

Domain Adaptation of Foundation Language Models for Legal Document Generation: A Systematic Review and Implications for Analogous Regulated Institutional Contexts

Ricardo Rodrigues Barcelar 
Universidade Federal de Mato Grosso
Cuiabá/MT, Brasil
ricardo.barcelar@sou.ufmt.br

Thiago Meirelles Ventura 
Universidade Federal de Mato Grosso
Cuiabá/MT, Brasil
thiago@ic.ufmt.br

Abstract—The adaptation of foundation language models for legal document generation has become a relevant problem in Legal Natural Language Processing, particularly because generated texts must preserve factuality, normative coherence, traceability and control over sensitive information. This systematic literature review maps, classifies and critically analyzes primary studies on domain adaptation techniques for foundation language models applied to legal and normative-argumentative document generation. It also examines the implications of the resulting evidence for analogous regulated institutional contexts that share requirements related to reliability, privacy, auditability and institutional control. The review followed guidelines for systematic reviews in Software Engineering and the applicable items of PRISMA 2020. Searches were conducted in Scopus, Web of Science, IEEE Xplore and ACM Digital Library, covering studies published from January 2021 to December 2025 and complemented by backward snowballing. From 200 initial records, 32 primary studies were included. The corpus is predominantly concentrated in the legal domain, with only one study directly addressing document production in Public Security. The results show a recurrent use of hybrid adaptation strategies involving parameter-efficient fine-tuning, retrieval-augmented generation and structured inference, although the literature does not establish a universally superior combination. Data engineering is marked by a trade-off between scale and control, with frequent use of semi-synthetic datasets. Traditional Natural Language Processing metrics remain useful as initial indicators, but are insufficient to assess factuality, normative validity and argumentative consistency. Ethical safeguards, privacy controls and bias mitigation metrics remain underdeveloped. The findings indicate that the institutional adoption of foundation language models for legal document generation should be treated as a problem of architecture, evaluation and governance, rather than only textual performance. These conclusions are directly grounded in the Legal NLP literature, while their transfer to analogous regulated institutional contexts should be understood as a functionally motivated implication requiring domain-specific empirical validation.

Index Terms—Large Language Models, Foundation Models, Legal NLP, Domain Adaptation, Parameter Efficient Fine-Tuning, Retrieval Augmented Generation, Systematic Literature Review

I. INTRODUCTION

The recent development of Large Language Models (LLMs) based on the Transformer architecture [33] has expanded the capabilities of Natural Language Processing (NLP) systems in tasks such as text generation, summarization, machine translation, information retrieval and decision support, thereby accelerating their adoption in specialized domains.

General-purpose foundation models, such as the LLaMA [35] and Qwen [46] families, exhibit broad generalization capabilities. However, their direct application to specialized domains remains constrained by issues related to factuality, model behavior control, computational cost, privacy and terminological adaptation [18], [19], [20], [25].

These limitations become particularly sensitive in the legal domain, which is the primary empirical focus of this review. In this domain, automatic text generation requires adherence to legal norms, conceptual precision, argumentative coherence and traceability of the information used, going beyond grammatical correctness or narrative fluency. A formally well-written response that is factually incorrect or normatively inadequate may have relevant effects on decisions, administrative procedures and individual rights.

Functionally analogous regulated institutional contexts may also impose operational constraints that are less pronounced in generic NLP applications. In many scenarios, the use of proprietary models accessed through external APIs may conflict with requirements of confidentiality, sensitive data protection and technological sovereignty. At the same time, local deployment of language models requires computational infrastructure that is not always available, especially when large-scale models are considered. In this setting, adaptation strategies such as supervised fine-tuning, parameter efficient fine-tuning, retrieval augmented generation and structured inference play a central role in attempts to balance performance, cost and control.

Recent literature in Legal NLP has examined different uses of language models in tasks such as classification, information

extraction, question answering, summarization and support for normative analysis. However, a significant portion of existing reviews focuses on discriminative tasks or broad legal applications, without specifically addressing the adaptation of foundation models for the generation of legal and normative-argumentative documents.

Existing surveys have mapped legal language models from complementary perspectives. Lai et al. [47] provide a broad overview of legal tasks, applications and model limitations, Hou et al. [48] organize the field around model families, frameworks, datasets and benchmarks and Dehghani et al. [49] synthesize applications, resources and risks across legal systems. These studies characterize the broader Legal NLP ecosystem, but treat document generation as one application among several. In contrast, the present review focuses on legal and closely related normative-argumentative document generation and examines adaptation techniques, data engineering, computational infrastructure, factuality assessment and governance from an integrated analytical perspective on the reliability of generated documents.

Table I summarizes the main differences between the previous surveys and the present review.

TABLE I
COMPARISON WITH PREVIOUS REVIEWS ON LEGAL LANGUAGE MODELS

Study	Primary Scope	Difference from the Present Review
Lai et al. [47]	Legal tasks, applications, models and limitations	Document generation is addressed within a broad mapping of Legal NLP tasks and applications.
Hou et al. [48]	Legal model families, frameworks, datasets and benchmarks	The analysis is organized around models and evaluation resources rather than the reliability pipeline of generated documents.
Dehghani et al. [49]	Applications, datasets, benchmarks and risks of LLMs in legal systems	Legal document drafting is examined as one application among several areas of legal-system adoption.
Present review	Adaptation of foundation models for legal and normative-argumentative document generation	Integrates adaptation, data engineering, infrastructure, factuality assessment and governance within a unified analysis of factors affecting the reliability of generated documents.

As Table I indicates, the literature still lacks an integrated synthesis of how adaptation, data engineering, computational infrastructure, factuality assessment and governance jointly affect the reliability of generated legal documents.

This review establishes the legal domain as the central empirical context for examining the adaptation of foundation models for document generation. This empirical focus is justified by the nature of legal document production, which involves control requirements also relevant to other institutional domains, such as public security, public administration and oversight bodies. By investigating the legal domain, this review maps the state of the art in technical solutions for challenges related to factuality, auditability and normative compliance.

Accordingly, its findings provide a technical basis for informing research in functionally analogous regulated environments, provided that any transfer is empirically validated within each operational context.

Against this background, this Systematic Literature Review aims to map, classify and critically analyze primary studies on the adaptation of foundation language models for legal and closely related normative-argumentative document generation. It also examines the implications of this evidence for functionally analogous regulated institutional contexts. Unlike broader Legal NLP reviews, this SLR integrates adaptation techniques, data engineering, computational infrastructure, factuality assessment and governance requirements.

A. Research Questions

The review was guided by the PICOC strategy (Population, Intervention, Comparison, Outcome, Context), in which the population comprises language models based on the Transformer architecture and employed in text generation tasks. The intervention encompasses domain adaptation techniques, model specialization and inference optimization. The comparison involves different technical strategies, model scales, computational resources and evaluation approaches. The outcomes include textual quality, factuality, normative coherence, bias mitigation and computational efficiency. The context corresponds to the legal domain and to regulated institutional environments with similar requirements regarding reliability, privacy and control.

Based on this framework, the following research questions were formulated:

RQ1 – Adaptation Strategies: Which model adaptation techniques and inference strategies have been employed to optimize the generation of legal and normative-argumentative texts, considering technical precision, control over model behavior and feasibility in computationally constrained environments? **RQ2 – Data Engineering and Curation:** Which methodologies for data curation, dataset construction, anonymization, protection of sensitive data and data augmentation have been used in the development of specialized datasets? **RQ3 – Quality and Factuality Evaluation:** To what extent are traditional NLP metrics suitable for evaluating the generation of legal documents and which alternative approaches have been proposed to measure factuality, normative coherence, argumentative consistency and textual quality? **RQ4 – Ethical Alignment and Bias Mitigation:** Which technical mechanisms have been implemented to reduce bias, preserve privacy and incorporate ethical safeguards into text generation by language models in regulated domains?

II. METHODOLOGY

The review was conducted based on the guidelines proposed by Kitchenham and Charters [40] for systematic literature reviews in Software Engineering and reported according to the applicable items of PRISMA 2020 [42]. The methodological process was organized into three phases, as illustrated in Fig. 1: (i) review planning, involving the definition of the

protocol; (ii) review conduction, including the search, selection and assessment of studies; and (iii) reporting of results, including synthesis and discussion.

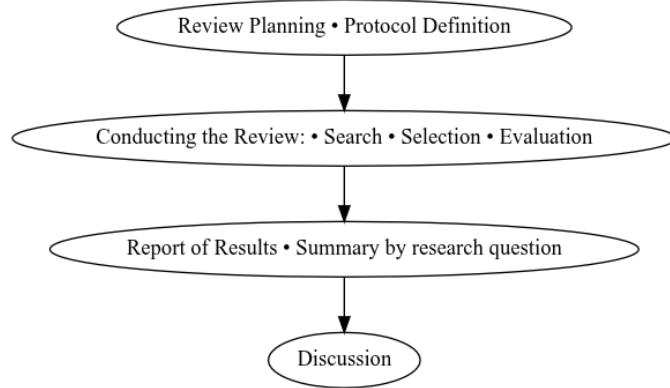


Fig. 1. Methodological process.

A. Review Protocol

The protocol of this review relied on the PICOC strategy to structure the research question and to guide the search, eligibility and extraction stages. To mitigate selection bias before document analysis, the protocol, eligibility criteria, quality assessment checklist and extraction fields were defined a priori and operationalized through standardized decision rules. Study-level screening decisions, exclusion reasons, quality scores and extracted data were recorded in supplementary matrices, providing an auditable trail of how the protocol was applied.

B. Search Strategies in Databases

The search strategy was designed to balance breadth and precision in the identification of relevant studies, following good-practice recommendations for systematic literature reviews in Software Engineering [40] and the PRISMA guidelines [42].

The automated search was conducted in January 2026 and covered studies published between January 2021 and December 2025. This time frame was defined because it corresponds to the consolidation period of foundation models based on Transformer architectures and to the intensification of their application in specialized domains. Four databases widely used in computer science were selected: Scopus, Web of Science, IEEE Xplore and ACM Digital Library. These databases provide broad coverage of peer-reviewed literature and are relevant to Artificial Intelligence and Natural Language Processing.

Guided by the PICOC framework (Population, Intervention, Comparison, Outcome, Context), the search string was structured into distinct conceptual blocks:

- **Population:** language models based on Transformer architectures, e.g., "Large Language Model", "Small Language Model", "Foundation Model", LLM and SLM;

- **Intervention:** model adaptation and specialization techniques, e.g., "Fine-tuning", "Instruction tuning", "Domain adaptation", PEFT, LoRA and "parameter efficient";
- **Comparison:** comparison among different adaptation strategies, model scales and evaluation approaches, considering trade-offs between technical precision, computational efficiency and behavioral control;
- **Outcome:** aspects related to text generation, quality and reliability, e.g., "Text generation", "Bias mitigation", "Fairness" and "Data curation";
- **Context:** regulated institutional domains, with a focus on Public Security and Criminal Justice, e.g., "Public safety", "Public security", "Law enforcement", "Police", "Criminal justice", "Legal domain" and "Legal NLP".

Based on these blocks, a generic search string was defined and subsequently adapted to the specific syntax of each database:

("Large Language Model" OR "Small Language Model" OR "Foundation Model" OR LLM OR SLM OR "Transformer model") AND ("Public safety" OR "Public security" OR "Law enforcement" OR Police OR "Criminal justice" OR "Legal domain" OR "Legal NLP") AND ("Fine-tuning" OR "Instruction tuning" OR "Domain adaptation" OR PEFT OR LoRA OR "parameter efficient" OR "Bias mitigation" OR Fairness OR "Data curation" OR "Text generation")

The adaptation of the string considered the specific characteristics of each database, including field restrictions and search operators. In ACM Digital Library, the search was restricted to the title and abstract fields to reduce noise. In Scopus, the TITLE-ABS-KEY field was used, whereas equivalent variations in title and abstract fields were applied in IEEE Xplore and Web of Science.

The terms associated with Public Security and Criminal Justice were included to capture institutionally regulated applications with functional requirements similar to those of the legal domain. After screening, a predominance of legal-domain studies was observed, which guided the interpretation of the findings as direct evidence for legal NLP and indirect evidence for analogous institutional contexts. The initial results totaled 200 records distributed across the selected databases. These records constituted the initial corpus submitted to the subsequent stages of deduplication, screening and eligibility assessment.

1) *Validation of the Search Strategy:* The robustness of the search strategy was assessed according to recommendations for systematic reviews [40].

An initial string composed of core terms related to language models and the legal domain, namely "Large Language Model" AND "Legal domain" AND "Fine-tuning", was used as the starting point. The analysis of the titles, abstracts and keywords of the retrieved studies allowed the identification of additional terms and relevant semantic variations, such as "Small Language Model", "SLM", "parameter efficient", "Data curation" and "Criminal justice", which were progressively incorporated into the PICOC blocks.

At each iteration, the string was rerun to verify its ability to retrieve additional relevant studies and was adjusted until the inclusion of new terms no longer produced significant gains in retrieval. This procedure made it possible to balance breadth and precision in the final composition of the string without compromising the thematic coverage of the investigated domain.

2) *Complementary Identification Through Backward Snowballing*: As a complementary strategy to the automated search, backward snowballing [45] was adopted. This technique examines the reference lists of eligible studies to identify relevant works not retrieved through database searches. Its use addresses an inherent limitation of automated searches in fields with heterogeneous terminology, where variations in nomenclature may compromise the retrieval of relevant studies.

