Article

Not peer-reviewed version

Trustworthy Legal Reasoning: A Comprehensive Survey

Sirui Han [†] , Zhizhuo Kou , Ruoxi Li , Yuyao Zhang , Yujin Zhou , Chuxue Cao , Han Zhu , Kunhao Pan , Haoran Li , Conghui He , Haitian Lu , Yike Guo [*]

Posted Date: 12 February 2026

doi: 10.20944/preprints202602.0870.v1

Keywords: LLMs; law applications

 doi:10.20944/preprints202602.0870.v1

Article

Trustworthy Legal Reasoning: A Comprehensive Survey

Sirui Han [1,2,†], Zhizhuo Kou [1], Ruoxi Li [1], Yuyao Zhang [1], Yujin Zhou [1], Chuxue Cao [1,3], Han Zhu [1,3], Kunhao Pan [2,4], Haoran Li [1], Conghui He [3], Haitian Lu [5] and Yike Guo[1,2,*]

[1] Hong Kong University of Science and Technology
[2] Hong Kong Generative AI R&D Center
[3] Shanghai Artifcial Intelligence Laboratory
[4] The LexiHK Project
[5] Hong Kong Polytechnic University
[*] Correspondence: yikeguo@ust.hk
[†] Project leader.

Abstract

As large language models are increasingly used for contract drafting, case research and even judicial work, a central question is how to make their outputs trustworthy. This survey addresses that question through the lens of verified generation for legal AI, focusing on systems that are robust against hallucinations and traceable to authoritative legal sources. First, we propose a unified framework for verified generation in legal AI, linking reasoning, retrieval, and validation around factual reliability. Second, we cast reliability methods into two paradigms of epistemic negotiation, *by failure* and *by conflict*, enabling models to recognize and act on their competence limits. Third, we survey the legal-AI landscape and identify challenges for verifiable, governance-native systems. This survey outlines a roadmap for trustworthy legal AI and for reliable reasoning beyond the legal domain.

Keywords: LLMs; law applications

1. Introduction

Verified generation for Legal AI aims to ensure that outputs from Large Language Models (LLMs) satisfy standards of legal reasoning, factual accuracy and professional accountability in high-stakes contexts.As we enter the era of LLM, advances in model scale, data quality and neural architectures are transforming natural language understanding Smith et al. (2023); Zhao et al. (2023). However, legal practice exposes specific requirements for *verifiability*, *transparency*, and *responsibility*, as emphasized in global regulatory frameworks such as the GDPR(2016), the EU AI Act (2024), and related international guidelines Han et al. (2025). These frameworks require outputs that are plausible, provably correct and traceable to authoritative sources, making law a testbed for methods for broader LLM use across domains.

In LegalAI, verifiability underpins well-defined scope, robust evidence grounding, and context-aware, accountable outputs.

Can generative systems deliver verifiable, up-to-date, and jurisdiction-aware answers with auditable reasoning steps?

Verified generation is a methodological shift, analogous to reinforcement learning with verifiable rewards on reasoning tasks X. Liu et al. (2025). Instead of maximizing preference alignment, it focuses on verifiable correctness via retrieval of legal sources and rule- or logic-based checks. In Legal AI, systems cross-check claims against statutes and case law before finalizing outputs, yielding legally sound text and generalizable methods for verifiable reasoning. General-purpose models such as GPT-4 Achiam et al. (2023) and Claude 3 Anthropic (2024) have demonstrated baseline reasoning abilities, while domain-specific systems, including ChatLaw Cui et al. (2023) and SaulLM Colombo et al. (2024a)

 doi:10.20944/preprints202602.0870.v1

extend these capabilities through targeted retrieval and formal-logic integration. New architectures pair large pretrained LLMs with verifiability modules that check output quality and citations, analogous to verifiable reward functions in recent Large Reasoning Models (LRMs) Valmeekam et al. (2024). Yet there is no standard way to measure factual reliability in law, and automated legal evaluation remains challenging, so legal knowledge is hard to leverage for the broader LLM community. Most existing work targets at general QA, leaving domain-specific legal issues largely unaddressed, such as jurisdictional limits, doctrinal variation, and strict formatting and citation rules Hepworth et al. (2024); South (2025); K. Sun et al. (2025); R. Zhang et al. (2025).

This survey reviews verified generation in Legal AI, covering formal reasoning, retrieval-based generation, verification architectures and evaluation, and outlines a roadmap for trustworthy systems for other safety-critical domains.

2. Constructions of LegalAI

2.1. Legal Reasoning

Faithfulness and Verifiability

Ensuring the faithfulness of generated rationales and adopting hypothesis-verification cycles have become paramount for developing interpretable and robust AI reasoning systems. Lyu et al. (2023) address the inherent unfaithfulness of vanilla Chain-of-Thought (CoT) by introducing a framework that translates natural language queries into executable symbolic reasoning chains. By delegating the final inference step to a deterministic solver, this approach enforces a *causal alignment* between the stated explanation and the model's decision logic, thereby grounding linguistic reasoning in formal semantics. Complementing this, Z. Yang et al. (2024) formalize inductive reasoning as an iterative process of generating latent hypotheses from evidence and systematically validating them. This structured decomposition not only facilitates auditable inference but also enables verifiable reasoning within high-dimensional natural language representations, bridging the gap between neural flexibility and symbolic rigour. Recent advances in decompositional prompting extend these ideas by partitioning complex problems into modular, verifiable sub-goals. Khot et al. (2023) propose decomposition strategies that embed symbolic functions or retrieval components into reasoning workflows, while Zhao et al. (2023) develop verify-and-edit mechanisms that iteratively validate intermediate steps. Together, these techniques show how structured decomposition and verification can enhance both correctness and transparency in reasoning.

Recent advances in decompositional prompting extend these ideas by partitioning complex problems into modular, verifiable sub-goals. Khot et al. (2023) propose decomposition strategies that embed symbolic functions or retrieval components into reasoning workflows, while Zhao et al. (2023) develop verify-and-edit mechanisms that iteratively validate intermediate steps. Together, these techniques show how structured decomposition and verification can enhance both correctness and transparency in model reasoning.

Verifiability and Neuro-Symbolic Integration

Within formal reasoning domains, recent research has transitioned toward a *generate-and-verify* paradigm to mitigate the stochastic nature of language models. Welleck et al. (2022a) and Poesia et al. (2023b) demonstrate that integrating external verifiers and logical constraints can effectively enforce stepwise validity and curtail *reasoning drift* in long-form proofs. These frameworks often leverage *autoformalization* to map natural language into symbolic representations, utilizing Satisfiability Modulo Theories (SMT) solvers and automated theorem provers to check the internal consistency of generated steps. By transforming probabilistic outputs into machine-checkable logical artifacts, these neuro-symbolic systems provide a robust mechanism for certifying the correctness of complex deductive chains (Bayless et al., 2025; J. P. Zhou et al., 2024). Shi et al. (2025) demonstrate that process-level verification with targeted remediation, using verifiers to inspect and refine intermediate steps, improves logical consistency in judgment prediction. By contrast, Magesh et al. (2025) report hallucination

 doi:10.20944/preprints202602.0870.v1

rates of 17–33% in leading legal RAG systems, driven largely by miscitations, and introduce an evaluation framework that separates correctness from groundedness across doctrine, jurisdiction, and factual recall. Persistent issues, including hallucination, data drift, and long-document degradation, underscore the need for domain-pretrained encoders, evidence-aligned retrieval systems, and human oversight to preserve jurisdictional accuracy, calibrated confidence, and auditable reasoning Guha, Nyarko, Ho, Ré, et al. (2023); Magesh et al. (2025); Zheng et al. (2021a). Accordingly, the emphasis shifts from linguistic plausibility towards legally grounded and verifiable correctness.

Conflict Calibration

In legal domain, models must communicate uncertainty to facilitate risk-aware decision-making. Beyond simple token probabilities, Lin et al. (2022a) demonstrate that verbalizing confidence directly improves transparency and human-AI collaboration. More recent techniques, such as semantic uncertainty, estimate confidence by measuring output invariance across semantically equivalent clusters, providing a more faithful characterization of model boundaries Kuhn et al. (2023). Theoretical analyzes further demonstrate that perfect calibration and zero hallucination cannot always be achieved simultaneously, motivating conservative behaviors in professional systems such as refusal thresholds, cautious defaults, and alert mechanisms Kalai and Vempala (2024). Complementary methods that output confidence sets or coverage-controlled predictions allow models to operate safely in high-stakes settings by combining candidate sets with calibrated refusal, ensuring predictable error Cherian et al. (2024).

Within in-context learning, recent work improves few-shot calibration by increasing sample diversity and decomposing sources of uncertainty, producing more stable alignment between confidence levels and empirical accuracy Ju et al. (2025); Ling et al. (2024). To counter systematic overconfidence introduced by reinforcement learning from human feedback (RLHF), prompt- and format-level interventions explicitly elicit and recalibrate confidence estimates, improving the correlation between stated confidence and correctness without altering model weights Tian et al. (2023). These strategies help users better assess the reliability of model output in high-stakes decision contexts.

Effective uncertainty quantification aligns probabilistic surrogates with domain-specific risk structures, enforcing stringent thresholds for high-stakes inference while permitting broader tolerances for low-risk tasks. By disentangling *epistemic* uncertainty (knowledge gaps) from *aleatoric* or contextual ambiguity (input noise), systems can adaptively trigger retrieval mechanisms, elicit clarification, or defer to human expertise. Recent advances in *conformal risk control* Oehri et al. (2025) provide distribution-free error bounds and calibrated refusal policies, offering a principled framework for selective prediction and rigorous risk management in sensitive domains such as LegalAI.

Constrained Generation

Constrained generation enforces the structural and logical consistency required in professional domains. Hard constraints, such as those in NeuroLogic Decoding X. Lu et al. (2021), and grammar-based decoding Geng et al. (2024); Shin et al. (2021) ensure rule satisfaction and format validity for structured outputs, including legal contracts and filings. Soft constraints complement this by preserving fluency through distribution-aware alignment Mudgal et al. (2024) and Pareto-optimized decoding Guttmann et al. (2025), balancing factuality, coherence, and domain conformity. Task-oriented and adaptive approaches, such as GeLaTo C. Li et al. (2024) and CAAD decoding M. Nguyen et al. (2025) integrate schema, knowledge, and contextual signals to enforce jurisdiction- and role-specific compliance. Invariant-based constraints further preserve numerical and logical consistency under paraphrase and other linguistic variation. Together, hard (logical and grammatical), soft (distributional), and adaptive (task-based) constraints form a coherent and auditable generation framework that supports reliability, interpretability, and regulatory compliance in professional LegalAI systems.

 doi:10.20944/preprints202602.0870.v1

2.2. Foundation Model

Foundation models underpin LegalAI by distilling large-scale legal pretraining into transferable knowledge, enabling high-performance adaptation for specialized downstream reasoning tasks.

Supervised Finetuning Supervised fine-tuning (SFT) adapts pretrained language models to legal domains using labeled datasets that encode domain-specific structures and reasoning conventions. This aligns models with the legal field's requirements for logical coherence, terminological precision, and interpretable justification, all of which are capabilities that general-purpose models typically lack Colombo et al. (2024b); J. Shi et al. (2024). Legal-LM J. Shi et al. (2024) illustrates how integrating legal knowledge graphs and external resources via soft prompting and Direct Preference Optimization (DPO) embeds structured reasoning patterns into model parameters, providing both semantic guidance and preference correction. Combining domain-adaptive pretraining with task-oriented fine-tuning on statutes, case law, and contracts further improves factuality, calibration, and robustness Chalkidis, Garneau, Goanta, Katz, and Søgaard (2023).

High-quality labeled corpora are essential for SFT because legal texts exhibit distinctive linguistic and structural properties, including dense citations, long-form documents, and formalized argumentative style Chalkidis, Garneau, Goanta, Katz, and Søgaard (2023). Benchmarks such as LexGLUE, MultiEURLEX, BillSum, LEDGAR, and COLIEE Carlini et al. (2021); Chalkidis et al. (2020a); Kornilova and Eidelman (2019); Rabelo et al. (2020); Tuggener et al. (2020) provide multi-task supervision for classification, retrieval, summarization, and case-based reasoning. Methodologically, legal SFT often employs text-to-text frameworks Raffel et al. (2020) with structured decoding, ranking losses, and schema constraints to enhance auditability and citation precision Rabelo et al. (2020). Parameter-efficient approaches such as LoRA Tran et al. (2024) enable scalable adaptation under resource constraints, while safety-aware fine-tuning (SaRFT) ZHAO et al. (2025) promotes role alignment and guarded behavior in professional legal interactions.

