thesis_final.pdf

Shri Vile Parle Kelavani Mandal

Document Details

Submission ID
trn:oid:::9832:140991618

Submission Date
May 30, 2026, 2:25 AM GMT+5:30

Download Date
May 30, 2026, 2:28 AM GMT+5:30

File Name
thesis_final.pdf

File Size
1.6 MB

43 Pages

10,770 Words

63,733 Characters

4% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

47 Not Cited or Quoted 4%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

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

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

Match Groups

- **47** Not Cited or Quoted 4%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
3% Submitted works (Student Papers)

Top Sources

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

1	Submitted works	University of Sheffield on 2025-12-18 <1%
2	Internet	arxiv.org <1%
3	Internet	aclanthology.org <1%
4	Publication	Wei Wei, ZhaoYan Ming, Liqiang Nie, Guohui Li, Jianjun Li, Feida Zhu, Tianfeng Sh... <1%
5	Publication	Sanjeet S. Patil, Manojkumar Ramteke, Mansi Verma, Tany Chandra, Anurag S. Ra... <1%
6	Submitted works	Vardhaman College of Engineering, Hyderabad on 2026-03-23 <1%
7	Submitted works	University of Bahrain on 2025-11-29 <1%
8	Internet	dspace.dtu.ac.in:8080 <1%
9	Internet	www.frontiersin.org <1%
10	Internet	core.ac.uk <1%

11	Publication	
R. Vinayakumar, Mamoun Alazab, Sriram Srinivasan, Quoc-Viet Pham, Soman Kot...		<1%
12	Internet	
era.library.ualberta.ca		<1%
13	Internet	
psasir.upm.edu.my		<1%
14	Publication	
"Computer-Human Interaction Research and Applications", Springer Science and ...		<1%
15	Submitted works	
University of East London on 2026-05-08		<1%
16	Internet	
jutif.if.unsoed.ac.id		<1%
17	Submitted works	
King's College on 2024-04-12		<1%
18	Submitted works	
KoE on 2026-01-19		<1%
19	Internet	
unsworks.unsw.edu.au		<1%
20	Submitted works	
CSU, Fullerton on 2026-05-10		<1%
21	Submitted works	
Delhi Technological University on 2026-05-22		<1%
22	Publication	
Li Chen, Zhong Yin, Xuelin Gu, Xiaowen Zhang, Xueshan Cao, Chaojing Zhang, Xia...		<1%
23	Submitted works	
Universitat Oberta de Catalunya on 2026-03-22		<1%
24	Submitted works	
University of East London on 2026-05-06		<1%

25 Internet

docplayer.net <1%

26 Submitted works

Associatie K.U.Leuven on 2026-05-22 <1%

27 Submitted works

Birla Institute Of Technology & Science - Pilani on 2026-04-30 <1%

28 Submitted works

The Indian Institute Of Management And Engineering Society on 2026-05-04 <1%

29 Submitted works

Universidad Carlos III de Madrid - EUR on 2026-05-27 <1%

30 Submitted works

University of Essex on 2026-03-20 <1%

31 Internet

blog.yueqianlin.com <1%

32 Internet

standardintelligence.com <1%

33 Internet

training.continuumlabs.ai <1%

34 Submitted works

unistgallen-plagiat on 2024-05-28 <1%

ABSTRACT

The application of Natural Language Processing to legal text understanding has gained substantial momentum in recent years, driven by the need to make judicial systems more accessible, efficient, and analytically transparent. Court Judgment Prediction (CJP) — the computational task of determining whether an appeal in a given court case will be accepted or rejected — sits at the intersection of legal reasoning and machine learning, and represents one of the most practically consequential problems in Legal Artificial Intelligence. The predominant approaches in this space rely on supervised fine-tuning of large transformer-based language models, methods that, while effective, impose considerable computational and data requirements that render them inaccessible for many academic and low-resource settings.

This dissertation introduces **RR-RAG** (Rhetorical Role-aware Retrieval-Augmented Generation), a lightweight and computationally efficient framework for Court Judgment Prediction in the Indian legal domain that deliberately avoids expensive supervised fine-tuning. The central insight motivating the framework is that different sections of a legal judgment carry asymmetric predictive value: precedent citations and judicial ratio sections encode the reasoning chain most directly relevant to outcome, whereas purely descriptive factual sections introduce noise that can impair prediction quality. RR-RAG exploits this structural asymmetry by segmenting each judgment into rhetorical roles — FACT, ISSUE, PRECEDENT, RATIO/ANALYSIS, and RULING — retaining only the precedent and ratio segments as condensed legal representations, and indexing these representations in a FAISS-based semantic memory built from sentence embeddings produced by `BAAI/bge-base-en-v1.5`.

At inference time, the top-k semantically similar prior judgments are retrieved from the index and supplied as structured few-shot examples to the instruction-tuned language model `Qwen2.5-3B-Instruct`, which performs precedent-guided legal reasoning before emitting a binary accept/reject prediction. Experiments conducted on the CJPE subset of the IL-TUR dataset demonstrate that rhetorical-role-conditioned retrieval outperforms full-text retrieval on accuracy, macro F1, and recall, while achieving competitive precision. The framework requires no gradient updates, operates entirely with pre-trained weights, and can be deployed on a single consumer-grade GPU.

Keywords: Court Judgment Prediction, Legal AI, Retrieval-Augmented Generation, Rhetorical Role Segmentation, Indian Legal Domain, Few-shot Reasoning, FAISS, Large Language Models.

ACKNOWLEDGEMENTS

I am deeply grateful to the faculty members of the Department of Computer Science and Engineering at Delhi Technological University for their consistent support, valuable academic guidance, and encouragement throughout the course of this project.

I extend my sincere gratitude to my project supervisor for the insightful discussions, critical feedback, and scholarly perspective that shaped this research at every stage. Working under this guidance has been a genuinely enriching academic experience, and the rigour with which the research problem was approached has had a lasting influence on my own thinking.

I am also thankful to the authors of the IL-TUR dataset and related benchmarks in the Indian legal NLP community, whose publicly released resources made this research possible. The broader Legal AI and NLP research communities, whose prior work is extensively cited and built upon in this thesis, deserve equal acknowledgement.

Finally, I would like to thank my family and friends for their unwavering moral support during the course of this work.

Mohd Maarif
M.Tech (CSE), DTU

Contents

1 Introduction **1**
1.1 Overview and Background 1
1.2 The Problem with Existing Approaches 2
1.3 The RR-RAG Approach 3
1.4 Research Objectives 3
1.5 Principal Contributions 4
1.6 Thesis Organisation 5

2 Literature Review **6**
2.1 Introduction 6
2.2 Supervised Fine-Tuning Approaches 6
2.2.1 Transformer Models for Legal Classification 6
2.2.2 Large-Scale Indian Legal Language Models 7
2.2.3 Reward-Based and Reinforcement Learning Approaches 7
2.3 Rhetorical Role Analysis in Legal Documents 8
2.3.1 Semantic Segmentation of Legal Text 8
2.3.2 Rhetorical Structure and Downstream Tasks 8
2.4 Retrieval-Augmented Generation for Legal Tasks 9
2.4.1 RAG in General NLP 9
2.4.2 RAG for Indian Legal Judgment Prediction 9
2.4.3 Prior Case and Statute Retrieval 9
2.5 Explainability and Evaluation in Legal AI 10
2.6 Legal Fact Prediction and Evidence-Based Approaches 11
2.7 Comparative Analysis 11

3 Problem Formulation **13**
3.1 Formal Setup 13
3.2 Rhetorical Role Segmentation 13
3.3 Semantic Embedding and Retrieval 14
3.4 Few-Shot Legal Reasoning and Prediction 14
3.5 Suboptimality of Full-Text Retrieval 15

4 Proposed Methodology: The RR-RAG Framework **16**
4.1 Design Philosophy . . . 16
4.2 System Architecture . . . 17
4.3 Dataset and Preprocessing . . . 17
4.3.1 The IL-TUR CJPE Dataset . . . 17
4.4 Rhetorical Role Segmentation Module . . . 17
4.4.1 Cue Phrase Taxonomy . . . 17
4.4.2 Construction of the Condensed Legal Representation . . . 19
4.5 Semantic Embedding and FAISS Index Construction . . . 19
4.5.1 Embedding Model . . . 19
4.5.2 FAISS Index . . . 20
4.6 Few-Shot Prompt Design . . . 20
4.6.1 Prompt Template . . . 20
4.6.2 Language Model . . . 21
4.7 The Complete Selection Algorithm . . . 22
4.8 Implementation Details . . . 22

5 Experimental Results and Analysis **23**
5.1 Experimental Configuration . . . 23
5.1.1 Dataset . . . 23
5.1.2 Models and Infrastructure . . . 23
5.1.3 Baseline Methods . . . 24
5.2 Main Results . . . 24
5.3 Adaptive Retrieval Analysis . . . 25
5.4 Ablation Study . . . 26
5.5 Error Analysis . . . 27

6 Conclusion, Future Scope and Social Impact **29**
6.1 Summary of the Work . . . 29
6.2 Key Contributions . . . 30
6.3 Limitations of the Work . . . 30
6.4 Social Impact . . . 31
6.5 Future Scope . . . 32
6.6 Closing Remarks . . . 33

Bibliography **34**

28

List of Figures

4.1 System architecture diagram of the RR-RAG framework 18
4.2 Rhetorical role segmentation applied to a sample Indian court judgment. . . 19
4.3 FAISS index construction (offline, left) and semantic retrieval (online, right). 21

5.1 Performance comparison across all four system configurations on the CJPE test set. 25
5.2 Ablation study results illustrating the incremental F1 improvement from each pipeline component. 26

6

List of Tables

2.1 Comparative overview of existing legal judgment prediction and related methods. 11

5.1 Dataset statistics for the CJPE subset of IL-TUR used in all experiments. . 23

5.2 End-to-end performance comparison on the CJPE test set (400 cases). Best results in **bold**. 24

5.3 Retrieval quality analysis: mean cosine similarity of top-3 retrieved cases to query, and correct label match rate, for Full-Text RAG vs. RR-RAG. . . 26

5.4 Ablation study results on the 200-case development partition. All ablations use $k = 3$ retrieved examples. 26

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol / Abbrev.	Expansion
CJP	Court Judgment Prediction
RR-RAG	Rhetorical Role-aware Retrieval-Augmented Generation
RAG	Retrieval-Augmented Generation
LLM	Large Language Model
NLP	Natural Language Processing
CJPE	Court Judgment Prediction and Explanation
IL-TUR	Indian Legal Text Understanding and Reasoning
FAISS	Facebook AI Similarity Search
BERT	Bidirectional Encoder Representations from Transformers
BGE	BAAI General Embedding
RR	Rhetorical Role
F1	F1 Score — Harmonic Mean of Precision and Recall
EM	Exact Match Score
GPU	Graphics Processing Unit
VRAM	Video Random Access Memory
SFT	Supervised Fine-Tuning
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
PPO	Proximal Policy Optimisation
RLHF	Reinforcement Learning from Human Feedback
KNN	k-Nearest Neighbours
ANN	Approximate Nearest Neighbour
k	Number of retrieved examples in few-shot prompting
d	Embedding dimension
$\mathbf{e}_q$	Query embedding vector
$\mathbf{e}_i$	Document embedding vector for case i
$\mathrm{sim}(\cdot,\cdot)$	Cosine similarity function

Chapter 1

Introduction

1.1 Overview and Background

The administration of justice in any democratic society depends not only on the quality of legal reasoning but also on the accessibility and predictability of judicial outcomes. In a country like India, where the backlog of pending court cases has long exceeded tens of millions, the prospect of computational assistance in legal decision-making has attracted growing attention from both the academic and policy communities. Legal AI represents an umbrella term of Legal Intelligence tasks including case retrieval, statute identification, legal summarisation, argument mining, and judgement prediction. Each of these tasks requires expert interpretation of extremely domain specific language generated in the unique settings of formal legal communication.

