International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I10P118

Volume 74 Issue 10, 250-265, October 2026
© 2026 Seventh Sense Research Group®

Original Article

Robust Cross-Lingual Legal Question Pair Similarity Analysis for English-Marathi

Hirkani Pathak[1], Milind Penurkar[2], Aatish Daryapurkar[3]

[1,2]*Department of Computer Science and Engineering, Indian Institute of Information Technology, Nagpur, Maharashtra, India*
[3]*Department of Basic Sciences and Engineering, Indian Institute of Information Technology, Nagpur, Maharashtra, India*

[1]*Corresponding Author: padwadhirkani10@gmail.com*

Received: 30 April 2026 Revised: 25 August 2026 Accepted: 29 August 2026 Published: 27 October 2026

Abstract - *Cross-lingual legal assistants in India process user queries in both English and Marathi while adhering to deployment speed requirements. A single model for cross-lingual and lightweight matching of legal question pairs is proposed, which consists of three modules: (i) contextualized semantics via a multilingual Transformer encoder, (ii) language-independent legal anchoring signals in the form of statutes/sections/articles in Latin and Devanagari scripts, and (iii) a sentence-embedding branch learned via knowledge distillation from teacher models. The model is trained using a joint classification and ranking loss in combination with a citation format invariance regularizer. The model's performance is tested using a deployment-faithful protocol in which the decision threshold is tuned on English validation and then kept fixed for cross-lingual evaluation. On the English benchmark set comprising 1,233 pairs, the model achieves an F1 score of 0.964 and an AUC score of 0.996 in 1.24 ms per pair. An LLM classifier achieves a higher F1 score of 0.986, but at a slower rate of 14.6 ms per pair. For the cross-lingual setting where a fixed English-selected threshold is used, the approach achieves an F1 value of 0.935 for English and 0.852 for Marathi ($\Delta F1 = 0.083$), while keeping a consistent rank quality (MRR ≈ 0.82). Given a small amount of labeled samples of Marathi, the F1-score for Marathi rises to 0.867, with $\Delta F1$ equaling 0.055. This result implies that the use of multilingual semantics along with light-weight legal anchoring leads to solid cross-lingual performance in real-world legal information retrieval and question-answering applications.*

Keywords - *English-Marathi, Legal citation normalization, Legal question pair similarity, Multilingual transformers, Semantic similarity.*

1. Introduction

Legal NLP has progressed from proof-of-concept text classification tasks to more useful tasks like precedent search, statute navigation, legal issue spotting, and the first level of legal question answering. An important common element in these tasks is question pair similarity: it must be determined if two questions are asking the same legal thing, even if they look different on the surface. This is necessary for de-duplicating user queries, grouping complaints, searching precedents, and routing queries into the correct legal category.

What makes legal language different from other language tasks is not just vocabulary, but structure as well. Legal questions are full of statutory references like IPC, CrPC, and CPC, as well as section and article references, and formulaic language usage. This means two legal questions may contain a lot of shared words, yet still be different, as a section number may change between two queries, or two queries may be identical in meaning with very few shared words. In addition, in an Indian context, multilingual usage must be dealt with: users may pose queries in Marathi (or other Indian languages), even when the underlying legal resources are mostly in English. A model that performs well on English-only benchmark sets but fails under language shift will not be useful in a real-world context.

It is intended to fill the aforementioned research gap through the development of a composite legal similarity framework consisting of attention-pooled legal embeddings (LTEAM), Concept-Centric Representation Learning (GRLC), and Heterogeneous Transfer and Distillation (HETLM). A cross-lingual extension of this proposed framework is also considered, with an evaluation strategy closer to practical use cases - i.e., training on English, tune a threshold on English validation, and then test on Marathi and mixed-language inputs without further fine-tuning. Current methods treat legal-domain semantics, cross-lingual transfer, translation, and efficient retrieval as separate issues. LEGAL-BERT and InLegalBERT have been developed to improve the representation of legal text [3, 4], whereas XLM-R, LaBSE, and multilingual E5 enable cross-lingual transfer [7-9]. Although translation-based approaches like IndicTrans2 are

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

feasible, they can change legally significant terminology and citation formats [11, 22]. There is, therefore, still a need for a lightweight legal similarity model capable of performing English-to-Marathi translation while retaining information about statutes, sections, and articles.

The combination of multilingual semantic representations, explicit legal anchors, compact distillation, citation-format invariance, and ranking supervision, alongwith the evaluation done in fixed-threshold English-to-Marathi and mixed-language scenarios contribute to the novelty of this work.

Legal question pair matching is at the heart of practical legal AI applications like query deduplication, issue retrieval, precedent identification, and legal QA. The domain of law is marked by great rigidity; the slightest change in the statute family (for instance, between IPC and CrPC/CPC), in the identifiers (section number/article number), or even in the procedural setting (for instance, bail/quash/maintainability) could change the entire meaning of the provision despite little or no change in the substance of the provision itself.

This is particularly challenging for the Indian legal domain since users tend to input queries in Marathi or a mix of Marathi and English, while the majority of the searchable data is available only in English. In the experiments, one common heuristic tried was to translate the input from Marathi to English and then perform matching. However, it was found that this was not reliable for legal similarity matching primarily because the process of translation tends to alter the very words that are legally significant (citations, abbreviations, section numbers, etc.).

The contributions of this work are listed below.

1. A large leakage-safe legal question-pair dataset - A binary legal similarity question-pairs dataset containing 94,133 question pairs, drawn from two sources of legal discourse on YouTube and question-like content derived from EUR-Lex, is provided. The dataset is de-duplicated, clustered to remove near-duplicate question-pairs, and question-pairs are strictly split to avoid leakage. Expert validation of label quality is supported by a balanced expert set of 1,200 question pairs with a high agreement rate $\kappa/\alpha \approx 0.933$.
2. A robust, lightweight cross-lingual legal similarity model - A composite matcher is introduced, that integrates multilingual encoding with legal anchor-aware representations and efficiency-oriented training (including distillation) to address the failure modes that are most important in the legal domain, including citation surface variation, statute and section anchors, and procedural cues.
3. A deployment-consistent evaluation of robustness for English-Marathi legal matching- A strict evaluation process is used for the EN-EN, EN→MR, and mixed language setups. The threshold value for English is selected and kept fixed for Marathi. Performance evaluation includes F1, AUC, MRR, and P@5/10, in addition to a performance drop under the influence of shift (ΔF1 and ΔMRR). Furthermore, a headroom evaluation technique based on the oracle threshold is also performed.
4. Evidence-based Analysis - Baseline Comparison, Ablations, and Legal Error Taxonomy - the experiments include baseline comparisons for sparse, neural, and Large Language Models (LLMs); ablative experiments (such as the omission of invariance); and qualitative analysis, including error typology classification for legal errors-namely statute mismatch, section/article confusion, procedural/substantive mismatch, and topical/true paraphrase.

2. Critical Review of Existing Models Used for Enhancing Question Pair Similarity Analysis

Legal NLP, in terms of judgment prediction, summarization, retrieval, and most recently, LLM-based legal assistant tools, has evolved rapidly in recent years. Nevertheless, in legal QA systems and legal precedent retrieval, a specific task keeps proving to be a bottleneck: semantic matching of legal question pairs, i.e., determining if two questions are asking about the same legal issue, as opposed to asking about different issues with common vocabulary or topic.

The problem becomes particularly difficult to deal with in law, as the concept of equivalence is grounded in solid anchors such as the statute family, procedural posture, section, and article numbering, which are usually expressed using inconsistent abbreviations or in mixed languages.

