Incorporating Legal Structure in Retrieval-Augmented Generation: A Case Study on Copyright Fair Use

Justin Ho[1,*], Alexandra Colby[2] and William Fisher[2]

[1]Harvard Business School, Boston MA, 02163, United States of America
[2]Harvard Law School, Boston MA, 02138, United States of America

Abstract

This paper presents a domain-specific implementation of Retrieval-Augmented Generation (RAG) tailored to the Fair Use Doctrine in U.S. copyright law. Motivated by the increasing prevalence of DMCA takedowns and the lack of accessible legal support for content creators, we propose a structured approach that combines semantic search with legal knowledge graphs and court citation networks to improve retrieval quality and reasoning reliability. Our prototype models legal precedents at the statutory factor level (e.g., purpose, nature, amount, market effect) and incorporates citation-weighted graph representations to prioritize doctrinally authoritative sources. We use Chain-of-Thought reasoning and interleaved retrieval steps to better emulate legal reasoning. Preliminary testing suggests this method improves doctrinal relevance in the retrieval process, laying groundwork for future evaluation and deployment of LLM-based legal assistance tools.

Keywords

Retrieval-Augmented Generation, Legal Knowledge Graphs, Legal Citation Networks, Fair Use Doctrine, Legal AI

1. Introduction

The Digital Millennium Copyright Act (DMCA) provides platforms such as YouTube with safe harbor protection from copyright infringement claims related to content uploaded by their users, as long as they offer copyright holders the ability to take down content that infringes their rights [1].

In theory, an uploader whose content was removed under the DMCA can submit a counter-notice to challenge the takedown, with the most commonly used defense being the Fair Use Doctrine under 17 U.S. Code § 107 [2]. This doctrine considers four factors: the purpose and character of the use (e.g., whether it is transformative, commercial, and if the work serves a different purpose to the original), the nature of the copyrighted work, the amount and substantiality of the portion used, and the effect of the use on the market for the original or its derivatives. It is designed to permit use of copyrighted material without permission for purposes such as criticism, comment, news reporting, teaching, scholarship, or research.

However, in practice, DMCA takedowns are often abused to suppress valid criticism protected under free speech, or are issued through automated systems that generate invalid and duplicative claims. This results in chilling effects and self-censorship, especially among creators without legal representation, who may be ill-equipped to assess whether they have a colorable fair use defense. Although the courts held in the *Lenz v. Universal Music Corp.*, 801 F.3d 1126 (9th Cir. 2015), decision that copyright holders must consider fair use in good faith before issuing a takedown notice, enforcement of this standard is weak. Users must prove the copyright holder acted in bad faith, a subjective mental state that is difficult to determine, rendering the safeguard largely ineffective in practice [3, 4].

1.1. LLMs in Legal Assistance

With the advancement of Large Language Models (LLMs), particularly in the legal domain, there is growing potential for these technologies to offer legal assistance to content creators who might otherwise lack representation to assert fair use claims [5, 6, 7]. While LLMs can automate tasks such as annotation, issue-spotting, interpretation of short legal texts, and even generating legally plausible conclusions, they still fall short in areas that require precise rule recall, multi-step reasoning, and the explanation of legal inferences [8, 9]. Even Retrieval-Augmented Generation (RAG) models from major legal research platforms are prone to hallucinations, including fabricating case law and misinterpreting precedents [10, 11].

Following the typology of RAG-based hallucinations proposed in [11], persistent issues arise from a combination of naive retrieval, inapplicable authority, sycophancy (i.e., the tendency to agree with a given text even when it is inaccurate), and reasoning errors. We hypothesize that local domain improvements—specifically, building expertise within a narrow subfield of legal doctrine—can improve the deployment performance of LLMs in certain legal contexts. Such narrowly focused local subfield experts can potentially be combined to provide more general automated legal assistance. Currently, we focus on the Fair Use Doctrine in copyright law as a case study. This is conceptually similar to Mixture of Experts (MoE) models [12]. However, our focus is on improving the non-parametric memory component of RAG by combining knowledge graphs and granular retrieval strategies in the Fair Use Doctrine [13, 14, 15, 16].

