LEGAL-UQA: Testing Constitutional Question-Answering Systems for Urdu Speakers with a Parallel English–Urdu Benchmark

Faizan Faisal[1,*], Umair Yousaf[2,*], Muhammad Hassnain[1], Zeerak Babar[1]
[1]University of California, Davis, Davis, CA, USA
fznfaisal@ucdavis.edu, mhassnain@ucdavis.edu, zebabar@ucdavis.edu
[2]University of Colorado Boulder, Boulder, CO, USA
umair.yousaf@colorado.edu

***Abstract*—Legal-information assistants could help Urdu speakers understand their constitutional rights, but building reliable systems requires test data and evaluation methods suited to the language. We present LEGAL-UQA, a parallel English–Urdu benchmark of 619 question–answer pairs linked to passages from Pakistan's Constitution. Each example includes the question, reference answer, and source passage in both languages. We evaluate two components separately: answering Urdu questions when the source passage is supplied, and retrieving the passage from which each question was generated. We also compare JEV, a decision model used as an automated judge, with human grades. The findings show that model rankings depend on how answer similarity is measured, while a simple word-matching method retrieves source passages more effectively than the embedding methods tested. At the evaluated cutoff, JEV agrees with human grades less often than a baseline that gives every answer full credit. An audit also identifies cases where generated references omit information supported by the source. Together, these findings highlight the need to test retrieval and answering separately, include simple retrieval baselines, and validate automated judges and reference answers with human review when engineering legal QA software for Urdu speakers.**

***Index Terms*—inclusive software engineering, language accessibility, legal question answering, Urdu, benchmark, test oracle, LLM-as-a-judge**

I. Introduction

Software that helps people understand their legal rights must work in the languages they use. Language barriers in Pakistani legal services motivate accessible Urdu resources [1]. Article 251 of the Constitution identifies Urdu as the national language and provides for continued official use of English pending arrangements for its replacement [2]. AI-enabled assistants that answer constitutional questions in Urdu could widen access to legal information, but only if their behavior can be tested in Urdu. Urdu legal QA combines limited task-specific data with the need to preserve institutions, conditions, exceptions, and procedural details in answers.

From a software-engineering perspective, a legal QA assistant combines at least two components: one finds a relevant source passage, and one writes an answer from it. Testing such a system needs inputs with known sources, expected outputs, and an *oracle* that decides whether an output is acceptable. For free-form Urdu answers, the oracle is the difficult part. Word-overlap metrics measure similarity to a reference rather than correctness, and model-based judges must be checked for the language and task at hand [3].

LEGAL-UQA provides such test data for questions about Pakistan's Constitution. Each example contains a question, an answer, and the source passage in both English and Urdu. We call each manually delimited constitutional article entry an *article chunk*. The questions imitate layperson wording but do not come from real consultations. The reference answers were generated and translated, not written as legal opinions.

We test the two components separately. In *supplied-article QA*, a model reads the Urdu question and its source passage, then writes an Urdu answer. In *article retrieval*, a method searches the stored passages for the one from which a question was generated. A correct retrieval result is not itself a correct answer; the separation locates whether a failure concerns finding the source or answering from it. We ask three research questions:

- **RQ1 (answering):** How do Urdu answer sets compare when the source passage is supplied, under reference-similarity metrics and final human grades?
- **RQ2 (retrieval):** How often do word-matching and embedding methods recover the originating passage?
- **RQ3 (oracle):** Can a decision model stand in for human grading as an automated test oracle for Urdu answers?

The contributions are:

1) A parallel dataset of 619 English–Urdu QA pairs linked to 305 article chunks, with both language versions of each source passage.
2) A component-level comparison of nine Urdu answer sets on 124 questions, alongside article retrieval results from searches based on numerical text representations and on word matching.
3) A test of a decision model's agreement with final human grades on 1,116 Urdu answers, together with an audit

[*]Equal contribution.

of dataset coverage, article overlap between data splits, answer references, and uncertainty, and the implications for engineering inclusive legal QA software.

II. Related Work

A. Urdu QA and Legal QA Resources

The UQA dataset translates SQuAD2.0, a reading-comprehension dataset, into Urdu and preserves answer spans [4]. UrFQuAD-v1 asks questions for which a system must find a relevant Wikipedia passage, including questions with no answer [5]. Kazi et al. [6] compare reading comprehension across Urdu and English questions and contexts. These datasets establish Urdu QA methods, but their passages do not test whether a model preserves the offices, conditions, and exceptions in Pakistan's Constitution.