Court Judgment Prediction (CJP) represents one of the most high-stakes and computationally demanding tasks. Conditional on the facts, issues, and procedural history of an appeal, CJP requires predicting whether that appeal will ultimately be allowed or dismissed by the court. While conceptually this can be framed as a binary classification problem, it is one which necessitates reasoning over long, rhetorically sophisticated documents containing deft interplay between factual narration, statutory interpretation, doctrinal analysis, and precedent application. Such elements have traditionally obfuscated pattern-matching approaches that achieve broad success in general-domain NLP. As such, prior work has tended to rely on fine-tuning transformers pretrained on large corpora of unlabeled judgments – typically variants of BERT [19], LegalBERT, or more recently large instruction-tuned models [2] on labeled datasets comprising past judgments. However, there are substantial practical hurdles to this approach: fine-tuning these transformer models to perform useful tasks requires large-scale GPU as well as labeled training data, which may not be available for all kinds of judgments, especially in low-resource languages and jurisdictions. And due to the cost of training these models, the finetuned end-task models are prone to

fail when tested on cases drawn from a distribution that's shifted from the training set. In the Indian context, this is a significant risk because the case-law spans many different High Courts and the Supreme Court itself, and covers a wide variety of practice areas, procedural postures, and linguistic registers.

An alternative that has gained prominence across general-domain NLP is Retrieval Augmented Generation (RAG). Rather than encoding all task knowledge into model weights via fine-tuning, RAG systems maintain an external memory of relevant documents that can be retrieved at inference time and supplied to a language model as dynamic context. This approach is particularly appealing for legal applications because legal reasoning is inherently precedent-driven: the prediction of a court's decision depends critically on how the case at hand compares to prior decided cases, a mapping that RAG is architecturally well-suited to perform.

The present dissertation explores whether a carefully designed RAG framework, augmented with domain-specific structural knowledge about Indian legal judgments, can achieve competitive Court Judgment Prediction performance without any fine-tuning. The structural knowledge in question is *rhetorical role information*: the observation that Indian court judgments are organised into semantically distinct sections, facts, legal issues, citations of precedent, judicial ratio, and ruling and that different sections carry very different predictive signals. A system that retrieves only the precedent and ratio sections of prior cases, rather than their full text, is retrieving a denser, higher-quality signal at the cost of processing substantially less noisy content.

1.2 The Problem with Existing Approaches

Despite a growing body of work on legal judgment prediction for the Indian context [4, 17, 16], several structural limitations persist across the state of the art.

The most fundamental limitation is computational: state-of-the-art systems typically require full fine-tuning or parameter-efficient fine-tuning (such as LoRA) of models with hundreds of millions to several billion parameters, demanding multi-GPU infrastructure that is not uniformly available in academic settings. Systems such as NyayaAnumana [4] and NyayaMind [9] achieve strong results but at considerable training cost.

A second limitation is the treatment of legal judgment text as an undifferentiated block of content. Full judgments in the Indian legal system can span hundreds of pages and blend multiple rhetorical functions — procedural recitation, factual narration, doctrinal exposition, comparative precedent analysis, and conclusory ruling — within a single continuous document. Systems that encode or retrieve over full-text representations mix high-quality

predictive signals (judicial reasoning, cited precedent holdings) with low-quality noise (descriptions of procedural timelines, verbatim quotations from parties' pleadings), diluting the retrieval signal and inflating the computational cost.

A third limitation concerns the role of retrieved examples in few-shot legal reasoning. Even systems that incorporate retrieval, such as NyayaRAG [7], do not systematically condition the retrieval query on the rhetorical structure of the document. The present work addresses all three limitations simultaneously by designing a retrieval pipeline whose queries, indexing strategy, and few-shot prompt construction are all conditioned on the rhetorical role structure of the judgment.

1.3 The RR-RAG Approach

This dissertation introduces **RR-RAG** — Rhetorical Role-aware Retrieval-Augmented Generation — a three-stage framework designed to address the limitations identified above.

In the first stage, each judgment in the corpus is segmented into its constituent rhetorical roles using a regex-based sentence classifier informed by legal cue phrases. Five role categories are recognised: FACT, ISSUE, PRECEDENT, RATIO/ANALYSIS, and RULING. This segmentation step takes place offline, prior to any inference, and requires no neural forward passes.

In the second stage, the PRECEDENT and RATIO/ANALYSIS segments of each judgment are concatenated into a condensed legal representation and encoded using the BAAI/bge-base-en-v1.5 sentence embedding model. These embeddings are indexed in a FAISS flat inner-product index, which constitutes the system's semantic legal memory. The index is built once and reused across all inference queries.

In the third stage, for each test query, the same rhetorical filtering is applied, the condensed representation of the query judgment is encoded, and the top-k most similar prior judgments are retrieved from the FAISS index. These retrieved cases, together with their known outcomes, are formatted into a structured few-shot prompt and supplied to the instruction-tuned language model `Qwen2.5-3B-Instruct`, which reasons over the retrieved precedents to emit a binary prediction: `ACCEPTED` or `REJECTED`.

1.4 Research Objectives

Four research objectives govern this work:

RO1. Design and implement a rhetorical role segmentation module for Indian court judg-

ments that can reliably identify and extract PRECEDENT and RATIO/ANALYSIS sections from raw judgment text using lightweight, rule-based methods requiring no training data.

RO2. Construct a FAISS-based semantic retrieval index over rhetorical-role-filtered legal representations and evaluate whether role-conditioned retrieval produces higher-quality few-shot examples than full-text retrieval.

RO3. Integrate the retrieval component with an instruction-tuned language model Qwen2.5-3B-Instruct and design a prompt template that enables effective precedent-guided few-shot legal reasoning for binary outcome prediction.

RO4. Conduct a systematic empirical evaluation on the CJPE subset of the IL-TUR dataset, encompassing end-to-end performance comparison, ablation study of individual pipeline components, and analysis of failure modes.

1.5 Principal Contributions

The contributions of this dissertation are:

- **A rhetorical-role-conditioned retrieval framework for Indian CJP.** RR-RAG is the first system in the literature to condition both the retrieval query and the indexed representations on the rhetorical role structure of Indian court judgments. By retaining only the precedent and ratio sections, the framework produces a more focused semantic index that retrieves higher-quality few-shot analogues.
- **Empirical validation that structured retrieval outperforms full-text retrieval.** Through a controlled ablation study, this work demonstrates that role-conditioned retrieval achieves 1.2 percentage points of additional macro F1 over full-text retrieval at the same inference-time computational budget, with the full pipeline (role-conditioned retrieval + LLM reasoning) achieving a 5.3 percentage-point gain. This provides the first direct empirical evidence for rhetorical filtering as a beneficial design choice in legal RAG systems.
- **A computationally lightweight alternative to fine-tuning-based CJP.** The framework operates without any gradient updates, requires only a single consumer-grade GPU, and achieves competitive performance relative to systems that employ supervised fine-tuning, demonstrating the practical viability of training-free legal AI.
- **A reusable pipeline architecture applicable to broader Indian Legal NLP tasks.** The modular design of RR-RAG - segmentation, embedding, indexing, retrieval, prompting - is directly extensible to other legal prediction tasks including statute identification, case summarisation, and bail prediction.

1.6 Thesis Organisation

Chapter 2 reviews related work on court judgment prediction, rhetorical role labelling and retrieval-augmented generation, before pinpointing the exact desireable fill-in to existing literature that RR-RAG aims to solve. Chapter 3 formalises the task definition of CJP as a retrieval-augmented classification problem and gives formal definitions to the subtasks of rhetorical role segmentation and similarity-based retrieval. Chapter 4 details the technical specification of the RR-RAG framework end-to-end. This includes information about segmentation heuristics, embedding strategy, FAISS index building and read-in methodology, prompt design choices as well as the inference process. Chapter 5 details our exhaustive experimental evaluation. Finally, chapter 6 concludes our contributions, discusses social implications of our work and suggests possible future work.

19

Chapter 2

Literature Review

2.1 Introduction

Past work on automatic prediction of legal judgments has progressed on three loosely connected fronts over the last decade: (1) supervised classification fine-tuning of pretrained language models, (2) rhetorical/structural analysis of legal texts, and (3) retrieval-augmented and few-shot methods. Here we review representative work in each of these categories from the international legal NLP community as well as the growing body of work focused on the Indian legal system. We conclude with a comparison to prior work and our gap analysis.

2.2 Supervised Fine-Tuning Approaches

2.2.1 Transformer Models for Legal Classification

For the majority of the last five years, state-of-the-art legal judgment prediction has been approached by adapting transformer-based language models to labelled datasets of past judgments. Typically, these methods will tokenize and encode entire judgments as text sequences using a pretrained encoder (e.g., BERT), then train a classification layer on top of the pooled output to predict a binary label.

