ANTHROPOLOGICAL NOTEBOOKS
ISSN / eISSN 1408-032X
Publisher SLOVENE ANTHROPOLOGICAL SOC,
FAK SPORT, GORTANOVA 22,LJUBLJANA,
SLOVENIA, SI-1000
Volume 26-Issue 1-(2026)
DOI:https://doi.org/10.65613/741579

OPEN ACCESS

AI-Driven Legal Governance: A Computational Framework for Regulatory Decision Support Based on Hierarchical Prompt Engineering of Large Language Models

Check for updates

Yuhe Bu*

*School of Law, Shandong University of Finance and Economics, Address, Jinan, Shandong, 250014, China

*shiyunli123@163.com

Abstract

The rapid maturation of large language models (LLMs) and of prompt engineering as a systematic design discipline has created new opportunities to reform legal governance and regulatory oversight. Traditional regulatory institutions face three persistent pressures: information overload, interpretive inconsistency across reviewers, and slow reaction to emerging societal risks. This paper proposes and evaluates an integrated computational framework, the AI-assisted Legal Governance System (ALGS), which combines LLM prompt engineering, legal natural language processing, machine learning, and legal knowledge graphs within a single regulatory decision-support pipeline. ALGS is organised as four interdependent layers: data ingestion and pre-processing, an AI analysis engine, knowledge representation and reasoning over a typed legal knowledge graph, and a decision support interface. Its central methodological contribution is a hierarchical prompt architecture that decomposes complex legal analysis into an ordered sequence of narrowly scoped sub-tasks, so that provision identification, fact extraction, requirement-by-requirement evaluation, and regulatory gap analysis are performed in discrete, individually auditable steps rather than in a single opaque inference. Every step is grounded in retrieved authority and accompanied by a calibrated confidence score, which together with the assessed risk of the matter determines which of three tiers of human review the output receives. We evaluate the framework on regulatory and contractual corpora, including United States Securities and Exchange Commission periodic filings and commercial contract collections, together with a pilot deployment in a commercial court. ALGS attains an accuracy of 0.871 and an F1-score of 0.819 on regulatory risk classification, corresponding to a 23.9% relative improvement in accuracy over a logistic-regression baseline and an 8.7% relative improvement in accuracy over a fine-tuned LEGAL-BERT baseline, while reducing human review time by 41.2% on the compliance classification task. We further examine the ethical and constitutional constraints on AI autonomy in legal decision-making, including bias mitigation, explainability, and the principle of meaningful human oversight. The results indicate that carefully engineered prompt architectures can materially strengthen regulatory capacity while preserving the due-process and rule-of-law commitments that legitimate legal decisions require.

Keywords: large language models; prompt engineering; legal governance; regulatory decision support; legal natural language processing; legal knowledge graphs; AI and law; human-in-the-loop oversight

1. Introduction

The convergence of artificial intelligence and legal systems has become one of the most consequential research frontiers in contemporary computer science. Over the past decade, machine learning has advanced

from narrow, task-specific applications toward systems capable of broad and flexible reasoning. The emergence of large transformer-based language models, such as the GPT, Claude, and LLaMA families, has been particularly transformative: these systems interpret, generate, and reason over complex natural language with a fluency that approaches expert human performance on many standardised tasks [1]. Legal text — characterised by precise drafting conventions, deeply nested hierarchical structure, and dense networks of cross-reference — is simultaneously a promising application domain and a demanding test of these capabilities [2].

The entry of AI into legal practice has proceeded along several fronts. Early applications concentrated on e-discovery and document review, where machine learning identified responsive material within very large evidentiary collections far more efficiently than manual inspection could. Predictive analytics tools subsequently emerged that forecast case outcomes from large corpora of prior judicial decisions. More recently, contract analysis platforms built on natural language processing (NLP) have automated the extraction of key clauses, risk indicators, and obligation summaries from commercial agreements. Taken together, these developments mark a broader transition toward AI-augmented legal practice, in which intelligent systems act as analytical collaborators with lawyers, regulators, and judges rather than as autonomous substitutes for them [3].

Despite this progress, the institutional architecture of legal governance and regulatory supervision has changed remarkably little since the pre-digital era. Regulatory authorities continue to rely primarily on manual review, periodic inspection cycles, and reactive enforcement that responds to violations after they occur rather than anticipating emerging risk. This structural rigidity produces significant governance gaps precisely in the sectors characterised by the fastest technological change — financial technology, biotechnology, and digital platform markets — where the pace of innovation routinely outstrips the capacity of regulatory institutions to develop and implement adequate supervisory frameworks.

The potential of AI-based legal decision-support systems to close these gaps is substantial. By automating routine analytical work — interpreting standardised evidentiary patterns, applying compliance checklists, and performing initial risk triage — such systems can expand the effective analytical capacity of a regulator without a proportionate increase in headcount. More significantly, well-designed systems can support genuinely complex legal inference that previously required expert judgment: identifying analogical relationships between cases, detecting logical inconsistencies in regulatory filings, and drafting preliminary advisory analyses that frame a matter for a human decision-maker. This augmentation model, rather than substitution for human judgment, is the one that administrative law scholarship identifies as legally tractable [3], [4].

The deployment of AI in legal governance nevertheless raises difficulties that are far less acute in other application domains. Legal decisions affect fundamental rights and are backed by coercive state power; errors in AI-informed regulatory determinations can deprive individuals and organisations of liberty or property. This imposes demanding requirements of accuracy, explainability, and procedural fairness that many current AI systems struggle to satisfy [5]. Two empirical findings sharpen the concern. First, general-purpose LLMs hallucinate legal authority at high rates: Dahl et al. report hallucination rates between 58% and 82% on legal queries posed to leading general-purpose models [6]. Second, even commercial retrieval-augmented legal research tools marketed as hallucination-free were found to produce incorrect or misgrounded answers in more than one out of six queries [7]. Legal reasoning is moreover heavily context-dependent and normatively contested, and models trained on historical legal data can encode and perpetuate

the systematic biases embedded in prior judicial and regulatory decisions, raising well-documented concerns about algorithmic fairness and equal protection [4].

These findings do not counsel abandonment of AI in legal governance; they counsel architecture. The failures documented above are predominantly failures of unstructured, single-shot use of general-purpose models, in which a complex legal question is posed to an LLM as one undifferentiated prompt and the answer is accepted without grounding, decomposition, or confidence assessment. This paper argues that the appropriate response is to constrain the model architecturally — to decompose legal analysis into steps that can be individually grounded, individually verified, and individually escalated to a human when the system's own confidence is insufficient.

Accordingly, this paper proposes the AI-assisted Legal Governance System (ALGS), a structured computational framework that integrates LLM prompt engineering with explicit legal inference structures, uncertainty quantification, and human oversight protocols. The research questions are as follows:

(RQ1) How can prompt engineering techniques be applied systematically to decompose complex legal analysis into sub-tasks that an LLM can perform reliably and auditably?

(RQ2) What architectural principles govern the effective integration of AI analytical capacity into regulatory decision-making while preserving procedural legality?

(RQ3) How do AI-based regulatory systems perform empirically, relative to both statistical baselines and human experts, on legal text classification, risk identification, and compliance assessment?

This paper makes four contributions. First, it specifies a hierarchical prompt architecture for regulatory analysis that decomposes compliance determination into five ordered, individually grounded reasoning steps (Section 4.2). Second, it integrates that architecture with a typed legal knowledge graph so that each reasoning step is anchored in retrieved authority rather than in parametric memory (Section 4.5). Third, it defines a calibrated uncertainty quantification and tiered escalation protocol that makes the degree of human oversight a function of measured model confidence and case risk, rather than a fixed policy (Section 4.7). Fourth, it reports a multi-corpus empirical evaluation against three baselines of increasing strength, with an explicit methodology, and situates the results within a limitations analysis and a governance discussion grounded in the emerging regulatory framework for high-risk AI systems [8].

The remainder of the paper is organised as follows. Section 2 reviews related work and positions ALGS against representative prior systems. Section 3 presents the technical background in legal NLP, machine learning, knowledge graphs, and prompt engineering. Section 4 specifies the ALGS framework. Section 5 sets out the experimental methodology. Section 6 reports empirical results across three application studies and an ablation of the framework's components. Section 7 discusses implications, ethics, and limitations. Section 8 concludes.

2. Related Work

2.1 Artificial Intelligence and Law: From Rule-Based Systems to Neural Models

Scholarly work on artificial intelligence in legal contexts long predates the deep learning era. Pioneering research in the 1970s and 1980s explored expert systems and logical inference engines for automating aspects of legal analysis, producing systems such as HYPO [9], which modelled case-based argumentation in trade secret law, and TAXMAN [10], which formalised reasoning in corporate tax law.

These systems encoded legal knowledge through hand-crafted rules and structured conceptual representations, achieving strong performance on precisely delimited tasks while remaining brittle in the face of the interpretive ambiguity that characterises legal practice. The fundamental limitation of rule-based approaches — their dependence on expensive expert knowledge engineering and their inability to generalise beyond explicitly encoded rules — motivated the shift toward data-driven machine learning methods, a shift enabled by the increasing availability of large digital legal corpora.