Legal QA work shows the value of linking an answer to its source. BSARD links Belgian citizens' questions to statutory articles [7]; LLeQA adds detailed answers and paragraph-level evidence [8]. GerLayQA pairs German lay questions with lawyer answers and cited paragraphs [9], while EQUALS links Chinese consultation questions to answer spans in law [10]. KoBLEX asks legal scenario questions in Korean and English that can require several provisions [11], and LexCLiPR tests multilingual queries against English European Court of Human Rights paragraphs [12]. These resources motivate our question–answer–passage structure, but their questions and laws address other jurisdictions.

For Pakistani law, Zafar et al. [13] provide data for finding names of people, organizations, and other entities in legal text. Urdu Legal Instruct describes 1,067 QA pairs across family, criminal, and religious law, but its public description does not specify paired English records or a source passage for each question [14]. PakLegalQA covers 300 questions about federal statutes and Lahore High Court judgments, with evidence citations, unanswerable questions, and a 60-question Roman Urdu version [15]. It studies citation support and when a system should decline to answer across a wider range of law. LEGAL-UQA instead focuses on Urdu-script constitutional passages and provides paired English–Urdu questions, answers, and source passages.

B. Evaluating AI-Generated Answers

Software-engineering research has evaluated language-model answers in other domain-constrained QA settings: Babar et al. [16] evaluate open-source models on technical StackExchange questions, using retrieval-augmented generation and fine-tuning. Such evaluations depend on a way to decide whether an answer is acceptable. Word-overlap scores cannot decide whether a free-form answer conveys the same meaning as a reference. A model-based judge can make that decision, but its agreement with people needs to be checked for the language and task at hand. Fu and Liu [3] find that model-based judgments vary across languages, with weaker consistency in lower-resource settings. In this lower-resource Urdu legal setting, we compare JEV, a decision model configured to judge reference equivalence, with final human grades. Here, *decision model* denotes a model that returns structured judgments and probabilities for predefined questions, enabling software to apply explicit decision rules. TypeSafe calls JEV a *System One model*; it returns typed outputs directly without generating free-form text [17].

III. Benchmark Construction

A. Source Documents and Text Extraction

The source documents are National Assembly English and Urdu versions of the Constitution [2, 18]. Both source editions state an amendment cutoff of 31 May 2018; their prefaces are dated 31 May 2018 (English) and 12 July 2018 (Urdu). Our evaluations concern the stored texts and make no claim about current law.

The English PDF has extractable text, while the Urdu PDF is image-based and requires optical character recognition (OCR). YOLOv8 detects Urdu text-line regions using the UrduDoc resources, and UTRNet recognizes text from the cropped regions [19]. Model-assisted cleanup with GPT-4o uses the English article and OCR-derived Urdu text to correct formatting, spelling, and sentence order. No independent transcription was available to measure OCR error, so the cleaned Urdu text is not a verified transcription.

After conversion to editable text, we manually segmented the English document into 312 article-level passages, usually containing an article heading and its clauses. For example, Article 1's heading and three clauses form one chunk; Article 2's single sentence forms another. Some chunks contain only an omitted-provision notice. Each passage received a unique identifier, distinct from the Constitution's article number. The paired Urdu passages were OCR-derived and cleaned, without an independent check of exact article boundaries.

The 25th and 75th percentiles of English chunk lengths were used to separate roughly the shortest quarter of chunks, the middle half, and the longest quarter. In the saved categories, fewer than 45 whitespace-separated words was small, 45–238 was medium, and 239 or more was large; the percentile calculation and category assignment used different ways of counting words. The prompt requested up to one, two, or three English QA pairs for small, medium, and large chunks, respectively; these counts were not enforced limits. The three categories contain 79, 157, and 76 chunks, average 31.05, 114.76, and 446.68 words, and contribute 79, 312, and 228 of the final QA pairs.

B. Question Generation, Translation, and Filtering

An OpenAI GPT-4 model generated English questions and answers from each article chunk, using a prompt that asks for layperson wording and forbids article numbers in questions. Parsing the saved output yielded 621 English QA pairs. Seven small chunks yielded no parsed pair, while seven other small chunks yielded two despite the one-pair request. Two rows repeated an English question already present, so removing them left 619 pairs with distinct English questions.

