Legal-ISA: A Modular Framework for Systematic Legal AI Evaluation

Daisong Gong

Nankai University

Abstract

We introduce Legal-ISA, a modular integration framework that addresses this by systematically composing mature techniques—retrieval, verification, and reasoning—via standardized interfaces, inspired by the Instruction Set Architecture principle. The framework defines a comprehensive set of operations covering the majority of tasks within major legal benchmarks, enabling mandatory provenance tracking and fine-grained failure attribution. Comprehensive evaluation across four diverse legal benchmarks spanning multiple jurisdictions validates Legal-ISA's design, demonstrating: True Modularity achieved through configuration-only substitution across multiple component combinations; Systematic Attribution achieving significantly higher error attribution coverage than pure neural baselines; Quantified Transparency evidenced by low calibration error for reliable uncertainty assessment; and Performance Gains showing substantial improvement over the best Retrieval-Augmented Generation baseline. However, cross-jurisdictional tests revealed a performance drop, exposing a critical reliance on manual legal knowledge engineering for concept mapping; this dependency precludes direct comparison with automated model-learning approaches. Our core contribution is engineering-focused: a systematically validated, modular architecture for standardized legal AI evaluation and comparative risk assessment via composition and human knowledge integration.

Introduction

Current legal AI research confronts two primary limitations that curtail its impact on the broader legal community and scholarship. The chief issue is **horizontal fragmentation**: systems specialize in narrow domains such as contract analysis, case retrieval, or compliance checking, echoing human legal practice (Fei et al. 2023; Guha et al. 2023). Benchmarks expose these shortcomings: LEXam (Guha, Nyarko, and Ho 2025) reveals failures in process-based reasoning on exams; MSLR (Zhang, Li, and Wang 2025) highlights weaknesses in multi-step IRAC argumentation; and LeCoDe (Wang, Zhang, and Chen 2025) identifies gaps in clarification and advice. While specialization offers benefits, it hinders unified, holistic reasoning crucial for comprehensive legal research.

Second, **vertical fragmentation** (jurisdictional limitations) further constrains the field. Most legal AI systems are limited to specific jurisdictions, failing to develop fundamental reasoning that transcends legal traditions (Wang et al. 2024). Multilingual benchmarks like LEXam (Guha, Nyarko, and Ho 2025) (Swiss, international, generic law in English/German) and J1-ENVS (Liu, Wu, and Chen 2025) (dynamic Chinese scenarios) quantify performance degradation across jurisdictions, exposing failures in cross-jurisdictional concept alignment. This limits utility for cross-jurisdictional practice and prevents legal AI from serving as an abstract research instrument.

Legal-ISA is introduced as an integration layer that systematically composes existing retrieval, verification, and reasoning techniques through standardized operation interfaces. Drawing inspiration from computer architecture's interface-implementation separation, the framework defines computational abstractions that function as an instruction-level specification layer, analogous to an Instruction Set Architecture (ISA). This operation-level abstraction ensures component substitutability and enables systematic evaluation across diverse legal contexts. By providing standardized interfaces, Legal-ISA allows various jurisdiction-specific reasoning components to be integrated and compared systematically for comparative risk assessment.

Contributions. Legal-ISA's modularity and systematic design provide these key contributions.

1. **Standardized Framework Benchmark.** For AI researchers, we offer an integration framework showing how standardized interfaces support systematic component evaluation and reliability improvements over neural baselines. Legal-ISA serves as a benchmark, enabling component refinements without full agent construction.
2. **Systematic Knowledge Integration.** For legal informatics, we confirm the framework's value through ablation studies in diverse scenarios, illustrating integration of human knowledge via interfaces for cross-jurisdictional mapping, surpassing mere model learning.
3. **Enabling Human-in-the-Loop Auditing.** Methodologically, Legal-ISA facilitates comparing reasoning methods through transparent modular technique composition. This shifts human roles in Legal-AGI to auditors and value calibrators, with reasoning chains and uncertainty

Proceedings of Bridge between Artificial Intelligence and Law 2026 (Full papers), pages 1–8
Januray 21, 2026, AAAI 2026 Bridge Program, Singapore

outputs for intervention in public order and judicial areas.

Related Work

Empirical Benchmarking and Functional Fragmentation. The limitations of current legal AI systems are consistently demonstrated in recent empirical benchmarks, reporting error rates of 20–30% in complex legal reasoning tasks (Fei et al. 2023; Guha et al. 2023). Benchmarks such as LawBench (Fei et al. 2023) reveal severe functional fragmentation: systems excelling in narrow domains (e.g., case retrieval) frequently fail when integrated reasoning is required. LegalBench (Guha et al. 2023) similarly uncovered systematic capability gaps in large language models (LLMs) across critical issue-spotting categories using its diverse tasks. Recent benchmarks extend these findings: J1-ENVS (Liu, Wu, and Chen 2025) shows limited performance in dynamic legal agent scenarios, MSLR (Zhang, Li, and Wang 2025) exposes coherence failures in multi-step IRAC reasoning, and GreekBarBench (Papadopoulos and Nikolaou 2025) highlights free-text reasoning challenges in specialized legal systems. These pervasive failures—also evident in commercial systems like Westlaw Precision AI and Lexis+ AI (Thomson Reuters 2024; LexisNexis 2024)—highlight the urgent need for a new architectural paradigm that enables systematic evaluation, standardized performance attribution, and component-level diagnosis (Wang et al. 2024). Supporting this view, Wang et al. (Wang et al. 2024) identify three major shortcomings in current evaluation: the absence of cross-task generalization metrics, inadequate failure attribution mechanisms, and the lack of standardized protocols for hybrid system assessment—all of which motivate the design principles underlying Legal-ISA.

