Grounded or Silent: Citation-Faithful Legal Question Answering for Pakistani Law

DOI 10.5281/zenodo.22305084

Muhammad Kashif Irshad

Correspondence: dev@irshados.com – Lahore, Pakistan

ORCID 0009-0008-9161-9875

Preprint v1.1 | 2026-09-04 (v1: 2026-08-21)

IrshadOS Research

irshados.com

Cite as: Irshad, Muhammad Kashif. "Grounded or Silent: Citation-Faithful Legal Question Answering for Pakistani Law." IrshadOS Research, 2026. DOI: 10.5281/zenodo.22305084.

All versions (resolves to the latest): 10.5281/zenodo.22037182

Web edition, data and evaluation harness: https://irshados.com/ebooks/grounded-or-silent-paklegalqa/

Contents

1. Abstract

Legal AI tools are known to hallucinate: audits of commercial legal research products report fabricated or misgrounded authorities in 17–33% of responses (Magesh et al., 2024). Existing legal QA benchmarks cover well-resourced jurisdictions; for Pakistan — a mixed common-law system of 240 million people — no benchmark spans its statutes and case law. We present **PakLegalQA**, the first citation-grounded question-answering benchmark for Pakistani law: 300 questions over 946 federal statutes (Pakistan Code) and 2,503 reported Lahore High Court judgments (2022–2026), with gold statute sections and neutral citations, an unanswerable subset whose only correct response is a refusal, and a 60-question code-switched Roman-Urdu overlay reflecting how users actually type. Using PakLegalQA we evaluate a deployed, production grounded-or-refuse RAG system and ablations of its retrieval stack against closed-book and vanilla-RAG baselines. The production system attains 74.5% gold-source retrieval — 14.5 points above vanilla dense retrieval, with title-affinity re-ranking and lexical rescues (citation lookup, name matching, deep reading of named judgments) contributing roughly 7 and 9 points respectively — while over-refusing on only 5.3% (Sol) to 0.6% (gpt-4o) of answerable questions. The closed-book baseline refuses just 7 of 84 unanswerable questions, answering the rest from parametric memory — the grounding-discipline gap the benchmark is designed to expose; grounded systems refuse or honestly signal insufficiency on the large majority. A 24-item stratified manual audit finds 87.5% correct answers and zero fabricated citations: the grounded system's failure mode is mis-selection or silence, never invention. We additionally document a query-rewrite-stage hallucination (the rewrite model injecting a wrong-jurisdiction statute), a failure stage named but unmeasured in prior work, and an infrastructure-failure episode that masqueraded as a calibration collapse — motivating positional sanity checks in LLM evaluation harnesses. The benchmark, harness, and corpus manifest are released.

Keywords: legal question answering; retrieval-augmented generation; benchmark; Pakistani law; refusal calibration; citation grounding; Roman Urdu; legal NLP

2. Introduction

When a lawyer asks a machine what the law says, there are only two honest outcomes: an answer grounded in a citable text, or an admission that the text does not answer. Everything else is a hallucination wearing a confident voice. Audits of commercial legal research tools have shown how common that third outcome is: Magesh et al. (2024) found purpose-built, retrieval-augmented products from the largest legal publishers hallucinating in 17–33% of responses, and Dahl et al. (2024) documented the taxonomy of ways large language models invent law. The stakes are not abstract. A fabricated citation in a pleading is a sanctionable event; a fabricated answer to a layperson may shape a decision they cannot afford to get wrong.

These audits, and the benchmarks that inform them, concentrate on well-resourced jurisdictions. Pakistan — a mixed common-law system serving 240 million people, whose statutes and reported judgments are published in English while its users increasingly ask their questions in Urdu or romanised Urdu — has no benchmark spanning its statutes and case law. The nearest prior work, LEGAL-UQA (Faisal & Yousaf, 2024), covers the Constitution alone: 619 QA pairs, without statute-section or case-citation gold labels, without an unanswerable subset, and in formal parallel Urdu rather than the code-switched

Roman Urdu that real queries use. India, by contrast, is comparatively well served (IL-TUR, ILDC, IndicLegalQA, IL-PCSR), underscoring that the gap is jurisdictional, not regional.