While most approaches are based on a variant of the Transformer, with a common starting point being contextual encoders like those in BERT-style approaches, pretraining on a legal dataset improves performance in the legal domains, leading to LEGAL-BERT and InLegalBERT-style pretraining on Indian law corpora [1, 3, 4].

Cross-encoders are commonly employed for pair similarity, as they are more accurate, while bi-encoders are more frequently used for retrieval, with SBERT as a widely adopted framework for bi-encoder training on sentence embeddings for pair similarity [2]. Cross-encoder re-ranking with BERT further confirmed the power of joint encoding for pair similarity, especially when accuracy is a priority [6]. From a retrieval viewpoint, lexical approaches are surprisingly strong in law, as they are able to perform citation and statute matching, with BM25 still a relevant first-stage comparator, and sparse neural retrieval approaches like SPLADE acting as a bridge between lexical matching and learned expansions [10].

These approaches offer different accuracy-efficiency trade-offs. The legal-domain models improve the representation of legal terms, the cross-encoders are best for classification accuracy, the bi-encoders are good for efficient retrieval, and the lexical/sparse methods preserve the exact legal citations. But these advantages do not by themselves guarantee the preservation of legal equivalence across different languages or scripts.

LexGLUE serves as a benchmark that proves the notion that there is significant variation in the legal task performance, thus making the existence of strong baselines necessary [20]. COLIEE, which is a practical legal competition, focuses on retrieval and shows that ranking metrics fit naturally to the system's behavior [21]. This retrieval-oriented perspective is also supported by recent work on cost-efficient legal precedent retrieval, which shows that retrieval quality must be studied together with computational feasibility in practical legal systems [14].

Cross-lingual and mixed-language matching poses another level of difficulty. The multi-lingual encoder models, such as XLM-R, are an excellent base for transfer where there is not much labeled data available in the target language, while, on the other hand, multi-lingual sentence embedding models, including LaBSE and multilingual E5, have become standard baselines for cross-lingual similarity and retrieval [7-9].

Translation-based approaches like “translate then match” are still a popular pragmatic approach, and for Indian languages, IndicTrans2 provides a viable open translation option [11]; however, existing studies on legal translation have also highlighted brittleness in terminology, citation styles, and legal phrasing, which can cause distortion in meaning even with fluent translations [22].

This is in accordance with existing robustness-focused legal NLP studies, where models are shown to be brittle under realistic conditions; existing studies on adversarial legal text generation directly demonstrate that high accuracy in the domain does not directly translate into robustness under stress [15, 16]. While translate-then-match systems begin by translating the query into English, multilingual encoders directly align the cross-language semantics.

In the case of legal text, translation can lead to changes in abbreviations, section numbers, citation formats, procedural terms, or negation. It is for this reason that cross-lingual matching must also involve the preservation of legally important surface information. At the same time, LLM-based approaches are being explored for legal question generation and interpretation classification, which indicates a move toward generative legal AI [17, 18]. Use of efficient finetuning approaches like QLoRA makes these baselines more accessible for reproduction, and existing open LLM families like Qwen2.5 are considered typical baselines for instruction tuning [12, 13]. Yet, latency and deployment factors are what drive lightweight architectures, which are robust without relying on heavy inference.

The recent advancements achieved by the retrieval and representation learning approaches provide a rich context for the present work. ColBERT proposed the concept of late interaction as a powerful alternative between the efficiency of the bi-encoder and the discriminative ability of the cross-encoder, especially for the matching tasks related to the retrieval process [23].

Dense Passage Retrieval (DPR) proved the effectiveness of the dual encoder approach for the improvement of the first-stage retrieval process, going beyond the lexical matching [24]. MPNet and MiniLM are closely related to the current work, as they proposed powerful teacher and efficient student representations for the semantic matching process [25, 26].

Another important contribution, related to the current work, is the effectiveness of the sample-efficient pretraining for the smaller encoder, as proposed by the ELECTRA model [27]. ColBERT and DPR demonstrate efficient retrieval through late interaction and dual encoding, while MPNet and MiniLM support strong compact semantic representations and teacher-student transfer. However, these approaches do not specifically combine efficiency with bilingual legal-anchor preservation.

Finally, the Longformer model is a powerful reference for the long context representation, especially for the legal NLP tasks, as the legal texts contain a lot of structure and are often long [28]. The recent works related to the legal judgment prediction [29] and the automated consistency analysis of contracts [30] emphasize the importance of precise legal AI systems, considering the sensitivity and the feasibility of the proposed approaches.

Related evidence from Indian legal judgment classification also shows that embeddings and ensemble strategies can materially improve legal-domain performance under realistic constraints [19].

Overall, the literature offers excellent building blocks in the form of legal domain PLMs [3, 4], similarity encoders [2, 6], multi-lingual embeddings [7-9], translation baselines [11, 22], retrieval baselines [5, 10, 21], and robustness studies [15, 16]-but still leaves an important gap for deployable bilingual legal similarity models that are accurate on English data, stable under EN↔MR shift, and stable under mixed languages.

The work fills this gap directly by using cross-lingual drop, mixed language stress tests, and ablations as key evidence. There are still four limitations when it comes to

matching legal questions between English and Marathi: legal-domain models do not ensure cross-script robustness; multilingual encoders do not explicitly maintain the identity of statutes and sections; translation might change legally important forms; and cross-encoders or LLMs could involve a higher inference cost.

The framework suggested addresses this combined issue by means of multilingual semantics, the use of explicit legal anchors, compact distillation, citation invariance, and the application of a fixed threshold for cross-lingual evaluation.

3. Proposed Model Design Analysis

Figure 1 shows the general architecture of the proposed framework. Specifically, when given a specific legal question pair (Q1, Q2), the framework produces a similarity probability based on (a) the semantic information derived from the Transformer, (b) the representation of legal anchors, and (c) a compact student embedding through teacher-student distillation. Each of these components and the training objective is further elaborated below.

3.1. Input Normalization and Extraction of Legal-Anchors

Every input question T is subjected to minor normalization, which involves trimming and merging consecutive white spaces. Extraction of legal anchors is done using Unicode-aware tokenization, preserving numbers, Latin tokens, and Devanagari tokens. Legal anchors contain references such as statute family (IPC/CrPC/CPC), sections/article numbers, and recurring legal markers.

To the extent possible, a conservative mapping is made from Marathi legal terms to their English equivalents. Such anchors are then used to create the concept sequence {c1,…, cm} for use by the concept encoder (Section 3.3). Citation normalization NormCit(·) is used only to define the optional invariance regularizer in Equation (21).

3.2. LTEAM: Contextual Encoder and Attention Pooling

Given a question T, the contextualized representations of the tokens are calculated using a multilingual encoder (Equation (1)), then projecting through a linear transformation (Equation (2)) and refining through multi-head self-attention (Equation (3)). Additive attention pooling is used for weighting the importance of each token and also for aggregating the states of tokens into an attention-pooled vector, according to Equations (4) to (6).

In order to improve robustness in spelling variation and cross-linguistic morphologies, a deterministic hashed character n-gram vector is added to the process before layer normalization takes place, resulting into the final LTEAM representation (Equation (7)).

$$H = \text{Encoder}(T; \theta_e) \tag{1}$$

$$X = HW_p + b_p \tag{2}$$

$$X' = \text{MHA}(X, X, X) \tag{3}$$

$$U = \tanh(X'W_a + b_a), \quad s_j = U_j \cdot v \tag{4}$$

$$\alpha_j = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \tag{5}$$

$$r = \sum_{j=1}^{n} \alpha_j X'_j \tag{6}$$

$$u = W_u g(T) + b_u, \quad \tilde{r} = \text{LayerNorm}(r + u) \tag{7}$$

3.3. GRLC-Light: Legal-Anchor Concept Encoder

Let {c1,…, cm} be the extracted concept tokens for a question. Each concept is mapped to an embedding (Equation (8)). A masked mean summary of concepts is computed (Equation (9))), and update concept states using a lightweight self + mean aggregation layer (Equation (10)).