1.2. Local Expertise and Structured Reasoning

Problems like naive retrieval and inapplicable authority may have stemmed from the general-purpose Question-Answering (QA) design of AI models deployed by major legal research platforms, but due to their proprietary nature, it is difficult to verify the true source of the problems [11]. Legal concepts may appear semantically similar in common usage, yet differ significantly in terms of legal doctrine (e.g., the distinction between 'moral turpitude' and the 'moral-wrong doctrine' in Criminal Law, or the differing meanings of 'neg-

Proceedings of the Seventh Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2025), June 16, 2025, Chicago, USA.
*Corresponding author.
jho@g.harvard.edu (J. Ho); acolby@jd26.law.harvard.edu (A. Colby); tfisher@law.harvard.edu (W. Fisher)
https://justinhjy1004.github.io/ (J. Ho); https://hls.harvard.edu/faculty/william-w-fisher/ (W. Fisher)
0009-0005-0751-9504 (J. Ho)

ligence' and 'reasonable person' across various areas of law). Although our focus on the Fair Use Doctrine offers some topical constraint, we show that retrieval can be improved by incorporating legal information as a citation-weighted knowledge graph. This graph encodes court hierarchy, citation relationships, and the statutory factors specific to Fair Use. As a result, the retrieval process prioritizes documents that are not only semantically relevant but also doctrinally authoritative.

We also include methods used by "reasoning models," such as Chain-of-Thought (CoT), to improve multi-step reasoning in legal cases since CoT has been shown to reduce reasoning errors in LLMs [17]. This is especially important for decisions under the Fair Use Doctrine, which is a multi-factor test requiring contextual considerations. Additionally, we implement a one-step Interleaving Retrieval CoT, where the LLM first analyzes how a complaint or case relates to the four fair use factors, which then guides the retrieval process [18]. This may reduce sycophancy by anchoring the model's reasoning in the structure of the doctrine itself. However, the issue of sycophancy is perhaps better addressed during the information elicitation stage.

We developed a functioning prototype to demonstrate the core features of our system, which is available at https://fairuselegalbot-main.streamlit.app/. The source code of the prototype, construction of the knowledge graph, and the results of the preliminary analysis is publicly available on GitHub: https://github.com/justinhjy1004/FairUseLegalBot.

2. Literature Review and Related Work

Our work integrates ideas from Retrieval-Augmented Generation (RAG), knowledge graphs, and information retrieval and representation, adapting them to the legal domain.

2.1. Similarity is Not All You Need

A widely adopted strategy to mitigate hallucinations in language models is grounding them through Retrieval-Augmented Generation (RAG). RAG retrieves external documents based on vector similarity to the user's query, typically measured using cosine similarity [19].

In legal applications, the retrieved documents often include case law, legal opinions, statutes, and regulatory codes. The core assumption is that retrieving semantically similar documents will produce more factually accurate and contextually relevant outputs by anchoring them in authoritative sources. However, the quality of the model's output is only as strong as the documents retrieved. In practice, LLMs deployed within legal research tools still exhibit hallucinations—especially when the retrieval corpus is noisy, outdated, or lacks contextual metadata, such as indicators that a precedent has been overruled [10].

Recent research in information retrieval and domain-specific indexing has aimed to improve the reliability of RAG-based systems. Still, the notion of similarity remains highly nuanced and context-dependent [15]. For instance, "Dracula" might refer to either the character or the novel, and a summarization task could be misled by retrieving content about Nosferatu, an unauthorized adaptation, despite the surface-level similarity[1].

Additionally, the granularity of the retrieval unit, whether at the document, sentence, or sub-sentence level, is important in downstream retrieval performance. Often, only a small portion of a document is relevant to the query, and this is particularly true in the context of legal reasoning and analysis [16, 20]. To address this, we structure our underlying data store at the level of statutory factors, allowing retrieval to operate at a finer granularity aligned with the specific analysis of the Fair Use Doctrine.

2.2. Incorporating Legal Structure

Legal reasoning relies on more than surface-level textual similarity. Documents in the legal domain carry contextual structure that flat document representation, as commonly used in standard RAG implementations, often fail to capture. Legal opinions are authored by courts of differing authority, and in common law systems, precedents shape legal interpretation. Determining which precedents are most relevant requires understanding the legal hierarchy, interpretive weight, and the frequency and influence of citation. Representing this information as a knowledge graph, where relationships between cases are explicitly modeled, can improve retrieval quality [21, 22].

Our approach builds on this idea by encoding legal structure directly by modeling court hierarchies, citation flows, and the interpretive weight of specific paragraphs with respect to statutory factors under consideration. This improves both the doctrinal relevance of retrieved material and the accuracy of subsequent inference tasks.