Contemporary practice combines supervised fine-tuning with domain-adaptive pretraining, retrieval grounding, and rationale supervision Chalkidis et al. (2020a); Dettmers et al. (2023a); Guha, Nyarko, Ho, Ré, et al. (2023); Gururangan et al. (2020). Legal knowledge graphs and safety constraints further improve factual robustness, ethical behavior, and role alignment Colombo et al. (2024b); Gao et al. (2025); J. Shi et al. (2024). Domain-specific instruction tuning enhances controllability through structured prompts and synthetic task construction Sanh et al. (2022); Y. Wang et al. (2023); Wei et al. (2022). Parameter-efficient approaches such as LoRA, QLoRA, and adapter layers Dettmers et al. (2023b); Hu et al. (2021); Pfeiffer et al. (2021) support modular and privacy-preserving updates, while long-context architectures like Longformer and BigBird Beltagy et al. (2020); Zaheer et al. (2020) enable efficient processing of statutes, contracts, and case reports. Hierarchical SFT and retrieval-augmented training further reduce hallucination and strengthen explainability by grounding intermediate rationales in evidence DeYoung et al. (2020); Lewis et al. (2021).

Evaluation relies on domain benchmarks such as LexGLUE and LegalBench Chalkidis et al. (2022); Guha, Nyarko, Ho, Ré, et al. (2023). COLIEE assesses retrieval and entailment tasks, while MultiEURLEX probes cross-jurisdictional generalization Chalkidis et al. (2021). Responsible governance additionally requires dataset deduplication, anonymization, red-teaming, and citation verification to mitigate privacy and memorization risks Carlini et al. (2021); Chalkidis et al. (2022).AIRA employs activation-informed singular value decomposition (SVD) initialization, dynamic rank allocation, and activation-aware training to efficiently adapt large models, achieving 53.66% accuracy on legal benchmarks while significantly reducing training time L. Li, Li, et al. (2025). NoRA introduces a nested low-rank structure where an outer frozen LoRA layer preserves pre-trained knowledge, while an inner trainable LoRA layer captures task adaptations, demonstrating strong performance on legal tasks such as legal text analysis & generation L. Li, Lin, et al. (2025).

Taken together, an effective LegalAI pipeline unifies domain-adaptive pretraining, high-quality multi-task SFT, structured instruction tuning, parameter-efficient adaptation, long-context modeling, and retrieval-guided rationale supervision. This integrated approach improves factual accuracy,

 doi:10.20944/preprints202602.0870.v1

citation precision, interpretability, and safety compliance across legal applications Beltagy et al. (2020); Chalkidis et al. (2020c); Dettmers et al. (2023a); Guha, Nyarko, Ho, Ré, et al. (2023); Lewis et al. (2021).

Reinforced Learning Reinforcement learning (RL) is moving LegalAI beyond surface-level text generation toward procedurally grounded reasoning by optimizing models with external signals of consistency, completeness, and legal compliance. These signals strengthen procedural alignment, adaptive reasoning, and logical structure, shifting LegalAI from retrieval-heavy pattern matching toward systems capable of structured deliberation. At scale, models such as SaulLM-54B and SaulLM-141B Colombo et al. (2024b) combine continual domain pretraining with multi-level supervised fine-tuning to balance semantic grounding, legal robustness, and general reasoning capacity.

To mitigate over-specialization, C. Liu et al. (2024) introduce General Capability Integration (GCI) and ALoRA, balancing domain-specific and general inference via dynamic attention routing. Research on alignment indicates that response diversity, rather than data volume, is critical for stable legal performance (Song et al., 2024), while Gao et al. (2025) demonstrate that short-instruction SFT under mixed supervision generalizes effectively to 128K-token sequences. While early efforts like LexGPT 0.1 (Lee, 2023) established legal instruction tuning and RLHF infrastructure, subsequent models like LexPam and LexNum integrate RL with curriculum learning to generate program-aligned, verifiable reasoning paths (K. Zhang et al., 2025). Most recently, Unilaw-R1 (Cai et al., 2025) combines structured-knowledge SFT with reinforcement learning to reach superior logical coherence and cross-domain generalization.

Efficiency and inference control form a complementary axis of progress. ASRR X. Zhang et al. (2025) modulates the trade-offs between accuracy and cost by adjusting the reasoning depth on demand, while Wu et al.Wu et al. (2024) develop D3LM, an RL-based diagnostic model that integrates legal knowledge graphs to guide questioning and improve consultation accuracy. Similarly, Yao et al.Yao et al. (2025) introduce LSIM, aligning fact–rule chains via RL to enhance interpretability and reduce hallucinations. Collectively, these developments mark a shift from static knowledge adherence to dynamic, procedurally compliant reasoning Cai et al. (2025); Wu et al. (2024); Yao et al. (2025); K. Zhang et al. (2025), establishing a multidimensional optimization paradigm that balances accuracy, efficiency, and regulatory compliance for interpretable and legally reliable AI systems.

2.3. Retrieval and Agentic Methods

Retrieval-augmented generation Advances in verifiable generation and retrieval-augmented generation (RAG) have substantially improved factual grounding by coupling neural generation with evidence retrieval from large corpora. This design mirrors the legal system's "citation–argumentation" paradigm and enables auditing of evidence chains Lewis et al. (2020a); Zheng et al. (2021a). This field has progressed from sparse retrieval heuristics to differentiable retrievers and seq2seq marginalization over latent documents, supporting per-token evidence switching, hot-swappable indices, and improved factuality and diversity Lin et al. (2022b); Poesia et al. (2023a). Modern verifiable-generation approaches further embed explicit checking stages, such as chain-of-checks and agentic debate, and introduce benchmarks that enforce evidence alignment and subsentence-level attribution C. Cao, Li, et al. (2025); C. Cao, Zhu, et al. (2025).

RAG integrates LLMs with external knowledge bases, grounding model outputs in verifiable evidence and mitigating knowledge staleness and traceability challenges that affect purely parametric models Lewis et al. (2020b). In professional services, these systems retrieve statutes, regulations, or procedural documents to produce accurate, auditable, and compliant responses.

Within LegalAI, RAG enhances factual accuracy, reasoning quality, and regulatory alignment. In judicial document generation, Su et al. (2025) introduce the JuDGE benchmark for Chinese case opinions, showing that RAG models surpass pure generators yet still struggle with logical coherence and structural fidelity. Xie et al. (2024) develop DeliLaw, an advisory system that retrieves statutes and precedents to reduce hallucination and improve multilingual consultation. On the retrieval side, Zhang et al. (2023) propose CFGL-LCR, enriching causal reasoning through graph-structured retrieval and counterfactual augmentation, while Ye et al. (2024) present MileCut, which uses multi-view modeling

 doi:10.20944/preprints202602.0870.v1

and information compression to capture semantic hierarchy more effectively. Su et al. (2025) also introduce Caseformer, an unsupervised dense pretraining framework for cross-case retrieval, and Askari et al. (2023) design CLosER, a conversational RAG system with expertise-aware ranking for precise legal responses. In common-law contexts, Nigam et al. (2025) propose NyayaRAG, combining factual narratives with semantically retrieved sources to emulate judicial reasoning. Collectively, legal RAG research now forms a cohesive pipeline that unifies retrieval-side knowledge modeling with generation-side reasoning alignment. This evolving framework moves LegalAI from information-augmented generation toward knowledge-aligned, verifiable reasoning Askari et al. (2023); Nigam et al. (2025); Su et al. (2025); Xie et al. (2024); Ye and Li (2024); K. Zhang et al. (2023).

However, precision in retrieval-augmented systems remains constrained by retrieval fidelity: missing or partially relevant passages propagate downstream errors, qualifiers disappear under aggressive compression, and performance deteriorates on tail facts or domain-specific corpora Hong Kong Productivity Council (HKPC) (2024); The Law Society of Hong Kong (LSHK) (2024). Fine-grained attribution metrics and benchmarks further reveal brittle subsentence attribution, noisy entailment on long documents, and limited robustness under distribution shift H. Li, Chen, et al. (2025a); H. Li, Hu, et al. (2025a). Achieving professional-grade verifiability therefore requires domain-calibrated retrieval, authority-aware entailment with explicit scope controls, and fully auditable pipelines that log micropropositions and supporting evidence end-to-end X.-W. Yang et al. (2025).

Tool-Augmented Generation Hybrid systems achieve a balance between accuracy and fluency by delegating fact selection to rule-based logic while utilizing language models for natural language synthesis (Fang et al., 2023). To further enforce correctness, logic-constrained decoding integrates formal invariants and theorem provers directly into the generation process. This approach significantly mitigates hallucinations and bolsters trustworthiness without compromising computational efficiency (Alpay & Alakkad, 2025).

LegalAI is thus evolving from single-model generation to multi-agent, tool-integrated systems that enhance transparency and interpretability in reasoning. Progress increasingly depends not on model size but on the integration of external knowledge, formal logic, and domain-specific tools. Wang et al. (2024) exemplify this shift with LegalReasoner, a multi-stage framework that orchestrates contrastive learning, graph neural networks (GNNs), and generative adversarial networks (GANs) across legal knowledge injection, precedent retrieval, multi-hop reasoning, and judgment generation. This design yields notable gains in judgment-prediction accuracy. Xu et al. (2025) further advance tool-embedded reasoning with the CLEAR framework, which employs a rule retriever and rule-insight generator to enforce statutory constraints during inference, significantly improving the interpretation of ambiguous provisions.

Continuing this trajectory, Petros et al. (2025) propose PAKTON, a multi-agent tool-augmented system that integrates RAG with specialized agents coordinating fact extraction, clause interpretation, and compliance assessment for contract review. Bendová et al. (2025) develop a hybrid extraction method combining regular expressions with LLMs to identify judicially recognized facts in Slovak criminal judgments. Their system produces structured factual annotations that strengthen downstream RAG pipelines, underscoring the central role of structured knowledge extraction in legal reasoning workflows.

Agentic-based Generation LegalAI research is shifting from isolated task models to adaptive multi-agent systems that mirror the deliberative structure of real-world legal reasoning. Chen et al. (2025) introduce the Debate-Feedback framework, in which multiple LLM agents engage in iterative debate and critique to improve judgment prediction while reducing reliance on annotated data. Similarly, Yuan et al. (2024) propose the MALR framework, showing that inter-agent collaboration deepens understanding of legal theories and accusatory structures, translating abstract jurisprudence into operational reasoning pathways.

Nguyen et al. (2025) present a unified framework integrating rule-based, abductive, and case-based reasoning, providing a logical foundation for coordinated multi-agent systems. Yue et al. (2025)

 doi:10.20944/preprints202602.0870.v1

introduce MASER, a multi-agent simulation environment that generates law-intensive interactive data to mitigate scarce supervision in dynamic legal dialogues. Chen G. et al. (2025) develop AgentCourt, which models courtroom adversarial processes through evolutionary agent interactions, enabling lawyer agents to acquire professional expertise. Xu et al. (2024) design LeGen, enabling controllable and consistent legal text generation via modular decomposition and concept-level verification. Klisura et al. (2025) extend multi-agent collaboration to multilingual and dialectally diverse settings through a cooperative privacy-policy QA system that improves reasoning accuracy and inclusivity. Together, these works advance coordinated, tool-integrated LegalAI ecosystems in which specialized agents perform retrieval, argumentation, and validation, fostering self-calibrating and interpretable legal reasoning.

3. Evaluation of LegalAI

3.1. Merits of Legal Reasoning

LegalAI benchmarking is shifting from narrow task-specific evaluations towards theoretically grounded, multidimensional frameworks that integrate knowledge, logic, and semantics. Early benchmarks such as CaseHOLD quantified legal reasoning through citation prediction, demonstrating the benefits of domain-specific pretraining for legal language and logic Zheng et al. (2021b). LeCaRDv2 extended this paradigm to Chinese law, applying expert-defined relevance standards that reflect judicial reasoning H. Li, Shao, et al. (2024). Cross-lingual evaluations such as LeXFiles and LegalLAMA examined the transferability of legal knowledge across jurisdictions Chalkidis, Garneau, Goanta, Katz, and Søgaard (2023), while the Cambridge Law Corpus introduced ethical and privacy-aware methodologies for large-scale historical legal data research Östling et al. (2023).