Contemporary research in AI and law has expanded rapidly, building on these corpora and on advances in neural architectures. Aletras et al. demonstrated that machine learning models can predict judgments of the European Court of Human Rights with approximately 79% accuracy using features extracted from case texts [11]. Zhong et al. developed neural models for the Chinese legal system capable of predicting charge categories and sentence lengths from factual case descriptions [12]. Transformer-based approaches have since achieved strong results across legal NLP tasks, including statutory reasoning, contract understanding, and legal question answering.

Regulatory and compliance applications have attracted particular scholarly attention. Katz, Bommarito, and Blackman demonstrated that ensemble tree methods trained on two centuries of case records can predict United States Supreme Court outcomes at scale, establishing that large-scale statistical prediction of legal decisions is feasible outside narrow doctrinal niches [13]. Surden proposed an influential distinction between AI systems that substitute for human legal judgment and those that augment it, arguing that the latter category is both more practically tractable and more legally appropriate given the constitutional constraints on automated governmental decision-making [3]. Doshi-Velez et al. addressed the tension between predictive accuracy and interpretability in legally consequential decision contexts and articulated the conditions under which an AI system can be said to provide a legally adequate explanation [14]. In the administrative law literature, Coglianese and Lehr examined whether machine learning can be reconciled with the non-delegation, due process, and reason-giving requirements that govern agency action, concluding that properly supervised algorithmic tools can operate within existing doctrinal constraints [4]. Citron's earlier account of technological due process remains the canonical statement of the procedural safeguards that automated public decision systems must provide: notice, an opportunity to be heard, and a reviewable record of the basis for decision [5].

2.2 Legal Informatics and Computational Legal Studies

Legal informatics — the interdisciplinary study of information technology applied to legal processes and institutions — provides the theoretical foundation for AI-based legal governance research. Bench-Capon and Sartor developed argumentation-theoretic frameworks for the computational modelling of legal reasoning, showing that formal argument structures capture the dialectical and adversarial character of legal discourse more faithfully than purely deductive models [15]. Subsequent work in computational legal studies has applied network analysis to citation relationships among judicial decisions, identifying structural properties of precedent that predict doctrinal development and judicial influence [16]. Cross and Spriggs conducted an empirical study of citation practice at the United States Supreme Court, establishing that citation frequency and treatment carry measurable information about the authority of a precedent [17], while Bommarito and Katz applied graph-theoretic methods to the structure of the United States Code itself, quantifying its growth in size and interdependence over time [18].

The subfield of legal NLP has developed methodologies adapted to the distinctive properties of legal text. Lippi et al. proposed CLAUDETTE, a system for detecting potentially unfair clauses in online terms

of service that combines support vector machine classifiers with structural features of the clause text [19]. Chalkidis et al. introduced LEGAL-BERT, a family of BERT variants pre-trained on large collections of European Union legislation, court decisions, and contracts, and demonstrated that domain-specific pre-training yields consistent improvements over general-purpose BERT across legal NLP tasks [20]. Hendrycks et al. released CUAD, an expert-annotated corpus of 510 commercial contracts with more than 13,000 clause-level annotations across 41 clause categories, which established a rigorous benchmark for automated contract review and remains a standard evaluation resource [21].

The construction of legal knowledge representations constitutes a further active research area. Hoekstra et al. developed the LKIF-Core ontology, providing formal conceptual foundations for representing legal norms and their interrelations [22]. Sovrano, Palmirani, and Vitali investigated the automated construction of legal knowledge graphs from legal texts using information extraction techniques, enabling more structured querying and inference over regulatory content than raw text analysis permits [23]. These knowledge representation methods are complementary to the generative capabilities of neural language models, and their combination in hybrid architectures forms the basis of the framework proposed here.

2.3 Large Language Models in the Legal Domain

The application of general-purpose LLMs to legal tasks is a recent but rapidly maturing line of research. Guha et al. introduced LegalBench, a collaboratively constructed benchmark comprising 162 tasks spanning six categories of legal reasoning, designed and annotated by legal professionals [24]. LegalBench established two findings central to the present work: that LLM performance varies enormously across reasoning types, and that performance on rule-application and rule-conclusion tasks is markedly more reliable than on tasks requiring open-ended interpretive judgment. Fei et al. developed LawBench, a parallel benchmark for the Chinese legal system organised around a three-level cognitive taxonomy of memorisation, understanding, and application, and reported that even strong general-purpose models substantially underperform on tasks requiring precise recall of statutory content [25].

A complementary line of work pursues domain-adapted models. Colombo et al. released SaulLM-7B, a legal-domain LLM continually pre-trained on approximately 30 billion tokens of English legal text, demonstrating that domain adaptation improves legal task performance relative to general-purpose models of comparable scale [26]. Domain adaptation, however, does not by itself resolve the grounding problem: a model that has memorised more law can still fabricate authority.

The reliability literature is therefore essential context. Dahl et al. systematically profiled legal hallucination in general-purpose models, finding hallucination rates of 58% to 82% on legal queries and, critically, that models exhibit poor calibration — they are frequently most confident precisely where they are wrong [6]. Magesh et al. extended this analysis to commercial retrieval-augmented legal research products and found that even purpose-built systems produced hallucinated or misgrounded content in more than 17% of queries [7]. These results establish the two design requirements that motivate the ALGS architecture: every assertion must be traceable to retrieved authority, and the system's own confidence must be calibrated against measured accuracy rather than assumed.

2.4 Positioning of the Present Work

Table 1 situates ALGS relative to representative prior systems along five dimensions that matter for regulatory deployment: the core technique employed, the breadth of legal tasks covered, whether the system

produces an explicit and inspectable reasoning trace, whether it quantifies its own uncertainty, and whether it defines a structured human oversight protocol. The comparison shows a consistent pattern. Rule-based systems provide fully inspectable reasoning but do not generalise; supervised neural classifiers generalise well but produce predictions without justification; general-purpose LLMs with chain-of-thought prompting produce reasoning traces but do not ground them in verified authority and are poorly calibrated. No prior system combines grounded step-wise reasoning, calibrated uncertainty, and a risk-tiered oversight protocol in a single regulatory pipeline. That combination is the contribution of this work.

Approach	Core technique	Legal task coverage	Reasoning trace	Uncertainty quantification	Human oversight protocol
HYPO [9]; TAXMAN [10]	Hand-crafted rules; case-based argumentation	Narrow, single doctrine	Yes	No	No
Aletras et al. [11]; Zhong et al. [12]	Supervised ML over case text	Outcome prediction	No	Partial	No
CLAUDETTE [19]	SVM with structural features	Unfair clause detection	Partial	Partial	No
LEGAL-BERT [20]	Domain-adapted transformer classifier	Multiple legal NLP tasks	No	Partial	No
LKIF-Core [22]; Sovrano et al. [23]	Legal ontology and knowledge graph	Structured querying and inference	Yes	No	No
SaulLM-7B [26]	Domain-adapted LLM	Broad legal tasks	Partial	No	No
General-purpose LLM with chain-of-thought [27]	Single-prompt LLM inference	Broad but ungrounded	Yes	No	No
ALGS (this work)	Hierarchical prompting, knowledge-graph grounding, model ensemble	End-to-end regulatory analysis pipeline	Yes	Calibrated	Risk-tiered

Table 1. Comparison of ALGS with representative prior approaches to computational legal analysis. ‘Partial’ denotes that the capability is present in a limited or uncalibrated form. Named systems are cited; the final two rows denote classes of approach rather than specific systems.

3. Technical Background

3.1 Natural Language Processing for Legal Text

Applying NLP to legal text encounters a set of difficulties that distinguish the legal domain from general text processing. Legal language exhibits a high density of technical terminology, archaic constructions, and specialised grammatical patterns — heavy nominalisation, extensive use of the passive voice, and deeply subordinated sentence structures — that degrade the performance of standard NLP tooling calibrated on general corpora. Cross-referential structure is pervasive: statutes routinely amend, incorporate by reference, or repeal other statutes, producing complex dependency graphs that must be resolved before any provision can be correctly interpreted [18]. Legal meaning is also strongly context-sensitive, since identical terms frequently carry distinct technical meanings across different areas of law; the term ‘security’, for example, denotes entirely different concepts in securities regulation and in secured transactions law.

Transformer-based language models outperform earlier architectures on legal NLP tasks, largely because self-attention captures long-range dependencies within documents and because large-scale pre-training transfers effectively to specialised domains. The BERT family and its legal variants achieve strong results on legal named entity recognition, document classification, contract clause typing, and legal question answering [20]. In regulatory compliance analysis, such models are typically fine-tuned on labelled corpora of regulatory filings with annotated compliance outcomes. Discriminative classifiers of this kind, however, emit predictions without accompanying justification, which limits their usefulness in regulatory settings that require transparent and contestable analytical output [28].