After L layers, the final legal-anchor vector g is obtained by masked mean pooling and normalization (Equation (11)). The statute/section/article anchors are explicitly captured by this representation in a compact form suitable for deployment.

$$h_j^0 = \text{Emb}(c_j) \tag{8}$$

$$\mu^t = \frac{\sum_{j=1}^{m} m_j h_j^t}{\sum_{j=1}^{m} m_j} \tag{9}$$

$$h_j^{t+1} = \text{ReLU}\left(W_t\left[h_j^t; \mu^t\right] + b_t\right) \tag{10}$$

$$g = \text{LayerNorm}\left(\frac{\sum_{j=1}^{m} m_j h_j^L}{\sum_{j=1}^{m} m_j}\right) \tag{11}$$

3.4. Compact Branch and Distillation

In parallel, a compact sentence encoder produces an embedding. e_c (Equation (12)) which is projected into the fusion space as s (Equation (13)).

As part of the training procedure, the teacher embedding model produces a soft similarity target t, calculated by scaling the cosine similarity into [0,1]. The student output is then aligned to this target using the distillation loss function shown in Equation (19).

$$e_c = \text{Compact}(T) \tag{12}$$

$$s = \text{MLP}_s(e_c) \tag{13}$$

$$\cos(a, b) = \frac{a \cdot b}{\|a\|\|b\|}, \quad t = \sum_{m=1}^{M} w_m \frac{\cos(e_1^m, e_2^m) + 1}{2} \tag{14}$$

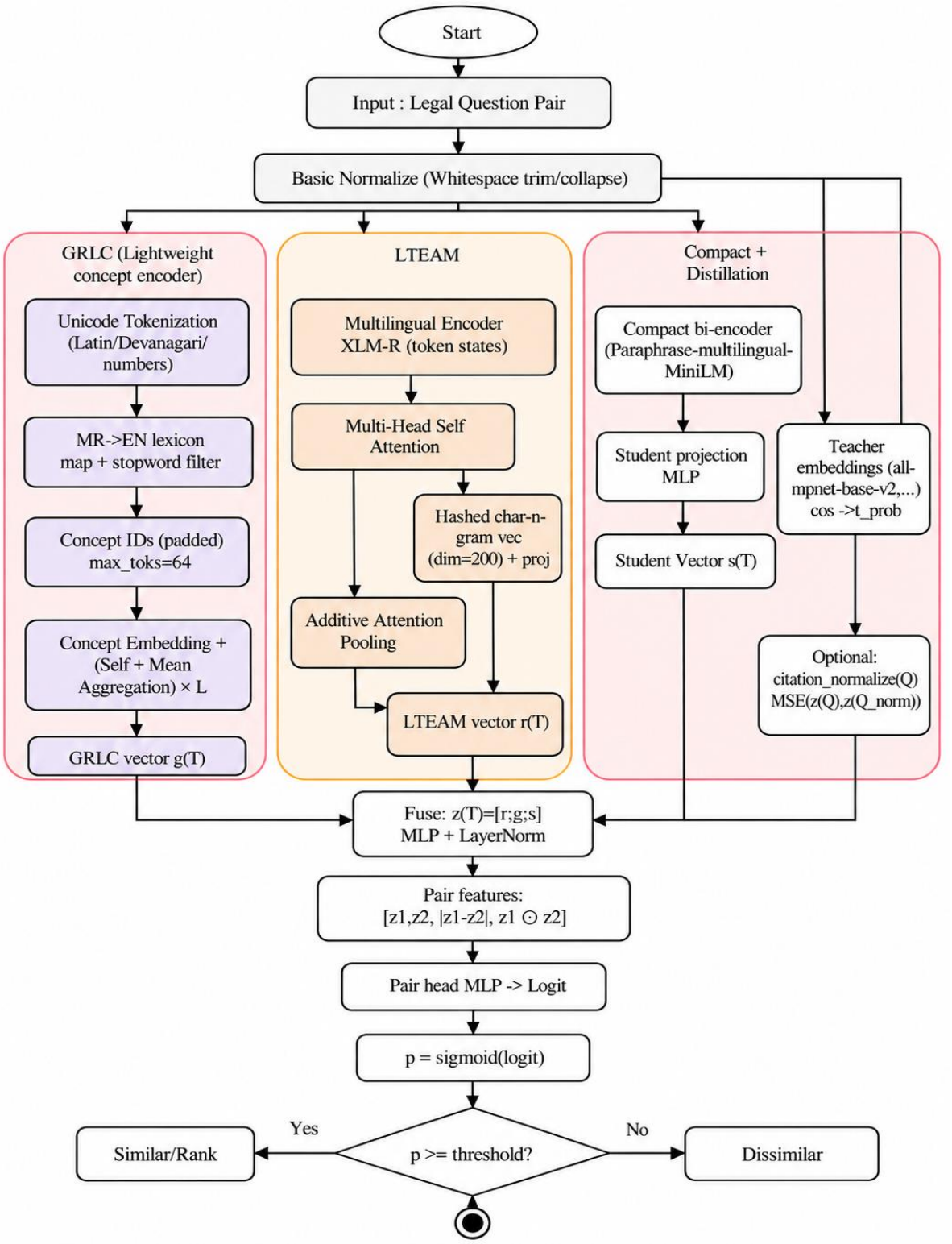

Fig. 1 Architecture of the proposed methodology

3.5. Fusion and Pairwise Decision Head

r̃, g, and s are combined to create the combined vector representation z for each question (Equation (15)). For a tuple (Q1, Q2), The symmetric features of the tuple are calculated using concatenation, absolute differences, and element-wise product (Equation (16)). The light MLP model generates the

pair logit ℓ and similarity probability p (Equation (17)). Following consistent deployment-reliable practices, the threshold is fixed on English validation and used in cross-lingual and mixed language experiments.

$$z = \text{LayerNorm}\left(\text{MLP}_f([\tilde{r}; g; s])\right) \quad (15)$$

$$\phi(Q_1, Q_2) = [z_1; z_2; |z_1 - z_2|; z_1 \odot z_2] \quad (16)$$

$$\ell = \text{MLP}_p\big(\phi(Q_1, Q_2)\big), \quad p = \sigma(\ell) \quad (17)$$

3.6. Training Objective

The training objective includes supervision through classification, teacher-guided distillation, ranking, and optionally citation invariance. The supervision is performed by binary cross entropy (Equation (18)), while the predictions are matched against that of the teacher target using mean squared error (Equation (19)). In order to encourage consistent rankings for retrieving queries, the margin ranking loss approach is used such that positive pairs have higher scores compared to negative pairs for the same query (Equation (20)). To reduce sensitivity to citation surface-form variation, optionally penalize the representation changes under citation normalization (Equation (21)). The total objective is the weighted sum in Equation (22).

