Question Answering Techniques for Portuguese Legal Documents: A Systematic Literature Review

Maurício Rodrigues Lima[1], **Vinícius Teles**[1], **Sávio Teles**[1], **Elisângela Silva Dias**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74.001-970 – Goiás – GO – Brasil

{mauricio.rodrigues,viniteles}@discente.ufg.br

{savioteles,elisangelasd}@ufg.br

Abstract. *The exponential growth of Portuguese-language legal documents has renewed interest in Question Answering (QA) systems capable of returning concise, legally sound answers to natural-language queries. This study presents a systematic literature review, conducted according to PRISMA 2020 guidelines, that synthesises current evidence on QA techniques applied to Lusophone legal texts. Searches, without temporal restrictions, were executed in nine databases (ACM, El Compendex, ISI Web of Science, Periódico Capes, Scielo, Science@Direct, Scopus, Sol SBC and Springer Link) using a string that combine jurisprudential, linguistic and methodological terms. After duplicate removal, independent screening and quality appraisal, ten primary studies met the inclusion criteria (peer-reviewed publications developing or evaluating QA pipelines over Brazilian or Portuguese legislation). Publication activity is recent: more than 70% of the papers appeared between 2023 and 2025 and focus on Brazilian statutes and court decisions. Most pipelines adopt hybrid retrieval—BM25 or symbolic regex filters coupled with BERT-family dense encoders fine-tuned on legal corpora, while Retrieval-Augmented Generation with GPT-class models emerges in the latest research. Reported exact-match scores range from 0.60 to 0.83 and F1 from 0.75 to 0.87; however, only a quarter of the studies release code or data, hindering reproducibility. Common gaps include limited handling of the temporal validity of norms, scarce evaluation by legal specialists, and the absence of benchmark datasets for Portuguese. Overall, QA research for Lusophone law is accelerating yet remains fragmented; future work should prioritize shared resources, temporally aware models, and metrics that capture legal soundness beyond lexical overlap.*

1. Introduction

The volume of legal documents in Portuguese has grown exponentially since the digitalization of Brazilian and Portuguese courts, with particular emphasis on the organizational role of these institutions [Filho 2019]. Laws, judicial decisions, contracts, and legal opinions now form vast repositories that make it difficult to locate specific information. The increase in international legal cooperation has also contributed to this scenario [Perlingeiro and Ghio 2020].

Question Answering (QA) systems, capable of responding to questions posed in natural language, have emerged as a promising solution to democratize access to legal information and support decision-making [Filho 2019]. However, applying QA to the

Portuguese-speaking legal domain poses challenges such as the linguistic peculiarities of Portuguese, the hierarchical normative structure, and the need to account for the temporal validity of legal provisions [Ramos 2017]. Generic models tend to fail in these aspects, compromising legal certainty.

Recent literature shows advances in adapting dense retrieval techniques, neural reranking, and answer generation to the Portuguese context, with the use of corpora such as the *eSafeguard* of the Brazilian Supreme Court and models like *BERTimbau*. Nevertheless, there remains a lack of benchmarks, appropriate legal metrics, and studies on temporal reasoning and integration with citation graphs [Ransolin and Baruffi 2022, Perlingeiro and Ghio 2020].

This article presents a systematic literature review on legal QA in Portuguese, mapping data types, techniques, and metrics, while identifying gaps and future directions. The study follows the PRISMA 2020 protocol [Page et al. 2021], including searches in ten national and international databases and assessing the methodological quality of the included studies. The contributions include: (i) a quantitative and qualitative overview of the scientific production on legal QA in the Portuguese-speaking context; (ii) a synthesis of current pipeline practices and limitations; and (iii) the provision of a replicable protocol for future research [Ramos 2017, Filho 2019].

The remainder of the article is organized as follows: Section 2 discusses the related work and highlights how this review differs from previous studies. Section 3 describes the methodology adopted, including the PRISMA 2020 protocol [Page et al. 2021], the databases consulted, and the study selection criteria. Section 4 presents the review planning, research questions, and search string. Section 6 reports the results and answers the research questions. Finally, Section 7 concludes the paper, summarizes the main contributions, and suggests directions for future research.

2. Related Works

The survey conducted by [Martinez-Gil 2023] consolidates nearly three decades of research in Legal Question Answering (LQA), categorizing classical Information Retrieval solutions, approaches based on knowledge graphs, and more recently, neural methods. Although the study provides a useful overview, it focuses on multiple jurisdictions and languages, and does not delve into issues specific to Portuguese (pt-BR/pt-PT), nor does it adopt a systematic review methodology (e.g., PRISMA). The present study complements this overview by focusing exclusively on legal documents in the Portuguese language and by adopting a reproducible systematic literature review protocol, revealing specific gaps such as the lack of Portuguese-language benchmarks and metrics sensitive to normative validity, which are absent in [Martinez-Gil 2023].

