What the Scores Miss: Semantic Retrieval and the Limits of Automated Evaluation in Legal AI

ANONYMOUS AUTHOR(S)

Current AI tools for legal professionals focus on search and summarization, leaving a gap in systems that support the compositional, analogical work of judicial drafting. We present the design and early-stage evaluation of a corpus-level AI drafting tool for UK judges, built on a corpus exploration framework [5] and a 4,000-judgment corpus from the UK National Archives. A four-version retrieval evaluation, graded by a legal domain expert using a collaboratively developed rubric, revealed that results with the highest cosine similarity scores frequently reproduced the input text verbatim rather than surfacing genuine legal analogies. This is a known property in RAG systems [4, 8, 13]. What this work contributes is a characterization of why its consequences differ in expert professional writing: a judge who receives a near-verbatim suggestion drawn from their own prior writing or a case already in hand may mistake self-referential retrieval for independent analogical support. Automated metrics cannot surface this gap. Expert grounding, along with algorithms that enhance retrieval quality and interfaces that support information analysis, can. We motivate a planned user study, informed by findings on how legal professionals engage with example-based writing tools [17], to test whether corpus-level example surfacing supports or erodes judicial drafting skill, and call for discussion on how to evaluate AI tools in high-stakes professional domains where the meaning of a good result cannot be captured by similarity scores alone.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; • **Computing methodologies → Artificial intelligence**; • **Applied computing → Law, social and behavioral sciences**.

Additional Key Words and Phrases: legal AI, judicial drafting, retrieval-augmented generation, corpus exploration, deskilling, RAG evaluation, professional AI tools

ACM Reference Format:

Anonymous Author(s). 2018. What the Scores Miss: Semantic Retrieval and the Limits of Automated Evaluation in Legal AI. In *Proceedings of CHIWORK '26: Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '26), (CHIWORK'26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

AI adoption in legal work is accelerating. Practitioners use general-purpose language models for drafting, summarization, and research [10], while major platforms like LexisNexis and Westlaw have embedded AI retrieval features into routine workflows [12]. This proliferation has arrived alongside documented failures: attorneys have submitted briefs with fabricated citations [11], and judges have flagged concerns about the transparency of tools on which they are expected to rely [3].

These failures share a root cause. Current legal AI is optimized for retrieval and summarization, not for the higher-order cognitive work that legal writing demands. Writing a judicial opinion requires understanding how prior courts constructed their reasoning, what argumentative conventions govern a given jurisdiction, and how a new decision

Manuscript submitted to ACM

situates itself within evolving doctrine. These are corpus-level cognitive tasks, and no widely deployed tool addresses them directly.

This paper presents the design and early evaluation of a corpus-level AI drafting system for UK judges, built on CorpusStudio [5] a corpus exploration framework. Unlike generative AI tools, the system never produces new text; it retrieves sentences from real prior judgments and surfaces multiple examples simultaneously, supporting comparison rather than anchoring on a single AI-produced suggestion. We contribute: (1) a system design that operationalizes corpus-level support for judicial drafting without risk of hallucination; (2) a retrieval evaluation that reveals how verbatim retrieval in RAG systems takes on distinct professional consequences in legal writing contexts; and (3) a planned user study to test skill support versus skill substitution in AI-assisted professional reasoning. We frame this as a prequel. The system is under development and is being iteratively refined based on feedback from legal and HCI experts. We submit this work to provoke conversation about how the research community should evaluate AI tools designed for high-stakes professional domains.

2 Related Work

2.1 AI in Legal Work

Prior HCI research on AI in legal work has examined practitioner attitudes [3, 15], precedent retrieval [14], and the risks of over-reliance [11]. Research with UK judges found consistent priorities across participants: transparency, verifiability of AI outputs, and protection of the drafting skills developed through years of practice [15]. These findings ground the design principles of our system. The deskilling concern is well-established in HCI literature: when AI tools mediate professional tasks, practitioners may adapt their working practice around the tool's outputs rather than their own judgment [1, 7], and empirical studies show that over reliance on AI assistance can degrade performance on tasks requiring expert reasoning [6, 18]. Tools that anchor users on single AI-generated outputs reduce deliberation and may degrade the cognitive practices that constitute professional expertise. In law specifically, where junior practitioners develop critical reasoning through repetitive drafting and research, this risk is particularly acute [15, 17]. Our system responds by surfacing multiple ranked examples [5] from real prior opinions rather than generating new text.