Using GPT-4, we translated the English QA pairs with both language versions of the article as context; the prompt asks for the Urdu constitutional register. We checked the

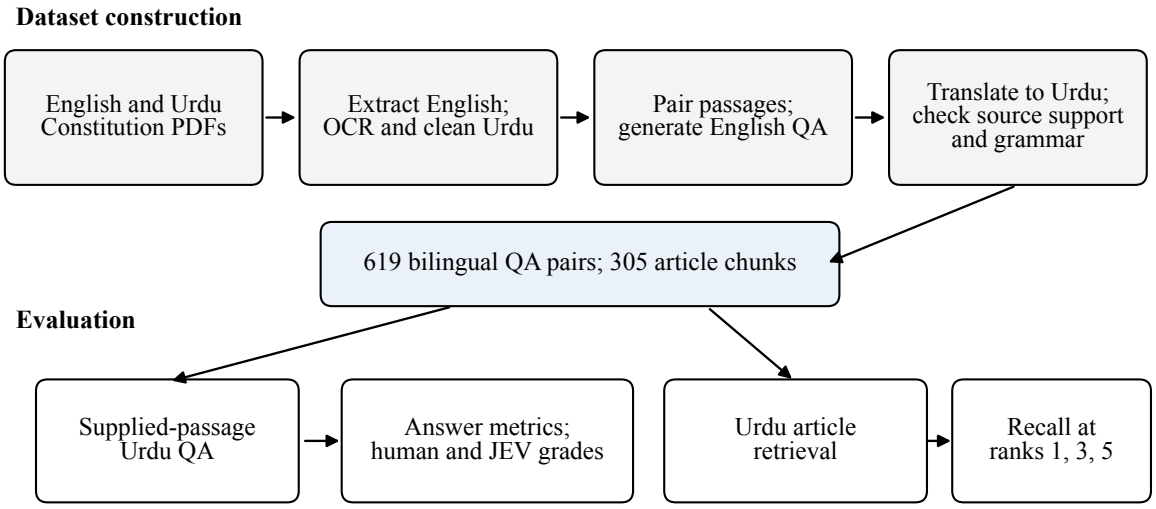


Fig. 1. Dataset construction and supplied-article QA evaluation. The diagram shows the English and Urdu source documents, question generation, translation, review, and answer evaluation. Answering models receive the question and source passage; JEV also receives the reference and model answer.

generated question–answer pairs against their source passages for contextual validity and checked that both the Urdu question and answer in each translated pair were grammatically correct.

We linked each English–Urdu QA pair to its bilingual source passage. The 619 final pairs cover 305 article chunks. The seven chunks without pairs contain omitted-provision notices or placeholder text. Each record contains English and Urdu questions, answers, and passages, plus an internal chunk identifier. The answers are free-form references, with no marked answer spans. Automated checks confirmed complete bilingual records, unique English questions, and consistent links between questions, passages, and model responses, as well as coverage of all records by the data split. These checks do not certify legal meaning. Figure 1 summarizes the construction and the supplied-article QA evaluation.

C. Coverage and Split

Of the 305 chunks with questions, 67 have one QA pair, 162 have two, and 76 have three. Simple English-prefix counts find 195 yes/no forms, 175 "what" questions, and 140 "who" questions; the remaining 109 use other openings. These counts describe wording, not legal topics or reasoning demands. English questions average 13.97 whitespace-separated words and Urdu questions 15.64; English and Urdu answers average 25.93 and 28.71, respectively.

The dataset was split at the *question* level (80% training, 20% validation; seed 25), giving 495 questions for model adaptation and 124 held aside for evaluation. The split was not made by article: 95 chunks appear in both sets. Thus 101 validation questions (81.45%) concern a source passage seen in training, although the English questions themselves do not repeat. The results measure answers to new questions in a largely shared article collection.