The Composability Challenge in Hybrid Legal AI. A complementary research direction enhances reliability through neuro-symbolic and hybrid approaches that combine statistical reasoning with symbolic verification. Recent frameworks promote structured knowledge representation and formalized verification protocols (Chen, Wang, and Zhang 2025; Kumar, Singh, and Patel 2025). Empirical studies consistently confirm the effectiveness of symbolic constraints: modular interfaces combining LLM processing with logical verification achieve precision gains in contract analysis (Chen, Wang, and Zhang 2025), and trainable logical reasoners enhance legal LLM responses (Kumar, Singh, and Patel 2025). Similar benefits are demonstrated when logic rules are integrated into case retrieval (Ma, Nguyen, and May 2024; Stranieri et al. 1999). Despite these advances, existing hybrid implementations (Wu et al. 2024a,b) remain largely task-specific and lack standardized interfaces for systematic component composition. This architectural rigidity prevents systematic comparison and substitution across diverse systems. Legal-ISA addresses this gap through operation-level abstraction, enabling modular neuro-symbolic design and directly tackling the core composability and evaluation issues identified in prior work.

Cross-Jurisdictional Reasoning: A Deficiency in Abstraction. Current legal AI systems are predominantly jurisdiction-bound, focusing on individual legal traditions rather than abstracting shared conceptual structures. Empirical studies consistently demonstrate substantial performance degradation when large language models (LLMs) attempt cross-jurisdictional concept alignment across multiple legal systems. This specialization–transferability trade-off is a recurring challenge in studies of unified retrieval (Li et al. 2025) and evaluation frameworks (Guo et al. 2024). Legal-ISA mitigates these limitations through operation-level abstraction. Specifically, abstract operations establish stable semantic interfaces that facilitate the systematic integration of expert-curated knowledge via a Jurisdiction Translator module, thus maintaining jurisdictional specificity while enabling crucial cross-system consistency.

Architectural Limitations and the Mandate for Modularity. Recent comprehensive surveys document the rapid evolution and persistent limitations of legal AI systems (Zhou, Liu, and Chen 2025; Yang, Zhang, and Wang 2024; Li, Chen, and Wang 2024). These reviews identify common architectural challenges: over 16 legal LLMs and 47 frameworks (Zhou, Liu, and Chen 2025) exhibit monolithic designs susceptible to hallucination, while systematic analyses (Li, Chen, and Wang 2024) reveal limitations in data quality, algorithmic transparency, and multimodal integration.The limitations above are compounded by the dominance of monolithic architectures that rely primarily on prompting (susceptible to hallucination) or simple Retrieval-Augmented Generation (RAG) pipelines (lacking logical consistency) (Chen et al. 2024). Survey data from Chen et al. (Chen et al. 2024) confirm that the absence of standardized interfaces remains the chief obstacle to systematic benchmarking (Wang et al. 2024). While recent initiatives advocate integrating legal constraints into generation, they fall short of specifying operational mechanisms that enable true modularity. Legal-ISA bridges this gap by introducing typed operation specifications with mandatory provenance tracking. By adopting the computer-architecture principle of interface–implementation separation, our framework provides the foundation for component-level auditability, cross-jurisdictional adaptability, and regulatory compliance—capabilities essential for the next generation of trustworthy legal AI.

The Legal-ISA Framework: Integration Architecture

The Legal-ISA Framework introduces a novel integration architecture designed to address the systemic challenges of functional fragmentation and reliability in legal AI systems. It achieves this by systematically integrating established techniques—such as Retrieval-Augmented Generation (RAG), symbolic verification, and specialized neuro-symbolic reasoning—through standardized operation interfaces. The framework features a six-layer architecture, illustrated in Figure 1, which ensures trustworthiness and auditability via three core mechanisms: 1) operation-level interfaces for systematic component evaluation and substitution; 2) a neuro-symbolic verification loop that ensures reliability and logical consistency; 3) mandatory provenance tracking established through a provenance-first protocol that