$$L_{sup} = -\big(y\log p + (1-y)\log(1-p)\big) \quad (18)$$

$$L_{dist} = \| p - t \|_2^2 \quad (19)$$

$$L_{rank} = \max(0, \gamma - \Delta\ell) \quad (20)$$

$$L_{inv} = \| z(T) - z\big(\text{NormCit}(T)\big) \|_2^2 \quad (21)$$

$$L = \sum_{q \in \mathcal{S}} w_q \, L_q \quad (22)$$

4. Dataset & Preprocessing

A binary legal question-pair matching dataset is compiled, where each data point is a pair of legal questions (Q1, Q2) labelled as 1 if the two questions are semantically equivalent/paraphrastic in a legal context (same legal issue, similar facts, along with nearly identical legal provisions and remedies) and labelled as 0 otherwise. The dataset is derived from two sources of publicly available legal text. The sources are legal discourse extracted from the transcriptions of YouTube videos of the Supreme Court of India [31] and question-like content extracted from the EUR-Lex dataset [32]. The questions are transformed into a consistent question form, paired to generate candidate question-pairs (Q1, Q2), followed by labelling. GPT-5.2 is used to annotate the pairs of questions for each candidate based on a binary schema. During the annotation process, labels are provided with values {0, 1} using a conservative canonicalizer. The canonicalizer treats standard phrases like “similar", “same”, "yes", and "true" as 1, and "dissimilar", "different", "no", and "false" as 0. Unknown labels are discarded. To assess the quality of the labels and the overall dataset, a balanced gold set of 1,200 question-pair instances is created. This set comprises 600 positive instances and 600 negative instances. This set is created to maximize diversity by limiting the number of instances to those drawn from a given cluster-pair. Two legal domain experts annotate the balanced set independently. The experts are provided with a common guideline to annotate the question pairs. Any disagreement is resolved through adjudication to arrive at the final gold set. A fully reproducible pipeline is implemented to pre-process the raw question pairs to remove noise, manage redundancy, and generate leakage-safe splits. The steps are:

1. Standardization & minimal cleaning - Normalize whitespace (trim and collapse repeated spaces), drop pairs with missing/invalid labels, keep only {0,1}, and filter out trivially short questions (length < 5 characters).
2. De-duplication (order-invariant) - Remove (i) exact duplicates of (q1_norm, q2_norm, label) and (ii) swapped duplicates, treating (Q1, Q2) and (Q2, Q1) as the same pair.
3. Question-based near-duplicate clustering: The method groups small-surface-variant instances of the same legal question. This entails embedding all the questions using the sentence-transformers/all-MiniLM-L6-v2 model and creating an HNSW index based on cosine similarity with K=10 neighbors. If two questions have a cosine similarity score of at least 0.92, they are merged using the union-find clustering. Each question receives a cluster_id, and each pair is mapped to a cluster_pair_key = (minCluster, maxCluster, label).
4. Repetition pruning (cap frequently appearing templates). In order to prevent the over-representation of repetition templates in weakly labelled data, two caps are applied via a stratified sampling method: (i) a maximum of three positive pairs where Q1 and Q2 belong to the same cluster (or near-identical paraphrases), and (ii) a maximum of ten pairs per cluster_pair_key.
5. Train/validation/test split with leakage control - Two split protocols are implemented:
   - PAIRKEY split: assigns entire cluster_pair_key groups to a split, thus preventing the same cluster-pair type from appearing across splits.
   - STRICT split: assigns connected question groups to a split, ensuring that related questions do not appear across splits and avoiding post-split removal of pairs. In the experiments, the STRICT partition is used.
   - Due to the strict leakage-control protocol, the achieved split was 77.70%, 10% and 12.30% across train, validation, and test. Because related question groups are kept intact, the validation set is more label-skewed than the train and test sets. Split sizes, along with the question inventory, are presented in Tables 1-3.

Table 1. Dataset split sizes and label distribution

Split	#Pairs	#Label=1	#Label=0	%Label=1
Train	73,137	33,866	39,271	46.30%
Validation	9,413	2,351	7,062	24.98%
Test	11,583	5,144	6,439	44.41%
Total	94,133	41,361	52,772	43.94%

Table 2. Question inventory, clustering, and reuse

Statistic	Value
Unique near-duplicate clusters (cluster_id)	22,576
Avg. pair occurrences per question	4.25
Median pair occurrences per question	4
Max pair occurrences per question	44
Clusters with size > 1 (unique questions)	5,300 (23.48%)
Largest cluster size (unique questions)	410
Pairs with Q1 and Q2 in the same cluster	3.33%
Split leakage check (question-cluster overlap)	0 (train∩val), 0 (train∩test), 0 (val∩test)

Table 3. Text length statistics (Word counts)

Field	Min	P25	Median	P75	P95	Max	Mean
Q1	5	19	27	30	42	154	26.76
Q2	3	19	26	30	42	154	26.66

Table 4. Expert gold set (For label validation)

Item	Value
Gold subset size	1,200
Pos/Neg balance	600 / 600
Sampling constraint	max 1 per cluster-pair key (diversity)
Annotators	2 legal-domain experts (independent)
Agreement metric	Cohen's κ = 0.933, Krippendorff's α = 0.933

4.1. Expert Agreement and Gold-Set Reliability

Prior to analyzing agreement, first, both expert annotation sheets were standardized onto the same labeling system (1 for similar, 0 for dissimilar). In all, two experts independently annotated 1,200 questions using the same criteria.

The agreement was very high with 96.83% exact agreement, Cohen's κ = 0.933 (bootstrap 95% confidence interval = 0.911 - 0.954), and Krippendorff's alpha = 0.933 (nominal). Only 38 out of 1,200 questions (3.17%) were under dispute or needed further analysis.

Table 5. Inter-Annotator Agreement (IAA) summary

Gold subset size	Raw agreement	Disagreements	Cohen's κ	95% CI (κ)	Krippendorff's α
1,200	96.83%	38 (3.17%)	0.933	[0.911, 0.954]	0.933

Many of the disputes involved close paraphrases that were similar on the surface but were different in terms of their legal equivalence because of subtle differences in framing (e.g., interpretation standard vs. application; forum competence vs. merits; remedy/procedural posture).

The remainder dealt with transcript artifacts. The following table shows the breakdown of the most common disagreement topics.

The composition of the gold dataset and annotation quality are presented in Tables 4-6.

Table 6. Disagreement taxonomy (N=38)

Disagreement category	Count	%
Legal standard / interpretive framing (e.g., "scope/test/standard/purposive interpretation")	10	26.32
Other borderline paraphrase (high topical overlap, unclear legal equivalence)	8	21.05
Transcript noise / incomplete phrasing	6	15.79
Forum & jurisdiction framing (competence/maintainability/entertainability)	6	15.79
Outcome wording (validity/arbitrariness/unreasonableness)	4	10.53
Procedural posture/remedy (writ/bail/quash/FIR/proceedings)	4	10.53

5. Results and Discussion

5.1. Experimental Setup

The proposed model is evaluated in the following scenarios: in-domain English (EN-EN), cross-lingual transfer (EN→MR) using translated data on the MR-MR test set, and bilingual and mixed-language settings as stress-testing scenarios that mimic real use cases. Experiments are implemented in Python (PyTorch/HuggingFace) and run on an i9 (14th Gen) + RTX 4090 (24 GB) workstation with AMP (bfloat16) enabled for efficiency, as shown in Table 7.

The experiment employs 3 epochs, a batch size of 32, a learning rate of 1e-5, and a maximum length of 192 when applied to the cross-lingual suite. A grid search is carried out in order to establish the decision threshold on the English validation data to maximize F1, after which the threshold is frozen for use in Marathi and mixed-language testing. This procedure is intended to prevent test leakage and to account for deployment constraints. Classification metrics (F1, AUC), ranking metrics (MRR, Precision@5/10), and robustness measures (ΔF1, ΔMRR) are used to compare the performance of the proposed model. An ablation study is conducted by removing one robustness component at a time by setting its weight to zero. An oracle Marathi threshold is used in order to estimate the calibration headroom. Error analysis is conducted by dividing high-confidence FP/FN into legal-relevant categories, such as section/article mismatch, statute-family mismatch, negation/exception handling, etc.