IV. Evaluation Design

A. Answer Generators under Test

Although the benchmark is bilingual, we evaluate answering only in Urdu; retrieval is evaluated separately in Urdu and English. Each QA model reads the Urdu question and its source article, then produces an Urdu answer. Only mT5 is adapted on the LEGAL-UQA training split; the other models are evaluated without task-specific fine-tuning on this dataset. We compare nine models with one saved answer per question: Claude-3.5-Sonnet, Gemini-1.5-Flash-002, Gemini-1.5-Pro-002, Mistral-Large, mT5, GPT-5.4, DeepSeek-V4-Pro, Claude-Sonnet-5, and Gemma-4-E4B. The last four runs requested temperature 0 and at most 1,024 output tokens, with reasoning set to `none` (GPT-5.4), thinking disabled (DeepSeek-V4-Pro and Claude-Sonnet-5), or no reasoning setting (Gemma-4-E4B). These four newer models are also the focus of the additional uncertainty and normalization analyses. They share an instruction to answer only in Urdu script from the given context, but prompt roles and wording differ among them.

For mT5, a multilingual text-generation model, we started with a checkpoint trained on Urdu UQA [4, 20] and adapted it for eight passes through the LEGAL-UQA training split (epochs), with inputs limited to 512 tokens and target answers to 128. The same mT5 answers are used for automatic metrics, human grading, and JEV evaluation.

B. Reference-Similarity Metrics

We first compare each answer with its single reference using token-overlap F1 [21]. Tokens are words separated by spaces. Before counting, we lowercase text, remove ASCII punctuation and the words "a", "an", and "the", and normalize spaces. Let P and R be the prediction and reference token lists, and let $c_P(w)$ and $c_R(w)$ count occurrences of word w in each. The per-answer score is $2\sum_w \min(c_P(w), c_R(w))/(|P|+|R|)$; two

empty answers score 1, while only one empty answer scores 0. We report the mean on a 0–100 scale.

We also report METEOR [22] on a 0–1 scale and SacreBLEU [23] on a 0–100 scale. Our METEOR implementation uses the Natural Language Toolkit's English-focused word splitting, stemming, and WordNet synonym data, so it is not an Urdu-specific semantic measure. SacreBLEU uses its `13a` word-splitting scheme, preserves capitalization, and uses one reference with exponential smoothing. As sensitivity checks, we calculate chrF [24] over runs of 1–6 characters with $\beta = 2$, and Urdu-normalized token F1, which applies Unicode NFC, unifies common "kaf" and "yeh" forms and decimal digits, removes combining marks, tatweel, and invisible formatting characters, and replaces punctuation with spaces. All answer sets are scored with the same implementation. These metrics measure reference similarity, not whether an answer is legally correct.

C. Human Grading

Human grading covered all 124 responses from each of nine model answer sets (1,116 responses). Three proficient speakers of Urdu as their first language participated. Two evaluators graded each response, and the third resolved disagreements to determine the final grade; we use these final adjudicated grades throughout. Graders saw the question, source passage, reference answer, and model answer. Under our binary rubric, an answer receives 1 only if it conveys the reference's meaning without omitting details, allowing paraphrases. Correct additional information supported by the source passage is not penalized; incorrect or partial answers receive 0. The evaluators had limited legal expertise.

D. Retrieval Components

We search all 305 distinct source passages from the combined training and validation data. A method succeeds when it retrieves the passage used to generate the question. Recall@k is the percentage of 124 questions for which that passage appears in the first k results. It measures recovery of the recorded source, not whether other passages are legally relevant.

Two embedding models convert questions and passages into numerical vectors and rank passages by cosine similarity; we report their results from saved evaluation summaries. The lexical method BM25 [25] searches by matching words within each language, using the stored Urdu and English questions separately. We apply the Urdu normalization above, then split text at spaces, without stemming or stop-word removal. BM25 uses $k_1 = 1.2$, $b = 0.75$, and the positive-smoothed inverse document frequency $\log(1 + (N - df + 0.5)/(df + 0.5))$. Each distinct query term counts once; ties break by internal chunk identifier.

E. Decision Model as Test Oracle

We used JEV (version 1.13.0) to compare each model answer with its reference. Its input contains the Urdu question, source passage, generated reference, and model answer. It returns a score between 0 and 1 for reference equivalence; we count scores of at least 0.5 as positive. The judge prompt accepts paraphrases and harmless additions but rejects contradictions, missing essential details, and unsupported substantive additions. It calls the reference "written by a human expert", although the references were generated and translated. Although the source passage is supplied, the audit examines cases where JEV's judgments appear to prioritize the reference over source-supported information.