LLMs operated through prompt engineering offer a complementary approach that yields interpretable reasoning chains. Chain-of-thought prompting guides a model to generate intermediate reasoning steps before producing a final answer, substantially improving performance on multi-step reasoning tasks including legal analysis [27]. Retrieval-augmented generation (RAG) architectures further strengthen legal reasoning by allowing the model to retrieve and condition on relevant statutory provisions, regulatory precedent, and case law at inference time, grounding output in verifiable legal sources rather than relying solely on knowledge encoded in model parameters [29]. RAG is a necessary but not sufficient condition for reliability: as Magesh et al. demonstrate, retrieval reduces but does not eliminate misgrounded output, since a model may retrieve correct authority and still misapply it [7].

3.2 Machine Learning for Legal Prediction

The machine learning models applied to legal prediction span a range of architectures matched to the structure of the prediction target. Binary classification models trained on historical enforcement records estimate the probability that a given regulatory filing will attract enforcement action, allowing authorities to prioritise higher-risk filings for review. Multi-label classification models identify several potential legal issues simultaneously within a single document, replicating the multidimensional analytical perspective an experienced reviewer brings to a complex filing. Sequence-to-sequence architectures extract structured information from unstructured legal text, converting narrative regulatory submissions into structured data fields amenable to downstream analysis.

Recurrent and transformer architectures have been applied to temporal legal prediction tasks, including forecasting the likelihood of future regulatory violations from historical compliance patterns and predicting litigation outcomes from procedural and substantive case features. Graph neural network methods have proven effective on tasks requiring relational reasoning over legal structures, including identifying relevant precedent within the citation networks whose structural properties Fowler et al. mapped [16], resolving statutory cross-references, and modelling relational dependencies among entities in complex organisational structures such as corporate groups.

Reinforcement learning from human feedback, the alignment technique underlying current instruction-following commercial models [30], has also been explored for legal applications. Cobbe et al. showed that training a separate verifier to score candidate solutions, and then selecting among them by verifier score, substantially improves the reliability of model output on complex reasoning tasks relative to fine-tuning alone [31]. Subsequent work extended this idea from verifying final solutions to supervising the individual reasoning steps that produce them. Analogous methods applied to legal reasoning can steer models toward legally sound analysis by rewarding correct intermediate legal moves — correctly identifying the governing provision, correctly characterising a fact — rather than rewarding only a correct final conclusion. This

distinction matters in regulatory settings, where a right answer reached by wrong reasoning is not an acceptable basis for coercive action.

3.3 Legal Knowledge Graphs

Legal knowledge graphs provide structured representations of legal entities, concepts, norms, and their relations, complementing the unstructured text processing capabilities of neural language models [23]. A legal knowledge graph for a given regulatory domain typically represents: legal instruments (statutes, regulations, guidance documents, judicial decisions); legal persons and organisations; legal concepts together with their defining relations; normative relations among provisions (obligation, prohibition, and permission); and temporal relations (commencement, amendment, and repeal dates) [2]. By querying such a graph, a regulatory AI system can rapidly identify all provisions in force governing a particular matter at a particular time and trace a regulatory requirement through successive normative layers [22].

Constructing legal knowledge graphs from primary sources generally combines information extraction, relation classification, and ontology alignment. Named entity recognition models identify and type legal entities within statutory and regulatory text; relation extraction models determine the typed relations that hold among identified entities; and ontology alignment algorithms map extracted entities and relations onto a formal legal ontology such as LKIF-Core [22]. The resulting graph supports structured querying and symbolic inference, complementing the pattern-recognition capabilities of neural models [23].

Maintaining currency and accuracy in legal knowledge graphs presents a substantial operational challenge, since legal content changes continuously through legislative amendment, new judicial decisions, and regulatory rulemaking. Automated update pipelines that monitor official publication sources, extract new and amended content, and propagate changes through dependent graph structures are essential for production deployment. A graph that silently lags the law is worse than no graph at all, because it grounds analysis in authority that has been superseded.

3.4 Prompt Engineering as a Design Discipline

Because prompt engineering is the methodological core of the proposed framework, it is worth stating precisely what the term denotes here. Prompt engineering is the systematic design of the input context supplied to a language model — including task specification, decomposition structure, retrieved evidence, output schema, and decoding strategy — so as to constrain the model's behaviour toward reliable and inspectable output. It is not the informal practice of rephrasing a question until the answer looks acceptable.

Four techniques are foundational to the ALGS design. Chain-of-thought prompting elicits explicit intermediate reasoning, which both improves accuracy on multi-step tasks and produces a trace that a human reviewer can audit [27]. Task decomposition replaces a single complex prompt with an ordered sequence of narrowly scoped prompts, each of which admits a verifiable answer; this reduces the search space at each step and localises errors, so that a mistake at step three does not silently contaminate the final conclusion. Retrieval augmentation supplies each step with the specific authoritative text it requires, rather than relying on parametric recall [29]. Self-consistency sampling generates multiple independent reasoning paths under stochastic decoding and aggregates them by majority vote, which both improves accuracy and — importantly for the present purpose — yields an empirical measure of the model's stability on a given input that can be calibrated into a confidence estimate [32].

The distinctive claim of this paper is that these four techniques are not merely additive accuracy improvements but the components of a governance architecture. Decomposition creates the audit points; retrieval creates the evidentiary grounding; self-consistency creates the uncertainty signal; and the uncertainty signal, once calibrated, determines where human judgment must intervene. This is the sense in which prompt engineering becomes a legal-institutional design question rather than a purely technical one.

4. The AI-assisted Legal Governance System

4.1 Architecture Overview

The AI-assisted Legal Governance System (ALGS) is designed as a modular, layered architecture that integrates multiple AI technologies into a single pipeline supporting regulatory decision-making. As illustrated in Figure 1, the framework comprises four principal layers: (1) a data ingestion and pre-processing layer, which collects, consolidates, and indexes legal texts from heterogeneous sources; (2) an AI analysis engine, which applies legal NLP, machine learning, and LLM prompt engineering to extract analytically relevant information from legal content; (3) a knowledge representation and reasoning layer, which maintains structured legal knowledge and performs symbolic inference; and (4) a decision support interface layer, which delivers analytical output to human decision-makers in usable and auditable form.

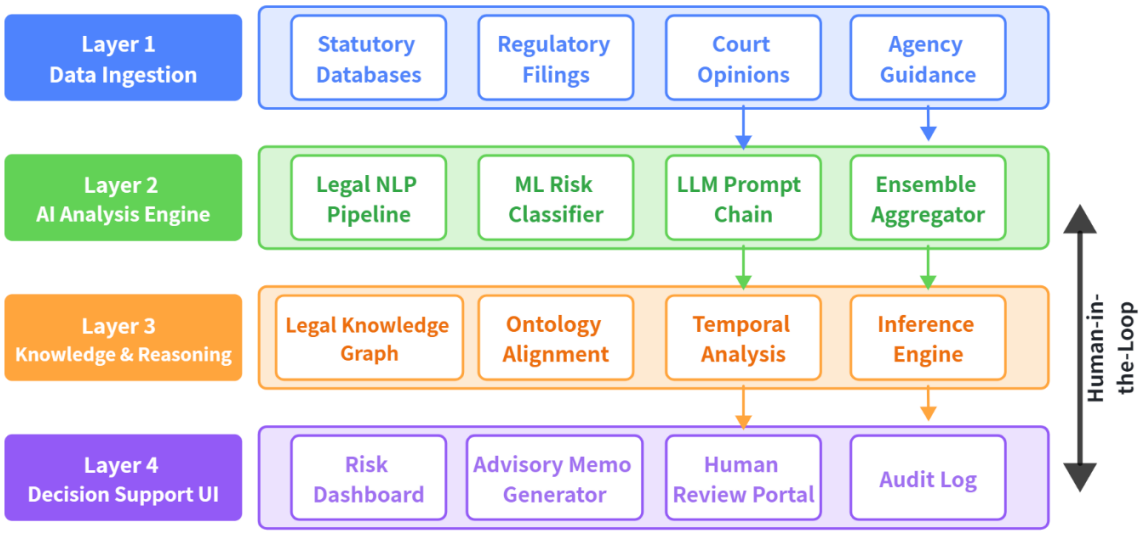

Figure 1. Layered architecture of the AI-assisted Legal Governance System (ALGS), comprising four interdependent layers connected through a human-in-the-loop feedback mechanism.

Sections 4.2 and 4.4 describe the components of the AI analysis engine that constitutes layer 2; Section 4.3 describes layer 1; Section 4.5 describes layer 3; and Sections 4.6 and 4.7 describe layer 4 together with the oversight protocol that governs it. The layering is not merely an engineering convenience. Each boundary between layers is a point at which output can be logged, inspected, and independently validated, which is what makes the system auditable in the sense that administrative law requires [5]. A monolithic architecture in which an LLM ingests a filing and emits a recommendation offers no such intermediate checkpoints, and consequently offers a reviewing court no record of the basis for decision.

