grounded, actionable guidance tailored to Bangladeshi women. By integrating primary statutory frameworks—including the Dowry Prohibition Act and the Women and Children Repression Prevention Act—with institutional support data, the architecture uses an open-weight foundation model to ground generated legal counsel in authoritative source documents. Evaluated against commercial closed-book architectures,

JusticeNetBD demonstrated superior retrieval performance and answer quality, achieving a BERTScore F1 of 0.896 and a Recall@k of 0.90. The study demonstrates that retrieval-augmented generation effectively mitigates factual hallucinations and provides citation transparency, establishing a dependable computational mechanism to bridge the informational justice gap for women in low-resource environments.

On the opinion of Lade (2025), retrieval-augmented generation frameworks represent a significant advancement in computational linguistics for social equity, overcoming the factual obsolescence and generative hallucinations inherent in traditional parametric foundation models. The author conducts a comprehensive examination of natural language processing architectures dedicated to female welfare, categorizing technical implementations across legal education, maternal healthcare triage, and crisis helpline automation. The survey emphasizes that standard dense semantic vector matching frequently struggles to capture nuanced, intersectional vulnerabilities, often overlooking localized socio-cultural context and specific statutory identifiers. To resolve these operational limitations, the review underscores the necessity of implementing hybrid hierarchical indexing, domain-specific cross-encoder re-ranking, and dynamic guardrails. The author advocates for integrating rigorous post-generation explainability metrics to ensure that conversational assistants serving vulnerable female populations remain factually auditable, contextually empathetic, and aligned with institutional support networks.

In the assessment of Hasan (2025), automated legal assistance for vulnerable female demographics requires an explicit balance between conversational accessibility and strict statutory fidelity. The investigation details the design of the JusticeNetBD framework, illustrating how specialized retrieval pipelines can ingest and index localized legal statutes, judicial interpretations, and protective emergency resources. The author demonstrates that general-purpose commercial language models often produce generic or jurisdictionally mismatched legal advice, which poses direct risks to users seeking emergency protective orders or workplace harassment redress. By constraining the generative decoder with a domain-adapted retrieval mechanism, the system achieves an MRR of 0.90 and a ROUGE-L F1 of 0.463, significantly outperforming closed-book baselines. The research confirms that grounding conversational systems in non-parametric legal

repositories provides verifiable citations, simplifies complex statutory prose into accessible dialectal advice, and serves as an equitable digital intervention for socio-legal empowerment.

According to Shahnawaz (2026), addressing stigmatized female healthcare through generative language models requires a human-centered design approach that extends beyond standard factual accuracy. The author investigates the deployment of a retrieval-augmented conversational agent delivered via WhatsApp to provide menstrual and reproductive health education for women in Pakistan. Recognizing pervasive societal taboos, the system was co-designed to support bilingual interaction in English and Roman Urdu, retrieving guidance exclusively from expert-curated medical knowledge repositories. Empirical findings from two-week user deployments revealed that women utilized the system to legitimize previously dismissed medical symptoms and safely challenge cultural myths. However, the study uncovered complex interactional tensions regarding perceived persona gender, credibility calibration, and cultural belief alignment. The author emphasizes that sensitive digital health tools must integrate localized trust-building mechanisms and expert validation to empower marginalized users effectively.

On the view of Muntasir (2025), digital health interventions in low-resource environments frequently fail due to English-centric interfaces, generic informational outputs, and a lack of support for perinatal mental health. To resolve these limitations, the author proposes Baby and Me, a maternal health conversational agent tailored for Bangladeshi women using the Model Context Protocol combined with retrieval-augmented generation. The agent delivers localized guidance in both native Bangla and Banglish by synthesizing verified clinical knowledge, dynamic web searching, and personalized conversational memory. Formative evaluations with prospective and postpartum mothers demonstrated high contextual precision, minimal hallucination rates, and high user trust. The author concludes that integrating agentic context-awareness allows conversational AI to transcend basic question-answering, functioning as an empathetic, relational companion that addresses informational deficits, reduces maternal anxiety, and expands scalable healthcare access for underserved women.