Backward snowballing was applied to the studies selected after the eligibility stage of the automated search, with the inclusion and exclusion criteria being applied to the works identified in their reference lists.

The combination of automated search and snowballing contributed to increasing the breadth of the identification strategy and reducing the risk of omitting relevant studies, especially works that use alternative or non-standardized terminology.

C. Inclusion and Exclusion Criteria

The eligibility criteria were defined to select studies with technical contributions related to the adaptation, specialization, evaluation, or control of language models in text generation tasks within the legal domain or in regulated institutional contexts with similar requirements. The criteria were designed to preserve the focus of the review on generative applications, avoiding the inclusion of studies exclusively focused on discriminative tasks or generic NLP uses.

Studies were considered eligible when they met the following inclusion criteria (IC):

- **IC1**: Studies addressing techniques for the adaptation, tuning or specialization of LLMs or SLMs, including fine-tuning, instruction tuning, domain adaptation, PEFT, RAG, structured prompting, external knowledge integration, or equivalent strategies for controlling model behavior.
- **IC2**: Studies investigating text generation tasks, including the generation of legal documents, reports, descriptive, explanatory, argumentative, narrative texts, or other textual artifacts with normative or institutional structure.
- **IC3**: Studies applied to the legal domain or to functionally similar regulated institutional contexts, in which text generation involves requirements related to factuality, normative coherence, privacy, traceability, or bias control.
- **IC4**: Studies addressing the construction, curation, augmentation, anonymization, or protection of specialized datasets, as well as evaluation metrics, factuality, bias mitigation, privacy, or ethical safeguards in texts generated by language models.

- **IC5**: Studies presenting technical, experimental, or empirical contributions described with sufficient detail to allow analysis of the adaptation strategies, data used, evaluation procedures or methodological limitations.

Studies were excluded when they met at least one of the following exclusion criteria (EC):

- **EC1 – Discrepant Domain**: Studies whose central topic is generic NLP or application in domains without functional relation to the legal domain, public security, criminal justice, or other regulated institutional environments.
- **EC2 – Publication Type**: Opinion articles, essays, editorials, secondary reviews, posters, extended abstracts, or publications without original technical or empirical contribution.
- **EC3 – Non-generative Task**: Studies in which model customization is intended exclusively for discriminative or extractive tasks, such as classification, regression, named entity recognition, or purely extractive question answering.
- **EC4 – Absence of Adaptation or Technical Control**: Studies based solely on the direct use of general-purpose models, basic prompt engineering, simple summarization, purely extractive RAG, or generic benchmarks, without a substantive strategy for adaptation, specialization, evaluation, or control.
- **EC5 – Incompatible Methodology**: Preliminary, exploratory, or incomplete studies that do not provide sufficient description of the method, data, adaptation strategy, or evaluation procedure.

Eligibility was defined in a combined manner. A study was retained in the corpus when it presented a technical contribution related to the adaptation, specialization, evaluation, or control of language models for generative tasks in the legal domain or in a regulated institutional context, met the inclusion criteria applicable to its scope and did not fall under any exclusion criterion.

D. Study Selection Process

The study selection process followed the four stages recommended by PRISMA 2020 [42]: (i) identification, (ii) deduplication, (iii) screening and (iv) eligibility. The automated search was complemented by backward snowballing in the reference lists of the included studies, with the purpose of identifying relevant works not retrieved from the consulted databases.

The automated search returned 200 records distributed across the consulted databases. Deduplication was conducted in two subphases. First, Rayyan [43] was used for the automatic detection of duplicate records, resulting in the removal of 20 duplicates. Then, during the inspection of the remaining records, 31 additional duplicates were identified, mainly associated with variations in titles, metadata, or indexing across different databases. In total, 51 duplicate records were removed, leaving 149 unique studies for screening.

In the screening stage, the titles, abstracts and keywords of the 149 studies were analyzed. At this stage, 112 records

were excluded for not meeting the eligibility criteria. The exclusions were distributed as follows: 42 were related to exclusion criterion EC1, Discrepant Domain; 16 were related to exclusion criterion EC2, Publication Type; 36 were related to exclusion criterion EC3, Non-generative Task; and 18 were related to exclusion criterion EC4, Absence of Adaptation or Technical Control. After this stage, 37 studies were submitted to full-text reading.

In the eligibility stage, the 37 remaining articles were read in full and reassessed according to the inclusion and exclusion criteria. Fourteen studies were excluded: 2 articles under exclusion criterion EC1, Discrepant Domain; 5 under exclusion criterion EC3, Non-generative Task; 2 under exclusion criterion EC4, Absence of Adaptation or Technical Control; and 5 under exclusion criterion EC5, Incompatible Methodology. At the end of this phase, 23 studies retrieved through the automated search were considered eligible for inclusion in the review corpus.

To reduce the risk of omitting relevant studies, backward snowballing [45] was conducted from the reference lists of the 23 included studies. The cited works were initially examined by title and, when necessary, by abstract. This procedure identified 14 potentially relevant studies that were not present in the automated-search corpus. After applying the same eligibility criteria, 5 studies were excluded: 1 under exclusion criterion EC1, Discrepant Domain; 1 under exclusion criterion EC2, Publication Type; 2 under exclusion criterion EC3, Non-generative Task; and 1 under exclusion criterion EC5, Incompatible Methodology. As a result, 9 studies were incorporated into the final corpus.

The final review corpus comprised 32 primary studies, of which 23 were identified through the automated database search and 9 through backward snowballing.

The review protocol and supplementary materials, including search records, screening decisions, quality assessments and data extraction matrices, are publicly available at <https://doi.org/10.5281/zenodo.22865728>.

Fig. 2 presents the PRISMA 2020 flow diagram detailing the selection process.

E. Methodological Quality Assessment of the Studies

The methodological quality assessment was conducted after the full-text reading of the studies considered eligible. The objective of this stage was to examine the technical robustness, methodological transparency and analytical usefulness of the included studies for answering the research questions. For this purpose, a checklist adapted to the scope of this review was developed based on the recommendations of Kitchenham and Charters [40] for systematic literature reviews in Software Engineering.

The checklist comprised seven criteria, organized around dimensions considered central to the analysis of language models in generative tasks within specialized domains: description of the architecture and adaptation technique, data transparency, computational efficiency, adequacy of evaluation

metrics, verification of factuality or coherence, bias mitigation and ethics and reproducibility.

Each criterion was scored on a trichotomous scale: 0, when the information was absent or insufficient; 0.5, when the information was partially reported; and 1, when the information was described with sufficient detail to support the analysis conducted in the review. The maximum possible score for each study was 7 points. The quality assessment was not used as an automatic exclusion criterion, but as an instrument for characterizing the corpus and qualitatively contextualizing the robustness of the findings.

Table II presents the criteria used and their practical interpretation.

TABLE II
QUALITY ASSESSMENT CRITERIA

Code	Criterion	Practical interpretation
QA1	Architecture and adaptation description	Does the study describe the base model used and the adaptation, specialization, or control technique employed?
QA2	Data quality and transparency	Does the study report the source, size, domain, characteristics and preprocessing, curation or anonymization procedures of the data?
QA3	Computational efficiency	Does the study consider computational constraints, such as memory consumption, training or inference time, quantization, infrastructure used, or feasibility in resource-constrained environments?
QA4	Metrics suitable for text generation	Are the automatic, qualitative, or hybrid metrics suitable for evaluating text generation tasks, going beyond purely surface-level measures when necessary?
QA5	Factuality/coherence assessment	Does the study evaluate factuality, textual coherence, normative consistency, validity of sources, or alignment between facts and conclusions?
QA6	Bias mitigation and ethics	Does the study discuss or implement strategies for bias mitigation, privacy preservation, sensitive-content control, or ethical safeguards?
QA7	Reproducibility	Does the study describe parameters, experimental settings, procedures, data, code, prompts, or limitations in sufficient detail to allow reproduction or partial verification of the method?

F. Data Extraction and Included Studies

Data extraction aimed to collect technical, methodological and contextual information from the included studies in a standardized and traceable manner. For this purpose, a structured spreadsheet was used, with fields previously defined based on the research questions and methodological quality criteria.

The extracted data were organized into six analytical dimensions: (i) study identification, including reference, year and source database; (ii) model characteristics, including base model, size, adaptation technique and computational infrastructure; (iii) data engineering, including dataset type, generation method, source, scale and treatment of sensitive data; (iv) ethical aspects, including bias mitigation strategies and

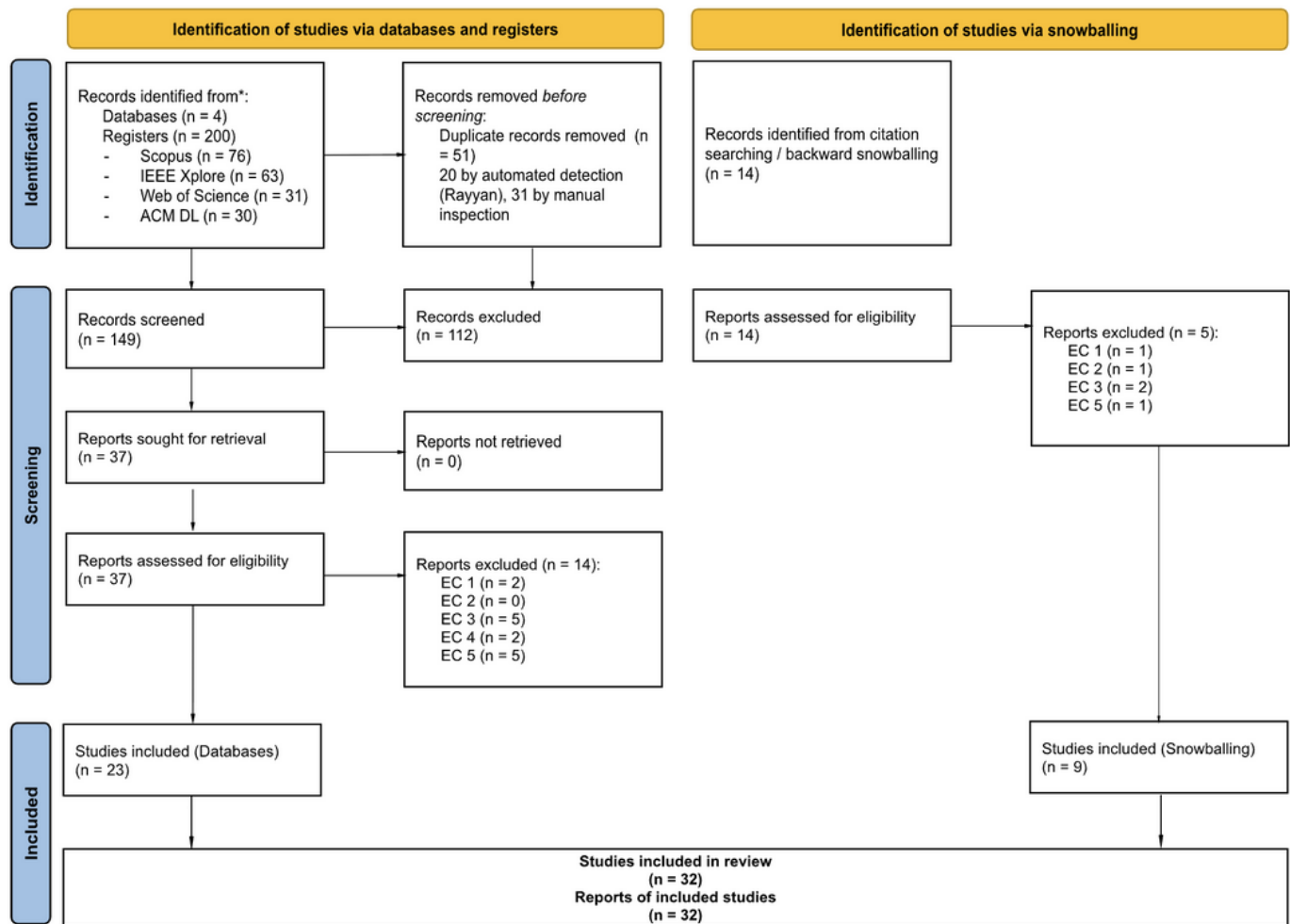


Fig. 2. PRISMA 2020 flow diagram of the study selection process.

associated metrics; (v) evaluation, including task, automatic metrics, human validation and reported performance; and (vi) synthesis, including methodological summary, limitations and main contribution.

Groups of extraction fields were associated with a research question to preserve traceability between the collected data and the synthesis presented in the results. Table III presents this relationship.

Of the 37 studies submitted to full-text reading in the automated search, 23 met the eligibility criteria. After backward snowballing, 9 additional studies were incorporated, resulting in a final corpus of 32 primary studies. These studies were then submitted to standardized data extraction and methodological quality assessment. Table IV lists the primary studies included in the review, with year, title, authors, identification method, quality score and associated research questions.