At the empirical level, LegalBench formalized 162 reasoning-oriented tasks, encompassing statutory interpretation, analogical reasoning, and inference, thereby enabling a systematic decomposition of legal cognition Guha, Nyarko, Ho, Ré, et al. (2023). Within the Chinese context, CLAW and LAiW implemented clause-level and layered evaluation frameworks that assess retrieval fidelity, reasoning consistency, and interpretability Y. Dai et al. (2025); X. Xu et al. (2025). Complementing these benchmarks, δ-Stance modeled judicial argumentation using polarity–intensity variables to capture fine-grained stance dynamics Gupta et al. (2025).

Evaluation has also expanded cross-jurisdictionally. Bilingual frameworks now measure alignment between translation fidelity and legal logic kui Sin et al. (2025). And enhanced Chinese models integrate statutes, precedents, and legal graphs to support end-to-end consulting Z. Zhou et al. (2024). Multi-agent architectures support modular evaluation by assigning retrieval, reasoning, and verification to distinct agents, forming a "checks-and-balances" configuration analogous to institutional governance structures C. Cao, Zhu, et al. (2025); J. Sun et al. (2024).

Governance and transparency have become central to benchmark design. Model cards codify disclosure standards regarding model scope and limitations Mitchell et al. (2019), while auditing frameworks emphasize traceability, reproducibility, and ethical accountability Birhane et al. (2024a). Regulatory frameworks, particularly in finance, further mandate interpretable, trustworthy, and auditable AI systems J. Luo et al. (2025); Zöller et al. (2025).

This evolution shifts assessment from task performance to *reasoning coherence*, verifying whether models reconstruct expert-justifiable logic (Y. Dai et al., 2025; X. Xu et al., 2025). Emerging certification standards now codify requirements for robustness, fairness, and post-deployment governance (Schweighofer et al., 2025), reinforced by compliance emphasizing adversarial resilience and log integrity (Schöning & Kruse, 2025).

Research on adversarial safety integrates gradient-based attack modeling with defense mechanisms to mitigate vulnerabilities such as prompt injection, enabling safer LegalAI deployment in risk-sensitive settings C. Guo et al. (2021). Collectively, these advances articulate a trust architecture that grounds LegalAI in epistemic rigor, modular reasoning, and verifiable governance, supporting a

 doi:10.20944/preprints202602.0870.v1

shift from mere functionality to reliability, interpretability, and institutional trust Birhane et al. (2024a); Guha, Nyarko, Ho, Ré, et al. (2023); kui Sin et al. (2025); Schweighofer et al. (2025); Zöller et al. (2025).

The Chain-of-Verification (CoVe) method mitigates factual errors by iteratively drafting, interrogating, and revising outputs against retrieved evidence (Dhuliawala et al., 2024). Similarly, "retrieve-rerank-verify" pipelines integrate external knowledge for multi-hop reasoning, while faithful Chain-of-Thought (CoT) validation audits intermediate steps to bolster factual fidelity and mechanistic transparency (Lyu et al., 2023). In legal applications, unifying detection metrics, evidence alignment, and controlled decoding into a single pipeline can preempt high-risk hallucinations while ensuring compliance and auditability, substantially improving the reliability and accountability of LegalAI systems Agrawal et al. (2024a).

LegalAI is shifting from raw text generation toward process-based evaluation centered on reliability, supported by knowledge alignment, open benchmarking, and explainable assessment. Early scrutiny focused on factual inaccuracies: Magesh et al. (2025) found 17–33% hallucination rates in commercial LegalAI tools, even under RAG setups, challenging "hallucination-free" claims and motivating standardized evaluation infrastructures. LegalBench Guha, Nyarko, Ho, Ré, et al. (2023) offers 162 fine-grained tasks to evaluate legal reasoning across models and jurisdictions, while InternLM-Law Fei et al. (2025) extends large-scale benchmarking to Chinese and multilingual settings. Luo et al. (2025) integrate evaluation into the generation process with ATRIE, which measures variance in legal-concept entailment and signals a broader move toward task-embedded evaluation.

At the same time, definitions of reliability are expanding. Alsagheer et al. (2025) identify trade-offs between model scale, logical consistency, and fairness, highlighting the need to evaluate ethical robustness alongside predictive accuracy. Eval-RAG Ryu et al. (2023) verifies generated outputs using external retrieval to enhance factual grounding, while DeCE Yu et al. (2025) decomposes single-score metrics into precision and coverage dimensions to improve semantic resolution.

3.2. Knowledge Alignment

Evaluation paradigms increasingly internalize retrieval to determine the timing and scope of evidence acquisition. *Self-RAG* enables models to autonomously trigger retrieval, evaluate evidence quality, and critique outputs, significantly enhancing factual accuracy and citation reliability (Asai et al., 2023). Complementing this, *Active RAG* leverages uncertainty signals to proactively retrieve information for multi-hop inference segments (Jiang et al., 2023). Finally, the *Verify-and-Edit* framework audits reasoning chains at the step level, iteratively validating and refining intermediate logic with external knowledge to balance interpretability and robustness (Zhao et al., 2023).

In professional domains, retrieval-augmented training and knowledge organization are advancing together. RAFT (Retrieval-Augmented Fine-Tuning) annotates "gold" and "distractor" documents during fine-tuning, training interference-resistant evidence selection and improving domain adaptation in fields such as law and healthcare T. Zhang et al. (2024). Adaptive RAG balances dialogue history and new retrievals to maintain multi-turn consistency and reduce context drift X. Wang et al. (2025).

For highly structured legal settings, knowledge-graph–enhanced RAG unifies entity retrieval and relation-based inference along graph paths, capturing complex "provision–exception–condition" dependencies and improving both the completeness and auditability of evidence recall S. Wang et al. (2025). Collectively, advances in retrieval strategy learning, evidence quality control, verifiable reasoning, and structured knowledge integration make alignment in professional AI systems more discoverable, accurate, and explainable S. Wang et al. (2025); T. Zhang et al. (2024).

3.3. Hallucination Control

Achieving hallucination-free legal reasoning requires moving from post-generation verification to controlled generation. Hallucinations are categorized as intrinsic vs. extrinsic, guiding interventions at the data, model, and decoding levels Z. Ji et al. (2023); and along factuality vs. faithfulness, structuring detection and mitigation strategies L. Huang et al. (2025). Detection methods now go beyond surface-level checks to task- and decision-level analysis, where internal consistency can be

 doi:10.20944/preprints202602.0870.v1

probed via natural language inference for contradiction-aware self-correction Mündler et al. (2024) Low-quality or fabricated citations are often associated with low token-level confidence, enabling uncertainty-based detectors Agrawal et al. (2024a). Domain-specific evaluation has emerged for high-stakes areas. In law, even "hallucination-free" models may still commit citation errors, motivating workflow-oriented legal evaluations Magesh et al. (2024); in medicine, clinical safety frameworks quantify factual, omission, and reasoning errors Asgari et al. (2025). For embodied or sequential decision-making, hallucinations are classified as perceptual (environment misreadings) or planning (infeasible action sequences), with world models used to verify whether perceptions and planned actions are physically and procedurally feasible Chakraborty et al. (2025).

4. Challenge and Future

Safety and Privacy The rapid expansion of LegalAI has intensified concerns over safety, privacy, and governance, especially as larger models increase risks of hallucination, misgrounding, and sensitive data disclosure in regulated settings Agrawal et al. (2024b); Magesh et al. (2025). Legal practice requires confidentiality and auditability, yet current fine-tuning and reinforcement pipelines remain vulnerable to data leakage and overconfidence J. Ji, Hong, et al. (2025). Governance-native frameworks tackle these issues by embedding compliance constraints directly into system architectures and integrating verifiable reasoning, calibrated refusal, and privacy-preserving methods such as federated and differential learning to keep models within verifiable, authoritative bounds Fernsel et al. (2024); H. Li, Chen, et al. (2025b); H. Li, Hu, et al. (2025b). Process-level safeguards and traceable inference pipelines further support professional accountability, while embedding jurisdictional limits, confidentiality, and privilege constraints into retrieval and generation reduces compliance risk J. Ji, Chen, et al. (2025); Raji et al. (2020). The convergence of verifiable reasoning and privacy-aware computation thus defines the frontier of trustworthy Legal AI, making safety a foundational architectural principle rather than a post-hoc safeguard H. Li, Hu, et al. (2025b).

Personalization and Self-Evolution The next frontier for LegalAI is personalization—adapting reasoning and retrieval to specific practitioners and jurisdictions while ensuring compliance (X. Chen et al., 2025; H. T. Nguyen et al., 2025). While adaptive multi-agent and reinforcement learning frameworks, such as the *AdvEvol* approach in *AgentCourt*, significantly improve expert-level reasoning, they also introduce challenges for liability and auditability (G. Chen et al., 2025). To mitigate these risks, hybrid workflows use supervised checkpoints to ensure model updates remain reversible and verifiable (Yuan et al., 2024). Furthermore, self-evolving systems utilize incremental fine-tuning and rule-based memory anchoring to integrate new precedents without catastrophic forgetting (C. Liu et al., 2024). Addressing potential biases and jurisdictional drift in autonomous updates requires transparent safeguards, including auditable logs and model certification, to maintain alignment with professional standards (Birhane et al., 2024b).

5. Conclusion

Legal foundation models can transform research, drafting, and compliance when anchored by verifiable reasoning and controlled retrieval. Reliable systems must integrate citation-anchored outputs with calibrated uncertainty to enable risk-aware abstention. However, significant challenges persist in cross-jurisdictional adaptation, analogical reasoning, and multilingual benchmark coverage. Advancing trustworthy LegalAI requires open, rights-cleared datasets and modular *retrieval–reasoning–verification* pipelines governed by rigorous oversight. Ultimately, deep collaboration between technologists and legal experts is essential to ensure that every inference remains auditable and aligned with verifiable ethical and legal standards.

6. Limitations

This survey offers a broad, text-focused synthesis and may give less attention to multimodal evidence and some lower-resource jurisdictions and languages. The review primarily reflects publicly

 doi:10.20944/preprints202602.0870.v1

available studies and benchmarks, which are evolving and may not capture practitioner considerations such as authority hierarchy, conflict resolution, and time- or jurisdiction-specific validity. Most findings are based on current snapshots rather than long-term observations, and operational topics such as liability allocation, privilege, and e-discovery costs are only briefly touched. Discussions of ethics, fairness, security, and agentic systems emphasize concepts and emerging practices, with limited large-scale field evidence to date. Considerations around data provenance, rights, and cross-border transfers are framed as areas for continued clarification. Some recommendations may be most readily implemented by well-resourced institutions, suggesting opportunities to adapt approaches for smaller organizations. Finally, choices about scope and citation inevitably reflect selectivity, indicating avenues for complementary work to extend coverage.

Appendix A. Related Work: Factuality Verified and Scope-Limited Generation

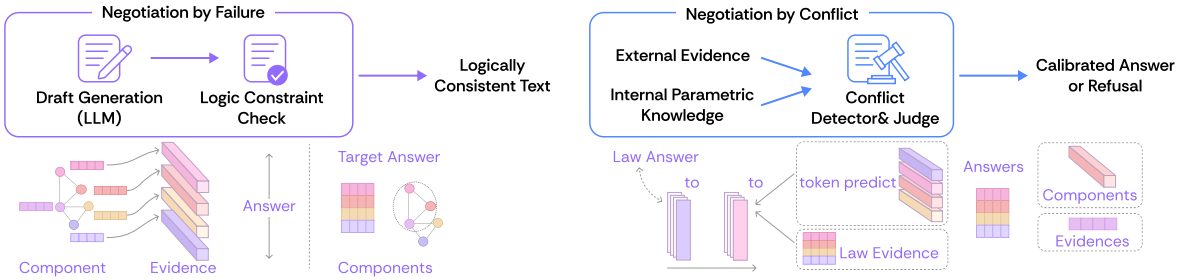

Figure A1. Comparison between two paradigms of verified generation: *Negotiation by Failure* and *Negotiation by Conflict*. The former ensures logical consistency through constraint checking, while the latter calibrates answers via conflict detection and adjudication.

From the perspective of computational logic, verified generation is not merely a technological device for reducing errors but an epistemic discipline aimed at developing models that can recognize the limits of their own competence Kaleeswaran et al. (2016). In this view, a verifiable system must implement an internal decision logic that differentiates between two fundamental forms of epistemic negotiation: *negotiation by failure*, in which the model identifies a contradiction and acknowledges that its inference is incorrect, and *negotiation by conflict*, in which the model detects insufficient support and withholds judgment Ni et al. (2023). The epistemological force of this negotiation lies in articulating a machine's capacity for self-reflective reasoning, establishing a structured relationship between what the system is capable of inferring and what the available evidence permits, as illustrates in Figure A1.