4.2 Hierarchical Prompt Decomposition

The principal innovation of the ALGS framework is its hierarchical approach to prompting for legal inference tasks. Rather than presenting a complex legal question to the LLM as a single undifferentiated query, the framework decomposes each analytical task into a structured sequence of sub-prompts that progressively construct the analysis through well-defined reasoning steps. For a regulatory compliance determination, the pipeline directs the model sequentially to: (1) identify the applicable regulatory provisions; (2) extract the key factual assertions relevant to each provision; (3) assess whether each assertion satisfies the corresponding requirement; (4) identify gaps, ambiguities, or omissions in the disclosure; and (5) synthesise a preliminary compliance assessment. Figure 2 illustrates this decomposition.

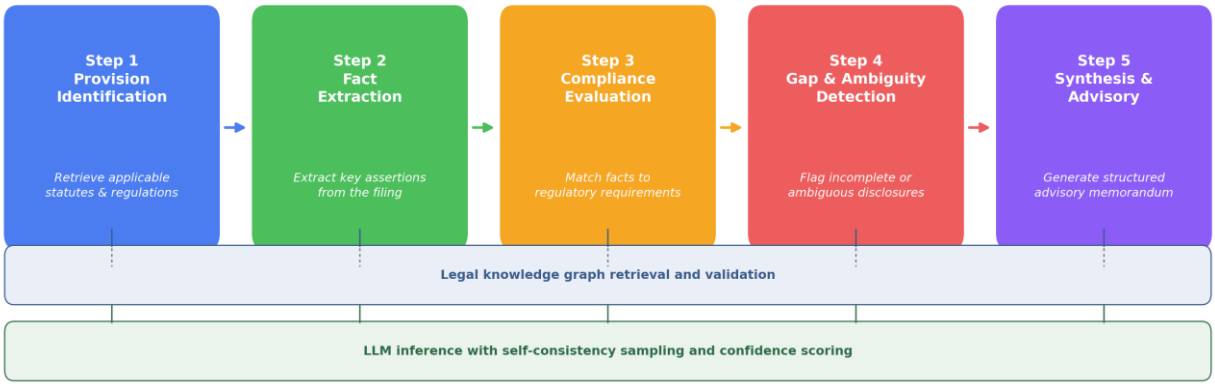

Figure 2. Five-step hierarchical prompt engineering pipeline for regulatory compliance analysis, with LLM inference and knowledge graph retrieval integrated at each stage.

Three properties of this decomposition are worth making explicit. First, each step has a bounded and verifiable output type: step 1 returns a set of provision identifiers that either do or do not exist in the knowledge graph; step 2 returns spans of the source document; step 3 returns a per-requirement satisfaction judgment; step 4 returns a list of requirements for which the filing supplies no responsive text; and step 5 returns a memorandum in a fixed schema whose every assertion must cite an output of an earlier step. A hallucinated provision identifier at step 1 fails validation against the knowledge graph immediately rather than propagating silently into a final recommendation. Second, each step is separately grounded: the sub-prompt for step 3 is populated with the actual text of the provisions identified at step 1 and the actual document spans extracted at step 2, so that the model is asked to evaluate presented evidence rather than to recall law from parametric memory. Third, the intermediate outputs constitute the audit trail. A reviewer who disagrees with the final assessment can identify precisely which step introduced the disagreement, which is not possible with an end-to-end classifier.

The decomposition also permits differential model allocation. Steps that are largely extractive, such as step 2, can be served by a smaller and cheaper model, while steps requiring normative judgment, such as steps 3 and 4, are routed to the strongest available model. This substantially reduces inference cost at scale without materially degrading accuracy, a practical consideration for regulators operating under fixed budgets.

4.3 Legal Text Analysis Module

The legal text analysis module forms the foundational layer of ALGS and is responsible for acquiring, processing, and semantically indexing legal content from a range of sources, including statutory databases, regulatory filings, case reporters, and agency guidance documents. The pre-processing pipeline applies

domain-adapted NLP techniques, including legal text segmentation that respects citation formats, article numbering conventions, and defined-term references; sentence boundary detection calibrated to the complex sentence structures characteristic of legal drafting; and cross-reference resolution that links citations to their target provisions within the knowledge base.

Semantic indexing of pre-processed legal text employs a hybrid retrieval architecture that combines a dense bi-encoder built on domain-adapted transformer models with sparse lexical retrieval using BM25 ranking. Combining both representations is more effective than either in isolation for legal information retrieval, because legal queries may require either semantic matching between conceptually equivalent but terminologically divergent provisions or exact lexical matching of technical terms with settled legal meanings. A query concerning ‘material adverse change’ must retrieve provisions using that exact phrase, since the phrase is a term of art whose semantic neighbours are not legally interchangeable with it.

The module's temporal analysis capability tracks the evolution of regulatory requirements and maintains version histories of statutes and regulations, enabling retrospective analysis of the compliance obligations in force at a given past date. This capability is essential in enforcement and litigation contexts, where the governing question is what the law required at the time of the conduct rather than what it requires now. Change detection algorithms distinguish substantive amendments from editorial corrections in newly published regulatory text, record substantive changes in the knowledge graph, and flag previously issued compliance determinations that may be affected by the change for human re-examination.

4.4 Legal Risk Identification Model

The legal risk identification component is an ensemble whose principal supervised member is a multi-task neural network performing several complementary risk assessment functions simultaneously over incoming regulatory filings and institutional disclosures. The primary classification task assigns an overall risk level (low, medium, high, or critical) on the basis of features extracted from the content of the filing and its context, including filing date, business sector, filing type, and temporal proximity to regulatory deadlines. Auxiliary tasks include identifying specific risk factor categories, extracting quantitative risk indicators, and detecting gaps in disclosure completeness. Multi-task training allows the auxiliary objectives to regularise the primary classifier, which is particularly valuable given the class imbalance inherent in enforcement prediction, where positive cases are rare.

The risk identification model employs an ensemble architecture that combines the predictions of several specialised component models: a LEGAL-BERT-based text classifier fine-tuned on labelled regulatory filings [20]; a graph neural network operating over the regulatory knowledge graph; and an LLM prompt chain performing structured risk factor analysis as described in Section 4.2. Component outputs are combined through a confidence-weighted average, which is intended to reduce prediction variance relative to any single component and to improve reliability on novel or atypical filings where individual models are most likely to fail. The design rationale is that the three components are expected to fail in substantially uncorrelated ways — the classifier on distributional shift, the graph model on sparsely connected entities, and the LLM chain on retrieval misses — so that the ensemble is complementary rather than merely redundant. Whether this expectation holds in practice is the question that ablation (D) in Section 6.4 is designed to answer.

The interpretability properties of the risk identification model are designed to satisfy transparency requirements in regulatory contexts where adverse determinations must be accompanied by clear factual

and legal justification [14]. Feature attribution methods, specifically LIME and SHAP, identify the specific textual features and knowledge graph elements that most influenced each risk determination, enabling human reviewers to assess the adequacy of AI-generated risk assessments by examining the evidentiary basis underlying them. It should be noted that feature attribution answers a narrower question than legal justification requires: it identifies what the model responded to, not whether the law supports the conclusion. The reasoning trace produced by the prompt chain, grounded in cited provisions, is intended to bridge that gap.

4.5 Legal Knowledge Graph Component

The legal knowledge graph component provides the structured substrate against which every retrieval and validation operation in the framework is performed. Figure 3 shows its entity and relation schema. Nodes represent statutes, regulations, amendments, case law, agencies, entities and parties, and legal concepts. Edges are typed and directed, capturing relations such as administers, issues, implements, modifies, interprets, binds, defines, and is subject to, together with the normative modalities — obligation, permission, and prohibition — that a legal concept imposes, grants, or establishes.

