Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrieval-Augmented Generation for Taxpayer Advisory in an LMIC Setting: A Proof-of-Concept Evaluation of Statutory Legal Retrieval

Omara Innocent
Universal Technology and management University (UTAMU)
Okasio Emmanuel
Universal Technology and management University (UTAMU)
Akello Matilda Sandra
Universal Technology and management University (UTAMU)
Fred Paul Mark Jjunju
fpmjjunju@utamu.ac.ug

Universal Technology and management University (UTAMU)
Denis Bbosa
Mbarara University of Science and Technology

Short Report

Keywords: retrieval-augmented generation, legal information retrieval, taxpayer services, tax administration, low- and middle-income countries, large language models, semantic retrieval, sustainable development goals

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11181452/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective

Tax administrations in low- and middle-income country settings are expanding digital services, yet reliable, legally grounded taxpayer guidance remains difficult to scale. This study developed and evaluated a proof-of-concept retrieval-augmented generation framework for taxpayer advisory using national income-tax legislation as an authoritative knowledge source. The objective was to test whether semantic retrieval could identify statutory provisions accurately enough to support grounded downstream language-model prompting. The legislation was transformed into 198 knowledge units, embedded with all-MiniLM-L6-v2, indexed in ChromaDB, and evaluated using 900 automatically generated questions linked to expected provisions.

Results

The retriever achieved 72.00% Top-1 accuracy, 86.78% Top-3 accuracy, 28.93% Precision@3, 86.78% Recall@3, and mean reciprocal rank 0.7880. Many Top-1 errors were recovered within the first three results, indicating ranking ambiguity among semantically related provisions rather than complete retrieval failure. The prototype integrated legal document processing, vector search, provision retrieval and structured prompt construction in an end-to-end workflow. The results support technical feasibility for grounded taxpayer information, but generation quality, legal correctness, usability and compliance effects were not evaluated. Validation should include additional tax instruments, real taxpayer queries, professional legal review, multilingual support and cross-jurisdiction testing before deployment in resource-constrained public-service settings.

Introduction

Retrieval-augmented generation (RAG) combines large language models with external retrieval so that responses can be conditioned on an explicit knowledge source rather than relying only on model parameters [1, 5]. Dense retrieval and sentence-embedding methods provide the semantic-search layer on which many RAG systems depend [2–4]. Recent evaluation work stresses that retrieval quality should be measured separately from generation quality because a fluent answer can still be unsupported if the wrong evidence is retrieved [6, 7]. This distinction is especially important in legal applications, where benchmark studies show both progress and persistent limitations in large-model legal reasoning and retrieval [7, 8].

Tax law is a demanding setting for grounded AI because statutes are structured, interdependent, frequently amended, and economically consequential. Prior work has shown that supplying relevant legal authority can improve large-language-model performance on tax questions, but performance remains below expert reliability in important settings [9, 10]. Statutory information retrieval is therefore a necessary, though not sufficient, component of any defensible tax-advisory RAG pipeline. Work on legal

information retrieval similarly shows that finding the correct statutory text is itself a substantive technical problem [11].

The policy context extends beyond a single jurisdiction. Tax administrations are rapidly digitalising taxpayer services and increasingly experimenting with artificial intelligence and virtual assistants [12, 13]. Evidence from African tax administrations also shows that digital technologies can facilitate compliance while introducing adoption, capacity and inclusion challenges [14, 15]. In developing economies, taxpayer facilitation forms part of the broader domestic-resource-mobilisation agenda [16]. This gives the present proof of concept potential relevance to Sustainable Development Goal (SDG) 17.1 on tax and revenue capacity, SDG 16.6 on effective public institutions, and SDG 8.3 on the formalisation and growth of micro-, small- and medium-sized enterprises [17–19]. These are relevance pathways rather than measured outcomes of this study.

Design and statutory knowledge base

The study used Design Science Research Methodology to guide problem definition, artefact development, demonstration and evaluation [20]. The empirical legal source was the Uganda Income Tax Act, Cap. 338 [21], used here as one low-income-country statutory case rather than as evidence of cross-country generalisability. The Act was parsed into 198 legal knowledge units. Each unit contained a unique Knowledge ID and associated Part, Section, Title and Content metadata. Validation found no missing values or duplicate Knowledge IDs. Full design-science stages, system components and knowledge-base descriptive statistics are provided in Additional file 1 (Tables S1-S2).

