LLM Agents in Law: Taxonomy, Applications, and Challenges

Shuang Liu[1], Ruijia Zhang[2], Ruoyun Ma[3], Yujia Deng[4], Lanyi Zhu[5], Jiayu Li[6], Zelong Li[7], Zhibin Shen[8], Mengnan Du[9]

[1]Carnegie Mellon University, [2]National University of Singapore, [3]Stanford University, [4]Independent Researcher, [5]University of Washington, [6]The University of Chicago, [7]Rutgers University, [8]Columbia University, [9]The Chinese University of Hong Kong, Shenzhen

Abstract

Large language models (LLMs) have precipitated a dramatic improvement in the legal domain, yet the deployment of standalone models faces significant limitations regarding hallucination, outdated information, and verifiability. Recently, LLM agents have attracted significant attention as a solution to these challenges, utilizing advanced capabilities such as planning, memory, and tool usage to meet the rigorous standards of legal practice. In this paper, we present a comprehensive survey of LLM agents for legal tasks, analyzing how these architectures bridge the gap between technical capabilities and domain-specific needs. Our major contributions include: (1) systematically analyzing the technical transition from standard legal LLMs to legal agents; (2) presenting a structured taxonomy of current agent applications across distinct legal practice areas; (3) discussing evaluation methodologies specifically for agentic performance in law; and (4) identifying open challenges and outlining future directions for developing robust and autonomous legal assistants.

1 Introduction

In recent years, large language models (LLMs) have precipitated a dramatic improvement in the legal domain, fundamentally altering how legal professionals approach complex information processing. These models have demonstrated exceptional capability across a variety of specialized legal tasks, ranging from routine document processing to sophisticated reasoning challenges. The integration of LLMs has improved distinct application areas, including legal judgment prediction (Shu et al., 2024), legal question answering (Louis et al., 2024), and contract review (Liu et al., 2025), offering new efficiencies in workflows that were previously manual and labor-intensive.

However, the deployment of standalone LLMs for legal tasks faces significant limitations, primarily due to persistent issues such as hallucination (Farquhar et al., 2024; Sriramanan et al., 2024) and the generation of outdated information. Given that legal practice constitutes a high-stakes environment where precision is mandatory and errors can have severe consequences, these reliability issues severely restrict the application of standard LLMs in real-world scenarios. To address these critical challenges, the field is shifting toward the development of LLM Agents (Li et al., 2025a), which utilize advanced capabilities such as planning, memory, and tool usage, to mitigate the shortcomings of base LLM models and meet the rigorous standards required by the legal profession.

In this paper, we present a comprehensive survey of LLM Agents for legal tasks, structuring our analysis to bridge the gap between technical capabilities and domain-specific needs. We first examine why LLM Agents are particularly promising for this domain by contrasting their advantages against the limitations of standalone models, followed by a detailed review of current application areas. Furthermore, we assess the performance of existing agents and conclude by highlighting open challenges and future directions to guide the next phase of research in legal AI.

1.1 Contribution and Uniqueness

Our Contributions. This paper provides a comprehensive overview of LLM agents in the legal domain, with four major contributions: (1) We systematically analyze the technical transition from standard legal LLMs to legal agents, detailing how agentic characteristics address critical deficiencies in legal applications. (2) We present a structured taxonomy of current agent applications across five distinct legal practice areas. (3) We discuss evaluation methodologies specifically for agentic performance in law. (4) We identify open challenges and outline future directions for developing robust, autonomous legal assistants.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15768–15792

Differences with Existing Surveys. While several existing surveys examine AI application in law (Lai et al., 2024; Hou et al., 2025; Chen et al., 2024), they differ significantly in scope and focus compared to our agent-centric approach. For instance, Lai et al. (Lai et al., 2024) focuses on standalone model capabilities and general ethical challenges in the judicial system, largely overlooking agentic architectures that utilize planning and tools to address identified limitations. In contrast, our survey uniquely and exclusively focuses on the depth of LLM Agents in the legal field, providing a detailed analysis of agentic principles, workflows, and evaluation criteria specific to legal practice.

2 Why LLM Agents for Legal Tasks?

2.1 Why LLMs by themselves are insufficient

LLMs are reshaping legal workflows through the Transformer architecture's dynamic semantic capture and generative reasoning (Shao et al., 2025). They excel at fundamental legal tasks (Davenport, 2025), with further improvements achieved via judicial-syllogism-based Chain-of-Thought (CoT) (Song et al., 2025). Models like LawLLM (Shu et al., 2024) demonstrates promise in simulating professional legal reasoning, while newest multimodal models extends analysis to audio and video modalities (vLex Team, 2025). However, LLMs do not resolve all legal challenges and even introduce new risks (Lai et al., 2024; Dehghani et al., 2025), which we categorize into three classes:

- *Class A: Persistent Traditional Challenges.* Long-cycle, multi-stage workflows remain difficult for LLMs to navigate (Mohsin et al., 2025). Current LLMs struggle with extended task consistency and procedural depth (Fei et al., 2024; Liu et al., 2024) for legal tasks requiring complex multi-steps (Guha et al., 2023).
- *Class B: LLM-Specific Challenges.* Hallucination is inherent in current generative architectures (Xu et al., 2025; Kalai et al., 2025; Huang et al., 2025). In law, fabricated citations or documents pose severe risks and significant real-world consequences (Dahl et al., 2024; Magesh et al., 2025; Scott, 2023).
- *Class C: Exacerbated or Emergent Risks.* Fixed post-training weights hinder adaption to evolving regulations(Ferraris et al., 2025), creating hidden risks where erroneous conclusions are hard to detect or trace(Dahl et al., 2024). Further, the "black-box" nature of LLMs complicates legal accountability, ethical frameworks, and societal trust(Anthropic, 2023; Weidinger et al., 2022).

2.2 LLM Agents Address These Limitations

Although comprehensive lifecycle management best supports responsible LLM development (Wang et al., 2025a), it is often cost-prohibitive, prompting a shift toward agentic frameworks as a practical alternative to standalone models. Unlike standalone models, LLM agents serve as reasoning engines that orchestrate external modules and iterative workflows (Anthropic, 2024; Xi et al., 2025b). We summarize the core agentic capabilities and corresponding remedies as follows:

- *External Grounding and Knowledge Freshness.* To mitigate hallucinations (Class B) and knowledge expiration (Class C), agents employ Tool Use and Retrieval-Augmented Generation (RAG) (Weng, 2023; Larochelle et al., 2020). By offloading "authority" from internal parameters to verifiable systems such as legal databases, statutes, and case law APIs, gents anchor conclusion to explicit evidence (Cui et al., 2023; Guha et al., 2023; Huang et al., 2023). This is essential for cross-jurisdictional tasks and rapidly evolving regulatory updates that static models fail to reliably trace (Vu et al., 2024).
- *Procedural Orchestration and Long-term Consistency.* To address persistent traditional challenges (Class A), agents employ Planning and Memory modules. Legal workflows are inherently multi-stage and long-cycle (Li et al., 2025a; Yang et al., 2025), and planning decomposes complex tasks, such as due diligence or multi-step litigation strategy, into manageable sub-goals (Chen et al., 2025c). Meanwhile, memory preserves context over time, preventing the "lost-in-the-middle" phenomenon and maintaining consistent legal reasoning across a case's lifecycle (Liu et al., 2024; Cui et al., 2023).
- *Multi-layer Verification and Governance.* To tackle accountability and ethical risks (Class C), agentic frameworks introduce Reflection, Multi-agent Collaboration, and Human-in-the-loop (HITL) protocols (Scanlon et al., 2025; Xi et al., 2025b). Reflection enables post-generation consistency checks to detect logical contradictions or evidence gaps, effectively mitigating bias (Shinn et al., 2023; Zhang and Ashley, 2025). Multi-agent Systems adopt specialized agents

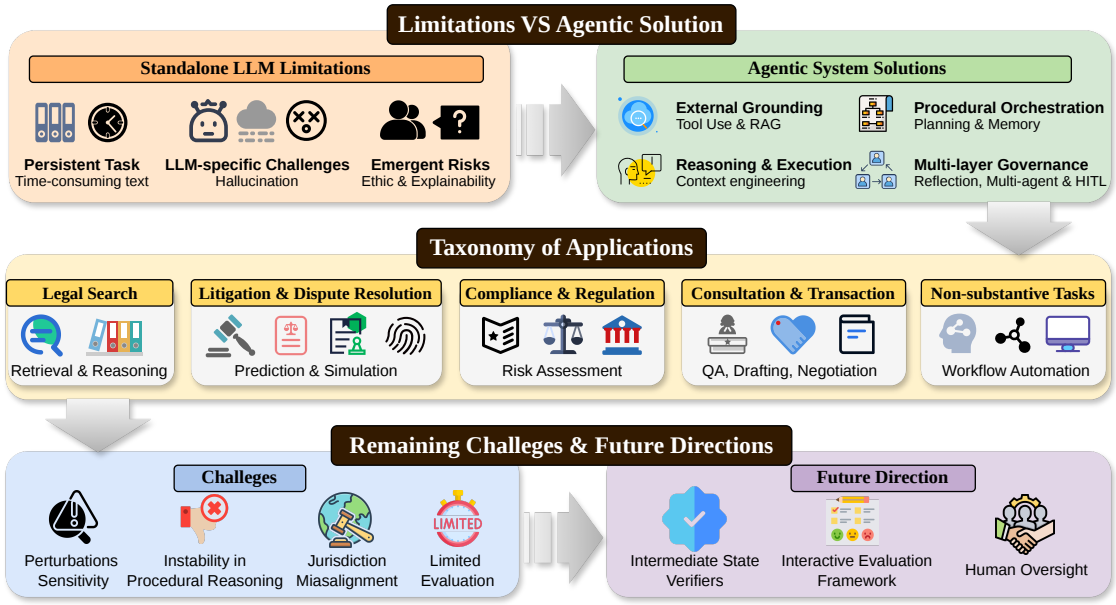

Figure 1: Overview of paper structure and LLM agent application.

to cross-examine outputs, mimicking the peer-review process in law firms to enhance interpretability (Jiang and Yang, 2025; Sun et al., 2024; Jing et al., 2025). HITL positions human as the final arbiter, ensuring data privacy control and ultimate legal accountability(School, 2025; Bommarito et al., 2025).