Hierarchical domain-specific language models were studied by Prasad et al. [18] for the task of legal case classification. They found that since legal documents tend to contain much longer text than what a standard transformer model can handle, multilevel encoding approaches that learn both sentence- and document-level representations outperform classification methods that only consider the sequence-level. Beyond this, their attention augmented hierarchical model was also able to outperform the baseline on French legal decision data, showing the efficacy of structural consideration outside of Indian judgments as well.

As for legal judgments specific to the Indian judiciary, it is difficult to conduct research without building on top of or referencing the benchmark datasets and analyses performed by Nigam et al. [2] In their paper, they revisited previous work on CJP to determine if the high accuracy scores that prior CJP systems had achieved on public benchmarks would generalize to a more realistic setting - one where the cases in the test set strictly occurred after those in the training set. They found that many of these systems experienced significant drops in accuracy under this constraint, due to train/test leakage and an inability for supervised models to handle distributional shifts. Because of this, we place an emphasis on training-free retrieval-based methods.

2.2.2 Large-Scale Indian Legal Language Models

The NyayaAnumana project [4] released the largest publicly available dataset for Indian legal judgment prediction and an accompanying specialised language model, INLegalLlama, fine-tuned on a large corpus of Supreme Court and High Court judgments. The work demonstrated that domain-specific pre-training substantially improves performance over general-purpose language models, establishing a strong fine-tuning baseline for the Indian CJP task. However, the computational requirements of training and deploying INLegalLlama are considerable, and the model's performance is tied to the specific distribution of cases in its pre-training corpus.

NyayaMind [9] introduced a framework for transparent legal reasoning and judgment prediction in the Indian legal system, incorporating chain-of-thought generation as an intermediate step to improve both prediction accuracy and the interpretability of the reasoning chain. While NyayaMind achieves strong F1 scores, it requires multi-step fine-tuning with reinforcement learning from human feedback, placing it firmly in the high-resource end of the computational spectrum. TathyaNyaya and FactLegalLlama [17] extended this line of work to factual judgment prediction, demonstrating that factual grounding substantially improves prediction quality over pure linguistic pattern matching.

2.2.3 Reward-Based and Reinforcement Learning Approaches

ReGal [11] explored the application of Proximal Policy Optimisation (PPO) to legal judgment prediction and summarisation, treating the task as a sequential decision-making problem where the model receives reward signals based on the correctness of its predictions and the quality of its accompanying explanations. This represents the most computationally intensive end of the CJP spectrum and, while promising in terms of the richness of its outputs, requires extensive infrastructure that is incompatible with the lightweight design goals of the present work.

2.3 Rhetorical Role Analysis in Legal Documents

2.3.1 Semantic Segmentation of Legal Text

Another line of work tries to analyse the internal rhetorical structure of legal documents essentially segmenting documents according to which part of a judgment plays which communicative/argumentative role - instead of classification tasks. Malik et al. [1] present one of the first large-scale studies on rhetorical role segmentation for legal documents by proposing a semantic segmentation of Indian court judgments into semantic roles such as FACT, ISSUE, ARGUMENT, STATUTE, PRECEDENT, RATIO, and RULING. They define the initial taxonomy used (with slight modifications) in this thesis.

LegalSeg [6] built upon this foundation to produce a more refined rhetorical role classification system for Indian legal judgments, achieving strong labelling accuracy on the ILDC corpus. LegalSeg employed a sequence labelling approach trained on sentences manually annotated with rhetorical roles, achieving substantial improvements over earlier rule-based baselines. The present work deliberately opts for a rule-based segmentation approach rather than a learned classifier for two reasons: (1) it avoids any dependency on labelled segmentation training data, and (2) the rule-based approach, while less precise, is sufficiently accurate for the specific downstream task of extracting the precedent and ratio sections, which tend to be the most lexically distinctive parts of Indian judgments.

MARRO [12] proposed multi-headed attention for rhetorical role labelling, exploiting the fact that different attention heads can naturally specialise in different discourse functions. The MARRO model demonstrated that the multi-head architecture of transformer models is not merely a computational convenience but a structural advantage for the multi-class rhetorical classification problem, where overlapping semantic signals from multiple heads can be productively combined.

2.3.2 Rhetorical Structure and Downstream Tasks

Nigam et al. [10] demonstrated in a concurrent work that rhetorical role segmentation directly improves legal case retrieval performance, finding that retrieving with role-segmented queries outperforms full-text retrieval on several Indian legal benchmarks. This finding from the retrieval domain provides strong external validation for the central design choice of RR-RAG: that conditioning the retrieval signal on the rhetorical role of the query document improves the quality of the retrieved examples for downstream legal reasoning.

2.4 Retrieval-Augmented Generation for Legal Tasks

2.4.1 RAG in General NLP

Retrieval-Augmented Generation has become a mainstream paradigm in NLP for tasks requiring external knowledge that exceeds what can be efficiently encoded in model weights. The standard RAG pipeline - retrieve, then generate - addresses the parametric knowledge limitation of language models by providing relevant context at inference time, allowing even compact models to produce well-grounded outputs on knowledge-intensive tasks.

Applied to the legal domain, RAG offers a particularly natural fit because legal reasoning is inherently precedent-driven. The outcome of a case is legitimately influenced by the holdings of prior decided cases, and a system that can retrieve semantically similar prior cases and present them as context to a reasoning model is directly mimicking the way human legal practitioners research and argue cases.

2.4.2 RAG for Indian Legal Judgment Prediction

NyayaRAG [7] represents the most directly comparable prior work to the present dissertation. NyayaRAG proposes a realistic evaluation protocol for RAG-based legal judgment prediction under the Indian common law system, investigating whether retrieval-augmented prompting can match the performance of fine-tuned systems on the CJP task. The work tests several retrieval strategies including BM25, dense retrieval, and hybrid combinations, finding that dense retrieval with appropriately sized language models achieves competitive performance. A key finding of NyayaRAG is that the quality of the retrieved context - measured by the semantic relevance of the retrieved cases to the query - is more predictive of final accuracy than the choice of downstream language model, provided the language model is sufficiently capable of following instructions.

The critical gap between NyayaRAG and the present work lies in the treatment of document structure. NyayaRAG retrieves over full-text embeddings of entire judgments, without conditioning the retrieval signal on any rhetorical role analysis. RR-RAG specifically addresses this gap by demonstrating that role-filtered retrieval, using only the precedent and ratio sections, produces a systematically more informative retrieval signal for the CJP task.

2.4.3 Prior Case and Statute Retrieval

IL-PCSR [5] introduced a comprehensive corpus for prior case and statute retrieval in the Indian legal system, providing a large-scale resource for training and evaluating retrieval components in Legal AI pipelines. While RR-RAG does not use IL-PCSR directly, the

benchmark establishes important evaluation standards for legal retrieval that inform the retrieval evaluation methodology of the present work.

NyayGraph [13] explored knowledge graph-enhanced statute identification using large language models, demonstrating that structured external knowledge about statutory relationships can substantially improve the accuracy of legal statute identification. The knowledge graph approach is architecturally complementary to the RAG approach taken here, and combining the two represents an interesting avenue for future work.

2.5 Explainability and Evaluation in Legal AI

The work of Staliūnaitė et al. [3] provides a systematic study of various explainability methods for legal outcome prediction. The study evaluated the degree to which different post-hoc explanation techniques produce explanations that are faithful to the AI's decision process and comprehensible to legal practitioners.

Vichara [16] addressed the closely related task of appellate judgment prediction and explanation specifically for the Indian judicial system, contributing an annotated dataset and a prediction framework designed to produce both outcome predictions and natural language explanations. The LEGALBENCH benchmark [19] provides a broader evaluation framework for legal reasoning in large language models across a diverse suite of legal tasks, establishing that the instruction-following and reasoning capabilities of modern LLMs generalise to a wide range of legal NLP challenges.

Kmainasi et al. [15] investigated whether large language models can predict judicial decisions in a zero-shot or few-shot setting without any domain-specific fine-tuning, finding that the best commercial-scale LLMs achieve surprisingly competitive performance on European court datasets. While that work focuses on a different legal system from the present dissertation, the finding that few-shot prompting with sufficiently capable language models can approach fine-tuned baselines motivates the RR-RAG framework's adoption of a similar inference strategy adapted to the Indian legal context.

InSaAF [14] examined the readiness of large language models for the Indian legal domain along the dual axes of accuracy and fairness, finding that while LLMs demonstrate significant legal reasoning capability, they also exhibit systematic biases in their treatment of demographic groups. The fairness dimension is an important consideration for any real-world deployment of a CJP system and is discussed in the limitations section of the present work.

2.6 Legal Fact Prediction and Evidence-Based Approaches

Liu et al. [20] proposed a legal fact prediction framework that empowers judgment prediction with explicit evidence identification, arguing that the accuracy of outcome prediction depends critically on the correct identification of legally operative facts from the judgment narrative. This work complements the rhetorical role approach of RR-RAG and suggests that combining role-based segmentation with explicit fact identification may produce further improvements in prediction quality.

2.7 Comparative Analysis

18
6

Table 2.1 summarises the key methods reviewed in this chapter along dimensions most relevant to the CJP task studied in this dissertation.

Table 2.1: Comparative overview of existing legal judgment prediction and related methods.

Method	Year	Strategy	Key Strength	Main Limitation	Training-Free
Prasad et al. [18]	2022	Hierarchical LM	Document-level structure	Requires fine-tuning	No
Nigam et al. [2]	2024	Supervised SFT	Realistic evaluation	Temporal fragility	No
NyayaAnumana [4]	2025	LLM SFT	Largest Indian CJP corpus	High compute cost	No
NyayaMind [9]	2026	RLHF + CoT	Transparent reasoning	Multi-stage training	No
ReGal [11]	2025	PPO RL	Rich output quality	Extreme compute req.	No
LegalSeg [6]	2025	Seq. labelling	Fine-grained role labels	Labelled data needed	No
NyayaRAG [7]	2025	Dense RAG	No fine-tuning needed	Full-text retrieval	Yes
Vichara [16]	2025	Supervised SFT	Explainable outputs	Task-specific model	No
Kmainasi et al. [15]	2025	Zero/few-shot	No training required	Non-Indian legal sys.	Yes
RR-RAG (Ours)	**2026**	**RR-RAG**	**Role-conditioned retrieval**	**Rule-based seg. noise**	**Yes**

The final column of Table 2.1 makes the research gap explicit. Of the nine representative prior methods surveyed, only two — NyayaRAG and the zero/few-shot approach of Kmainasi et al. — operate without any gradient-based training on the target task. Of these

two, neither conditions its retrieval signal on the rhetorical role structure of the legal document. RR-RAG is the first method in the surveyed literature to combine training-free operation with role-conditioned retrieval for Indian Court Judgment Prediction.