Prior work in U.S. and EU legal systems demonstrates the value of citation networks, particularly when nodes represent cited paragraphs rather than entire opinions. Prior work shows that paragraph-level modeling captures the 'grammar of repetition' in judicial reasoning. This illustrates how interpretive principles gain authority through repeated citation. Such granularity also enables detection of indirect influence chains and improves our understanding of how legal doctrines evolve [20].

In this spirit, we structure our dataset around statutory factor-level modeling. By explicitly annotating legal opinions according to the statutory factors from the Fair Use Doctrine. This enables context-sensitive retrieval that has the potential of improving performance. For instance, two copyright disputes involving unauthorized film use might seem similar, but diverge sharply depending on whether the use is non-expressive or a parody of the original material. This distinction is important in fair use analysis, and modeling the data in this granularity helps reflect and align how courts often extract legal principles from specific parts of a ruling rather than relying on the full opinion [20].

We also incorporate the citation network structure of legal precedents by modeling court hierarchies and citation relationships to include not only the semantic similarity of the dispute at hand, but also the doctrinal relevance and importance. We apply PageRank [23], commonly used in citation analysis, as a way to incorporate the doctrinal relevance in our retrieval and ranking process. While basic degree centrality offers insight, legal reasoning can drift or 'un-anchor' from original sources over time [20]. PageRank,

[1] The original Nosferatu (1922) was found by German courts to infringe the Stoker Estate's copyright. A German judge ordered all copies of Nosferatu to be destroyed and it survives today because of a single copy that found its way to the United States.

by contrast, accounts for the authority of citing sources—i.e., a citation from a widely cited opinion carries more doctrinal weight than one from a marginal case [23]. This allows our system to better reflect the practical significance of legal authority in ranking the retrieved judicial opinions.

3. Methods

This section provides the details regarding the implementation of the automated analysis of Fair Use cases, particularly in data representation and the retrieval process.

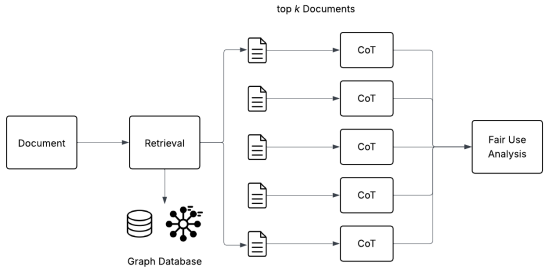


Figure 1: Overview of the Automated Analysis of Fair Use Cases.

Figure 1 shows an overview of the methods used. The system starts by retrieving the top-k most relevant legal cases using a graph-based database that takes into account not just text similarity, but also court authority and how often cases are cited. Each case is then analyzed step by step using Chain-of-Thought reasoning to assess how it applies to the four Fair Use factors. Finally, these analyses are combined to generate a structured Fair Use evaluation based on the user's document.

The Large Language Model used is Google's Gemini Flash 2.0 [24], and the embedding model for semantic-based vector search is Google's Gecko [25].

3.1. Data Corpus

Using WestLaw Precision's fact pattern search, we located all legal precedents relevant to the Fair Use Doctrine in copyright law. We then sourced the legal corpus relevant to the Fair Use Doctrine in copyright from Court Listener [26] and Hein Online. Furthermore, we used EyeCite, an Open Source Software developed to identify case law citations in documents to construct our citation network [27].

Table 1
Corpus Overview

Total Number of Cases	209
Total Number of Opinions	283
Time Range Coverage	1976-2025
Number of Unique Courts	51

The number of opinions exceeds the number of cases because a single case may generate multiple judicial opinions, including appellate decisions, as well as concurring or dissenting views authored by individual judges.

Additionally, we sourced complaints related to Copyright infringement that were not resolved in court from Public Access to Court Electronic Records (PACER) [28]. These serve as a preliminary test dataset for our working prototype, as they represent real Fair Use disputes that were unresolved, and hence provide a way to measure how well our model performs in unresolved cases. We sourced a total of 20 cases.