V. Results

A. RQ1: Answering with the Source Supplied

a) Scores depend on the metric: Table I shows that GPT-5.4 leads token F1 (71.45), Claude-Sonnet-5 leads METEOR (0.703), and DeepSeek-V4-Pro leads SacreBLEU (53.98). We estimate uncertainty for GPT-5.4, DeepSeek-V4-Pro, Claude-Sonnet-5, and Gemma-4-E4B with 10,000 paired bootstrap resamples (seed 25) of the 113 validation article chunks, where each sampled chunk contributes all its questions. GPT-5.4's F1 is 2.75 points above Claude-Sonnet-5, but the paired 95% interval for this difference runs from -1.74 to 7.38 points, so this sample does not establish a clear F1 difference between them. Its F1 leads over DeepSeek-V4-Pro (5.41, interval 1.18 to 9.90) and Gemma-4-E4B (10.05, interval 5.46 to 14.72) exclude zero. These intervals reflect variation across sampled articles, not annotation errors or repeated model runs, and are exploratory, with no adjustment for testing several pairs.

b) Sensitivity analyses: Urdu spelling normalization raises token F1 for GPT-5.4, DeepSeek-V4-Pro, Claude-Sonnet-5, and Gemma-4-E4B by 1.40–1.91 points. Claude-Sonnet-5 leads chrF (69.86) and exact match (11.29%), the share of answers identical to their references under the token-F1 normalization; GPT-5.4 leads Urdu-normalized F1 (73.36). Gemini Flash and Pro answers average 14.95 and 18.08 words, versus 28.88 for validation references. Their answers are shorter, but these data do not show that length causes their lower overlap scores.

c) Human grades: Claude-3.5-Sonnet, GPT-5.4, DeepSeek-V4-Pro, and Claude-Sonnet-5 each received full credit on 123 of 124 answers (99.19%). The Gemini, Mistral, and Gemma answer sets received full credit on 95.97–98.39% of answers; the mT5 set reached 74.19%. Each of the three leading reference-similarity models has the same human full-credit rate, so the final grades do not separate them. One example makes an answer difference concrete: asked who administers an oath, GPT-5.4 names the Chief Justice, as the reference does, whereas Gemma-4-E4B answers "Before the President," which states before whom the oath is taken rather than who administers it. Their final human grades are 1 and 0; one example does not establish a model-wide difference.

B. RQ2: Retrieving the Source Passage

As Table II shows, Urdu BM25 retrieves the originating passage first for 71.77% of questions and within the first five for 87.90%. The strongest embedding run, text-embedding-3-large, reaches 53.23% and 77.42%, respectively. English BM25 reaches 62.90% at rank 1 and 89.52% at rank 5. The mean reciprocal rank, the average of one divided by the correct

TABLE I

SUPPLIED-ARTICLE URDU ANSWERING ON 124 VALIDATION QUESTIONS PER ANSWER SET. F1, METEOR, AND SACREBLEU MEASURE SIMILARITY TO GENERATED REFERENCES; COLUMN MAXIMA ARE BOLD. "HUMAN FULL" IS THE FULL-CREDIT RATE AFTER THE THIRD REVIEWER RESOLVED GRADING DISAGREEMENTS; "JEV POS." IS THE PERCENTAGE JEV JUDGED EQUIVALENT AT THE 0.5 CUTOFF; AGREEMENT AND COHEN'S κ COMPARE JEV WITH THE HUMAN GRADES.

	Reference similarity			Human grades and JEV judge			
Answer set	F1	METEOR	SacreBLEU	Human full (%)	JEV pos. (%)	Agreement (%)	κ
Claude-3.5-Sonnet	66.57	0.690	44.87	99.19	93.55	94.35	0.211
Gemini-1.5-Flash-002	35.17	0.279	18.00	98.39	83.87	83.87	0.063
Gemini-1.5-Pro-002	41.32	0.342	24.60	98.39	91.13	92.74	0.288
Mistral-Large	65.91	0.615	51.45	96.77	84.68	84.68	0.127
mT5	62.22	0.605	42.81	74.19	65.32	79.84	0.527
GPT-5.4	**71.45**	0.701	46.86	99.19	91.94	92.74	0.170
DeepSeek-V4-Pro	66.05	0.638	**53.98**	99.19	91.94	92.74	0.170
Claude-Sonnet-5	68.71	**0.703**	45.88	99.19	91.94	92.74	0.170
Gemma-4-E4B	61.41	0.592	45.90	95.97	87.90	88.71	0.255
Pooled (nine human-graded sets)				95.61	86.92	89.16	0.336

TABLE II

ORIGINATING-ARTICLE RETRIEVAL FOR 124 VALIDATION QUESTIONS OVER 305 PASSAGES (RECALL@k, %). EMBEDDING SCORES COME FROM SAVED URDU RETRIEVAL RESULTS; BM25 SCORES ARE CALCULATED FROM THE DATASET.