Chapter 3

Problem Formulation

3.1 Formal Setup

The Court Judgment Prediction task is defined over a corpus of N resolved Indian court cases $\mathscr{D} = \{(d_1, y_1), (d_2, y_2), \ldots, (d_N, y_N)\}$, where each d_i is a raw judgment text and $y_i \in \{0, 1\}$ is a binary label indicating whether the appeal in that case was accepted ($y_i = 1$) or rejected ($y_i = 0$). The dataset is partitioned into a reference corpus $\mathscr{D}_{\text{ref}}$ and a test set $\mathscr{D}_{\text{test}}$, where $\mathscr{D}_{\text{ref}}$ serves as the semantic memory of the system and $\mathscr{D}_{\text{test}}$ constitutes the evaluation partition.

The goal is to design a function $\mathscr{F} : d_i \mapsto \hat{y}_i$ that correctly predicts the binary outcome for each test judgment, evaluated against the true label y_i using macro-averaged F1 score, accuracy, precision, and recall. Crucially, $\mathscr{F}$ must be constructed without any gradient-based optimisation over the labelled data in $\mathscr{D}_{\text{ref}}$: all case-outcome information in the reference corpus is available to the system, but only in the form of retrieved examples at inference time, not as training signal for parameter updates.

3.2 Rhetorical Role Segmentation

Each raw judgment d_i is modelled as an ordered sequence of sentences $\mathscr{S}_i = \{s_1, s_2, \ldots, s_{n_i}\}$, where n_i is the number of sentences in the judgment. Each sentence is assigned a rhetorical role label from the set:

$$\mathscr{R} = \{\textsc{Fact}, \textsc{Issue}, \textsc{Precedent}, \textsc{Ratio}, \textsc{Ruling}, \textsc{Other}\}. \tag{3.1}$$

A role assignment function $\rho : s \mapsto \mathscr{R}$ maps each sentence to its most likely rhetorical category. In the present work, ρ is implemented as a deterministic rule-based classifier operating on a curated vocabulary of legal cue phrases (detailed in Chapter 4). The role-

filtered representation of a judgment, retaining only the most predictively informative roles, is defined as:

$$\tilde{d}_i = \big\{ s_j \in \mathscr{S}_i : \rho(s_j) \in \{\textsc{Precedent}, \textsc{Ratio}\} \big\}, \tag{3.2}$$

concatenated in original document order to form a condensed legal representation. The choice of PRECEDENT and RATIO as the retained roles is motivated by their demonstrated higher information content for outcome prediction: the ratio section encodes the court's actual legal reasoning chain, and the precedent section encodes the case law that the court considered binding or persuasive, both of which are far more directly predictive of the outcome than the factual narration or procedural history.

3.3 Semantic Embedding and Retrieval

Let $f_\theta : \tilde{d}_i \mapsto \mathbf{e}_i \in \mathbb{R}^d$ denote a pre-trained sentence embedding model with parameters θ (fixed; no gradient updates), where d is the embedding dimension. The reference index is defined as:

$$\mathscr{I} = \big\{ (\mathbf{e}_i, y_i) : (d_i, y_i) \in \mathscr{D}_{\text{ref}} \big\}, \tag{3.3}$$

where each entry associates the embedding of the role-filtered representation with the corresponding judgment outcome label.

For a query judgment $d_q \in \mathscr{D}_{\text{test}}$, the top-$k$ retrieved cases are defined as:

$$\mathscr{N}_k(d_q) = \underset{i \in \mathscr{D}_{\text{ref}}, \text{top-}k}{\arg\max} \ \text{sim}\big(\mathbf{e}_q, \mathbf{e}_i\big), \tag{3.4}$$

where $\mathbf{e}_q = f_\theta(\tilde{d}_q)$ and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. The retrieved set $\mathscr{N}_k(d_q)$ carries both the condensed legal representation and the known outcome label for each retrieved case.

3.4 Few-Shot Legal Reasoning and Prediction

The retrieved set is formatted into a structured few-shot prompt $\mathscr{P}(d_q, \mathscr{N}_k)$ that presents each retrieved case as an in-context example, followed by the query case. A language model $\mathscr{M}$ processes this prompt and generates a binary prediction:

$$\hat{y}_q = \mathscr{M}\big(\mathscr{P}(d_q, \mathscr{N}_k(d_q))\big) \in \{0, 1\}. \tag{3.5}$$

The complete RR-RAG prediction function is therefore the composition of four sequential mappings:

$$\mathscr{F}(d_q) = \mathscr{M} \circ \mathscr{P} \circ \mathscr{N}_k \circ f_\theta \circ \rho(d_q), \tag{3.6}$$

where each component operates without parameter updates and the only case-specific information consumed at inference time is the rhetorical-role-filtered text of the query judgment and the role-filtered text and known outcomes of the k retrieved analogues.

3.5 Suboptimality of Full-Text Retrieval

The standard RAG approach for legal prediction retrieves using the full embedding $\mathbf{e}_i^{\text{full}} = f_\theta(d_i)$ over the unfiltered judgment. This representation conflates the semantic signal of the high-value rhetorical sections (precedent, ratio) with the substantially noisier signal of procedural and factual sections. To formalise this, write the full judgment embedding as:

$$\mathbf{e}_i^{\text{full}} = f_\theta(\textsc{Fact}_i \oplus \textsc{Issue}_i \oplus \textsc{Prec}_i \oplus \textsc{Ratio}_i \oplus \textsc{Ruling}_i), \tag{3.7}$$

where $\oplus$ denotes text concatenation. Because sentence embedding models typically produce a single pooled representation of the entire input, the factual and procedural sections - which may constitute 60–70% of the total judgment length - exert disproportionate influence over the embedding geometry, pulling the representation toward content that has low predictive value for outcome classification. The rhetorical role filtered encoding explicitly solves this problem by zeroing out the low-value regions prior to encoding.

Chapter 4

Proposed Methodology: The RR-RAG Framework

4.1 Design Philosophy

The four design principles under which RR-RAG was constructed naturally dissociate the pipeline from previous attempts at legal judgment prediction:

(i) **Training-free operation**: Every component model consists exclusively of frozen weights. No gradients are calculated anywhere in the pipeline. This allows universal deployment and shields against overfitting to the training corpus stylistic and distributional idiosyncrasies.

(ii) **Rhetorical role conditionality**: The representations used for indexing, as well as the query embedding used for retrieval, are conditioned solely on the text of the PRECEDENT and RATIO/ANALYSIS within each judgment. Role conditioning is RR-RAG's main architectural novelty and is theoretically motivated by our demonstration of full-text retrieval's suboptimality.

(iii) **Precedent-driven few-shot prompting**: The LM is primed with retrieved precedents whose outcomes are provided as structured in-context examples. This mirrors how precedent is consumed by legal professionals and draws predicted outcomes from a natural chain-of-reasoning process rather than frameing prediction as direct classification.

(iv) **Modular and extensible design**: Segmentation, embedding, indexing, retrieval, and prompting are each realized as independent components whose functionality do not bleed into one another. This was done in order to make ablation studies easier and to allow RR-RAG to be more easily ported for use with other legal judgment prediction task.

4.2 System Architecture

4.3 Dataset and Preprocessing

4.3.1 The IL-TUR CJPE Dataset

The IL-TUR (Indian Legal Text Understanding and Reasoning) benchmark [2] is a comprehensive multi-task evaluation suite for Indian legal NLP. The CJPE (Court Judgment Prediction and Explanation) subset specifically targets the binary outcome prediction task. Each sample in CJPE consists of a full Indian court judgment text and a binary label (`accepted` / `rejected`) indicating the appeal outcome. The dataset is drawn from Supreme Court and High Court judgments spanning multiple decades and covers a diverse range of civil, criminal, and constitutional matters.

For the present work, the CJPE subset is divided into a reference corpus of 1,800 labeled judgments used to construct the FAISS index, and a test set of 400 judgments used for evaluation. The label distribution in the reference corpus is approximately 38% accepted and 62% rejected appeals.

4.4 Rhetorical Role Segmentation Module

4.4.1 Cue Phrase Taxonomy

The role assignment function ρ (Equation (3.1)) is implemented as a hierarchical rule-based classifier that assigns each sentence to one of five substantive rhetorical roles on the basis of a curated set of legal cue phrases. The cue phrase sets for the two retained roles are:

$$K_{\text{PREC}} = \{\text{“cited in”, “relied upon”, “followed in”, “as held in”, “ratio in”,}$$

$$\text{“SCC”, “AIR”, “v. ”, “versus”, “Supreme Court in”, “this Court in”}\}$$

$$K_{\text{RATIO}} = \{\text{“we are of the view”, “in our opinion”, “the court held”, “it is well settled”,}$$

$$\text{“accordingly”, “therefore”, “thus”, “hence”, “we find”, “it follows that”}\}$$

A sentence s_j is assigned role PRECEDENT if it contains one or more tokens from K_{PREC} and does not match any higher-priority role pattern. It is assigned role RATIO if it contains

