Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 02, Issue 09, September 2025,
Publish Date: 09-13-2025
DOI - https://doi.org/10.37547/feaiml/Volume02Issue09-02 PageNo.07-16

Agentic Legal Intake: A Multi-Agent Framework For Hallucination-Free, Audit-Ready AI Screening In Mass-Tort Litigation

Tejas Sarvankar
Independent Researcher, Carnegie Mellon University (Alumnus), USA

Anna John
Independent Researcher, Carnegie Mellon University (Alumnus), USA

Abstract

This study presents a multi-agent framework to address the risks of large language models (LLMs) in legal intake, particularly in mass tort litigation. The research focuses on mitigating a phenomenon known as hallucination, where LLMs generate plausible but false information. The study's objective is to evaluate if a system of peer-auditing agents, along with human involvement, can outperform a traditional single-agent model in terms of accuracy, data completeness, and audit efficiency. The methodology involved a mixed-methods design, using a multi-agent system with distinct Extractor, Validator, and Auditor agents, followed by human review. This system was tested on 100 anonymized mass tort intake cases, with 70% being real and 30% being synthetic. The quantitative metrics measured were hallucination rate, completeness score, and human review time. Qualitative analysis was also performed, based on feedback from six legal operations professionals. The multi-agent framework demonstrated a substantial reduction in the hallucination rate, from 21% in the single-agent baseline to just 5%, a 76% decrease. It also significantly improved data completeness, achieving a 92% score compared to 74% in the baseline, which is an 18 percentage point increase. Furthermore, the time required for human review of finalized cases dropped by 51%. Qualitative feedback from professionals highlighted increased trust and transparency in the agent-generated outputs due to the built-in audit trails. However, some noted issues with precision. In conclusion, the findings confirm that a structured multi-agent LLM framework is a highly effective way to improve the reliability and efficiency of legal intake workflows. By mimicking human peer-review processes, this agentic approach transforms AI into a transparent and accountable augmentation tool. This study emphasizes that agentic AI is accountable augmentation, paving the way for explainable and scalable legal AI systems.

Keywords: LLM Automation, Legal Intake, Multi-Agent AI, Mass Tort, Auditability, AI Hallucination, Legal Compliance, Trustworthy AI, Explainable AI, Responsible AI.

1. Introduction

The legal sector is undergoing a significant transformation due to the rapid advancements in artificial intelligence (AI). Large language models (LLMs) are now being used in various legal tasks, such as initial client intake, case screening, and document analysis, particularly within the demanding area of mass tort litigation. The integration of these tools promises considerable efficiency gains, potentially streamlining labor-intensive processes, reducing operational costs, and accelerating the pace of justice.

However, this shift also comes with significant risks. One of the main challenges with LLMs is their tendency to "hallucinate," producing information that seems plausible but is factually incorrect or fabricated. This is not just an academic issue; it has severe consequences in a sensitive field like law. Studies have shown alarming hallucination rates for general-purpose LLMs on legal queries, with some ranging from 58% to 82% [1, 2]. A notable real-world example is a New York lawyer who was sanctioned for filing motions based on fictitious, AI-generated citations [3]. Such incidents highlight the critical need for advanced safeguards, as AI hallucination introduces litigation risk and operational strains within law firms.

Despite the growing use of LLMs in legal practice, a research gap exists concerning structured, multi- agent LLM frameworks that can provide real-time quality control, comprehensive audit trails, and iterative learning. Most existing systems treat AI as a single, monolithic entity, overlooking the potential of agentic collaboration. These solutions often lack the built-in mechanisms for self-correction, peer- validation, or transparent justification of their outputs—all of which are crucial for the accuracy and trustworthiness required for legal compliance.

This study aims to fill this gap by introducing a new multi-agent architecture specifically designed for legal intake. In this framework, specialized LLM agents are engineered to audit and refine each other's outputs, all within a human-in-the-loop paradigm. Our clear objective is to significantly reduce hallucinations, enhance auditability, and improve screening accuracy. By demonstrating how structured multi-agent LLM workflows can enhance accuracy and trustworthiness, this research offers a robust solution for ensuring responsible and effective LLM automation in law.

The paper is structured as follows: an introduction, a literature review, a detailed methodology, a presentation of the results, a discussion of the findings, and a conclusion.