2.3 Common Architecture of Legal LLM Agents

Legal agents commonly adopt architectures that decompose legal reasoning into explicit intermediate stages, specialized roles, or iterative refinement loops. Reflection-based agents improve outputs through self-critique and revision (Zhang and Ashley, 2025), debate-style agents compare competing reasoning paths under a judge (Chen et al., 2025e), and plan-and-execute frameworks separate high-level coordination from subtask execution (Nguyen et al., 2025a). Tool-augmented and role-based multi-agent systems further improve grounding, controllability, and specialization by combining LLMs with retrieval, external APIs, and role-specific collaboration (Zhu et al., 2025c; Cui et al., 2023). More recently, agent frameworks combined with reinforcement learning extend these designs by enabling procedural strategy discovery and adaptive behavior through simulation and self-play (Badhe, 2025). Table 1 in Appendix A summarizes representative agentic architectures used in legal-domain systems and their core properties.

3 Current LLM Agent Application Areas in the Legal Domain

In this section, we organize existing work into a taxonomy consisting of five core categories: (1) Legal search, which support the application of legal authorities; (2) Litigation and dispute resolution, where agentic systems enable role specialization and long-horizon interaction; (3) Compliance and regulation, where verification and auditability are essential; (4) Advisory consultation and transactional practice, emphasizing client interaction and document-centric workflows; (5) Non-substantive tasks such as administration automation. Figure 2 shows the structure of the taxonomy with representative agentic systems. Table 2 and Table 3 summarizes the legal agentic systems proposed by existing literature and commercial products.

This categorization reflects the structure of real-world legal practice, closely mirroring how work is organized within practice groups of legal institutions. Each category maps to distinct technical demands on agents, ranging from retrieval and reasoning primitives, to long-horizon interaction, procedural compliance, and workflow orchestration.

3.1 Legal Search & Research

Legal retrieval, legal theory understanding, and legal reasoning form the foundation of legal search and research system, jointly supporting how users identify, interpret, and apply legal sources. Le-

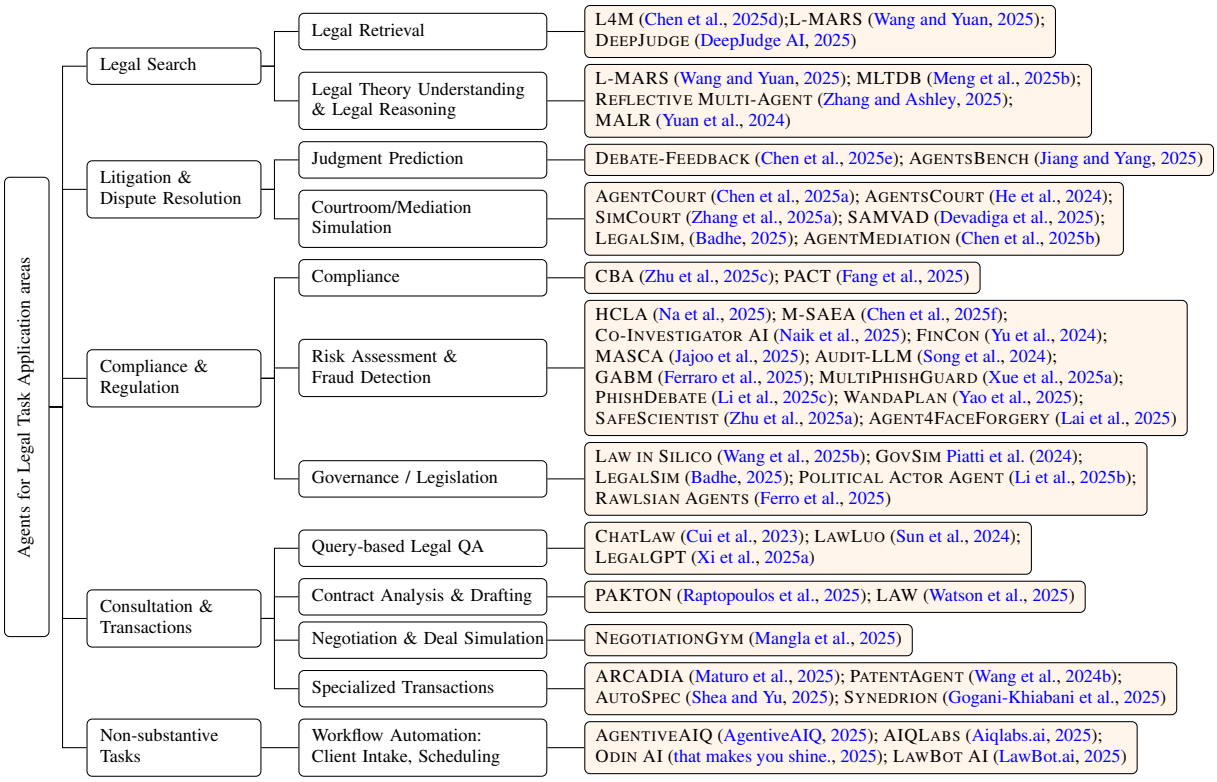

Figure 2: Taxonomy of LLM agents for legal tasks. This framework categorizes five core legal domains and maps them to representative academic and commercial agentic systems.

gal retrieval connects user information needs to cases, statutes, and regulations, while legal theory understanding captures the doctrinal concepts that govern their interpretation. Legal reasoning integrates these components through structured deduction, analogy, and rule application, supporting legally valid and defensible conclusions.

Legal Retrieval. Legal retrieval often maps informal queries to authoritative cases, statutes, and regulations (Liu et al., 2022), and must ensure high precision, traceability of retrieved evidence, jurisdictional and temporal correctness, while providing interfaces that support professional inspection and control (Liu et al., 2023; Chen et al., 2025d; Wang and Yuan, 2025). Case retrieval identifies prior judicial decisions relevant to a fact pattern, statutory retrieval locates applicable legislative provisions, and regulation retrieval seeks relevant administrative rules (Feng et al., 2024; Savelka and Ashley, 2022). Recent work adopts agentic system designs to support high-quality legal retrieval. First, structured query generation enable LLM agents to translate underspecified user inputs into legally grounded retrieval queries, with mechanisms such as candidate buffers rewriting queries and preserving intermediate results across turns to improve retrieval stability and success rates (Liu et al., 2022). Second, robustness is improved through multi-granular and logic-aware fusion of legal information, where documents are modeled at different levels such as facts, reasoning, and rulings, enabling LLM agent systems to tolerate paraphrase and noisy expressions while maintaining legal relevance (Meng et al., 2025a). Finally, multi-agent and neural-symbolic workflows are introduced to enhance verification and trustworthiness. Orchestrated agent systems explicitly check jurisdictional and temporal validity before producing answers (Wang and Yuan, 2025), while neural-symbolic approaches formalize statutes and use symbolic solvers to provide machine-checkable support for retrieved and applied legal rules (Chen et al., 2025d).

Legal Theory Understanding & Legal Reasoning. Legal theory defines the set of concepts and rules that underpin legal claims and their justification, including elements of offenses, duties, rights, and how these elements relate to one another (Yuan et al., 2024; Sadowski and Chudziak, 2025). Legal reasoning is the process of applying legal theory to facts using deduction, analogy, and rule application. Recent work shows that standalone LLMs do not reliably grasp legal theories or their formal struc-

ture (Yuan et al., 2024; Jing et al., 2025), and they also struggle to perform reliable legal reasoning without extra structure or checks (Jing et al., 2025; Sadowski and Chudziak, 2025).

To improve legal reasoning and theory understanding, recent work follows complementary agentic design strategies. First, multi-agent decomposition is used to split complex tasks into focused sub-tasks, enabling role specialization. Systems such as MALR and ChatLaw show that a multi-agent design helps models focus on theory extraction, fact analysis, and synthesis separately (Yuan et al., 2024; Cui et al., 2023). Second, some systems decouple retrieval and reasoning into atomic primitives that a planner agent composes; the Deterministic Legal Agents work proposes a canonical primitive API for auditable, point-in-time retrieval and deterministic steps (de Martim, 2025). Third, a reflective multi-agent framework is used for legal argument generation in which specialized agents iteratively analyze facts and refine arguments, improving factual grounding, ethical reliability, and robustness of legal reasoning compared to single-agent approaches (Zhang and Ashley, 2025).

3.2 Litigation & Dispute Resolution

Litigation and dispute resolution covers legal tasks that arise in adversarial proceedings, which involve analyzing case facts, applying relevant law, and engaging in strategic interactions between opposing parties and judges. Core litigation-related tasks include predicting judicial outcomes, drafting litigation documents such as motions and briefs, and simulating courtroom or mediation processes.

Judgment Prediction. On the task of legal judgment prediction, Shui et al. (2023) suggests that effective legal judgment prediction requires more sophisticated mechanisms beyond naive LLM–IR pipelines. Building on this limitation, Jiang and Yang (2025) propose a multi-agent LLM framework that explicitly models judicial deliberation through role-based agents, where individual agents independently reason over cases and iteratively reach a consensus through discussion, leading to consistent improvements in both predictive accuracy and decision quality. Further advancing agent-based legal reasoning, Chen et al. (2025e) introduce a Debate-Feedback framework that augments multi-agent debate with explicit reliability evaluation, mitigating the instability of uncontrolled debate and yielding more stable and effective legal judgment prediction.

Courtroom & Dispute Resolution Simulation. Courtroom simulation must handle dynamic courtroom interactions, including multi-round adversarial debates, real-time responses, and evolving argument strategies, motivating agent-based frameworks with role-specialization and long-horizon interactions. Early works such as AGENTCOURT introduces adversarial lawyer agents to simulate courtroom debate and enable automated knowledge evolution through interaction (Chen et al., 2025a). Subsequent work further integrate egal knowledge augmentation and structured court debate to improve judicial decision-making fidelity (He et al., 2024). Related systems explore courtroom reasoning from different perspectives: MASER simulates scalable synthetic legal scenario data by simulating interactions between roles such as client, lawyer, and supervisor (ShengbinYue et al., 2025).