3.2. Data Representation

Our knowledge graph is implemented using Neo4j [29]. In order to more faithfully represent the data in legal precedents related to the Fair Use Doctrine in copyright, we modeled the cases with a knowledge graph, and the schema of the graph database is shown in Figure 2. A schema defines the structure of the graph—it specifies what types of entities (nodes) exist, such as *Case*, *Court*, *Opinion*, and Fair Use Factor (e.g., *Purpose*, *Market*), and how they are connected through relationships like *CITED*, *DECIDED_IN*, or *HAS_OPINION*. For example, a *Case* node from the Supreme Court is connected to a *Court* node labeled "SCOTUS" and is cited by multiple lower-court *Opinion* nodes. This incorporates important features of the legal system in the United States where legal precedents issued by a higher court have higher authority over others, as well as the citation network formed between the cases.

Furthermore, in every given opinion, we extract the verbatim paragraphs related to the *Facts* of the case, the four factors: (1) *Purpose* and character of the use (2) *Nature* of the copyrighted work, (3) *Amount* and substantiality of the portion used, and (4) Effect of the use on the potential *Market*. We also included the *Conclusion* of the opinion to reflect how the court balanced the four factors to arrive at their opinion. The extraction is done using the LLM to identify and extract verbatim paragraphs in which each of the four Fair Use factors were discussed, along with the factual background and the court's conclusion. We designed specific prompts to direct the LLM to focus on legal reasoning and factor-specific content. These prompts instruct the model to return direct quotations from the opinion text corresponding to each factor, rather than paraphrased summaries.

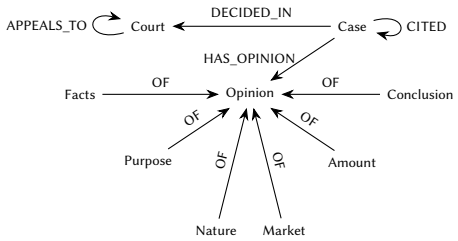


Figure 2: Domain Specific Knowledge Graph Schema of Judicial Opinions of the Fair Use Doctrine in Copyright 17 U.S. Code § 107.

The choice of our data representation is that it is not only a more faithful representation of the data, which allows us to use the structure of the data (i.e. citations) to improve our retrieval process, but it also provides the ability to retrieve based on contextual similarity. For instance, a complaint on copyright infringement might be similar with respect to the medium in which the work was distributed (print or via video recordings), but might differ substantially based on the purpose of the use (ie. parody, criticism, or for educational reasons).

The choice of using a knowledge graph representation allows for more granular, context-specific similarity comparison during the retrieval process which has shown to be

effective [16, 15]. Moreover, the interpretability of such representation might increase the interest, trust, and therefore adoption of LLMs in the legal space since this mimics how legal experts might reason in the context of Fair Use [30].

Each case is modeled as a node connected to its issuing court and to the legal opinion(s) it contains. Opinions are linked to factor-specific paragraph nodes (e.g., Purpose, Market, Nature), enabling granular retrieval by legal reasoning dimensions. Citations between cases form directed edges within the graph. The schema is implemented in Neo4j using labeled nodes (e.g., `Case`, `Court`, `Opinion`, `Fact`) and relationship types (e.g., DECIDED_IN, HAS_OPINION, CITED, APPEALS_TO). This representation supports both structural queries (e.g., retrieving appellate court opinions) and vector search via LLM embeddings.

3.3. Retrieval and Reranking

Our retrieval process combines semantic-based vector search by computing similarity scores. However, we extend this by incorporating two features from the data representation by determining the authoritativeness of a legal precedent using the PageRank algorithm as well as the cited opinions of the opinions that were retrieved.

As discussed in Section 3.2, the verbatim passages that discuss the facts of the case, the four factors of Fair Use, and the conclusion of the case were extracted using an LLM. We then chunk the passages and embed them using Gecko. We use cosine similarity as our method in computing the similarity between the documents. Furthermore, to incorporate the citation metrics as well as the court hierarchy, we used the PageRank algorithm to quantify the relative importance of each court decision within the legal citation network. This is done for both the legal opinions (based on the citations) and the courts (based on appellate relationships).

We used the PageRank algorithm to quantify two distinct but complementary aspects of legal authority: citation authority, calculated from the inter-opinion citation network, and court hierarchy, based on appellate relationships among courts. While these dimensions capture different sources of legal relevance—influence through citation versus institutional authority by position in the judiciary, they are often correlated in practice, as higher courts tend to issue opinions that are more frequently cited. However, we include this dual representation to allow our model to consider both the structural and reputational weight of each legal source.