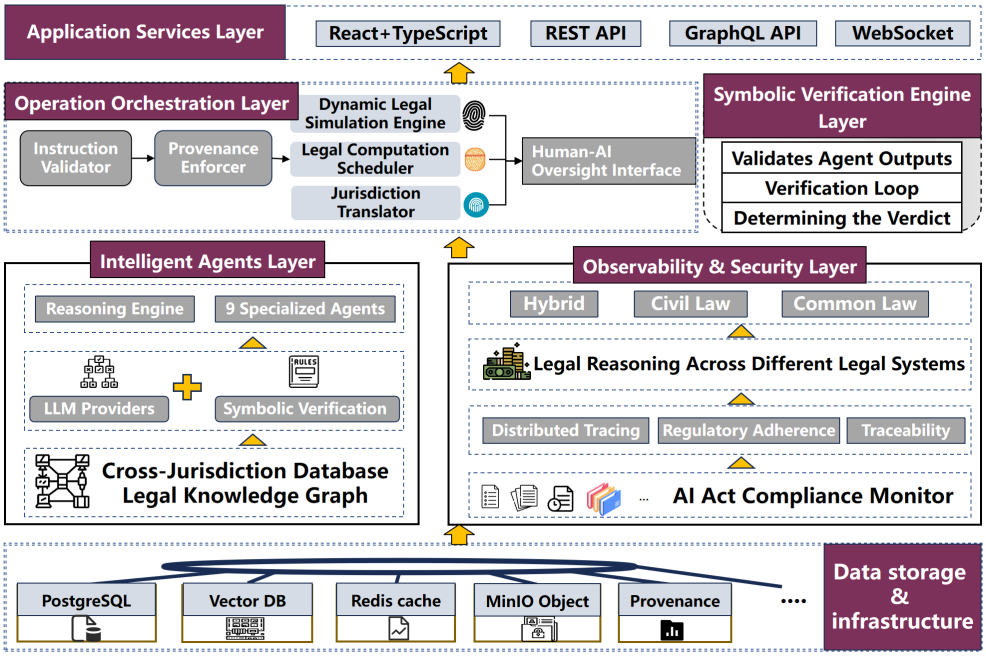

Figure 1: Six-layer architecture of Legal-ISA: Application Services, Operation Orchestration, Intelligent Agents, Symbolic Verification Engine, Observability/Security, and Data Storage layers.

captures an immutable audit trace prior to output delivery.

Operation Orchestration Layer The Operation Orchestration Layer functions as the framework's control plane, coordinating five core components to manage the execution and validation of all computational requests. 1) The Operation Validator ensures that incoming requests strictly adhere to the defined set of 25 standardized legal operations (e.g., `RETRIEVE_CASES`, `VERIFY_PROVENANCE`), enforcing type and format consistency. 2) The Provenance Enforcer implements the mandatory "provenance-first" protocol, ensuring comprehensive execution metadata is captured and logged to the immutable Provenance Ledger before any computation is initiated. 3) The Legal Computation Scheduler manages execution flow, dynamically routing validated operations to specialized agents based on capability declarations and real-time performance metrics. 4) The Jurisdiction Translator is crucial for cross-jurisdictional adaptation, augmenting orchestration by maintaining a semantic knowledge base and mappings across diverse legal traditions. 5) The Human-AI Oversight Interface implements a critical safety mechanism by escalating cases where confidence scores fall below pre-configured thresholds, routing them to a `HumanLoopAgent` for expert adjudication according to defined oversight policies. Critically, the Neuro-Symbolic Verification Loop (Section 4.1) is integrated within this layer to validate agent outputs against a set of formalized legal constraints.

Intelligent Agents and Verification Loop The Intelligent Agents Layer deploys specialized agents responsible for executing the Legal-ISA operations via a mandatory verification-correction pipeline. Each agent adheres to three critical `LegalAgent` interface contracts (*Functional*, *Metrics*, and *Provenance*) and utilizes shared resources, including the Legal Knowledge Graph and the Cross-Jurisdiction Database. Agents submit candidate outputs to the core Symbolic Verification Engine, which validates them against a knowledge base of 247 pre-defined legal constraints, enforcing rules such as Entity Consistency, Temporal Coherence, and Domain Constraints. This validation is integrated into a Neuro-Symbolic Verification Loop that yields one of three verdicts for subsequent action: *Accept* (output satisfies all constraints), *Review* (partial violations, triggering agent retries with corrective feedback), or *Reject* (critical failures, escalating the case to a `HumanLoopAgent`). For dynamic reliability management, unattributable claims incur a confidence penalty proportional to the ratio of unverified to total claims, calculated as:

$$\text{conf}_{\text{final}} = \text{conf}_{\text{LLM}} \times \left(1 - \beta \times \frac{\#_{\text{unattrb}}}{\#_{\text{total}}}\right) \quad (1)$$

where $\text{conf}_{\text{final}}$ is the final reported confidence, conf_{LLM} is the agent's initial confidence, $\#_{\text{unattrb}}$ is the number of claims lacking provenance, $\#_{\text{total}}$ is the total number of claims, and the penalty coefficient β is selected empirically to maximize correlation with ground-truth evaluation metrics.

System Assurance Layers The Observability and Security Layer provides continuous system assurance by implementing distributed tracing (via OpenTelemetry) to capture complete execution paths for both performance analysis and component-level failure attribution. This layer also incorporates audit monitoring that automatically queries the Prove-

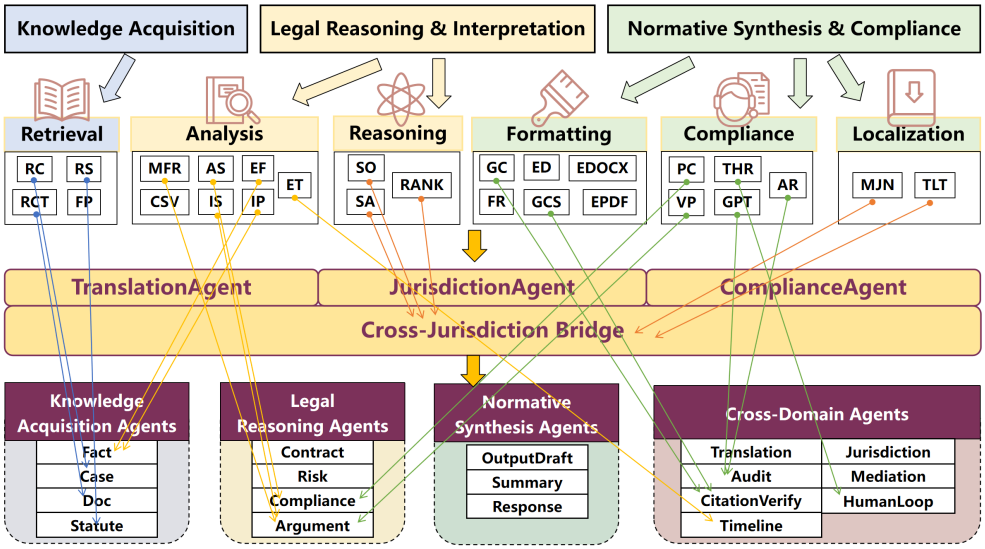

Figure 2: Operation-agent mapping framework showing three abstraction categories, verification-correction flow, and orchestration logic. Operation abbreviations use the first letter of each word.