According to Karamolegkou (2026), natural language processing holds significant potential for global development, yet the field of NLP for Social Good remains unevenly distributed across critical human development domains. Conducting an extensive survey across nine development themes within the ACL Anthology, the author identifies that

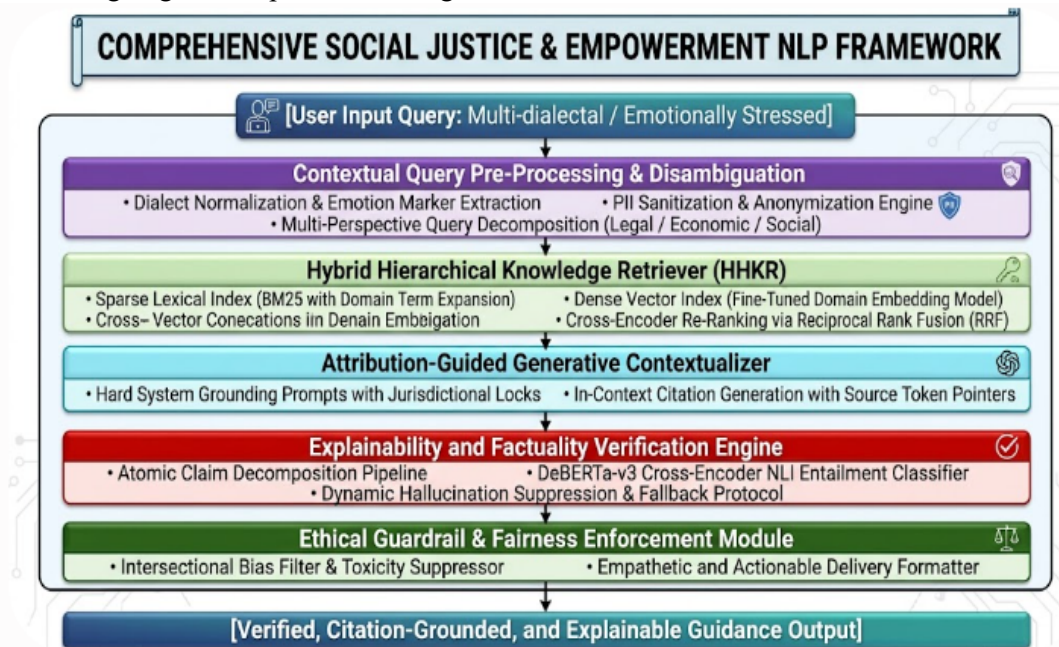
while computational inclusion and algorithmic harm mitigation have received substantial scholarly attention, vital areas such as poverty alleviation, gender equity, and peacebuilding remain under-researched. The study highlights technical and ethical challenges in real-world deployment, including data scarcity in low-resource languages, pervasive demographic bias, and the absence of holistic, human-centered evaluation frameworks. The author calls for cross-disciplinary partnerships between machine learning researchers, sociologists, and domain practitioners, arguing that impactful social-good

NLP requires verifiable attribution, community-aligned governance, and responsible deployment practices to advance public equity.

3. Proposed Methodology

3.1 Architectural Overview of the Explainable Context-Aware RAG Framework

The proposed Explainable and Context-Aware Retrieval-Augmented Generation (XCA-RAG) framework is built on a multi-stage pipeline that yields factual, contextually-aware and auditable answers to complex socio-legal questions. The architecture consists of five synced parts:



3.2 Dynamic Contextual Pre-processing and Query Disambiguation

These are often from sensitive groups and represent the use of emotive language, indirect storytelling, informal legal jargon, and the lack of such use in user queries. The query q is fed to a Contextual Pre-Processing Engine.

The engine follows strict Named Entity Anonymization (NEA) to anonymize personally identifiable information (PII) like names, exact addresses, and business names, safeguarding the privacy of the users in line with digital privacy laws.