Each provision was encoded as a 384-dimensional semantic vector using the sentence-transformers all-MiniLM-L6-v2 model, consistent with the broader sentence-embedding and MiniLM literature [3, 4]. The 198 x 384 embedding matrix was indexed in a persistent ChromaDB collection. A natural-language query was embedded with the same model, used for semantic similarity search, and mapped to the three highest-ranked statutory provisions. The retrieved provisions and metadata were then assembled into a structured prompt for downstream language-model use.

Retrieval evaluation

Evaluation was deliberately restricted to the retrieval component; the study did not score language-model answers. A benchmark of 900 automatically generated natural-language questions was created from the statutory knowledge base. Each question was linked to one expected Knowledge ID, which served as the study ground truth. For every question, the system returned three candidate provisions. Top-1 Accuracy measured whether the expected provision ranked first; Top-3 Accuracy and Recall@3 measured whether it appeared within the first three; Precision@3 reflected one designated relevant item among three returned items; and mean reciprocal rank (MRR) captured the rank position of the expected provision. Metric definitions and the evaluation workflow are detailed in Additional file 1 (Table S3; Fig. S2).

Results

The legal corpus comprised 198 knowledge units spanning 24 Parts. The semantic-retrieval benchmark achieved 72.00% Top-1 Accuracy and 86.78% Top-3 Accuracy. Precision@3 was 28.93%, Recall@3 was 86.78%, and MRR was 0.7880 (Table 1). Because each query had one designated relevant provision, Precision@3 is mathematically constrained by that single-relevant-item evaluation design and should not be interpreted as a conventional multi-relevance precision estimate.

Approximately 28% of evaluation questions were not matched to the expected provision at rank one. A substantial share of these were nevertheless recovered within the top three. The source analysis reported that many rank-one misses involved semantically related legal concepts, suggesting ranking ambiguity rather than complete retrieval failure. The prototype also demonstrated the end-to-end sequence from legal document processing through embedding, vector indexing, retrieval and prompt construction. Detailed implementation parameters and study boundaries are provided in supporting information (Tables Summary of System and Technical Components S1 and Table S2 Summary of Implementation and Evaluation Results).

Table 1
Retrieval evaluation results

Measure	Result
Legal knowledge units	198
Evaluation questions	900
Top-1 Accuracy	72.00%
Top-3 Accuracy	86.78%
Precision@3	28.93%
Recall@3	86.78%
Mean Reciprocal Rank	0.7880

Values are based on 900 automatically generated questions, each linked to one expected Knowledge ID.

Interpretation and LMIC relevance

The 86.78% Top-3 result indicates that the retriever often placed the designated statutory provision within the context window that would be supplied to a downstream generator. This is encouraging for proof-of-concept use, but it is not equivalent to 86.78% advisory accuracy. LegalBench-RAG similarly emphasizes that retrieval should be evaluated as a distinct stage of legal RAG [7], while tax-law studies show that access to legal context can improve model performance without eliminating reasoning errors

[9, 10]. The present study therefore supports a retrieval layer, not a claim that the system can safely provide autonomous tax advice.

For LMIC tax administrations, the more defensible contribution is methodological: a relatively compact statutory corpus can be converted into a machine-retrievable knowledge base and evaluated with explicit ranking metrics. This may be relevant where administrations seek scalable taxpayer information services while facing limited specialist capacity. However, transfer to other countries would require jurisdiction-specific corpora, version control for legislative amendments, professional legal validation, language adaptation, governance and security controls, and evaluation with real taxpayer questions. Additional file 1 (Table S4) maps these potential policy links to SDGs 8.3, 16.6 and 17.1 while explicitly distinguishing them from outcomes measured in this study.

Limitations

• The study used one national income-tax statute and therefore does not establish cross-jurisdiction generalisability across LMICs or other tax domains.

• The 900 evaluation questions were automatically generated from the same legal knowledge base used to construct the retriever. They may be less linguistically diverse and less ambiguous than questions posed by real taxpayers.

• Each question had one designated ground-truth provision even though real tax questions can legitimately require multiple interacting provisions, regulations, administrative guidance or case law.

• Only semantic retrieval was quantitatively evaluated. The legal correctness, completeness, faithfulness and safety of downstream large-language-model answers were not assessed.

• No comparison was reported against sparse retrieval, hybrid retrieval, cross-encoder re-ranking, alternative embedding models or alternative chunking strategies, so the chosen architecture cannot be claimed to be optimal.

• The study did not evaluate user experience, professional acceptance, tax compliance, revenue mobilisation, institutional effectiveness or SME formalisation. The SDG links described are therefore prospective policy relevance rather than demonstrated impact.