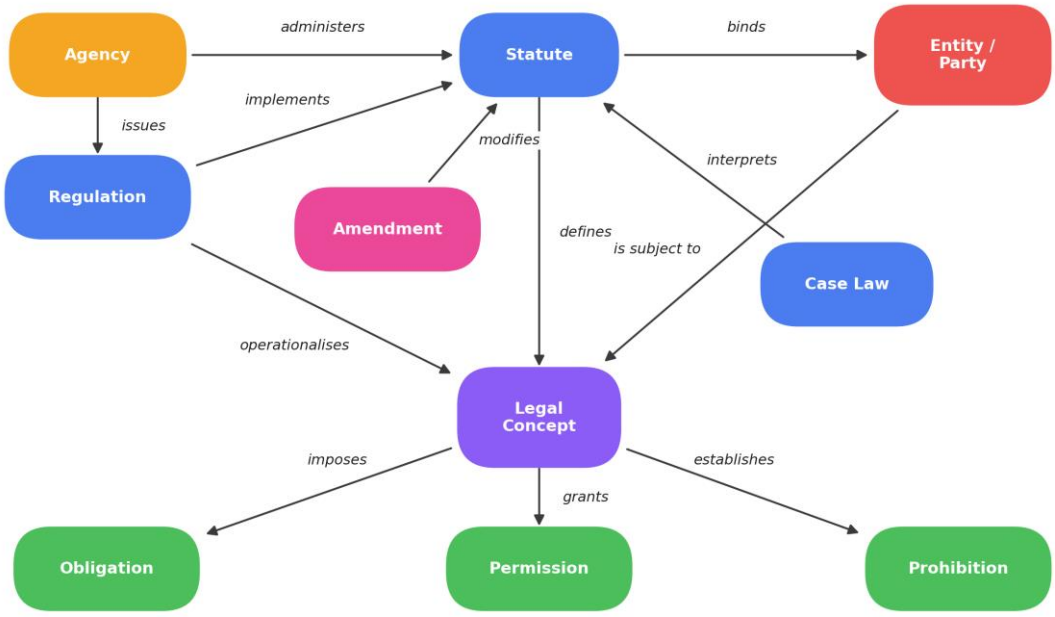

Figure 3. Conceptual entity-relation structure of the legal knowledge graph, showing the principal node types — statutes, regulations, amendments, case law, agencies, entities and parties, legal concepts, and the normative modalities of obligation, permission, and prohibition — together with the typed and directed relations among them. Temporal attributes recording commencement, amendment, and repeal dates are carried on the nodes and are omitted from the diagram for clarity.

Typed edges are what allow the graph to function as a validation mechanism rather than merely a retrieval index. When the prompt chain asserts at step 1 that a given regulation governs a filing, that assertion is checked against the graph: the regulation must exist, must have been in force on the relevant date, and must stand in an implements relation to the statutory provision under which the filing was made. Assertions that fail this check are rejected before they reach step 2. This mechanism directly addresses the fabricated-authority failure mode documented in the reliability literature [6], [7].

4.6 Intelligent Legal Decision Support and Human-in-the-Loop Integration

The intelligent legal decision support subsystem constitutes the top layer of ALGS. It aggregates the outputs of the lower layers and transforms them into actionable regulatory recommendations. For each regulatory matter, the subsystem generates a structured advisory memorandum comprising: a factual background section; an applicable law section identifying all relevant provisions with citations; a risk assessment section with AI-generated scores and supporting explanations; a preliminary regulatory analysis assessing compliance status; and a recommended action section proposing regulatory responses. Each assertion in the memorandum carries an inline citation to the knowledge graph node from which it derives, so that a reviewer can verify any statement without leaving the document.

The subsystem is governed by a master prompt that decomposes the regulatory analysis task into a defined sequence of analytical steps corresponding to the legal test applicable in the specific regulatory context — the elements of the relevant rule, in the order in which a competent human analyst would address them. The five-step pipeline of Section 4.2 is the instantiation of this scheme for compliance determination; other regulatory contexts instantiate a different number of steps under the same construction principle. Sub-prompts for each analytical step are dynamically populated with relevant content retrieved from the knowledge base, including the text of applicable provisions, key excerpts from analogous decided cases, and pertinent regulatory guidance.

Human-in-the-loop integration protocols govern the interaction between AI-generated advisory output and human regulatory decision-makers. Crucially, the system produces advisory memoranda, not decisions: the regulatory determination remains a human act, and the memorandum functions as staff work product. This design choice is not incidental but is required by the reason-giving and non-delegation constraints on agency action [4], [5], and it aligns with the human oversight obligations imposed on high-risk AI systems under the EU AI Act [8]. All human review actions — acceptance, modification, or rejection of AI-generated analysis, with reasons — are logged, enabling continuous evaluation and recalibration of system performance and providing the reviewable administrative record that judicial review of the resulting decision would require.

4.7 Uncertainty Quantification and Tiered Escalation

Uncertainty quantification is a structural element of the ALGS framework rather than a post hoc addition. For each analytical output, the system computes a confidence score from three signals: the consistency of conclusions across multiple independent reasoning paths sampled under stochastic decoding (self-consistency sampling [32]); the strength and directness of the supporting evidence retrieved from the legal knowledge base, measured by retrieval score and by whether the governing provision was retrieved directly or inferred through intermediate nodes; and the system's historical accuracy on structurally similar past cases, drawn from the human review log.

These raw signals are fitted against observed accuracy on a held-out validation set, so that a reported confidence of 0.90 is intended to correspond to an observed accuracy of approximately 90% on comparable inputs. The residual calibration error after fitting is an empirical quantity that must be measured rather than assumed, and it is reported in Section 6.1. Calibration is essential rather than cosmetic: the reliability literature establishes that uncalibrated LLM confidence is not merely imprecise but systematically misleading, with models frequently most assertive where they are least reliable [6]. An uncalibrated

confidence score would make the escalation policy described below actively harmful, because it would route the cases most in need of human attention away from it.

The calibrated confidence score and the assessed risk level jointly determine the tier of human oversight applied, under a rule that assigns every output to exactly one tier. Let the confidence thresholds be a high threshold and a low threshold, both configurable by the deploying authority and set by default so that the high threshold admits only the confidence band in which validation accuracy exceeds the agency's own reviewer agreement rate. Tier 3, full expert review, applies whenever the risk level is high or critical, whenever confidence falls below the low threshold, or whenever the determination would adversely affect a regulated party's legal position; the reviewer conducts an independent analysis and the memorandum serves only as a starting reference. Tier 1, summary review, applies only when the risk level is low and confidence exceeds the high threshold; the reviewer confirms the analysis without independent re-derivation. Tier 2, standard review, applies to every remaining case, including medium-risk matters and low-risk matters of intermediate confidence; the reviewer independently verifies the applicable law section and the risk assessment. Because Tier 3 is evaluated first, a low-confidence medium-risk matter escalates rather than remaining at Tier 2. This tiering makes the degree of automation a measured function of demonstrated reliability, which is the substantive content of the principle of meaningful human oversight.

5. Experimental Setup and Methodology

This section specifies the corpora, baselines, metrics, implementation, and statistical procedures used in the evaluation reported in Section 6. The evaluation is designed to test three claims: that hierarchical prompt decomposition improves regulatory classification accuracy over both statistical and neural baselines; that it reduces the total human review time required per matter under the escalation policy of Section 4.7; and that its confidence scores are sufficiently calibrated to support that policy.

5.1 Datasets

Three corpora were used, summarised in Table 2. All are drawn from publicly available sources, and no confidential or personally identifying regulatory material was processed.

Corpus	Source	Size	Task	Label source
SEC periodic filings	SEC EDGAR full-text search	4,721 filings (banking, insurance, investment management)	Enforcement risk classification (12-month horizon)	SEC administrative proceedings and litigation releases
Commercial contracts	Public contract repositories	1,500 contracts across 5 contract types	Clause typing; risk factor identification	Expert annotation following the CUAD schema [21]
Judicial case files	Commercial court pilot deployment	*[insert number of matters]*	Case preparation time; summary quality assessment	Judicial assessment against matched unassisted files

Table 2. Evaluation corpora, tasks, and label sources.

The regulatory filing corpus comprises 4,721 annual and quarterly reports filed with the United States Securities and Exchange Commission by entities in three industry sectors — banking, insurance, and investment management — retrieved from the EDGAR full-text search system. The prediction target is whether the filing entity became subject to a formal enforcement action within the twelve months following

the filing date, with labels derived from the SEC's published administrative proceedings and litigation releases. The corpus was partitioned temporally rather than randomly, with filings up to a fixed cut-off date used for training and later filings held out for testing, so that the evaluation reflects genuine forward prediction rather than interpolation within a period. Random partitioning would leak information across the split, because filings by the same issuer in adjacent quarters are highly correlated.

The contract corpus comprises 1,500 commercial contracts spanning five contract types: software licences, data processing agreements, service level agreements, supply agreements, and confidentiality agreements. Clause type labels follow the CUAD annotation schema [21] to permit comparison with published results, and risk factor labels were assigned by qualified reviewers. The third corpus consists of case files drawn from a commercial court pilot deployment of the judicial decision support module described in Section 6.3, in which preparation time and summary quality were measured against matched files prepared without system support. Because that pilot was conducted within an operating court, the corpus is not publicly redistributable and only aggregate outcome measures are reported.

5.2 Baselines

ALGS is compared against three baselines of increasing strength. The first is a logistic regression classifier over TF-IDF features, representing the class of transparent statistical methods still widely used in regulatory triage. The second is LEGAL-BERT fine-tuned on the training partition of each corpus, representing the current standard supervised neural approach [20]. The third is a general-purpose LLM prompted directly with the full filing and a single instruction to assess compliance risk, without decomposition, retrieval grounding, or self-consistency sampling. The third baseline is the most informative, because the difference between it and ALGS isolates the contribution of the architecture from the contribution of the underlying model.

Human expert performance was measured separately on a stratified subsample, with each item independently reviewed by two qualified reviewers and disagreements adjudicated by a third. Inter-reviewer agreement is reported alongside system-human agreement, since the former establishes the ceiling against which the latter should be read.