This paper makes four contributions. **First, PakLegalQA** — the first citation-grounded QA benchmark for Pakistani law: 300 questions over 946 federal statutes (the Pakistan Code) and 2,503 reported Lahore High Court judgments (2022–2026), with gold statute sections and neutral citations, seven question types including false-premise and unanswerable subsets, and a 60-question code-switched Roman-Urdu overlay. **Second, a refusal-calibration evaluation of a deployed system.** The system under study is not a laboratory pipeline but a production grounded-or-refuse RAG service with live users; to our knowledge no prior legal-QA evaluation measures a deployed system's abstention behaviour. **Third, component ablations of a domain retrieval stack**, quantifying what category pooling, title-affinity re-ranking, and lexical rescues (citation lookup, name matching, deep reading of named judgments) each contribute beyond vanilla dense retrieval. **Fourth, two methodological findings**: a hallucination at the query-generation stage of RAG — the retrieval-rewrite model injecting a wrong-jurisdiction statute into the search — which prior work names as a possible failure stage but does not measure; and an infrastructure-failure episode in which silent upstream API exhaustion produced a plausible-looking calibration collapse, detectable only by positional analysis — an argument for position- and time-aware sanity checks in any LLM evaluation harness.

3. Related Work

Hallucination in legal AI. Magesh et al. (2024) audited proprietary legal research tools with a preregistered query set and a correctness × groundedness coding scheme that we adopt; Dahl et al. (2024) provide the underlying taxonomy. Our setting differs in that the system is open to us: we can attribute failures to retrieval, ranking, or generation, and intervene.

Legal QA benchmarks. LegalBench (Guha et al., 2023) and LexGLUE (Chalkidis et al., 2022) established construction conventions for legal task suites; COLIEE's yearly competitions and the LeCaRD/LeCaRDv2 datasets (Ma et al., 2021; Li et al., 2024) cover case retrieval and entailment for Japanese and Chinese law. For Pakistan, LEGAL-UQA (Faisal & Yousaf, 2024) is the closest work and, notably, also originates in Lahore; PakLegalQA extends the jurisdiction's coverage from the Constitution to statutes and case law, and from answer strings to citation-level gold labels with an abstention axis.

Citation-faithful generation and abstention. ALCE (Gao et al., 2023) supplies the citation precision/recall framing we adapt to statutes and judgments. Self-RAG (Asai et al., 2023) and calibration studies (Kadavath et al., 2022) treat refusal as a trainable, measurable behaviour. Recent legal-RAG evaluations — LegalBench-RAG, Legal RAG Bench (Butler & Butler, 2026), ClaimRAG-LAW, LegalCiteBench — measure retrieval precision and citation reliability on common-law corpora; SearchFireSafety (Chae et al., 2026) is the closest methodological neighbour, probing hierarchical statutory retrieval and safe abstention in a single regulatory domain. None evaluates a deployed system, and none covers Pakistan.

Retrieval. BM25 (Robertson & Zaragoza, 2009) and late-interaction dense retrieval (Khattab & Zaharia, 2020) anchor the baseline families; SAILER (Li et al., 2023) motivates structure-aware legal retrieval, which our statute/caselaw pooling and judgment deep-reading instantiate in production.

4. The PakLegalQA Benchmark

4.1 Corpus

The corpus is a frozen snapshot (paklegalqa-corpus-v1, SHA-256 manifest released) of 3,449 documents: 946 federal statutes from the official Pakistan Code and 2,503 reported judgments of the Lahore High Court, 2022–2026, each carrying its neutral citation and official source URL. One coverage gap was found and is disclosed: the Court Fees Act 1870 is absent. A second gap reported in v1 of this paper — that the ingested Limitation Act 1908 lacks its First Schedule — was misdiagnosed, and is corrected here. The schedule text was present in the ingested document throughout the evaluation window, but no article of it was addressable by the retrieval stack, so its behaviour was indistinguishable from absence. The distinction matters: the cause is a fixable indexing defect on our side, not a missing source document. The ingested Pakistan Penal Code reflects amendments at least through the Criminal Laws (Amendment) Act 2022 (XXXVII of 2022), verified via the decriminalised s.325 (attempted suicide).