The article by [Jerónimo 2025] examines, from the perspective of Criminal Procedural Law and fundamental rights, the obstacles imposed by language barriers and deficiencies in the appointment of interpreters in Portugal. Although it does not directly address computational systems, the discussion on the quality of judicial translation and its impact on procedural fairness reinforces the importance highlighted in our review of incorporating legal expert evaluation and measuring the validity of answers.

In summary, while [Martinez-Gil 2023] provides a broad but not specific view of LQA and [Jerónimo 2025] emphasizes practical language challenges in legal settings, the

present work distinguishes itself by limiting its scope to the Lusophone corpus, applying the PRISMA protocol, quantifying reproducibility gaps, metrics, and code availability, and outlining a future agenda that integrates translation quality, normative temporality, and expert validation.

3. Methodology

This Systematic Literature Review (SLR) was conducted to ensure transparency, traceability, and replicability, adopting the PRISMA 2020 protocol [Page et al. 2021] as the methodological reference and using the *Parsif.al* platform to plan, register, and document each stage of the study. The six research questions (RQ1–RQ6) guided the formulation of the search strategies, for which we defined Portuguese and English terms covering Question Answering systems, Portuguese language variants, the legal domain (law, jurisprudence, contracts), and pipeline components (retrieval, embedding, RAG, graphs).

To ensure international and regional coverage, searches were conducted across nine indexing databases: ACM Digital Library, EI Compendex, ISI Web of Science, Periódico Capes, Scielo, ScienceDirect, Scopus, Sol SBC and Springer Link. The search strings were composed of logical blocks interlinking QA, Portuguese, and legal domain terms using AND/OR operators, respecting the 255-character limit where necessary; the final versions of the strings were recorded in *Parsif.al*.

The search was conducted in May 2025. Records were imported into the platform in BibTeX format and subsequently processed by the automatic deduplication module. Study selection involved a dual screening of titles and abstracts, followed by full-text reading. Two reviewers independently assessed each record according to predefined inclusion and exclusion criteria, peer-reviewed articles, written in legal Portuguese, and focused on QA systems. Any conflicts were resolved by a third reviewer.

4. Planning

This research focuses primarily on systems, models, or frameworks for Question Answering (QA) applied to legal documents written in Portuguese. The intervention considered includes methods, architectures, or pipelines developed or evaluated, encompassing approaches such as BM25, Dense Passage Retrieval (DPR), Retrieval-Augmented Generation (RAG), fine-tuned large language models (LLMs), knowledge graphs, and ontologies. The comparison between studies involves the types of legal data used, as well as the techniques adopted throughout the pipeline, covering preprocessing steps, semantic representation, retrieval strategies, re-ranking, and answer generation or selection. The outcomes analyzed include reported metrics and results, such as F1-score, Exact Match (EM), precision, coverage, legal validity, explainability, and response time. The scope of the study is limited to the Portuguese-speaking legal domain, including laws, decrees, case law, court summaries, contracts, legal opinions, and documents from Brazilian and Portuguese courts.

4.1. Research Questions

RQ1 What types of legal documents in Portuguese (laws, decrees, case law, contracts, legal opinions, etc.) have been used as datasets in Question Answering studies, and what are their characteristics in terms of size, annotation, and availability?

RQ2 Which retrieval and re-ranking strategies (lexical, dense, hybrid, LLM-based, or graph-based) show better performance in Portuguese legal QA scenarios, considering precision, coverage, and computational cost?

RQ3 What evaluation metrics and protocols are reported to measure answer quality in Portuguese legal QA, and what gaps or challenges remain in the development of benchmarks and the generalization across legal subdomains?

4.2. Search String

("question answering" OR "question-answer" OR "QA system*" OR "pergunta-resposta" OR "sistema de respostas" OR "sistema de pergunta-resposta" OR "resposta automática a perguntas" OR "QA jurídico" OR "legal QA")*
AND
(Portuguese OR "língua portuguesa" OR Português OR "português brasileiro" OR "Brazilian Portuguese" OR Brazil OR Brasil OR Brazilian OR brasileiro OR lusophone OR "pt-BR" OR "pt-PT" OR lusófona)*
AND
(law OR legal OR legislação OR "documento jurídico" OR lei OR jurisprudência OR jurisprudence OR "court decision*" OR "decisão judicial*" OR contrato* OR contratual)*
AND
(pipeline OR model OR framework OR architecture OR transformer OR embedding OR "knowledge graph" OR "open information extraction" OR "extração de informação aberta" OR "tripl" OR "retrieval-augmented generation" OR RAG OR "dense retrieval" OR rerank* OR "pré-processamento" OR lemmatização OR stemming)*