Method	Query language	R@1 (%)	R@3 (%)	R@5 (%)
text-embedding-3-large	Urdu	53.23	72.58	77.42
mistral-embed	Urdu	29.03	45.97	52.42
BM25	English	62.90	83.87	89.52
BM25	Urdu	71.77	84.68	87.90

passage's rank, is 0.793 for Urdu BM25 and 0.751 for English BM25. These are within-language searches over generated questions; the results do not isolate a language effect or measure end-to-end QA.

C. RQ3: The Judge as an Automated Oracle

Across nine answer sets for the same 124 questions, JEV matched 995 of 1,116 final adjudicated human grades (89.16%; Cohen's $\kappa = 0.336$). Cohen's κ measures agreement after adjusting for agreement expected by chance. Because 1,067 human grades were positive, always assigning full credit would agree on 95.61%. This baseline benefits from the predominance of full-credit grades but detects none of the human-negative answers. JEV's ability to identify those answers therefore matters alongside its overall agreement; raw agreement alone does not establish reliable legal correctness. Per answer set, agreement ranges from 79.84% for the mT5 set to 94.35% for Claude-3.5-Sonnet (Table I); the human full-credit rate is also the agreement of an always-positive decision for that set, and it exceeds JEV's agreement for every set except mT5.

Of the judgments, 958 are positive under both JEV and the human grade, and 37 are negative under both. JEV marks 12 human-negative answers positive and 109 human-positive answers negative. It identifies 89.78% of human-positive answers and 75.51% of human-negative answers. An exploratory sweep selects a cutoff of 0.20 by maximizing pooled κ on these same labels (96.15% agreement, $\kappa = 0.578$), but this cutoff has not been tested on held-out judgments. Answers to the same question recur across models, so the 1,116 judgments are not independent observations. We release the 121 JEV/human disagreements as review candidates; they are not an exhaustive error set.

D. Reference and Source Audit

An answer can differ from the generated reference and still be supported by the source. In one appeals question, the article permits an appeal when the High Court takes over a lower-court case, tries it, and imposes a death or life sentence. The reference omits this route. GPT-5.4 and DeepSeek-V4-Pro include it and receive full human credit, but JEV rejects them despite receiving the source passage. These judgments appear to follow the incomplete reference rather than the supporting passage; the disagreement does not show that the answers are wrong.

Translation can also change the reference itself. In a question defining law, the Urdu reference limits an exclusion for fiscal and banking laws to ten years, but the English reference omits the limit. These selected cases call for expert bilingual review; they do not measure error frequency across the dataset.

VI. IMPLICATIONS FOR ENGINEERING INCLUSIVE LEGAL QA SOFTWARE

a) Test components separately: With the source passage supplied, final human grades give full credit to 95.97–99.19% of answers from eight of the nine sets, whereas the best retrieval method ranks the originating passage first for 71.77% of questions. These measures come from different tasks and do not predict end-to-end accuracy, which needs its own test. They do show that a single system-level score would not reveal which component failed, and that component-level test sets with recorded sources make that diagnosis possible.

b) Keep simple lexical baselines: Urdu BM25 outperformed both embedding runs at rank 1 in our data. Teams building Urdu assistants should not assume that embedding retrieval is the stronger default; a word-matching baseline is cheap to include in regression tests. Our questions were

generated from the passages, so independently written questions are needed to confirm this pattern for real users' wording.

c) Validate automated oracles per language: A raw agreement rate of 89.16% looks high but falls below a trivial always-full-credit baseline when most answers are acceptable. Before an automated judge gates releases or ranks models for a new language, its decisions should be compared with human grades and with that baseline, reported with chance-corrected agreement and class balance, and its cutoff selected on held-out questions or articles, keeping all answers to a question in the same group. Supplying the source passage is insufficient unless the judge also uses it to recognize correct information absent from the reference.

d) Keep humans in the loop for references: Generated and translated references can omit routes or add limits that the other language lacks. Parallel English–Urdu records make such discrepancies visible, but finding and correcting them requires bilingual reviewers with legal expertise. Benchmark maintenance for under-served languages should budget for that review rather than relying on generated references alone.