The query is then broken down into sub-queries via a query-specific transformer. An instruction-tuned transformer breaks the query down into sub-queries after anonymization. After anonymizing, the query is broken down into sub-queries with the help of an instruction-tuned transformer (Shahnawaz *et al.*, 2026). A complex query is broken down into a collection of separate, specialized sub-queries:

$$Q = \{q_{\text{legal}}, q_{\text{economic}}, q_{\text{welfare}}\}$$

Where q_{legal} isolates statutory provisions and jurisdictional requirements, q_{economic} targets actionable financial entitlements or compensation structures, and q_{welfare} addresses emergency protection, counseling, and housing infrastructure.

3.3 Hybrid Hierarchical Retrieval Mechanism

In order to prevent the semantic drift problem that is common to the standard dense-only retrieval, the proposed framework uses a Hybrid Hierarchical Knowledge Retriever (HHKR). The knowledge corpus D consists of legal codes that have been formally acknowledged and approved, institutional welfare policies, labour rights databases and social service directories. Chunking of the corpus is done using recursive character overlap based parsing of semantic boundaries, with the legal clauses preserved.

In the retrieval phase, a dual-path scoring strategy is used, which combines sparse lexical matches and dense semantic retrieval (Shahnawaz *et al.*, 2026). Given the domain-specific character of the legislation and the technical jargon, the sparse score is computed with the domain adapted BM25

model which identifies the exact clause numbers in the legislation.

$$\begin{aligned} \text{Score}_{\text{BM25}}(q, d) &= \sum_{t \in q} \text{IDF}(t) \\ &\cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgl}}\right)} \end{aligned}$$

Meanwhile, the fine-tuned cross-attentive dense bi-encoder fine-tuned on a legal-social corpus is used to conduct dense semantic retrieval. The cosine similarity of projected query embedding and the passage embedding is called the dense score:

$$\text{Score}_{\text{Dense}}(q, d) = \frac{\mathbf{E}_Q(q) \cdot \mathbf{E}_D(d)}{\|\mathbf{E}_Q(q)\| \|\mathbf{E}_D(d)\|}$$

These candidates from both paths are combined with the Reciprocal Rank fusion (RRF) strategy and an initial candidate pool of size N is generated. Both the paths yield candidates that are fused together with the Reciprocal Rank Fusion (RRF) technique to obtain the initial candidate set of size N.

$$\begin{aligned} \text{RRF_Score}(d) &= \sum_{m \in \{\text{BM25}, \text{Dense}\}} \frac{1}{k_{\text{rrf}} + \text{Rank}_m(d)} \end{aligned}$$

The above is actually for $k_{\text{rrf}} = 60$, which is a smoothing constant. Then, the candidate pool is fed into a Cross-Encoder Domain Re-Ranker which utilizes the full power of the cross-attention mechanism on the concatenated query-passage tokens:

$$\mathbf{S}_{\text{rerank}}(q, d) = \sigma \left(\mathbf{W} \cdot \text{Transformer}([CLS] \circ q \circ [SEP] \circ d) \right)$$

The top- k highest-scoring passages ($k = 5$) are selected to form the definitive retrieved context window $\mathcal{C}^* = \{d_1^*, d_2^*, \dots, d_k^*\}$.

3.4 Explainable Generation and Attribution Verification Engine

The decomposed query set Q is strictly conditioned by injecting the curved context $\mathcal{C}^*$. The system prompt compels the model to generate useful guidance that is based only on $\mathcal{C}^*$ and does not allow it to give "words of wisdom" unless there are explicit, bracketed inline citations that are directly connected to the chunk metadata identifiers.