Beyond generic courtroom debate, several works focus on jurisdiction- or process-specific simulation. SIMCOURT introduces a multi-agent framework to simulate real and well-structured trial procedures of Chinese criminal courts, showing potentials to outperform human legal practitioners (Zhang et al., 2025a). Broader institutional simulations further extend this line of work: SAMVAD models judicial deliberation dynamics in India (Devadiga et al., 2025), while LEGALSIM simulates legal systems to uncover procedurally valid yet harmful strategies, enabling stress-testing of legal processes (Badhe, 2025). Besides court simulation, agentic simulation has also been applied to adjacent dispute-resolution settings. AGENTMEDIATION, simulates full five-stage civil dispute mediation with configurable roles, enabling experiments that reveal patterns like group polarization and surface consensus (Chen et al., 2025b).

3.3 Compliance, Governance, & Regulation

Compliance, governance and regulation aim at ensuring that individuals and organizations adhere to applicable laws, regulations, and governance standards. These tasks are largely proactive and continuous, and often involve analyzing legislation and regulatory guidance, translating legal requirements into organizational policies, and ensuring accountability through audits and oversight.

Compliance. In financial crime prevention, agentic architectures assign bounded roles to autonomous agents, emphasizing compliance-by-

design through traceability, explainability, and audit logging (Axelsen et al., 2025). For enterprise compliance, conversational agentic assistant CBA routes user queries between a fast RAG-based mode and a full agentic mode, balancing latency and capability by dynamically selecting execution paths (Zhu et al., 2025c). Embeddings-driven graph enable agentic systems such as PACT (Fang et al., 2025) to link heterogeneous enterprise artifacts over metadata, ownership, and compliance context and reason over privacy-relevant connections at scale. Complementary work proposes specialized agentic systems that function as ethics counsel within legal workflows, providing accountable guidance and mitigating bias in professional decision-making (O'Grady and OG, 2024).

Agentic approaches have also been explored in data-intensive compliance settings. For instance, a multi-agent system assists drug asset in due diligence, combining a web-browsing agent with an LLM-as-a-judge to suppress hallucinations and improve precision (Vinogradova et al., 2025). For data protection and privacy compliance, multi-agent systems decompose complex regulatory obligations into planning, execution, and verification roles to support end-to-end governance and data transfer validation (Nguyen et al., 2025a; Videsjorden et al., 2025). At the regulatory level, multi-agent RAG frameworks construct structured representations of regulations to enable traceable, ontology-free compliance question answering (Agarwal et al., 2025). Beyond enterprise settings, LLM agent moderation system has been applied to identifying non-compliant content in decentralized social platforms (La Cava and Tagarelli, 2025), and to context-aware, jurisdiction-specific child safety moderation (Fillies et al., 2025).

Additionally, commercial AI agents are emerging to automate regulatory compliance workflows. REGOLOGY provides agents for regulatory research and change management (Regology, 2025), while V7LABS offers regulatory cross-referencing and compliance verification (V7 Labs, 2025b). AKIRA AI automates regulatory monitoring, risk analysis, and policy adjustments (Akira AI, 2025).

Risk Assessment & Fraud detection. Recent research utilizes multi-agent frameworks to break down risk assessment and fraud detection tasks into specialized roles. For example, Park (2024) and HCLA (Na et al., 2025) assign agents to data transformation, anomaly identification, and reporting. Other systems, such as FINCON and MASCA, integrate risk-control agents directly into financial decision-making and credit assessment pipelines, embedding risk modeling into the agent's workflow (Yu et al., 2024; Jajoo et al., 2025). Beyond detection, agentic designs are used to enforce security and evaluate systemic hazards. A robotic security framework (Shah and Deshpande, 2025) combines LLM agents with blockchain and monitoring to safeguard online transactions. To ensure system safety, M-SAEA employs safety-aware auditor agents to probe multi-agent teams for coordination and deployment risks (Chen et al., 2025f). CO-INVESTIGATOR AI divides Anti-Money Laundering tasks into planning, detection, and validation stages, coordinate through shared memory and use an Agent-as-a-Judge loop to ensure regulatory accuracy with human oversight (Naik et al., 2025).

For insider threats and phishing, AUDIT-LLM (Song et al., 2024) and GABM (Ferraro et al., 2025) use Decomposer and Supervisor agents to analyze logs for insider threats. For cyber-attacks, MULTIPHISHGUARD (Xue et al., 2025a) and PHISHDEBATE (Li et al., 2025c; Nguyen et al., 2025b) utilize debating and judge agents to aggregate evidence against deceptive URLs and emails. Additionally, such agentic internal control extends to other fields: Gu et al. (2025) detects greenwashing, WANDAPLAN (Yao et al., 2025) identifies misinformation, SAFESCIENTIST (Zhu et al., 2025a) monitors ethical risks in science, AGENT4FACEFORGERY (Lai et al., 2025) uses memory-equipped agents to detect forged facial data, and Islayem et al. (2025) introduces a blockchain–LLM agentic framework for detecting health insurance fraud through smart-contract enforcement and retrieval-grounded analysis.

Governance, Regulation & Legislation. In the governance domain, Agentic legal assistants integrate retrieval-augmented generation with LLM agents to support interactive querying and reasoning over EU GDPR legislation and case law (Mamalis et al., 2024). Further, multi-agent simulations are used to model governance dynamics: LAW IN SILICO combines individual and institutional agents to simulate lawmaking, adjudication, and enforcement (Wang et al., 2025b). Other frameworks simulate cooperative governance in shared-resource settings, showing how planning, communication, and moral reasoning shape collective outcomes (Piatti et al., 2024; Zhou et al., 2025).

For legislative analysis and simulation, prior work models U.S. Senate committee discussions using LLM-driven agents to study debate, reflection, and bipartisan decision-making under controlled conditions (Baker and Azher, 2024). LEGALSIM introduces a multi-agent legal simulation to uncover procedurally valid yet harmful strategies, enabling stress-testing of legislation (Badhe, 2025). Other systems simulate legislative behavior to predict roll-call votes using role-based LLM agents with interpretable reasoning (Li et al., 2025b).

3.4 Advisory Consultation & Transactions

Advisory consultation focuses on responding to consultation-style inquiries, requiring the translation of legal rules into practical guidance. Transactional practice, by contrast, centers on the analysis and negotiation of exchanges of rights, obligations, and assets among multiple parties. Core transactional tasks include contract review and drafting and negotiation of deal terms. Transactional work often extends to specialized domains such as patent registration, real estate, bankruptcy, and merger and acquisition (M&A), where domain-specific rule and procedural constraints are critical.

Query-Based Advisory Consultation. Recent research adopts agentic, multi-agent LLM systems to support reliable responses to user-posed legal queries and consultation-style questions. CHATLAW (Cui et al., 2023) treats legal consultation as a series of structured question-answering steps, using role-specialized agents and a knowledge-graph–enhanced mixture-of-experts to reduce hallucinations in practical legal Q&A. Similarly, LAWLUO (Sun et al., 2024) uses role-based agents to support multi-round Chinese legal consultations, iteratively clarifying and answering user queries to produce structured consultation reports. To improve trustworthiness in high-stakes question answering, Xi et al. (Xi et al., 2025a) propose an agentic hybrid RAG framework that dynamically routes legal queries between retrieval-based answering and multi-model generation, using a selector agent to choose the most reliable response. L-MARS (Wang and Yuan, 2025) combines agentic search with multi-agent reasoning to answer user queries with improved factuality and reduced uncertainty. Finally, LEGALGPT (Shi et al., 2024) formalizes legal chain-of-thought reasoning within a multi-agent framework for query answering.

Contracts, Negotiation, and Transactions. For contract review and drafting, systems such as PAKTON (Raptopoulos et al., 2025) and LAW (Watson et al., 2025) frame contract question answering and analysis as role-structured multi-agent pipelines, enabling traceable reasoning over long agreements. Complementary retrieval-augmented multi-agent systems automate the drafting of transactional documents by combining drafting, validation, and compliance agents (Suravarjhula et al., 2025). RAWLSIAN AGENTS applies agentic reasoning grounded in fairness principles to evaluate and draft bilateral agreements (Ferro et al., 2025). In addition to academic advances, industrial practice has increasingly adopted AI agents for contract review and drafting. Top international law firms are collaborating with legal AI companies to incorporate agentic AI for complex legal tasks such as reviewing loan documents for leveraged finance transactions (A&O Shearman, 2025; Pearson Labs, 2024). Similarly, SPELLBOOK AI ASSOCIATE provides agentic services covering streamlined agreement revision for commercial lawyers (Spellbook, 2025). Other non-legal focused AI agent startups also increasingly expand their services and products into contract review and analysis domains by providing customizable legal agents to assist in contract drafting (Harvey AI, 2025), extracting key contract clauses to streamline contract reviews (that makes you shine., 2025), and autonomous contract revision and drafting (ZBrain AI, 2025; Flank AI, 2025; Aline, 2025; Legora, 2025; Spellbook, 2025).

In the context of negotiation, agentic systems investigate autonomous bargaining and deal-making behaviors. Work such as NEGOTIATIONGYM (Mangla et al., 2025) simulates iterative negotiation dynamics among self-optimizing agents. Other studies use large-scale simulations to study strategic dynamics and risks in agent-to-agent negotiations and consumer transactions (Vaccaro et al., 2025; Zhu et al., 2025b). In addition to contract and business negotiation, some recent works focus on specialized transactional domains including agentic systems for automated patent analysis, filing, and document drafting (Wang et al., 2024b,a; Shea and Yu, 2025; Sakhinana et al., 2024), real estate transactions and housing services (Haurum et al., 2024), corporate bankruptcy analysis (Maturo et al., 2025), tax preparation (Gogani-Khiabani et al., 2025), and M&A evaluation (Mirzayev et al., 2025).