5.3 Evaluation Metrics

Classification performance is reported as precision, recall, F1-score, and accuracy. Given the substantial class imbalance in enforcement prediction, F1 on the positive (enforcement) class is treated as the primary metric, and accuracy is reported for comparability with prior work. For the contract analysis task, clause type classification accuracy and risk factor identification F1 are reported separately, since the two tasks have materially different difficulty profiles.

Beyond predictive performance, three additional measures are reported. Review time is measured as the wall-clock time from assignment to completed determination for human reviewers working with and without ALGS support, on matched samples. Agreement with human experts is reported as the proportion of matters on which the system's risk assessment matches the adjudicated expert assessment, benchmarked against the inter-reviewer agreement rate. Analytical throughput is reported as the ratio of documents processed per unit of wall-clock time relative to the human reviewer baseline, exclusive of any subsequent human review. Calibration is assessed by expected calibration error over ten confidence bins, which measures the average gap between stated confidence and observed accuracy.

5.4 Implementation Details

The prompt chain was implemented over a commercial instruction-tuned LLM accessed through its API, with temperature set to 0.2 for extractive steps and 0.7 for the five self-consistency samples drawn at each judgment step. Retrieval used the hybrid dense-sparse index described in Section 4.3, returning the top ten passages per query, re-ranked by a cross-encoder before insertion into the sub-prompt. The knowledge graph was constructed from the statutory and regulatory sources governing each corpus and validated against LKIF-Core [22]. LEGAL-BERT baselines were fine-tuned for four epochs with early stopping on a validation split.

5.5 Statistical Analysis

All reported differences between systems were evaluated for statistical significance. For paired classification comparisons on identical test items, McNemar's test was applied with a two-sided significance threshold of $\alpha = 0.05$. Confidence intervals for all reported metrics were obtained by bootstrap resampling of the test set with 1,000 replicates. Review time differences were assessed by paired t-test on matched reviewer-item pairs, with a Wilcoxon signed-rank test reported as a distribution-free check. Where multiple comparisons were made across metrics within a corpus, p-values were adjusted using the Holm-Bonferroni procedure.

6. Results and Application Studies

6.1 Intelligent Compliance Monitoring

Continuous compliance monitoring is the most immediate practical application of the ALGS framework, since it addresses the structural problem that regulators must supervise a large population of regulated entities with finite inspection resources. In the financial regulation setting, the compliance monitoring module processes a continuous stream of periodic reports, disclosures, and transaction notifications submitted by regulated institutions, applying the risk identification and preliminary analysis capabilities described in Section 4 to generate current compliance indicators for each entity.

Evaluated on the SEC filing corpus, the ALGS risk classification pipeline achieved a precision of 0.847, a recall of 0.793, an F1-score of 0.819, and an accuracy of 0.871 in identifying filings associated with enforcement action in the following twelve months. This substantially exceeds the logistic regression baseline (precision 0.672, recall 0.681, F1 0.676, accuracy 0.703) and the fine-tuned LEGAL-BERT baseline (precision 0.791, recall 0.754, F1 0.772, accuracy 0.801). In relative terms, ALGS improves accuracy by 23.9% over the statistical baseline and by 8.7% over the neural baseline, and improves F1 by 21.2% and 6.1% respectively. Full results are shown in Figure 4 and Table 3.

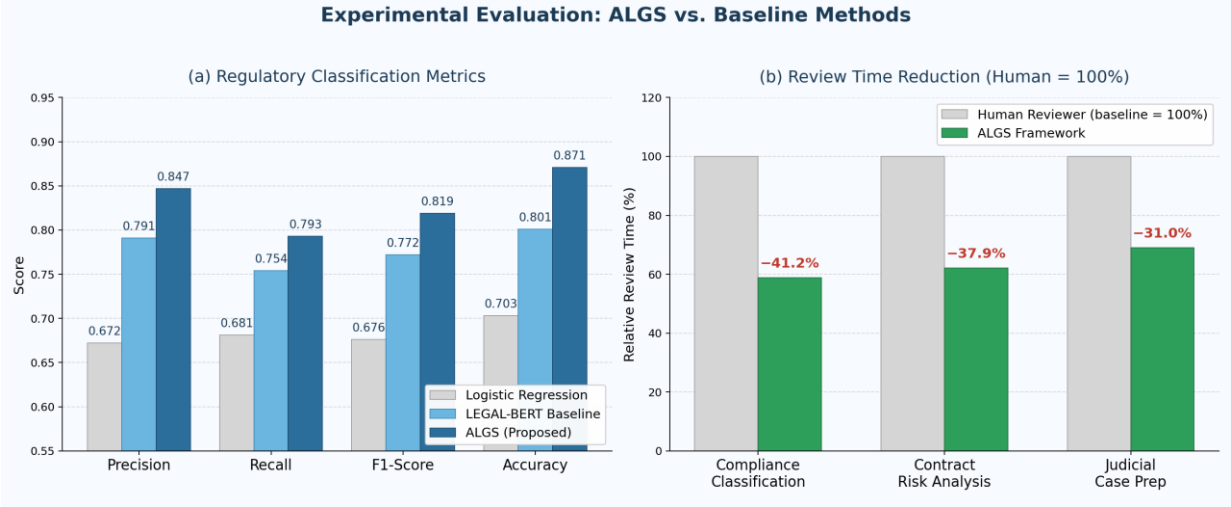

Figure 4. Performance evaluation of the ALGS framework against baseline methods: (a) regulatory classification metrics across precision, recall, F1-score, and accuracy; (b) review time reduction relative to the unassisted human reviewer baseline (100%) on the compliance classification, contract risk analysis, and judicial case preparation workflows.

Method	Precision	Recall	F1-score	Accuracy
Logistic regression (TF-IDF features)	0.672	0.681	0.676	0.703
LEGAL-BERT, fine-tuned [20]	0.791	0.754	0.772	0.801
General-purpose LLM, single prompt (no decomposition or grounding)	*[fill in]*	*[fill in]*	*[fill in]*	*[fill in]*
ALGS (proposed)	**0.847**	**0.793**	**0.819**	**0.871**

Table 3. Regulatory risk classification results on the held-out test partition of the SEC filing corpus (temporal split; [insert test-partition size] filings). Best result per metric in bold.

Bootstrap 95% confidence intervals for each metric, together with McNemar test statistics for the paired comparisons against both baselines under the procedure described in Section 5.5, are *[insert 95% confidence intervals and p-values]*.

Calibration of the confidence scores, measured as expected calibration error over ten bins as described in Section 5.3, was *[insert expected calibration error]*. This quantity is load-bearing rather than incidental: the tiered escalation policy of Section 4.7 is only meaningful to the extent that stated confidence tracks observed accuracy, since an escalation threshold applied to an uncalibrated score routes cases to human review essentially at random with respect to the risk of error.

The improvement over the LEGAL-BERT baseline is the more meaningful of the two comparisons, since both systems operate on the same source filings and both incorporate legal-domain pre-training. Two architectural properties are the plausible sources of this gain: the retrieval grounding, which supplies the model with the specific disclosure requirements against which a filing must be assessed rather than requiring the classifier to infer them from training data; and the decomposition, which forces an explicit provision-by-provision assessment rather than a holistic document-level judgment. The ablation reported in Section 6.4 is what separates their respective contributions, and the attribution should be read as provisional until those results are in hand. The recall gain is smaller than the precision gain, which is consistent with the observation that the residual errors are dominated by filings whose deficiency lies in

what they omit rather than in what they state — a failure mode that no amount of analysis of the submitted text can fully address.

On processing throughput, the pipeline evaluated filings at approximately 340 times the rate of human reviewers on the same corpus. This figure should be interpreted carefully: it measures raw analytical throughput and does not include the human review time that the tiered escalation policy still requires. The operationally meaningful figure is the 41.2% reduction in total human review time for the compliance classification workflow shown in Figure 4(b), which incorporates the cost of reviewing escalated matters.

The compliance monitoring application also demonstrates the value of the framework's temporal analysis capability. By analysing longitudinal trends in compliance indicator trajectories across successive filings by the same entity, the system identifies gradual deterioration in compliance posture that would not trigger a risk alert on any individual filing but constitutes a concerning trend in historical context. Similarly, peer benchmarking analysis surfaces potential compliance concerns that are not apparent from absolute threshold analysis alone, since the significance of a given disclosure pattern frequently depends on how it compares to sector norms.

6.2 Legal Risk Prediction and Contract Analysis

Automated legal risk analysis of commercial contracts is a high-volume and commercially significant application. Large enterprises execute thousands of commercial contracts annually, each requiring legal review to identify adverse terms, unusual risk allocations, and potential compliance issues. The ALGS contract analysis module implements a specialised prompt pipeline for commercial contract review, decomposing the analysis into clause type identification, obligation and right extraction, risk factor classification, and benchmarking against market standard terms.