4.3. Selection Criteria

The inclusion criterion required that only peer-reviewed journal articles or full conference papers addressing Question Answering systems or pipelines developed or evaluated on legal texts in Portuguese were considered. The applied exclusion criteria were as follows: EC1, the study does not address Question Answering; EC2, the work is not a primary study; EC3, the study is not accessible; EC4, the study addresses Question Answering in non-legal domains; EC5, the work is a duplicate; EC6, the study belongs to the grey literature, such as dissertations, theses, preprints, technical reports, extended abstracts, posters, or book chapters; and EC7, the work represents an earlier version of another study already pre-selected.

5. Conducting

After applying the search string to the selected sources, candidate studies for inclusion were identified. Specifically, 2 studies were found in ACM Digital Library, 6 studies in EI Compendex, 4 studies in ISI Web of Science, none in Periódico Capes, none in SciELO, none in ScienceDirect, 7 studies in Scopus, 17 studies in Sol SBC, and 592 studies from Springer Link. During the Identification stage, 628 studies were retrieved from the databases. Subsequently, 32 duplicate records were manually removed. In the Screening stage, 596 studies remained, of which 69 records belonging to grey literature were excluded, resulting in 527 studies for full-text analysis. From these, 46 secondary studies were removed, leaving 481 studies for eligibility assessment. During the eligibility assessment, 471 studies were excluded for not addressing QA techniques in the context

of legal Portuguese documents. Finally, at the Inclusion stage, 10 studies were included in the systematic literature review. Figure 1 presents the process following the PRISMA 2020 diagram [Page et al. 2021].