3.5 Workflow Automation

Non-substantive tasks focus on supporting internal operations rather than legal analysis itself. Legal practitioners and vendors use AI agents to automate repetitive and high-volume steps around intake and internal measures. For instance, intelligent intake agents collect case information and schedule consultations by interacting with calendaring systems (Voiceflow, 2025). Email-management agents support inbox triage, summarization, and draft generation (Jace AI, 2025; Relevance AI, 2025). Legal AI agents also assist in monitoring system states and triggering actions across tools without requiring constant user prompts (Matt, 2025). Workflow automation agents are also applied to internal coordination tasks such as document routing and version control. AGENTIVEAIQ enable firms to build branded agents that combine retrieval-augmented generation with knowledge graphs to autonomously perform client triage and internal insights delivery (AgentiveAIQ, 2025). AIQLABS provides agentic services on operational services including invoice automation, knowledge management, and recruiting automation (Aiqlabs.ai, 2025). ODIN AI offers a business AI agent focusing on meeting notes summarization and task automation (that makes you shine., 2025). LAWBOT AI provides legal-specific voice assistant agents for automatically handling client inquiries, scheduling consultation, and delivering case information (LawBot.ai, 2025).

4 Legal LLM Agent Evaluation

Evaluating LLM agents for legal tasks requires criteria that go beyond standard language modeling metrics. This section reviews recent work on common evaluation dimensions, benchmarks, and empirical results, and summarizes current progress of legal agent systems. Table 4 in Appendix C summarizes the commonly used legal benchmarks and their key properties.

4.1 Common Evaluation Criteria for the Legal Domain

Substantive Correctness. Substantive correctness measures whether the final conclusion is legally correct under the applicable law. Existing work operationalizes this notion mainly through two paradigms: (i) task-based paradigms that assesses whether LLMs or agents successfully complete legal tasks with varying complexity using task completion or final-answer accuracy as the primary metric (Fei et al., 2024; Li et al., 2024, 2025a; Jia et al., 2025). and (ii) legal Q&A settings, in which models' answers to legal questions and are evaluated against expert-authored reference answers, using fine-grained relational judgments, such as equivalence or contradiction, to capture outcome-level correctness (Bhambhoria et al., 2024).

Legal Reasoning Correctness. Existing work evaluates the reasoning capabilities of legal agents by validating intermediate steps and structured argumentation, including keyword-matching rates in function-calling outputs (Li et al., 2025a), and the Issue–Rule–Application–Conclusion (IRAC) structure (Guha et al., 2023). Other studies assess syllogistic reasoning by examining the logical coherence and alignment of LLM-generated responses, premises, and legal principles, jointly evaluating both reasoning correctness and citation quality (Zhang et al., 2025b). Similarly, Dai et al. (2025) categorize legal reasoning capabilities into fundamental information retrieval, legal principle inference, and advanced legal applications.

Ethics. Existing work also emphasizes the need to evaluate whether LLMs behave in ethically appropriate ways in legal contexts, and several studies argue that such evaluation requires domain-specific metrics rather than generic AI ethics principles. For instance, Wright (2020) calls for context-sensitive, measurable criteria that reflect how AI systems are deployed and used. Zhang et al. (2024) propose an operational framework for ethical evaluation in legal LLMs, covering legal instruction following, legal knowledge consistency, and robustness to misleading or adversarial prompts.

4.2 Performance of Legal Agent

Performance across legal tasks Across legal tasks, agent-based systems show strong but uneven performance, with the clearest gains appearing in legal search, reasoning, and compliance. In legal search and reasoning, systems such as L4M (Chen et al., 2025d) achieve strong results, including an F1 score of 0.75 for specific provision prediction. In litigation, agents such as LegalSim (Badhe, 2025) deliver more moderate performance, with reported win rates around 0.742, indicating meaningful capability but also the greater difficulty of adversarial and strategy-intensive tasks. Compliance-oriented agents also perform well: Multi-Agent Legal Verifier Systems (Nguyen et al., 2025a) reach an F1 score of 0.725 and accuracy of 0.72. By

contrast, performance on commercial and transactional tasks is more variable: Synedrion (Gogani-Khiabani et al., 2025) achieves high pass rates on simple tasks but drops to 45% in complex settings, while ChatLaw (Cui et al., 2023) attains only a moderate average score of 60.08, with several subcategories below 50. Overall, these results suggest that legal agents currently perform best on structured, grounded, and verification-friendly tasks, while more open-ended and complex legal problem-solving remains comparatively challenging.

Improvements over vanilla LLMs Across the reported studies, legal agents consistently outperform vanilla LLMs across multiple evaluation dimensions. In final output correctness, agentic methods improve over strong general-purpose baselines, with Debate-Feedback (Chen et al., 2025e) increasing F1 by about 12% over GPT-4o and PAKTON (Raptopoulos et al., 2025) improving accuracy by about 19.7%. Agents also substantially mitigate hallucination-related reasoning failures, with MALR (Yuan et al., 2024) improving reasoning accuracy by 67% over Qwen-2-1.5B and 4.1% over Qwen-2-72B. Preference-based evaluations show a similar trend: ChatLaw (Cui et al., 2023) achieves a 66% win rate over GPT-4, while LawLuo (Sun et al., 2024) attains a 52% win rate over GPT-3.5.

5 Challenges and Future Directions

5.1 Remaining Challenges

Despite recent progress, legal agents still face substantial domain-specific challenges. In legal search and reasoning, a central difficulty is reasoning under interdependent legal constraints dispersed across long documents (Raptopoulos et al., 2025); these systems also remain sensitive to factual perturbations, often failing to revise conclusions reliably under counterfactual edits (Han et al., 2025). In litigation and dispute resolution, current agents still struggle to reproduce realistic courtroom conditions and to remain procedurally stable across multi-turn interactions, especially in dynamic, long-horizon settings involving evolving evidence (Jia et al., 2025; Chen et al., 2025a). In consultation and transactional practice, practical usability can be constrained by system latency (Watson et al., 2025) and by the need to satisfy professional-responsibility requirements such as competence and confidentiality.

Current evaluation of legal LLMs and agent-based systems remains limited along several core dimensions. First, many benchmarks still undermeasure law-specific properties, especially procedural compliance, fairness, and bias (Jia et al., 2025; Xue et al., 2025b). Second, current datasets still provide limited coverage beyond Chinese and English-speaking jurisdictions. Third, agent-based evaluation often relies on simulated legal actors and prompt-sensitive setups whose realism, stability, and comparability remain insufficiently validated (Chen et al., 2025a; Jia et al., 2025), while principled evaluation frameworks for multi-agent coordination and workflow coherence are still largely missing (Jiang and Yang, 2025; Li et al., 2025a).

5.2 Future Directions

Future research should design legal agents with stronger legal-specific structure and validation. For legal search and reasoning, long-context retrieval should be coupled with intermediate representations that explicitly track cross-references and temporal validity, together with step-level verifiers (Wang and Yuan, 2025). For litigation, future systems should monitor the procedural state over multi-turn interactions (Jia et al., 2025; Chen et al., 2025a). In consultation and transactional settings, low-latency architectures with bounded tool use and compliance-aware safeguards will be important for practical deployment (Watson et al., 2025). Finally, evaluation should move beyond final-answer accuracy toward legal-specific protocols that measure procedural legality, citation faithfulness, fairness, jurisdictional robustness, and multi-agent coordination under realistic, adversarial, and cross-jurisdiction settings (Xue et al., 2025b; Li et al., 2025a; Jing et al., 2025).

6 Conclusions

In this survey, we provided a comprehensive examination of LLM agents as a solution for complex legal tasks. We systematically analyzed the technical transition toward agentic architectures and presented a structured taxonomy of applications ranging from litigation assistance to workflow automation. While these agents effectively mitigate persistent issues like hallucination and outdated information through modular execution and tool usage, several challenges regarding long-horizon reliability, procedural correctness, and citation fidelity remain. Lastly, we outline promising directions for future research.

Limitations

This survey has several limitations. First, our survey focuses primarily on LLM-based agents for legal tasks, and our coverage skews toward English-language publications and systems. Although we tried to cover varying jurisdictions, including common law jurisdictions (e.g. the US and UK), India, and Chinese legal systems. However, the scope may potentially under represents non-English paper or products in other regions. Second, our coverage of commercial legal AI agents relies primarily on publicly available information such as press releases, product documentation, and vendor websites. We lack access to proprietary system architectures, internal evaluations, or user adoption data. Consequently, our characterization of commercial systems may be incomplete, as we cannot independently verify vendor claims about system capabilities or performance. Third, given the complexity of legal work, distinctions between task categories are not always clear-cut, and real-world practice often encompasses multiple task types simultaneously. Although our categorization of legal domains and tasks is grounded in established legal practice, the boundaries between categories such as litigation support and legal consultation, or compliance and fraud detection, are inherently fluid, and certain systems may reasonably be classified under multiple categories.

References

Bhavik Agarwal, Hemant Sunil Jomraj, Simone Kaplunov, Jack Krolick, and Viktoria Rojkova. 2025. Ragulating compliance: A multi-agent knowledge graph for regulatory qa. *arXiv preprint arXiv:2508.09893*.

AgentiveAIQ. 2025. Ai agents that actually work for your business. Company website; Accessed December 29, 2025.

Aiqlabs.ai. 2025. Your ai transformation partner. Company website; Accessed December 29, 2025.

Akira AI. 2025. Compliance officer ai agents. Company website; Accessed December 30, 2025.

Aline. 2025. Run trusted legal ai agents. Company website; Accessed December 30, 2025.

Anthropic. 2023. Core Views on AI Safety: When, Why, What, and How. Accessed: December 30, 2025.

Anthropic. 2024. Building effective ai agents. Accessed December 30, 2025.

A&O Shearman. 2025. A&o shearman and harvey to roll out agentic ai agents targeting complex legal workflows. Company website; Accessed December 30, 2025.

Henrik Axelsen, Valdemar Licht, and Jan Damsgaard. 2025. Agentic ai for financial crime compliance. *arXiv preprint arXiv:2509.13137*.

Sanket Badhe. 2025. Legalsim: Multi-agent simulation of legal systems for discovering procedural exploits. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 370–381.

Zachary R Baker and Zarif L Azher. 2024. Simulating the us senate: An llm-driven agent approach to modeling legislative behavior and bipartisanship. *arXiv preprint arXiv:2406.18702*.

Rohan Bhambhoria, Samuel Dahan, Jonathan Li, and Xiaodan Zhu. 2024. Evaluating AI for Law: Bridging the Gap with Open-Source Solutions. *Preprint*, arXiv:2404.12349.

Jillian Bommarito, Daniel Martin Katz, and Michael James Bommarito. 2025. Governing AI Agents: Risk, Compliance, and Accountability in Law and Finance.