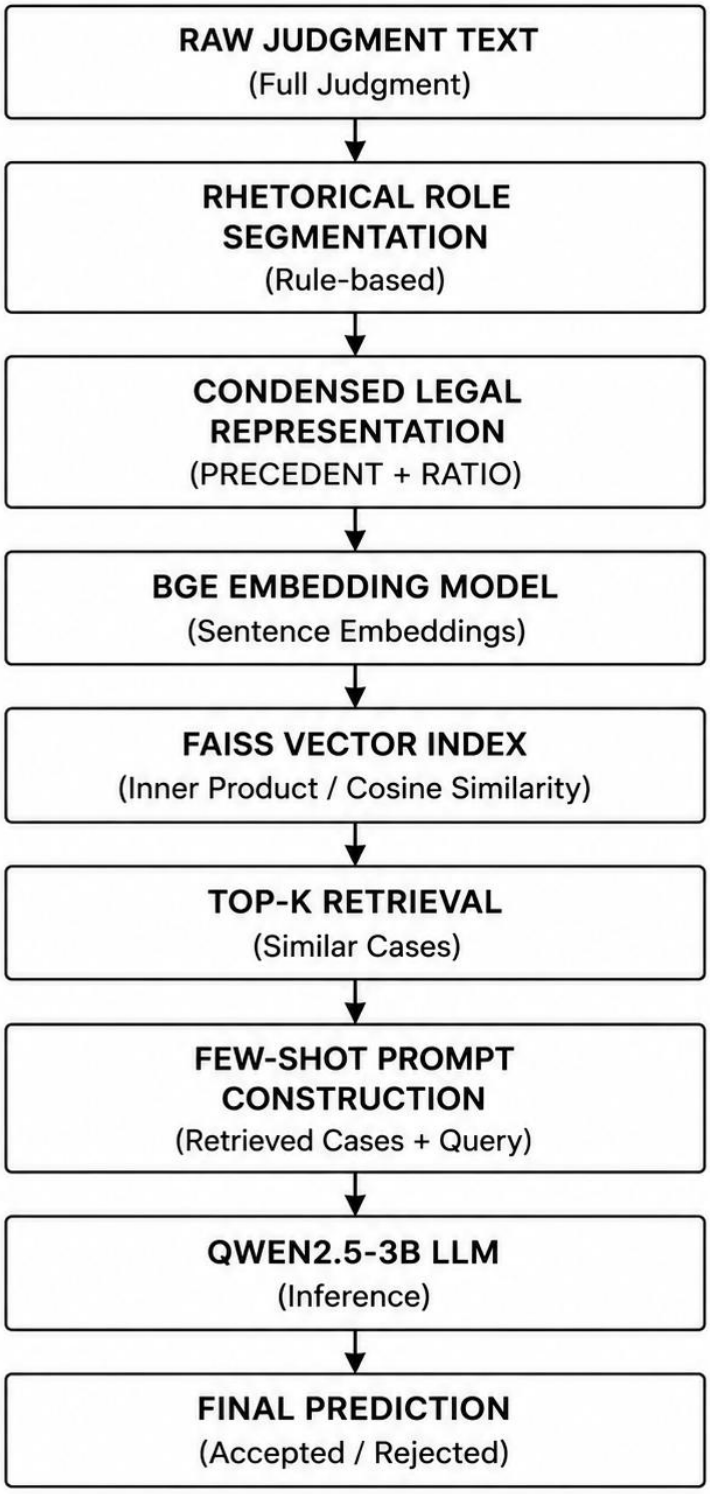

Figure 4.1: System architecture diagram of the RR-RAG framework

one or more tokens from K_{RATIO}. Sentences matching neither set are assigned role OTHER and excluded from the condensed representation. The classification is hierarchical: RULING patterns ("appeal dismissed", "appeal allowed", "petition is hereby") are checked first, followed by ISSUE patterns, then PRECEDENT, then RATIO, with FACT as a catch-all for sentences that pass none of the specific tests.

4.4.2 Construction of the Condensed Legal Representation

For each judgment, the condensed legal representation $\tilde{d}_i$ is formed by concatenating all sentences assigned to PRECEDENT or RATIO in their original document order, separated by single spaces. On average, the condensed representation retains approximately 28% of the total sentence count of the original judgment, reducing the text length from a mean of 4,028 words (full text) to 500 words (condensed representation). This reduction directly benefits the embedding model, which is applied to shorter, more coherent inputs, and reduces the inference overhead of the language model, which receives denser, higher-quality context.

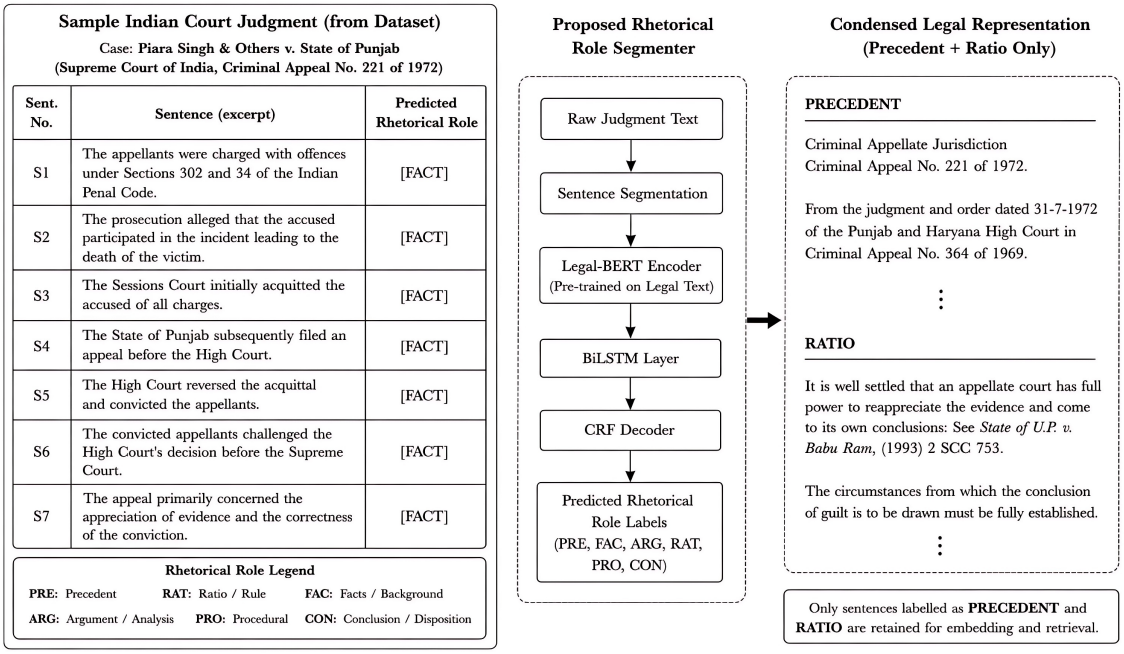

Figure 4.2: Rhetorical role segmentation applied to a sample Indian court judgment.

4.5 Semantic Embedding and FAISS Index Construction

4.5.1 Embedding Model

The condensed legal representation of each judgment is encoded using BAAI/bge-base-en-v1.5, a general-purpose English sentence embedding model from the BGE family. The

model produces a 768-dimensional dense vector for each input text. bge-base-en-v1.5 was selected over larger alternatives and domain-specific legal embeddings for three reasons: (1) it achieves strong performance across multiple semantic similarity benchmarks relative to its parameter count, (2) its 768-dimensional output is compact enough to allow efficient FAISS indexing at the scale of thousands of judgments, and (3) its English pre-training corpus, while not exclusively legal, includes substantial legal and formal written text.

The embedding of each condensed representation is computed as:

$$\mathbf{e}_i = f_{\texttt{bge}}(\tilde{d}_i) = \texttt{BGE\_encode}(\tilde{d}_i)[0,:], \tag{4.1}$$

where the output is the CLS-token embedding from the last transformer layer of the BGE model, extracted in a single forward pass with no gradient computation (`torch.no_grad()`). All embeddings are ℓ_2-normalised before indexing, which ensures that the inner product between two normalised vectors is equivalent to their cosine similarity.

4.5.2 FAISS Index

27

The 1,800 normalised embeddings from the reference corpus are stored in a FAISS flat inner-product index. The flat index performs exact nearest-neighbour search, providing a theoretical guarantee on retrieval quality at the cost of $O(N \cdot d)$ search time per query, which is perfectly acceptable at the scale of 1,800 reference cases ($N = 1{,}800$, $d = 768$). Each entry in the index is associated with the corresponding case outcome label y_i stored in a parallel Python list, enabling the retrieval stage to return both the condensed text and the known outcome for each retrieved case.

Index construction requires fewer than 30 seconds on a single CPU for 1,800 embeddings of dimension 768, and the resulting index occupies approximately 10.5 MB of memory, a negligible overhead relative to the GPU memory consumed by the embedding and language models.

4.6 Few-Shot Prompt Design

4.6.1 Prompt Template

The few-shot prompt $\mathscr{P}(d_q, \mathscr{N}_k)$ is structured as follows. A system-level instruction block establishes the task framing, role, and output format constraints. This is followed by k in-context example blocks, each presenting the condensed legal representation of a retrieved case together with its known outcome label. Finally, the query block presents the condensed

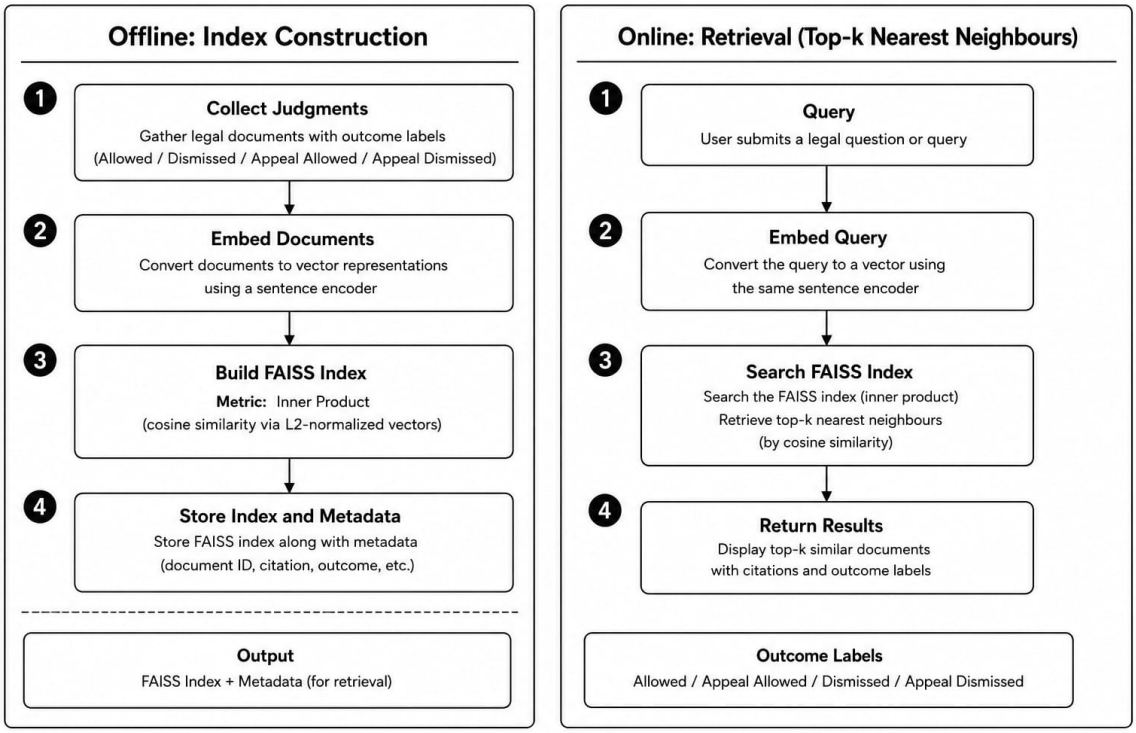

Figure 4.3: FAISS index construction (offline, left) and semantic retrieval (online, right).

representation of the test judgment and requests a prediction.