2.2 RAG Evaluation and Retrieval Quality

Verbatim or near-verbatim retrieval in dense embedding systems is a documented challenge in RAG evaluation research. RAGAS [8] and ARES [13] both identify faithfulness and context precision as core evaluation dimensions precisely because retrieved content does not always provide the substantive grounding it appears to. Cuconasu et al. [4] show that retrieval quality is a primary determinant of RAG performance and that standard similarity metrics are insufficient proxies for what users actually need. RAG systems are designed to retrieve semantically similar text, not analogies, and verbatim matches represent retrieval working as intended by its own metric. The problem we identify is not that the system malfunctions but that its success metric and the domain's criteria for quality suggestions diverge in a way that neither the RAG system nor standard automated evaluation through semantic scores can detect.

Our work extends this line of inquiry to a domain where the stakes of retrieval failure are qualitatively different from those in general information retrieval. In general retrieval, a verbatim result is simply not useful and the user notices and moves on. In judicial drafting, a near-verbatim suggestion may reflect a commonly quoted passage from a landmark case that recurs across the corpus, surfacing as a high confidence result not because it is analogous to the text in the draft but because of the similar language. In legal systems in the common law tradition, evidence of a

passage's recurrence in the corpus will tend to signal its authoritativeness as a judicial maxim or dictum, 'Lawyers speak in terms of the law being settled by virtue of long standing and long followed authorities' [9]. Such information is independently valuable in the legal context, and may also be captured by a corpus-level tool in summary form, but does not represent analogical thinking through an appreciation of which, the writer's own legal reasoning might benefit in distinctive ways. Moreover, retrieval of a list of near verbatim matches may give the misleading impression of a monolithic understanding of the issue when in fact only one analogous example has been surfaced. Addressing this issue is important for legal AI tools that aim to facilitate enhanced professional drafting standards rather than introduce new complications into judicial workflows.

2.3 Corpus-Level Exploration

The system builds on CorpusStudio [5], an interactive writing environment that surfaces emerging writing patterns from a corpus of prior documents as the user drafts, using spatial retrieval to return contextually relevant sentence examples without generating new text. CorpusStudio was designed for academic writing within the HCI field, surfacing many real examples from a community's prior work rather than generating new text. This core functionality transfers naturally to legal drafting, where writing is similarly precedent-driven and governed by common conventions. Swoopes et al. found that legal professionals responding to a CorpusStudio demo were receptive to the idea of orienting their drafting to the structure and phrasing of similar texts, describing it as useful for learning stylistic norms and understanding audience expectations [17]. This motivates our core hypothesis that corpus-level suggestion surfacing may support professional skill development, unlike many AI approaches that risk deskilling. The planned user study is designed to test this hypothesis.

3 System Design

3.1 Architecture and Corpus

The system ingests UK Supreme Court and Upper Tribunal judgments from the National Archives Case Law API, approximately 4,000 in total. An automated pipeline using GPT-4 Mini classifies each judgment as Majority, Concurring, or Dissenting by analyzing the final 1,000 characters of each document. Each judgment is then segmented at the sentence level and assigned one of six functional role labels: Introduction, Facts, Authority, Doctrine & Policy, Reasoning, and Judgment. Sentence-level labeling outperformed paragraph-level labeling because individual sentences within a paragraph frequently serve different functions, and propagating a single label down to all sentences in a paragraph introduced noise during retrieval.

Sentences are embedded using a legal-domain model ranked first on the Massive Legal Embedding Benchmark for UK law retrieval [2] and stored in a Supabase PostgreSQL backend using the pgvector extension for K-Nearest Neighbors (KNN) similarity search. The writing interface presents a text editor alongside a suggestion panel showing ranked sentences from the corpus. As the user drafts, they press tab when they want a relevant suggestion surfaced and the last 100 characters are used as a query to retrieve the most contextually similar sentence from the corpus, returning results ranked by cosine similarity to the draft text.

3.2 Design Principles

Three principles guided the design. First, the system surfaces real sentences from real prior opinions with traceable citations. It never generates text. Second, it shows multiple ranked examples simultaneously rather than a single best

Ver.	Enrichment Strategy	Avg. Score	Expert Rank
V1	Functional role label only	0.6489	3rd
V2	Context tag (case name, year, judge prepended)	0.6689	2nd
V3	Full case summary prepended	0.5751	4th
V4	Hybrid: role label + context tag	0.6542	1st

Table 1. Retrieval versions, average cosine similarity scores, and expert ranking. V4 ranked highest in expert evaluation despite not having the highest automated score.

match, consistent with prior findings that comparison across examples deepens understanding of writing conventions [5]. Third, users can apply filters by opinion type (Majority, Concurring, or Dissenting) if they want more control over what they are seeing, which prevents the system from inadvertently surfacing dissenting arguments as models for majority opinions.