G. Methodological Limitations

The interpretation of the results of this review requires consideration of methodological limitations inherent to the research design. The selection, full-text reading, quality as-

TABLE III
TRACEABILITY BETWEEN RESEARCH QUESTIONS AND EXTRACTED DATA FIELDS

RQ	Analytical dimension	Extracted fields	Nature
RQ1	Adaptation Strategies	Model type; adaptation technique, including FT, PEFT, RAG and ICL; architectural modifications; computational infrastructure	Categorical
RQ2	Data Engineering	Dataset type; source; size; synthetic vs. real data; curation method; treatment of sensitive data	Mixed
RQ3	Evaluation and Metrics	Evaluation type; metrics, including BLEU, F1, factuality and related measures; human validation; benchmarks	Mixed
RQ4	Ethics and Bias Mitigation	Bias mitigation strategy; privacy treatment; ethical safeguards	Categorical

essment and data extraction stages were conducted by the first author, without independent duplicate assessment or a formal inter-rater reliability analysis. Throughout the review process, the eligibility criteria and methodological decisions

TABLE IV
PRIMARY STUDIES INCLUDED IN THE SYSTEMATIC REVIEW

ID	Year	Article title	Authors	Method	QA	RQ
[1]	2025	Alkafi-LLaMA3: fine-tuning LLMs for precise legal understanding in Palestine	Al-Qaesm et al.	DB extraction	4/7	RQ1, RQ2
[2]	2024	SaulLM-54B and SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain	Colombo et al.	DB extraction	5/7	RQ1, RQ2, RQ3
[3]	2025	Legal AI for All: Reducing Perplexity and Boosting Accuracy in Normative Texts With Fine-Tuned LLMs and RAG	García-Montero et al.	DB extraction	6/7	RQ1, RQ2, RQ3, RQ4
[4]	2024	Vietnamese Legal Question Answering: An Experimental Study	Ha et al.	DB extraction	5/7	RQ1, RQ2
[5]	2025	Synthetically generated text for supervised text analysis	Halterman	DB extraction	4.5/7	RQ2, RQ4
[6]	2024	InLegalLLaMA: Indian Legal Knowledge Enhanced Large Language Model	Ghosh et al.	DB extraction	5/7	RQ1, RQ2
[7]	2024	LAPIS: Language Model-Augmented Police Investigation System	Kim et al.	DB extraction	5.5/7	RQ1, RQ2, RQ4
[8]	2024	More Than Catastrophic Forgetting: Integrating General Capabilities for Domain-Specific LLMs	Liu et al.	DB extraction	5/7	RQ1, RQ2, RQ3, RQ4
[9]	2024	Enhancing Legal Expertise in Large Language Models through Composite Model Integration	Liu et al.	DB extraction	4/7	RQ1, RQ4
[10]	2025	Learning to Follow Domain-specific Instruction with Verifiable Rewards	Huang et al.	DB extraction	4.5/7	RQ1, RQ3
[11]	2025	LEGALMIND: A Fine-Tuned Gemma-2-Based Legal Assistant for Indian Judiciary	Pimpale et al.	DB extraction	4/7	RQ1, RQ2
[12]	2025	Generating Legal Arguments using LLM and Vector Database	Sinha and Sharma	DB extraction	4.5/7	RQ1
[13]	2025	Legal text summarization via judicial syllogism with large language models	Song et al.	DB extraction	4/7	RQ1, RQ3
[14]	2025	JuDGE: Benchmarking Judgment Document Generation for Chinese Legal System	Su et al.	DB extraction	4.5/7	RQ2, RQ3
[15]	2025	FairCauseSyn: Towards Causally Fair LLM-augmented Synthetic Data Generation	Nagesh et al.	DB extraction	3.5/7	RQ2, RQ4
[16]	2025	Fine-tuning a Large Language Model for the Indian Legal System	Ta et al.	DB extraction	5/7	RQ1, RQ2, RQ3
[17]	2024	Reasoning before Responding: Towards Legal Long-form QA with Interpretability	Ujwal et al.	DB extraction	4.5/7	RQ1, RQ3
[18]	2024	Adapting a Large Language Model to the Legal Domain: A Case Study in Italian	Valerio et al.	DB extraction	4.5/7	RQ1, RQ2, RQ3
[19]	2024	DeliLaw: A Chinese Legal Counselling System Based on a Large Language Model	Xie et al.	DB extraction	4/7	RQ1, RQ3
[20]	2025	Domain-Specific Legal Language Modeling through Knowledge Fusion and Structured Attention	Yu	DB extraction	4/7	RQ1, RQ3
[21]	2025	Optimizing Answer Generator in Vietnamese Legal Question Answering Systems Using Language Models	Le et al.	DB extraction	4/7	RQ1, RQ3
[22]	2024	LawLLM: Law Large Language Model for the US Legal System	Shu et al.	DB extraction	5.5/7	RQ1, RQ3
[23]	2024	Lawyer GPT: A Legal Large Language Model with Enhanced Domain Knowledge and Reasoning Capabilities	Yao et al.	DB extraction	5.5/7	RQ1, RQ2, RQ3, RQ4
[24]	2023	Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model	Cui et al.	Snowballing	4/7	RQ1, RQ2
[25]	2024	LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model	Zhou et al.	Snowballing	3.5/7	RQ1, RQ2
[26]	2024	Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models	Dahl et al.	Snowballing	5.5/7	RQ3
[27]	2023	Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction	Jiang and Yang	Snowballing	4/7	RQ1, RQ2, RQ3
[28]	2022	Legal Prompting: Teaching a Language Model to Think Like a Lawyer	Yu et al.	Snowballing	4/7	RQ1, RQ3
[29]	2023	ChatLaw: Open-source legal large language model with integrated external knowledge bases	Cui et al.	Snowballing	4.5/7	RQ1, RQ3
[30]	2023	Syllogistic Reasoning for Legal Judgment Analysis	Deng et al.	Snowballing	4/7	RQ1, RQ3
[31]	2021	Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents	Xiao et al.	Snowballing	3.5/7	RQ1
[32]	2022	LegalBench: Prototyping a Collaborative Benchmark for Legal Reasoning	Guha et al.	Snowballing	3.5/7	RQ1, RQ3

were jointly discussed and defined by both authors and studies requiring additional interpretation of the eligibility criteria were discussed jointly. This collaborative process contributed to the operational interpretation of the criteria and to the consistency of methodological decisions throughout the review.

The use of a single primary reviewer may introduce some degree of reviewer-dependent judgment during study eligibility assessment and data extraction, which may influence the composition of the final corpus and the synthesis of the evidence. To mitigate this limitation, the review used predefined eligibility criteria, a standardized quality checklist, a structured extraction instrument and documented selection and exclusion decisions throughout the review process. The complete selection records, quality assessments and extraction matrices are available in the Zenodo repository described in Section II-D, supporting transparency, consistency and auditability. These procedures strengthen the reliability of the review process, while recognizing that they do not substitute independent duplicate assessment.

Regarding retrieval scope, the automated strategy prioritized indexed databases with peer-reviewed publications, which favors the reliability of the captured evidence but may have excluded preprints, technical reports, government documents and other forms of gray literature from direct retrieval. Although this is relevant in a rapidly evolving field such as LLMs, this risk was partially mitigated by complementing the search with backward snowballing in the reference lists of the included studies, allowing the identification of relevant works that had not been directly retrieved from the databases.

In addition, the literature on model adaptation in the legal domain presents marked terminological heterogeneity. Studies on the adaptation of language models to the legal domain use different terms for related techniques, such as fine-tuning, instruction tuning, domain adaptation, PEFT, RAG, structured prompting and external knowledge integration. To mitigate this issue, the search strategy combined broad and specific terms, organized according to the PICOC strategy and was complemented by backward snowballing.

At the analytical level, the corpus presents relevant variations in tasks, architectures, model sizes, adaptation techniques, datasets, languages and evaluation metrics. Moreover, not all studies report information with the same level of detail regarding data composition, prompt configuration, computational infrastructure, anonymization strategies, baselines, dispersion measures and validation procedures. This variation prevented the conduction of a meta-analysis and justified the use of narrative synthesis of the results [38], [39].

The temporal scope constitutes another boundary condition. The main search covered studies published between January 2021 and December 2025. Since the literature on LLMs evolves rapidly, studies published after this period may not have been covered. This delimitation, however, was adopted to maintain methodological consistency across the search, screening, extraction and synthesis of results.

Finally, regarding external validity, the consolidated corpus is concentrated in the legal domain. Consequently, the con-

clusions of this review are directly grounded in the Legal NLP literature, whereas their relevance to other regulated institutional contexts should be interpreted through functional analogy and requires domain-specific empirical validation.

III. RESULTS

This section presents the results of the analysis of the 32 primary studies included in the review. The exposition is organized according to the four research questions. First, the corpus is characterized in terms of application domain, temporal distribution, source databases, models used, geographical scope and methodological quality. Then, the results are presented regarding adaptation strategies, data engineering and curation, quality and factuality evaluation and mechanisms for ethical alignment and bias mitigation.

The organization of the results distinguishes two levels of analysis. The first describes the composition and methodological robustness of the corpus, while the second presents the technical patterns identified in the literature, with emphasis on trade-offs among domain adaptation, computational cost, factuality control and governance.

A. General Characterization of the Included Studies

The final corpus of the review comprises 32 primary studies. Most works are concentrated in the legal domain, covering tasks such as legal question answering, legal text summarization, argument generation, judicial document generation, model adaptation to national legal systems and legal reasoning evaluation. Only one study explicitly addresses document production in the context of public security.

This discrepancy indicates that the literature on the adaptation of foundation models for document generation remains strongly concentrated in the legal domain. Public security appears as an emerging and still underexplored application field in primary studies, although it shares relevant technical requirements with the legal domain, such as factuality, normative grounding, control of sensitive information, traceability and argumentative coherence. For this reason, the findings of this review are interpreted as directly supported by the Legal NLP literature, while their relevance to other regulated institutional contexts is considered through functional analogy rather than direct empirical evidence.

This asymmetry in the corpus also guides the interpretation of the results. The conclusions regarding adaptation techniques, data engineering, evaluation metrics and bias mitigation predominantly reflect evidence from legal-domain studies. The implications for other institutional environments should therefore be understood as inferences based on the functional compatibility between document types and control requirements, rather than as consolidated empirical evidence in those specific domains.

1) *Methodological Quality of the Included Studies:* The assessment of the 32 included studies indicates a corpus with moderate methodological robustness. The mean score was 4.45 points, with a median of 4.50 out of a maximum possible score of 7.0. The distribution was concentrated in the intermediate

range: 11 studies scored between 4.0 and 4.4; 7 studies scored between 4.5 and 4.9; and 10 studies reached scores equal to or higher than 5.0. Only 4 studies were located in the range between 3.5 and 3.9.

The highest-scoring studies stand out for combining adequate architecture description, methodological transparency, data detailing and evaluation in tasks closer to real-world usage scenarios. In contrast, the lowest-scoring studies tend to present gaps in the description of computational infrastructure, factuality assessment, bias mitigation or discussion of limitations.

The criterion-level analysis shows a relevant difference between the technical maturity of the studies and the maturity of evaluative and ethical dimensions. Criteria QA1, QA2 and QA7 presented higher mean values, indicating that most studies adequately describe models, data and experimental procedures. By contrast, criteria QA3 and QA6 presented the lowest values, evidencing limited attention to computational efficiency and ethical safeguards. This result is particularly relevant when considering analogous regulated institutional contexts, where infrastructure constraints, privacy, auditability and bias mitigation may constitute important operational requirements. Table V presents the distribution of scores by quality criterion.

TABLE V
DISTRIBUTION OF SCORES BY QUALITY CRITERION

Criterion	Mean	Score = 0	Score = 0.5	Score = 1	Interpretation
QA1	0.98	0 (0%)	1 (3%)	31 (97%)	Almost all studies adequately describe the model and the technique employed.
QA2	0.84	0 (0%)	10 (31%)	22 (69%)	Most studies report data transparently; gaps are concentrated in dataset size or pre-processing.
QA3	0.31	21 (66%)	2 (6%)	9 (28%)	Two thirds of the studies do not report computational constraints, representing a reproducibility gap.
QA4	0.72	2 (6%)	14 (44%)	16 (50%)	Half of the studies employ suitable metrics; the other half use partially surface-level metrics.
QA5	0.52	8 (25%)	15 (47%)	9 (28%)	One quarter of the studies do not assess factuality; the largest group addresses it only partially.
QA6	0.22	21 (66%)	8 (25%)	3 (9%)	Sixty-six percent of the studies do not address bias mitigation or ethics, representing the most critical gap in the corpus.
QA7	0.86	1 (3%)	7 (22%)	24 (75%)	Most studies provide sufficient elements for partial verification of the method, although full reproduction remains limited by the absence of data, code, infrastructure or parameters in part of the corpus.

Overall, the quality assessment suggests that recent litera-

ture presents reasonable maturity in the description of architectures, adaptation techniques and experimental procedures, but remains limited regarding the discussion of computational cost, factuality, privacy, bias mitigation and governance. This methodological asymmetry reappears in the four research questions analyzed in the following subsections.

2) *Temporal Distribution and Publication Databases*: The temporal distribution of the studies confirms the recent nature of the literature on the adaptation of foundation models for generative tasks in the legal domain. The studies identified through automated search are mainly concentrated in 2024 and 2025, a period in which there is a greater presence of works on LLM fine-tuning, the use of open models, integration with RAG and the construction of benchmarks for legal tasks. This concentration is consistent with the recent expansion of adaptable foundation models and the increasing availability of parameter efficient fine-tuning techniques.