Caseflood.ai. 2024. Caseflood.ai: The ai support staff for law firms. Accessed December 30, 2025.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Hamid Alinejad-Rokny, Shiwen Ni, and Min Yang. 2025a. AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5850–5865, Vienna, Austria. Association for Computational Linguistics.

Junjie Chen, Haitao Li, Minghao Qin, Yujia Zhou, Yanxue Ren, Wuyue Wang, Yiqun Liu, Yueyue Wu, and Qingyao Ai. 2025b. Simulating dispute mediation with llm-based agents for legal research. *arXiv preprint arXiv:2509.06586*.

Junjie Chen, Haitao Li, Jingli Yang, Yiqun Liu, and Qingyao Ai. 2025c. Enhancing LLM-Based Agents via Global Planning and Hierarchical Execution. *arXiv preprint*.

Linze Chen, Yufan Cai, Zhe Hou, and Jinsong Dong. 2025d. Towards Trustworthy Legal AI through LLM Agents and Formal Reasoning. *Preprint*, arXiv:2511.21033.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. 2025e. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 462–470, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*.

Zichen Chen, Jianda Chen, Jiaao Chen, and Misha Sra. 2025f. From tasks to teams: A risk-first evaluation framework for multi-agent llm systems in finance. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, pages 2454–2469, Rio de Janeiro Brazil. ACM.

Clearbrief. 2025. Cite facts, not fake cases. Company website; Accessed December 30, 2025.

ContractPodAi. 2025. Leah legal. Company website; Accessed January 2, 2026.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese Legal Large Language Models Benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE. Association for Computational Linguistics.

Mark J. Davenport. 2025. Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation. *Open Journal of Modern Linguistics*, 15(2):232–280.

Hudson de Martim. 2025. Deterministic legal agents: A canonical primitive api for auditable reasoning over temporal knowledge graphs. *arXiv preprint arXiv:2510.06002*.

DeepJudge AI. 2025. Precision ai search for legal teams. Company website; Accessed December 30, 2025.

Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. 2025. Large Language Models in Legal Systems: A Survey. *Humanities and Social Sciences Communications*, 12(1):1977.

Prathamesh Devadiga, Omkaar Jayadev Shetty, and Pooja Agarwal. 2025. Samvad: A multi-agent system for simulating judicial deliberation dynamics in india. *Preprint*, arXiv:2509.03793.

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Cha, Arijit Ghosh Chowdhury, Prashanth Kallur Ramaswamy, Jeremy Roghair, Hannah R. Marlowe, Carina Suzana Negreanu, Kitty Boxall, and Diana Mincu. 2025. LeMAJ (Legal LLM-as-a-Judge): Bridging Legal Reasoning and LLM Evaluation. *Preprint*, arXiv:2510.07243.

Chenhao Fang, Yanqing Peng, Rajeev Rao, Matt Sarmiento, Wendy Summer, Arya Pudota, Alex Goncalves, Jordi Mola, and Hervé Robert. 2025. Privacy artifact connector (pact): Embedding enterprise artifacts for compliance ai agents. *arXiv preprint arXiv:2507.21142*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Filippo Ferraris, Davide Audrito, Luigi Di Caro, and Cristina Poncibò. 2025. The architecture of language: Understanding the mechanics behind LLMs. *Cambridge Forum on AI: Law and Governance*, 1:e11.

Antonino Ferraro, Gian Marco Orlando, and Diego Russo. 2025. Generative agent-based modeling with large language models for insider threat detection. *Engineering Applications of Artificial Intelligence*, 157:111343.

Sergio Ferro, Jenny Tai, Raywen Tsai, Salone Verma, Jenny Ma, Nora Skjerdal, and Martin Lopatka. 2025. Rawlsian agents: An application of large language models (llm) to forge fairer bilateral agreements.

Jan Fillies, Theodoros Mitsikas, Ralph Schäfermeier, and Adrian Paschke. 2025. Scalable, context-aware nlp moderation for child safety: A multi-agent ethical and legal compliance framework. In *International Joint Conference on Rules and Reasoning 2025*.

Sean Fitzpatrick. 2025. The next chapter in legal tech innovation: Introducing protégé™. Company website; Accessed January 2, 2026.

Flank AI. 2025. Delegate routine legal processes to autonomous ai agents. Company website; Accessed December 30, 2025.

Sina Gogani-Khiabani, Ashutosh Trivedi, Diptikalyan Saha, and Saeid Tizpaz-Niari. 2025. An llm agentic approach for legal-critical software: A case study for tax prep software. *arXiv preprint arXiv:2509.13471*.

Yu Gu, Lanxin Jiang, Jun Dai, and Miklos Vasarhelyi. 2025. An llm-based agentic system for greenwashing detection. SSRN Electronic Journal.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Preprint*, arXiv:2308.11462. Comment: 143 pages, 79 tables, 4 figures.

Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K. Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. 2025. CourtReasoner: Can LLM agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35279–35294, Suzhou, China. Association for Computational Linguistics.

Harvey AI. 2025. Introducing agents in harvey. Company website; Accessed December 30, 2025.

Kasper Raupach Haurum, Ruiqi Ma, and Wen Long. 2024. Real estate with ai: An agent based on langchain. *Procedia Computer Science*, 242:1082–1088.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.

Zhitian Hou, Zihan Ye, Nanli Zeng, Tianyong Hao, and Kun Zeng. 2025. Large language models meet legal artificial intelligence: A survey. *arXiv preprint arXiv:2509.09969*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer LLaMA Technical Report. *arXiv preprint*.

Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. 2025. TRIDENT: Benchmarking LLM Safety in Finance, Medicine, and Law. *Preprint*, arXiv:2507.21134.

Ruba Islayem, Senay Gebreab, Walaa AlKhader, Ahmad Musamih, Khaled Salah, Raja Jayaraman, and Muhammad Khurram Khan. 2025. Using large language models for enhanced fraud analysis and detection in blockchain based health insurance claims. *Scientific Reports*, 15(1):29763.

Jace AI. 2025. Emails so good, you just press send! Company website; Accessed December 29, 2025.

Gautam Jajoo, Pranjal A Chitale, and Saksham Agarwal. 2025. Masca: Llm based-multi agents system for credit assessment. *arXiv preprint arXiv:2507.22758*.

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Yun Song, and Zhongyu Wei. 2025. Ready Jurist One: Benchmarking Language Agents for Legal Intelligence in Dynamic Environments. *Preprint*, arXiv:2507.04037.

Cong Jiang and Xiaolei Yang. 2025. AgentsBench: A Multi-Agent LLM Simulation Framework for Legal Judgment Prediction. *Systems*, 13(8):641.

Huihao Jing, Wenbin Hu, Hongyu Luo, Jianhui Yang, Wei Fan, Haoran Li, and Yangqiu Song. 2025. Maslegalbench: Benchmarking multi-agent systems in deductive legal reasoning. *arXiv preprint arXiv:2509.24922*.

Juro. 2025. Intelligent contracting is here. Company website; Accessed January 2, 2026.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *arXiv preprint*.

Lucio La Cava and Andrea Tagarelli. 2025. Safeguarding decentralized social media: Llm agents for automating community rule compliance. *Online Social Networks and Media*, 48:100319.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Yingxin Lai, Zitong Yu, Jun Wang, Linlin Shen, Yong Xu, and Xiaochun Cao. 2025. Agent4faceforgery: Multi-agent llm framework for realistic face forgery detection. *Preprint*, arXiv:2509.12546.

H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors. 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, volume 33. Curran Associates, Inc.

LawBot.ai. 2025. Your ai-powered legal voice assistant. Company website; Accessed December 29, 2025.

Legora. 2025. Orchestrate complex legal tasks. Company website; Accessed December 30, 2025.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2025a. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Preprint*, arXiv:2409.20288.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023. LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset. *Preprint*, arXiv:2310.17609.

Hao Li, Ruoyuan Gong, and Hao Jiang. 2025b. Political actor agent: simulating legislative system for roll call votes prediction with large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Wenhao Li, Selvakumar Manickam, Yung-wey Chong, and Shankar Karuppayah. 2025c. Phishdebate: An llm-based multi-agent framework for phishing website detection. *arXiv preprint arXiv:2506.15656*.

Yunhan Li and Gengshen Wu. 2025. LegalEval-Q: A New Benchmark for The Quality Evaluation of LLM-Generated Legal Text. *Preprint*, arXiv:2505.24826.

Litera. 2025. Litera introduces lito: The next evolution of agentic ai in legal fully built on your firm's trusted experience. Accessed January 2, 2026.

Bulou Liu, Yiran Hu, Yueyue Wu, Yiqun Liu, Fan Zhang, Chenliang Li, Min Zhang, Shaoping Ma, and Weixing Shen. 2023. Investigating conversational agent action in legal case retrieval. In *European Conference on Information Retrieval*, pages 622–635. Springer.

Bulou Liu, Yueyue Wu, Fan Zhang, Yiqun Liu, Zhihong Wang, Chenliang Li, Min Zhang, and Shaoping Ma. 2022. Query generation and buffer mechanism: Towards a better conversational agent for legal case retrieval. *Information Processing & Management*, 59(5):103051.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shuang Liu, Zelong Li, Ruoyun Ma, Haiyan Zhao, and Mengnan Du. 2025. Contracteval: Benchmarking llms for clause-level legal risk identification in commercial contracts. *EMNLP 2025 workshop on Natural Legal Language Processing (NLLP)*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Luminance. 2025. Luminance launches "agent lumi", automating legal work across the enterprise with agentic ai. Company website; Accessed January 2, 2025.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Marios Evangelos Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. 2024. A large language model agent based legal assistant for governance applications. In *International Conference on Electronic Government*, pages 286–301. Springer.

Shashank Mangla, Chris Hokamp, Jack Boylan, Demian Gholipour Ghalandari, Yuuv Jauhari, Lauren Cassidy, and Oisin Duffy. 2025. Negotiationgym: Self-optimizing agents in a multi-agent social simulation environment. *Preprint*, arXiv:2510.04368.

Mohammad Matt. 2025. Ai agents in legal: What they are and why they matter now. Accessed December 29, 2025.