VII. Threats to Validity

a) Construct validity: Reference similarity measures overlap with generated references, and METEOR relies on English-focused resources. Our final human grades assess answer completeness relative to the references and permit correct additions supported by the source passages. The evaluators had limited legal expertise; independent review by legal experts is needed to validate legal correctness. References may omit or mistranslate essential facts; selected examples do not establish how often this occurs. Despite receiving source context, the judge can reject source-supported answers that differ from the reference, and its prompt inherits an erroneous expert-reference description.

b) Internal validity: Prompt roles and reasoning settings differ across GPT-5.4, DeepSeek-V4-Pro, Claude-Sonnet-5, and Gemma-4-E4B. Each evaluated answer set contains one answer per question. Metric intervals capture corpus sampling variability rather than variation across repeated runs or differences in prompting. The 0.20 judge cutoff was selected on the evaluated labels.

c) External validity: The 619 QA pairs arise from one collection of constitutional passages and generated question style. They support a pilot benchmark, not a claim that this size suffices for broad legal training or robust model ranking. Neither question prefixes nor length statistics establish semantic diversity. The row-based split shares most validation articles with training, limiting unseen-article generalization for the adapted model. OCR cleanup and QA translation have not been independently measured for preservation of legal meaning. Although both documents state the same amendment cutoff, their provision-by-provision correspondence has not been independently verified, and the resource does not cover later amendments. Our results do not establish performance when a system must both find a source passage and answer from it, reasoning across several legal provisions, or suitability for legal advice.

VIII. Ethics and Data Availability

The resource is intended for research, including the evaluation of QA software. Generated or translated answers can misstate legal rights and procedures, and should not replace qualified legal counsel. The construction uses public constitutional documents and generated questions, with no collection of private legal consultations. The English and Urdu contexts and references must retain their provenance when redistributed. We use the MIT license for our project code and dataset contributions; third-party constitutional documents and model assets retain their own rights and terms, and this designation does not grant rights the project does not hold. An analysis manifest records dataset and response checksums, scoring library versions, metric settings, and resampling settings. The dataset, code, and analysis artifacts are available here: https://github.com/muhammad-hassnain/LEGAL-UQA.

IX. Conclusion and Future Work

LEGAL-UQA links English and Urdu questions and answers to constitutional source passages, supporting separate tests of answering from a supplied passage and finding that passage. The leading QA model depends on the text similarity metric; Urdu BM25 retrieves the originating passage more often than the reported embedding runs at rank 1. Selected omissions in reference answers show that a source-supported answer can differ from its reference. At the 0.5 cutoff, JEV's agreement with final human grades is below an always-full-credit baseline, so decision models used as judges need per-language validation before they serve as test oracles for Urdu answers.

Future work should independently review the legal content, split training and evaluation by article, and compare answering in English and Urdu with both supplied and retrieved passages. Independently written questions should test conditions, exceptions, insufficient evidence, and when a system should decline to answer. Formal methods have been applied to statutory reasoning [26] and to abstract notions such as software trust [27]; similar encodings could strengthen this pipeline, for example by checking answers against the source provision.

References

[1] A. Ahmad, S. Hussan, and A. Butt, "English in the domain of legal services: Problems of access in pakistan," *Pakistan Journal of Criminology*, vol. 12, no. 1, pp. 109–127, 2020. [Online]. Available: https://nja.pastic.gov.pk/PJCIS/index.php/PJC/article/download/34828/33920/37330

[2] National Assembly of Pakistan, "The constitution of the islamic republic of pakistan," 2018, eighth edition, as modified through 31 May 2018. Archived source PDF.