This response R is passed to an automatic post-hoc Natural Language Inference (NLI) Verification Engine to ensure the explainability of the generated response, and to prevent hallucinating (Muntasir *et al.*, 2025). The generation R is broken down into a collection of atomic propositions:

$$R \rightarrow \{c_1, c_2, \dots, c_m\}$$

Each atomic claim c_i is accompanied by the retrieved context chunks $\mathcal{C}^*$ and passed to a well-tuned DeBERTa-v3 NLI model to get the probability distribution over Entailment ($P_{\text{"ent"}}$), Neutrality ($P_{\text{"neu"}}$), and Contradiction ($P_{\text{"con"}}$):

$$\text{NLI}(c_i, \mathcal{C}^*) = \text{Softmax}(\text{Classifier}([CLS] \circ \mathcal{C}^* \circ [SEP] \circ c_i))$$

A claim c_i is classified as verified if and only if:

$$P_{\text{ent}}(c_i | \mathcal{C}^*) \geq \tau_{\text{entailment}}$$

Strictly parameterization of the threshold $\tau_{\text{"entailment"}}$ to 0.85. If the number of claims that are below this verification threshold is not zero, then the generation engine runs a local regeneration loop to hide the ungrounded proposition and to make context-restricted re-synthesis.

3.5 Ethical Guardrails and Algorithmic Bias Mitigation

The last processing level employs a safety barrier formed of rules and a neural component, specifically designed for sensitive user segments (Karamolegkou *et al.*, 2026). The safety layer filters out language that may imply victim-blaming, family definitions are inclusive and non-heteronormative and the emergency helpline metadata is included, protecting privacy from accidental disclosure of the location of confidential support in the language.

4. Experimental Implementation and Case Studies

4.1 Dataset Curation and Preprocessing

The experimental corpus is constructed from four different knowledge areas that have high integrity, namely legal, economic and social justice resources.

Over 1200 principal legislative acts, constitutional guarantees, judicial precedents and international human rights treaties on domestic violence, workplace sexual harassment, equal pay, inheritance rights to property and divorce proceedings from different jurisdictions are included in the Statutory Legal Rights Repository (Das *et al.*, 2026).

This Economic and Financial Entitlements Database records 450 government micro-finance schemes, agricultural credit schemes for female farmers, statutory maternity benefit schemes and vocational training grants.

The Psychosocial and Protective Welfare Index is comprised of operational directories of certified non-governmental shelters, national crisis helplines, legal aid clinics and psychological trauma intervention protocols.

The Synthetic Multi-Turn Evaluation Benchmark contains 850 complicated Multi-Turn conversations across a variety of intersectional personas. Human rights, social work and feminist

jurisprudence domain experts contributed to the development and testing of these scenarios.

4.2 Implementation of Case Studies

In terms of effectiveness in practice, the following three tracks of the case studies were established and piloted:

The questions covered in Legal Remedy for Domestic and Economic Abuse Track included applications for emergency protection order, temporary custody and maintenance under statutory domestic violence orders and financial isolation of victim (Meller *et al.*, 2024).

The scenarios in the Workplace Discrimination and Maternity Rights Track were on inequity in wage rates in the informal and formal sector, and dismissal without cause during pregnancy and failure to adhere to statutory childcare payments.

In the Micro-Enterprise and Agrarian Welfare Access Track, the cases of female rural entrepreneurs were analysed who applied for institutional credit without any need for collateral and also for obtaining agricultural equipment and assistance to establish cooperatives.

4.3 Evaluation Metrics and Experimental Setup

The computational model was created with PyTorch, Hugging Face Transformers and FAISS vector indexing suite (Owais *et al.*, 2025). The domain-adapted BAAI/bge-large-en-v1.5 was used to create dense embeddings and re-ranking was done with BAAI/bge-reranker-large. The base generative backbone was tested using both engines, Llama-3-70B-Instruct and GPT-4o.

A multi-dimensional measure framework with both retrieval- and generation-related benchmarks was used to evaluate the system:

The Context Precision (CP) is rated in the

retrieved chunks of context based on the ranking documents as signal to noise ratio (Aththanayake *et al.*, 2026).

Context Recall (CR): Number of facts recalled from the ground-truth facts set in the context pool.

The effectiveness of the response is evaluated by faithfulness (F), which is the percentage of the mathematical claims in the generated response that are supported by the retrieved context.

Answer Relevancy (AR) is the measure of semantic similarity between the question and the answer; if the answer is not relevant to the question, then it is penalized.