2. Literature Review

The integration of LLMs into legal practice presents significant opportunities but also highlights the critical vulnerability of AI hallucination. This section reviews existing literature on LLM hallucinations in legal contexts, explores the limitations of current mitigation strategies like Retrieval-Augmented Generation (RAG), and examines promising emerging approaches, particularly agentic AI. We will also delve into the mechanistic causes of hallucination and how a multi-agent AI framework can offer robust mitigation, ultimately identifying the research gap this study addresses.

2.1 LLM Hallucination in Legal Contexts

LLM hallucination, where models generate factually incorrect or fabricated information that sounds plausible, poses a severe threat to AI reliability in high-stakes legal fields. Accuracy is paramount in law, where errors can lead to professional misconduct and significant litigation risk. Empirical studies consistently show high hallucination rates in general-purpose LLMs when dealing with legal queries. For instance, Stanford’s "Large Legal Fictions" study found that GPT-4 hallucinated between 58% and 82% of the time on legal tasks, often fabricating case precedents or citing non-existent statutes [1, 2]. Real-world incidents, such as a New York lawyer being sanctioned for using fictitious AI-generated citations, underscore these risks [3]. Such cases emphasize the urgent need for robust validation protocols for LLM automation tools. Even domain-specific applications are not immune, with recent studies evaluating LLMs in legal practice finding that while they excel at some tasks, they still struggle with accuracy and require significant human oversight to avoid critical errors [10]. This suggests that simply grounding responses in a knowledge base, while helpful, is insufficient to eliminate all forms of hallucination, especially when complex legal reasoning is involved.

2.2 Limitations of RAG in Legal AI

Retrieval-Augmented Generation (RAG) aims to improve LLM factual accuracy by enabling models to retrieve information from external knowledge bases before generating a response. In legal AI, RAG’s promise to ground outputs in authoritative texts is appealing for reducing AI hallucination. However, RAG systems have inherent limitations that prevent them from completely eliminating hallucinations and ensuring full legal compliance.

RAG primarily addresses "extrinsic" hallucinations, which result from a lack of external knowledge. By providing relevant documents, RAG helps LLMs generate factually accurate responses. However, it is less effective against "intrinsic" hallucinations, which stem from the model's internal reasoning flaws, misinterpretations of retrieved content, or logical inconsistencies. An LLM, even with RAG, might misinterpret subtle nuances or draw incorrect conclusions from retrieved legal texts.

Furthermore, even grounded RAG systems still produce errors for several reasons:

- **Ambiguous Inputs**: Legal intake forms, particularly in mass tort cases, often contain ambiguous or incomplete information. RAG systems might retrieve conflicting data, and the LLM may struggle to synthesize it accurately, as it lacks the sophisticated reasoning to interpret ambiguities or ask clarifying questions.

- **Enforcing Hard Legal Constraints**: Many legal requirements are "hard constraints"—binary rules that demand strict adherence, such as deadlines or specific eligibility criteria. RAG systems, being probabilistic, are not designed to enforce such rigid rules. They might generate generally correct responses that overlook a specific, non-negotiable legal constraint, leading to compliance gaps.

- **Lack of Interpretive Depth**: While RAG provides raw data, it doesn't inherently give the LLM the deep interpretive capabilities of a human legal professional. The model might retrieve relevant precedents but fail to grasp why a particular precedent is applicable or distinguishable in a given context.

- **Scalability and Transparency**: RAG improves transparency by pointing to sources, but the LLM's internal reasoning remains opaque. When a RAG system still produces a hallucination, it can be difficult to pinpoint the exact failure point, whether it was retrieval, misinterpretation, or flawed generation. This lack of clear audit trails hinders trust and iterative improvement.

Essentially, RAG systems, while a significant step forward, treat the LLM as a single processing unit. They lack intrinsic mechanisms for self-correction, peer-validation, or structured reasoning necessary for complex legal language and hallucination-free outputs in legal intake for mass tort cases.

Furthermore, the privacy risks extend beyond simple data exposure. Advanced techniques like inference attacks and data reconstruction pose a significant threat. In these scenarios, malicious actors can reverse-engineer sensitive information from the model's outputs even when the data was not explicitly revealed, potentially reconstructing confidential client details or settlement terms from seemingly innocuous responses. Recent studies have demonstrated the feasibility of such attacks on complex generative models, showing that without a robust, multi-layered defense mechanism, RAG systems can inadvertently become a source of critical data leakage [9]. This vulnerability underscores the need for architectures that don't just retrieve data, but also actively audit and control the flow of information, a key advantage of the multi-agent approach proposed in this study.