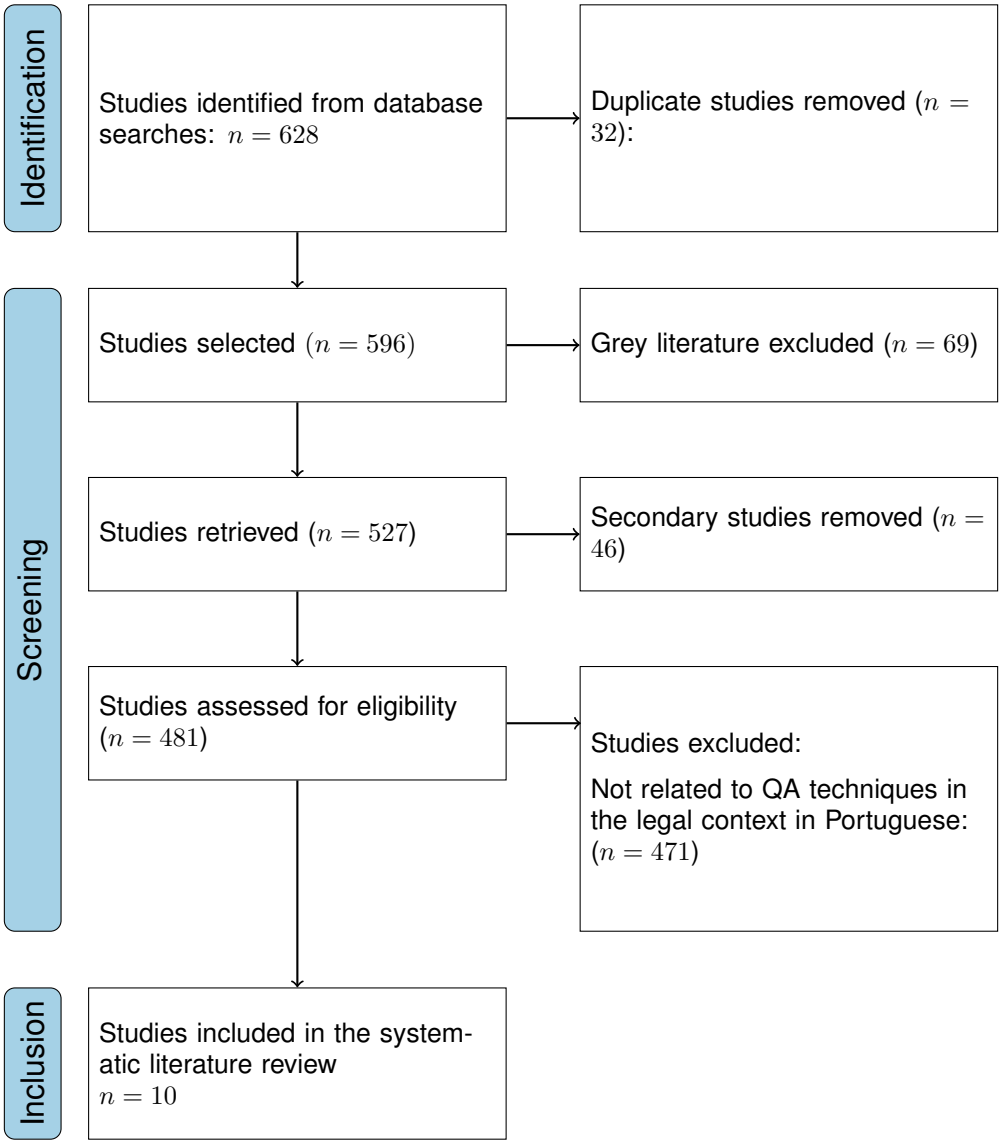

Figure 1. Study selection process according to PRISMA 2020 [Page et al. 2021]

6. Results

The extraction phase resulted in **10 primary studies**[1] published between 2012 and 2025 (Figure 2). The studies cover more than a decade, with a concentration in the 2023–2025 period. However, evaluation standardization remains limited: seven articles do not report metrics ([Ferneda et al. 2012, Barcellos et al. 2020, de Vargas Feijó and Moreira 2018,

[1]Refs.: [Barcellos et al. 2020, Viegas et al. 2023, Sakiyama et al. 2023, Nunes et al. 2025, de Vargas Feijó and Moreira 2018, Ferneda et al. 2012, Barros et al. 2023, Athaydes et al. 2024, Bertalan and Ruiz 2024, Costa et al. 2025]

Barros et al. 2023, Bertalan and Ruiz 2024, Athaydes et al. 2024, Costa et al. 2025]), revealing the need for common evaluation protocols.

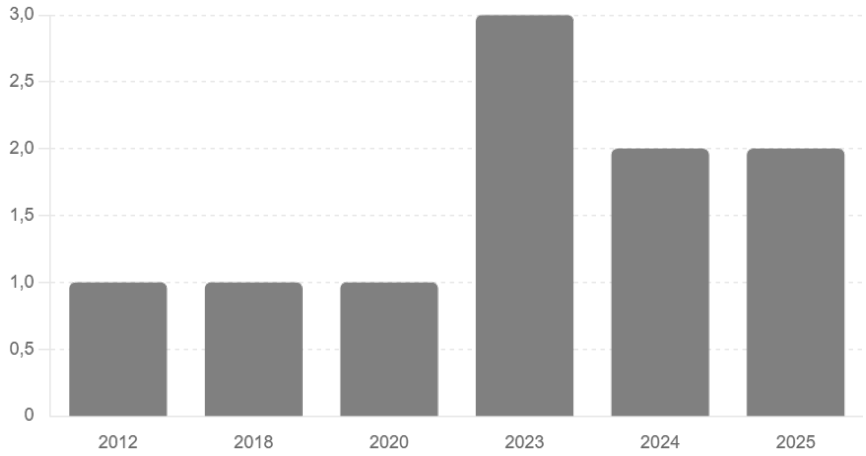

Figure 2. Distribution of Publication Years

Nine studies describe at least one reproducible step of cleaning or segmentation; only [Viegas et al. 2023] limits itself to general mentions without concrete steps. No article presents diagrams or pseudocode covering all stages (retrieval, reranking, and generation); descriptions generally stop at the main employed phase. Only [Bertalan and Ruiz 2024] includes a lexical baseline (BM25) for comparison, highlighting the lack of reference models in the field.

Only three studies report appropriate metrics (e.g., F1 and EM in [Sakiyama et al. 2023, Nunes et al. 2025, Viegas et al. 2023]); the others offer only qualitative examples or no measurable results. No study submitted answers for legal expert review or compared results to official databases, indicating a critical reliability gap. None of the papers explicitly present a limitations section, hindering the understanding of the real scope of applicability.

Nine articles claim to consider the temporal validity or revocation of legal norms ([Barcellos et al. 2020, Sakiyama et al. 2023, Nunes et al. 2025, Barros et al. 2023, Bertalan and Ruiz 2024, Athaydes et al. 2024, Costa et al. 2025, de Vargas Feijó and Moreira 2018, Viegas et al. 2023]); only [Ferneda et al. 2012] does not address this aspect. No study reports response time or resource consumption, preventing scalability analyses. Although all works deal with Portuguese corpora, none thoroughly discuss specific linguistic issues (such as regional variations or orthographic differences).