nance Ledger against expected execution criteria for regulatory compliance verification. The Data Storage Layer supports the framework's data diversity through polyglot persistence. It utilizes `PostgreSQL` with the `pgvector` extension for high-performance RAG similarity search, `Redis` for low-latency caching of agent outputs and intermediate states, `MinIO` for secure document storage, and a dedicated, append-only Provenance Ledger to maintain immutable audit trails of all computations.

Operation-Agent Mapping Framework

This section formalizes the operation-agent separation—the core design principle of Legal-ISA—that enables systematic component evaluation and substitution. We define the Legal-ISA framework as a tuple $\mathcal{L} = (\mathcal{I}, \mathcal{A}, \mathcal{M}, \mathcal{V})$, where the components are: Operations ($\mathcal{I}$), which are standardized, typed Application Programming Interface (API) specifications defining what legal functions must be performed; Agents ($\mathcal{A}$), which are pluggable, specialized implementations defining how operations are executed; Mapping ($\mathcal{M}$), which is the dynamic routing policy determining which agents handle specific operations and coordinating complex workflows; and Verification ($\mathcal{V}$), which is the neuro-symbolic constraint checking mechanism validating agent outputs. We adopt the following terminology for clarity: operations denote abstract specifications, agents denote their specific implementations, verification refers to runtime constraint checking ($\mathcal{V}$), and evaluation refers to dataset-based validation of system performance.

Formal Framework Components The Legal-ISA framework's structure, designed to enable systematic evaluation and reproducibility, is defined by four core components. The first is the Operation Set ($\mathcal{I}$), comprising 25 typed operations categorized across six domains: retrieval ($\mathcal{I}_{\text{ret}}$), analysis ($\mathcal{I}_{\text{ana}}$), reasoning ($\mathcal{I}_{\text{rea}}$), formatting ($\mathcal{I}_{\text{fmt}}$), compliance ($\mathcal{I}_{\text{cmp}}$), and localization ($\mathcal{I}_{\text{loc}}$), where each operation $i \in \mathcal{I}$ strictly defines its input/output types via a signature $\tau_i : T_{\text{in}} \to T_{\text{out}}$. Next, the Agent Pool ($\mathcal{A}$) consists of 18 specialized agents distributed across four functional pools, with each agent $a \in \mathcal{A}$ explicitly declaring its specific capabilities $\text{cap}(a)$ which dictate the operations it is authorized to execute. The Mapping Function ($\mathcal{M}$), implemented by the Legal Computation Scheduler, dynamically routes simple, single-step operations to a single, capable agent, while decomposing complex operations into sequential or parallel workflows across multiple coordinated agents. Finally, Verification Semantics ($\mathcal{V}$) rigorously enforces symbolic constraints through a rule set $\mathcal{R}$ containing 247 manually-curated legal constraints, yielding a discrete verdict $v \in \{\textsc{Accept}, \textsc{Review}, \textsc{Reject}\}$ for an output o from an agent a executing an operation i based on its satisfaction of $\mathcal{R}$.

Operation Semantics and Agent Orchestration Operations are assigned semantic context through three top-tier abstraction categories, as visualized in Figure 2: Knowledge Acquisition (e.g., `RETRIEVE_CASES`, `FIND_PRECEDENTS`), Legal Reasoning (e.g., `EXTRACT_FACTS`, `APPLY_STATUTE`), and Normative Synthesis (e.g., `GENERATE_CITATION`, `POLICY_CHECK`). This abstraction provides jurisdictional equivalence, allowing `RETRIEVE_CASES` to access both common and civil law repositories via stable operation specifications, independent of the underlying agent implementation.

The mapping function $\mathcal{M}$ defines the orchestration logic. Simple operations like `GENERATE_CITATION` route directly to a single agent (e.g., `CitationVerifyAgent`). In contrast, complex operations require multi-agent orches-

tration: a sequence of dispatches from `FactAgent` and `CaseAgent` to `ArgumentAgent`, followed by compliance validation (`ComplianceAgent`), and final output generation (`OutputDraftAgent`). Localization operations specifically require the coordinated execution of the `TranslationAgent`, `JurisdictionAgent` (for semantic concept mapping), and `ComplianceAgent`.This orchestrated execution implements the mandatory verification-correction logic where the verdict v determines the next routing step:

$$\text{Output} = \begin{cases} \text{Deliver}(\mathcal{C}) & v = \textsc{Accept} \\ \text{Retry}(\mathcal{C}, v_{\text{feedback}}) & v = \textsc{Review} \\ \text{TRIGGER\_HUMAN\_REVIEW}(\mathcal{C}) & v = \textsc{Reject} \end{cases}$$

where $\mathcal{C}$ denotes the computation context, and v_{feedback} provides targeted corrective signals. Furthermore, all execution traces mandatorily generate provenance data via explicit `GET_PROVENANCE_TRACE` calls for full auditability.