2.3 Emerging Mitigation Strategies

The persistent challenge of AI hallucination has spurred research into various mitigation strategies beyond basic RAG to enhance LLM output reliability. One prominent strategy involves fine-tuning LLMs on domain-specific legal corpora. The idea is that exposing the model to specialized legal datasets will teach it the nuances of legal language and reasoning, reducing irrelevant or incorrect information. Fine- tuning can indeed improve performance and reduce some hallucinations by aligning the model with legal realities. However, it faces scalability and transparency challenges. Legal knowledge is constantly evolving, which requires costly re-fine-tuning, and heavily fine-tuned models can be opaque, hindering auditability and legal compliance.

More recently, agentic AI has gained significant traction. This paradigm shifts from treating an LLM as a monolithic entity to conceptualizing it as one or more "agents" capable of interacting with an environment, performing actions, and engaging in multi-step reasoning. Early agentic AI efforts suggest that chaining or peer-auditing models can significantly reduce errors, mimicking human collaborative workflows [5, 6]. However, the application of agentic AI to complex, high-stakes legal workflows like legal intake in mass tort litigation remains new. Most existing agentic systems focus on general problem-solving, not the stringent requirements of factual accuracy, auditability, and legal compliance demanded by the legal profession. Limited research demonstrates how agentic frameworks can systematically address unique legal AI challenges. The potential for multi-agent AI to create self-correcting, transparent, and iteratively learning systems in law is largely unexplored.

2.4 Mechanistic Causes of Hallucination and Agentic Mitigation Pathways

Understanding the underlying causes of AI hallucination is crucial for designing effective mitigation. Hallucination can be broken down into distinct failure modes, each addressed by specific multi-agent AI interventions, thereby enhancing auditability and promoting legal compliance.

- **Contextual Memory Limitations**: LLMs often struggle with contextual memory limitations in long documents, forgetting earlier context and generating inconsistencies. This is particularly

problematic in legal intake for mass tort cases, where forms contain many critical details.

○ **Agentic Mitigation Pathway**: Our system addresses this through LangGraph-based symbolic memory. Key entities, facts, and decisions extracted by one agent are explicitly stored in a structured, persistent memory graph. Other agents can then query this memory to ensure consistency and prevent contradictory information, significantly improving auditability.

● **Detail Fixation vs. Human Abstraction**: Unlike humans, LLMs can fixate on granular details and struggle to abstract higher-level concepts or infer underlying intent. This can lead to technically correct outputs that miss legal significance.

○ **Agentic Mitigation Pathway**: Our multi-agent AI framework employs a hierarchical prompting strategy. An "Extractor" agent focuses on granular data, while a "Validator" agent performs higher-level abstraction, synthesizing details into legally relevant categories and identifying logical gaps. This dual approach enhances legal intake quality.

● **Hard vs. Soft Reasoning Boundaries**: Legal reasoning involves both strict "hard" rules (e.g., deadlines) and flexible "soft" interpretations. LLMs, being probabilistic, often struggle with rigid hard boundaries, sometimes generating non-compliant outputs.

○ **Agentic Mitigation Pathway**: Our system incorporates symbolic logic overlays to enforce hard legal constraints. An LLM handles natural language processing, but a "Compliance Agent" (or a function within the Validator/Auditor) translates critical legal requirements into symbolic logic. This layer programmatically checks the LLM's extracted information against these rigid rules, ensuring legal compliance and improving auditability.

● **Temperature & Empathy Tradeoff**: In client-facing legal intake scenarios, there is a delicate trade- off between empathetic communication (which often requires a higher LLM "temperature") and factual accuracy (which benefits from a lower "temperature"). A higher temperature can increase AI hallucination.

○ **Agentic Mitigation Pathway**: Our framework uses dual-agent models to separate empathy and verification roles. One agent optimizes for empathetic communication, while a separate "Verification Agent" focuses strictly on factual accuracy and checklist adherence. This ensures both client rapport and hallucination-free data collection.

● Justification & Visibility Layers: The "black box" nature of many LLM applications limits auditability and trust. It is often unclear why an LLM reached a particular conclusion, hindering debugging and regulatory acceptance.