The studies identified through backward snowballing cover the period from 2021 to 2024. This distribution suggests that the complementary procedure retrieved works of a more foundational nature, especially those related to pre-trained models for long legal documents, legal prompting, syllogism-based reasoning and legal reasoning benchmarks. Although these studies do not always directly address complete document generation, they help contextualize the evolution of adaptation and evaluation strategies later incorporated by generative models.

Table VI presents the distribution of studies by year and source database. Scopus, IEEE Xplore and ACM Digital Library concentrated the eligible studies retrieved through the automated search. Although Web of Science returned records in the identification stage, no study indexed exclusively in this database met the eligibility criteria without falling under an exclusion criterion. This result indicates that, for the adopted scope, the effective contribution of Web of Science occurred mainly through overlap with records also found in other databases.

Overall, the temporal distribution and source of the studies indicate that the analyzed literature is in a consolidation stage. More recent works tend to emphasize the adaptation of foundation models and hybrid strategies, whereas older studies contribute to the methodological basis of legal representation, structured prompting and legal reasoning evaluation.

TABLE VI
DISTRIBUTION OF ARTICLES BY YEAR AND SOURCE DATABASE

Source	2021	2022	2023	2024	2025	Total	% (n=32)
Scopus	0	0	0	5	5	10	31.3%
IEEE Xplore	0	0	0	1	5	6	18.8%
ACM DL	0	0	0	5	2	7	21.9%
Snowballing	1	2	4	2	0	9	28.1%

Although Web of Science contributed 31 records in the identification stage, no study indexed exclusively in this database met the inclusion criteria without falling under an exclusion criterion.

3) *Overview of Language Models:* The included studies employ both Large Language Models and smaller-scale models, generally adapted to specific tasks or jurisdictions. The most recurrent families are LLaMA, Qwen, Mistral, Gemma, BERT-derived models, GPT versions and specialized or less frequent architectures, such as ChatGLM, Baichuan and Longformer. This diversity indicates that the literature does not converge on a single model family, but rather on adaptation strategies that can be applied to different architectures and scales.

Since several studies evaluate more than one architecture, the counts by model family are not mutually exclusive. The LLaMA family appears most frequently in the corpus, being used in 12 studies [1], [3], [6], [7], [10], [16], [17], [18], [20], [21], [26], [29]. Models from the Qwen family appear in 5 studies [8], [9], [13], [14], [23], while Mistral appears in 2 studies [2], [3] and Gemma is employed in 3 studies [3], [11], [22]. BERT-derived models appear in 4 studies [4], [12], [19], [29], mainly in tasks related to the representation or processing of legal documents. GPT versions are used in 7 studies [5], [17], [26], [27], [28], [30], [32], often as a reference model, evaluator, data generator or comparison component. Other models, such as ChatGLM, Baichuan and Longformer, appear in 11 studies [3], [4], [7], [8], [14], [19], [21], [24], [26], [30], [31], reflecting the presence of solutions adapted to specific linguistic and legal contexts.

The recurrence of open or tunable models suggests a preference for architectures that allow greater control over the adaptation process. In the analyzed applications, this choice appears to be associated not only with performance, but also with the possibility of local execution, pipeline inspection, tuning on specialized data and reduced dependence on external services. This aspect is especially relevant when the data involve court cases, official documents, sensitive information or records that cannot be freely transferred to external services.

Although GPT versions appear in part of the studies, the analyzed literature points to limitations associated with dependence on proprietary models accessed through APIs. These limitations include lower control over data handling, dependence on external infrastructure, difficulty in auditing the model and potential privacy risks. For this reason, several studies justify the use of open, smaller or locally executable models as an alternative more compatible with institutional, budgetary and regulatory constraints.

In summary, the characterization of the models shows that the field is organized less around a dominant architecture and more around a combination of adaptability, computational availability and operational control. This trend anticipates one of the central patterns observed in RQ1: the preference for hybrid pipelines that combine open models, parameter efficient fine-tuning, information retrieval and structured inference.

4) *Application Context and Geographical Scope:* The included studies cover different national legal contexts and legal traditions, with tasks ranging from assisted document generation to the benchmark construction for legal reasoning. These applications share the need for adaptation to specific ter-

minology, language, normative structure, argumentative style and legal sources, which reinforces the dependence between model performance and application context.

From a geographical perspective, there is a higher concentration of studies in Asia, with 20 works distributed across China [8], [9], [10], [13], [14], [19], [23], [24], [25], [27], [29], [30], [31], India [6], [11], [12], [16], South Korea [7] and Vietnam [4], [21]. Europe appears with 3 studies, involving Italy [18], France [2] and the United Kingdom [17]; the Americas comprise 7 studies, related to Ecuador [3] and the United States [5], [15], [20], [22], [26], [32]; and the Middle East appears with 2 studies, associated with Palestine [1] and the United Arab Emirates [28]. This distribution suggests that the adaptation of foundation models to the legal domain is conditioned less by the architecture in isolation than by the ability to integrate the model into the linguistic, normative and documentary context of each legal system, including its own evaluation criteria.

B. Results Related to RQ1: Adaptation Strategies

This section analyzes the adaptation techniques and inference strategies employed to adjust foundation models to legal and closely related normative-argumentative generation tasks. The analysis considers not only the frequency of each technique, but also its function in the adaptation process, its relationship with computational constraints and its role in controlling factuality, coherence and normative adherence.

Overall, the studies do not indicate the predominance of a single technique. The most recurrent pattern is the combination of complementary approaches.

1) *Adaptation Techniques Employed:* Given the terminological heterogeneity observed in the primary studies, the techniques reported in Table VII were grouped into six functional categories: model parameter tuning; parameter efficient fine-tuning; feedback-based alignment; adaptation without weight updates; external knowledge integration; and structural modifications or compression. Since several studies combine more than one technique, the categories are not mutually exclusive.

Adaptation strategies without weight updates constitute the most frequent category, appearing in 15 studies, corresponding to 46.9% of the corpus [5], [7], [12], [13], [14], [15], [17], [19], [22], [23], [26], [27], [28], [30], [32]. This group includes approaches such as zero-shot, few-shot, prompt engineering and Chain-of-Thought. The recurrence of these techniques occurs in contexts involving cloud-based models or APIs, but also suggests their usefulness in scenarios in which training cost or infrastructure unavailability limits direct model adaptation. However, their effectiveness depends on prompt quality, task structure and the ability of the base model to follow complex instructions.

Model parameter tuning appears in 13 studies, representing 40.6% of the corpus [2], [4], [5], [9], [10], [11], [12], [13], [14], [17], [21], [23], [31]. This category includes supervised fine-tuning and continued pre-training, mainly employed to approximate the model to the terminology, style and argumentative structures of the legal domain. The same proportion was

TABLE VII
DISTRIBUTION OF MODEL ADAPTATION TECHNIQUES

Adaptation category	Included techniques	Studies (n)	% (n=32)
Parameter tuning, complete or partial fine-tuning	Supervised Fine-Tuning, Continued Pre-training	13	40.6%
parameter efficient Fine-Tuning (PEFT)	LoRA, QLoRA, ALoRA, Prompt Tuning	13	40.6%
Feedback-based alignment	RLHF (Reinforcement Learning from Human Feedback), DPO (Direct Preference Optimization), VDPO (Verified Direct Preference Optimization)	4	12.5%
Adaptation without weight updates, Prompting/ICL	Zero-shot, Few-shot, Prompt Engineering, Chain-of-Thought	15	46.9%
External knowledge integration	RAG, vector databases, knowledge integration	11	34.4%
Structural modifications and compression	Pruning, knowledge distillation, AMSI (Adaptive Multi-Stage Knowledge Integration)	2	6.3%

observed for parameter efficient fine-tuning techniques, also present in 13 studies, equivalent to 40.6% of the corpus [1], [3], [6], [7], [8], [16], [17], [18], [19], [21], [22], [25], [29]. Methods such as LoRA [36], QLoRA [37], ALoRA and prompt tuning indicate an attempt to reduce memory and processing costs without giving up domain adaptation, an aspect relevant to institutional environments with limited infrastructure.

External knowledge integration appears in 11 studies, corresponding to 34.4% of the corpus [3], [7], [11], [12], [14], [16], [19], [21], [22], [23], [24]. This category includes RAG [41], vector databases and knowledge integration mechanisms, used to increase response traceability, update information and contribute to hallucination reduction. Less frequently, feedback-based alignment strategies, such as RLHF, DPO [44] and VDPO, were identified in 4 studies, corresponding to 12.5% of the corpus [2], [9], [10], [23], while structural modification or compression techniques appeared in 2 studies, equivalent to 6.3% [16], [20]. This pattern shows that the literature has prioritized more accessible adaptation and knowledge retrieval solutions, whereas alignment and compression approaches still appear only occasionally.

2) *Model Size Distribution*: The analysis of model size considered the 25 studies that explicitly reported the number of parameters. In some cases, the same study evaluated models of different scales, which is why the size ranges are not mutually exclusive. Fig. 3 presents the distribution of studies by parameter range.

The results indicate a predominant concentration in intermediate-scale models. Models with 7 to 14 billion parameters appear in 20 studies, corresponding to 62.5% of the total corpus and 80% of the studies that reported model size [3], [6], [7], [8], [9], [10], [11], [12], [13], [14], [16], [17], [18], [21], [22], [23], [25], [26], [29], [32]. This predominance suggests

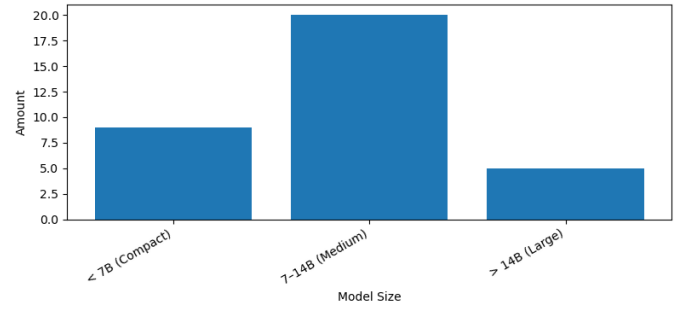


Fig. 3. Distribution of studies by model size range.

a preference for architectures that balance representational capacity and computational cost, especially in tasks that require normative interpretation, text generation, legal summarization and integration with information retrieval mechanisms.

Compact models, with fewer than 7 billion parameters, account for 28.1% of the corpus [1], [4], [8], [14], [17], [19], [21], [30], [32]. These models tend to appear in scenarios with greater computational constraints, often combined with techniques such as PEFT, structured prompting or RAG to compensate for capacity limitations. By contrast, models with more than 14 billion parameters appear in 5 studies, corresponding to 15.6% of the total corpus [2], [9], [10], [30], [32], generally associated with higher training, inference and infrastructure costs.

In the context of RQ1, the data indicate that model size is a decision associated with the adaptation strategy and the execution environment. The analyzed literature does not indicate a linear relationship between larger scale and greater adequacy to the legal domain. Instead, the studies suggest a preference for intermediate-scale models combined with adaptation techniques, knowledge retrieval and structured inference as a way to balance performance, cost and control.

3) *Computational Infrastructure*: The analysis of computational infrastructure considered the 32 included studies, of which 24, corresponding to 75% of the corpus, explicitly reported the resources used. Fig. 4 presents the relationship between adaptation techniques and the identified hardware categories. The absence of information in 8 studies indicates a reproducibility gap, as it limits the assessment of the actual computational cost of the proposed strategies.

High-performance environments based on HPC clusters were used in 7 studies, equivalent to 21.9% of the corpus [2], [9], [13], [14], [24], [25], [31]. These environments appear mainly in larger-scale experiments, such as continued pre-training, full fine-tuning and feedback-based alignment strategies. In these cases, the potential gain in model specialization is accompanied by higher training costs, dependence on distributed infrastructure and lower feasibility for institutions with restricted computational resources.

Intermediate infrastructures, composed of servers with 1 to 4 GPUs, also appear in 7 studies, corresponding to 21.9% [3], [4], [7], [8], [16], [22], [29]. This profile is frequently associated with the use of PEFT, RAG and intermediate-scale

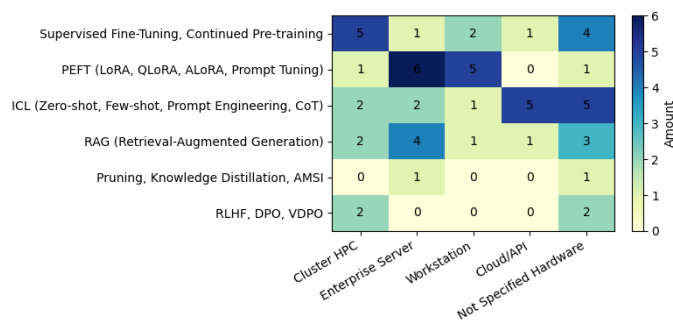


Fig. 4. Relationship between adaptation techniques and computational infrastructure categories.

models, indicating an attempt to balance performance, cost and the possibility of local deployment. The recurrence of LoRA and QLoRA in this group reinforces the suitability of parameter efficient techniques for scenarios in which full training is operationally infeasible.