The retrieved documents are ranked based on a convex combination, where

$$s_i = w_{\text{text}}\text{TextSim}_i + w_{\text{cit}}\text{Citation}_i + w_{\text{court}}\text{Court}_i \quad (1)$$

such that $w_{\text{text}}, w_{\text{cit}}, w_{\text{court}} \in [0, 1]$ and $w_{\text{text}} + w_{\text{cit}} + w_{\text{court}} = 1$. The weights hence can be interpreted as hyperparameters in which one can adjust for optimal retrieval. We applied `min-max` scaling to the scores individually to ensure that each score is between 0 and 1. In the current prototype, the weights are manually specified by the user, which allows legal experts to adjust the retrieval behavior based on the characteristics of the query or dispute. For instance, a legal expert may prioritize citations and court hierarchy in appellate-heavy disputes, while another may favor textual similarity in novel or atypical cases. In future work, we plan to systematically evaluate the effect of different weight configurations using ablation studies.

Lastly, based on the top k legal precedents that are retrieved, an additional parameter n can be specified to retrieve the cited opinions by the retrieved cases based on the citation and court rankings. These cited cases are included directly in the inference step to provide broader legal context and to simulate how a legal practitioner might draw from precedent when reasoning about a novel dispute.

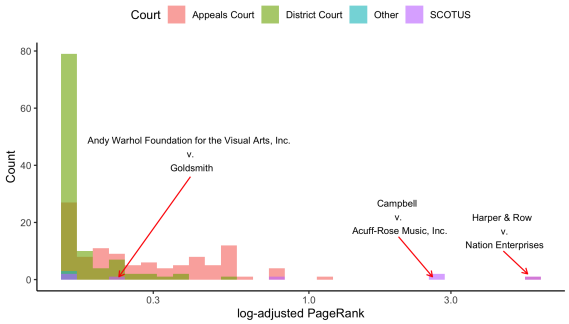


Figure 3: Distribution of Legal Case Influence by Court and PageRank

Figure 3 displays the distribution of legal cases by their log-adjusted PageRank, a measure of influence within the legal citation network. Most cases, particularly from District Courts, cluster at lower PageRank values, while a few landmark Supreme Court decisions like *Campbell v. Acuff-Rose Music, Inc.* and *Harper & Row v. Nation Enterprise* have disproportionally high PageRank values. This is, of course, not surprising, as many natural networks exhibit power-law distributions [31].

Although *Warhol v. Goldsmith*, 598 U.S. 508 (2023), is considered to be highly significant by most legal scholars, its recency means it has had limited time to accumulate citations. This is a limitation of PageRank which does not account for time. Future work could explore time-adjusted measures to better capture the emerging influence of newer cases.

3.4. Current Progress and Implementation

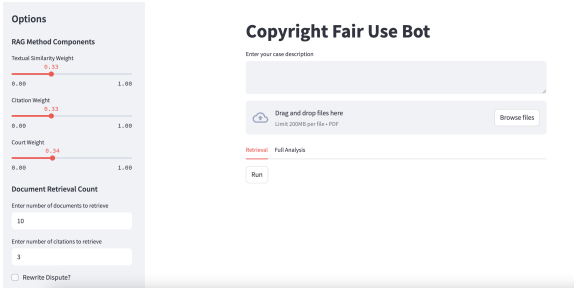
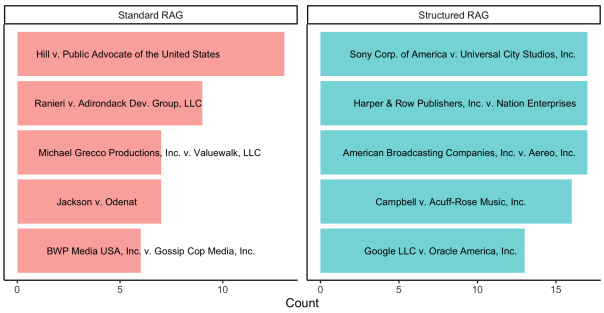
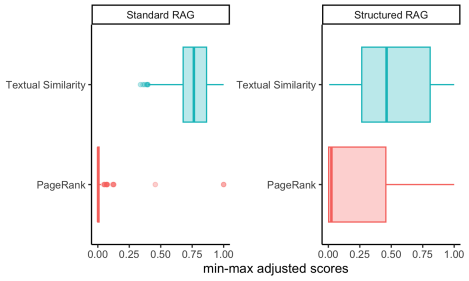


Figure 4: Interface of the Fair Use Legal Bot

As of April 2025, we have developed a functional prototype of the application. The current prototype of the Fair Use Legal Bot can be found here. Figure 4 shows the interface where users can provide their case or dispute description for fair use analysis. Users can upload relevant documents in PDF format or enter a written description directly into the input box. They can also customize the retrieval algorithm by adjusting the weights for textual similarity, citation frequency, and court relevance, as well as specify the number of documents and citations to retrieve.