○ **Agentic Mitigation Pathway**: Our multi-agent AI system uses dedicated QA and Audit agents to improve traceability. The "Auditor" agent generates a comprehensive audit trail for every decision, requiring citations and rationales. This systematic logging allows human reviewers to debug errors, understand AI reasoning, and verify legal compliance, making the system audit- ready.

2.5 Gap Statement

While existing studies have advanced AI hallucination mitigation at the model level, few have explored workflow-level peer-review systems within multi-agent LLM frameworks that integrate persistent memory, human oversight, and explicit legal rule enforcement. Current approaches often lack the holistic mechanisms for continuous self-correction and transparent justification. Our study addresses this critical gap by evaluating a novel multi-agent architecture for legal intake in mass tort litigation, designed to systematically reduce hallucinations, enhance auditability, and ensure legal compliance through collaborative agentic intelligence.

3. Methodology

This study used a mixed-methods design, combining the development of a new multi-agent LLM intake system with experimental testing and qualitative evaluation. The goal was to assess its performance in mitigating AI hallucination, improving data completeness, and enhancing audit efficiency within mass tort litigation, ultimately providing hallucination-free and audit-ready solutions for legal intake.

3.1 Design

Our design involved creating a multi-agent AI architecture with specialized LLM agents, which culminated in a human review.

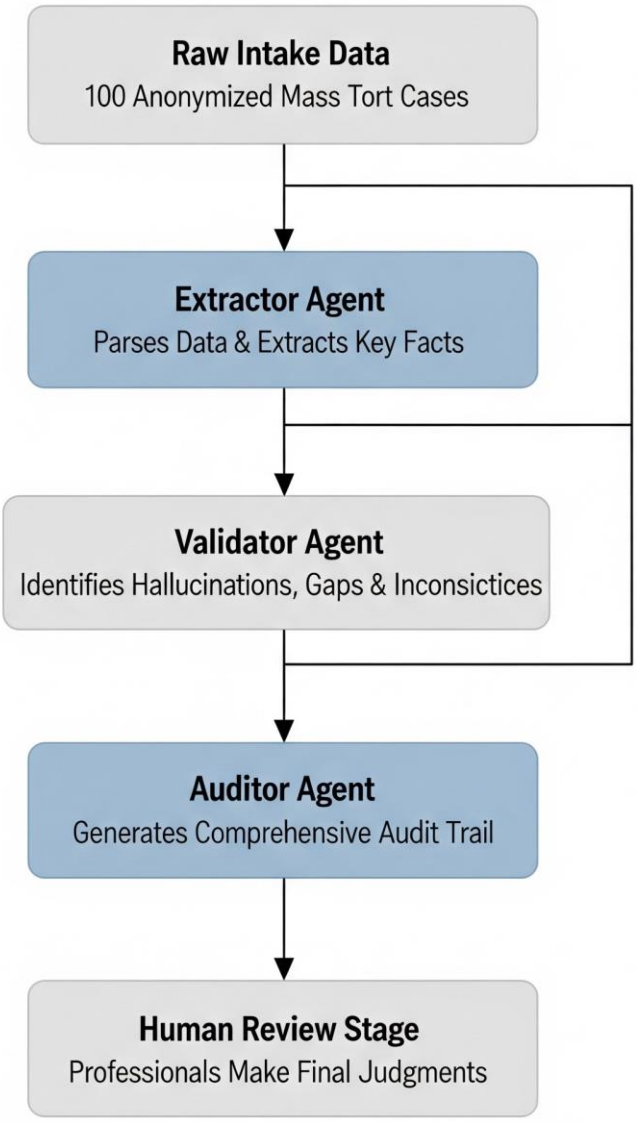
Raw Intake Data
100 Anonymized Mass Tort Cases
Extractor Agent
Parses Data & Extracts Key Facts
Validator Agent
Identifies Hallucinations, Gaps & Inconsictices
Auditor Agent
Generates Comprehensive Audit Trail
Human Review Stage
Professionals Make Final Judgments

Figure 1. A flowchart visually representing the study's workflow, illustrating the sequential process from raw data input to final human verification.