Workstations with RTX GPUs were used in 5 studies, equivalent to 15.6% of the corpus [1], [6], [17], [18], [21]. In these cases, lightweight approaches predominate, such as PEFT, structured prompting and smaller-scale adaptations. Public cloud or API-based solutions appear in 5 studies, also corresponding to 15.6% [12], [27], [28], [30], [32], generally associated with in-context learning, the use of proprietary models or integration with previously trained models. From an architectural perspective, public cloud or API-based solutions reduce the need for in-house infrastructure, but introduce dependence on external providers and require specific analysis of privacy, auditability and data handling.

In the context of RQ1, the reported computational environment was associated with the adaptation strategies employed. Broad parameter tuning was more frequently reported in HPC environments, whereas PEFT, RAG and structured prompting were associated with intermediate servers or workstations. In the analyzed legal applications, this pattern links adaptation strategy to model scale, the reported computational environment, computational cost and requirements for institutional control.

4) *Structured Inference*: The inference strategies identified in the studies show that the adaptation of LLMs to the legal domain is not limited to tuning model parameters. A relevant part of the literature employs structured inference mechanisms to guide text generation, organize reasoning and approximate the response to argumentative formats used in legal contexts. These approaches include structured prompting, Chain-of-Thought, legal syllogism and argumentative models such as IRAC (Issue, Rule, Application, Conclusion).

In the analyzed corpus, strategies based on legal syllogism appear in 3 studies [13], [27], [30], while formulations based on the IRAC model appear in 2 studies [28], [32]. Chain-of-Thought-based methods were employed in 5 studies [13], [17], [23], [29], [32]. These techniques may improve argumentative organization and facilitate human verification of the response,

although they do not, by themselves, guarantee that the explicit steps faithfully correspond to the model's internal decision process.

The analyzed studies report gains in metrics such as BLEU, ROUGE, F1-score, precision and recall in specific tasks [13], [17], [27], [28]. However, these results should be interpreted with caution. The explicit presentation of intermediate steps does not, by itself, guarantee factuality, normative correctness or legal validity of the response. A model may produce a formally coherent argumentative chain, but based on incorrect premises, non-existent rules or inadequate interpretations. Thus, structured inference strategies should be understood as mechanisms for organizing and partially controlling generation, and not as substitutes for factual verification, legal validation or expert evaluation.

In the context of RQ1, structured inference approaches appear as a complement to adaptation and knowledge retrieval strategies. Their main value lies in guiding the form of the response and making the relationship between facts, rules and conclusions more explicit. When combined with RAG, PEFT or specialized data curation, these strategies may contribute to more controlled pipelines. Even so, their effectiveness depends on the quality of the base model, the prompt structure, the available sources and the evaluation criteria adopted.

5) *Synthesis of RQ1*: The results of RQ1 indicate that the adaptation of foundation models for text generation in the legal domain occurs predominantly through technical combinations and not through the isolated application of a single strategy. The corpus points to three recurrent arrangements: supervised or parameter efficient fine-tuning associated with terminological specialization; external knowledge integration, especially through RAG, for factual reinforcement and traceability; and structured inference, through Chain-of-Thought, legal syllogism or argumentative models, to organize the relationship between facts, rules and conclusions.

This hybrid configuration reflects the practical trade-offs of the domain. Parameter tuning contributes to approximating the model to legal language and legal structures, while RAG expands the possibility of updating and linking responses to external sources and structured inference offers greater control over the form of the response. Together, these strategies seek to balance technical precision while reducing computational cost and maintaining normative adherence and auditability. The predominance of models in the range of 7 to 14 billion parameters, combined with PEFT, RAG and structured prompting, reinforces this attempt to reconcile performance and deployment feasibility in resource-constrained environments.

Despite this pattern, the analyzed literature does not yet support the conclusion that one architecture or combination is generally superior. Results vary according to task, language, infrastructure, data quality and evaluation criteria. Thus, the main evidence from RQ1 is a recurrent technical direction toward hybrid pipelines and adaptable models, rather than the definitive validation of a dominant solution for legal document generation.

C. Results Related to RQ2: Data Engineering and Curation

The results related to RQ2 presented in this section explore the intersection between data engineering and privacy in the construction of specialized training datasets. The analysis considers both the construction of datasets and the reported practices of anonymization, privacy and quality control.

Overall, the studies indicate that data engineering for adapting LLMs to the legal domain is marked by a trade-off between scale and control. The observed distribution suggests that the scarcity of annotated legal data and access restrictions may favor semi-synthetic and synthetic strategies, although not all studies explicitly state this motivation. On the other hand, tasks that require greater normative fidelity still depend on manual curation, human validation or careful use of real legal databases.

1) *Dataset Typology*: The form of dataset construction, data sources and scale of the datasets used were analyzed jointly to understand how engineering and curation practices influence the adaptation of LLMs in legal applications. For this analysis, each study was classified according to the predominant dataset construction strategy, so that the categories in Table VIII are mutually exclusive.

The distribution of datasets, as shown in Table VIII, demonstrates that the adaptation of LLMs to the legal domain has relied mainly on strategies that seek to increase scale without completely abandoning control mechanisms. The predominance of semi-synthetic approaches in 14 studies [2], [4], [7], [8], [9], [10], [13], [17], [21], [24], [25], [29], [30], [32] suggests that automated generation by LLMs has been used as a practical response to the scarcity of annotated legal data. In contrast, studies using synthetic data without systematic validation [1], [3], [5], [11], [15], [22], [23], [26] indicate greater emphasis on scale and availability, whereas manual curation of real data [6], [14], [16], [18], [19], [20], [31] represents a more controlled but less scalable strategy. The use of existing datasets without modification [12], [27], [28] appears residually, reinforcing the perception that ready-made datasets rarely fully meet the terminological, normative and documentary requirements of generative legal tasks.

Regarding sources, the concentration on generic legal texts [2], [6], [7], [8], [9], [11], [13], [14], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32] indicates that the literature prioritizes broad corpora capable of capturing language, argumentative structure and documentary patterns of the legal domain. The lower presence of codified legislation [1], [3], [4], [12], [16] suggests that, although formal rules are relevant to factuality and normative coherence, they appear to be used more selectively, possibly because they require greater alignment between rule, context and application. The use of generic non-legal texts [5], [10], [15] points to strategies aimed at preserving general capabilities or constructing synthetic data, but with lower direct adherence to the domain.

Regarding dataset scale, a heterogeneous distribution can be observed. The results presented in Table IX allow the inference that a larger data volume is not, in isolation, a sufficient

TABLE VIII
RELATIONSHIP BETWEEN DATASET CONSTRUCTION FORM AND DATA SOURCE

Construction form	Generic legal texts	Codified legislation	Generic texts	Studies (n)	% (n=32)
Semi-synthetic, with validation	12	1	1	14	43.8%
Synthetic, without validation	4	2	2	8	25.0%
Manual curation, real data	6	1	0	7	21.9%
Existing dataset, without modification	2	1	0	3	9.4%

condition for better domain adaptation. Studies with moderate-sized datasets, when subjected to adequate curation, may achieve competitive performance, especially when the data preserve terminological consistency and normative adherence, as discussed by Valerio et al. [18]. Thus, the analyzed literature suggests that data quality, representativeness and control play a role as relevant as scale, especially in legal tasks that require factual precision and argumentative coherence.

TABLE IX
DISTRIBUTION OF STUDIES BY DATASET SIZE RANGE

Record range	Studies (n)	% (n=32)
Up to 10,000	11	34.4%
10,001 to 100,000	6	18.8%
100,001 to 1,000,000	7	21.9%
More than 1,000,000	5	15.6%
Not reported	3	9.4%

2) *Treatment of Sensitive Data and Privacy*: The analysis of the studies shows that the treatment of sensitive data is still poorly detailed in the literature on the adaptation of LLMs to the legal domain. Only a reduced portion of the corpus reports explicit practices for data protection, anonymization, control of sensitive attributes or prevention of information leakage [8], [14], [15], [23], [24], [27]. This result indicates a relevant gap, since legal and institutional documents often involve personal data, confidential information, procedural records and narratives with potential impact on individual rights.

The reported strategies can be grouped into three main approaches: anonymization of real data, generation of synthetic or semi-synthetic data and control mechanisms during model operation. Anonymization appears in studies that use procedural documents or real legal cases, in which personal identifiers are removed before training or evaluation [14], [27]. However, these studies do not always detail the technical de-identification procedures, the tools employed or the criteria used.

Synthetic data generation appears as an alternative to reduce direct exposure of sensitive information, especially when real data are scarce, confidential or difficult to access [15]. Even so, this strategy does not automatically eliminate privacy or bias

risks, since it depends on the sources used and the restrictions imposed on the generation process. In some cases, synthetic data may reproduce problematic patterns present in the original data, affecting normative fidelity.

Some studies propose control mechanisms in system operation, such as editing agents, filters, firewall strategies, safe dialogues or human-feedback alignment to prevent disclosure of sensitive information and inappropriate responses [8], [23], [24]. These approaches expand operational control in a targeted manner. In addition, few studies evaluate the effectiveness of these mechanisms in adversarial scenarios, prompt injection situations, contextual leakage or input manipulation.

In the context of RQ2, the results indicate that the literature recognizes the relevance of privacy, but still presents low methodological maturity in the treatment of sensitive data. This gap is particularly relevant to legal applications involving sensitive or confidential information, where data protection constitutes not only a technical requirement, but also an important condition for responsible institutional use.

3) *Synthesis of RQ2*: The analysis of the studies indicates that dataset construction for adapting LLMs in legal applications is marked by the combination of strategies that seek to balance scale, quality and restrictions on data access. There is a predominance of semi-synthetic generation methods, frequently associated with generic legal texts, which points to an attempt to overcome limitations in the availability of real data without completely abandoning quality control mechanisms [2], [4], [7], [8], [9], [10], [13], [17], [21], [24], [25], [29], [30], [32]. In contrast, manual curation remains more restricted to scenarios that require greater normative fidelity, while synthetic data without systematic validation are used mainly to increase volume.

The results also show that data volume is not determinative of adaptation quality. Studies with moderate-sized datasets may present competitive performance when the data are well selected, consistent and adherent to the domain [18]. This finding suggests that the selection, structuring and quality of examples play a relevant role in model adaptation, especially in legal tasks that require terminological precision, argumentative coherence and preservation of the normative context.

Another relevant finding is the low maturity of practices for treating sensitive data. Only a reduced portion of the corpus reports explicit protection mechanisms, concentrated on anonymization of real data, synthetic data generation and model control or alignment strategies [8], [14], [15], [23], [24], [27]. Even in these cases, the procedures are often described in a limited manner, without sufficient detail on de-identification techniques, residual risk assessment, privacy validation or resistance to contextual leakage.

The results indicate that data engineering in the analyzed scope applications is guided by the need to reconcile dataset representativeness with methodological rigor in data curation. While semi-synthetic strategies can expand dataset scale, manual curation remains relevant in higher-criticality tasks. The limited reporting of privacy and data governance practices

also reveals an important methodological gap, particularly for applications involving sensitive information.

D. Results Related to RQ3: Quality and Factuality Evaluation

This section details the findings of RQ3, which investigates the adequacy of traditional NLP metrics for evaluating legal text generation and the alternative approaches proposed.

Overall, the studies indicate that traditional metrics remain widely used, although they are insufficient to evaluate central dimensions of legal generation, since in normative-argumentative tasks text quality does not depend only on fluency, lexical similarity or classification accuracy. It also involves adherence to facts, validity of normative sources, coherence between premises and conclusion, argumentative consistency and adequacy to the legal context.

1) *The Role and Limits of Traditional NLP Metrics*: Table X summarizes the categories of metrics identified in the included studies. There is recurrent use of traditional classification metrics combined to a lesser extent with semantic metrics, qualitative assessments, specific benchmarks and approaches focused on factuality. This distribution shows that the evaluation of generative models in the legal domain still depends heavily on metrics originally designed for more structured tasks or for surface-level textual comparison, although part of the literature already recognizes the need for criteria closer to the requirements of the domain.

TABLE X
CATEGORIES OF EVALUATION METRICS

Category	Included metrics	Studies (n)	% (n=32)
Traditional lexical overlap metrics, n-grams	BLEU, ROUGE-1, ROUGE-2, ROUGE-L	9	28.1%
Traditional classification/accuracy metrics	Accuracy, F1-score, recall, exact match	21	65.6%
Semantic metrics, advanced	BERTScore, METEOR	5	15.6%
Specialized metrics: factuality, knowledge and probabilistic	Hallucination rate, penalty accuracy, perplexity, MMLU	7	21.9%
Qualitative/expert assessment	Human evaluation, LLM-as-a-judge, ELO rating	15	46.9%

The literature reports the use of an ecosystem of traditional NLP metrics, ranging from exact lexical matching, such as BLEU and ROUGE, to the capture of semantic equivalence and probabilistic fluency, such as BERTScore, METEOR and perplexity [3], [11], [13], [18], [20]. However, the usefulness of these automatic indicators is restricted to the structural validation of text generation and summarization. By design, these metrics primarily capture lexical overlap, fluency or semantic similarity, without verifying whether the output adheres to facts, rules or legal logic.

This limitation is particularly sensitive in legal tasks, in which a text may present adequate linguistic form and still

contain material errors. A model may produce a fluent response, with technical vocabulary and a plausible argumentative structure, but cite a non-existent rule, distort facts, omit relevant premises or support a conclusion incompatible with the case. Studies on factuality and legal hallucinations show that traditional metrics may fail to capture this type of failure, since they do not directly assess the existence of cited sources, the correctness of premises or the validity of normative reasoning [3], [26].