4.2 Question types and construction

Type	Count	Description
S-REC	80	Statute recall — punishments, definitions, procedures
S-INT	60	Interpretation — real-life scenarios mapped to provisions
C-HOLD	50	Case holdings — what a named judgment decided
C-FACT	20	Judgment facts — judge, bench, dates
TIME	20	Amendment- and currency-sensitive probes
FALSE	20	False premises — wrong years, wrong codes, fictional sections
UNANS	50	Unanswerable from the corpus — refusal is the only correct output
RU	60	Roman-Urdu overlay — code-switched twins sharing gold labels

Questions were authored in six batches against three sourcing streams (statute-derived, headnote/disposition-derived, and production-query-derived phrasings), under a corpus-verification rule: no statute gold was recorded unless the section's text was located in the ingested corpus (115 automated checks across five verification waves), and no holding gold unless the disposition sentence was read in the judgment's text (a regex harvester over all chunks, since final pages are often signature blocks). False-premise items encode realistic confusions; unanswerable items span provincial law, other courts, current affairs, and outcome predictions, so that the correct behaviour is refusal precisely because the corpus — not the world — cannot answer.

The Roman-Urdu overlay re-asks 60 questions sampled across types in code-switched romanised Urdu as users actually type ("Cheque bounce hone par kya saza hai?"), sharing gold labels with their English twins so outcome divergence isolates language handling.

4.3 Gold labels and verification status

Each item carries gold sources (statute title + section, or neutral citation), short gold answer points for automated scoring, an answerable flag, and provenance metadata. All labels were authored by the first

author and corpus-verified automatically; independent legal review is planned, and its absence in v1 is stated as a limitation (§9).

5. System Under Study

The evaluated system is Irshad AI Employee, a deployed multi-tenant RAG service whose answering contract is *grounded or silent*: answers must come from retrieved documents with page-cited sources, and the model must emit an insufficiency sentinel — surfaced to users as an honest refusal — when the context does not answer. The retrieval stack layers, over dense embeddings with a relevance floor and per-document diversity caps: (i) statute/caselaw category pooling, so terse statutory prose is not crowded out by discursive judgments; (ii) title-affinity re-ranking toward acts the question names; (iii) lexical rescues — per-term title keyword matching for party names, and regex citation lookup ("2025 LHC 846") fetching the cited judgment directly; and (iv) deep reading of named judgments at both ends, because dispositions live in closing pages. A query-rewrite model normalises Roman-Urdu and colloquial phrasings into English statutory vocabulary before embedding. The production answer model is a reasoning-class model ("Sol"); ablations use gpt-4o with the generator held constant within every comparison.

6. Experimental Setup

Six systems: PROD (full stack) and CB (closed-book, no retrieval, prose refusals permitted) on the production model; PROD, DENSE (vanilla dense retrieval — no pools, boosts, or rescues), −TITLE (full minus title affinity) and −RESCUE (full minus lexical rescues and deep reading) on gpt-4o. Runs are executed offline against the corpus snapshot through an evaluation harness that bypasses tenant billing but not the real APIs; ablation switches are evaluation-only context flags that production code paths never set, with a regression test asserting the modes differ. Automatic metrics: answer/refusal rates, refusal calibration (correct refusal on unanswerables versus over-refusal on answerables), gold-source retrieval hit, and Roman-Urdu/English twin outcome parity. A stratified manual coding pass (correctness × groundedness, per Magesh et al., 2024, with the §8-F5 rubric extensions) complements the automatic layer.

7. Results

System	Answered	Correct refusal (of 84)	Over-refusal	Gold-source hit	RU twins same
PROD (Sol)	84.7%	36/84	5.3%	74.5%	52/60
Closed-book (Sol)	81.7%	7/84	16.4%	–	60/60
PROD (gpt-4o)	89.7%	35/84	0.6%	74.5%	59/60
DENSE (gpt-4o)	84.2%	35/84	6.1%	60.0%	58/60
− title affinity (gpt-4o)	90.3%	34/84	0.3%	67.3%	59/60
− rescues/deep-read (gpt-4o)	83.6%	36/84	6.4%	65.8%	59/60

Table 1: automatic metrics over all 360 items (2,160 evaluations). "Gold-source hit": a required gold source appears among the answer's cited sources. "Correct refusal" counts explicit refusals only; grounded

non-answers and premise corrections — which manual coding shows dominate the remainder on unanswerable items — are analysed in §7.1.