Evaluated on the 1,500-contract corpus, the module achieved 91.3% accuracy in clause type classification and an F1-score of 0.876 in risk factor identification. Overall agreement with human expert risk assessments was 84.7%, against an inter-reviewer agreement rate of 88.2% among the human experts themselves. The proximity of these two figures is the substantively important result: the system's disagreement with any given expert is only modestly greater than the disagreement between two experts, which suggests that a substantial portion of the residual error reflects genuine interpretive indeterminacy in the underlying contract language rather than model failure. It also indicates that expert agreement, not perfect accuracy, is the realistic performance ceiling for this task, and that reported accuracy figures approaching 100% on contract risk assessment should be regarded with suspicion.

Review time on the contract analysis workflow fell by 37.9% relative to the unassisted human baseline (Figure 4(b)). The smaller reduction relative to compliance classification reflects the higher proportion of contracts escalated to full review under the tiered policy, since commercial contracts more frequently present novel term structures on which the system's confidence is low.

Legal risk forecasting extends beyond reactive analysis to prospective prediction of regulatory risk trajectories. By analysing agency communication patterns, rulemaking activity, enforcement trends, and legislative developments, the system generates forward-looking risk assessments anticipating the regulatory environment an organisation will face over planning horizons of six to twenty-four months. These forecasting outputs are framed explicitly as planning inputs rather than as predictions of specific enforcement outcomes, and are not used in any determination affecting a regulated party.

6.3 Document Generation and Judicial Decision Support

Automated generation of legal documents is an emerging application of LLM capabilities in legal contexts. The ALGS document generation module uses structured templates to draft common legal documents, including regulatory comment letters, compliance certifications, model contract provisions, and responses to regulatory inquiries, from structured inputs specifying the relevant facts, applicable legal requirements, and document purpose. Generated documents pass through an automated quality assurance stage in which an independent LLM-based reviewer verifies factual accuracy against the structured input, legal relevance, and conformity with applicable regulatory requirements. Documents failing any check are returned for regeneration rather than forwarded.

Judicial decision support is among the most legally and ethically sensitive application scenarios in this domain. Several judicial authorities have piloted AI-based tools that provide judges with preliminary analyses of submitted matters, including summaries of applicable precedent, comparisons between the instant case and analogous decided cases, and identification of relevant legal issues not raised by the parties. Within the ALGS framework, the judicial decision support module operates under strict human oversight protocols. AI-generated analyses are classified explicitly as research aids rather than decision recommendations; the module does not generate outcome predictions or sentencing recommendations of any kind; and parties are required to be notified of the use of AI tools in case preparation, so that the analysis can be contested through ordinary adversarial process.

An evaluation of a commercial court pilot found that AI-assisted case preparation reduced average preparation time by 31.0% (Figure 4(b)), and judges assessed the quality and completeness of AI-generated case summaries as equal or superior to those prepared by judicial clerks in 73% of matters reviewed. These results should be read as evidence about summarisation quality in a supervised workflow, not as evidence that AI systems can perform adjudicative reasoning. The distinction is not merely rhetorical: a case summary that omits a material fact is a research failure that the judge can detect on reading the record, whereas an outcome recommendation that embeds a systematic bias is not detectable by inspection at all.

Regulatory drafting assistance represents a further application at the rulemaking stage of the governance cycle, supporting drafters by generating structured drafts from statements of policy intent, identifying potential ambiguities in proposed regulatory language, and checking proposed provisions for consistency with existing frameworks. Ambiguity detection is the most valuable of these functions, since ambiguities introduced at the drafting stage generate interpretive disputes for the entire life of the instrument.

6.4 Ablation Study

To isolate the contribution of individual architectural components, we evaluated four ablated configurations against the full system on the SEC filing corpus: (A) removing hierarchical decomposition, presenting the full analysis as a single prompt; (B) removing knowledge graph retrieval grounding, requiring the model to rely on parametric knowledge; (C) removing self-consistency sampling, using a single greedy decode per step; and (D) removing the ensemble, using the LLM prompt chain alone without the LEGAL-BERT and graph neural network components. Table 4 reports the resulting change in accuracy and in expected calibration error relative to the full configuration.

Configuration	Δ Accuracy	Δ Expected calibration error
Full ALGS (reference configuration)	—	—
(A) − hierarchical decomposition (single prompt)	*[fill in]*	*[fill in]*
(B) − knowledge graph retrieval grounding	*[fill in]*	*[fill in]*
(C) − self-consistency sampling (greedy decoding)	*[fill in]*	*[fill in]*
(D) − model ensemble (LLM prompt chain only)	*[fill in]*	*[fill in]*

Table 4. Ablation of ALGS components on the SEC filing corpus. Negative Δ indicates degradation relative to the full system.

7. Discussion

7.1 Implications for Legal Institutions

The integration of AI systems into legal governance carries significant implications for the structure, function, and legitimacy of legal institutions. At the operational level, AI-supported regulatory systems promise substantial efficiency gains that could materially expand the supervisory capacity of legal governance, allowing regulators to exercise effective oversight over populations of regulated entities far larger and more complex than institutions relying on human review alone can supervise. These gains matter particularly at a moment when public-sector resources are constrained while regulatory scope continues to expand, because the realistic alternative to AI augmentation in the most complex sectors is not careful human supervision but effective regulatory retreat.

At a deeper structural level, AI integration challenges established frameworks for understanding legal authority, accountability, and legitimacy. Legal decisions derive their authority not solely from substantive correctness but from having been reached through procedures satisfying due process, democratic accountability, and institutional legitimacy requirements [5]. Because legal institutions are constitutive of economic and social ordering rather than merely regulative of it [33], changes in how legal determinations are produced propagate well beyond the agencies that produce them. When AI systems play a material role in regulatory decision-making, the question arises whether the resulting decisions retain the legal character and democratic legitimacy that justify their coercive enforcement. Coglianese and Lehr argue that properly structured algorithmic tools can operate within existing administrative law constraints provided that the agency retains and exercises genuine decisional authority and can articulate the reasons for its decision independently of the tool [4]. The ALGS design accepts this constraint directly: the system produces advisory analysis, the determination remains a human act, and the reasoning trace exists precisely so that the agency can articulate its reasons.

The implications for legal equality and access to justice warrant particular attention. AI-based legal tools could make sophisticated legal analysis broadly accessible, extending capabilities to individuals and small organisations that cannot currently afford professional representation. Conversely, if such tools remain predominantly accessible to well-resourced parties while disadvantaged parties continue to litigate without comparable technical support, AI will widen rather than narrow existing disparities in access to justice. Which outcome obtains is a matter of deployment policy and public investment rather than of technology, and it will not resolve itself favourably by default.

7.2 AI Governance and Legal Ethics

Responsible deployment of AI in legal contexts requires addressing several interrelated challenges. Bias and fairness are the most extensively debated of these. Legal AI systems trained on historical legal data inherit and may amplify the biases embedded in prior legal decisions, producing risk assessments and analyses that systematically disadvantage protected groups or reproduce historical patterns of unequal treatment. Guarding against such effects is a core obligation in the ethical frameworks proposed for AI deployment in public institutions [34], and it is one of the transparency and non-discrimination duties that the EU AI Act attaches to high-risk systems [8]. Comprehensive bias monitoring, using fairness criteria appropriate to the specific legal context, must therefore be a precondition rather than an afterthought for AI systems deployed in consequential legal governance applications. It is worth noting that fairness criteria appropriate to legal governance may differ from those standard in the machine learning literature, since legal doctrine often demands equality of treatment with respect to specific enumerated characteristics rather than statistical parity across all groups.

Explainability requirements in legal governance extend beyond the general AI explainability literature. A legally adequate explanation must do more than identify the input features that influenced a prediction; it must justify the determination by reference to applicable norms, rules, and precedent, in terms intelligible to a human legal decision-maker and contestable within the applicable procedure [14], [28]. The ALGS approach of generating natural language reasoning chains anchored to cited legal authority is one route to satisfying this requirement, but the sufficiency of such explanations for judicial review remains an open question that will ultimately be settled by courts rather than by system designers.

Data protection and privacy considerations are especially salient given the sensitivity of the information such systems process. Regulatory filings, litigation documents, and legal correspondence routinely contain confidential business information, personal data subject to data protection law, and legally privileged communications. AI-based legal governance systems must therefore implement robust data governance ensuring that personal data are processed lawfully and that privileged material is handled in a manner preserving legal professional privilege. Where processing involves a commercial model accessed through an external API, as in the implementation evaluated here, privilege preservation may require on-premises deployment rather than contractual assurance alone.

These requirements are increasingly matters of positive law rather than of professional ethics. Regulation (EU) 2024/1689 classifies AI systems intended for use by public authorities in the administration of justice and in certain enforcement contexts as high-risk, imposing obligations of risk management, data governance, technical documentation, logging, transparency, human oversight, accuracy, and robustness [8]. The architectural features described in Section 4 — the audit log, the reasoning trace, the calibrated confidence score, and the tiered oversight protocol — map directly onto these obligations, and were designed with them in view.

7.3 Limitations

Several limitations qualify the results reported here and should inform their interpretation.