In addition, evaluations based on lexical matching are limited because equivalent legal arguments may be formulated in different ways without sharing the same textual surface. A legally adequate reasoning may differ from the reference in vocabulary, argumentative order or writing style, without this representing an error. This limitation is pointed out in studies discussing judgment generation, legal responses and normative justifications, in which quality depends less on term repetition and more on coherence between facts, applicable rules and conclusion [14], [21].

For these reasons, part of the studies resort to qualitative assessments, human validation, specific benchmarks, LLM-as-a-judge and metrics focused on factuality or penalization of inconsistencies [1], [3], [5], [8], [13], [16], [17], [19], [21], [23], [24], [26], [29], [30], [32]. These approaches seek to complement traditional metrics with criteria more adherent to legal generation, such as factual consistency, normative coherence and response completeness. Even so, the lack of standardization among evaluation protocols hinders direct comparison across studies and reveals a relevant methodological gap in the field.

2) *Multidimensional and Domain-Specific Criteria in the Legal Domain*: The studies show a transition from general evaluation metrics to alternative methods designed specifically to address the limitations discussed in the previous subsection, with a focus on qualitative dimensions and criteria specific to the legal domain. These approaches seek to capture aspects that generic metrics do not adequately evaluate, such as the validity of normative sources and coherence between facts and conclusion.

To capture the complexity of the studied domain, the literature proposes evaluations that approximate the analysis performed by legal experts. The JuDGE benchmark, for example, moves away from generic metrics and introduces a framework focused on penalty accuracy and correct classification of accusations [14]. Other studies propose the direct measurement of legal hallucination through the “Not-found” metric, which tracks the proportion of responses entirely fabricated by the model and unsupported by the provided context [22].

The use of expert-guided evaluations is also observed, focusing on dimensions such as depth of reasoning, applicability of legal principles and understanding of the implicit intent in user queries [17], [23]. Taken together, these approaches represent progress compared with the exclusive use of generic metrics, but still appear heterogeneously, without a consolidated protocol for evaluating legal generation.

3) *Evaluation Automation, LLM-as-a-Judge, Logical Consistency and Calibration*: Given the subjectivity and time cost of human evaluation by experts, some studies resort to the LLM-as-a-judge approach, in which more robust models, such as ChatGPT and Gemini, are used as automated evaluators or pseudo-references to score attributes such as logical coherence, reliability, conciseness and accuracy of responses generated by smaller models [21], [23], [26]. This strategy increases the scale of evaluation and reduces part of the operational cost, but also introduces a relevant dependence on the judging model. In particular, the quality of the judgment depends on the criteria defined in the evaluation prompt, the stability of the evaluator’s responses and its own biases, which limits its use as a direct substitute for expert evaluation.

Another identified approach uses the ELO ranking system to compare the performance of different LLMs in legal tasks, such as multiple-choice judicial examinations [29]. This strategy shifts evaluation from an absolute measure of quality to a relative ordering among models, allowing paired comparisons and dynamic updating of observed performance. Its use is suitable for comparative scenarios, but does not, by itself, resolve the evaluation of substantive dimensions of legal generation, such as factuality, normative validity, completeness of reasoning and argumentative quality.

In the context of hallucination detection and consistency verification, methods based on Natural Language Inference (NLI) have been employed to assess whether the final response generated by the model is logically compatible with the premises and facts of the case [17]. In scenarios without an available gold standard, self-contradiction-based strategies propose generating multiple responses to the same question under non-deterministic conditions, based on the premise that factual responses tend to remain more stable, whereas hallucinatory responses show greater variation [26]. These approaches are relevant because they shift part of the evaluation from textual surface to internal coherence and factual stability of generation.

Finally, confidence calibration appears as a complementary evaluation dimension. Expected Calibration Error (ECE) measures the discrepancy between the confidence declared by the model and the observed accuracy, penalizing incorrect responses accompanied by a high degree of certainty [26]. In the legal domain, this point deserves attention because an incorrect response, when presented assertively, may lead the user to trust non-existent, incomplete or inadequate normative grounds. Thus, confidence evaluation does not replace factual or legal verification, but contributes to measuring the risk associated with the institutional use of generative models.

4) *Synthesis of RQ3*: RQ3 produced results showing that traditional NLP metrics, although useful as initial performance indicators, are insufficient to meet the factuality requirements of the legal domain. It became evident that traditional metrics may assign good scores to responses that still contain generated text incompatible with the case. In addition, dependence on exact term matching ignores that valid legal arguments may be expressed through different linguistic constructions.

To address these limitations, the analyzed literature has proposed multidimensional and domain-specific evaluation approaches aimed at measuring concrete legal competencies. These approaches incorporate criteria related to the validity of normative sources, coherence between facts, completeness of reasoning and adequacy in interpreting user intent. Expert-guided evaluations also become relevant in this scenario, as they allow the examination of qualitative dimensions that automatic metrics hardly capture in isolation.

In the context of evaluation automation, the use of LLM-as-a-judge appears as a scalable alternative to human evaluation, especially when there is a need to compare multiple responses or models. In parallel, approaches based on NLI, self-contradiction across multiple generations and confidence calibration expand the evaluative focus beyond the textual surface, examining logical consistency, factual stability and the degree of certainty associated with responses. Taken together, these advances indicate a gradual change in the way legal generative models are evaluated, with greater attention to factuality, normative validity and the risks of plausible but incorrect responses.

E. Results Related to RQ4: Ethical Alignment and Bias Mitigation

This section presents the results related to RQ4, which investigates which technical mechanisms have been implemented as ethical safeguards to mitigate biases in text generation. The analysis considers strategies applied at different stages of the model lifecycle, from data curation to behavioral alignment, and examines the extent to which these strategies are integrated into the pipeline or implemented as isolated control measures.

The main finding is the low density of the literature on bias mitigation. Although factuality, privacy and security are frequently mentioned as requirements for the use of LLMs in legal applications, few studies implement specific mitigation mechanisms. When these mechanisms are reported, they generally appear as isolated interventions rather than as components of an integrated governance approach.

1) *Technical Strategies and Ethical Safeguards for Bias Mitigation*: The studies that actively address bias mitigation indicate that safeguards can be inserted at different stages of the LLM lifecycle [3], [5], [7], [8], [9], [15], [22], [23]. The strategies listed in Table XI are not directly comparable in granularity or purpose, but represent intervention points in the model lifecycle, from data composition to behavioral alignment.

In data curation, human-in-the-loop approaches appear as an attempt to contain the propagation of biases before they are incorporated into the model. The literature warns that texts and synthetic data produced by LLMs may reproduce or amplify biases from the original data, requiring specialized supervision [5], [7]. The Legal AI for All system adopts diversified example selection and expert validation to reduce implicit biases in generated data [3]. In LAPIS, human supervision also fulfills an accountability function, as it reviews ambiguous

TABLE XI
MAPPING OF BIAS MITIGATION STRATEGIES BY DEVELOPMENT STAGE

Lifecycle stage	Technical strategy and ethical safeguard	Studies (id)
Data curation	Human-in-the-loop, specialized supervision and validation	[3], [7]
Data curation	Statistical balancing of the training dataset	[22]
Data generation	Causal fairness constraints and safe prompts	[5], [15]
Behavior and alignment	RLHF, PPO and training with adversarial or contaminated dialogues	[8], [9], [23]

points before the final decision in a system designed to support investigative work [7]. In the first case, mitigation acts mainly on dataset quality, whereas in the second, it approaches a mechanism of institutional control over system use.

The second front shifts ethical control to the data generation process itself. In FairCauseSyn, mitigation does not appear as a posterior filter, but as a constraint incorporated into the synthetic generation process through Structural Causal Models and iterative prompt refinement [15]. This design can be considered methodologically stronger than simple posterior review, since it attempts to interfere with the formation of generated distributions. However, by imposing statistical constraints on sensitive attributes, the method improves the measurable fairness of the dataset without, by itself, resolving latent biases associated with context, language or source selection.

Regarding model behavior, studies resort to alignment techniques as well as adversarial training. Lawyer GPT, for example, uses RLHF with PPO to penalize the generation of sensitive information and align outputs with predefined ethical norms [23]. Other studies use safe dialogues and contaminated data, such as malicious requests or false legal premises, to teach the model to refuse inappropriate behaviors and identify fallacies [8]. These strategies operate closer to user interaction and therefore respond to risks that data curation cannot fully anticipate, since their effectiveness depends on the coverage of adversarial scenarios, the formulation of rewards and the subsequent evaluation of refusals.

Statistical balancing of the training dataset also appears as an additional measure associated with reducing asymmetries before model learning [22]. With limited effectiveness, a statistically balanced dataset may reduce certain discriminatory patterns without necessarily guaranteeing model neutrality.

Taken together, the studies suggest that bias mitigation depends less on a single technique and more on the articulation among curation, controlled generation, behavioral alignment and continuous evaluation.

2) *Ethical Safety and Algorithmic Bias Metrics*: The measurement of algorithmic bias and ethical safety still appears in a limited way in the analyzed literature. Among the 32 studies included in this review, only 4 methodological approaches reported specific metrics to quantify fairness, harmlessness

or ethical vulnerability of the models. This result indicates that bias mitigation tends to be mentioned as a desirable requirement, but is not always incorporated as a measurable dimension of evaluation.

The identified metrics evaluate different dimensions of LLM behavior. Some focus on response safety, such as refusal of malicious requests and the ability to legally justify such refusal. Others seek to measure output harmlessness, the presence of inductive biases or the causal influence of sensitive attributes on data generation. Table XII consolidates these strategies. Although they indicate an initial advance in the evaluation of ethical risks, these metrics still appear in a fragmented manner, without a common taxonomy that allows direct comparison of safety, fairness and reliability across studies.

TABLE XII
APPROACHES AND METRICS FOR QUANTIFYING BIAS AND ETHICAL SAFETY

Methodological approach	Specific metric	Metric objective	Studies (id)
Automated evaluation, LLM-as-a-Judge	Safety Score and ScoreE	Quantify refusal of malicious intentions and verify the legal grounding of the refusal.	[8]
Causal decomposition	Total Variation (TV), Direct Effect (DE), Indirect Effect (IE), Spurious Effect (SE)	Audit structural algorithmic fairness and mathematically isolate the impact of sensitive attributes.	[15]
Automated evaluation, LLM-as-a-Judge	Harmlessness	Measure the model's ability to avoid responses that cause harm, offense or legal violations.	[23]
Counterfactual stress testing and inductive bias analysis	Prominence bias and counterfactual bias, sycophancy	Measure the model's vulnerability to agreeing with false legal premises introduced by the user, sycophancy and quantify the overrepresentation of famous judges or cases, which generates "legal monoculture".	[26]

Safety and harmlessness metrics indicate that the ethical evaluation of legal LLMs still focuses mainly on the observable behavior of the response, especially on the ability to refuse improper requests. In Yao et al. [23], the Harmlessness metric evaluates whether the model avoids harmful, offensive or legally improper responses, while the study on catastrophic forgetting in legal LLMs employs Safety Score to measure refusal in the face of malicious intentions, such as fraud or criminal evasion [8]. The inclusion of ScoreE adds a relevant dimension to the legal domain, as it verifies whether the refusal is accompanied by adequate legal grounding. This difference suggests that, in legal applications, safety is not limited to blocking the response. The refusal must be technically adequate and legally justifiable, otherwise the system may

be safe from an operational perspective, but not sufficiently transparent or institutionally robust.

Approaches focused on algorithmic fairness and legal biases broaden this picture by shifting evaluation to less visible causes of model behavior. FairCauseSyn uses causal decomposition metrics to isolate the impact of sensitive attributes [15]. Dahl et al. [26] add another layer by showing that bias may also arise from the way the model handles authority, precedents and false premises. Sycophancy reveals vulnerability to user influence, while prominence bias indicates a tendency to concentrate on better-known cases. Thus, bias mitigation in legal LLMs cannot be reduced to the control of sensitive attributes. It also requires evaluating whether the model preserves jurisprudential diversity, resists legally false premises and avoids converting pre-training patterns into apparent normative criteria.

3) *Synthesis of RQ4*: The analysis of the results of RQ4 indicates that bias mitigation and ethical alignment still occupy a secondary position in the literature on LLM adaptation for legal generation. Most studies focus on performance, domain adaptation and textual quality. Safeguards appear less frequently and, when present, rarely structure the complete design of the system.

This asymmetry deserves attention because ethical risks do not arise only in the final output of the model. They may be present in source selection, dataset composition, synthetic data generation, preferences incorporated during alignment and the criteria used to evaluate responses. The identified strategies operate at different points in this trajectory, such as human curation, causal constraints, statistical balancing, RLHF, adversarial training, safe dialogues and safety metrics. In this sense, these layers appear poorly articulated with each other. In many cases, mitigation is treated as a localized solution to a specific risk and not as a continuous governance process.

The scarcity of metrics reinforces this diagnosis. Harmlessness, Safety Score, ScoreE, causal measures and inductive bias tests represent advances, but they evaluate distinct and still poorly comparable dimensions of the problem.

Finally, the results show that terminological adaptation and textual performance are not sufficient to characterize the suitability of a model for institutional use. This suitability depends on the integration of data governance, alignment, evaluation, monitoring and human accountability.