Table 7. Experimental setup

Item	Setting
Hardware	Intel Core i9 (14th Gen), NVIDIA RTX 4090 (24 GB VRAM), 64 GB RAM, 2 TB SSD
Framework	Python, PyTorch, HuggingFace Transformers
Precision	Automatic Mixed Precision (AMP), bfloat16 when supported
Train/Val/Test protocol	Train on EN-EN, tune threshold on EN validation, evaluate on EN test, MR-MR translated test, and mixed-language tests.
Core training hyperparameters	Epochs 3, batch size 32, learning rate 1e-5, max sequence length 192
Threshold selection	Grid search over ([0,1]), choose threshold maximizing F1 on EN val; freeze threshold for MR/mixed evaluation
Metrics	Classification: Accuracy/Precision/Recall/F1/AUC; Ranking: MRR, Precision@5, Precision@10
Robustness reporting	$\Delta F1 = F1_{(EN)} - F1_{(shift)}$, $\Delta MRR = MRR_{(EN)} - MRR_{(shift)}$
Ablations	Remove one component by setting its loss weight to zero (e.g., w_inv = 0, w_rank = 0, w_distill = 0) while keeping other settings unchanged.
Few-shot adaptation	Add a small labeled Marathi subset during training; evaluated under the same fixed-threshold protocol.
Calibration (upper bound)	"Oracle MR threshold": best threshold on MR scores (analysis only) to estimate headroom
Error analysis	Categorize FP/FN by: section/article mismatch, statute family mismatch, procedural vs substantive, negation/exception handling.

5.2. Results

5.2.1. English Performance

On the large English benchmark, the model achieves an F1 of 0.96 on the test split, with a similar validation score. This proves once again that the architecture is not simply a robustness method that undermines the quality of the English language, but rather a robust similarity-detection model on its own. On the controlled English benchmark (n=1,233), the

model continues to perform strongly (F1 = 0.964, AUC = 0.996) while being efficient. More powerful interaction-heavy baselines (fine-tuned legal cross-encoders) and the LLM baseline can indeed achieve higher F1 scores; however, they do so at the cost of a much higher inference cost.

The LLM classifier, for instance, has F1 ~ 0.986 but requires an order of magnitude more time for each sample compared to the lightweight matcher, as shown in Figure 2. The full results for all baselines are shown in Table 8.

Table 8. Results on EN-EN benchmark (N=1,233)

Model	F1	AUC	MRR	P@5	P@10	Time_ms
LLM baseline (Qwen2.5-7B-Instruct, QLoRA) [12,13]	0.986	-	-	1	1	14.597
LegalBERT-based Cross-Encoder (fine-tuned) [3]	0.978	0.998	0.839	0.635	0.632	0.499
InLegalBERT-based Cross-Encoder (fine-tuned) [4]	0.974	0.997	0.839	0.634	0.632	0.629
SBERT Bi-Encoder [2]	0.972	0.992	0.841	0.634	0.632	0.172
Cross-Encoder (generic) [6]	0.968	0.994	0.836	0.635	0.632	0.406
Proposed Model	0.964	0.996	0.586	1	1	1.238
LegalBERT Bi-Encoder [3]	0.95	0.982	0.834	0.634	0.632	2.179
BM25 [5]	0.946	0.977	0.839	0.634	0.632	2.944
InLegalBERT Bi-Encoder [4]	0.941	0.971	0.839	0.634	0.632	2.644

5.2.2. Cross-Lingual Robustness

A straightforward baseline for the bilingual legal pipeline is translating from Marathi to English and then using the English model. The English model achieves very high in-domain results (test F1 = 0.94), but the translation-based approach fails to compete on Marathi evaluation (F1 = 0.40, AUC = 0.61). This makes sense for legal text, as translation pipelines can change abbreviations, spacing, citation format, and other important "anchor tokens" in legal texts. In general domain, it may be suitable, but in the legal domain, it is often pivotal.

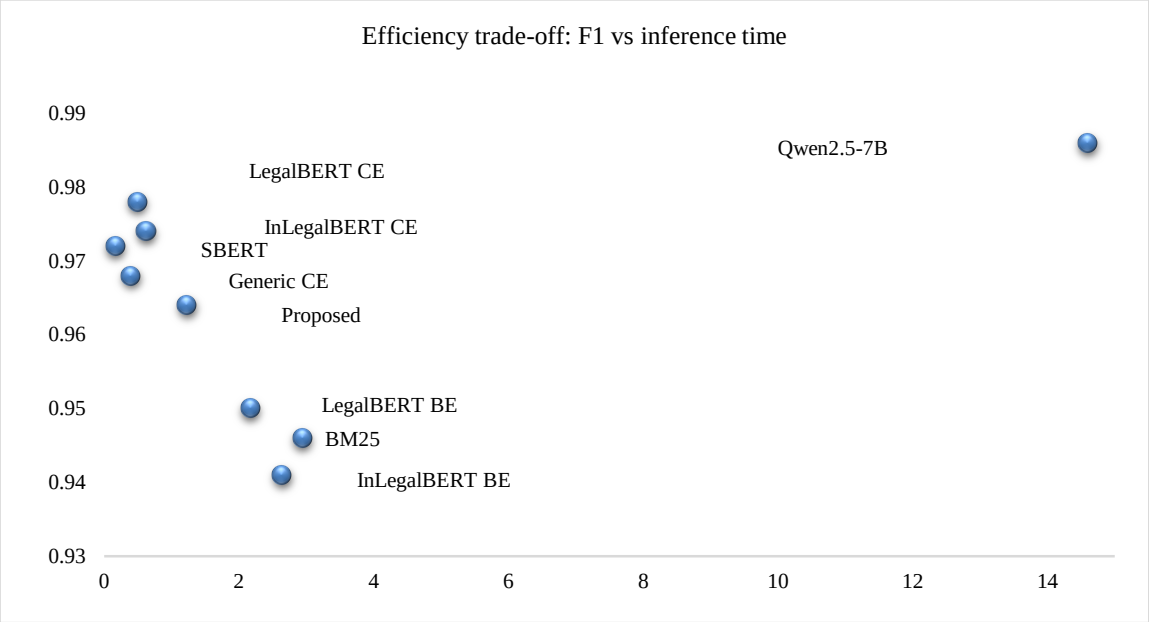

Fig. 2 Efficiency trade-off: F1 vs inference time

5.2.3. Fixed Threshold EN→MR Transfer

Cross-lingual transfer is evaluated using a realistic deployment setup where the threshold is chosen on the English validation set and then utilized for testing on Marathi. For zero-shot cross-lingual transfer, namely EN→MR (seed=42), it is observed that there is a controlled decrease in performance with the model getting an F1 score for English in the lower range of 0.93 and an F1 score for Marathi in the middle range

of 0.85, which is a drop of 0.08. The drop in ranking metrics is smaller than classification; this may indicate that language shift affects score calibration more than semantic ordering. More practical findings, however, is what happens when a small Marathi few-shot component is added to the pipeline. Under the few-shot EN→MR (with seed=42), the Marathi F1 improves to 0.867, with the cross-lingual F1 drop decreasing from 0.083 to 0.055.

Thus, a small Marathi few-shot component reduces the performance drop while retaining the fixed English-selected threshold. The model does not need a fully translated Marathi training set to improve; a modest few-shot MR dataset is enough to yield substantial improvement. The cross-lingual and mixed-language results are summarized in Table 9.