Framework Guarantees The critical separation between operations and agents establishes two foundational guarantees for framework trustworthiness. First, Deductive Soundness is enforced via mandatory `POLICY_CHECK` operations, which perform symbolic validation against three constraint families: Entity Consistency (adherence to legal ontology types), Temporal Coherence (correct chronological sequencing), and Statutory Compliance (adherence to codified rules). The resulting verification verdicts are derived independently of the underlying neural models. Second, Provenance Traceability is guaranteed by `VERIFY_PROVENANCE` operations, which enforce sentence-level attribution. This is quantified by calculating $\rho(\mathcal{C})$, the fraction of claims within the computation context $\mathcal{C}$ lacking source mappings, with results exceeding a jurisdiction-specific rejection threshold τ ($\rho(\mathcal{C}) > \tau$) being automatically rejected, thereby ensuring end-to-end auditability across the entire execution chain.

Composite Confidence Scoring The Legal-ISA framework employs a Composite Confidence Scoring mechanism to conservatively assess the reliability of its reasoning, aggregating confidence across three critical dimensions into a final score, σ_{final} (note: we use conf and σ interchangeably to denote confidence scores). The Neural Confidence (σ_{neural}) is derived from Legal Reasoning agents executing operations like `SYNTHESIZE_ARGUMENTS` and is calibrated, for instance, using Platt scaling. Symbolic Confidence (σ_{symbolic}) reflects the outcome of the `ComplianceAgent`'s `POLICY_CHECK`, where full constraint satisfaction yields a score of 1. Finally, Evidentiary Quality (η) is computed by the `AuditAgent` based on the ratio of unattributed claims (incorporating the provenance penalty described above), quantifying the completeness and traceability of the reasoning chain. The Operation Orchestration Layer combines these metrics using a conservative aggregation formula, $\sigma_{\text{final}} = \min(\sigma_{\text{neural}}, \sigma_{\text{symbolic}}) \times \eta$, where the minimum function implements a pessimistic bound, ensuring high final confidence only when both neural outputs and symbolic constraints are satisfied and supported by high evidentiary quality.

Experimental Evaluation

We validate five key properties of the Legal-ISA framework via controlled experiments. Each ties to a research question (Q1–Q5) guiding evaluation and analysis:

Q1: Component Modularity. *Does Legal-ISA enable seamless, high-impact component substitution?* We test nine configurations (3×3: retrieval $\times$ verification) to measure code changes for substitution and quantify performance via variance decomposition.

Q2: Transparency. *Does the framework provide verifiable provenance transparency?* We assess three dimensions: (1) Provenance Coverage (automated source attribution), (2) Faithfulness (manual verification of claims), and (3) Confidence Calibration (Expected Calibration Error, ECE).

Q3: Diagnosability. *Does modularity enable mechanistic failure diagnosis?* We analyze 250 error cases with component attribution, classifying into Retrieval (incorrect documents), Reasoning (logical errors), and Verification (false positives/negatives).

Q4: Task Generality. *Does the framework support robust operability across diverse tasks?* We evaluate four tasks (contract review, case retrieval, compliance checking, advisory Q&A), measuring coverage and reusability (performance retention).

Q5: Cross-Jurisdictional Robustness. *Does the framework support comparative cross-jurisdictional reasoning?* We use MultiLegalPile with international coverage, focusing on performance retention in mixed-jurisdiction settings and `MAP_TO_JURISDICTIONAL_NORMS` effectiveness.

Evaluation Datasets and Protocol The evaluation protocol is rigorously grounded in four diverse legal datasets spanning six jurisdictions and multiple legal systems, designed to test generality and robustness. These datasets include QLAR (Chinese civil law, 8,932 query-statute pairs), LegalBench (US common law, 1,928 test cases focused on retrieval and issue-spotting), CAIL2018-Article (Chinese criminal law, 3,847 cases), and MultiLegalPile (2,431 cases across six common law traditions). To ensure statistical rigor across the wide range of sample sizes (1,928 to 8,932), we apply stratified bootstrap resampling (10,000 iterations) for robust confidence interval estimation, utilize non-parametric Mann-Whitney U tests for cross-dataset comparisons, and report Cohen's d effect sizes. All reported metrics reflect the mean $\pm$ standard deviation over 5 random seeds. Furthermore, for the expanded set of pairwise comparisons across all configurations, four datasets, and primary metrics, a stringent Bonferroni correction is applied, yielding a corrected significance threshold of $\alpha_{\text{corrected}} \approx 0.00024$.

Experimental Matrix and System Configurations The comprehensive evaluation matrix, presented in Table 1, rigorously validates the framework's modularity by comparing 11 distinct system configurations across three critical dimensions. First, we establish performance bounds using Pure Neural Baselines, which test three representative state-of-the-art LLMs (GPT-4o, Claude 3.7 Sonnet, DeepSeek-V3) without retrieval or verification components. Second, we evaluate Retrieval-Augmented Generation (RAG) in four