Fabrizio Maturo, Donato Riccio, Andrea Mazzitelli, Giuseppe Bifulco, Francesco Paolone, and Iulia Brezeanu. 2025. Arcadia: Scalable causal discovery for corporate bankruptcy analysis using agentic ai. *arXiv preprint arXiv:2512.00839*.

Chunyun Meng, Cheng Tang, Yuki Todo, and Weiping Ding. 2025a. Multi-granular legal information fusion with adversarial compensation: A hierarchical and logic-aware framework for robust case retrieval. *Knowledge-Based Systems*, page 113964.

Lingyi Meng, Maolin Liu, Hao Wang, Yilan Cheng, Qi Yang, and Idlkaid Mohanmmed. 2025b. Building from scratch: a multi-agent framework with human-in-the-loop for multilingual legal terminology mapping. *Artificial Intelligence and Law*, pages 1–40.

Emil Mirzayev, Bart Vanneste, and Marco Testoni. 2025. Artificial agents and the evaluation of m&as. *Available at SSRN*.

Muhammad Ahmed Mohsin, Muhammad Umer, Ahsan Bilal, Zeeshan Memon, Muhammad Ibtsaam Qadir, Sagnik Bhattacharya, Hassan Rizwan, Abhiram R. Gorle, Maahe Zehra Kazmi, Ayesha Mohsin, Muhammad Usman Rafique, Zihao He, Pulkit Mehta, Muhammad Ali Jamshed, and John M. Cioffi. 2025. On the Fundamental Limits of LLMs at Scale. *arXiv preprint*.

Gyuyeon Na, Minjung Park, Hyeonjeong Cha, and Sangmi Chai. 2025. Human-centered llm-agent system for detecting anomalous digital asset transactions. *arXiv preprint arXiv:2510.20102*.

Prathamesh Vasudeo Naik, Naresh Kumar Dintakurthi, Zhanghao Hu, Yue Wang, and Robby Qiu. 2025. Coinvestigator ai: The rise of agentic ai for smarter, trustworthy aml compliance narratives. *arXiv preprint arXiv:2509.08380*.

Ha-Thanh Nguyen, Wachara Fungwacharakorn, and Ken Satoh. 2025a. Multi-agent legal verifier systems for data transfer planning. *Preprint*, arXiv:2511.10925.

Ngoc Tuong Vy Nguyen, Felix D Childress, and Yunting Yin. 2025b. Debate-driven multi-agent llms for phishing email detection. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE.

Catherine Gage O'Grady and Casey OG. 2024. Agentic workflows in the practice of law—ai agents as ethics counsel. *Arizona Legal Studies Discussion Paper*, pages 25–03.

Taejin Park. 2024. Enhancing anomaly detection in financial markets with an llm-based multi-agent framework. *arXiv preprint arXiv:2403.19735*.

Pearson Labs. 2024. Pearson labs - we build ai agents to help law firms execute corporate transactions. Accessed December 30, 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. In *Advances in Neural Information Processing Systems*, volume 37, pages 111715–111759. Curran Associates, Inc.

Ravi Shanker Raju, Swayambhoo Jain, Bo Li, Jonathan Lingjie Li, and Urmish Thakker. 2024. Constructing Domain-Specific Evaluation Sets for LLM-as-a-judge. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 167–181, Miami, Florida, USA. Association for Computational Linguistics.

Petros Raptopoulos, Giorgos Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2025. Pakton: A multi-agent framework for question answering in long legal agreements. *arXiv preprint arXiv:2506.00608*.

Regology. 2025. Purpose-built ai agents for enterprise compliance. Company website; Accessed December 30, 2025.

Relevance AI. 2025. Recruit your ai inbox manager agent. Company website; Accessed December 29, 2025.

Thomson Reuters. 2025. Cocounsel legal. Company website; Accessed January 2, 2026.

Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137, Singapore. Association for Computational Linguistics.

Albert Sadowski and Jaroslaw A Chudziak. 2025. On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2535–2545.

Sagar Srinivas Sakhinana, Venkataramana Runkana, and 1 others. 2024. Towards automated patent workflows: Ai-orchestrated multi-agent framework for intellectual property management and analysis. In *NeurIPS 2024 Workshop on Open-World Agents*.

Jaromir Savelka and Kevin Ashley. 2022. Legal information retrieval for understanding statutory terms. *Artificial Intelligence and Law*, 30.

Rachel Scanlon, Miklos Orban, and D2 Legal Technology. 2025. A new era of artificial intelligence with agentic AI. Accessed: December 30, 2025.

Stanford Law School. 2025. Rethinking Human – AI Agent Collaboration for the Knowledge Worker.

Benjamyn Scott. 2023. A case of 'AI hallucination' in the air. Accessed: January 1, 2026.

Shraddha Pradipbhai Shah and Aditya Vilas Deshpande. 2025. Enforcing cybersecurity constraints for llm-driven robot agents for online transactions. *arXiv preprint arXiv:2503.15546*.

Peizhang Shao, Linrui Xu, Jinxi Wang, Wei Zhou, and Xingyu Wu. 2025. When Large Language Models Meet Law: Dual-Lens Taxonomy, Technical Advances, and Ethical Governance. *arXiv preprint*.

Ryan Shea and Zhou Yu. 2025. Autospec: An agentic framework for automatically drafting patent specification. *arXiv preprint arXiv:2509.19640*.

ShengbinYue ShengbinYue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing Huang, and Zhongyu Wei. 2025. Multi-agent simulator drives language models for legal intensive interaction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6537–6570.

Juanming Shi, Qinglang Guo, Yong Liao, and Shenglin Liang. 2024. Legalgpt: legal chain of thought for the legal large language model multi-agent framework. In *International Conference on Intelligent Computing*, pages 25–37. Springer.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. LawLLM: Law Large Language Model for the US Legal System. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. 2024. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *Preprint*, arXiv:2408.08902.

Yumei Song, Yongbin Qin, Ruizhang Huang, Yanping Chen, and Chuan Lin. 2025. Legal text summarization via judicial syllogism with large language models. *Journal of King Saud University Computer and Information Sciences*, 37(5):111.

Spellbook. 2025. Beyond chat: Ai that works for you. Company website; Accessed December 30, 2025.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation. *arXiv preprint arXiv:2407.16252*.

Amulya Suravarjhula, Rashi Chandrashekhar Agrawal, Sakshi Jayesh Patel, and Rahul Gupta. 2025. Retrieval-augmented multi-agent system for rapid statement of work generation. *Preprint*, arXiv:2508.07569.

ODIN AI AI that makes you shine. 2025. din ai. Company website; Accessed December 29, 2025.

V7 Labs. 2025a. Ai compliance verification agent ensure every document is compliant. Company website; Accessed December 30, 2025.

V7 Labs. 2025b. Ai legal workflow automation agent automate the operational side of legal work. Company website; Accessed December 29, 2025.

Michelle Vaccaro, Michael Caosun, Harang Ju, Sinan Aral, and Jared R Curhan. 2025. Advancing ai negotiations: New theory and evidence from a large-scale autonomous negotiations competition. *arXiv preprint arXiv:2503.06416*.

Adela Nedisan Videsjorden, Nikolay Nikolov, Carl-Henrik Lien, Arda Goknil, Sagar Sen, Hui Song, Ahmet Soylu, and Dumitru Roman. 2025. Positioning llm-enabled agents as legal compliance aides for data pipelines. In *International Joint Conference on Rules and Reasoning*, pages 227–236. Springer.

Vincent. 2025. Ai engineered for lawyers. Company website; Accessed January 2, 2026.

Alisa Vinogradova, Vlad Vinogradov, Dmitrii Radkevich, Ilya Yasny, Dmitry Kobyzev, Ivan Izmailov, Katsiaryna Yanchanka, Roman Doronin, and Andrey Doronichev. 2025. Llm-based agents for competitive landscape mapping in drug asset due diligence. *arXiv preprint arXiv:2508.16571*.

vLex Team. 2025. vLex: Transforming Legal Workflows: Vincent AI Multimodal A. Accessed: December 31, 2025.

Voiceflow. 2025. Ai agent for law firms. Voiceflow company webstie; Accessed December 29, 2025.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. 2025a. A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy. *arXiv preprint*.

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, and 1 others. 2024a. Autopatent: a multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*.

Xin Wang, Yifan Zhang, Xiaojing Zhang, Longhui Yu, Xinna Lin, Jindong Jiang, Bin Ma, and Kaicheng Yu. 2024b. PatentAgent: Intelligent agent for automated pharmaceutical patent analysis. *arXiv preprint arXiv:2410.21312*.

Yiding Wang, Yuxuan Chen, Fanxu Meng, Xifan Chen, Xiaolei Yang, and Muhan Zhang. 2025b. Law in silico: Simulating legal society with llm-based agents. *arXiv preprint arXiv:2510.24442*.

Ziqi Wang and Boqin Yuan. 2025. L-mars: Legal multi-agent workflow with orchestrated reasoning and agentic search. *arXiv preprint arXiv:2509.00761*.

William Watson, Nicole Cho, Nishan Srishankar, Zhen Zeng, Lucas Cecchi, Daniel Scott, Suchetha Siddagangappa, Rachneet Kaur, Tucker Balch, and Manuela Veloso. 2025. LAW: Legal agentic workflows for custody and fund services contracts. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 583–594, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.

Lilian Weng. 2023. LLM Powered Autonomous Agents. Accessed December 28, 2025.

Steven A. Wright. 2020. AI in the Law: Towards Assessing Ethical Risks. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2160–2169, Atlanta, GA, USA. IEEE.

Yueqing Xi, Yifan Bai, Huasen Luo, Weiliang Wen, Hui Liu, and Haoliang Li. 2025a. Hybrid retrieval-augmented generation agent for trustworthy legal question answering in judicial forensics. *Preprint*, arXiv:2511.01668.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 10 others. 2025b. The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences*, 68(2):121101.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint*.

Yinuo Xue, Eric Spero, Yun Sing Koh, and Giovanni Russello. 2025a. Multiphishguard: An llm-based multi-agent system for phishing email detection. *arXiv preprint arXiv:2505.23803*.

Zongyue Xue, Siyuan Zheng, Shaochun Wang, Yiran Hu, Shenran Wang, Yuxin Yao, Haitao Li, Qingyao Ai, Yiqun Liu, Yun Liu, and Weixing Shen. 2025b. JustEva: A Toolkit to Evaluate LLM Fairness in Legal Knowledge Inference. *Preprint*, arXiv:2509.12104.