[3] X. Fu and W. Liu, "How reliable is multilingual LLM-as-a-judge?" in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics,

Nov. 2025, pp. 11 040–11 053. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.587/

[4] S. Arif, S. Farid, A. Athar, and A. A. Raza, "UQA: Corpus for Urdu Question Answering," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, 2024, pp. 17 237–17 244. [Online]. Available: https://aclanthology.org/2024.lrec-main.1497/

[5] M. Shakeel and R. M. A. Nawab, "Open-Domain Factoid Question-Answering in Urdu: Data and Methods," *IEEE Access*, vol. 13, pp. 30 167–30 185, 2025. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3540939

[6] S. Kazi, M. Rahim, and S. A. Khoja, "Crossing Language Boundaries: Evaluation of Large Language Models on Urdu-English Question Answering," in *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*. Abu Dhabi: Association for Computational Linguistics, 2025, pp. 141–151. [Online]. Available: https://aclanthology.org/2025.indonlp-1.17/

[7] A. Louis and G. Spanakis, "A statutory article retrieval dataset in French," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6789–6803. [Online]. Available: https://aclanthology.org/2022.acl-long.468/

[8] A. Louis, G. van Dijck, and G. Spanakis, "Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 22 266–22 275. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30232

[9] M. Büttner and I. Habernal, "Answering legal questions from laymen in German civil law system," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2015–2027. [Online]. Available: https://aclanthology.org/2024.eacl-long.122/

[10] A. Chen, F. Yao, X. Zhao, Y. Zhang, C. Sun, Y. Liu, and W. Shen, "Equals: A real-world dataset for legal question answering via reading chinese laws," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ser. ICAIL '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 71–80. [Online]. Available: https://doi.org/10.1145/3594536.3595159

[11] J. Lee, D. Kim, S. Hwang, H. Kim, and G. Lee, "KoBLEX: Open Legal Question Answering with Multi-hop Reasoning," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, nov 2025, pp. 4019–4053. [Online]. Available: https://aclanthology.org/2025.emnlp-main.200/

[12] R. Upadhya and S. T.Y.S.S, "LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, jul 2025, pp. 13 971–13 993. [Online]. Available: https://aclanthology.org/2025.acl-long.683/

[13] A. Zafar, S. Ashraf, and S. Nowaczyk, "From Courtroom to Corpora: Building a Name Entity Corpus for Urdu Legal Texts," in *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2025, pp. 1396–1405. [Online]. Available: https://aclanthology.org/2025.ranlp-1.161/

[14] Proxima AI, "Urdu Legal Instruct," Mozilla Data Collective, Jul. 2026, dataset card. Released 23 July 2026; accessed 11 October 2026. [Online]. Available: https://mozilladatacollective.com/datasets/cmrxgqrj8014znv07t3j3psj3

[15] M. K. Irshad, "Grounded or Silent: Citation-Faithful Legal Question Answering for Pakistani Law," sep 2026, preprint, version 1.1, 4 September 2026. [Online]. Available: https://irshados.com/ebooks/grounded-or-silent-paklegalqa/

[16] Z. Babar, N. I. Khan, M. Hassnain, and V. Filkov, "Open-source LLMs for technical Q&A: Lessons from stackexchange," in *SEET—Software Engineering for Emerging Technologies*, ser. Communications in Computer and Information Science, vol. 2725. Cham: Springer Nature Switzerland, 2026, pp. 615–626. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-032-08977-9_41

[17] D. Almeida, "Introducing system one models & jev," TypeSafe AI Blog, 2026. [Online]. Available: https://typesafe.ai/blog/introducing-system-one-models-and-jev

[18] National Assembly of Pakistan, "Constitution of pakistan (in urdu)," Online, 2018, eighth edition, as modified through 31 May 2018; preface dated 12 July 2018. [Online]. Available: https://na.gov.pk/uploads/documents/1575018297_955.pdf

[19] A. Rahman, A. Ghosh, and C. Arora, "UTRNet: High-resolution urdu text recognition in printed documents," in *Document Analysis and Recognition - ICDAR 2023*, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds. Cham: Springer Nature Switzerland, 2023, pp. 305–324. [Online]. Available: https://doi.org/10.1007/978-3-031-41734-4_19

[20] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[21] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[22] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[23] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319/

[24] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049

[25] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. [Online]. Available: https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf

[26] N. Holzenberger, A. Blair-Stanek, and B. Van Durme, "A dataset for statutory reasoning in tax law entailment and question answering," in *Proceedings of the Natural Legal Language Processing Workshop 2020*, 2020. [Online]. Available: https://ceur-ws.org/Vol-2645/paper5.pdf

[27] M. Hassnain, A. Basu, E. Ng, and C. Stanford, "Cargo Sherlock: An SMT-based checker for software trust costs," in *Proceedings of the IEEE/ACM 14th International Conference on Formal Methods in Software Engineering (FormaliSE '26)*. New York, NY, USA: Association for Computing Machinery, 2026, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3793656.3793687