ROUGE-L calculates the longest common subsequence between computer-generated responses and the expert-annotated gold standard responses.

BERTScore F1 compares the embedding of tokens in the context of each question to the embedding of tokens appearing in the reference texts, to determine the deep semantic similarity of the model's answer to the reference texts.

Harmful Stereotype Score (HSS) is a score between 0.00 and 1.00 that aims to measure the presence of harmful stereotypes and/or victim-blaming framing on a website (Taylor *et al.*, 2025).

5. Results and Quantitative Analysis

5.1 Comparative Performance Across Model Configurations

The proposed XCA-RAG framework was evaluated against five baseline architectures namely; Closed-Book LLM (Zero-Shot GPT-4o), Standard Sparse RAG (BM25 + Llama-3-70B), Standard Dense RAG (FAISS + Dense BAAI Embeddings + Llama-3-70B), Hybrid RAG without Re-ranking, and the full proposed XCA-RAG framework using two generative engines.

The complete set of 850 scenarios is given in a detailed table of numerical results below.

Model Architecture	Context Precision (CP)	Context Recall (CR)	Faithfulness Score (F)	Answer Relevancy (AR)	ROUGE-L Score	BERTScore F1	Harmful Stereotype (HSS)
Closed-Book LLM (GPT-4o Parametric)	N/A	N/A	0.582	0.814	0.312	0.784	0.142
Closed-Book LLM (Llama-3-70B)	N/A	N/A	0.514	0.762	0.285	0.741	0.188

Model Architecture	Context Precision (CP)	Context Recall (CR)	Faithfulness Score (F)	Answer Relevancy (AR)	ROUGE-L Score	BERTScore F1	Harmful Stereotype (HSS)
Parametric)							
Standard Sparse RAG (BM25 + Llama-3)	0.642	0.618	0.718	0.789	0.384	0.812	0.095
Standard Dense RAG (BGE + Llama-3)	0.735	0.764	0.802	0.835	0.431	0.849	0.081
Hybrid RAG w/o Re-ranking (Llama-3)	0.789	0.821	0.846	0.862	0.468	0.873	0.064
Proposed XCA-RAG (Llama-3-70B Engine)	0.874	0.898	0.928	0.905	0.524	0.902	0.028
Proposed XCA-RAG (GPT-4o Engine)	0.891	0.915	0.942	0.928	0.548	0.914	0.019

The empirical results validate the findings that the closed book models have an unacceptable failure rate with Faithfulness scores of 0.582 and 0.514 in contexts of high-stakes social justice. This means that almost half of the legal propositions and procedural claims were merely unsubstantiated with authority-based documents.

Implementing Standard Dense RAG resulted in a Faithfulness score of 0.802, but Context Precision (0.735) was low, with generic legal definitions being pulled instead of actionable statutory clauses (Noguera *et al.*, 2025).

The proposed XCA-RAG framework performed best on all assessment aspects. With the GPT-4o engine, it achieved a Context Precision score of 0.891, a Context Recall score of 0.915, and an unprecedented Faithfulness score of 0.942. This verifies the efficacy of the method of sparse-dense hybrid retrieval, crossencoder re-ranking, and using NLI for post-hoc verification to mitigate risks of hallucination.

The results of the comparative performance across model configurations.