Contemporary LLMs generate text through autoregressive next-token prediction over learned probability distributions Vaswani et al. (2017). This computational paradigm presents three fundamental challenges. First, models must develop capability boundary awareness: the ability to distinguish interpolation within their training distribution from extrapolation beyond it Q. Guo et al. (2025). Because embedding spaces are continuous, models can produce fluent yet unfounded responses when queried outside their epistemic range. Second, standard maximum-likelihood training conflates statistical prevalence with factual correctness, motivating training-distribution alignment so that factual precision, rather than frequency, governs generation J. Ji et al. (2023); H. Lu et al. (2025). Third, the fixed computational depth of feedforward architectures limits a model's capacity for iterative deliberation, making reflective reasoning dependent on auxiliary memory structures, external tools, and multi-pass generation strategies Nye et al. (2021); Schick et al. (2023). While traditional approaches impose correctness through *post-hoc* validation, verified generation advances toward meta-reasoning, wherein models explicitly evaluate the validity of their own inferences S. Cao and Wang (2024).

Techniques such as Chain-of-Verification (CoVe) implement this idea by having models propose provisional hypotheses, issue internal verification queries that operate as logical checks over their representations, and revise outputs when inconsistencies are detected Dhuliawala et al. (2024). Architecturally, this inserts a verification loop into the generative process: whereas conventional mod-

 doi:10.20944/preprints202602.0870.v1

els compute $p(x_t \mid x_{<t})$, verified generation realizes $p(x_t \mid x_{<t}, \text{Verify}(x_{<t}))$, where the verification function acts as a learned consistency constraint X. Li et al. (2024).

The LLM-as-a-judge paradigm operationalizes this principle by introducing secondary evaluators that function as proof validators T. Wang et al. (2023), conceptually paralleling logic-programming systems in which inference and verification act as dual computational processes. Verified generation therefore requires a shift from single-pass token prediction to architectures capable of global consistency reasoning, recasting generation not as a one-shot feedforward computation but as an iterative proof search over factually grounded hypotheses. At the implementation level, pipelines such as generate–verify–edit ritun16 (2024) instantiate these principles by embedding verification directly into model workflows, treating each generated statement as a propositional claim subject to falsification and revision. This design constitutes an applied form of computational epistemology that rather than training models to approximate omniscience, verified generation constrains reasoning within the boundaries of what can be formally demonstrated and empirically validated, thereby establishing a principled framework for epistemic accountability in language generation.

Appendix B. Formulation of Verified Generation

Complementing the Syllogistic definition, we formalize the probabilistic bounds required for *Negotiation by Conflict* and *Uncertainty Calibration*.

Standard autoregressive generation maximizes the likelihood $P(Y|X)$ for a prompt X. In a verified LegalAI framework, we introduce a verification function $V(\cdot)$ and an external knowledge base $\mathcal{K}$ (e.g., statutes, case law). The generation objective is modified to maximize the joint probability of the token sequence y_t and its logical validity:

$$y_t^* = \underset{y_t}{\text{argmax}} \Big[\log P_\theta(y_t|y_{<t}, X) + \lambda \cdot \log P_{\text{ver}}(V(y_t, \mathcal{K})|y_{<t}) \Big] \tag{A1}$$

Where:

- P_θ is the base LLM probability distribution.
- P_{ver} is the probability assigned by the verification module.
- λ is a hyperparameter weighing factual adherence against linguistic fluency.

To satisfy the requirement that models must "know when not to answer", generally define a Conformal Risk Control set $\mathcal{C}(X)$ for a legal query X. For a user-specified error rate $\alpha \in [0, 1]$ (e.g., $\alpha = 0.05$ for high-stakes advice), we require:

$$P(Y_{\text{true}} \in \mathcal{C}(X)) \geq 1 - \alpha$$

The construction of $\mathcal{C}(X)$ relies on a calibration score $s(X, Y)$ (e.g., the refusal threshold). If the model's confidence $c(Y) < \tau$, where τ is the calibrated threshold derived from a hold-out set of legal experts, the system outputs the empty set $\emptyset$ (abstention) rather than a hallucination.

The loss function for optimizing the retriever R and generator G in this context minimizes the Kullback-Leibler divergence Kullback (1951) between the generated rationale and the authoritative legal reasoning path π:

$$\mathcal{L}_{\text{verify}} = -\mathbb{E}_{q \sim \mathcal{D}} \left[\sum_{t=1}^{T} \log P_G(y_t|y_{<t}, R(q), \pi_{\text{logic}}) \right]$$

where π_{logic} represents the logical invariants derived from formal theorem provers to ensure no contradiction exists in the generated chain.

 doi:10.20944/preprints202602.0870.v1

Appendix C. Detailed Discussions on the Construction of LegalAI

A comprehensive list of the existing domain LLMs for law is provided in Table A1.

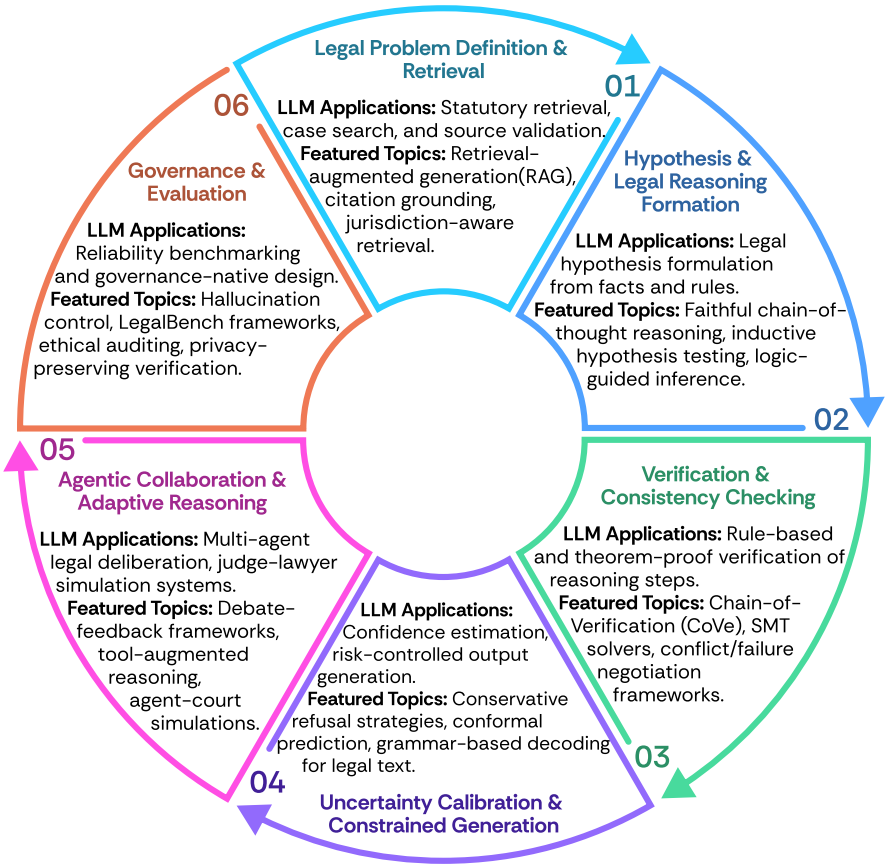

Figure A2. Six Stages of the Trustworthy Legal AI

Table A1. Statistics comparisons among Legal LLMs (R&D: Research and Development; PT: Pre-Train; SFT: Supervised Fine-Tune; RL: Reinforcement Learning).

Model / Framework	Country / Region	Parameters	Training Resources	PT	SFT	RL	Date
Legal-BERT (Chalkidis et al., 2020b)	Greece, United Kingdom	LEGAL-BERT-BASE(110M)	US contracts, EU legislation, ECHR cases	✓	-	-	2020
German BERT Base	Germany	bert-base-cased	Wiki, OpenLegalData, News	✓	✓	-	2020
CaseLaw-BERT	United States	bert-base-uncased(110M)	ILSI, ISS, ILDC Dataset	✓	✓	-	2021
CodeGen (Nijkamp et al., 2023)	United States	CodeGen1/CodeGen2/CodeGen2.5(350M,1B,3B,7B,16B)	Jaxformer	✓	-	-	2022
ChatLaw (Cui et al., 2024)	China	InternLM(4x7B, MoE)	93,000 court case decisions	✓	-	-	2023
DISC-LawLLM (S. Yue et al., 2023)	China	Qwen2.5-instruct 7B	DISC-Law-SFT	✓	✓	-	2023
HanFei	China	HanFei-1.0(7b)	Cases, regulations, indictments, legal news	✓	✓	-	2023
JurisLMs	China	LLaMA, GPT2	Chinese legal corpus	✓	✓	-	2023
LaWGPT (Z. Zhou et al., 2024)	China	Chinese-alpaca-plus-7B	Awesome Chinese Legal Resources	✓	✓	-	2023
LexiLaw	China	ChatGLM-6B	BELLE 1.5M	-	✓	-	2023
WisdomInterrogatory	China	Baichuan-7B	40G legal-related data	✓	✓	-	2023
Lawyer LLaMA (Q. Huang et al., 2023)	China	quzhe/llama_chinese_13B, Chinese-LLaMA-13B	Alpaca-GPT4 (52k Chinese + 52k English)	✓	✓	-	2023
LAWGPT-zh	China	ChatGLM-6B	CrimeKgAssitant	-	✓	-	2023
BaoLuo LawAssistant	China	ChatGLM, sftglm-6b	Legal dataset	-	✓	-	2023
FedJudge (L. Yue et al., 2024)	China	baichuan-7b	C3VG, Lawyer LLaMA	-	✓	-	2023
Law-GLM	Germany, China	GLM-10B	30GB of Chinese legal data	-	✓	-	2023
LexLM (Chalkidis, Garneau, Goanta, Katz, & Søgaard, 2023)	Denmark	RoBERTa large	LeXFiles corpus	✓	-	-	2023
legal-xlm-roberta-large (Niklaus et al., 2024)	Switzerland, Denmark, USA	XLM-R large	Multi Legal Pile	✓	-	-	2023
bert-large-portuguese-cased-legal	Brazil	BERTimbau large	assin, assin2, stsb_multi_mt pt, IRIS STS datasets	✓	✓	-	2023
SaulLM-7B (Colombo et al., 2024a)	United States, France, Portugal	Mistral-7B	Specialized data in the legal field	✓	-	-	2024
LawLLM (Shu et al., 2024)	United States	Gemma-7B	American legal data from the case.law platform	-	✓	-	2024
LegalΔ (X. Dai et al., 2025)	China	Qwen2.5-14B-Instruct	Lawbench, Lexeval, Disclaw	-	✓	✓	2025

Appendix C.1. Formal Reasoning for Law: A Taxonomy

Definition. Let the legal reasoning space be $L = (R, F, C, T, J)$, where R is the set of rules, F is the set of facts, C is the space of conclusions, $T : R \to \mathcal{T}$ is the temporal validity function, and $J : R \to \mathcal{P}(\mathcal{J})$ is the jurisdiction function. A legal rule is represented as $r_i : \varphi_i(x) \to \psi_i(x)$, and the validity constraint is defined as:

$$\text{Valid}(r_i, t, j) \equiv t \in T(r_i) \wedge j \in J(r_i) \tag{A2}$$

 doi:10.20944/preprints202602.0870.v1

Syllogistic reasoning is formalized as:

$$\frac{r : \varphi(x) \to \psi(x), \quad f : \varphi(a), \quad \text{Valid}(r, t, j)}{\therefore \psi(a)} \tag{A3}$$

Verifiability Theorem. An argument $A = \langle R', F', c, t, j \rangle$ is valid if and only if:

$$\bigwedge_{r_i \in R'} \text{Valid}(r_i, t, j) \ \wedge \ (F' \cup R') \models c \tag{A4}$$

where

$R' \subseteq R$ denotes the rules applied, $F' \subseteq F$ denotes the facts invoked, and $\models$ indicates logical entailment. The verification complexity is $O(|R'| \times |F'| \times d)$, where d is the reasoning depth.

Feasibility of SMT Solvers. Satisfiability Modulo Theories (SMT) solvers such as *Z3* are well-suited for legal reasoning verification. Legal reasoning is essentially logical deduction under multiple constraints, which aligns precisely with the constraint satisfaction capabilities of SMT. *Z3* supports combinations of theories such as first-order logic, temporal constraints, and set operations, naturally expressing the logical structure and spatiotemporal constraints of legal rules. By applying *reductio ad absurdum*, it efficiently determines validity and provides a strong technical foundation for constructing trustworthy legal reasoning evaluation frameworks.