IV. DISCUSSION

The results of this review indicate that the adaptation of foundation models for legal generation increasingly depends on the coordination of multiple technical mechanisms rather than on model architecture alone. However, advances in adaptation are not accompanied by equally mature approaches to factual evaluation, privacy, bias mitigation and governance.

The main issue identified is the asymmetry between adaptation capacity and validation capacity. The studies show that there are already relatively consolidated strategies for adapting models to legal vocabularies, styles and tasks, especially with

open intermediate-scale models. The same does not occur with factuality evaluation, bias measurement, protection of sensitive data and validation in operational scenarios. Thus, there is faster progress in the construction of legally fluent systems than in the demonstration that these systems are reliable, auditable and safe for institutional use.

This asymmetry is particularly relevant to legal document generation. In the analyzed applications, response quality cannot be assessed solely through linguistic adequacy or performance on general benchmarks. Generated texts must preserve the relationship between facts, applicable rules and conclusions, avoid unsupported references and maintain appropriate behavior when confronted with sensitive data, false premises or improper requests. In this context, technical adaptation remains a necessary but insufficient condition.

Therefore, the adaptation of LLMs for document generation in the context of this review should be understood as a multi-dimensional system-design problem, in which model specialization, knowledge retrieval, inference structuring, evaluation and governance perform complementary functions.

A. Hybrid Adaptation Strategies and their Trade-offs

The recurrence of hybrid strategies observed in RQ1 should not be interpreted as evidence that combining multiple adaptation mechanisms is inherently superior. Rather, the analyzed studies suggest that architecture choices reflect trade-offs among domain specialization, computational cost, knowledge updating, traceability and control over generation [2], [3], [12], [13], [17], [27]. These trade-offs become clearer when the limitations and operational requirements of each strategy are considered separately.

Full fine-tuning offers broader parameter adaptation and may provide stronger domain specialization, but at the cost of substantially higher computational and data requirements [2]. PEFT methods such as LoRA and QLoRA reduce memory and training costs by updating only a small subset of parameters. This efficiency, however, introduces a different trade-off: PEFT preserves most of the base model unchanged and therefore remains more dependent on its original capabilities and limitations. In both cases, adaptation quality remains strongly dependent on the representativeness and curation of the training data and neither strategy, by itself, guarantees factual or normative correctness.

These trade-offs are not exclusively computational. As shown in Fig. 4, broader parameter tuning is concentrated in HPC environments, whereas PEFT, RAG and structured prompting are more frequently associated with intermediate servers and workstations. This pattern suggests that adaptation choices are also conditioned by institutional requirements. In regulated settings, confidentiality, data-protection and technological-sovereignty requirements may limit the use of proprietary models accessed through external APIs, while auditability requirements may favor approaches that provide greater control over data handling, information sources and system operation. Consequently, a computationally less demanding strategy may also be institutionally preferable for rea-

sons beyond performance, particularly when sensitive records must remain within the institutional environment.

Retrieval-augmented generation addresses a different trade-off by externalizing part of the knowledge required for generation. In legal applications, this is particularly relevant because normative and case-specific information may change over time or depend on particular documents. Retrieval can reduce exclusive dependence on parametric memory and improve source traceability [3], [12]. However, this advantage shifts part of the system's reliability to the retrieval layer. Incomplete, outdated or irrelevant documents may be retrieved, and even relevant evidence may be interpreted or incorporated incorrectly by the model. Thus, RAG improves access to external evidence without eliminating the need to evaluate retrieval quality and the model's use of the retrieved information.

Structured inference introduces a different form of control by constraining how the response is organized. Approaches based on Chain-of-Thought, legal syllogism and argumentative structures can make the relationship between premises, rules and conclusions more explicit and facilitate human review [13], [17], [27]. The trade-off is that greater structural transparency does not imply greater substantive correctness. A model may produce a coherent and well-structured argument while relying on false premises, irrelevant retrieved evidence or inapplicable legal rules. Structured inference should therefore be understood as a mechanism for organizing and inspecting generation rather than as a substitute for factual or legal validation.

To make these relationships explicit, Table XIII organizes the main trade-offs into deployment scenarios defined by local computational capacity, restrictions on external data processing and requirements for institutional control. The scenarios provide a qualitative synthesis of the implications identified across the reviewed evidence and were not coded as mutually exclusive study-level categories during data extraction.

As summarized in Table XIII, the relevant trade-offs extend beyond computational performance. When external processing is permitted, prompting, ICL and API-based solutions can reduce local infrastructure requirements, but increase dependence on external providers and reduce institutional control over data handling and system auditability. When sensitive records must remain within the institutional environment, locally deployable open models, PEFT and controlled RAG configurations become more compatible with confidentiality, traceability and technological-sovereignty requirements, although they transfer infrastructure, maintenance and governance costs to the institution. Where both computational capacity and data mobility are restricted, compact models and parameter-efficient or inference-based approaches improve deployment feasibility but do not eliminate limitations related to model capacity, retrieval quality, data curation and evaluation.

Taken together, these scenarios indicate that hybridization should be driven by task and deployment requirements rather than treated as a default architecture. The reviewed evidence does not establish a universally superior combination; strategy suitability depends on task characteristics, computa-

TABLE XIII
QUALITATIVE SYNTHESIS OF ADAPTATION STRATEGIES UNDER
DEPLOYMENT CONSTRAINTS

Deployment scenario	Suggested adaptation strategies	Main trade-off
Substantial local computational capacity and strong specialization requirements	Full or continued parameter tuning, feedback alignment and hybrid architectures	Greater specialization and institutional control, at the cost of higher requirements for training data, computation, evaluation and maintenance.
Limited local infrastructure and external processing permitted	Prompting or in-context learning and RAG supported by externally hosted models or APIs	Lower local computational requirements, at the cost of provider dependence and reduced control over data handling and system auditability.
Sensitive or confidential data must remain local, with adequate local computational resources	Locally deployed open models, parameter-efficient fine-tuning, controlled parameter tuning and local RAG	Greater confidentiality, traceability and technological sovereignty, at the cost of local infrastructure, maintenance and governance requirements.
Sensitive or confidential data must remain local, with limited computational resources	Compact local models, parameter-efficient fine-tuning or quantization, structured prompting and constrained local RAG	Greater deployment feasibility and institutional control, at the cost of reduced model capacity, additional pipeline complexity and greater dependence on data curation and evaluation quality.

tional conditions, data-governance requirements, data quality, knowledge-update needs and evaluation criteria. Table XIII should therefore be interpreted as a qualitative design synthesis rather than as a comparative ranking of adaptation strategies.

B. Semi-Synthetic Data, Curation and Privacy

The results of RQ2 indicate that data engineering is one of the most sensitive layers in the adaptation of LLMs to the legal domain. The predominance of semi-synthetic datasets suggests an attempt to overcome the scarcity, confidentiality and curation cost of real legal data, without completely abandoning control and validation mechanisms [7], [24]. The gain lies in increasing scale and variety, but the risk lies in the possibility of reproducing biases, consolidating weak argumentative patterns or creating examples that are formally plausible and normatively inaccurate [5], [15].

For this reason, curation deserves particular attention at this stage. A smaller but well-selected dataset may be more useful than a broad and poorly controlled corpus, especially in tasks that require terminological precision, argumentative coherence and normative adherence [16], [18]. This shifts the focus from data quantity to the quality and validation of the datasets used.

Privacy-related aspects add another point of concern. Real documents bring the model closer to the application domain, but increase the risk of exposing personal or confidential data. Synthetic data reduce this direct risk, but do not eliminate indirect leakage, reproduction of sensitive patterns or statistical distortions. Thus, in legal applications involving sensitive data, anonymization should be accompanied by residual risk

assessment, access control, traceability of data use and explicit criteria for excluding sensitive information.

C. Weaknesses in Evaluation Protocols

While adaptation strategies have advanced through fine-tuning, PEFT, RAG and structured inference, evaluation methods still frequently depend on traditional NLP metrics, such as BLEU, ROUGE, BERTScore, METEOR and perplexity [3], [11], [18], [20]. These metrics are useful for assessing textual similarity, fluency or semantic approximation, but they do not adequately capture decisive dimensions for legal texts, such as the validity of normative sources, adherence to facts, correctness of reasoning and compatibility between premises and conclusion, as shown in the results of RQ3.

This limitation does not mean that automatic metrics should be discarded, since they remain relevant for screening, initial comparison between models and operational performance monitoring. The problem arises when they are treated as sufficient evidence of reliability. In legal generation, a response that is lexically close to a reference may contain a non-existent citation, apply an inadequate rule or construct reasoning incompatible with the case.

For this reason, studies proposing multidimensional criteria, expert evaluation, hallucination verification and approaches such as LLM-as-a-Judge represent an important methodological advance, although they remain heterogeneous and poorly standardized [14], [17], [22], [23], [26]. Special attention should be given to LLM-as-a-Judge. Although it is a scalable alternative to complement or reduce the cost of expert evaluations [21], [23], the method should be treated as an auxiliary evaluation, subject to bias from the evaluator model, prompt variation and the need for calibration with specialized human judgment.

The implication is that evaluation must cease to be only a posterior stage and become part of the adaptation architecture itself. For legal document generation, it is not sufficient to demonstrate that the model produces coherent text. Evaluation should also examine whether the response is factually consistent, grounded in applicable legal sources, appropriately calibrated regarding uncertainty and resistant to false premises or poorly retrieved evidence. Methods based on NLI, self-contradiction across multiple generations and confidence calibration point in this direction, although they do not yet constitute a consolidated protocol for institutional validation [17], [26].

Therefore, an important methodological gap remains. It is not enough to adapt models more effectively; it is also necessary to verify robustly when their responses can be considered reliable. The methodological consequence is that evaluation should combine automatic metrics, source verification, factual consistency analysis, expert judgment and confidence calibration.

D. Governance and Ethical Safeguards

The analysis of the results of RQ4 shows that privacy, bias, safety and accountability are frequently mentioned as requirements for institutional use. However, few studies implement

specific and measurable safeguards throughout the pipeline. The results show that, when these measures are reported, they tend to concentrate on isolated points, such as human curation, data balancing, controlled synthetic generation, RLHF, adversarial training or harmlessness metrics [3], [8], [15], [22], [23].

It is important to observe that ethical risks are not limited to the moment of the final response. They may arise in source selection, unequal representation of social groups, incorporation of historical patterns from the data, prompt formulation, the way the model handles false premises and the degree of confidence with which it presents a legally inadequate conclusion. In this sense, localized safeguards can reduce specific risks, but do not replace a governance architecture capable of connecting data, training, inference, evaluation and monitoring.

In legal applications involving sensitive or institutionally consequential documents, system design should specify which data may be used, which behaviors should be restricted, how responses will be audited and when human intervention is required. A model with good linguistic performance may still exhibit latent biases, expose sensitive information, improperly refuse legitimate requests or produce responses with an appearance of technical authority without adequate normative support. Governance, therefore, should not be understood only as an external compliance requirement, but as a technical component of the system itself.

The analyzed literature points to relevant directions, such as human-in-the-loop approaches that reinforce curation and institutional accountability [3], [7]; causal methods that allow auditing and reduction of biased associations in synthetic data [15]; and alignment and safety strategies that seek to limit harmful or legally inadequate responses [8], [23]. The challenge, however, lies in integrating these layers into reproducible protocols. Without this integration, the adaptation of LLMs for legal generation tends to remain technically promising, but methodologically vulnerable regarding ethical safety, auditability and accountability in real-world use.

E. Implications for Analogous Regulated Institutional Contexts

The implications of this review go beyond the legal domain in the strict sense, but should not be interpreted as automatic generalization to any institutional context. The analyzed corpus is predominantly composed of legal studies and only a limited portion directly addresses applications in Public Security. Even so, the results are relevant to other regulated environments because these contexts share similar constraints, such as the use of sensitive data, the need for grounding, the production of documents with institutional effects, traceability requirements and the possibility of accountability for decisions supported by automated systems.

The findings also suggest that, in functionally analogous regulated environments, the adoption of LLMs should not be guided only by textual performance. Where document generation has institutional or legal effects, systems may require adherence to verifiable facts, authoritative sources, traceability

and domain-specific formal standards. A model capable of producing fluent text, but unable to support the provenance of the information used or to preserve the relationship between facts, rules and conclusions, may introduce additional risks. In such contexts, the combination of model adaptation, knowledge retrieval, structured inference and expert validation represents a potentially useful architectural direction, but requires domain-specific empirical validation before operational adoption.

Police and administrative documents, for example, may involve incomplete factual narratives, specific technical language, multiple subjects, personal data, confidential information and relevant legal consequences. In such settings, model adaptation should preserve the individualization of conduct, consistency between evidence and legal classification, separation between narrated facts and model inference and the possibility of human review before official use. These requirements are functionally similar to those identified in legal generation, while also demanding controls specific to each institutional environment.

Thus, the central implication for regulated institutional contexts is that LLMs should be treated as supporting components for document production and not as substitutes for technical or decision-making responsibility. The usefulness of these models depends less on the ability to generate long and formally adequate texts and more on the existence of mechanisms for curation, grounding, auditing, data protection and human validation. Without these layers, automation may merely shift the risk from incomplete or inconsistent documents to documents that appear sophisticated, but are factually fragile, legally inadequate or institutionally opaque.