Se Yang, Zhe Yang, Yutong Liu, and Hongtao Wang. 2025. From single-agent to multi-agent: a comprehensive review of llm-based legal agents. Accessed December 7, 2025.

Junchi Yao, Jianhua Xu, Tianyu Xin, Ziyi Wang, Shenzhe Zhu, Shu Yang, and Di Wang. 2025. Is your llm-based multi-agent a reliable real-world planner? exploring fraud detection in travel planning. *arXiv preprint arXiv:2505.16557*.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, and 1 others. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7577–7597.

ZBrain AI. 2025. Zbrain ai agents: Streamlining enterprise operations. Company website; Accessed December 29, 2025.

Kaiyuan Zhang, Jiaqi Li, Yueyue Wu, Haitao Li, Cheng Luo, Shaokun Zou, Yujia Zhou, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025a. Chinese court simulation with llm-based agent system. *arXiv preprint arXiv:2508.17322*.

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2025b. CitaLaw: Enhancing LLM with Citations in Legal Domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11183–11196, Vienna, Austria. Association for Computational Linguistics.

Li Zhang and Kevin D Ashley. 2025. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation. *arXiv preprint arXiv:2506.02992*.

Ruizhe Zhang, Haitao Li, Yueyue Wu, Qingyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024. Evaluation Ethics of LLMs in Legal Domain. *Preprint*, arXiv:2403.11152. Comment: 10 pages, in processing of ACL 2024.

Deyu Zhou, Yuqi Hou, Xiao Xue, Xudong Lu, Qingzhong Li, and Lizhen Cui. 2025. Llm-empowered agents simulation framework for scenario generation in service ecosystem governance. *arXiv preprint arXiv:2509.01441*.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. 2025a. Safescientist: Toward risk-aware scientific discoveries by llm agents. *arXiv preprint arXiv:2505.23559*.

Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. 2025b. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Shitong Zhu, Chenhao Fang, Derek Larson, Neel Reddy Pochareddy, Rajeev Rao, Sophie Zeng, Yanqing Peng, Wendy Summer, Alex Goncalves, Arya Pudota, and 1 others. 2025c. Compliance brain assistant: Conversational agentic ai for assisting compliance tasks in enterprise environments. *arXiv preprint arXiv:2507.17289*.

A Representative Legal Agent Structure

Table 1: Representative agentic architectures used in legal-domain systems and their core properties.

Agentic Architecture	Example Agents	Author, Year	Properties
Reflection	REFLECTIVE MULTI-AGENT	Zhang and Ashley (2025)	Generate → Self-critique → Revise
Debate Agent	DEBATE-FEEDBACK	Chen et al. (2025e)	Multiple proposing agents + centralized judge agent
Plan-and-Execute	LEGAL VERIFIER	Nguyen et al. (2025a)	Planner coordinates → Executor agents perform subtasks
Agent + RL	LEGALSIM	Badhe (2025)	Multi-agent simulation + RL policy learning
Tool-Augmented Agent	COMPLIANCE BRAIN ASSISTANT	Zhu et al. (2025c)	LLM + Retrieval/Tool Router + External APIs
Role-based Multi-Agent	CHATLAW	Cui et al. (2023)	Multiple agents with distinct roles collaborating

B Legal Agentic Systems

Table 2: Comprehensive overview of representative LLM agents for legal tasks (literature and academic report).

Domain	Agentic System	Author, Year	Agentic Design Pattern	Task
Legal Retrieval	L-MARS	Wang and Yuan (2025)	Reasoning–search–verification loop	Multiple choice legal question
	L4M	Chen et al. (2025d)	Neural-symbolic workflow	Chinese legal case retrieval
Legal Reasoning	L-MARS	Wang and Yuan (2025)	Reasoning–search–verification loop	Multiple choice legal question
	MLTDB	Meng et al. (2025b)	Human-in-the-loop on multi-agent	Multilingual legal terminology extraction
	REFLECTIVE MULTI-AGENT	Zhang and Ashley (2025)	Reflective multi-agent	Legal Argument Generation
	DETERMINISTIC LEGAL AGENTS	de Martim (2025)	Canonical primitive API, point-in-time retrieval	Legal Q&A
	MALR	Yuan et al. (2024)	Non-parametric learning multi-agent	Confusing Charge Prediction Task
Judgment Prediction	DEBATE-FEEDBACK	Chen et al. (2025e)	Multi-agent debate & verification	Trial Prediction & Article Prediction
	AGENTSBENCH	Jiang and Yang (2025)	Multi-agent simulation of judicial bench's discussion	Criminal Prison-Term Prediction
Litigation & Dispute Resolution	AGENTCOURT	Chen et al. (2025a)	Adversarial evolutionary multi-agent simulation	Civil Court Debate & Legal Reasoning
	AGENTSCOURT	He et al. (2024)	Multi-agent simulation of trial procedures	Judicial Decision-Making Process
	SIMCOURT	Zhang et al. (2025a)	Multi-agent simulation of trial procedures	Criminal Judgement Prediction & Trial Process
	SAMVAD	Devadiga et al. (2025)	Multi-agent simulation of trial procedures	Criminal Judgement Prediction & Trial Process
	LEGALSIM	Badhe (2025)	Multi-agent simulation of adversarial legal proceedings	Litigation procedural strategy discovery
	AGENTMEDIATION	Chen et al. (2025b)	Multi-agent simulation of dispute mediation process	Civil Dispute Mediation

Continued on next page

Table 2 – *Continued from previous page*

Domain	Agentic System	Author, Year	Agentic Design Pattern	Task
	MASER	ShengbinYue et al. (2025)	Multi-agent simulation of legal service interactions	Legal consultation & Complaint drafting
Compliance	CBA	Zhu et al. (2025c)	Combination of RAG and agentic mode	Legal Q&A
	PACT	Fang et al. (2025)	Embeddings-driven graph	Enterprise search
Fraud Detection	HCLA	Na et al. (2025)	Multi-role conversational agents	Anomalous transaction detection
	M-SAEA	Chen et al. (2025f)	Multi-agent system with auditor role	Financial risk evaluation
	CO-INVESTIGATOR AI	Naik et al. (2025)	Multi-role agents and Agent-as-a-Judge loop	Anti-Money Laundering compliance
	FINCON	Yu et al. (2024)	Multi-agent system with risk control module	Financial decision making
	MASCA	Jajoo et al. (2025)	Multi-agent system with risk control module	Credit assessment
	AUDIT-LLM	Song et al. (2024)	Multi-agent system for detection simulation	Insider threat detection
	GABM	Ferraro et al. (2025)	Hierarchical pipeline with specialists and supervisor	Insider threat detection
	MULTIPHISHGUARD	Xue et al. (2025a)	Multi-role agents and adversarial simulation	Phishing detection
	PHISHDEBATE	Li et al. (2025c)	Multi-role debating agents	Phishing detection
	WANDAPLAN	Yao et al. (2025)	Multi-role agentic pipeline	Misinformation and fraud detection
	SAFESCIENTIST	Zhu et al. (2025a)	Research simulation with defense and attack modules	Scientific discovery safety
	AGENT4FACEFORGERY	Lai et al. (2025)	Behavior simulation and adaptive rejection sampling	Face forgery detection
Regulation	LAW IN SILICO	Wang et al. (2025b)	Multi-agent law system simulation	Legal theory study
	GOVSIM	Piatti et al. (2024)	Cooperation simulation and moral reasoning	Social sustainability study
	LEGALSIM	Badhe (2025)	Multi-agent pipeline for legal system simulation	Legislation improvement
	POLITICAL ACTOR AGENT	Li et al. (2025b)	Legislative behavior simulation with role-playing agents	Roll-call vote prediction
	RAWLSIAN AGENTS	Ferro et al. (2025)	Bilateral agent negotiation under law theory	Contract negotiation
Legal Consultation	CHATLAW	Cui et al. (2023)	Role-specialized workflow (MoE)	Legal Q&A
	LAWLUO	Sun et al. (2024)	Multi-role conversational agents	Legal Q&A
	LEGALGPT	Shi et al. (2024)	Chain-of-thought orchestration with role-based agents	Legal Q&A
Contract Analysis	PAKTON	Raptopoulos et al. (2025)	Document-centric agent pipeline	Contract / legal agreements
	LAW	Watson et al. (2025)	Legal agentic workflow for contracts	Custody and fund services contract analysis
Transaction	NEGOTIATIONGYM	Mangla et al. (2025)	Multi-agent negotiation simulation	Business negotiation
	ARCADIA	Maturo et al. (2025)	Agentic causal reasoning pipeline	Corporate bankruptcy analysis
	PATENTAGENT	Wang et al. (2024b)	Role-specialized patent analysis agents	Patent application & filing

Continued on next page

Table 2 – *Continued from previous page*

Domain	Agentic System	Author, Year	Agentic Design Pattern	Task
	AUTOPATENT	Wang et al. (2024a)	Multi-agent patent drafting pipeline	Patent application & filing
	AUTOSPEC	Shea and Yu (2025)	Agentic specification generation workflow	Patent application & filing
	SYNEDRION	Gogani-Khiabani et al. (2025)	Multi-agent deliberation and synthesis	Collaborative decision-making and agreement synthesis

Table 3: Commercial AI agent products for legal tasks (industrial products).