The fact that 70% of publications appear in conference proceedings suggests a dynamic and experimental ecosystem still in transition toward consolidation in journals ([Viegas et al. 2023, Sakiyama et al. 2023, Nunes et al. 2025]). Half of the studies employ multiple document types (e.g., administrative acts and contracts in [Barros et al. 2023]), reinforcing the importance of robust preprocessing, although described superficially. Transformer models dominate (4/10) ([Viegas et al. 2023, Sakiyama et al. 2023, Nunes et al. 2025, Costa et al. 2025]), coexisting with classical approaches (pure BM25) and emerging methods (GNN in [Bertalan and Ruiz 2024]; LLM

in [Athaydes et al. 2024]). This diversity suggests an exploratory phase without technological consensus.

Only [Sakiyama et al. 2023] provides code, but the link is currently inactive, and six studies publish datasets ([Barcellos et al. 2020, Viegas et al. 2023, Nunes et al. 2025, Barros et al. 2023, Bertalan and Ruiz 2024, Costa et al. 2025]). The lack of artifacts limits replication and practical adoption. Nine articles claim to address the temporal validity and revocation of legal norms ([Barcellos et al. 2020, Viegas et al. 2023, Sakiyama et al. 2023, Nunes et al. 2025, de Vargas Feijó and Moreira 2018, Barros et al. 2023, Bertalan and Ruiz 2024, Athaydes et al. 2024, Costa et al. 2025]), but without detailed methodology, making it difficult to assess effectiveness.

6.1. Answering RQ1

The ten studies analyzed rely, to varying degrees, on four major families of documents: digitized case law, consolidated legislation, contracts and petitions, and administrative opinions or acts. Table 1 summarizes the most frequently cited datasets, indicating their volume, annotation schema, and availability policy.

Table 1. Main Portuguese-language legal datasets used in QA tasks

Type	Dataset / Study	Volume	Annotation	Availability
Case Law	*RulingBR* [de Vargas Feijó and Moreira 2018]	10,623 rulings	Section (Summary, Vote)	GitHub (MIT)
Legislation	*CDC-QA* [Barcellos et al. 2020]	1,050 articles + 10,504 rulings	5,217 Q–A pairs	GitHub (CC-BY)
Legal Pretraining	Legal-BERT-PT Corpus [Viegas et al. 2023]	1.5M paragraphs	Raw text + metadata	GitHub
Contracts	*Clauses-PT* [Athaydes et al. 2024]	315 contracts	Clause–obligation	Restricted (NDA)
Petitions	*IP-Dataset* [Costa et al. 2025]	270 PDFs	Word-level tagging	Under agreement

Case law dominates the landscape, with six articles employing rulings from Brazil's superior Courts, driven by open publication policies that facilitate scraping and OCR. Laws and decrees appear in five studies; although smaller in volume, these texts are often structured in legislative XML or HTML, easing the extraction of articles and paragraphs. Datasets involving contracts or petitions are still incipient: only two works employ such genres, and for confidentiality reasons, these datasets are either partially available or shared under non-disclosure terms. Administrative opinions appear sporadically and lack standardized annotation, which hampers cross-study comparison.

For extractive evaluation, only three studies publish reference question–answer pairs. The others provide raw text, supporting excerpts, or BM25 indices, which limits the construction of standardized benchmarks. Most legislative and case law datasets are open (MIT or CC-BY licenses); in contrast, contractual and petition-related materials remain protected by NDAs, reflecting concerns about confidentiality and privacy.

Despite progress, three gaps stand out. First, there is a lack of multimodal datasets: no corpus combines text with images or tables appended to legal cases, a critical requirement for exploring multimodal LLMs capable of interpreting stamps, signatures, or cover pages. Second, there is still no bilingual pt-BR/pt-PT benchmark to assess knowledge transfer between Portuguese variants or evaluate multilingual models such as XLM-R. Lastly, there is a lack of versioned legal corpora, i.e., collections with yearly captures that support ex-tunc queries and reasoning about legal validity changes over time.

Recent studies point to the gradual integration of temporal metadata and citation graphs into legislative corpora, paving the way for answering queries about repeals or legal consolidations. Initial efforts to handle unstructured case law using OCR pipelines and vote zonation are also emerging, but segmented texts are not yet public, making replication unfeasible. In light of these findings, the recommended next steps are: to release annotated datasets that combines text, images, and metadata to support research on multimodal LLMs; to coordinate Lusophone initiatives aimed at building parallel Brazil–Portugal benchmarks for fine-tuning multilingual models; and to version legislative corpora, enabling QA evaluation in historical contexts and reducing the risk of legally outdated answers.