V. LIMITATIONS OF THE EVIDENCE AND FUTURE DIRECTIONS

The findings of this review allow the characterization of relevant trends in the Legal NLP literature, especially the recurrent use of hybrid adaptation strategies, the centrality of data engineering, the insufficiency of traditional metrics for evaluating legal generation and the still peripheral position of ethical safeguards. The scope of these conclusions, therefore, does not support direct empirical generalization to all regulated institutional contexts. The approximation with areas such as Public Security, public administration and institutional oversight should be understood by functional analogy, based on shared requirements of grounding, traceability, protection of sensitive data and decision-making accountability.

The heterogeneity of the included studies also limits the type of conclusion that can be drawn. Since the analyzed studies vary in tasks, languages, models, datasets, adaptation techniques, metrics and validation methods, this review identifies recurrent patterns and methodological gaps, but does not allow ranking approaches or affirming the general superiority of an architecture, fine-tuning technique, RAG configuration or structured inference method. Comparisons of this nature would require controlled studies, common datasets, shared evaluation protocols and expert validation.

The same caution applies to the interpretation of the identified gaps. The low frequency of factuality, privacy, bias and ethical safety metrics may indicate immaturity of the field, but it also reflects the level of detail reported by the primary studies and the temporal scope of a rapidly evolving literature. Thus, the gaps described in this review should be read as evidence of the current state of the analyzed production and not as a definitive absence of technical or methodological solutions.

Further progress in this area requires overcoming technical challenges, beginning with the structuring of legal-institutional benchmarks grounded in realistic generative tasks. In parallel, methodological maturity requires privacy, safety and bias mitigation safeguards to move from theoretical requirements to active metrics in the training pipeline. Addressing these fronts would provide a stronger methodological basis for future empirical validation in regulated institutional contexts that remain underrepresented in the corpus, including Public Security.

VI. FINAL CONSIDERATIONS

This systematic review identified a recurrent use of hybrid adaptation strategies for legal document generation, in which PEFT, RAG and structured inference perform complementary functions. The analyzed corpus indicates that the technical maturity of these strategies is not yet accompanied by equally consolidated protocols for factual evaluation, data governance, bias mitigation and institutional validation. Thus, the central contribution of the review lies in showing that the adoption of LLMs for legal document generation should be treated as a problem of architecture, evaluation and governance, and not only as a problem of textual performance. This integrated synthesis distinguishes the present review from broader surveys organized primarily around legal applications, model families, datasets or benchmarks.

Taken together, the four research questions were answered with different levels of evidential density. The corpus provides stronger evidence regarding adaptation strategies and data engineering, while evaluation, privacy and ethical safeguards remain less mature and more heterogeneous. This asymmetry is itself one of the main findings of the review.

For analogous regulated institutional contexts, the findings suggest that LLMs should be treated as supporting components for document production rather than substitutes for technical or decision-making responsibility. The adoption of these models depends on technical governance throughout the entire system lifecycle. Future research should prioritize the development of legal-institutional benchmarks, the standardization of multidimensional evaluation protocols and empirical validation in analogous regulated institutional contexts that remain underrepresented in the corpus, including Public Security.

REFERENCES

- [1] R. Al-Qaesm, M. Hendi and B. Tantour, "Alkafi-llama3: Fine-tuning LLMs for precise legal understanding in Palestine," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 107, 2025, doi: 10.1007/s44163-025-00313-w.
- [2] P. Colombo, T. Pires, M. Boudiaf, R. Melo, D. Culver, S. Morgado, E. Malaboeuf, G. Hauteux, J. Charpentier and M. Desa, "SaulLM-54B & SaulLM-141B: Scaling up domain adaptation for the legal domain," *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024.
- [3] P. S. García-Montero, P. Vizcaíno, I. G. Reyes-Chacón and M. E. Morocho-Cayamcela, "Legal AI for all: Reducing perplexity and boosting accuracy in normative texts with fine-tuned LLMs and RAG," *IEEE Access*, vol. 13, pp. 179759–179775, 2025, doi: 10.1109/ACCESS.2025.3622138.
- [4] N. T. Ha, T.-P. Nguyen, K. T. Trung, H.-L. Le, L. T. V. Huong, C. T. Nguyen and M.-T. Nguyen, "Vietnamese legal question answering: An experimental study," in *Proc. 2024 16th Int. Conf. Knowl. Syst. Eng. (KSE)*, 2024, pp. 440–446, doi: 10.1109/KSE63888.2024.11063637.
- [5] A. Halterman, "Synthetically generated text for supervised text analysis," *Political Analysis*, vol. 33, no. 3, pp. 181–194, 2025, doi: 10.1017/pan.2024.31.
- [6] S. Ghosh, D. Verma, B. Ganesan, P. Bindal, V. Kumar and V. Bhatnagar, "InLegalLLaMA: Indian legal knowledge enhanced large language model," *CEUR Workshop Proc.*, vol. 3818, pp. 37–46, 2024.
- [7] H. Kim, D. Kim, J. Lee, C. Yoon, D. Choi, M. Gim and J. Kang, "LAPIS: Language model-augmented police investigation system," in *Proc. Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 4637–4644, doi: 10.1145/3627673.3680044.
- [8] C. Liu, Y. Kang, S. Wang, L. Qing, F. Zhao, C. Wu, C. Sun, K. Kuang and F. Wu, "More than catastrophic forgetting: Integrating general capabilities for domain-specific LLMs," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2024, pp. 7531–7548, doi: 10.18653/v1/2024.emnlp-main.429.
- [9] Z. Liu, Y. Zhu and M. Lu, "Enhancing legal expertise in large language models through composite model integration: The development and evaluation of Law-Neo," in *Proc. Nat. Leg. Lang. Process. Workshop (NLLP)*, 2024, pp. 33–41.
- [10] S. Huang, Y. Liu, J. Qi, H. Yang, Z. Luan and D. Qian, "Learning to follow domain-specific instruction with verifiable rewards," in *Proc. 2025 Int. Joint Conf. Neural Netw. (IJCNN)*, 2025, pp. 1–8, doi: 10.1109/IJCNN64981.2025.11228901.
- [11] H. D. Pimpale, A. Raut, Y. Patil, P. Parpol, P. Yadav and J. Sangoi, "LEGALMIND: A fine-tuned Gemma-2-based legal assistant for Indian judiciary with RAG and embedding integration," *Journal of Engineering and Technology for Industrial Applications*, vol. 11, no. 55, pp. 105–117, 2025, doi: 10.5935/jetia.v11i55.1925.
- [12] D. Sinha and O. Sharma, "Generating legal arguments using LLM and vector database to support precedents," in *Proc. 2025 Int. Conf. Next Generation Inf. Syst. Eng. (NGISE)*, vol. 1, 2025, pp. 1–5, doi: 10.1109/NGISE64126.2025.11085307.
- [13] Y. Song, Y. Qin, R. Huang, Y. Chen and C. Lin, "Legal text summarization via judicial syllogism with large language models," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 5, 2025, doi: 10.1007/s44443-025-00113-3.
- [14] W. Su, B. Yue, Q. Ai, Y. Hu, J. Li, C. Wang, K. Zhang, Y. Wu and Y. Liu, "JUDGE: Benchmarking judgment document generation for Chinese legal system," in *Proc. 48th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2025, pp. 3573–3583, doi: 10.1145/3726302.3730295.
- [15] N. Nagesh, Z. Wang and A. M. Rahmani, "FairCauseSyn: Towards causally fair LLM-augmented synthetic data generation," in *Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–6, doi: 10.1109/EMBC58623.2025.11252705.
- [16] R. Ta, R. Salunke, R. R. R. Nv and S.R. Upadhyaya, "Fine-tuning a large language model for the Indian legal system," in *Proc. Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, 2025, doi: 10.1109/INFOTEH64129.2025.10959207.
- [17] U. Ujwal, S. S. H. Surampudi, S. Mitra and T. Saha, "Reasoning before responding: Towards legal long-form question answering with interpretability," in *Proc. Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 4922–4930, doi: 10.1145/3627673.3680082.
- [18] F. Valerio, P. Basile and M. de Gemmis, "Adapting a large language model to the legal domain: A case study in Italian," *CEUR Workshop Proc.*, vol. 3877, 2024.
- [19] N. Xie, Y. Bai, H. Gao, F. Fang, Q. Zhao, Z. Li, Z. Xue, L. Zhu, S. Ni and M. Yang, "DeliLaw: A Chinese legal counselling system based on a large language model," in *Proc. 33rd ACM Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 5299–5303, doi: 10.1145/3627673.3679219.

- [20] Y. Yu, "Domain-specific legal language modeling through knowledge fusion and structured attention," in *Proc. Int. Conf. Electr. Inf. Technol. Comput. Eng. (EITCE)*, 2025, pp. 685–689, doi: 10.1145/3766671.3766790.
- [21] H. Le, N. Luu, T. Nguyen, T. Dao and S. Dinh, "Optimizing answer generator in Vietnamese legal question answering systems using language models," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 6, 2025, doi: 10.1145/3732938.
- [22] D. Shu, H. Zhao, X. Liu, D. Demeter, M. Du and Y. Zhang, "LawLLM: Law large language model for the US legal system," in *Proc. 33rd ACM Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 4882–4889, doi: 10.1145/3627673.3680020.
- [23] S. Yao, Q. Ke, Q. Wang, K. Li and J. Hu, "Lawyer GPT: A legal large language model with enhanced domain knowledge and reasoning capabilities," in *Proc. 2024 3rd Int. Symp. Robotics, Artificial Intelligence and Information Engineering*, 2024, pp. 108–112, doi: 10.1145/3689299.3689319.
- [24] J. Cui, Z. Li, Y. Yan, B. Chen and L. Yuan, "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2023.
- [25] Z. Zhou, J.-X. Shi, P.-X. Song, X.-W. Yang, Y.-X. Jin, L.-Z. Guo and Y.-F. Li, "LawGPT: A Chinese legal knowledge-enhanced large language model," *arXiv preprint arXiv:2406.04614*, 2024.
- [26] M. Dahl, V. Magesh, M. Suzgun and D. E. Ho, "Large legal fictions: Profiling legal hallucinations in large language models," *Journal of Legal Analysis*, vol. 16, no. 1, pp. 64–93, 2024, doi: 10.1093/jla/laae003.
- [27] C. Jiang and X. Yang, "Legal syllogism prompting: Teaching large language models for legal judgment prediction," in *Proc. Nineteenth Int. Conf. Artificial Intelligence and Law*, 2023, pp. 417–421, doi: 10.1145/3594536.3595170.
- [28] F. Yu, L. Quartey and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2022, doi: 10.48448/sdt7-nc75.
- [29] J. Cui, Z. Li, Y. Yan, B. Chen and L. Yuan, "ChatLaw: Open-source legal large language model with integrated external knowledge bases," *arXiv preprint arXiv:2306.16092*, 2023.
- [30] W. Deng, J. Pei, K. Kong, Z. Chen, F. Wei, Y. Li, Z. Ren, Z. Chen and P. Ren, "Syllogistic reasoning for legal judgment analysis," in *Proc. 2023 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2023, pp. 10283–10298.
- [31] C. Xiao, X. Hu, Z. Liu, C. Tu and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021, doi: 10.1016/j.aiopen.2021.06.003.
- [32] N. Guha, D. E. Ho, J. Nyarko and C. Ré, "LegalBench: Prototyping a collaborative benchmark for legal reasoning," *arXiv preprint arXiv:2209.06120*, 2022.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems*, 2017, pp. 6000–6010.
- [34] J. Crawford, "Systematic literature reviews: Why I rejected your review," *Journal of University Teaching & Learning Practice*, vol. 22, no. 2, 2025, doi: 10.53761/10vb5076.
- [35] H. Touvron et al., "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [36] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations*, 2022.
- [37] T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Proc. 37th Int. Conf. Neural Information Processing Systems*, 2023.
- [38] J. Popay, H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, et al., *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews*. ESRC Methods Programme, 2006.
- [39] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, et al., *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2019, doi: 10.1002/9781119536604.
- [40] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *Keele University and Durham University, Tech. Rep. EBSE-2007-01*, 2007.
- [41] P. Lewis et al., "retrieval augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.
- [42] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.
- [43] M. Ouzzani, H. Hammady, Z. Fedorowicz and A. Elmagarmid, "Rayyan: A web and mobile app for systematic reviews," *Systematic Reviews*, vol. 5, no. 210, 2016, doi: 10.1186/s13643-016-0384-4.
- [44] R. Rafailov et al., "Direct preference optimization: Your language model is secretly a reward model," *Adv. Neural Inf. Process. Syst.*, 2023.
- [45] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Evaluation and Assessment in Software Engineering*, 2014, doi: 10.1145/2601248.2601268.
- [46] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [47] J. Lai, W. Gan, J. Wu, Z. Qi and P. S. Yu, "Large language models in law: A survey," *AI Open*, vol. 5, pp. 181–196, 2024, doi: 10.1016/j.aiopen.2024.09.002.
- [48] Z. Hou, Z. Ye, N. Zeng, T. Hao and K. Zeng, "Large language models meet legal artificial intelligence: A survey," *arXiv preprint arXiv:2509.09969*, 2025.
- [49] F. Dehghani, R. Dehghani, Y. Naderzadeh Ardebili and S. Rahnamayan, "Large language models in legal systems: A survey," *Humanities and Social Sciences Communications*, vol. 12, no. 1, Art. no. 1977, 2025, doi: 10.1057/s41599-025-05924-3.