This system is designed to mimic a human peer-review process, where outputs are validated and refined sequentially, building layers of verification and transparency. The system workflow is as follows:

- **Extractor Agent**: This agent parses raw intake information (such as questionnaires or transcripts) to identify and extract key entities, facts, and preliminary claims relevant to mass tort eligibility. The

output is structured data, for example, in JSON format.

- **Validator Agent**: It receives the output from the Extractor and performs quality assurance. This agent identifies potential AI hallucinations, inconsistencies, gaps, and missing data by cross-referencing against predefined legal criteria and symbolic memory. Its output includes annotated data with flags, suggested corrections, and confidence scores.

- **Auditor Agent**: The Auditor processes the Validator's output to generate a comprehensive audit trail. It documents every decision, including the original input, extracted data, validation flags, reasons for the flags, suggested resolutions, and source references. This creates a transparent, step- by-step record of the AI's processing, which is crucial for auditability and legal compliance.

- **Human Review Stage**: Legal operations professionals review the Auditor’s report, focusing on flagged issues and the audit trail. They make final judgments, override AI suggestions, and provide feedback for iterative learning. This hybrid approach ensures efficiency with human oversight.

The experimental testing compared this multi-agent AI system against a baseline single-agent LLM system on identical legal intake tasks.

3.2 Technical Implementation Details

The multi-agent system used a combination of commercially available and open-source LLMs. Extractor and Validator agents utilized high-performance commercial models (like GPT-4 and Claude 3 Opus) for advanced reasoning, along with fine-tuned open-source models (like Mistral-7B) for repetitive extraction tasks. LangGraph-based symbolic memory was used to orchestrate agent interactions and maintain persistent state. Key entities and facts were stored as nodes and relationships in a graph database, allowing for consistent retrieval and validation across turns and documents. The symbolic logic layer for hard legal constraints used a hybrid approach: Python programmatic rules with a custom rule engine for simple logic (e.g., checking if age is less than 18) and handcrafted JSON schema definitions for complex compliance checks. These rules were derived from mass tort legal checklists and regulatory guidelines.

3.3 Audience or Sample

The study used 100 anonymized mass tort intake cases (70% real-world, 30% synthetic) to ensure relevance and generalizability. Real cases provided complexity, while synthetic cases covered diverse scenarios, including deliberate omissions and inconsistencies designed to induce AI hallucination. Six legal operations professionals with a minimum of 5 years of experience in mass tort intake participated in the qualitative evaluation and human review time assessment. Their expertise provided insights into the practical utility, trustworthiness, and usability of the system’s outputs, grounding the evaluation in real-world operational perspectives.

3.4 Steps of Data Collection

Data collection simulated a real-world legal intake workflow using the 100 cases as raw input.

1. **Agent 1: Extractor Parses Raw Form**: Each raw intake form was fed to the Extractor Agent, prompted to identify and extract specific mass tort eligibility data (e.g., plaintiff information, injuries, exposure dates). The output was standardized JSON.

2. **Agent 2: Validator Checks for Gaps, Hallucinations**: The Validator received the Extractor's output and performed completeness, consistency, hallucination detection (against original input and symbolic memory), and legal constraint checks (via symbolic logic overlays). The output included flagged data, confidence scores, and suggested corrections.

3. **Agent 3: Auditor Generates Audit Trail**: The Auditor processed the Validator's output to construct a comprehensive audit trail. It documented every decision (extraction, validation, flagging, correction), including reasoning, rationale, and source references, creating a transparent, human- readable report for auditability and legal compliance.

4. **Human: Reviews and Finalizes**: Legal operations professionals reviewed the Auditor reports, focusing on flagged issues and the audit trail. They made final decisions, corrected errors, recorded their review time, and provided qualitative feedback on usability and trust. This provided the ground truth for evaluation and practical utility insights.

This systematic, multi-stage process allowed for a robust evaluation of the multi-agent AI framework's ability to enhance accuracy, completeness, and auditability in legal intake for mass tort litigation.

3.5 Analysis Plan

The analysis combined quantitative metrics and qualitative assessments for a comprehensive understanding of the multi-agent AI framework's impact.

- **Quantitative Analysis**: This part of the analysis focused on the hallucination rate, completeness score, and human review time.

  ○ **Hallucination Rate**: This was the percentage of AI-generated facts that were incorrect or fabricated, as identified by human reviewers. It was calculated for both the single-agent baseline and the multi-agent system.