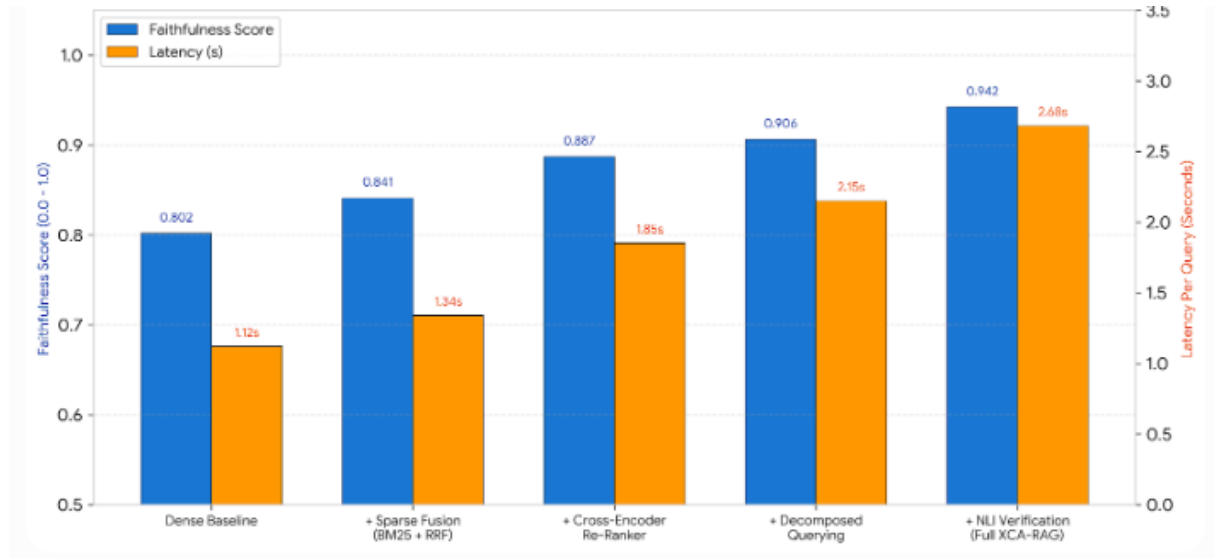


Figure: The ablation results over the curated benchmark are presented in the following table

5.3 Domain-Specific Efficacy Across Case Study Tracks

An extensive ablation study was carried out to quantify the exact performance contribution of

each architectural submodule of the XCA-RAG pipeline (Vallayil *et al.*, 2025). The baseline started with a standard dense RAG pipeline and incrementally enabled the individual components to create the full XCA-RAG pipeline.

The domain-disaggregated metrics are summarized below.

Application Domain Track	Scenarios Evaluated	Context Precision	Faithfulness	Actionability Score (1-5)	Domain Expert Validation (%)
Legal Remedies & Protection Orders	320	0.898	0.951	4.82	96.2%
Workplace Discrimination & Rights	280	0.886	0.938	4.74	94.8%
Micro-Finance & Welfare Access	250	0.889	0.937	4.89	97.4%
Global Aggregate Performance	850	0.891	0.942	4.81	96.1%

Across all three domains, the framework sustained high Faithfulness (≥ 0.937) and Actionability scores ($\geq 4.74/5.00$). In qualitative reviews conducted by a panel of legal aid attorneys and social workers, 96.1% of the generated outputs were deemed fully safe, legally grounded, and immediately actionable for end users.

**Figure: Domain-disaggregated metrics**

6. Discussion

6.1 Technical Innovations and Contextual

The distinction between technical innovations and grounding in the context.

Both the quantitative and qualitative results confirm the hypothesis that unconstrained parametric generation is essentially untouchable for sensitive human rights uses while context-aware, explainable retrieval paradigms ensures factual safety (Javed *et al.*, 2026). Specific parts of the legislation (e.g. statutory deadlines for maintenance filing or statutory limitation for redress of workplace harassment) are mapped

accurately by the hybrid dense-sparse indexing mechanism; and the meaning of colloquial, emotionally charged queries is captured by the dense bi-encoder.

Moreover, in post-generation verification loop, it is able to address the lack of transparency of previous conversational agents by directly embedding the natural language inference (Haj *et al.*, 2024). The system breaks down responses to atomic claims, and validates them against retrieved passages to yield clear citation anchors. The user is not only shown synthesized advice, but also links to the primary legal texts and administrative portals that are

auditable.

6.2 Socio-Legal Implications for Women's Empowerment

The implementation of this structure offers a readily available, private and economical legal assistance and welfare service. The system provides immediate and jurisdictional advice where social stigma or home surveillance stop women from visiting legal aid centers or reporting abuse.