R1 — Grounding discipline (PROD vs closed-book). Grounded systems refuse 34–36 of the 84 unanswerable/false-premise items outright; the closed-book baseline refuses 7, answering 77 from parametric memory — including Supreme Court holdings, current-affairs facts, and provincial rules outside any provided corpus. Notably, closed-book *is* honest where its ignorance is self-evident (it declines specific LHC citation lookups); its failure is answering questions whose premise requires a corpus it does not have. This is the benchmark's central contrast: unanswerable items measure discipline, not knowledge.

R2 — What the retrieval stack buys. The full stack retrieves a required gold source for 74.5% of scorable items against 60.0% for vanilla dense retrieval (+14.5 points). Removing title affinity costs 7.2 points of gold-source accuracy while *raising* the answer rate slightly — it changes what is cited, not whether the system speaks. Removing the lexical rescues and deep reading costs 8.7 points of gold-source accuracy and 6.1 points of answer rate (≈22 recovered answers), concentrated in the citation-lookup and case-holding classes that embeddings cannot serve.

R3 — Refusal calibration. Over-refusal on answerable items is 5.3% for the production reasoning model and 0.6% for gpt-4o. Per-type analysis places the residual over-refusals overwhelmingly in scenario-phrased interpretation questions — the class with the weakest gold-source retrieval — indicating the remaining silence is retrieval-caused, not policy-caused.

R4 — Roman-Urdu parity. After the translation-first rewrite (§8, F4), outcome parity between Roman-Urdu twins and their English counterparts is 59/60 (gpt-4o) and 52/60 (Sol). The benchmark's paired design is what makes this measurable at all: every divergent pair is an actionable defect report.

R5 — Model temperament. Sol errs toward caution (higher over-refusal, slightly lower RU parity); gpt-4o toward helpfulness (near-zero over-refusal, a softer refusal edge on unanswerables). Manual inspection (§7.1) shows gpt-4o's extra "answers" on unanswerable items are largely honest: explicit insufficiency statements and implicit premise corrections rather than fabrications.

7.1 Manual audit

A stratified 24-item audit of PROD outputs (Magesh-style coding with the §8-F5 rubric extensions) finds 21/24 (87.5%) correct or better — including four grounded non-answers and three implicit premise corrections — and 3/24 incorrect, all scenario questions where retrieval selected a plausible but wrong provision (a sister doctrine, a neighbouring section, a different act). **No fabricated citation appears anywhere in the audit**: every cited source exists and is quoted faithfully. The grounded system's observed failure mode is mis-selection or silence — never invention.

8. Discussion: Findings from a Deployed System

F1 — Query-rewrite hallucination. The retrieval-rewrite model expanded "punishment for dishonour of a cheque" into "…under the Negotiable Instruments Act" — the *Indian* location of that offence — steering both dense retrieval and title-affinity toward the wrong statute while PPC s.489-F sat unretrieved. The fix

(a jurisdiction guard with corrected worked examples) and its measurement are, to our knowledge, the first documentation of hallucination at the query-generation stage of a legal RAG pipeline.

F2/F3 — What lexical rescues buy. Neutral citations tokenize below keyword-length floors and carry no embedding signal; party names collide across thousands of titles; dispositions live in closing pages. Each defect was found by a benchmark question and repaired by a rescue component; the DENSE ablation quantifies their joint value.

F4 — Cross-lingual regressions are silent. A one-line prompt improvement for jurisdiction correctness silently disabled Roman-Urdu translation; only the benchmark's language twins exposed it. Prompt changes need cross-lingual regression suites.

F5 — The rubric needs an honesty category. Real outputs forced two coding refinements: implicit premise correction (answering the real statute behind a false premise) and the grounded non-answer ("the documents do not state…"), both of which are correct behaviours that naive answered/refused coding miscounts.

F6 — Infrastructure failures masquerade as findings. Upstream credit exhaustion mid-run caused embeddings to fail silently; the engine's honest refusals then looked like a calibration collapse (windowed refusal rates: 10, 10, 13, 19, 60, 60 per 60 items). Positional sanity checks belong in LLM evaluation harnesses; we release ours.

9. Limitations

Single-annotator gold labels (corpus-verified, not lawyer-verified — planned for v2); one High Court's reported judgments (2022–2026) and federal statutes only; the evaluated system is built by the author (mitigated by releasing the benchmark, harness, and manifest so anyone can rerun and extend); automatic metrics validated by a bounded manual audit rather than exhaustive coding; BM25/hybrid baselines deferred.