(a) Most Commonly Retrieved Cases in Each Retrieval Method



(b) Boxplot of scores for retrieved cases under Standard RAG and Structured RAG.

Figure 5: Comparison of Retrieval Methods. The Standard RAG achieves higher textual similarity but lacks doctrinal authority as measured by PageRank, and the most common cases retrieved tend to be less authoritative.

This prototype was developed mainly for internal testing and refinement of the retrieval process, but is accessible to external users for testing and evaluation.

The current version supports uploading a complaint or a text description of a dispute, and retrieve the most relevant documents based on the hyperparameters configured in the left panel ("RAG Component Methods"). We have currently implemented the manual weighting of the three parameters, and users can specify the k number of documents as well as n number of cited cases.

While formal evaluations and user studies have not yet been conducted, the current version of the application establishes a strong foundation for future experimentation and ablation studies.

4. Preliminary Testing on Retrieval

In this section, we describe the preliminary experiments conducted to evaluate our retrieval system. The primary goal was to compare the baseline Standard RAG approach with our proposed Structured RAG method, which incorporates additional legal structure in the form of citation authority and court hierarchy.

Table 2
Comparison of Retrieval Methods with Grouped Metrics. All scores are min-max scaled.

Retrieval Method	PageRank		Text Similarity	
	Mean	SD	Mean	SD
Standard RAG	0.026	0.114	0.753	0.169
Structured RAG	0.213	0.315	0.521	0.305

We used the unresolved copyright complaints from PACER for our preliminary testing and the experiments were run under two configurations:

- **Standard RAG:** The retrieval process relies solely on textual similarity. The hyperparameters are set $w_{\text{text}} = 1$ and $w_{\text{cit}} = w_{\text{court}} = 0$
- **Structured RAG:** The retrieval incorporates legal structural elements by assigning uniform weights to each component $w_{\text{text}} = w_{\text{cit}} = w_{\text{court}} = 0.333$.

We compared the PageRank score and cosine similarity score (reflecting textual similarity) of the retrieved legal opinions. Unsurprisingly, the Standard RAG achieves high textual similarity (Mean = 0.753, SD = 0.169) but retrieves cases with low doctrinal authority as reflected by its low PageRank scores (Mean = 0.026, SD = 0.114). In contrast, the Structured RAG yields higher doctrinal relevance with significantly increased PageRank scores (Mean = 0.213, SD = 0.315), though its textual similarity is somewhat lower (Mean = 0.521, SD = 0.305) - Figure 2. These findings support our hypothesis that adding legal structural data enhances the retrieval of legally significant cases, and could be a way to reduce problems arising from naive retrieval and inapplicable authority [11].

However, we want to note that this preliminary testing is limited since (1) PageRank is an imperfect estimate of the doctrinal authority as noted in Section 3 and (2) the tradeoff between textual similarity and doctrinal relevance might lead to worse Fair Use analysis.

5. Limitations and Future Work

While our prototype demonstrates promising initial results, there are several limitations that must be addressed in future work to ensure robust and reliable deployment.

5.1. Future Evaluation

Our current evaluation is limited to internal testing using unresolved copyright complaints and a curated set of legal precedents. To rigorously assess the effectiveness of our prototype, future work should include user studies with legal practitioners and creators, as well as quantitative metrics such as retrieval precision, argument validity, citation relevance, and user trust. We also plan to perform ablation studies to evaluate the individual contributions of textual similarity, citation authority, and court hierarchy in the retrieval scoring function, along with more granular retrieval methods. Informal, preliminary testing has been promising.

Additionally, it is important to evaluate not only the factual and doctrinal accuracy of generated analyses, but also the quality and persuasiveness of the legal arguments. Since legal reasoning involves a degree of subjectivity and contextual nuance, human-in-the-loop evaluations will be essential for understanding the viability of the prototype.