6.2. Answering RQ2

The selected studies employ five main strategies for retrieval and re-ranking: lexical methods based on BM25, dense retrievers using embeddings (DPR or SBERT), hybrid combinations that integrate lexical and vector-based ranking, re-rankers leveraging knowledge graphs and large language models, and finally, approaches that explore multilingual resources. Table 2 provides an overview of the retrieval and re-ranking strategies adopted by the analyzed studies.

Table 2. Retrieval and re-ranking strategies in legal QA

Study	Main Strategy
[Ferneda et al. 2012]	Lexical BM25
[de Vargas Feijó and Moreira 2018]	Lexical BM25
[Barcellos et al. 2020]	BM25 + fastText dense (RRF)
[Sakiyama et al. 2023]	DPR Legal BERT + BM25
[Nunes et al. 2025]	DPR BERTimbau + MonoT5
[Barros et al. 2023]	BM25 + GNN over RDF graph
[Bertalan and Ruiz 2024]	DPR + Graph features
[Costa et al. 2025]	SBERT + BM25 learning-to-rank

The results indicate that pure BM25 provides reasonable initial coverage but loses precision on questions requiring synonymy or semantic reasoning. Dense retrievers trained specifically in the legal domain, such as DPR Legal BERT in [Sakiyama et al. 2023], improve precision by approximately nine percentage points but require high-dimensional vector indexes and approximate search on GPU. Hybrid methods mitigate this cost by invoking the dense component only on a lexically ranked subset and maintaining latency around thirty milliseconds, which is suitable for low-concurrency interactive applications.

The use of legal-normative graphs introduces additional gains when legal coherence is the goal. [Barros et al. 2023] and [Bertalan and Ruiz 2024] show that centrality weights and revocation relationships help prioritize valid responses, yielding an average increase of eight percentage points in F1 compared to standalone BM25. However, these methods require preprocessing to build the graph and additional memory to traverse neighborhoods at query time.

Re-ranking models based on LLMs, such as the adapted MonoT5 in [Nunes et al. 2025], achieve the best coverage and precision, but their latency is four times higher than that of lightweight hybrids. Practical adoption depends on asynchronous orchestration or limiting simultaneous queries. No study has yet evaluated the incorporation

of multimodal LLMs in the re-ranking stage, which is an increasingly relevant gap as courts begin releasing documents in PDF format with graphical elements.

Experiments involving multilingual resources are limited. When present, they employ embeddings from models like XLM-R without fine-tuning to legal Portuguese. Thus, there is a lack of a systematic protocol for cross-testing pt-BR and pt-PT queries. The potential of cross-lingual transfer, which is common in English QA studies, to reduce annotation costs in Portuguese also remains unexplored.

6.3. Answering RQ3

Table 3 summarizes the frequency with which each evaluation metric and system-level measure is reported across the ten primary studies addressing Lusophone legal question answering.

Table 3. Coverage of evaluation metrics and system measures in the ten primary studies analyzed (RQ3).

Metric	N	Studies reporting it
F1 (token level)	4	[Barcellos et al. 2020]; [Sakiyama et al. 2023]; [Nunes et al. 2025]; [Viegas et al. 2023]
Exact Match	3	[Sakiyama et al. 2023]; [Nunes et al. 2025]; [Athaydes et al. 2024]
Precision@k / Recall@k	4	[Sakiyama et al. 2023]; [de Vargas Feijó and Moreira 2018]; [Ferneda et al. 2012]; [Costa et al. 2025]
Mean Reciprocal Rank	2	[Sakiyama et al. 2023]; [Bertalan and Ruiz 2024]
Legal Validity Score	1	[Bertalan and Ruiz 2024]
Average query latency	5	[de Vargas Feijó and Moreira 2018]; [Ferneda et al. 2012]; [Sakiyama et al. 2023]; [Barros et al. 2023]; [Nunes et al. 2025]
Index memory consumption	2	[Sakiyama et al. 2023]; [Costa et al. 2025]

The evaluation of answer quality in Lusophone legal QA remains heterogeneous. Among the ten studies analyzed, four report the F1 score calculated at the token level ([Barcellos et al. 2020], [Sakiyama et al. 2023], [Nunes et al. 2025], [Viegas et al. 2023]). Exact Match appears in three works, namely [Sakiyama et al. 2023], [Nunes et al. 2025], and [Athaydes et al. 2024]. Retrieval indicators such as Precision@k or Recall@k are presented in four studies: [Sakiyama et al. 2023], [de Vargas Feijó and Moreira 2018], [Ferneda et al. 2012], and [Costa et al. 2025]. Only two authors, [Sakiyama et al. 2023] and [Bertalan and Ruiz 2024], report Mean Reciprocal Rank. A single article, [Bertalan and Ruiz 2024], introduces the Legal Validity Score metric, which penalizes answers based on repealed legal provisions, although this proposal has not yet been replicated by the research community.