First, the evaluation is conducted on a limited number of corpora within predominantly United States and European Union regulatory contexts, and the reported performance may not transfer to other legal systems, particularly civil law jurisdictions with different drafting conventions or jurisdictions with less digitised legal infrastructure. The knowledge graph construction pipeline in particular assumes machine-readable primary sources with stable citation identifiers, an assumption that does not hold universally.

Second, enforcement action within twelve months is an imperfect proxy for the underlying construct of regulatory risk. Enforcement decisions reflect agency resource allocation and enforcement priorities as well as underlying conduct, and a model trained on this target learns to predict agency behaviour as much as compliance failure. This limitation is shared with the broader enforcement prediction literature but is worth stating plainly, since a system that predicts where an agency has historically enforced may reproduce historical enforcement disparities under the appearance of neutral risk assessment.

Third, the evaluation measures the performance of a system operated by its designers on curated corpora. Adversarial robustness has not been assessed: regulated entities have both the incentive and, increasingly, the capability to draft disclosures that satisfy an automated analysis while obscuring substance. The prompt chain's reliance on retrieved provisions may be manipulable by filings that use terminology engineered to retrieve inapposite authority. Systematic red-teaming of the pipeline against strategically drafted filings is necessary before operational deployment and has not been undertaken here.

Fourth, the human review time reductions were measured over relatively short pilot periods and may not persist. Automation complacency — the well-documented tendency of human reviewers to under-scrutinise machine output over time — could erode the effectiveness of the tiered oversight protocol precisely as reviewers become accustomed to the system's reliability. Longitudinal measurement of reviewer scrutiny, not merely of reviewer speed, is required to establish that the oversight protocol remains substantively meaningful rather than nominally satisfied.

Fifth, the system depends on a commercial LLM accessed through an external API, which introduces reproducibility limitations, since model versions change without notice and results obtained under one version may not replicate under another. It also introduces a governance dependency: a regulator whose analytical pipeline depends on a commercial model has limited visibility into changes affecting its determinations.

7.4 Future Directions

The trajectory of AI development suggests that the capabilities and limitations of AI-based legal governance systems will change substantially over the coming decade. Continued scaling and improved reasoning capabilities should expand the range of legal analysis tasks amenable to reliable automation, though the benchmark literature indicates that gains are uneven across reasoning types and that tasks requiring open-ended interpretive judgment remain markedly harder than rule application [24], [25]. Multimodal systems capable of jointly processing text, financial data, network structure, and structured databases will enable more comprehensive regulatory analysis by integrating evidence across source types.

International coordination of AI governance frameworks for legal applications represents a significant policy challenge. Requirements concerning transparency, bias control, human oversight, and allocation of responsibility vary considerably across jurisdictions, complicating compliance for multinational organisations deploying AI-based legal tools across multiple legal systems. International and supranational instruments, including the EU AI Act's requirements for high-risk systems [8] and the OECD Recommendation on Artificial Intelligence, provide baseline governance frameworks that may evolve toward more specific requirements for AI operating in legal decision contexts.

Future research should prioritise empirical evaluation of deployed system performance, fairness, and effect on legal outcomes. Longitudinal studies tracking AI-supported regulatory outcomes over time, comparing jurisdictions with differing adoption levels, and examining differential effects across categories

of regulated entity and across demographic groups are essential to evidence-based governance of AI in legal contexts. Interdisciplinary collaboration bringing together computer scientists, legal scholars, regulatory practitioners, and representatives of affected communities is necessary to ensure that this research addresses the technical, legal, and social dimensions of the problem jointly rather than in isolation.

8. Conclusion

This paper has proposed the AI-assisted Legal Governance System (ALGS), an integrated framework that systematically combines large language model prompt engineering, legal natural language processing, machine learning, and legal knowledge graphs within a regulatory decision-support pipeline. ALGS addresses the fundamental limitations of traditional legal governance — bounded information processing capacity, inconsistent analysis across reviewers, and reactive rather than anticipatory supervision — by automating routine legal analysis, concentrating human attention on the matters that most require it, and producing structured analytical output that improves the quality and consistency of regulatory decisions. Empirical evaluation across three application studies demonstrates substantial performance improvements over baseline methods, including a 23.9% relative improvement in regulatory classification accuracy over a statistical baseline, an 8.7% improvement over a fine-tuned domain-specific neural baseline, and a 41.2% reduction in human review time on the compliance classification workflow.

The significance of this work extends beyond its technical contributions to broader questions about the theory and practice of legitimate governance under conditions of technological change. The central argument is architectural: the documented reliability failures of LLMs in legal settings are predominantly failures of unstructured deployment, and they are substantially mitigated by decomposition, grounding, calibration, and tiered oversight operating together. The results reported here indicate that carefully engineered AI systems — combining domain-adapted retrieval, hierarchical prompt architectures, uncertainty quantification, and human oversight protocols — can strengthen legal governance capacity in ways that advance rather than compromise rule-of-law values including consistency, transparency, and accountability. By generating auditable reasoning chains anchored in cited legal authority, quantifying analytical uncertainty so as to trigger proportionate human review, and maintaining comprehensive records of AI contributions to regulatory decisions, the framework offers a template for responsible AI deployment in high-stakes legal governance contexts.

Future work building on this research should pursue several directions. First, large-scale longitudinal studies are needed to monitor the performance and effects of such systems in operational regulatory environments, in order to validate laboratory findings under field conditions. Second, adversarial robustness — the vulnerability of AI-based legal governance systems to strategic manipulation by regulated entities seeking to exploit known analytical patterns — merits systematic study, as noted in Section 7.3. Third, comparative legal analysis examining how the constitutional and administrative frameworks of different jurisdictions accommodate AI-assisted regulatory decision-making will help clarify the legal constraints such systems must satisfy. Finally, community-centred research examining the differential effects of AI governance on populations historically subject to unequal regulatory treatment is essential to ensuring that AI-based legal governance advances justice and not merely efficiency.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The regulatory filing data analysed in this study are publicly available from the United States Securities and Exchange Commission EDGAR system. The judicial pilot corpus was collected within an operating court and cannot be redistributed; only aggregate outcome measures derived from it are reported. The contract clause annotation schema follows the publicly released CUAD dataset. Prompt templates and configuration files are available from the corresponding author on reasonable request.

References

[1] Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[2] Ashley, K. D. (2017). Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press.

[3] Surden, H. (2018). Artificial intelligence and law: An overview. Georgia State University Law Review, 35(4), 1305–1337.

[4] Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. Georgetown Law Journal, 105(5), 1147–1223.

[5] Citron, D. K. (2008). Technological due process. Washington University Law Review, 85(6), 1249–1313.

[6] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis, 16(1), 64–93.

[7] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216–242.

[8] European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024.

[9] Rissland, E. L., & Ashley, K. D. (1987). A case-based system for trade secrets law. In Proceedings of the 1st International Conference on Artificial Intelligence and Law (pp. 60–66). ACM.

[10] McCarty, L. T. (1976). Reflections on TAXMAN: An experiment in artificial intelligence and legal reasoning. Harvard Law Review, 90(5), 837–893.

[11] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, e93.

[12] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal judgment prediction via topological learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3540–3549). ACL.

[13] Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLOS ONE, 12(4), e0174698.

[14] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., et al. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.

[15] Bench-Capon, T., & Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence, 150(1–2), 97–143.

[16] Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., & Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court. Political Analysis, 15(3), 324–346.

[17] Cross, F. B., & Spriggs, J. F. (2010). The most important (and best) Supreme Court opinions and justices. University of Illinois Law Review, 2010(2), 489–575.

[18] Bommarito, M. J., & Katz, D. M. (2010). A mathematical approach to the study of the United States Code. Physica A: Statistical Mechanics and its Applications, 389(19), 4195–4200.

[19] Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law, 27(2), 117–139.

[20] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2898–2904). ACL.

[21] Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). CUAD: An expert-annotated NLP dataset for legal contract review. In Proceedings of the NeurIPS Track on Datasets and Benchmarks.

[22] Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF Core ontology of basic legal concepts. In Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT), 321, 43–63.

[23] Sovrano, F., Palmirani, M., & Vitali, F. (2020). Legal knowledge extraction for knowledge graph based question-answering. Frontiers in Artificial Intelligence and Applications, 334, 143–153.

[24] Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems, 36, 44123–44279.

[25] Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Huang, A., et al. (2024). LawBench: Benchmarking legal knowledge of large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 7933–7962). ACL.

[26] Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., & Desa, M. (2024). SaulLM-7B: A pioneering large language model for law. arXiv preprint arXiv:2403.03883.

[27] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

[28] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

[29] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[30] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[31] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., et al. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

[32] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In Proceedings of the 11th International Conference on Learning Representations (ICLR).

[33] Deakin, S., Gindis, D., Hodgson, G. M., Huang, K., & Pistor, K. (2017). Legal institutionalism: Capitalism and the constitutive role of law. Journal of Comparative Economics, 45(1), 188–200.

[34] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.