5.2. Limitations of Current Work

Despite our focus on grounding retrieval in legal structure, our prototype still exhibits known weaknesses of LLMs, in-

cluding hallucination and sycophancy. For instance, when presented with vague or generic inputs, the model may generate speculative or overly confident legal conclusions. This is especially problematic in scenarios where users are not legally trained and may rely too heavily on the prototype's output without independent verification.

While the prototype is designed with legal structure in mind, its interface and guidance mechanisms are not yet optimized for lay users. Since the goal is to support individuals subjected to unfair DMCA takedowns, there is a need for an appropriate information elicitation phase—where an LLM prompts users to describe their dispute, provide specific details relevant to a Fair Use defense, and potentially disclose points that might disqualify them from Fair Use protection.

Furthermore, as noted in Section 3, the use of PageRank—which has a bias against recency—may result in the omission of relevant judicial opinions that reflect evolving doctrine. Empirical legal work that studies how courts interpret and apply legal doctrines can be integrated into the model to complement the limitations of citation-based metrics by capturing nuanced shifts in judicial reasoning and doctrinal emphasis [32].

5.3. Extension to Other Legal Doctrines

The current prototype assumes the input case pertains to Fair Use and does not include functionality for classifying the applicability of legal doctrines. Expanding the prototype to determine whether Fair Use is even the appropriate legal framework for a given dispute remains an important next step. The choice to use knowledge graphs was made with the intent of enabling future integration of other legal doctrines.

Future work can build on and extend the current prototype by constructing modules of local expertise that integrate into a larger system. This will likely require a routing mechanism—for instance, training a classifier to determine which legal doctrine applies to a case, and then routing it to the relevant 'expert'.

6. Discussion and Conclusion

This paper introduces a structured approach to Retrieval-Augmented Generation (RAG) for legal analysis, using the Fair Use Doctrine in copyright law as a case study. By incorporating knowledge graphs that model citation networks, court hierarchies, and statutory factor-level reasoning, our system aims to address persistent issues in legal LLM applications—namely hallucination, irrelevant retrieval, and inadequate legal inference.

Our method aligns with how legal professionals approach multi-factor tests, providing a more interpretable and granular framework that improves both retrieval and downstream reasoning. The integration of citation-based authority metrics and Chain-of-Thought reasoning supports more grounded and nuanced analysis than traditional vector-based approaches alone.

While our prototype remains in an early stage, the foundational design lays the groundwork for both academic study and practical applications. Future work will focus on empirical validation, interface development for non-expert users, and potential generalization to other areas of law. We believe that structuring AI systems around legal doctrines and reasoning patterns holds significant promise for improving access to justice and legal assistance.

Acknowledgments

We would like to thank the Berkman Klein Center for Internet & Society at Harvard University for their support of this research.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword, Generate literature review, and Improve writing style. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] U.S. Congress, 17 u.s.c. § 512 - limitations on liability relating to material online, https://www.law.cornell.edu/uscode/text/17/512, 2024. Retrieved from https://www.law.cornell.edu/uscode/text/17/512.

[2] U.S. Congress, 17 u.s.c. § 107 - limitations on exclusive rights: Fair use, https://www.law.cornell.edu/uscode/text/17/107, 2024. Retrieved from https://www.law.cornell.edu/uscode/text/17/107.

[3] J. D. Matteson, Unfair misuse: How section 512 of the dmca allows abuse of the copyright fair use doctrine and how to fix it, Santa Clara high-technology law journal 35 (2018) 1.

[4] S. M. Blythe, Freedom of speech and the dmca: Abuse of the notification and takedown process, European intellectual property review 41 (2019) 70–88.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[6] J. Lai, W. Gan, J. Wu, Z. Qi, P. S. Yu, Large language models in law: A survey, AI Open 5 (2024) 181–196. URL: https://www.sciencedirect.com/science/article/pii/S2666651024000172. doi:https://doi.org/10.1016/j.aiopen.2024.09.002.

[7] H. Westermann, S. Meeùs, M. Godet, A. Troussel, J. Tan, J. Savelka, K. Benyekhlef, Bridging the gap: Mapping layperson narratives to legal issues with language models, in: Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), CEUR Workshop Proceedings, Braga, Portugal, 2023. URL: https://ceur-ws.org/Vol-3441/, available under CC BY 4.0 license.