Table 9. Cross-lingual and mixed-language results (Fixed EN-selected threshold)

Scenario	seed	EN-selected threshold	EN F1	MR F1	F1 drop	EN MRR	MR MRR	MRR drop	Oracle threshold	Oracle F1	Oracle F1 gain
Few-shot EN→MR (MR-MR)	42	0.86	0.922	0.867	0.055	0.82	0.809	0.011	0.711	0.897	0.030
Few-shot mixed-language (MR→EN / EN→MR)	42	0.91	0.927	0.877	0.05	0.819	0.82	-0.001	0.684	0.886	0.009
Zero-shot EN→MR (MR-MR)	42	0.84	0.935	0.852	0.083	0.827	0.806	0.021	0.91	0.864	0.012
Zero-shot mixed-language (MR→EN / EN→MR)	42	0.955	0.917	0.851	0.066	0.827	0.824	0.003	0.817	0.856	0.005

5.2.4. Mixed-Language Testing: Robustness under Realistic Bilingual Usage

Mixed language evaluation is a more difficult test case than "clean" MR-MR testing, as it imposes a certain level of asymmetry on the two sides of a pair. In the zero-shot mixed language case, where seed=42, the model performs well on English and worse on Marathi, with a difference in F1 of about 0.06-0.07.

In the few-shot mixed language case, with seed=42, F1 of Marathi improves, and the difference with English decreases further, with a difference of about 0.05. It is interesting to note that MRR remains very consistent, even increasing in one direction in the few-shot mixed case.

This suggests a desirable real-world practicality: even if a fixed threshold does not perfectly capture a mixed-language case, it still maintains the correct ordering among the candidates. The ΔF1 results for each case are shown in Figure 3.

5.2.5. Ablation Study

An ablation study was done to assess the contributions of the different components to cross-lingual robustness. The effect of removing citation invariance is limited in the current single-seed setting: the displayed EN/MR F1 values remain 0.935/0.852 with ΔF1 = 0.083, while the MRR drop decreases from 0.021 for the full model to 0.005. The effect of the removal of the other robustness components is not as clear for a single seed, with seed=42.

The effect of removing the lexicon results in EN/MR F1=0.932/0.857 with ΔF1=0.075, the effect of removing the char-ngram branch results in 0.926/0.856 with ΔF1=0.070, the effect of setting w_rank=0 results in 0.935/0.857 with ΔF1=0.078, and the effect of setting w_distill=0 results in 0.932/0.867 with ΔF1=0.065. These results show non-monotonic effects across the individual components.

The smaller ΔF1 values observed in some ablations result from changes in both English and Marathi performance and therefore shoulsd not be interpreted as uniform improvements in robustness.

Since the analysis uses a single seed, the component-level results should be viewed with caution. Table 11 summarizes the ablation test results for the zero-shot EN→MR base configuration.

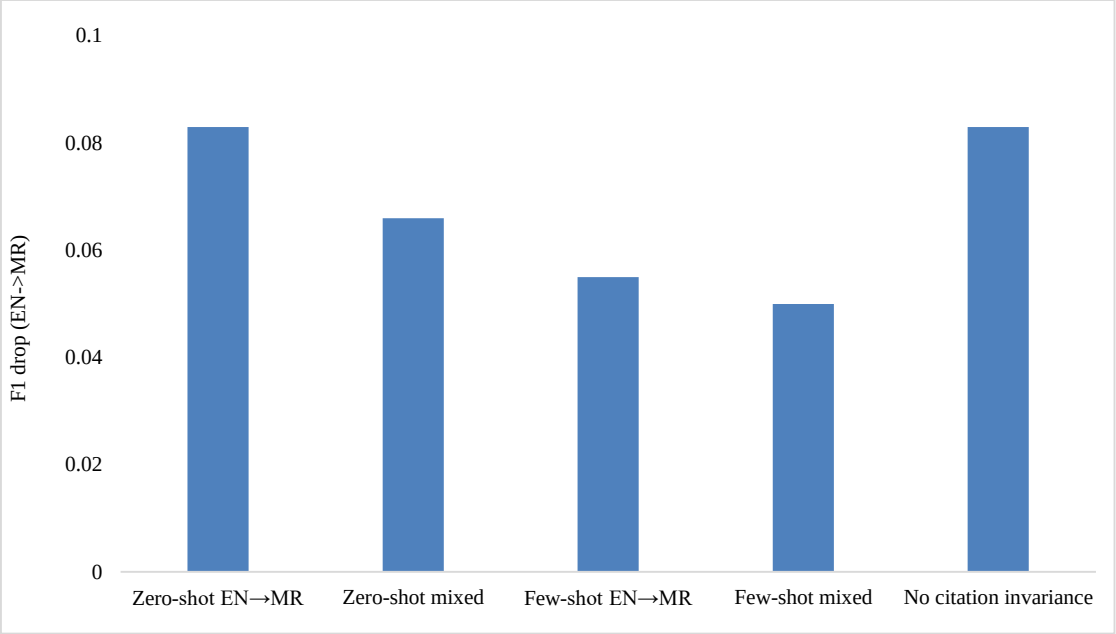

Fig. 3 ΔF1 across scenarios (zero-shot vs few-shot vs mixed vs ablation)

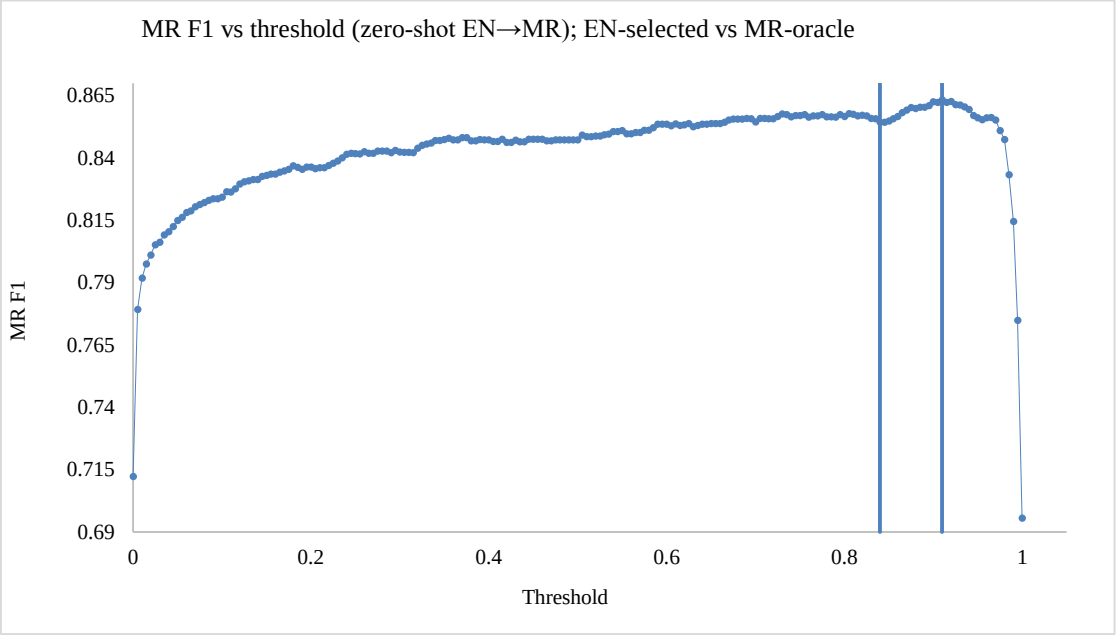

Fig. 4 MR F1 vs threshold (EN-selected threshold vs MR oracle threshold)

5.2.6. Calibration: Oracle Threshold Headroom

Since the threshold is fixed (the one chosen on English validation), an upper bound is also computed for the Marathi "oracle threshold. The oracle improvement is generally low in most cases (single-digit absolute improvement in F1, typically around +0.01), which is useful: it suggests that the remainder is not simply a matter of choosing a different threshold. The difficulty of cross-lingual legal semantics and representational mismatch is reflected by it. In adversarial cases, calibration may be more important, but in any case, a pattern consistent with the main claim is observed: robustness requires modeling decisions, not just subsequent threshold tuning. The MR calibration curve and thresholds are shown in Figure 4.