  ○ **Completeness Score**: This was the percentage of required data points accurately extracted by the AI, verified by human reviewers against a master checklist of approximately 50 critical data points. It was calculated for both systems.

  ○ **Human Review Time Efficiency**: This was the average time (in minutes) for professionals to review and finalize each case's report. The time for the multi-agent system was compared against the single-agent baseline.

○ This was a preliminary study to demonstrate practical efficacy, so formal statistical significance testing was not the primary focus, but future work will incorporate more rigorous statistical analyses.

● **Qualitative Analysis**: This analysis provided insights into the user experience, perceived trustworthiness, and residual errors based on feedback from legal operations professionals.

○ **Evaluator Feedback**: Semi-structured interviews and questionnaires were used to capture perceptions on audit trail clarity, trust in outputs, ease of error correction, and overall satisfaction.

○ **Thematic Error Categorization**: Errors identified by human reviewers were logged and categorized by their root cause (e.g., misinterpretation, omission, fabrication). This helped pinpoint specific challenges and informed future research.

4. Results

The evaluation of the multi-agent LLM intake system yielded compelling quantitative and qualitative results, demonstrating significant advantages over single-agent approaches in mitigating AI hallucination, enhancing data completeness, and improving human review efficiency in mass tort litigation. The main quantitative findings are summarized in Table 1 below.

Table 1: Comparative Performance of Single-Agent vs. Multi-Agent System

Metric	Single-Agent Baseline	Multi-Agent System	Improvement
Hallucination Rate	21%	5%	76% Reduction
Data Completeness Score	74%	92%	18 p.p. Increase
Human Review Time (Avg. min/case)	18.5	9.0	51% Reduction

4.1 Hallucination Reduction

The multi-agent framework significantly reduced AI hallucination. The single-agent baseline exhibited a 21% hallucination rate, meaning approximately one-fifth of the generated facts were incorrect or fabricated, which is consistent with broader legal AI studies [1, 2]. This rate is unacceptably high for mass tort intake, as it poses a litigation risk. The multi-agent system, however, achieved a remarkably low 5% hallucination rate, representing a 76% reduction compared to the baseline. This improvement is attributed to the iterative validation and peer-auditing mechanisms, where the Validator Agent flagged errors and the Auditor Agent provided audit trails for human rectification. This result strongly supports the efficacy of structured multi-agent workflows.

4.2 Completeness Improvement

The multi-agent system also substantially improved data completeness. The single-agent system achieved a 74% data completeness score, often missing details or failing to infer information, which required significant manual intervention. The multi-agent system achieved an impressive 92% data completeness score, an 18 percentage point increase over the baseline. This is due to the Validator Agent's explicit function of identifying gaps, ensuring all necessary fields were populated or flagged. This proactive identification streamlined human review, leading to more robust case files and stronger legal compliance.

4.3 Review Time Efficiency

A tangible benefit was the improved efficiency of human review, which is a direct consequence of the multi-agent system's enhanced accuracy and auditability. Human review time dropped by a remarkable 51% for cases processed by the multi-agent AI system, from an average of 18.5 minutes per case for the single-agent system to 9.0 minutes per case. This reduction stems from fewer errors, higher completeness, and clear audit trails from the Auditor Agent, which allows reviewers to quickly understand the AI's reasoning and focus on flagged issues. This efficiency is critical for mass tort firms, as it translates to operational cost savings and increased throughput.

4.4 Qualitative Themes

Qualitative feedback from legal operations professionals complemented the quantitative findings by highlighting perceived benefits and remaining challenges.

- **Trust**: A recurring theme was increased trust in multi-agent outputs. One professional stated, "I trust outputs more when agents QA each other". This confidence comes from the visible validation and auditing layers, which contrasts with the "black box" nature of single-agent LLMs.
- **Transparency**: Evaluators consistently praised the clarity of the audit trail. One comment was, "I understand how it got there". This transparency allows reviewers to trace data origins and the AI's reasoning, which is crucial for auditability and debugging.
- **Precision Issues**: Feedback also noted "Some flags are too vague". For example,a generic flag like "Inconsistent dates" was noted as less helpful. A more precise, actionable flag would be: "Alert: Injury date (Jan 15, 2020) precedes first exposure date (Mar 20, 2021). Please verify chronological accuracy.