[8] J. Savelka, K. D. Ashley, M. A. Gray, H. Westermann, H. Xu, Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise?, in: Proceedings of the Sixth Workshop on Automated

Semantic Analysis of Information in Legal Text (ASAIL 2023), CEUR Workshop Proceedings, Braga, Portugal, 2023. URL: http://ceur-ws.org/Vol-3441/, available under CC BY 4.0 license.

[9] N. Guha, J. Nyarko, D. E. Ho, C. Re, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G. M. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. J. Nay, J. H. Choi, K. Tobia, M. Hagan, M. Ma, M. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, Z. Li, Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: https://openreview.net/forum?id=WqSPQFxFRC.

[10] M. Dahl, V. Magesh, M. Suzgun, D. E. Ho, Large legal fictions: Profiling legal hallucinations in large language models, Journal of Legal Analysis 16 (2024) 64–93. URL: https://doi.org/10.1093/jla/laae003. doi:10.1093/jla/laae003.

[11] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, Hallucination-free? assessing the reliability of leading ai legal research tools, 2024. URL: https://arxiv.org/abs/2405.20362. arXiv:2405.20362.

[12] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87. doi:10.1162/neco.1991.3.1.79.

[13] B. J. Gutiérrez, Y. Shu, W. Qi, S. Zhou, Y. Su, From rag to memory: Non-parametric continual learning for large language models, 2025. URL: https://arxiv.org/abs/2502.14802. arXiv:2502.14802.

[14] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[15] S. Chen, H. Zhang, T. Chen, B. Zhou, W. Yu, D. Yu, B. Peng, H. Wang, D. Roth, D. Yu, Sub-sentence encoder: Contrastive learning of propositional semantic representations, 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), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1596–1609. URL: https://aclanthology.org/2024.naacl-long.89/. doi:10.18653/v1/2024.naacl-long.89.

[16] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, D. Yu, Dense X retrieval: What retrieval granularity should we use?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15159–15177. URL: https://aclanthology.org/2024.emnlp-main.845/. doi:10.18653/v1/2024.emnlp-main.845.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[18] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal, Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10014–10037. URL: https://aclanthology.org/2023.acl-long.557/. doi:10.18653/v1/2023.acl-long.557.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[20] G. Sartor, P. Santin, L. D. Caro, Chasing the invisible in the grammar of repetitions: A network analysis approach to fiscal state aids, in: Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), CEUR Workshop Proceedings, Braga, Portugal, 2023, pp. 1–10. URL: http://ceur-ws.org/Vol-3441/, use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[21] D. Sanmartin, Kg-rag: Bridging the gap between knowledge and creativity, 2024. URL: https://arxiv.org/abs/2405.12035. arXiv:2405.12035.

[22] Z. Sepasdar, S. Gautam, C. Midoglu, M. A. Riegler, P. Halvorsen, Enhancing structured-data retrieval with graphrag: Soccer data case study, 2024. URL: https://arxiv.org/abs/2409.17580. arXiv:2409.17580.

[23] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web., Technical Report, Stanford infolab, 1999.

[24] G. Team, Gemini: A family of highly capable multimodal models, 2024. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

[25] J. Lee, Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, K. Hui, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. H. Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, I. Naim, Gecko: Versatile text embeddings distilled from large language models, 2024. URL: https://arxiv.org/abs/2403.20327. arXiv:2403.20327.

[26] T. F. L. Project, Recap archive, https://www.courtlistener.com/recap/, 2020. Accessed January 23, 2020.

[27] J. Cushman, M. Dahl, M. Lissner, eyecite: A tool for parsing legal citations, Journal of Open Source Software 6 (2021) 3617. URL: https://doi.org/10.21105/joss.03617.

[28] Administrative Office of the U.S. Courts, Public access to court electronic records (pacer), https://pacer.uscourts.gov, ???? Original source of federal court records.

[29] Neo4j, Inc., Neo4j Graph Database, 2025. Version 5.26.2, Available at: https://neo4j.com/.

[30] A. Ferrario, M. Loi, How explainability contributes to trust in ai, in: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022, pp. 1457–1466.

[31] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512. doi:10.1126/science.286.5439.509.

[32] B. Beebe, An empirical study of u.s. copyright fair use opinions, 1978–2005, University of Pennsylvania Law Review 156 (2008) 549–634. URL: https://scholarship.law.upenn.edu/penn_law_review/vol156/iss3/2/.