5.2.7. Error Analysis

The error tagging also reveals that many of the leftover errors are those for which explicit legal anchors are absent-i.e., cases where the model had to rely on inference for equivalence based on procedural or doctrinal paraphrasing. The second common error bucket is partial overlap, where both questions contain legal words but are about different legal topics, i.e. false semantic friends. This is another well-

known error mode for legal similarity. The obvious way forward is to improve legal entity/citation detection, better modelling of exceptions/negation, and procedural/substantive distinctions. The category-wise error frequencies are also summarised in Table 10 and Figure 5. The representative errors for the EN test are provided in Table 12.

Table 10. Error analysis summary on MR test (zero-shot EN→MR) using fixed EN-selected threshold

Category	Count	FP	FN	Percent
Statute mismatch	0	0	0	0.00
Section/Article confusion	4	4	0	1.61
Procedural vs substantive	33	26	7	13.31
Negation/exception	46	34	12	18.55
Topical similarity/entailment	87	87	0	35.08
Other/unclear	78	13	65	31.45

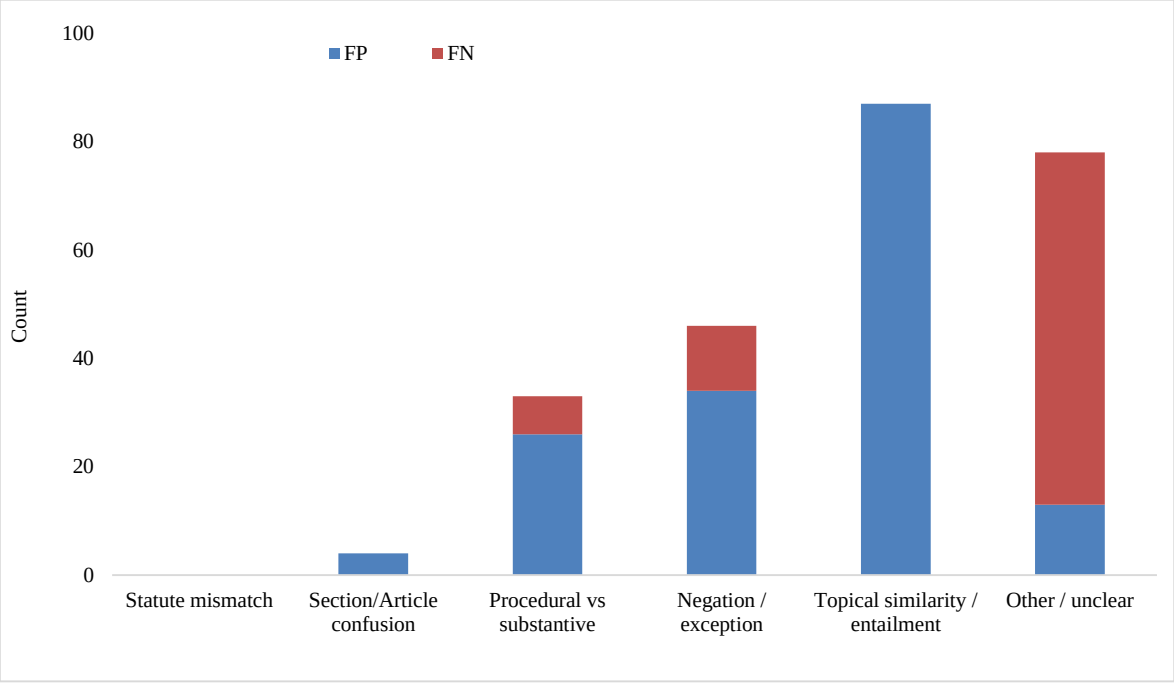

Fig. 5 Error category distribution over high-confidence FP/FN

Table 11. Ablation study results (Δ relative to base)

Base scenario	Ablation	EN F1	MR F1	F1 drop	EN MRR	MR MRR	MRR drop	Oracle F1 gain	ΔF1 drop	ΔMRR drop
Zero-shot EN→MR (MR-MR)	Full Model	0.935	0.852	0.083	0.827	0.806	0.021	0.012	0.000	0.000
Zero-shot EN→MR (MR-MR)	No citation invariance (w_inv=0)	0.935	0.852	0.083	0.824	0.820	0.005	0.016	0.000	-0.016
Zero-shot EN→MR (MR-MR)	Lexicon-off	0.932	0.857	0.075	0.823	0.809	0.014	-	-0.008	-0.007
Zero-shot EN→MR (MR-MR)	Char-ngram-off	0.926	0.856	0.070	0.826	0.820	0.007	-	-0.013	-0.014
Zero-shot EN→MR (MR-MR)	w_rank = 0	0.935	0.857	0.078	0.825	0.817	0.008	-	-0.005	-0.013
Zero-shot EN→MR (MR-MR)	w_distill = 0	0.932	0.867	0.065	0.825	0.817	0.008	-	-0.018	-0.013

Table 12. Qualitative error examples

Category	Subtype	Gold	Pred	Score	Q1	Q2
Section/Article confusion	number mismatch	0	1	0.988	Does Indian law allow the maintainability of a writ petition under Article 226?	Does Article 21 protect recognition of marriage choices as part of personal liberty under Article 21?
Negation/exception	negation asymmetry	0	1	1.0	What is the scope and legal effect of COMMISSION REGULATION (EEC) No 21/81 of 1 January 1981 fixing the minimum buying-in price applicable in Greece for lemons delivered to industry, THE COMMISSION OF THE EUROP…	What is the scope and legal effect of COMMISSION DECISION of 12 June 2007 on a harmonised monitoring of antimicrobial resistance in Salmonella in poultry and pigs (notified under document number C(2007) 2421) …
Negation/exception	negation asymmetry	0	1	0.996	What is the scope and legal effect of COUNCIL DECISION of 16 September 1996 on the provisional application of certain agreements between the European Community and certain third countries on trade in textile p…	What is the scope and legal effect of COMMISSION REGULATION (EC) No 1646/98 of 27 July 1998 fixing the quantities of banana imports for supply to the Community for the fourth quarter of 1998 (Text with EEA rel…
Paraphrase vs entailment vs topical similarity	topical/entailment FP	0	1	1.0	What is the scope and legal effect of COMMISSION REGULATION (EC) No 226/94 of 1 February 1994 laying down certain additional detailed rules for the application of the Supplementary Trade Mechanism (STM) to tra…	What is the scope and legal effect of COUNCIL REGULATION (EEC) No 3357/91 of 7 November 1991 amending Regulation (EEC) No 918/83 setting up a Community system of reliefs from customs duty THE COUNCIL OF THE EU…
Paraphrase vs entailment vs topical similarity	topical/entailment FP	0	1	1.0	Under the statute, is it legally correct that in cases where the seller of the specified things retains them at the place where he/she lives since he/she entered into the contract…	Under the statute, is it legally correct that (A) and (B) agreed that (A) sell the used bicycle, which (A) owned (hereinafter referred to as P), to (B) on April 1, with the provision for…
Other	unclear	1	0	0.0	Is it legally correct that if a person under curatorship performs, without getting the consent, an act that requires getting consent from his/her curator, the curator may ratify that act…	Is it legally correct that even if claims on negotiable instruments or checks are specified as claims that are to be secured by Revolving Mortgage, any claim on negotigable.
Other	unclear	1	0	0.0	Under the statute, is it legally correct that the period of prescription of a right to demand compensation for damages in tort established in an unappealable judgment shall be ten years?	Under the statute, is it legally correct that Regarding delay in performance of a claim that occurred due to a commercial transaction, when a civil conciliation…