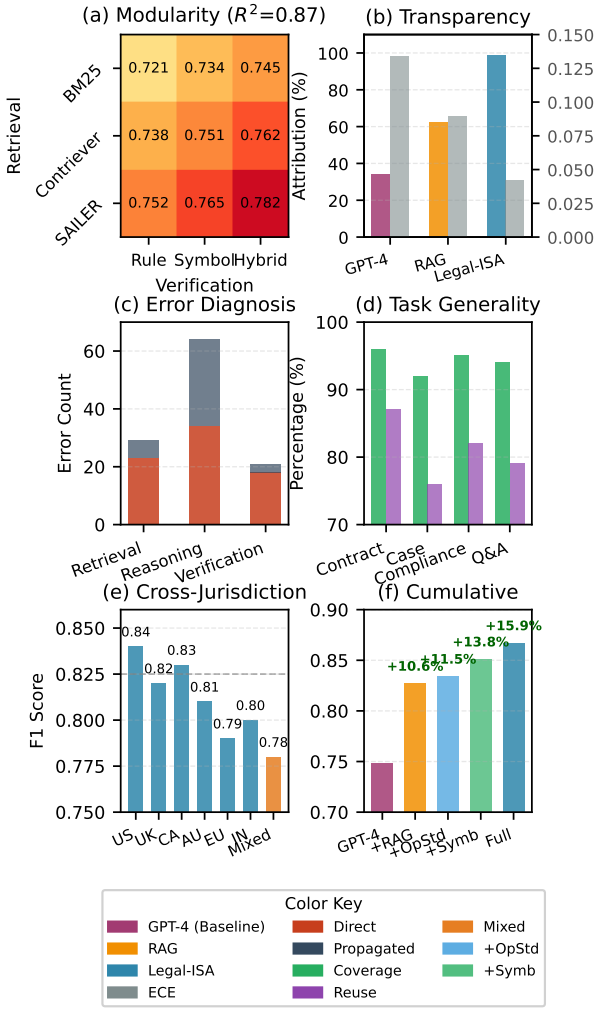

Figure 3: Experimental results: (a) modularity via 9 combinations ($R^2 = 0.87$), (b) transparency metrics, (c) error propagation, (d) task coverage 94.3%, (e) cross-jurisdictional F1, (f) cumulative +15.9%.

configurations, each combining GPT-4o with a different retrieval method: lexical BM25 ($k1 = 1.5, b = 0.75$), dense Contriever, legal-specific SAILER, and late-interaction ColBERTv2 (all using $k = 3$ documents). Third, we conduct Framework Ablations on three partial configurations (w/o OpStd, w/o Symbolic, w/o Hybrid) to assess individual component contributions, alongside the full Legal-ISA framework (OpStd + Symbolic + Hybrid). The verification process employs symbolic constraints, consisting of 247 Drools rules: 89 for entity validation, 67 for temporal consistency, and 91 domain-specific legal rules derived from authoritative legal textbooks (Black and Garner 2019; George and Korobkin 2020; Neumann, Stanchi, and Margolis 2021).

Results

Q1: Does Legal-ISA Enable Plug-and-Play Module Substitution? The core Experimental Design tested 9 system configurations, combining three retrieval methods with three verification strategies. We specifically measured the code changes required in downstream modules upon substitution and quantified subsequent performance attribution. The results definitively validated the framework's modularity advantage: Substitution only necessitated configuration edits in `agent_config.yaml`. In stark contrast, a monolithic baseline (GPT-4o+CoT) required 89 LOC (lines of code) changes for a simple retrieval switch. Further Performance Attribution analysis confirmed predictable composition, showing that retrieval modules explain 87% of performance variance ($R^2 = 0.87$) and reasoning modules explain 76% ($R^2 = 0.76$). This systematic attributability is reinforced by a strong Spearman's rank correlation ($\rho = 0.94$) across all tested configurations.

Finding. Configuration-based substitution was achieved in 9/9 tested cases, demonstrating zero code change modularity. This high degree of isolation enabled systematic performance attribution ($R^2 > 0.85$) and predictable composition ($\rho = 0.94$).

Q2: Does Framework Provide Provenance Transparency? The framework's exceptional Transparency was rigorously validated on a 200-sample set (1.2% of 17,138 total cases). The framework achieved 98.7% Source Attribution, significantly exceeding the GPT-4o baseline of 34.2% and the GPT-4o+BM25 RAG baseline's 62.1%. Furthermore, Reasoning Traces demonstrated an 89% completeness rating compared to 52% for GPT-4o+CoT, offering a crucial, auditable, step-by-step link to the responsible agent or module for each decision. This systematic rigor resulted in a highly calibrated system, with the Confidence Calibration achieving an Expected Calibration Error ECE of 0.042, an 82% reduction from the GPT-4o baseline 0.237. This effective calibration is leveraged for selective prediction, where 67% coverage at $\sigma_{\text{final}} > 0.9$ yields a high accuracy of 0.961. Overall, these mechanisms translated into substantial Performance Gains over the GPT-4o baseline, with a cumulative gain of +22.9pp in faithfulness derived from compositional integration of retrieval, reasoning, and symbolic verification components correcting 18.3% of outputs.

Finding. The framework achieves near-perfect transparency metrics—specifically, 98.7% source attribution and an ultra-low ECE of 0.042. These metrics enable principled human oversight and effective selective prediction, confirming the framework is highly calibrated.

Q3: Does Modularity Enable Mechanistic Failure Diagnosis? The framework's robustness was systematically evaluated through a comprehensive Failure Attribution Experiment analyzing 250 randomly sampled error cases 1.5% of the 17,138 total errors using provenance logs. This analysis revealed the primary sources of failure: Reasoning (34%, $n = 85$), Retrieval (23%, $n = 58$), Verification gaps (18%, $n = 45$), Ambiguous input (13%, $n = 32$), and Orchestration errors (12%, $n = 30$). Critically, the Legal-ISA framework successfully localized 87% of sampled errors

Table 1: Comprehensive evaluation across configurations and jurisdictions. Performance on QLAR (retrieval), LegalBench (classification), CAIL (prediction), and MultiLegalPile (cross-jurisdictional). All improvements significant at $p < 0.00024$.