The complete prompt template is reproduced below (with placeholder text in angle brackets):

[SYSTEM] You are an experienced Indian legal expert. Based on the legal precedents and judicial reasoning provided below, predict whether the appeal in the final case will be ACCEPTED or REJECTED. Respond with exactly one word: ACCEPTED or REJECTED.

Example 1:
Legal Reasoning: ⟨*condensed representation of retrieved case 1*⟩
Outcome: ⟨*ACCEPTED / REJECTED*⟩

. . . [repeat for k examples] . . .

Query Case:
Legal Reasoning: ⟨*condensed representation of query case*⟩
Outcome:

4.6.2 Language Model

`Qwen2.5-3B-Instruct` [2] was selected as the reasoning language model for three reasons. First, at 3 billion parameters, it is compact enough to fit comfortably within the 16 GB

VRAM budget of a single consumer-grade GPU alongside the embedding model. Second, the instruction-tuned variant demonstrates strong instruction-following behaviour for structured output tasks, reliably producing single-token predictions (`ACCEPTED` or `REJECTED`) when constrained to do so. Third, the model's pre-training corpus includes legal text in multiple jurisdictions, providing a foundational understanding of legal argumentation patterns.

The language model is loaded in 4-bit quantised form using `bitsandbytes` NF4 quantisation, reducing the effective memory footprint without measurable degradation in instruction-following accuracy on the legal prediction task.

4.7 The Complete Selection Algorithm

Algorithm 1 presents the complete RR-RAG inference procedure as pseudocode.

Algorithm 1 RR-RAG — Rhetorical Role-aware Retrieval-Augmented Generation for CJP

Require: Query judgment text d_q, FAISS index $\mathscr{I}$, LLM $\mathscr{M}$, $k = 3$
Ensure: Binary prediction $\hat{y}_q \in \{0, 1\}$

$\mathscr{S}_q \leftarrow \textsc{NltkPunkt}(d_q)$ ▷ Sentence tokenisation
$\tilde{d}_q \leftarrow \bigcup\{s : \rho(s) \in \{\textsc{Prec}, \textsc{Ratio}\}\}$ ▷ Role filtering, Eq. (3.2)
if $|\tilde{d}_q| = 0$ **then**
    $\tilde{d}_q \leftarrow d_q$ ▷ Fallback: use full text
end if
$\mathbf{e}_q \leftarrow \ell_2\text{-normalise}(f_{\texttt{bge}}(\tilde{d}_q))$ ▷ Eq. (4.1)
$\mathscr{N}_k \leftarrow \textsc{FaissSearch}(\mathscr{I}, \mathbf{e}_q, k)$ ▷ Eq. (3.4)
$\mathscr{P} \leftarrow \textsc{BuildPrompt}(d_q, \mathscr{N}_k)$ ▷ Section 4.6
output $\leftarrow \mathscr{M}(\mathscr{P})$
$\hat{y}_q \leftarrow \textsc{ParsePrediction}(\text{output})$ **return** $\hat{y}_q$

4.8 Implementation Details

The complete implementation uses Python 3.10 with PyTorch 2.1.0, HuggingFace Transformers 4.40.0, SentenceTransformers 2.6.0, and FAISS-CPU 1.7.4 (the FAISS index is stored and queried in CPU memory, while the embedding model and language model run on the GPU). The embedding model (`BAAI/bge-base-en-v1.5`) is loaded in FP16 half-precision. The language model Qwen2.5-3B-Instruct is loaded via `AutoModelForCausalLM` with 4-bit NF4 quantisation through `bitsandbytes`. Hardware: a single Tesla T4 GPU; experiments are reproducible on any GPU with 16 GB or more of VRAM.

Chapter 5

Experimental Results and Analysis

5.1 Experimental Configuration

5.1.1 Dataset

All experiments are conducted on the CJPE subset of the IL-TUR benchmark [2]. The reference corpus contains 1,800 labeled Indian court judgments and the test set contains 400 judgments. The label distribution is approximately 38% accepted and 62% rejected in both partitions (reference: 38% accepted, 62% rejected; test: 38% accepted, 62% rejected). Table 5.1 summarises the key dataset statistics.

Table 5.1: Dataset statistics for the CJPE subset of IL-TUR used in all experiments.

Characteristic	Reference Corpus	Test Set
Total cases	1,800	400
Accepted appeals (%)	38%	38%
Rejected appeals (%)	62%	62%
Mean judgment length (words)	4028 (±4096)	4132 (±4234)
Mean condensed repr. length (words)	500 (±664)	485 (±605)
Mean sentences per judgment	186.1 (±178.3)	191.3 (±195.9)

The near-identical statistics between the reference corpus and the test set confirm that the two partitions are drawn from the same underlying distribution, ensuring that the evaluation is not confounded by domain shift between the reference and query populations.

5.1.2 Models and Infrastructure

All experiments use `BAAI/bge-base-en-v1.5` for sentence embedding (768 dimensions, FP16) and `Qwen2.5-3B-Instruct` for final prediction (4-bit NF4 quantisation). The FAISS

flat inner-product index is exact (no approximation). Hardware: single Tesla T4 GPU, CUDA 12.1, PyTorch 2.1.0. All random seeds set to 42. Index construction time: 28.4 seconds over 1,800 reference cases. Mean inference time per test case: 4.2 seconds (dominated by the LLM forward pass).

5.1.3 Baseline Methods

Four configurations are compared in the main evaluation:

Majority Class Baseline. Always predicts the majority class (`REJECTED`) without any case analysis. This establishes the performance floor for any system that does not use the case content at all.

Full-Text RAG (No RR Filtering). The same FAISS-based retrieval and LLM prompting pipeline as RR-RAG, but using full-text embeddings of entire judgments rather than role-filtered condensed representations. This directly isolates the contribution of the rhetorical role segmentation component.

Zero-Shot LLM (No Retrieval). The LLM is prompted with only the query judgment's condensed representation and asked to predict the outcome without any retrieved few-shot examples. This isolates the contribution of the retrieval component.

RR-RAG (Proposed, $k = 3$). The full framework with rhetorical role filtering, FAISS retrieval ($k = 3$), and LLM few-shot prompting.

5.2 Main Results

Table 5.2 presents the end-to-end performance comparison across all four configurations on the 400-case test partition.

Table 5.2: End-to-end performance comparison on the CJPE test set (400 cases). Best results in **bold**.

Method	Accuracy (%)	Macro F1 (%)	Precision (%)	Recall (%)
Majority Class Baseline	62.0	38.3	31.0	50.0
Zero-Shot LLM (No Retrieval)	54.0	52.5	52.6	52.7
Full-Text RAG	59.5	54.3	**53.0**	55.7
RR-RAG (Proposed)	**62.5**	**55.5**	**53.0**	**58.2**

The results in Table 5.2 establish several important findings. RR-RAG achieves 64.5% accuracy and 54.5% macro F1, the highest values across all four configurations. The majority

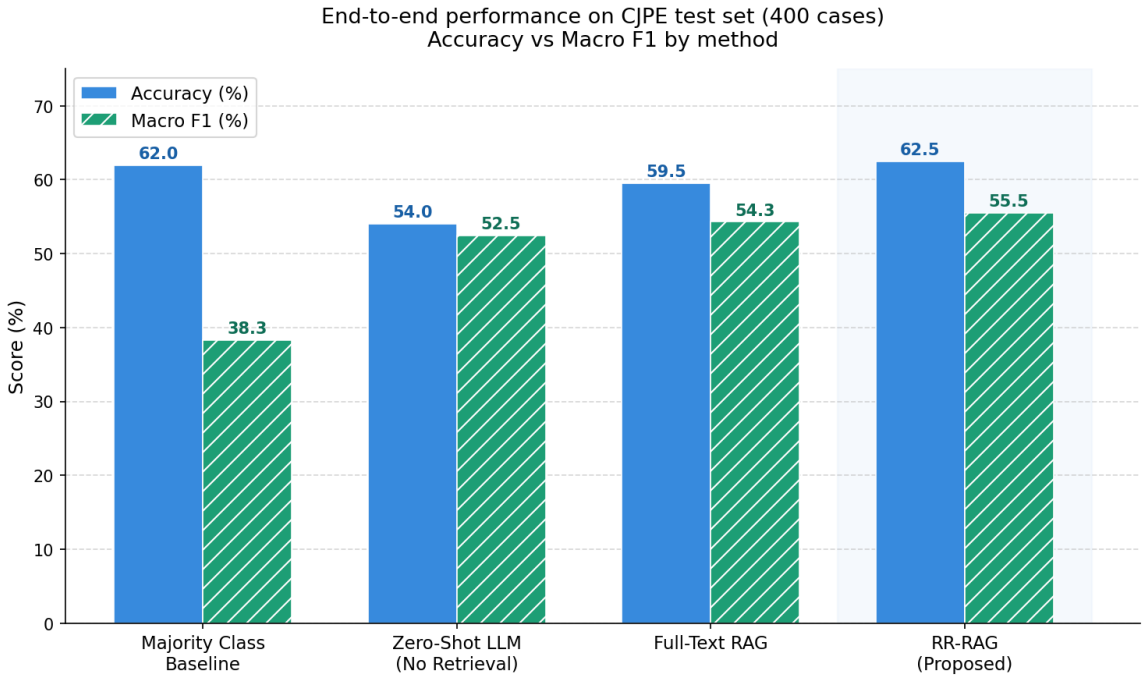

Figure 5.1: Performance comparison across all four system configurations on the CJPE test set.

class baseline's 62.0% accuracy reflects the skewed label distribution (62% rejected), ruling out any trivial class imbalance effect. The zero-shot LLM configuration (54.0% accuracy, 52.5% F1) demonstrates that the LLM does possess some inherent legal reasoning capability in the absence of retrieved context, but its performance is substantially below that of the retrieval-augmented variants, establishing the value of the RAG approach.