Paraphrase vs entailment vs topical similarity	topical/entailment FP	0	1	0.996	Is it legally correct that If A tenders the performance of the obligation of delivery of X and B refuses to accept without a legitimate reason, and X is lost due to gross negligence of A while in custody of X b	Is it legally correct that, unless a particular intention is manifested with respect to the place where the performance should take place, even if X was located at the domicile of D at the time of the conclusion
Negation/exception	negation asymmetry	0	1	0.996	What is the scope and legal effect of COMMISSION REGULATION (EEC) No 1316/93 of 28 May 1993 laying down detailed rules of application for the management of an annual quota of 1,000 tonnes of cheese and curds o…	What is the scope and legal effect of COMMISSION DECISION of 4 August 2006 fixing an indicative allocation by Member State of the commitment appropriations for the European territorial cooperation objective for…

5.3. Discussion

The experimental results reveal three real-world insights. The English results reflect an accuracy-efficiency trade-off rather than overall superiority. The proposed model achieves F1 = 0.964 and AUC = 0.996 with 1.238 ms inference per pair, compared with 14.597 ms for Qwen2.5-7B, which attains a higher English F1. Its main advantage is therefore the combination of strong discrimination, lower inference cost, and cross-lingual applicability.

The biggest gain is over the translate and match baseline, where Marathi F1 improves from 0.40 to 0.852 in the zero-shot fixed-threshold setup. This shows that the use of multilingual representations with explicit legal anchors minimizes the loss of legally relevant information during translation. The MRR stays fairly stable when switching from English to Marathi. Adding the Marathi oracle threshold gives a small improvement, which shows that ranking itself is not the main issue. The challenge lies in how the model assigns scores in cases that involve topical relevance, entailment, negation and the difference between procedural and substantive legal issues. These errors show that further improvement will need an understanding of the underlying semantics.

6. Conclusion

The work introduced a composite legal similarity framework and showed that the model performs well for English and is reasonably robust for the English-Marathi language shift. While translation-based baselines show poor performance in this setting, the cross-lingual approach shows controlled degradation and a strong recovery for the few-shot approach. Testing done on mixed-language input confirms that the approach is suitable for real-world bilingual usage. The overall results exhibit an efficient and robust solution for the multilingual semantic matching in the legal domain. Even though the proposed model doesn't have the highest English-only F1 score, it does offer a good balance in terms of classification performance, cross-lingual robustness, ranking stability, and inference efficiency. The fact that its Marathi performance is much better than that of the translation-based baseline shows that direct multilingual modelling with explicit preservation of legal anchors is feasible. The limitations are the use of single-seed ablations, the present design of the English-Marathi dataset, and the partial reliance on translated Marathi data. Future work should involve evaluating multiple seeds, using larger native Marathi datasets, and including other Indian languages.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

The authors received no financial support for the research.

Acknowledgments

The authors acknowledge the guidance provided by Dr. Amol sBhopale, Assistant Professor, IIIT Nagpur, in the design and implementation of the system.

References

[1] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, Minnesota, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nils Reimers, and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, Hong Kong, China, pp. 3982-3992, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ilias Chalkidis et al., "LEGAL-BERT: The Muppets Straight Out of Law School," *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, pp. 2898-2904, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shounak Paul et al., "Pre-Trained Language Models for the Legal Domain: A Case Study on Indian Law," *arXiv preprint arXiv*, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] M. Gatford et al., "Okapi at TREC-3," *Proceedings of the Third Text REtrieval Conference*, Published in NIST Special Publication 500-225, 1994. [Publisher Link]

[6] Rodrigo Nogueira, and Kyunghyun Cho, "Passage Re-ranking with BERT," *arXiv preprint arXiv*, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 8440-8451, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Fangxiaoyu Feng et al., "Language-Agnostic BERT Sentence Embedding," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Dublin, Ireland, pp. 878-891, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Liang Wang et al., "Multilingual E5 Text Embeddings: A Technical Report," *arXiv preprint arXiv*, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant, "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, United States, pp. 2288-2292, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jay Gala et al., "Indic Trans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages," *arXiv preprint arXiv*, pp. 1-95, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Binyuan Hui et al., "Qwen2.5 Technical Report," *arXiv preprint arXiv*, pp. 32, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Tim Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10088-10115, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Hugo Mentzingen, Nuno António, and Fernando Bacao, "Effectiveness in Retrieving Legal Precedents: Exploring Text Summarization and Cutting-Edge Language Models Toward a Cost-Efficient Approach," *Artificial Intelligence and Law*, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yunting Zhang et al., "Kalt: Generating Adversarial Explainable Chinese Legal Texts," *Machine Learning*, vol. 113, no. 9, pp. 6341-6371, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yunting Zhang et al., "Chinese Legal Adversarial Text Generation based on Interpretable Perturbation Strategies," *World Wide Web*, vol. 28, no. 2, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh, "Investigating Legal Question Generation using Large Language Models," *Artificial Intelligence and Law*, vol. 34, no. 3, pp. 775-813, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Gaspar Dugac, and Tilmann Altwicker, "Classifying Legal Interpretations using Large Language Models," *Artificial Intelligence and Law*, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Priyanka Prabhakar, and Peeta Basa Pati, "Enhancing Indian Legal Judgment Classification with Embeddings, Feature Selection, and Ensemble Strategies," *Artificial Intelligence and Law*, vol. 34, no. 2, pp. 523-564, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ilias Chalkidis et al., "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Dublin, Ireland, pp. 4310-4330. 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yoshinobu Kano et al., "COLIEE-2018: Evaluation of the Competition on Legal Information Extraction and Entailment," *JSAI International Symposium on Artificial Intelligence*, Yokohama, Japan, vol. 1, pp. 177-192, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Abbas Brashi, and Muhammad M.M. Abdel Latif, "Teaching Arabic-English Legal Translation using Model Texts: A Mixed-Method Study," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Omar Khattab, and Matei Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction Over BERT," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39-48, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Vladimir Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 6769-6781, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Kaitao Song et al., "MPNet: Masked and Permuted Pre-Training for Language Understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857-16867, 2020. [Google Scholar] [Publisher Link]

[26] Wenhui Wang et al., "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776-5788, 2020. [Google Scholar] [Publisher Link]

[27] Kevin Clark et al., "ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators," *arXiv preprint arXiv*, pp. 1-18, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Iz Beltagy, Matthew E. Peters, and Arman Cohan, "Longformer: The Long-Document Transformer," *arXiv preprint arXiv*, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] Qihui Zhao, "Legal Judgment Prediction Via Legal Knowledge Extraction and Fusion," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 3, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] Alan Khoja et al., "Automated Consistency Analysis for Legal Contracts," *Artificial Intelligence and Law*, vol. 34, no. 3, pp. 881-922, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[31] Supreme Court of India, YouTube Channel, 1950. [Online]. Available: https://www.youtube.com/@SupremeCourtofIndia-1950

[32] Ilias Chalkidis et al., "Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation," *Proceedings of the Natural Legal Language Processing Workshop 2019*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 78-87, 2019. [CrossRef] [Google Scholar] [Publisher Link]