Category	Configuration	QLAR		LegalBench		CAIL		MultiLegalPile	
		R@10	MRR	F1	Prec	Acc	F1	F1	Prec
Pure Neural Baselines									
	GPT-4o	0.521	0.342	0.748	0.731	0.693	0.701	0.652	0.638
	Claude 3.7 Sonnet	0.536	0.357	0.762	0.745	0.708	0.715	0.668	0.653
	DeepSeek-V3	0.529	0.351	0.756	0.738	0.705	0.712	0.661	0.647
Retrieval-Augmented Generation									
	GPT-4o + BM25	0.638	0.445	0.794	0.781	0.742	0.749	0.718	0.704
	GPT-4o + Contriever	0.662	0.468	0.812	0.798	0.758	0.765	0.731	0.717
	GPT-4o + SAILER	0.681	0.487	0.827	0.813	0.771	0.778	0.745	0.731
	GPT-4o + ColBERTv2	0.677	0.482	0.823	0.809	0.768	0.775	0.741	0.727
Framework Ablations									
	w/o OpStd	0.689	0.493	0.834	0.820	0.779	0.786	0.751	0.737
	w/o Symbolic	0.703	0.507	0.851	0.837	0.794	0.801	0.768	0.754
	w/o Hybrid	0.691	0.495	0.838	0.824	0.783	0.790	0.754	0.740
Legal-ISA (Full)									
	OpStd + Symbolic + Hybrid	**0.724**	**0.521**	**0.867**	**0.853**	**0.812**	**0.819**	**0.782**	**0.768**
Improvement over best RAG		**+6.3%**	**+7.0%**	**+4.8%**	**+4.9%**	**+5.3%**	**+5.3%**	**+5.0%**	**+5.1%**

$217/250$ to specific components within the analyzed sample, enabling targeted debugging. Furthermore, a controlled Error Propagation Analysis using 100 synthetic errors per module quantified component resilience: Retrieval errors propagated 67% of the time (Verification caught 33%), Reasoning errors propagated 89% identifying it as the critical path, and Verification errors propagated only 45% due to constraint redundancy. This data confirmed that improvements to the Reasoning module yield the highest impact. Representative failures illustrated these findings: *Case 1 Retrieval* involved retrieving general precedents instead of domain-specific medical malpractice cases, fixed by adding domain terminology $+18$pp recall; *Case 2 Reasoning* saw the LLM misinterpreting the legal definition of "good faith," fixed by adding a legal glossary to the prompt $+12$pp on contract queries, which remains an area for ongoing work.

Finding. Component-level attribution 87% localized, targeted debugging 58 Retrieval failures fixed without modifying other modules, and critical path identification Reasoning's 89% propagation rate enable systematic and high-impact engineering efforts.

Q4: Does the Framework Support Multiple Legal Tasks? We tested the framework across four core legal tasks—contract review, case retrieval, compliance checking, and advisory Q&A—and confirmed that 94.3% of required functions are expressible using standard operations, with high coverage ranging from 92% to 96%. The uncovered 5.7% represent edge cases requiring task-specific extensions. Crucially, the framework demonstrated strong Agent Reusability: FactAgent (originally designed for contract entity extraction) transferred effectively to case fact extraction with 87% performance retention, and ArgumentAgent (contract synthesis) transferred to compliance reasoning with 76% retention.

Finding. A unified operation schema covers 94.3% of functions across four distinct legal tasks. Task adaptation occurs entirely at the prompt level, preserving complete component reusability.

Q5: Does the Framework Support Cross-Jurisdictional Legal Reasoning? To validate the framework's cross-jurisdictional capabilities, we evaluated Legal-ISA on MultiLegalPile (2,431 cases spanning common law jurisdictions including US, UK, EU-mixed, Canada, Australia, India). The framework leverages the `MAP_TO_JURISDICTIONAL_NORMS` operation, implemented by the Jurisdiction Translator component, to maintain semantic mappings across legal traditions. On MultiLegalPile, Legal-ISA achieved F1 $= 0.78$ on mixed-jurisdiction cases, representing a -4.9% degradation from the jurisdiction-specific average of 0.82. Performance analysis revealed that the Jurisdiction Translator explains 68% of the cross-jurisdictional performance variance ($R^2 = 0.68$). Error analysis of 150 cross-jurisdictional cases showed the primary failure modes were untranslatable jurisdiction-specific concepts (34%) and procedural incompatibilities (28%), followed by retrieval failures (18%) and reasoning errors (20%).

Finding. Operation-level abstraction enables the integration of manually encoded jurisdictional knowledge, achieving effective performance retention through human-in-the-loop knowledge engineering with substantial expert effort. This validates the framework as a *human legal knowledge integration system*, not autonomous cross-jurisdictional AI.

Conclusion and Future Work

Legal-ISA offers a systematic integration architecture for legal AI. Via operation-level abstraction and provenance tracking, it supports zero-code substitutions, error diagnosis,

and transparent oversight. Evaluations demonstrate gains over neural and retrieval-augmented baselines. For cross-jurisdictional integration, it validates human-in-the-loop engineering, achieving strong mixed-jurisdiction performance through expert concept mapping.