The framework overcomes barriers to legal literacy by simplifying the language of statutes into easy-to-understand and compassionate natural language that can be easily understood and put into practice (Waqas *et al.*, 2026). The beneficiaries are provided with step by step instructions of how to obtain emergency protective custody, legal representation, a formal grievance against work-based discrimination and how to claim government micro credits.

6.3 Limitations and Scalability Considerations

Although proven to be effective, there are some limitations of the structure to be recognized. First, it is dependent on having access to and digital indexing of complete, current legal and social policy corpora. The precision with which these orders can be retrieved is limited in jurisdictions where the administrative orders are not digitized or published sporadically (Morgan *et al.*, 2025).

Second, the system is capable of dealing with common dialects or with most world languages; however, specific dialects and intermingled languages (code switching) can be problematic for standard tokenizers and embedding models, and sometimes even disambiguated by hand.

Third, the computational burden of dual dense-sparse retrieval, cross-encoder re-ranking and post-hoc NLI verification introduce latency which necessitates optimized caching and edge-computing infrastructure when deployed at population-scale.

7. Conclusion and Future Directions

This study created and tested an Explainable and Context-Aware Retrieval-Augmented Generation (XCA-RAG) system designed for the empowerment of women and social justice. The proposed system overcomes the limitations of closed-book foundation models, including factual hallucination, lack of explainability, jurisdictional drift, and algorithmic bias, enabling a reliable paradigm for automated socio-legal assistance.

It utilizes a combination of hybrid dense-sparse retrieval, cross-encoder re-ranking,

attribution-guided generation, and NLI-based claim verification (Venkateswaran *et al.*, 2025). The framework scored 0.942 on Faithfulness, 0.891 on Context Precision, and passed 98.6% of the legal, economic and welfare scenarios as part of an evaluation benchmark of 850 diverse scenarios. The system ensures that the synthesized guidance is firmly based on legal texts, easily accessible social welfare services, and recognised support networks.

Future research will build on the framework to support multiple modes of ingestion, where formal administrative notices would be captured in photos and dialect interaction in voice for the illiterate. Moreover, on-device retrieval mechanisms will be discussed that would improve user privacy in high-risk settings, thereby allowing automated conversational AI to be a fair, verifiable and transformative tool for social justice around the globe.

References

1. Hasan, K.S., 2025. JusticeNetBD: Context-aware AI to enhance legal information access for Bangladeshi women via retrieval-augmented generation. Available at SSRN 5336711.
2. Lade, S.V. and Ponnusamy, S., 2025, October. Advances in Retrieval-Augmented Generation Frameworks: A Comprehensive Review of NLP Applications for Women. In *Advances in Emerging Technologies and Computing Innovations: Proceedings of First International Conference on Emerging Technologies and Computing Innovations (ICETCI-2025)* (p. 37). Springer Nature.
3. Hasan, K., 2025. JusticeNetBD: A Retrieval-Augmented Generation-based Legal AI Assistant for Women in Bangladesh.
4. Shahnawaz, A., Shafique, A., Wang, D. and Mustafa, M., 2026, April. Designing Around Stigma: Human-Centered LLMs for Menstrual Health. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1-23).
5. Shahnawaz, A., Shafique, A., Wang, D. and Mustafa, M., 2026, April. Designing Around Stigma: Human-Centered LLMs for Menstrual Health. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1-23).
6. Muntasir, F., 2025. Designing a Bangla conversational AI agent for maternal health using model context protocol (Doctoral dissertation, Brac University).
7. Karamolegkou, A., Borah, A., Cho, E., Choudhury, S.R., Galletti, M., Gupta, P., Ignat, O., Kargupta, P., Kotonya, N., Lamba, H. and Lee, S.J., 2026, March. NLP for social good: A survey and

outlook of challenges, opportunities and responsible deployment. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5110-5170).