10. Ethics Statement

Statutes and reported judgments are public documents; the benchmark redistributes citations, sections, and short verified quotes with links to official sources rather than bundling texts. Production-derived questions are paraphrased phrasings with no user identifiers; raw logs are never released. No copyrighted law-report headnotes (PLD/PLJ/CLC/SCMR) are used. The system evaluated presents itself to users as research assistance, not legal advice.

11. AI Assistance Disclosure

This research was conducted and drafted with substantial AI assistance (Anthropic's Claude: experiment tooling, corpus verification scripts, analysis, and prose drafting). All experiments, data, gold labels, and claims were reviewed by the human author, who bears sole responsibility for the content.

12. Data and Code Availability

The PakLegalQA benchmark, the evaluation harness, the corpus manifest and the per-run result files are released at https://github.com/mkirshad/grounded-or-silent. The repository holds the 360 benchmark items with their gold labels (benchmark/all-360.jsonl), the SHA-256 manifest of the frozen paklegalqa-corpus-v1 snapshot (corpus/manifest-v1.jsonl), the scoring and run scripts that produced every table in this paper (scripts/), and the stratified manual-audit coding sheet (results/).

An archived copy of this preprint is deposited at https://doi.org/10.5281/zenodo.22305084 (all versions: https://doi.org/10.5281/zenodo.22037182), where the repository is recorded as supplementary material. Statute and judgment texts are not redistributed: the manifest carries each document's neutral citation and official source URL, so the snapshot can be reconstructed from the official Pakistan Code and Lahore High Court portals.

The released material is licensed for reuse: the benchmark items, gold labels, corpus manifest and result files under Creative Commons Attribution 4.0 International, and the harness and scoring scripts under the MIT Licence. The underlying statutes and judgments are public legal documents and are not ours to license, which is why the manifest carries citations, hashes and official URLs instead of the texts themselves.

Changes in v1.1. (i) This section was added: v1 stated that the benchmark, harness and corpus manifest were released without saying where. (ii) The corpus disclosure in §4.1 was corrected. The ingested Limitation Act 1908 does contain its First Schedule; v1 reported it as missing, when it was unreachable to retrieval rather than absent. The three benchmark items whose gold source is that schedule (TIME-012, TIME-057, TIME-127) are labelled answerable in both versions, so no gold label, score, or table changes — the correction is to a stated cause, not to any result. The Court Fees Act 1870 gap is unchanged and real. (iii) Licence terms for the released material were stated for the first time, and LICENSE and LICENSE-DATA files were added to the repository. (iv) A right-to-left rendering defect in the v1 PDF, which mirrored the running header, was fixed in the document source.

13. References

1. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools.* arXiv:2405.20362.
2. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models.* arXiv:2401.01301.
3. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). *Enabling Large Language Models to Generate Text with Citations.* arXiv:2305.14627.
4. Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* arXiv:2005.11401.
5. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.* arXiv:2310.11511.
6. Kadavath, S., et al. (2022). *Language Models (Mostly) Know What They Know.* arXiv:2207.05221.
7. Guha, N., et al. (2023). *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models.* arXiv:2308.11462.
8. Chalkidis, I., et al. (2022). *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.* arXiv:2110.00976.

9. Khattab, O., & Zaharia, M. (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.* arXiv:2004.12832.

10. Robertson, S., & Zaragoza, H. (2009). *The Probabilistic Relevance Framework: BM25 and Beyond.* Foundations and Trends in Information Retrieval, 3(4).

11. Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M., & Ma, S. (2021). *LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System.* SIGIR '21.

12. Li, H., et al. (2024). *LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset.* arXiv:2310.17609.

13. Li, H., et al. (2023). *SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval.* arXiv:2304.11370.

14. Faisal, F., & Yousaf, U. (2024). *LEGAL-UQA: A Low-Resource Urdu-English Dataset for Legal Question Answering.* arXiv:2410.13013.

15. Joshi, A., Paul, S., Sharma, A., Goyal, P., Ghosh, S., & Modi, A. (2024). *IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning.* ACL 2024.

16. Chae, K., et al. (2026). *Beyond Case Law: Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA.* arXiv:2604.06173.

17. Butler, A.-R., & Butler, U. (2026). *Legal RAG Bench: An End-to-End Benchmark for Legal RAG.* arXiv:2603.01710.

18. Goebel, R., et al. (2024). *Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024.* JURISIN 2024.