Task Category	Product / Company	Website	Use Cases
Comprehensive Legal Tasks	HARVEY AGENT	Harvey AI (2025)	Multi-model agents for legal research, contract review, due diligence, litigation support
	COCOUNSEL LEGAL	Reuters (2025)	Legal research by Deep Research with agentic planning, drafting with guided workflows, bulk document review
	LEXISNEXIS PROTÉGÉ	Fitzpatrick (2025)	Legal research, drafting, document analysis through multiple agents
	VINCENT	Vincent (2025)	Legal research, litigation, transactions through Agentic Workflow Engine with 20+ pre-built workflows, cross-jurisdictional reasoning
	LUMINANCE	Luminance (2025)	Autonomous contract review, compliance analysis, negotiation
Legal Search	DEEPJUDGE	DeepJudge AI (2025)	Legal reasoning and multi-step legal search or information retrieval
Litigation	CLEARBREIF	Clearbrief (2025)	Litigation focused tasks; offering Microsoft Add-ons for legal document processing
Compliance & Regulatory	REGOLOGY	Regology (2025)	AI agents for research, regulatory change, and compliance workflows
	V7LABS	V7 Labs (2025a)	Regulatory cross-referencing and compliance verification
	AKIRA	Akira AI (2025)	Automates regulatory monitoring, analyzes risks, and ensures timely policy adjustments
Contract & Transaction	FLANK AGENTS	Flank AI (2025)	Autonomous enterprise legal system covering reviewing contracts, answering compliance questions, and drafting legal documents
	SPEELBOOK AI ASSOCIATE	Spellbook (2025)	Streamlined legal agreement/contract revision for commercial transactions
	ALINE	Aline (2025)	Contract lifecycle and AI-assisted contract automation for in-house legal teams
	PEARSON	Pearson Labs (2024)	Automating corporate transactions, including M&A due diligence and financing
	LEGORA	Legora (2025)	Customizable agentic workflows for lawyers covering varying perspectives of commercial transactions
	LEAH	ContractPodAi (2025)	Contract lifecycle management, legal workflows, procurement through agentic AI platform
	JURO	Juro (2025)	Contract lifecycle management, contract drafting, negotiation through workflow-embedded playbook redlining, risk surfacing
	LITO	Litera (2025)	Contract review, due diligence, and contract summaries
Supporting Legal Tasks	AGENTIVEAIQ	AgentiveAIQ (2025)	Client triage & internal insights delivery
	AIQLABS	Aiqlabs.ai (2025)	Invoice automation, knowledge mgmt, recruiting
	ODIN AI	that makes you shine. (2025)	Meeting notes summarization & task automation
	LAWBOT AI	LawBot.ai (2025)	Client inquiries, scheduling, case info delivery

Continued on next page

Table 3 – *Continued from previous page*

Task Category	Product / Company	Website	Use Cases
	CASEFLOOD.AI	Caseflood.ai (2024)	Administrative tasks including phone intake and client follow-up.

C Extended Discussions of Legal LLM Agent Evaluation

C.1 Benchmarks Overview

Table 4 summarizes a diverse set of benchmarks for evaluating LLM-based agents on legal tasks. These benchmarks span procedural accuracy, legal reasoning, judgment quality, ethics, and fairness across simulated and real-world settings. Evaluation methods include rule-based metrics, statistical models, human annotation, and LLM-as-judge frameworks. Task formats range from single-turn classification to multi-agent simulations and adversarial court scenarios: CourtReasoner examines whether LLM agents can reason like judges through structured deliberation (Han et al., 2025); to evaluate agent-based court simulation systems, SimuCourt provides a benchmark comprising 420 real-world Chinese judgment documents across criminal, civil, and administrative cases in first- and second-instance settings (He et al., 2024); MASLegalBench provide strict, deductive testbeds to compare multi-agent designs and measure progress on formal legal reasoning tasks (Jing et al., 2025).

C.2 Fidelity of Evaluation Metrics

Many existing evaluation benchmarks rely heavily on human-authored reference data, which is costly and difficult to scale. To reduce annotation overhead, recent work has proposed *LLM-as-a-judge* approaches that automate evaluation by using LLMs to assess generated outputs. However, this paradigm raises concerns about *fidelity*, defined as the degree to which an evaluation metric aligns with how legal experts assess quality. Low-fidelity evaluation can undermine both the reliability and the practical effectiveness of benchmarking results.

Several studies have proposed methods to improve fidelity by better approximating expert legal judgment. For example, Enguehard et al. (2025) decompose long-form legal answers into fine-grained *legal data points*, enabling evaluation at a level that more closely reflects how lawyers assess correctness and omissions. Other work incorporates external legal context into the evaluation process: retrieval-based evaluation methods for legal QA, where model outputs are judged against retrieved supporting documents, improving correlation with lawyer assessments relative to purely model-based evaluators, and reducing reliance on surface-level similarity (Ryu et al., 2023). In addition, Raju et al. (2024) improve evaluation fidelity by constructing domain-specific benchmarks and data pipelines that better align model assessment with human preferences. Beyond that, Cheong et al. (2024) focus more on the practical side, and expert-centered studies show that appropriate LLM systems may prioritize helping users ask the right questions and avoid unauthorized practice of law, implying that efficacy must be assessed relative to realistic legal-use objectives rather than only answer correctness.

C.3 Challenges in Evaluation

We identify several key challenges in the current evaluation of legal LLMs and agent-based systems:

- *Metric-related limitations.* Existing evaluations largely lack coverage of law-specific, substantively important metrics, such as procedural compliance and bias or fairness assessment. Only a limited number of studies explicitly address these dimensions. For example, (Jia et al., 2025) evaluates procedural correctness, while (Xue et al., 2025b) examines judicial fairness and bias. Given the centrality of such properties to legal reasoning and decision-making, more systematic investigation of law-specific evaluation metrics is urgently needed.
- *Distributional bias across legal systems.* Current benchmarks and empirical studies exhibit limited diversity in legal systems and traditions. We observe that most existing datasets and evaluations are grounded in the Chinese legal system, with relatively few benchmarks derived from U.S., European, or other legal jurisdictions. Because legal rules, procedures, and normative standards vary substantially across jurisdictions, conclusions drawn from a single legal system may not generalize and may introduce systemic bias.
- *Evaluation fidelity in agent-based settings.* Many agent-based evaluations rely on LLM-driven role-play to simulate legal actors such as lawyers or judges, e.g., (Chen et al., 2025a; Jia et al., 2025). However, the realism, stability, and reliability of these simulated agents remain insufficiently validated. It is unclear to what extent their behavior reflects real-world legal practice, how sensitive they are to prompt variations, or how their limitations may confound evaluation outcomes.
- *Limited evaluation of multi-agent systems.* Although recent work has begun to explore multi-agent legal systems, e.g., (Jiang and Yang, 2025; Cui et al., 2023; Chen et al., 2025d), dedicated evaluation frameworks for multi-agent settings

are still lacking. In particular, reliability-oriented dimensions, such as workflow coherence, agent coordination, and inter-agent consistency, are rarely assessed in a principled manner.

- *Fairness in prompt tuning and agentic framework adaptation.* Many evaluations do not adequately account for model-specific prompt tuning or agentic framework adaptation, which can lead to suboptimal performance and undermine the validity of cross-model comparisons. While some studies explore zero-shot and few-shot prompting strategies for LLMs, e.g., (Fei et al., 2024; Li et al., 2024) or vary agent architectures to obtain a more comprehensive view (Li et al., 2025a), there is no widely accepted standard for ensuring fair and comparable evaluation across models. Given the sensitivity of LLM behavior to prompting and system design, standardized protocols are needed to support meaningful comparative conclusions.

Table 4: Legal Benchmarks Summary

Benchmark	Paper	Construct	Measurable	Metrics	Size	Simulation	Judge	Type
J1-EVAL / J1-ENVS	Jia et al. (2025)	Task performance and procedural compliance	Task success; procedural correctness	BIN; NBIN; PFS; JUD	6 environments, 3 levels, 508 scenarios	Multi-agent role-playing sandbox, multi-turn	LLM-as-Judge (GPT-4o) + Rule-based	Agent
LegalAgentBench	Li et al. (2025a)	Practical legal agent capability of tooling and interaction	Interaction success & Process rate	Success rate; BERT-Score	300 tasks	Single-turn, multi-step tool-augmented agent	Rule-based (Keyword/Program)	Agent
CourtReasoner	Han et al. (2025)	LLMs' legal reasoning quality based on real court opinion documents	Citation relevance, constraint extraction accuracy, argument validity evaluation	Human evaluation scores	292 expert-annotated meta-evalaution examples	Single-turn	Human annotators	Agent
SimuCourt	He et al. (2024)	Judicial decision-making	Legal article generation and judgment accuracy	Precision, recall and F1 scores	420 real-world judgment documents	Multi-agent simulated court debate embedded in decision-making process	LLM-as-Judge + human evaluation	Agent
MASLegalBench	Jing et al. (2025)	Deductive legal reasoning with multi-agent systems tailored to GDPR and rich real-world legal contexts	Performance on legal sub-tasks	Human evaluation scores	950 legal questions built from expert-authored court case contexts	Multi-agent deduction tasks with role specialization and collaborative reasoning configurations	Human evaluation	Agent
CourtBench	Chen et al. (2025a)	Interactive Reasoning Capability	Court performance over civil cases	Case winning rate; Prof score	1,000 civil cases	Multi-turn, Adversarial Evolution	LLM-as-judge	Agent
Trident-Bench	Hui et al. (2025)	Domain Safety & Ethics	Adherence to Professional Codes	Violation/Refusal Rate (%)	887 prompts	Single-turn	Rule-based	LLM
LAiW	Dai et al. (2025)	Legal Syllogism	Logic (FIR, LPI, ALP)	Accuracy, F1, Macro-F1	11,000 tasks	Single-turn	Rule-based	LLM
JudiFair	Xue et al. (2025b)	Judicial Fairness & Bias	Extra-legal factor influence	Inconsistency; Bias; Imbalanced Acc	65 extra-legal labels	Single-turn	Rule-based (Statistical)	LLM
LegalEval-Q	Li and Wu (2025)	Linguistic Quality	Clarity, coherence, terminology	AdjScore; 0–100%	946 annotated queries	Single-turn	Logistic Regression	LLM
LexEval	Li et al. (2025a)	Cognitive Ability	LexAbility Taxonomy	Accuracy, ROUGE-L	14,150 questions	Single-turn	Rule-based	LLM
LeCaRDv2	Li et al. (2023)	Legal Relevance	Characterization, Penalty, Procedure	Retrieval Recall	800 queries / 55K candidates	Single-turn	Rule-based	LLM
LawBench	Fei et al. (2024)	Legal Cognitive Ability	Memorization, Understanding, Applying	Accuracy, F1, ROUGE-L	20 tasks	Single-turn	Rule-based (Regex)	LLM
LegalBench	Guha et al. (2023)	Reasoning Breakdown	Six reasoning types (IRAC)	Accuracy; Balanced Acc	162 tasks	Single-turn	Rule-based	LLM