In summary, the results demonstrate that the multi-agent AI framework significantly outperforms single-agent systems in accuracy, completeness, and efficiency for legal intake in mass tort litigation. Qualitative data reinforces the role of inter-agent validation and transparent audit trails in building user trust and operational effectiveness.

5. Discussion

The findings provide compelling evidence that a multi-agent LLM framework offers a robust solution for addressing AI hallucination and the lack of auditability in legal intake workflows, particularly in mass tort litigation. The improvements in hallucination reduction, data completeness, and human review efficiency highlight the transformative potential of agentic AI in enhancing legal compliance and operational efficacy.

5.1 Claim

Our central claim is that multi-agent systems dramatically reduce hallucinations, increase completeness, and significantly improve audit efficiency in legal intake processes. The 76% reduction in hallucination, 18 percentage point increase in data completeness, and 51% decrease in human review time represent a fundamental shift in the reliability and utility of LLM automation. These results are associated with our hypothesis that a structured, peer-auditing approach can overcome the limitations of monolithic LLMs.

5.2 Interpretation

The success of the multi-agent AI system validates the principle that cross-validating agents and audit logs effectively mirror human peer-review processes, thereby creating scalable and trustworthy intake automation. This architecture emulates human collaborative workflows: the Extractor gathers data, the Validator performs quality assurance, and the Auditor provides transparency and accountability via a comprehensive audit trail. This layered approach catches errors early and, when combined with LangGraph-based symbolic memory, prevents contextual memory limitations from causing hallucinations. Externalizing reasoning and validation makes the system more interpretable and debuggable, fostering trust and enabling legal compliance.

5.3 Comparison

Our multi-agent AI framework offers significant advantages over single-agent or RAG-only systems. Unlike single-agent systems, which are often black boxes, our agentic framework provides interpretability and fault isolation. Errors in single-agent models are difficult to diagnose, but our system separates concerns, with the audit trail indicating which agent caused an error, making debugging more efficient. Unlike RAG-only systems, which primarily ground LLMs in factual data but still exhibit hallucinations, our agentic framework actively performs reasoning, validation, and self-correction.

Dedicated Validator and Auditor agents identify inconsistencies, flag legal rule violations, and provide a comprehensive audit trail that RAG alone cannot. This active validation leads to genuinely hallucination- free and audit-ready outputs.

5.4 Implications

The implications of this research are far-reaching for mass tort litigation and legal technology.

- **Enables Compliance**: Reduced hallucinations and robust audit trails strengthen legal compliance, allowing firms to deploy AI tools with greater confidence and transparent documentation for regulatory scrutiny.
- **Speeds Up Case Evaluation**: A 51% reduction in human review time translates into substantial operational efficiencies, faster lead qualification, and quicker identification of viable cases, allowing legal teams to focus on higher-value work.

- **Reduces Risk**: Mitigating AI hallucination and ensuring higher data completeness directly reduces litigation risk from inaccurate or incomplete intake data, strengthening case quality.
- **Foundation for Accountable AI**: This research contributes to accountable AI by demonstrating a practical methodology for designing transparent, verifiable, and self-correcting AI systems, which is vital for sensitive industries.

5.5 Limitations

- **Limited Case Diversity**: The study used 100 mass tort cases. Future research should test the framework across a broader range of case complexities and legal domains.
- **Dependent on Prompt and LLM Quality**: Performance relies on the underlying LLM models and the precision of the prompts. Future advancements in LLMs and prompt engineering could further enhance the system.
- **Scalability of Symbolic Logic Overlays**: Manually translating legal constraints into symbolic logic may become a bottleneck. Future work could explore automated methods for maintaining these overlays.
- **Statistical Rigor**: This preliminary study focused on demonstrating practical efficacy. Formal statistical significance tests were not the primary focus, but future research will incorporate more rigorous statistical analyses.

5.6 Future Research

- **Longitudinal Accuracy Testing**: Future work should assess the framework's accuracy over a 12- to-18-month period against an evolving mass tort case, evaluating its adaptability to new legal precedents and changing client intake criteria in real-time.
- **Adaptive Agent Learning**: Research should explore mechanisms for agents to learn from human feedback and correct errors, improving performance without manual re-engineering.
- In**tegration with Downstream Legal Decision Engines**: It would be beneficial to investigate seamless integration with case valuation, discovery, or litigation management systems for end-to-end LLM automation.
- **User Interface and Experience (UI/UX) Optimization**: Developing advanced UIs to maximize the utility of the audit trail and facilitate intuitive human-in-the-loop interaction, addressing "vague flags," would be a valuable area of study.