Abbreviations

AI: artificial intelligence; DSRM: Design Science Research Methodology; LLM: large language model; LMIC: low- and middle-income country; MRR: mean reciprocal rank; RAG: retrieval-augmented generation; SDG: Sustainable Development Goal; SME: small and medium-sized enterprise.

Declarations

Ethics approval and consent to participate

Not applicable. The study used publicly available statutory material and an automatically generated benchmark and did not involve human participants, human data, human tissue or animals.

Consent for publication

Not applicable.

Availability of data and materials

The data, source code, parser and evaluation script used in this study can be availed upon request.

Competing interests

None

Funding

None

Authors' contributions

All authors revised and approved the final manuscript.

Acknowledgements

None

Additional files

None

References

1. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv Neural Inf Process Syst. 2020;33:9459–74.

2. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S et al. Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of EMNLP 2020. Association for Computational Linguistics; 2020. pp. 6769–6781. 10.18653/v1/2020.emnlp-main.550
3. Reimers N, Gurevych I, Sentence-BERT. Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of EMNLP-IJCNLP 2019. Association for Computational Linguistics; 2019. pp. 3982–3992. 10.18653/v1/D19-1410
4. Wang W, Bao H, Huang S, Dong L, Wei F. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. In: Findings of ACL-IJCNLP 2021. Association for Computational Linguistics; 2021. pp. 2140–2151. 10.18653/v1/2021.findings-acl.188
5. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. 2023;arXiv:2312.10997.
6. Es S, James J, Espinosa-Anke L, Schockaert S, RAGAs. Automated Evaluation of Retrieval Augmented Generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024. pp. 150–158. 10.18653/v1/2024.eacl-demo.16
7. Pipitone N, Houir Alami G. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv. 2024;arXiv:240810343.
8. Guha N, Nyarko J, Ho DE, Re C, Chilton A, Chohlas-Wood A et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems. 2023.
9. Nay JJ, Karamardian D, Lawsky SB, Tao W, Bhat M, Jain R, et al. Large language models as tax attorneys: a case study in legal capabilities emergence. Philos Trans Math Phys Eng Sci. 2024;382:20230159. 10.1098/rsta.2023.0159.
10. Gogani-Khiabani S, Trivedi A, Chyi S, Tizpaz-Niari S. Performance of LLMs on VITA test: potential for AI-assisted tax returns for low income taxpayers. Artif Intell Law. 2025. 10.1007/s10506-025-09465-7.
11. Savelka J, Ashley KD. Legal information retrieval for understanding statutory terms. Artif Intell Law. 2021. 10.1007/s10506-021-09293-5.
12. OECD. Tax Administration 2024: Comparative Information on OECD and other Advanced and Emerging Economies. Paris: OECD Publishing; 2024. 10.1787/2d5fba9c-en
13. OECD. Tax Administration Digitalisation and Digital Transformation Initiatives. Paris: OECD Publishing; 2025. 10.1787/c076d776-en
14. Okunogbe O, Santoro F. Increasing tax collection in African countries: the role of information technology. J Afr Econ. 2023;32(Suppl 1):i57–83.
15. Tinta JK, Zerbo M, Santoro F, Diouf A, Pale K. Electronic Services and Tax Compliance: Evidence from Medium and Small Businesses in Burkina Faso. ICTD Working Paper 209. Brighton: Institute of Development Studies; 2024. 10.19088/ICTD.2024.099

16. Dom R, Custers A, Davenport S, Prichard W. Innovations in Tax Compliance: Building Trust, Navigating Politics, and Tailoring Reform. Washington, DC: World Bank; 2022. 10.1596/978-1-4648-1755-7.
17. United Nations Department of Economic and Social Affairs. Goal 17: Partnerships for the Goals - Target 17.1. https://sdgs.un.org/goals/goal17. Accessed 24 Sep 2026.
18. United Nations Department of Economic and Social Affairs. Goal 16: Peace, Justice and Strong Institutions - Target 16.6. https://sdgs.un.org/goals/goal16. Accessed 24 Sep 2026.
19. United Nations Department of Economic and Social Affairs. Goal 8: Decent Work and Economic Growth - Target 8.3. https://sdgs.un.org/goals/goal8. Accessed 24 Sep 2026.
20. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst. 2007;24(3):45–77. 10.2753/MIS0742-1222240302.
21. Uganda Revenue Authority. Income Tax Act, Cap. 338. Kampala: Uganda Revenue Authority. https://www.ura.go.ug/. Accessed 24 Sep 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- LMICTaxRAGSupplementary.docx