SMT solvers let legal norms be encoded as machine-checkable logical formulas, where obligations, permissions, and prohibitions are expressed as constraints over entities, actions, and time. This supports automated consistency checking, revealing conflicts, hidden exceptions, or loopholes by searching for models that satisfy or violate certain rule combinations. Their support for quantifiers and rich data types (integers, reals, arrays, sets) matches the arithmetic and set-theoretic needs of domains like taxation, social benefits, and quantitative compliance. As a result, SMT-based encodings can accurately determine whether fact patterns fall within a rule's scope and whether derived outcomes (such as sanctions or benefits) follow soundly from the premises.

SMT technology also supports counterfactual and what-if analysis. By systematically varying factual assumptions or modifying specific clauses, one can explore how outcomes change and thus assess the robustness, fairness, or potential bias of a legal framework. When a formula is unsatisfiable, modern SMT solvers like *Z3* can produce unsat cores, i.e., minimal subsets of constraints that cause the contradiction. These cores can be mapped back to the underlying legal provisions, providing interpretable explanations of why a certain combination of rules and facts is inconsistent.

In the context of automated legal reasoning systems, SMT solvers can serve as a trusted back-end for both verification and execution. High-level legal-rule languages or graphical modeling tools can be compiled into SMT formulas, while the solver ensures that any derived conclusions are logically entailed by the formalized norms and facts. This separation of concerns allows domain experts to work at a conceptual level while relying on a mature, rigorously tested solver to guarantee correctness and completeness within the expressiveness of the chosen theories.

Finally, because SMT solvers are optimized for scalability and incrementality, they are suitable for large regulatory corpora and dynamically evolving case data. Incremental solving allows new facts or amended regulations to be added without recomputing everything from scratch. Altogether, these features make *Z3* and related SMT technology a compelling backbone for building reliable, transparent, and auditable legal reasoning and evaluation frameworks.

Verification Algorithm.

 doi:10.20944/preprints202602.0870.v1

Table A2. Legal reasoning verification algorithm using SMT solver

Algorithm: LEGALREASONINGVERIFICATION_SMT(R', F', c, t, j)

Input: R' (rules), F' (facts), c (conclusion), t (time), j (jurisdiction)
Output: (validity, counterexample, violations)

solver ← new Z3Solver()
violations ← ∅

for each rule $r : \varphi \to \psi$ **in** R' **do**
    solver.add(ForAll$(x,$ Implies$(\varphi(x), \psi(x)))$)
    if $t \notin T(r)$ **or** $j \notin J(r)$ **then**
        violations.add(r)

if violations $\neq \emptyset$ **then**
    return (False, None, violations)

for each fact f **in** F' **do**
    solver.add(encode_fact(f))

solver.add(Not(c))
result ← solver.check()

if result == UNSAT **then**
    return (True, None, ∅)
    else if result == SAT **then**
        return (False, solver.model(), {"insufficient reasoning"})
    else
        return (Unknown, None, {"solver timeout"})

Appendix D. Products of LegalAI

Legal AI has evolved from experimental prototypes to operational systems, reshaping both investigative and judicial processes. In forensics, deep-learning microservices such as YOLO and NudeNet automate the analysis of seized digital evidence, while models like CriminalNet-228 leverage surveillance data for real-time offender recognition with high accuracy. Judicial applications employ NLP and predictive analytics to promote fairness and consistency in sentencing, while symbolic and machine learning methods enhance semantic search and legislative annotation. The integration of explainable AI and hybrid reasoning frameworks underscores the growing emphasis on accountability and interpretability in legal decision-making. Meanwhile, LLMs such as ChatGPT, Harvey, and CoCounsel are redefining legal drafting and client communication, fostering collaborative workflows between humans and machines. Collectively, these developments signal a maturing and increasingly integrated Legal AI ecosystem that bridges investigation, adjudication, and legal practice.

 doi:10.20944/preprints202602.0870.v1

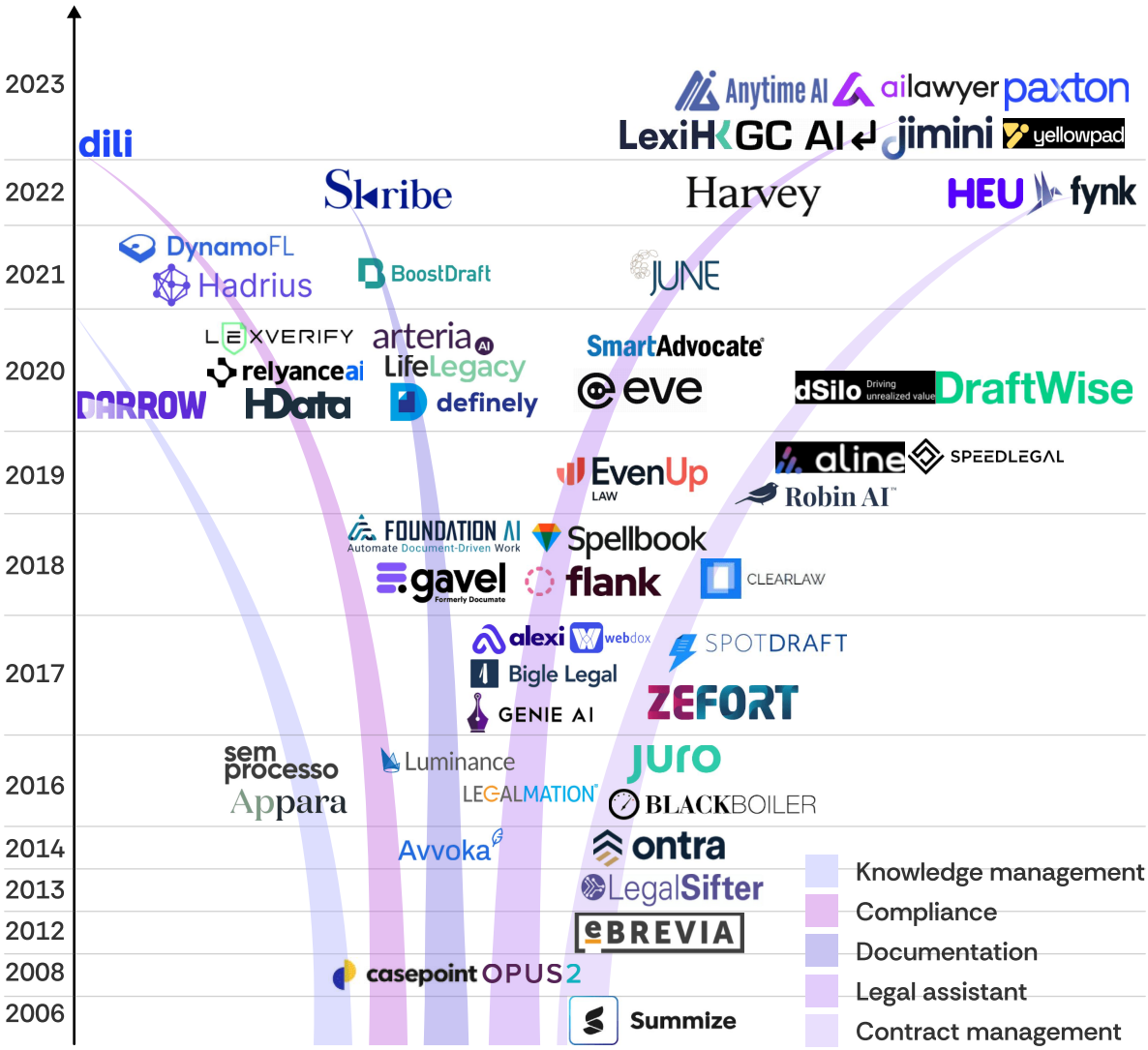

Figure A3. Evolution of generative AI-based legal products from 2006 to 2023, illustrating the emergence of different product categories such as knowledge management, documentation, legal assistance, and contract management.

Appendix E. Supplementary Materials

Appendix E.1. Ethical Concerns in Developing LegalAI

There has been an increasingly explicit call for governance-native frameworks in LLM-enhanced professional services, including LegalAI. A governance-native framework embeds governance constraints ex ante during model design and inference rather than relying on ex post correction. This shift reframes reliability, error resistance, and auditability as architectural commitments rather than downstream safeguards. Early rule-based tools were interpretable but brittle Zheng et al. (2021a). Deep learning vastly expanded capability but at the cost of opacity, motivating accountability-first designs that expose reasoning traces and intermediate steps. The push toward governance-native systems is driven by persistent failure modes such as hallucinations, misgrounding, and authority or temporal violations Magesh et al. (2025). In response, faithfulness pipelines separate linguistic realization from underlying problem solving, enforcing consistency between reasoning and output. Theorem proving and certified interfaces constrain intermediate steps within verifiable logic, turning reasoning into a sequence of machine-checkable commitments Welleck et al. (2022b). Modular decomposition architectures isolate retrieval, citation verification, and hierarchical legal interpretation into dedicated, auditable components Khot et al. (2022). Complementary auditability tooling, including explainability instruments and fact-checking modules, further strengthens error resistance and supports traceable,

 doi:10.20944/preprints202602.0870.v1

accountable decision-making Dathathri et al. (2019). Yet statistical structure imposes hard limits on generative models. Even when trained on error-free corpora, cross-entropy–optimized generators retain irreducible uncertainty on closely confusable facts, producing "confident guesses" that reflect objective-function bias and inherent data ambiguity rather than engineering deficiencies Kalai et al. (2025). Reinforcement-learning setups that score only right or wrong reward confidence and penalize uncertainty, therefore amplifying overconfidence and hallucination. In legal contexts, however, calibration—knowing when not to answer—is indispensable. Our framework therefore aims not at "a model that writes plausible legal language," but at "a system that declines without evidence." Concretely, it couples evidence-backed generation with calibrated confidence estimates and explicit abstention or deferral pathways for high-risk matters, favoring verifiable silence over eloquent error. The core hypothesis is that a governance-native architecture will raise sentence-level support accuracy, reduce out-of-scope reasoning and mis-citations, and improve task efficiency relative to standard approaches W. Shi et al. (2025). Metrics such as support accuracy, false support rate, verifier latency, and human audit time will measure the system's impact C. Cao, Zhu, et al. (2025). By embedding verifiability into the generation process itself, this project develops a unified architecture where "backed by law" becomes a default, machine-checkable property. As datasets scaled, privacy shifted from afterthought to architecture. Federated learning enables cross organization training without pooling raw data, preserving legal compliance and data sovereignty Raji et al. (2020). Paired with process verified reasoning and auditability by design, it yields privacy preserving, transparent systems Fernsel et al. (2024). Safeguards now reason about context: Li et al. (2025a; 2025a) embeds general policy of data protection into benchmarks and guardrail models, treating law as general safeguards for generative AI over roles, sensitivity, and jurisdiction. In legal AI, where leaking sealed records or enabling unlawful advice is unacceptable, these mechanisms enforce compliant generation, enabling calibrated refusals and end to end verification.

Appendix E.2. Differences with Existing Surveys

Several recent surveys have examined NLP and AI in legal contexts Katz et al. (2023); Zhong et al. (2020), but most treat technical methods and legal applications in isolation rather than integrating both perspectives. Our survey differs along four dimensions:

Holistic coverage of the LLM era. Earlier surveys predate the widespread adoption of large-scale pretrained models (e.g., GPT-3.5/4 Achiam et al. (2023); Hurst et al. (2024), Claude 3 Anthropic (2024), LLaMA Touvron et al. (2023)) and instruction tuning paradigms. While Zhong's works provide valuable taxonomies of legal NLP tasks, they primarily address pre-transformer methods or early BERT-style models Zhong et al. (2020). Our survey centers on modern LLM architectures, including RAG, tool-use frameworks, and agentic workflows that have emerged since 2022.

Integration of technical and legal perspectives. Existing technical surveys H. Li, Chen, et al. (2024) emphasize model architectures and benchmark performance but give limited attention to legal doctrine, professional responsibility, and regulatory compliance. Conversely, law-focused reviews Surden (2018) address ethical and policy dimensions but often lack technical depth. We explicitly bridge these gaps by co-developing technical and legal lenses, examining how architectural choices (e.g., citation-grounded generation, neuro-symbolic reasoning) align with legal requirements (e.g., explainability, precedent fidelity) and professional standards (e.g., Model Rules of Professional Conduct in the United States, Solicitors Regulation Authority standards in the United Kingdom).