Half of the studies report average query latency in an online environment, specifically [de Vargas Feijó and Moreira 2018], [Ferneda et al. 2012], [Sakiyama et al. 2023], [Barros et al. 2023], and [Nunes et al. 2025]. Information on index memory consumption is even rarer, limited to [Sakiyama et al. 2023] and [Costa et al. 2025]. No author conducts formal evaluation by legal experts or proposes an explainability score to assess the clarity of the supporting evidence.

Regarding experimental protocols, four articles use the standard 80-10-10 split for training, validation, and testing, while [Costa et al. 2025] adopts five-fold cross-validation due to the small number of petitions. There are no generalization experiments across domains, such as training on legislation and testing on contracts, nor are there benchmarks available for comparison between Brazilian and European Portuguese. Attempts to incorporate multilingual models are limited to using embeddings such as XLM-R without legal-domain adaptation and without dedicated evaluation metrics.

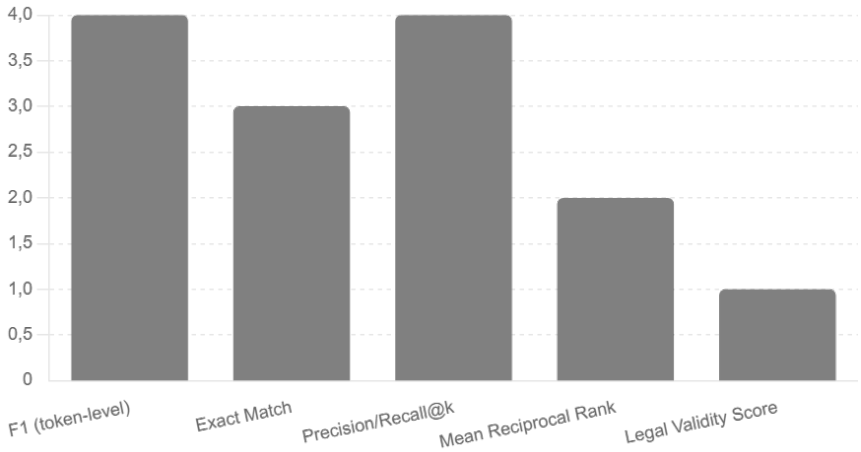

Figure 3. Evaluation Metrics Coverage

In light of these results, three gaps remain evident. First, the lack of metrics that combine textual accuracy, legal validity, and source transparency. The Legal Validity Score is a promising initiative, but it remains isolated. Second, the absence of standardized human validation, which prevents the assessment of practical utility and legal correctness in professional contexts. Third, the scarcity of public benchmarks covering multiple subdomains and linguistic variations, which is essential for evaluating the robustness of multilingual and multimodal systems. Addressing these challenges requires collaborative efforts between computer science researchers and legal experts, as well as protocols that emphasize both technical rigor and the normative relevance of the generated answers.

7. Conclusion

This systematic review analyzed the state of the art in Question Answering (QA) systems for legal documents in Portuguese, based on ten studies published between 2012 and 2025. The findings indicate a predominance of laws, decrees, and case law, with emerging initiatives in contracts and petitions. Most corpora are publicly available but lack standardized annotations.

In preprocessing, hierarchical segmentation and citation normalization are well-established practices, while the treatment of normative validity still varies considerably. Semantic representation has evolved from bag-of-words to contextual embeddings and the integration of knowledge graphs. Hybrid strategies (lexical + dense) have proven superior in balancing precision and performance. DPR and SBERT models outperformed BM25 in accuracy, but combined methods offer better latency control. Graph-based methods and LLM rerankers improve normative consistency, albeit with higher computational costs.

In answer generation, extractive span selection remains the most traceable technique; RAG provides more coherent answers but requires filtering to ensure normative validity. The combination of LLMs and graphs has shown promise in mitigating hallucinations. Regarding evaluation, F1 and EM metrics prevail, with only one study proposing a Legal Validity metric. Significant gaps remain, including the absence of standardized benchmarks, systematic human evaluations, and cross-domain legal comparisons. Despite the robust protocol, limitations include the possible exclusion of relevant prototypes (not peer-reviewed), difficulties in comparison due to heterogeneous reporting, and potential residual bias even with dual screening.