4 Retrieval Evaluation

4.1 Four-Version Experiment

To identify the optimal metadata enrichment strategy, the team built and evaluated four retrieval versions on an extended corpus that incorporated 14 additional HTML cases and six PDF cases provided by the domain expert. Each version prepended different metadata to sentences before embedding.

4.2 When Similarity Diverges from Analogy

Automated similarity scores were insufficient as the sole evaluation criterion. To grade retrieval quality on legal analogy, the team developed a grading rubric collaboratively with a domain expert in Irish and UK common law, then applied it independently to all suggestions across all four versions. The expert subsequently reviewed the grading and provided corrections. This process covered eight test texts per version with ten suggestions each, graded on legal analogy rather than surface overlap.

The evaluation confirmed a pattern well-documented in RAG research [4, 8, 13]: results with the highest cosine similarity scores (0.85 and above) frequently reproduced the input sentence near-verbatim. Automated metrics not only fail to detect this but actively rank these results at the top. Figure 1 shows heatmaps of manual analogy grades (top row per version) and semantic similarity scores (bottom row per version) across all four system versions. The visual comparison makes the divergence clear: the cells with the darkest similarity scores are often the cells with the lowest analogy grades.

What this evaluation adds beyond existing RAG research is a characterization of why this divergence between retrieval objective and domain requirements matters differently in judicial drafting. A near-verbatim suggestion might reflect the passage's authoritativeness as a judicial dictum but would fail to provide the judge with independent analogical support. Faced with a list of near verbatim suggestions, the judge might mistake the passage's legal authority qua formal dictum for a natural uniformity in judicial thought. This is not a retrieval precision problem. It is an interpretive problem that automated metrics alone cannot surface, with direct implication for how retrieval confidence should be communicated in legal AI tools.

V4 emerged as the most reliable version despite its lower average automated score. What matters in a drafting tool is the concentration of genuinely analogous results at positions 1 through 3, since those are the suggestions users engage

Fig. 1. Manual analogy grades (top row per version pair) and semantic similarity scores (bottom row) for all four retrieval versions across eight test texts and ten suggestions each. High-similarity results, shown in green in the bottom rows, frequently correspond to low analogy grades, shown in red in the top rows. Blue outlines mark near-verbatim matches.

with first. V4 achieved the strongest concentration there. V3 performed worst: prepending a full case summary diluted the specificity of individual sentence embeddings, pulling retrieval toward the general rather than the precise.

This finding motivates a design recommendation applicable to any high-stakes professional AI writing tool: similarity-based ranking should be accompanied by explicit detection and disclosure of near-verbatim matches. Treating them as high-confidence successes, as most retrieval systems currently do by default, actively misleads users who are most likely to engage with top-ranked results.

5 Planned User Study

The central open question is whether corpus-level suggestion surfacing supports judicial drafting skill or gradually substitutes for it. We have developed a study protocol targeting UK judges and law clerks using a counterbalanced task design: participants complete a 10-minute drafting task under two conditions, a standard text editor and the system's writing interface, with order randomized across participants.

Participants think aloud throughout. Post-task surveys assess ease of drafting, stylistic alignment with court expectations, and structural adherence on 7-point scales. A semi-structured interview examines workflow integration, concerns about professional voice and skill erosion, and a closing counterfactual: what would participants have done without the tool?

Two hypotheses will be tested. First, that participants using the corpus-level tool develop a clearer understanding of their courts' writing conventions than those drafting without it. Second, that participants who encounter high-similarity but verbatim results either recognize them as different from genuine analogies, or they do not, which would itself be a

significant finding about how practitioners calibrate trust in AI suggestions. We present a high-level summary of the protocol here to invite methodological feedback.

6 Discussion

This work raises three questions we believe are relevant to researchers working on AI in professional contexts.

How should the field evaluate skill support versus skill substitution in AI writing tools? The deskilling literature has documented that over-reliance on AI assistance can degrade performance on tasks requiring expert reasoning [6, 18], and that knowledge workers across industries identify deskilling as a primary concern when adopting generative tools [1, 6, 7, 18]. Our system does not generate text, yet the risk remains: a practitioner who consistently selects from a ranked list may develop different cognitive habits than one who composes from scratch. Measuring this distinction requires methods beyond task performance metrics. We need instruments that capture changes in deliberation, not just output quality.