Emphasis on deployment, governance, and real-world integration. Most surveys evaluate models on static benchmarks without addressing production deployment concerns. We systematically cover workflow integration patterns, human-in-the-loop review protocols, organizational controls (approval workflows, audit logs, version management), data governance (licensing, privacy, secure retrieval), and incident response, drawing on case studies from law firms, legal aid organizations, and courts.

 doi:10.20944/preprints202602.0870.v1

Cross-jurisdictional and multilingual scope. Prior work predominantly focuses on U.S. or English-language legal systems. We instead address multilingual legal NLP, comparative law considerations, and adaptation strategies for civil-law, common-law, and mixed jurisdictions. This includes multilingual corpora such as MultiLegalPile Niklaus et al. (2023)), translation-based transfer learning, and jurisdiction-specific fine-tuning.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. (2024a, March). Do Language Models Know When They're Hallucinating References? In Y. Graham & M. Purver (Eds.), *Findings of the association for computational linguistics: Eacl 2024* (pp. 912–928). St. Julian's, Malta: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.findings-eacl.62/ (accessed on).

Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. (2024b). Do language models know when they're hallucinating references? In *Findings of the association for computational linguistics: Eacl 2024* (pp. 912–928).

Alpay, F., & Alakkad, H. (2025). *Truth-aware decoding: A program-logic approach to factual language generation.* Available online: https://arxiv.org/abs/2510.07331 (accessed on).

Alsagheer, D. R., Kamal, A., Kamal, M., Wu, C. Y., & Shi, W. (2025). The Lawyer That Never Thinks: Consistency and Fairness as Keys to Reliable AI. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9943–9954).

Anthropic. (2024). *The claude 3 model family: Opus, sonnet, haiku.* Available online: https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf (accessed on).

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). *Self-rag: Learning to retrieve, generate, and critique through self-reflection.* Available online: https://arxiv.org/abs/2310.11511 (accessed on).

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025, may 13). A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. *NPJ Digital Medicine*, *8*(1), 274. (PMID: 40360677; PMCID: PMC12075489) https://doi.org/10.1038/s41746-025-01670-7.

Askari, A., Aliannejadi, M., Abolghasemi, A., Kanoulas, E., & Verberne, S. (2023). Closer: conversational legal longformer with expertise-aware passage response ranker for long contexts. In *Proceedings of the 32nd acm international conference on information and knowledge management* (pp. 25–35).

Bayless, S., Buliani, S., Cassel, D., Cook, B., Clough, D., Delmas, R., Diallo, N., Erata, F., Feng, N., Giannakopoulou, D., Goel, A., Gokhale, A., Hendrix, J., Hudak, M., Jovanović, D., Kent, A. M., Kiesl-Reiter, B., Kuna, J. J., Labai, N., … Yao, J. (2025). *A neurosymbolic approach to natural language formalization and verification.* Available online: https://arxiv.org/abs/2511.09008 (accessed on).

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer.* Available online: https://arxiv.org/abs/2004.05150 (accessed on).

Bendová, K., Knap, T., Černỳ, J., Pour, V., Savelka, J., Kvapilíková, I., & Drápal, J. (2025). What Are the Facts? Automated Extraction of Court-Established Facts from Criminal-Court Opinions. *arXiv preprint arXiv:2511.05320*.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024a). *Ai auditing: The broken bus on the road to ai accountability.* Available online: https://arxiv.org/abs/2401.14462 (accessed on).

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024b). AI auditing: The broken bus on the road to AI accountability. In *2024 ieee conference on secure and trustworthy machine learning (satml)* (pp. 612–643).

Cai, H., Zhao, S., Zhang, L., Shen, X., Xu, Q., Shen, W., Wen, Z., & Ban, T. (2025). Unilaw-R1: A Large Language Model for Legal Reasoning with Reinforcement Learning and Iterative Inference. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 18128–18142).

Cao, C., Li, M., Dai, J., Yang, J., Zhao, Z., Zhang, S., Shi, W., Liu, C., Han, S., & Guo, Y. (2025). Towards Advanced Mathematical Reasoning for LLMs via First-Order Logic Theorem Proving. In *Proceedings of the 2024 conference on empirical methods in natural language processing: Emnlp 2025.*

Cao, C., Zhu, H., Ji, J., Sun, Q., Zhu, Z., Wu, Y., Dai, J., Yang, Y., Han, S., & Guo, Y. (2025). SafeLawBench: Towards Safe Alignment of Large Language Models. In *Findings of the association for computational linguistics: Acl 2025* (pp. 14015–14048). Vienna, Austria: Association for Computational Linguistics.

Cao, S., & Wang, L. (2024, August). Verifiable Generation with Subsentence-Level Fine-Grained Citations. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 15584–15596). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.findings-acl.920/ (accessed on). https://doi.org/10.18653/v1/2024.findings-acl.920.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *30th usenix security symposium (usenix security 21)* (pp. 2633–2650).

Chakraborty, N., Ornik, M., & Driggs-Campbell, K. (2025, March). Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art. *ACM Computing Surveys*, *57*(7), 1–35. Available online: http://dx.doi.org/10.1145/3716846 (accessed on). https://doi.org/10.1145/3716846.

Chalkidis, I., Fergadiotis, M., & Androutsopoulos, I. (2021, November). MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6974–6996). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.emnlp-main.559/ (accessed on). https://doi.org/10.18653/v1/2021.emnlp-main.559.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020a). LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020b). *Legal-bert: The muppets straight out of law school.* Available online: https://arxiv.org/abs/2010.02559 (accessed on).

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020c, November). LEGAL-BERT: The Muppets straight out of Law School. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2898–2904). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.findings-emnlp.261/ (accessed on). https://doi.org/10.18653/v1/2020.findings-emnlp.261.

Chalkidis, I., Garneau, N., Goanta, C., Katz, D. M., & Søgaard, A. (2023). LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Chalkidis, I., Garneau, N., Goanta, C., Katz, D. M., & Søgaard, A. (2023). *Lexfiles and legallama: Facilitating english multinational legal language model development.* Available online: https://arxiv.org/abs/2305.07507 (accessed on).

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). *Lexglue: A benchmark dataset for legal language understanding in english.* Available online: https://arxiv.org/abs/2110.00976 (accessed on).

Chen, G., Fan, L., Gong, Z., Xie, N., Li, Z., Liu, Z., Li, C., Qu, Q., Alinejad-Rokny, H., Ni, S., et al. (2025). Agentcourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the association for computational linguistics: Acl 2025* (pp. 5850–5865).

Chen, X., Mao, M., Li, S., & Shangguan, H. (2025). Debate-feedback: A multi-agent framework for efficient legal judgment prediction. *arXiv preprint arXiv:2504.05358*.

Cherian, J. J., Gibbs, I., & Candès, E. J. (2024). *Large language model validity via enhanced conformal prediction methods.* Available online: https://arxiv.org/abs/2406.09714 (accessed on).

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., & Desa, M. (2024a). *Saullm-7b: A pioneering large language model for law.* Available online: https://arxiv.org/abs/2403.03883 (accessed on).

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F., Esposito, F., Raposo, V. L., Morgado, S., et al. (2024b). Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., & Yuan, L. (2023). Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., & Yuan, L. (2024). *Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.* Available online: https://arxiv.org/abs/2306.16092 (accessed on).

Dai, X., Xu, B., Liu, Z., Yan, Y., Xie, H., Yi, X., Wang, S., & Yu, G. (2025). *Legalδ: Enhancing legal reasoning in llms via reinforcement learning with chain-of-thought guided information gain.* Available online: https://arxiv.org/abs/2508.12281 (accessed on).

 doi:10.20944/preprints202602.0870.v1

Dai, Y., Feng, D., Huang, J., Jia, H., Xie, Q., Zhang, Y., Han, W., Tian, W., & Wang, H. (2025). LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st international conference on computational linguistics* (pp. 10738–10766).

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2019). *Plug and play language models: A simple approach to controlled text generation.* Available online: https://arxiv.org/abs/1912.02164 (accessed on).

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023a). QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023b). *Qlora: Efficient finetuning of quantized llms.* Available online: https://arxiv.org/abs/2305.14314 (accessed on).

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020, July). ERASER: A Benchmark to Evaluate Rationalized NLP Models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4443–4458). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.acl-main.408/ (accessed on). https://doi.org/10.18653/v1/2020.acl-main.408.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024, August). Chain-of-Verification Reduces Hallucination in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 3563–3578). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.findings-acl.212/ (accessed on). https://doi.org/10.18653/v1/2024.findings-acl.212.

Fang, H., Balakrishnan, A., Jhamtani, H., Bufe, J., Crawford, J., Krishnamurthy, J., Pauls, A., Eisner, J., Andreas, J., & Klein, D. (2023, July). The Whole Truth and Nothing But the Truth: Faithful and Controllable Dialogue Response Generation with Dataflow Transduction and Constrained Decoding. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 5682–5700). Toronto, Canada: Association for Computational Linguistics. Available online: https://aclanthology.org/2023.findings-acl.351/ (accessed on). https://doi.org/10.18653/v1/2023.findings-acl.351.

Fei, Z., Zhang, S., Shen, X., Zhu, D., Wang, X., Ge, J., & Ng, V. (2025). Internlm-law: An open-sourced chinese legal large language model. In *Proceedings of the 31st international conference on computational linguistics* (pp. 9376–9392).

Fernsel, L., Kalff, Y., & Simbeck, K. (2024). Assessing the Auditability of AI-integrating Systems: A Framework and Learning Analytics Case Study. *arXiv preprint arXiv:2411.08906*.

Gao, T., Wettig, A., Yen, H., & Chen, D. (2025). How to train long-context language models (effectively). In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7376–7399).

Geng, S., Josifoski, M., Peyrard, M., & West, R. (2024). *Grammar-constrained decoding for structured nlp tasks without finetuning.* Available online: https://arxiv.org/abs/2305.13971 (accessed on).

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, *36*, 44123–44279.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., … Li, Z. (2023). LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., … Li, Z. (2023). *Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.* Available online: https://arxiv.org/abs/2308.11462 (accessed on).

Guo, C., Sablayrolles, A., Jégou, H., & Kiela, D. (2021). *Gradient-based adversarial attacks against text transformers.* Available online: https://arxiv.org/abs/2104.13733 (accessed on).

Guo, Q., Dong, Y., Tian, L., Kang, Z., Zhang, Y., & Wang, S. (2025). BANER: Boundary-aware LLMs for few-shot named entity recognition. In *Proceedings of the 31st international conference on computational linguistics* (pp. 10375–10389).

 doi:10.20944/preprints202602.0870.v1

Gupta, A., Rice, D., & O'Connor, B. (2025). -Stance: A Large-Scale Real World Dataset of Stances in Legal Argumentation. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 31450–31467).

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.acl-main.740/ (accessed on). https://doi.org/10.18653/v1/2020.acl-main.740.

Guttmann, K., Charkiewicz, A., Rostek, Z., Pokrywka, M., & Nowakowski, A. (2025, November). Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the tenth conference on machine translation* (pp. 1276–1283). Suzhou, China: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.wmt-1.107/ (accessed on).

Han, S., Guo, Y.-K., & Huang, S. (2025). Hong Kong Generative Artificial Intelligence Technical and Application Guideline. *Hong Kong SAR Government Press Release*.

Hepworth, I., Olive, K., Dasgupta, K., Le, M., Lodato, M., Maruseac, M., Meiklejohn, S., Chaudhuri, S., & Minkus, T. (2024). *Securing the ai software supply chain* (Tech. Rep.). Google.

Hong Kong Productivity Council (HKPC). (2024, nov 21). *Hong kong enterprise cyber security readiness index and ai security survey.* Press Release. Available online: https://www.hkpc.org/en/about-us/media-centre/press-releases/2024/hong-kong-enterprise-cyber-security-readiness-index (accessed on).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *Lora: Low-rank adaptation of large language models.* Available online: https://arxiv.org/abs/2106.09685 (accessed on).

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025, January). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. Available online: http://dx.doi.org/10.1145/3703155 (accessed on). https://doi.org/10.1145/3703155.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., & Feng, Y. (2023). *Lawyer llama technical report.* Available online: https://arxiv.org/abs/2305.15062 (accessed on).

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ji, J., Chen, X., Pan, R., Zhang, C., Zhu, H., Li, J., Hong, D., Chen, B., Zhou, J., Wang, K., et al. (2025). Safe RLHF-V: Safe Reinforcement Learning from Multi-modal Human Feedback. *arXiv preprint arXiv:2503.17682*.

Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T. A., Zhou, J., Wang, K., Li, B., Han, S., Guo, Y., & Yang, Y. (2025, July). PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 31983–32016). Vienna, Austria: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.acl-long.1544/ (accessed on). https://doi.org/10.18653/v1/2025.acl-long.1544.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023, March). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. Available online: http://dx.doi.org/10.1145/3571730 (accessed on). https://doi.org/10.1145/3571730.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). *Active retrieval augmented generation.* Available online: https://arxiv.org/abs/2305.06983 (accessed on).