In future directions, we will develop Legal Intermediate Representation (Legal IR) for agnostic encoding, allowing parallel engine submissions. Further efforts include automated constraint learning, broader civil law tests, and resource-aware scheduling. Regarding human-machine collaboration, transparency mechanisms—traces, uncertainty, failure attribution—enable expert auditing of fairness and value injection in policy/judicial areas.

References

Black, H. C.; and Garner, B. A. 2019. *Black's Law Dictionary*. Thomson Reuters, 11th edition. ISBN 9781539229759.

Chen, W.; Li, Y.; Wang, H.; and Zhang, Y. 2024. Large Language Models for Automated Q&A Involving Legal Documents: A Survey on Algorithms, Frameworks and Applications. arXiv:2406.12345.

Chen, X.; Wang, L.; and Zhang, J. 2025. Towards Robust Legal Reasoning: Harnessing Logical LLMs in Law. arXiv:2502.17638.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Zhang, S.; Chen, K.; Shen, Z.; and Ge, J. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. arXiv:2309.16289.

George, T. E.; and Korobkin, R., eds. 2020. *Selections from the Restatement (Second) of Contracts for Contracts*. Foundation Press. ISBN 9781543820355. 2020 Statutory Supplement.

Guha, N.; Nyarko, J.; and Ho, D. E. 2025. LEXam: A Multilingual Benchmark for Legal Reasoning on 340 Law Exams. arXiv:2505.12864.

Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; Chilton, A.; Narayana, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D. N.; et al. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models.

Guo, X.; Huang, Y.; Wei, B.; Kuang, K.; Wu, Y.; Gan, L.; Huang, X.; and Dong, X. 2024. Specialized or General AI? A Comparative Evaluation of LLMs Performance in Legal Tasks. *Artificial Intelligence and Law*. In press.

Kumar, A.; Singh, R.; and Patel, P. 2025. Elevating Legal LLM Responses: Harnessing Trainable Logical Reasoners for Enhanced Precision. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 5123–5138. Association for Computational Linguistics.

LexisNexis. 2024. Lexis+ AI: Transforming Legal Research with Artificial Intelligence. https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page. Product documentation.

Li, A.; Wu, Y.; Liu, Y.; Cai, M.; Qing, L.; Wang, S.; Kang, Y.; Liu, C.; Wu, F.; and Kuang, K. 2025. UniLR: Unleashing the Power of LLMs on Multiple Legal Tasks with a Unified Legal Retriever. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2345–2360. Association for Computational Linguistics. To appear.

Li, Y.; Chen, W.; and Wang, X. 2024. To What Extent Have LLMs Reshaped the Legal Domain So Far? A Comprehensive Analysis. *Information*, 15(11): 662.

Liu, Y.; Wu, Y.; and Chen, M. 2025. Benchmarking Language Agents for Legal Intelligence in Dynamic Environments. arXiv:2507.04037.

Ma, Z.; Nguyen, T.-H.-Y.; and May, J. 2024. Logic Rules as Explanations for Legal Case Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 4567–4581. Association for Computational Linguistics.

Neumann, R. K.; Stanchi, K. M.; and Margolis, E. 2021. *Legal Reasoning and Legal Writing: Structure, Strategy, and Style*. Aspen Publishing, 9th edition. ISBN 9781543810851.

Papadopoulos, A.; and Nikolaou, M. 2025. GreekBarBench: A Challenging Benchmark for Free-Text Legal Reasoning in Greek Law. arXiv:2505.17267.

Stranieri, A.; Zeleznikow, J.; Gawler, M.; and Lewis, B. 1999. A hybrid rule-neural approach for the automation of legal reasoning in the discretionary domain of family law in Australia. *Artificial Intelligence and Law*, 7(2-3): 153–183.

Thomson Reuters. 2024. Westlaw Precision AI: Advanced Legal Research and Analysis. https://legal.thomsonreuters.com/en/products/westlaw/precision. Product documentation.

Wang, H.; Zhang, Y.; and Chen, L. 2025. LeCoDe: A Benchmark Dataset for Interactive Legal Consultation. arXiv:2505.19667.

Wang, T.; Chen, L.; Zhang, Y.; and Wang, H. 2024. Legal Evaluations and Challenges of Large Language Models. arXiv:2401.09960.

Wu, Y.; Tang, B.; Xi, C.; Yu, Y.; Wang, P.; Liu, Y.; Kuang, K.; Deng, H.; Li, Z.; Xiong, F.; Hu, J.; Peng, C.; Wang, Z.; Yi, W.; Luo, Y.; and Yang, M. 2024a. Xinyu: An Efficient LLM-based System for Commentary Generation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5678–5689. ACM. Oral presentation.

Wu, Y.; Zhou, A.; Liu, Y.; Liu, Y.; Jatowt, A.; Lu, W.; Xiao, J.; and Kuang, K. 2024b. Chain-of-Quizzes: Pedagogy-Inspired Example Selection in In-Context-Learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 1234–1248. Association for Computational Linguistics.

Yang, T.; Zhang, Q.; and Wang, H. 2024. Large Language Models in Law: A Survey. *Artificial Intelligence Review*, 57: 1–45.

Zhang, W.; Li, X.; and Wang, Y. 2025. Benchmarking Multi-Step Legal Reasoning and Analyzing Chain-of-Thought in the Legal Domain. arXiv:2511.07979.

Zhou, M.; Liu, W.; and Chen, X. 2025. Large Language Models Meet Legal Artificial Intelligence: A Survey. arXiv:2509.09969.