How should professional AI tools surface retrieval confidence? The RAG evaluation literature [4, 8, 13] has largely treated verbatim retrieval as a system-level problem to be solved through deduplication or filtering. Our finding suggests it also needs to be surfaced to the user, particularly where self-referential retrieval carries signals a specific salience in common law legal reasoning contexts. In practice, a system should be able to distinguish between a result ranked highly for cross-document similarity and one ranked highly due to lexical overlap with the query itself. Transparency about why a result was ranked highly, not just its score, is a design requirement that connects to the community's broader interest in fairness and accountability in AI-assisted work.

What does responsible AI design look like for judicial work specifically? Judges face unique accountability constraints. The legitimacy of a decision depends on both its correctness as well as the reasoning process behind it [16]. A tool that surfaces high-confidence suggestions that are actually self-referential could undermine that process without the judge ever knowing. Designing tools that support deliberative judicial reasoning rather than offering the appearance of intellectual uniformity is a challenge that sits within HCI's core competencies.

7 Conclusion

We have presented the design and retrieval evaluation of a corpus-level AI drafting tool for judicial professionals and a planned user study to test its effects on professional reasoning. The central contribution is not the identification of verbatim retrieval as a new problem. It is the demonstration that the tendency of RAG models to surface verbatim matches carries qualitatively different consequences in legal professional writing contexts, and that automated evaluation metrics can conflate distinct signals of legal salience – an argument's formal authority and its analogical coherence. Only expert grounding makes the gap visible. We argue this challenge has implications for any high-stakes professional AI tool where the meaning of a good result cannot be captured by surface similarity alone. How should AI tools designed to support professional expertise make their own limitations visible to the people using them? That is the question this work is built around, and we hope it is one with which the research community will engage.

References

[1] Claus Bossen and Randi Markussen. 2010. Infrastructuring and Ordering Devices in Health Care: Medication Plans and Practices on a Hospital Ward. *Computer Supported Cooperative Work* 19, 6 (2010), 615–637.

[2] Ulysse Butler, Angus R. Butler, and Anna L. Malec. 2025. The Massive Legal Embedding Benchmark (MLEB). *arXiv preprint* (2025). arXiv:2510.19365

[3] Courts and Tribunals Judiciary. 2023. Artificial Intelligence (AI) Guidance for Judicial Office Holders.

[4] Florin Cuconasu, Giovanni Trappolini, Andrea Santilli, Zhaoxi Noci, Alon Yehuda, Fabrizio Silvestri, and Nir Rosenfeld. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. ACM, New York, NY, USA.

[5] Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena L. Glassman. 2025. CorpusStudio: Surfacing Emergent Patterns in a Corpus of Prior Work While Writing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, New York, NY, USA, Article 1211, 19 pages. doi:10.1145/3706598.3713974

[6] Fabrizio Dell'Acqua, Edward III McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, Francois Candelon, and Karim R. Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality*. Technical Report 24-013. Harvard Business School. https://ssrn.com/abstract=4573321

[7] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2019. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[8] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint* (2023). arXiv:2309.15217

[9] High Court of Ireland. 2017. McCaffrey v. Central Bank of Ireland [2017] IEHC 546. at [95], per Noonan J..

[10] Jiaxi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large Language Models in Law: A Survey. *AI Open* 5 (2024), 181–196.

[11] Cyrus Moore, Margaret Hu, Clare Garvie, Kate Munger, Andrew Selbst, and Natasha Duarte. 2025. Opinion Concerning the Responsible Use of AI in the U.S. Criminal Justice System. *Commun. ACM* 68, 9 (2025), 41–44.

[12] Saad Nasir, Qasim Abbas, Siyuan Bai, and Rao Adeel Khan. 2025. A Comprehensive Framework for Reliable Legal AI: Combining Specialized Expert Systems and Adaptive Refinement. *arXiv preprint* (2025). arXiv:2412.20468v2

[13] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv preprint* (2023). arXiv:2311.09476

[14] Daniel Schwarcz, Suzanne Manning, Patrick Barry, David R. Cleveland, J. J. Prescott, and Bryce Rich. n.d.. AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice. Working paper.

[15] Erin Solovey, Brian Flanagan, and Daniel Chen. 2025. Interacting with AI at Work: Perceptions and Opportunities from the UK Judiciary. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25)*. ACM, New York, NY, USA, Article 5, 8 pages. doi:10.1145/3729176.3729192

[16] Lawrence B. Solum. 2004. Procedural Justice. *Southern California Law Review* 78 (2004), 181.

[17] Chelse Swoopes, Ziwei Gu, and Elena L. Glassman. 2025. Interface Design to Support Legal Reading and Writing: Insights from Interviews with Legal Experts. *arXiv preprint* (2025). arXiv:2509.24854

[18] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2024. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA. doi:10.1145/3613904.3642700

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009