Ju, C., Shi, W., Liu, C., Ji, J., Zhang, J., Zhang, R., Xu, J., Yang, Y., Han, S., & Guo, Y. (2025). Benchmarking multi-national value alignment for large language models. In *Findings of the association for computational linguistics: Acl 2025* (pp. 20042–20058).

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate.* Available online: https://arxiv.org/abs/2509.04664 (accessed on).

Kalai, A. T., & Vempala, S. S. (2024). Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th annual acm symposium on theory of computing* (p. 160–171). New York, NY, USA: Association for Computing Machinery. Available online: https://doi.org/10.1145/3618260.3649777 (accessed on). https://doi.org/10.1145/3618260.3649777.

Kaleeswaran, S., Santhiar, A., Kanade, A., & Gulwani, S. (2016). Semi-supervised verified feedback generation. In *Proceedings of the 2016 24th acm sigsoft international symposium on foundations of software engineering* (pp. 739–750).

Katz, D. M., Hartung, D., Gerlach, L., Jana, A., & Bommarito II, M. J. (2023). Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2022). *Decomposed prompting: A modular approach for solving complex tasks.* Available online: https://arxiv.org/abs/2210.02406 (accessed on).

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2023). *Decomposed prompting: A modular approach for solving complex tasks.* Available online: https://arxiv.org/abs/2210.02406 (accessed on).

Klisura, Đ., Torres, A. R. B., Gárate-Escamilla, A. K., Biswal, R. R., Yang, K., Pataci, H., & Rios, A. (2025). A Multi-Agent Framework for Mitigating Dialect Biases in Privacy Policy Question-Answering Systems. *arXiv preprint arXiv:2506.02998*.

Kornilova, A., & Eidelman, V. (2019). BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd workshop on new frontiers in summarization* (pp. 48–56).

Kuhn, L., Gal, Y., & Farquhar, S. (2023). *Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.* Available online: https://arxiv.org/abs/2302.09664 (accessed on).

kui Sin, K., Xuan, X., Kit, C., yan Chan, C. H., & kin Ip, H. H. (2025). *Solving the unsolvable: Translating case law in hong kong.* Available online: https://arxiv.org/abs/2501.09444 (accessed on).

Kullback, S. (1951). Kullback-leibler divergence. *Tech. Rep.*.

Lee, J.-S. (2023). Lexgpt 0.1: pre-trained gpt-j models with pile of law. *arXiv preprint arXiv:2306.05431*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K"uttler, H., Lewis, M., Yih, W.-t., Rockt"aschel, T., et al. (2020a). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020b). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive nlp tasks.* Available online: https://arxiv.org/abs/2005.11401 (accessed on).

Li, C., Wang, S., Zhang, J., & Zong, C. (2024, June). Improving In-context Learning of Multilingual Generative Language Models with Cross-lingual Alignment. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 8058–8076). Mexico City, Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.naacl-long.445/ (accessed on). https://doi.org/10.18653/v1/2024.naacl-long.445.

Li, H., Chen, Y., Ai, Q., Wu, Y., Zhang, R., & Liu, Y. (2024). Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, *37*, 25061–25094.

Li, H., Chen, Y., Zeng, J., Peng, H., Jing, H., Hu, W., Yang, X., Zeng, Z., Han, S., & Song, Y. (2025a). *Gspr: Aligning llm safeguards as generalizable safety policy reasoners.* Available online: https://arxiv.org/abs/2509.24418 (accessed on).

Li, H., Chen, Y., Zeng, J., Peng, H., Jing, H., Hu, W., Yang, X., Zeng, Z., Han, S., & Song, Y. (2025b). GSPR: Aligning LLM Safeguards as Generalizable Safety Policy Reasoners. *arXiv preprint arXiv:2509.24418*.

Li, H., Hu, W., Jing, H., Chen, Y., Hu, Q., Han, S., Chu, T., Hu, P., & Song, Y. (2025a). *Privaci-bench: Evaluating privacy with contextual integrity and legal compliance.* Available online: https://arxiv.org/abs/2502.17041 (accessed on).

Li, H., Hu, W., Jing, H., Chen, Y., Hu, Q., Han, S., Chu, T., Hu, P., & Song, Y. (2025b). Privaci-bench: Evaluating privacy with contextual integrity and legal compliance. *arXiv preprint arXiv:2502.17041*.

Li, H., Shao, Y., Wu, Y., Ai, Q., Ma, Y., & Liu, Y. (2024). Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 2251–2260).

Li, L., Li, D., Lin, C., Li, W., Xue, W., Han, S., & Guo, Y. (2025). AIRA: Activation-Informed Low-Rank Adaptation for Large Models. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1729–1739).

Li, L., Lin, C., Li, D., Huang, Y.-L., Li, W., Wu, T., Zou, J., Xue, W., Han, S., & Guo, Y. (2025). Efficient Fine-Tuning of Large Models via Nested Low-Rank Adaptation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 22252–22262).

Li, X., Cao, Y., Pan, L., Ma, Y., & Sun, A. (2024, August). Towards Verifiable Generation: A Benchmark for Knowledge-aware Language Model Attribution. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 493–516). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.findings-acl.28/ (accessed on). https://doi.org/10.18653/v1/2024.findings-acl.28.

Lin, S., Hilton, J., & Evans, O. (2022a). *Teaching models to express their uncertainty in words.* Available online: https://arxiv.org/abs/2205.14334 (accessed on).

Lin, S., Hilton, J., & Evans, O. (2022b). *Teaching models to express their uncertainty in words.* Available online: https://arxiv.org/abs/2205.14334 (accessed on).

Ling, C., Zhao, X., Zhang, X., Cheng, W., Liu, Y., Sun, Y., Oishi, M., Osaki, T., Matsuda, K., Ji, J., Bai, G., Zhao, L., & Chen, H. (2024, June). Uncertainty Quantification for In-Context Learning of Large Language Models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 3357–3370). Mexico City, Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.naacl-long.184/ (accessed on). https://doi.org/10.18653/v1/2024.naacl-long.184.

Liu, C., Kang, Y., Wang, S., Qing, L., Zhao, F., Wu, C., Sun, C., Kuang, K., & Wu, F. (2024). More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 7531–7548).

Liu, X., Liang, T., He, Z., Xu, J., Wang, W., He, P., Tu, Z., Mi, H., & Yu, D. (2025). Trust, But Verify: A Self-Verification Approach to Reinforcement Learning with Verifiable Rewards. *arXiv preprint arXiv:2505.13445*.

Lu, H., Fang, L., Zhang, R., Li, X., Cai, J., Cheng, H., Tang, L., Liu, Z., Sun, Z., Wang, T., et al. (2025). Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *arXiv preprint arXiv:2507.19672*.

Lu, X., West, P., Zellers, R., Le Bras, R., Bhagavatula, C., & Choi, Y. (2021, June). NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints. In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4288–4299). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.naacl-main.339/ (accessed on). https://doi.org/10.18653/v1/2021.naacl-main.339.

Luo, J., Kou, Z., Yang, L., Luo, X., Huang, J., Xiao, Z., Peng, J., Liu, C., Ji, J., Liu, X., et al. (2025). FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. *arXiv preprint arXiv:2505.24714*.

Luo, K., Huang, Q., Jiang, C., & Feng, Y. (2025). Automating Legal Interpretation with LLMs: Retrieval, Generation, and Evaluation. *arXiv preprint arXiv:2501.01743*.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., & Callison-Burch, C. (2023). *Faithful chain-of-thought reasoning.* Available online: https://arxiv.org/abs/2301.13379 (accessed on).

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). *Hallucination-free? assessing the reliability of leading ai legal research tools.* Available online: https://arxiv.org/abs/2405.20362 (accessed on).

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019, January). Model Cards for Model Reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 220–229). ACM. Available online: http://dx.doi.org/10.1145/3287560.3287596 (accessed on). https://doi.org/10.1145/3287560.3287596.

Mudgal, S., Lee, J., Ganapathy, H., Li, Y., Wang, T., Huang, Y., Chen, Z., Cheng, H.-T., Collins, M., Strohman, T., Chen, J., Beutel, A., & Beirami, A. (2024). *Controlled decoding from language models.* Available online: https://arxiv.org/abs/2310.17022 (accessed on).

Mündler, N., He, J., Jenko, S., & Vechev, M. (2024). *Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.* Available online: https://arxiv.org/abs/2305.15852 (accessed on).

Nguyen, H. T., Fungwacharakorn, W., Zin, M. M., Goebel, R., Toni, F., Stathis, K., & Satoh, K. (2025). LLMs for legal reasoning: A unified framework and future perspectives. *Computer Law & Security Review*, *58*, 106165.

Nguyen, M., Gupta, S., & Le, H. (2025). *Caad: Context-aware adaptive decoding for truthful text generation.* Available online: https://arxiv.org/abs/2508.02184 (accessed on).

 doi:10.20944/preprints202602.0870.v1

Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-t., Wang, S., & Lin, X. V. (2023). Lever: Learning to verify language-to-code generation with execution. In *International conference on machine learning* (pp. 26106–26128).

Nigam, S. K., Patnaik, B. D., Mishra, S., Thomas, A. V., Shallum, N., Ghosh, K., & Bhattacharya, A. (2025). NyayaRAG: Realistic Legal Judgment Prediction with RAG under the Indian Common Law System. *arXiv preprint arXiv:2508.00709*.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., & Xiong, C. (2023). *Codegen: An open large language model for code with multi-turn program synthesis.* Available online: https://arxiv.org/abs/2203.13474 (accessed on).

Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., & Ho, D. E. (2023). Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., & Ho, D. E. (2024). *Multilegalpile: A 689gb multilingual legal corpus.* Available online: https://arxiv.org/abs/2306.02069 (accessed on).

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models.

Oehri, M., Conti, G., Pather, K., Rossi, A., Serra, L., Parody, A., Johannesen, R., Petersen, A., & Krasniqi, A. (2025). *Trusted uncertainty in large language models: A unified framework for confidence calibration and risk-controlled refusal.* Available online: https://arxiv.org/abs/2509.01455 (accessed on).

Östling, A., Sargeant, H., Xie, H., Bull, L., Terenin, A., Jonsson, L., Magnusson, M., & Steffek, F. (2023). The Cambridge law corpus: A dataset for legal AI research. *Advances in Neural Information Processing Systems*, *36*, 41355–41385.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2021, April). AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 487–503). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.eacl-main.39/ (accessed on). https://doi.org/10.18653/v1/2021.eacl-main.39.

Poesia, G., Gandhi, K., Zelikman, E., & Goodman, N. D. (2023a). *Certified deductive reasoning with language models.* Available online: https://arxiv.org/abs/2306.04031 (accessed on).

Poesia, G., Gandhi, K., Zelikman, E., & Goodman, N. D. (2023b). Certified Reasoning with Language Models. *ArXiv*, *abs/2306.04031*. Available online: https://api.semanticscholar.org/CorpusID:259095869 (accessed on).

Rabelo, J., Kim, M.-Y., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2020). COLIEE 2020: Methods for Legal Document Retrieval and Entailment. In *New frontiers in artificial intelligence: Jsai-isai 2020 workshops, jurisin, lenls 2020 workshops, virtual event, november 15–17, 2020, revised selected papers* (p. 196–210). Berlin, Heidelberg: Springer-Verlag. Available online: https://doi.org/10.1007/978-3-030-79942-7_13 (accessed on). https://doi.org/10.1007/978-3-030-79942-7_13.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020, January). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(1).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33–44).

Raptopoulos, P., Filandrianos, G., Lymperaiou, M., & Stamou, G. (2025). PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements. *arXiv preprint arXiv:2506.00608*.

ritun16. (2024). *Chain-of-verification implementation.* GitHub repository. Available online: https://github.com/ritun16/chain-of-verification (accessed on).

Ryu, C., Lee, S., Pang, S., Choi, C., Choi, H., Min, M., & Sohn, J.-Y. (2023). Retrieval-based evaluation for LLMs: a case study in Korean legal QA. In *Proceedings of the natural legal language processing workshop 2023* (pp. 132–137).

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2022). *Multitask prompted training enables zero-shot task generalization.* Available online: https://arxiv.org/abs/2110.08207 (accessed on).

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, *36*, 68539–68551.

 doi:10.20944/preprints202602.0870.v1

Schöning, J., & Kruse, N. (2025). Compliance of AI Systems. *arXiv preprint arXiv:2503.05571*.