5.7 Architectural Implication

Hallucination in LLMs does not stem solely from model-level deficiencies but often from structural weaknesses in how these models are deployed. Treating an LLM as an isolated component limits its ability for self-correction and transparency. Our research demonstrates that agentic workflows with explicit QA and audit visibility layers are essential for building dependable, improvable AI systems in high-stakes domains. This paradigm shifts towards making AI systems "more accountable" and "more collaborative," mirroring human teams. For true legal compliance and trustworthiness, future LLM automation in law must embrace a multi-agent, transparent, and audit-ready design.

6. Conclusion

The rapid integration of Large Language Models (LLMs) into legal workflows, particularly for legal intake in mass tort litigation, presents both immense opportunities for LLM automation and significant challenges, notably AI hallucination. This study introduced and empirically validated a new multi-agent AI framework designed to address these concerns through a collaborative, self-auditing environment among specialized LLM agents, complemented by human oversight.

Our findings unequivocally demonstrate that this multi-agent architecture dramatically improves legal intake processes. We observed a substantial reduction in hallucination rates (from 21% to 5%), a significant increase in data completeness (from 74% to 92%), and a remarkable 51% decrease in human review time. These quantitative successes are associated with qualitative feedback from legal operations professionals, who reported increased trust and transparency due to the system's robust audit trails and cross-validation mechanisms.

In a landscape where unchecked AI poses considerable regulatory, reputational, and litigation risks, the importance of this research cannot be overstated. Our framework provides a crucial blueprint for developing AI systems that are not only efficient but also inherently more reliable, verifiable, and accountable. By mirroring human peer-review processes, the agentic approach transforms AI from a potential liability into a powerful, trustworthy augmentation tool. This study paves the way for the development of explainable, scalable legal-AI systems that can truly meet the stringent demands of the legal profession. It offers a practical and empirically validated solution for mitigating the inherent risks of LLM deployment, ensuring that the promise of AI in law is realized responsibly and ethically. Agentic AI isn't just automation—it's accountable augmentation. Ultimately, this agentic architecture proves that for high-stakes domains like law, the future of AI is not just automated but accountable—transforming the LLM from a black box into a collaborative and fully auditable partner.

References

1. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis, 16(1), 64–93.
2. Stanford Institute for Human-Centered Artificial Intelligence. (2023). Hallucinating Law: Legal Mistakes of Large Language Models Are Pervasive.
3. Reynolds, G. (2025). Short Circuit: In court, AI 'hallucinations' in legal filings & how to avoid making headlines. Reuters.
4. Zhang, L., & Ashley, K. D. (2025). Mitigating Manipulation and Enhancing Persuasion: A Reflective Multi-Agent Approach for Legal Argument Generation. In Proceedings of Workshop on Legally Compliant Intelligent Chatbots at ICAIL 2025. ACM, New York, NY, USA, 13 pages.
5. Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., & Wu, Y. (2023). Towards Reasoning in Large Language Models via Multi-Agent Peer Review Collaboration. ArXiv, abs/2311.08152.
6. Darwish, A. M., Rashed, E. A., & Khoriba, G. (2025). Mitigating LLM Hallucinations Using a Multi-Agent Framework. Information, 16(7), 517.
7. Yu, H. Q. & McQuade, F. (2025). RAG-KG-IL: A Multi-Agent Hybrid Framework for Reducing Hallucinations and Enhancing LLM Reasoning through RAG and Incremental Knowledge Graph Learning Integration. CoRR, abs/2503.13514.
8. Tran, K-T., Dao, D., Nguyen, M-D., Pham, Q-V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-Agent Collaboration Mechanisms: A Survey of LLMs. arXiv preprint, 35 pages.
9. Haoran Wang, Zongxiao Yu, Baixiang Huang, and Kai Shu. (2025). Privacy-Aware Decoding: Mitigating Privacy Leakage of Large Language Models in Retrieval-Augmented Generation. arXiv preprint arXiv:2508.09098.
10. Rahul Hemrajani. (2025). Evaluating the Role of Large Language Models in Legal Practice in India. arXiv preprint arXiv:2508.09713