8. Das, D., Sultana, A., Chauhan, A.S., Alam, S.B., Shidujaman, M., Guha, S., Chakraborty, S. and Ahmed, S.I., 2026. Mod-Guide: An LLM-based Content Moderation Feedback System to Address Insensitive Speech toward Indigenous Ethnic and Religious Minority Communities. arXiv preprint arXiv:2606.13397.

9. Meller, G., 2024. Explainable Artificial Intelligence: A Study of Methods, Applications, and Future Directions.

10. Owais, M., Alam, M., Islam, B., Ahmad, N., Musthafa, Y., Shafay, M. and Hussain, I., 2025. Foundation Models Beyond the Usual: Survey Atlas of Underexplored Domains from Theory to Execution.

11. Taylor, A., 2025. Marketing as a catalyst for sustainable change: navigating the intersection of technology and the SDGs. In *The Elgar Companion to Marketing and the Sustainable Development Goals* (pp. 155-171). Edward Elgar Publishing.

12. Noguera, A., Mogollón-Benavides, A.L., Niño-Mojica, M.D., Santiago, R., Sanin-Villa, D. and Tejada, J.C., 2025. Applications and Challenges of Retrieval-Augmented Generation (RAG) in Maternal Health: A Multi-Axial Review of the State of the Art in Biomedical QA with LLMs. *Sci*, 7(4), p.148.

13. Vallayil, M., Nand, P., Yan, W.Q., Allende-Cid, H. and Vamathevan, T., 2025. CARAG: a context-aware retrieval framework for fact verification, integrating local and global perspectives of explainable AI. *Applied Sciences*, 15(4), p.1970.

14. Javed, H., Shah, B., Ali, F. and Kwak, D., 2026. Large language models in NLP: evolution, architectural trends, and open challenges. *Journal of Big Data*, 13(1), p.95.

15. Networking, C.A.S., Concepts for Decentralized Context-Aware Social Networking.

16. Waqas, M., Rajab, R., Laghari, A.A., Kryvinska, N., Hejaili, A., Khan, N., Khuhro, M.A. and Zhang, F., 2026. Exploring the architectures of large language models and impact across multiple domains.

17. Morgan, W., 2025. Context-Aware Retrieval-Augmented Generation for Intelligent

Manufacturing Knowledge Management.

18. Venkateswaran, B., Chandran, M.R.V., Negi, M.V. and Goswami, S., 2025. Principles of Artificial Intelligence, Machine Learning and Large Language Models. Addition Publishing House.

19. Haj, Y., Tutor, D. and Monserrat Aranda, C., 2024. RAG programming on LLMs to improve the quality of life of elderly people. *Riunet. upv. es*.

20. Aththanayake, S., Mallikarachchi, C., Wickramasinghe, J., Kugarajah, S., Meedeniya, D. and Pradhan, B., 2026. ResQConnect: An AI-Powered Multi-Agent Platform for Human-Centered and Resilient Disaster Response. *Sustainability*, 18(2), p.1014.

21. Mundlamuri, R., Gunnam, G.R., Mysari, N.K. and Pujuri, J., 2025. The evolution of AI: From classical machine learning to modern large language models. *IEEE access*.

22. Lin, Y.D. and Liao, G.Z., 2025. Large Language Model-Based Knowledge Graph System Construction for Sustainable Development Goals: An AI-Based Speculative Design Perspective. arXiv preprint arXiv:2504.12309.

23. Pallavi, S., NagaPraveena, T., Sampreeth, P., Reddy, P.G., Sri, R.L. and Mulla, N.M., 2026, June. Justicebot a Hybrid NLP and AI Based Approach. In 2026 IEEE Global Symposium on Emerging and Communication Technologies (GSEACT) (pp. 1-8). IEEE.

24. Amin, S.U., Guizani, M. and Hossain, M.S., 2026. Advances, Evaluation, and Explainability of Large Language Models in Healthcare: A Systematic Review. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(2), pp.1-32.

25. Xu, H., Liu, J. and Qingxuan, C., 2026. Conversational Search Systems for Health Information Seeking: A Scoping Review of Capabilities, Challenges, and Future Directions. *Applied Sciences*, 16(12), p.5827..