Schweighofer, K., Brune, B., Gruber, L., Schmid, S., Aufreiter, A., Gruber, A., Doms, T., Eder, S., Mayer, F., Stadlbauer, X.-P., Schwald, C., Zellinger, W., Nessler, B., & Hochreiter, S. (2025). *Safe and certifiable ai systems: Concepts, challenges, and lessons learned.* Available online: https://arxiv.org/abs/2509.08852 (accessed on).

ShengbinYue, S., Huang, T., Jia, Z., Wang, S., Liu, S., Song, Y., Huang, X.-J., & Wei, Z. (2025). Multi-agent simulator drives language models for legal intensive interaction. In *Findings of the association for computational linguistics: Naacl 2025* (pp. 6537–6570).

Shi, J., Guo, Q., Liao, Y., Wang, Y., Chen, S., & Liang, S. (2024). Legal-LM: Knowledge graph enhanced large language models for law consulting. In *International conference on intelligent computing* (pp. 175–186).

Shi, W., Zhu, H., Ji, J., Li, M., Zhang, J., Zhang, R., Zhu, J., Xu, J., Han, S., & Guo, Y. (2025). LegalReasoner: Step-wised Verification-Correction for Legal Judgment Reasoning. *arXiv preprint arXiv:2506.07443*.

Shin, R., Lin, C., Thomson, S., Chen, C., Roy, S., Platanios, E. A., Pauls, A., Klein, D., Eisner, J., & Van Durme, B. (2021, November). Constrained Language Models Yield Few-Shot Semantic Parsers. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7699–7715). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.emnlp-main.608/ (accessed on). https://doi.org/10.18653/v1/2021.emnlp-main.608.

Shu, D., Zhao, H., Liu, X., Demeter, D., Du, M., & Zhang, Y. (2024, October). LawLLM: Law Large Language Model for the US Legal System. In *Proceedings of the 33rd acm international conference on information and knowledge management* (p. 4882–4889). ACM. Available online: http://dx.doi.org/10.1145/3627673.3680020 (accessed on). https://doi.org/10.1145/3627673.3680020.

Smith, J., Jones, A., & Williams, B. (2023). A Comprehensive Overview of the Development and Impact of Large Language Models. *Journal of Artificial Intelligence Research*, *70*, 1–50.

Song, F., Yu, B., Lang, H., Yu, H., Huang, F., Wang, H., & Li, Y. (2024). Scaling data diversity for fine-tuning language models in human alignment. *arXiv preprint arXiv:2403.11124*.

South, T. (2025). *Private, verifiable, and auditable ai systems.* Available online: https://arxiv.org/abs/2509.00085 (accessed on).

Su, W., Yue, B., Ai, Q., Hu, Y., Li, J., Wang, C., Zhang, K., Wu, Y., & Liu, Y. (2025). Judge: Benchmarking judgment document generation for chinese legal system. In *Proceedings of the 48th international acm sigir conference on research and development in information retrieval* (pp. 3573–3583).

Sun, J., Dai, C., Luo, Z., Chang, Y., & Li, Y. (2024). *Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation.* Available online: https://arxiv.org/abs/2407.16252 (accessed on).

Sun, K., Wu, J., Guo, M., Li, J., & Huang, J. (2025). *Accurate target privacy preserving federated learning balancing fairness and utility.* Available online: https://arxiv.org/abs/2510.26841 (accessed on).

Surden, H. (2018). Artificial intelligence and law: An overview. *Ga. St. UL Rev.*, *35*, 1305.

The Law Society of Hong Kong (LSHK). (2024, jan 20). *The impact of artificial intelligence on the legal profession: Position paper.* Position Paper. Available online: https://www.hklawsoc.org.hk/-/media/HKLS/Home/News/2024/LSHK-Position-Paper_AI_EN.pdf (accessed on).

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). *Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.* Available online: https://arxiv.org/abs/2305.14975 (accessed on).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tran, H. T. H., Chatterjee, N., Pollak, S., & Doucet, A. (2024). Deberta beats behemoths: A comparative analysis of fine-tuning, prompting, and peft approaches on legallensner. In *Proceedings of the natural legal language processing workshop 2024* (pp. 371–380).

Tuggener, D., Von Däniken, P., Peetz, T., & Cieliebak, M. (2020). LEDGAR: a large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 1235–1241).

Union, E. (2016). *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).* Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 (accessed on). (Official Journal of the European Union: L 119, 4 May 2016, pp. 1–88; Entered into force: 25 May 2018)

 doi:10.20944/preprints202602.0870.v1

Union, E. (2024). *Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act).* Available online: https://www.aiact-info.eu/full-text-and-pdf-download/ (accessed on). (Content consistent with Official Journal of the European Union (L 2024/1689, 12 July 2024); Entered into force: 1 August 2024; PDF and full text provided by EU-authorised platform (aiact-info.eu))

Valmeekam, K., Stechly, K., & Kambhampati, S. (2024). *Llms still can't plan; can lrms? a preliminary evaluation of openai's o1 on planbench.* Available online: https://arxiv.org/abs/2409.13373 (accessed on).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S., Fan, W., Feng, Y., Lin, S., Ma, X., Wang, S., & Yin, D. (2025). *Knowledge graph retrieval-augmented generation for llm-based recommendation.* Available online: https://arxiv.org/abs/2501.02226 (accessed on).

Wang, T., Yu, P., Tan, X. E., O'Brien, S., Pasunuru, R., Dwivedi-Yu, J., Golovneva, O., Zettlemoyer, L., Fazel-Zarandi, M., & Celikyilmaz, A. (2023). Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.

Wang, X., Sen, P., Li, R., & Yilmaz, E. (2025, April). Adaptive Retrieval-Augmented Generation for Conversational Systems. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: Naacl 2025* (pp. 491–503). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.findings-naacl.30/ (accessed on). https://doi.org/10.18653/v1/2025.findings-naacl.30.

Wang, X., Zhang, X., Hoo, V., Shao, Z., & Zhang, X. (2024). LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. *IEEE Access*.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). *Self-instruct: Aligning language models with self-generated instructions.* Available online: https://arxiv.org/abs/2212.10560 (accessed on).

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned language models are zero-shot learners.* Available online: https://arxiv.org/abs/2109.01652 (accessed on).

Welleck, S., Liu, J., Lu, X., Hajishirzi, H., & Choi, Y. (2022a). *Naturalprover: Grounded mathematical proof generation with language models.* Available online: https://arxiv.org/abs/2205.12910 (accessed on).

Welleck, S., Liu, J., Lu, X., Hajishirzi, H., & Choi, Y. (2022b). Naturalprover: Grounded mathematical proof generation with language models. In *Advances in neural information processing systems* (Vol. 35, pp. 4913–4927).

Wu, Y., Wang, C., Gumusel, E., & Liu, X. (2024). Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. *arXiv preprint arXiv:2406.03600*.

Xie, N., Bai, Y., Gao, H., Xue, Z., Fang, F., Zhao, Q., Li, Z., Zhu, L., Ni, S., & Yang, M. (2024). Delilaw: A chinese legal counselling system based on a large language model. In *Proceedings of the 33rd acm international conference on information and knowledge management* (pp. 5299–5303).

Xu, Q., Liu, Q., Fei, H., Yu, H., Guan, S., & Wei, X. (2025). CLEAR: A Framework Enabling Large Language Models to Discern Confusing Legal Paragraphs. In *Findings of the association for computational linguistics: Emnlp 2025* (pp. 8937–8953).

Xu, Q., Wei, X., Yu, H., Liu, Q., & Fei, H. (2024). Divide and Conquer: Legal Concept-guided Criminal Court View Generation. In *Findings of the association for computational linguistics: Emnlp 2024* (pp. 3395–3410).

Xu, X., Zhao, L., Xu, H., & Chen, C. (2025). CLaw: Benchmarking Chinese Legal Knowledge in Large Language Models-A Fine-grained Corpus and Reasoning Analysis. *arXiv preprint arXiv:2509.21208*.

Yang, X.-W., Shao, J.-J., Guo, L.-Z., Zhang, B.-W., Zhou, Z., Jia, L.-H., Dai, W.-Z., & Li, Y. (2025). *Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models.* Available online: https://arxiv.org/abs/2508.13678 (accessed on).

Yang, Z., Dong, L., Du, X., Cheng, H., Cambria, E., Liu, X., Gao, J., & Wei, F. (2024). *Language models as inductive reasoners.* Available online: https://arxiv.org/abs/2212.10923 (accessed on).

Yao, R., Wu, Y., Wang, C., Xiong, J., Wang, F., & Liu, X. (2025). Elevating legal LLM responses: harnessing trainable logical structures and semantic knowledge with legal reasoning. *arXiv preprint arXiv:2502.07912*.

Ye, F., & Li, S. (2024). MileCut: A multi-view truncation framework for legal case retrieval. In *Proceedings of the acm web conference 2024* (pp. 1341–1349).

Yu, F., Seedat, N., Herrmannova, D., Schilder, F., & Schwarz, J. R. (2025). Beyond Pointwise Scores: Decomposed Criteria-Based Evaluation of LLM Responses. In *Proceedings of the 2025 conference on empirical methods in natural language processing: Industry track* (pp. 1931–1954).

Yuan, W., Cao, J., Jiang, Z., Kang, Y., Lin, J., Song, K., Lin, T., Yan, P., Sun, C., & Liu, X. (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* (pp. 7577–7597).

Yue, L., Liu, Q., Du, Y., Gao, W., Liu, Y., & Yao, F. (2024). *Fedjudge: Federated legal large language model.* Available online: https://arxiv.org/abs/2309.08173 (accessed on).

Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Huang, X., & Wei, Z. (2023). *Disc-lawllm: Fine-tuning large language models for intelligent legal services.* Available online: https://arxiv.org/abs/2309.11325 (accessed on).

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: transformers for longer sequences. In *Proceedings of the 34th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Zhang, K., Chen, C., Wang, Y., Tian, Q., & Bai, L. (2023). Cfgl-lcr: A counterfactual graph learning framework for legal case retrieval. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining* (pp. 3332–3341).

Zhang, K., Xie, G., Yu, W., Xu, M., Tang, X., Li, Y., & Xu, J. (2025). Legal Mathematical Reasoning with LLMs: Procedural Alignment through Two-Stage Reinforcement Learning. *arXiv preprint arXiv:2504.02590*.

Zhang, R., Sullivan, D., Jackson, K., Xie, P., & Chen, M. (2025, April). Defense against Prompt Injection Attacks via Mixture of Encodings. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *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)* (pp. 244–252). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.naacl-short.21/ (accessed on). https://doi.org/10.18653/v1/2025.naacl-short.21.

Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., & Gonzalez, J. E. (2024). *Raft: Adapting language model to domain specific rag.* Available online: https://arxiv.org/abs/2403.10131 (accessed on).

Zhang, X., Ruan, J., Ma, X., Zhu, Y., Zhao, H., Li, H., Chen, J., Zeng, K., & Cai, X. (2025). When to continue thinking: Adaptive thinking mode switching for efficient reasoning. *arXiv preprint arXiv:2505.15400*.

Zhao, R., Li, X., Joty, S., Qin, C., & Bing, L. (2023). *Verify-and-edit: A knowledge-enhanced chain-of-thought framework.* Available online: https://arxiv.org/abs/2305.03268 (accessed on).

ZHAO, W., HU, Y., DENG, Y., GUO, J., SUI, X., HAN, X., ZHANG, A., ZHAO, Y., QIN, B., CHUA, T.-S., et al. (2025). Beware of your Po! Measuring and mitigating AI safety risks in role-play fine-tuning of LLMs.(2025). In *Proceedings of the 63rd annual meeting of the association for computational linguistics, vienna, austria* (pp. 5131–5157).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2).

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E. (2021a). When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law* (pp. 159–168).

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E. (2021b). When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law* (pp. 159–168).

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

Zhou, J. P., Staats, C., Li, W., Szegedy, C., Weinberger, K. Q., & Wu, Y. (2024). *Don't trust: Verify – grounding llm quantitative reasoning with autoformalization.* Available online: https://arxiv.org/abs/2403.18120 (accessed on).

Zhou, Z., Shi, J.-X., Song, P.-X., Yang, X.-W., Jin, Y.-X., Guo, L.-Z., & Li, Y.-F. (2024). *Lawgpt: A chinese legal knowledge-enhanced large language model.* Available online: https://arxiv.org/abs/2406.04614 (accessed on).

Zöller, M.-A., Iurshina, A., & Röder, I. (2025). Trustworthy Generative AI for Financial Services (Practitioner Track). In *Symposium on scaling ai assessments (saia 2024)* (pp. 2–1).