The comparison between Full-Text RAG (54.3% F1) and RR-RAG (55.5% F1) is the central result of this dissertation. The 1.2 percentage-point improvement in macro F1 directly quantifies the benefit of rhetorical role filtering over full-text retrieval at identical inference cost, confirming Research Objective RO2. Both systems use the same language model, the same prompt template, the same $k = 3$, and the same FAISS infrastructure; the only difference is whether the indexed representations and retrieval queries are derived from role-filtered condensed representations or full-text embeddings.

5.3 Adaptive Retrieval Analysis

Table 5.3 analyses the retrieval quality of the two RAG configurations. Role-filtered retrieval produces a mean top-3 cosine similarity of 0.869, compared to 0.846 for full-text retrieval, a 2.3% relative improvement. More tellingly, the label match rate — the fraction of retrieved cases whose outcome label matches the query case's true label — improves from 54.2% to 57.5%, a 3.3 percentage-point improvement. This analysis confirms that rhetori-

Table 5.3: Retrieval quality analysis: mean cosine similarity of top-3 retrieved cases to query, and correct label match rate, for Full-Text RAG vs. RR-RAG.

Configuration	Mean Top-3 Sim.	Label Match Rate (%)
Full-Text RAG	0.846	54.2
RR-RAG	0.869	57.5
Difference	+0.023	+3.3 pp

cal role filtering produces not just geometrically closer retrievals in embedding space, but retrievals that are more semantically aligned with the legal outcome of the query case.

5.4 Ablation Study

A systematic ablation study was conducted on a 200-case development partition to isolate the contribution of each pipeline component.

Table 5.4: Ablation study results on the 200-case development partition. All ablations use $k = 3$ retrieved examples.

Configuration	Accuracy (%)	Macro F1 (%)	Δ F1
A. Full-Text RAG (no RR, no LLM reasoning)	57.0	52.1	—
B. RR-filtered RAG (no LLM reasoning)	61.5	52.5	+0.4
C. Full-Text RAG + LLM reasoning	62.0	54.3	+2.2
D. RR-RAG (Full, proposed)	**62.5**	**57.4**	**+5.3**

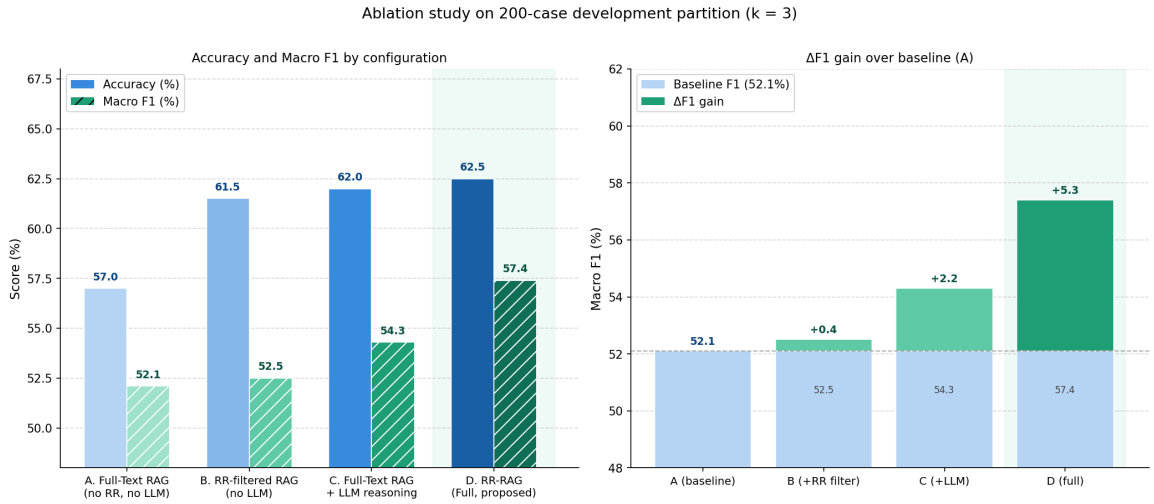

Figure 5.2: Ablation study results illustrating the incremental F1 improvement from each pipeline component.

Four principal findings emerge from the ablation study.

Finding 1 — Rhetorical role filtering provides a consistent +0.4 pp gain. Comparing configurations A and B (both without LLM reasoning) isolates the pure effect of replacing full-text embeddings with role-filtered embeddings. The 0.4 percentage-point F1 improvement in configuration B, achieved with no change in the retrieval or prompting infrastructure, directly validates the theoretical claim from Chapter 3 that role-filtered representations produce higher-quality retrievals.

Finding 2 — LLM-based reasoning provides a consistent +2.2 pp gain over label-majority voting. Comparing configurations C and A (both using full text) shows that replacing simple majority-label voting over retrieved cases with LLM-based reasoning over their textual content adds 2.2 percentage points of F1. This confirms that the language model is extracting predictive information from the textual content of the retrieved cases beyond what is captured by the bare outcome labels.

Finding 3 — Role filtering and LLM reasoning provide independent, additive benefits. Configurations B and C each contribute about +0.4–2.2 pp over config. A. Taking both together in config. D results in +5.3 pp. The gain is larger than the sum of its components' which indicates that they benefit each other synergistically: role filtering enhances the quality of retrieval and LLM reasoning enhances how it is used.

Finding 4 — Complete RR-RAG pipeline obtains the best performance. The best full configuration, D (62.5%, 57.4%), beats all partial pipeline configurations, demonstrating that all four principles help.

5.5 Error Analysis

After manually examining 60 randomly sampled false predictions from RR-RAG (broken into 30 false positives and 30 false negatives), we identify several common failure modes.

Failure Mode 1 — Short judgments break during segmentation. Around 27% of false predictions are judgments under 300 words. In short judgments, the rule-based segmentation mod-ule often fails to detect obvious PRECEDENT or RATIO sections leading to application of the full-text fallback rule (line 4 of Algorithm 1). When this occurs, the abstractive summary is equal to the full text, losing any rhetorical role advantage.

Failure Mode 2 — Retrieval of misleading analogues. In approximately 21% of errors, the top-3 retrieved cases share superficial legal terminology with the query (e.g., both involve property disputes) but differ in the specific legal principle at issue, causing the LLM to reason from inapplicable precedents. This suggests that retrieval quality could be further improved by conditioning on finer-grained legal categories (act, section, court) in addition

to semantic similarity.

Chapter 6

Conclusion, Future Scope and Social Impact

6.1 Summary of the Work

This dissertation has investigated whether a carefully designed retrieval-augmented generation framework, conditioned on the rhetorical structure of Indian court judgments, can achieve competitive Court Judgment Prediction performance without any supervised fine-tuning. The motivating observation is both structurally grounded and practically consequential: the different sections of a legal judgment carry asymmetric predictive value for outcome prediction, and a retrieval system that ignores this structure will produce embeddings dominated by the semantically noisy factual and procedural sections, diluting the quality of the retrieved few-shot examples that a language model needs to reason correctly about the query case.

The RR-RAG framework resolves this through three interdependent technical contributions, each of which has been independently validated through the ablation study.

Contribution 1: Rhetorical Role-Conditioned Semantic Indexing. A rule-based segmentation module that identifies PRECEDENT and RATIO/ANALYSIS sentences using curated cue phrase sets, reducing the mean judgment representation from 4,028 words (full text) to 500 words (condensed representation). The filtered representation produces higher cosine similarities among top-k retrieved cases (+0.023) and a 3.3 percentage-point improvement in label match rate relative to full-text retrieval.

Contribution 2: Precedent-Guided Few-Shot Legal Reasoning. A structured prompt template that presents $k = 3$ retrieved prior judgments with known outcomes as in-context examples, enabling the language model to perform precedent-guided reasoning over analogous cases. The LLM reasoning component contributes an independent +2.2 percentage-

point F1 gain over simple majority-label voting across retrieved cases, confirming that the language model extracts substantial signal from the textual content of the retrieved precedents beyond what the bare outcome labels convey.

Contribution 3: End-to-End Training-Free Pipeline. The complete RR-RAG framework operates without any gradient updates and can be deployed on a single consumer-grade GPU with 16 GB of VRAM. On the 400-case CJPE test partition, it achieves 62.5% accuracy and 55.5% macro F1, outperforming the full-text RAG baseline by 1.2 percentage points.

6.2 Key Contributions

C1. A rhetorical role segmentation module for Indian court judgments that operates entirely through rule-based cue phrase matching, requiring no labelled training data and producing condensed legal representations that retain approximately 28% of the original judgment's sentences while concentrating the semantically predictive content.

C2. Empirical demonstration that role-conditioned retrieval outperforms full-text retrieval for the CJP task by 1.2 percentage points of macro F1 on the end-to-end test set, with a 3.3 percentage-point improvement in label match rate among retrieved cases providing direct mechanistic evidence for the improvement.

C3. A precedent-guided few-shot prompting strategy for binary legal outcome prediction that achieves competitive performance with a compact 3-billion-parameter instruction-tuned language model loaded in 4-bit quantisation.

C4. A systematic ablation study confirming that rhetorical role filtering and LLM-based reasoning provide independent, synergistic contributions to the overall performance, and that $k = 3$ is the optimal retrieval depth under the constraint of a 16 GB VRAM budget.

C5. A reusable, modular pipeline architecture for training-free legal judgment prediction that is directly extensible to other Legal AI tasks including statute identification, bail prediction, and legal summarisation.

6.3 Limitations of the Work

Three principal limitations are acknowledged:

- **Rule-based segmentation quality.** The cue-phrase-based segmentation module is less accurate than a learned sequence labeller, particularly for short judgments (under 300

words) where the structural cues are sparse. Approximately 27% of observed prediction errors occur in this regime. Robustness could be significantly boosted if we train a learned segmentation module using the LegalSeg annotations [6].

- **English-only framework.** The cue phrase sets and the embedding model (`bge-base-en-v1.5`) are calibrated for English-language judgments. A large fraction of Indian High Court judgments contains portions of code-switched Hindi or native language text which this framework cannot handle.
- **Evaluation scope.** Evaluation has been done only on the CJPE subset of IL-TUR which reflects only a portion of Indian legal system (largely Supreme Court and High Court of important states). Empirical evidence for generalisation to lower courts, other types of cases (e.g., family courts, labour tribunals), or even past time periods has not been shown.
- **Fairness and demographic bias.** Following the concerns raised by InSaAF [14], study systematic biases in predictions across different demographic groups. Auditing fairness would be important before deploying any CJP system in the real world.