Future research directions include: developing benchmarks with legal validity labels across multiple subdomains and Portuguese variants (pt-BR, pt-PT); creating metrics that combine textual accuracy and normative validity; further integration of LLMs and graphs to reduce hallucinations and increase explainability; computational optimization via quantization, knowledge distillation, and hybrid indexes; user-centered evaluation with empirical studies involving legal professionals; and cross-domain testing to assess semantic and syntactic robustness across different document types.

8. Acknowledgment

This work has been funded by P&D CEMIG/ANEEL PD-04950-D0677/2023. It was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) [grant number408490/2024-1].

References

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (qa) models. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 493–500, Porto Alegre, RS, Brasil. SBC.

Barcellos, R., Bernardini, F., and Viterbo, J. (2020). A methodology for retrieving datasets from open government data portals using information retrieval and question and answering techniques. In Viale Pereira, G., Janssen, M., Lee, H., Lindgren, I., Rodríguez Bolívar, M. P., Scholl, H. J., and Zuiderwijk, A., editors, *Electronic Government*, pages 239–249, Cham. Springer International Publishing.

Barros, T. S., Pires, C. E. S., and Nascimento, D. C. (2023). Leveraging bert for extractive text summarization on federal police documents. *Knowledge and Information Systems*, 65(11):4873–4903.

Bertalan, V. G. F. and Ruiz, E. E. S. (2024). Using attention methods to predict judicial outcomes. *Artificial Intelligence and Law*, 32(1):87–115.

Costa, Y. D. R., Oliveira, H., Nogueira, V., Massa, L., Yang, X., Barbosa, A., Oliveira, K., and Vieira, T. (2025). Automating petition classification in brazil's legal system: a two-step deep learning approach. *Artificial Intelligence and Law*, 33(1):227–251.

de Vargas Feijó, D. and Moreira, V. P. (2018). Rulingbr: A summarization dataset for legal texts. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 255–264, Cham. Springer International Publishing.

Ferneda, E., do Prado, H. A., Batista, A. H., and Pinheiro, M. S. (2012). Extracting definitions from brazilian legal texts. In Murgante, B., Gervasi, O., Misra, S., Nedjah, N., Rocha, A. M. A. C., Taniar, D., and Apduhan, B. O., editors, *Computational Science and Its Applications – ICCSA 2012*, pages 631–646, Berlin, Heidelberg. Springer Berlin Heidelberg.

Filho, A. D. A. (2019). Do casamento às uniões sem selo: O alcance social e jurídico dos arranjos familiares no brasil e em portugal. *Revista Jurídica Portucalense*.

Jerónimo, P. (2025). Legal translation and the challenges of overcoming language barriers in court practice: Evidence from portuguese courts. *International Journal for the Semiotics of Law*. Advance online publication.

Martinez-Gil, J. (2023). A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Nunes, R. O., Santos, J., Spritzer, A., Balreira, D. G., Freitas, C. M. D. S., Olival, F., Cameron, H. F., and Vieira, R. (2025). Assessing european and brazilian portuguese llms for ner in specialised domains. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 215–230, Cham. Springer Nature Switzerland.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., and Moher, M. J. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, 18(3).

Perlingeiro, R. and Ghio, E. (2020). *Princípios Gerais da Cooperação Jurídica Internacional: uma abordagem temática e comparativa*. Núcleo de Pesquisa e Extensão sobre Ciências do Poder Judiciário (Nupej), Niterói, Brasil, 1 edition.

Ramos, M. (2017). Governo das sociedades e responsabilidade civil dos administradores: Algumas reflexões a partir da experiência jurídica portuguesa. *Social Science Research Network*.

Ransolin, M. and Baruffi, P. (2022). O direito ao esquecimento: A exclusão de notícias que ferem a integridade e intimidade da pessoa e a atual discussão do supremo tribunal federal. *Ponto de Vista Jurídico*.

Sakiyama, K., Montanari, R., Malaquias Junior, R., Nogueira, R., and Romero, R. A. F. (2023). Exploring text decoding methods for portuguese legal text generation. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 63–77, Cham. Springer Nature Switzerland.

Viegas, C. F. O., Costa, B. C., and Ishii, R. P. (2023). Jurisbert: A new approach that converts a classification corpus into an sts one. In Gervasi, O., Murgante, B., Taniar, D., Apduhan, B. O., Braga, A. C., Garau, C., and Stratigea, A., editors, *Computational Science and Its Applications – ICCSA 2023*, pages 349–365, Cham. Springer Nature Switzerland.