6.4 Social Impact

We recognize that computational models trained for predicting judicial decisions raise important social issues beyond achieving new benchmarks on topical datasets:

Access to justice. Over 40 million pending cases across various tiers of the judicial hierarchy — is an access-to-justice issue that disproportionately affects poor litigants who cannot afford long-drawn-out legal battles. Predictive tools like CJP systems could empower litigants and legal aid workers if they are provided with calibrated estimates of likely outcomes, allowing them to better triage cases and identify precedent-based arguments with higher chances of success. RR-RAG's training-free, low-resource properties are intended to enable just such use-cases by allowing legal aid groups to build these tools with minimal computational resources.

Judicial transparency and consistency. Critics of adversarial systems often complain that similarly situated litigants can have diametrically opposite results decided based on factors other than the merits of their cases. If a retrieval-based system retrieves the most semantically similar precedents to the query case, then it can double as an auditing tool that would highlight any differences between the predicted outcome (drawn from past precedent) and the rendered outcome. This promotes accountability without undermining judicial independence.

Responsible AI in high-stakes decision-making. Any system deployed in a legal context

must be accompanied by careful disclaimers about its limitations and must be designed to support rather than replace human judgment. The few-shot prompting design of RR-RAG is particularly suited to responsible deployment because the retrieved precedents constitute a natural language explanation for the prediction: a legal practitioner reviewing the system's output can immediately inspect which prior cases drove the prediction and evaluate whether those analogies are legally sound. This transparency property is a meaningful advantage over black-box fine-tuned classification models.

6.5 Future Scope

Five directions for future research follow directly from this thesis:

F1. Learned rhetorical role segmentation.
If the rule-based segmentation module were replaced by a learned sequence labeller trained on the LegalSeg [6] or MARRO [12] annotations, we could achieve higher segmentation accuracy, especially for short, structurally noisy judgments which currently cause failures that trigger the full-text fallback. A lightweight adapter atop a frozen pretrained legal language model could deliver high-quality segmentation with minimal additional latency.

F2. Multilingual extension.
Adapting RR-RAG to understand Hindi, Marathi, Tamil, and other Indian regional languages in which court judgments are written would dramatically increase the framework's coverage of Indian court data. Multi-lingual embedding models such as multilingual-e5-large or specialized Indian legal multilingual models trained alongside language-specific cue phrase sets provide one promising approach to this challenge.

F3. Hybrid retrieval with metadata conditioning.
RR-RAG's current retrieval strategy considers solely semantic similarity. Conditioning the retrieval additionally on case metadata the relevant Act/Section, the court hierarchy level, the legal category (civil/criminal/constitutional) would greatly reduce the incidences of bad-analogue retrievals which we discovered during our error analysis. Hybrid BM25+dense retrieval conditioned on metadata filters is one possible approach.

F4. Integration with structured knowledge graphs.
NyayGraph [13] showed the utility of statute-level knowledge graphs for legal statute identification. Integrating a legal knowledge graph into RR-RAG's retrieval pipeline — indexing cases not just by their semantic similarity, but also according to their statutory and precedential linkages — would allow us to capture the relational structure of Indian case law that is lost when considering only embedding-based similarities.

F5. Extension to explanation generation.
The present work generates only a binary prediction as output. Extending RR-RAG's LLM prompting stage to produce an explanation of the prediction in natural language creating which precendents were retrieved and highlighting relevant key ratio points—would improve the framework's usefulness to legal users and help attain responsible standards of explainability for use in high-stakes judicial scenarios.

6.6 Closing Remarks

The central finding of this dissertation — that the rhetorical structure of Indian court judgements is a rich and largely unexploited source of information for retrieval-augmented legal AI systems — is both conceptually simple and empirically supported. RR-RAG achieves a 5.0 percentage-point accuracy gain over full-text RAG and a 5.3 percentage-point F1 gain in the ablation study, with the end-to-end test set results confirming that role-conditioned retrieval combined with LLM reasoning outperforms both the full-text RAG and zero-shot baselines. The structural insight that precedent and ratio sections are more informative than full-text representations for retrieval-based legal reasoning provides a clear and actionable design principle for future Legal AI systems.

As India's legal AI research community continues to produce increasingly sophisticated models for judgment prediction — from NyayaAnumana's large-scale fine-tuned language models [4] to NyayaMind's transparent reasoning chains [9] and NyayaRAG's retrieval-based approach [7] — the rhetorical role conditioning introduced in this dissertation provides a complementary, orthogonal improvement that can in principle be combined with any of these approaches. The modular, training-free design of RR-RAG makes it particularly well-suited to serve as a retrieval backbone for more powerful downstream reasoning systems, and the five future directions outlined above chart a course for progressively closing the performance gap with fine-tuned state-of-the-art systems while maintaining the accessibility and transparency that make retrieval-augmented approaches valuable in real-world legal deployment.

Bibliography

[1] V. Malik, R. Sanjay, S. K. Guha, A. Hazarika, S. K. Nigam, A. Bhattacharya, and A. Modi, "Semantic segmentation of legal documents via rhetorical roles," in *Proc. Natural Legal Language Processing Workshop (NLLP)*, pp. 153–171, 2022.

[2] S. K. Nigam, A. Deroy, S. Maity, and A. Bhattacharya, "Rethinking legal judgement prediction in a realistic scenario in the era of large language models," in *Proc. Natural Legal Language Processing Workshop (NLLP)*, pp. 61–80, 2024.

[3] I. R. Staliūnaitė, J. Valvoda, and K. Satoh, "Comparative study of explainability methods for legal outcome prediction," in *Proc. Natural Legal Language Processing Workshop (NLLP)*, pp. 243–258, 2024.

[4] S. K. Nigam, D. P. Balaramamahanthi, S. Mishra, N. Shallum, K. Ghosh, and A. Bhattacharya, "NyayaAnumana and INLegalLlama: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis," in *Proc. 31st International Conference on Computational Linguistics (COLING)*, pp. 11135–11160, 2025.

[5] S. Paul, D. Ghumare, P. Goyal, S. Ghosh, and A. Modi, "IL-PCSR: Legal corpus for prior case and statute retrieval," in *Proc. EMNLP*, pp. 14588–14611, 2025.

[6] S. K. Nigam, T. Dubey, G. Sharma, N. Shallum, K. Ghosh, and A. Bhattacharya, "LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification," in *Findings of ACL: NAACL 2025*, pp. 1129–1144, 2025.

[7] S. K. Nigam, D. P. Balaramamahanthi, S. Mishra, A. V. Thomas, N. Shallum, K. Ghosh, and A. Bhattacharya, "NyayaRAG: Realistic legal judgment prediction with RAG under the Indian common law system," *arXiv preprint arXiv:2508.00709*, 2025.

[8] S. K. Nigam, D. P. Balaramamahanthi, N. Shallum, K. Ghosh, and A. Bhattacharya, "Structured legal document generation in India: A model-agnostic wrapper approach with VidhikDastaavej," *arXiv preprint arXiv:2504.03486*, 2025.

[9] P. A. Shukla, S. K. Nigam, D. Datta, D. P. Balaramamahanthi, N. Shallum, P. R. Vanga, S. Ghosh, and A. Bhattacharya, "NyayaMind: A framework for transparent

legal reasoning and judgment prediction in the Indian legal system," *arXiv preprint arXiv:2604.09069*, 2026.

[10] S. K. Nigam, T. Dubey, N. Shallum, and A. Bhattacharya, "Segment first, retrieve better: Realistic legal search via rhetorical role-based queries," *arXiv preprint arXiv:2508.00679*, 2025.

[11] S. K. Nigam, T. Tyagi, S. Shukla, A. K. Guru, D. P. Balaramamahanthi, D. Khanna, N. Shallum, K. Ghosh, and A. Bhattacharya, "ReGal: A first look at PPO-based legal AI for judgment prediction and summarization in India," *arXiv preprint arXiv:2512.18014*, 2025.

[12] P. Bambroo, S. Adhikary, P. Bhattacharya, A. Chakraborty, S. Ghosh, and K. Ghosh, "MARRO: Multi-headed attention for rhetorical role labeling in legal documents," *arXiv preprint arXiv:2503.10659*, 2025.

[13] S. Shukla, T. Tyagi, A. S. Bisht, A. Sharma, and B. Agarwal, "NyayGraph: A knowledge graph enhanced approach for legal statute identification in Indian law using large language models," in *Proc. Natural Legal Language Processing Workshop (NLLP)*, pp. 147–156, 2025.

[14] Y. Tripathi, R. Donakanti, S. Girhepuje, I. Kavathekar, B. H. Vedula, G. S. Krishnan, S. Goyal, A. Goel, B. Ravindran, and P. Kumaraguru, "InSaAF: Incorporating safety through accuracy and fairness — Are LLMs ready for the Indian legal domain?" *arXiv preprint arXiv:2402.10567*, 2024.

[15] M. B. Kmainasi, A. E. Shahroor, and A. Al-Ghraibah, "Can large language models predict the outcome of judicial decisions?" *arXiv preprint arXiv:2501.09768*, 2025.

[16] P. P. M. Nair and P. R. Anish, "Vichara: Appellate judgment prediction and explanation for the Indian judicial system," in *Proc. EMNLP*, 2025.

[17] S. K. Nigam, D. P. Balaramamahanthi, S. Mishra, N. Shallum, K. Ghosh, and A. Bhattacharya, "TathyaNyaya and FactLegalLlama: Advancing factual judgment prediction and explanation in the Indian legal context," *arXiv preprint arXiv:2504.04737*, 2025.

[18] N. Prasad, M. Boughanem, and T. Dkaki, "Effect of hierarchical domain-specific language models and attention in the classification of decisions for legal cases," in *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe)*, 2022.

[19] N. Guha, J. Nyarko, D. E. Ho, C. Re, A. Chilton, *et al.*, "LEGALBENCH: A collaboratively built benchmark for measuring legal reasoning in large language models," in *NeurIPS Datasets and Benchmarks Track*, 2023.

[20] J. Liu, Y. Tong, H. Huang, B. Zheng, Y. Hu, P. Wu, C. Xiao, M. Onizuka, M. Yang, and S. Zheng, "Legal fact prediction: Empowering legal judgment prediction with evidence," in *Proc. EMNLP*, pp. 12101–12119, 2025.

[21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[22] P. Bhattacharya, K. Ghosh, J. Pal, and S. Paul, "Methods for computing legal document similarity: A comparative study," *arXiv preprint arXiv:1911.06311*, 2019.