

Large Language Model-Based Semantic Search for Legal Document Retrieval and Case Analysis

Trevor R. Robles

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.

trevorrobes312@unr.edu

Xiuke Tian

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.

xtian@colostate.edu

Kanald Rnderson

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
KS, USA.

ronald.work@ku.edu

Herish Mair

Department of Computer Science, George Mason University, Fairfax, VA, USA.

harish.nair708@gmu.edu

Abstract

Large language models are reshaping the design of legal information systems by enabling semantic search over complex, heterogeneous corpora of statutes, judicial opinions, contracts, and case records. This article presents a system-level analysis of large language model-based semantic search for legal document retrieval and case analysis. It examines the architectural foundations that combine dense retrieval, reranking, retrieval-augmented generation, and persistent user modeling. The discussion emphasizes structural trade-offs between retrieval quality, latency, interpretability, and computational cost. It further addresses the legal domain constraints of authority, confidentiality, fairness, and explainability. The article argues that effective legal semantic search cannot be reduced to embedding quality or generative fluency; it requires an integrated socio-technical design that accounts for domain adaptation, evaluation, deployment infrastructure, robustness, and governance. By connecting technical mechanisms with institutional responsibilities, the paper offers a forward-looking perspective on building legal retrieval systems that are not only accurate but also auditable, sustainable, and aligned with the normative expectations of legal practice.

Keywords

large language models, semantic search, legal document retrieval, case analysis, retrieval-augmented generation, legal information systems, governance.

1. Introduction

Legal document retrieval and case analysis have historically been dominated by lexical search technologies that emphasize exact term overlap and statistical term weighting. Systems based on probabilistic relevance frameworks such as BM25 provide interpretable and efficient retrieval mechanisms [1]. However, lexical methods struggle with linguistic variation,

synonymy, complex legal phrasing, and the conceptual relationships that are common in legal reasoning. The introduction of deep bidirectional contextual language representations enabled models to capture semantic relationships beyond surface-level terms [2]. This shift created the possibility of retrieving legal documents based on meaning rather than strictly on shared vocabulary. Subsequent research on dense passage retrieval demonstrated that learned dense vector representations can outperform sparse lexical models in open-domain question answering by encoding semantic content into compact embeddings [3]. Retrieval-augmented generation further extended these capabilities by using retrieved passages as evidence for generative language models, thereby grounding outputs in specific source material [4]. More recent work on user memory systems has examined how query matching and intent modeling can be improved through persistent representations of prior user interactions, which is particularly relevant for longitudinal legal research tasks [5].

The legal domain introduces distinctive requirements that go beyond general-purpose search. Legal users must often distinguish between binding authority and persuasive authority, interpret statutory language in context, compare fact patterns across jurisdictions, and trace the precedential value of a case. A retrieval system that returns semantically similar but legally irrelevant documents may be more harmful than a system that returns fewer but more authoritative results. Moreover, legal document retrieval operates within institutional environments that impose confidentiality, data retention, and auditability obligations. These requirements create structural tensions between the flexibility of large language models and the need for predictable, reviewable behavior. This article therefore approaches legal semantic search as a system problem rather than a model problem. It analyzes architectural choices, retrieval workflows, robustness and fairness considerations, infrastructure constraints, evaluation frameworks, and governance mechanisms. The goal is to provide a comprehensive, interdisciplinary account of how large language model-based semantic search can be designed, deployed, and sustained in legal contexts.

2. Background and Related Work

Semantic search in legal information systems builds on advances in representation learning, retrieval, and generative language modeling. Sentence embedding methods such as Sentence-BERT introduced efficient semantic similarity computation by adapting transformer-based architectures to produce fixed-length sentence representations [6]. These representations enable large-scale comparison of legal paragraphs without requiring direct word overlap. In the legal domain, domain-specific pretraining has been shown to improve performance on tasks involving legal vocabulary and document structure. LEGAL-BERT demonstrated that continuing pretraining on legal corpora can produce representations that better capture statutory and contractual language [7]. Such domain adaptation is essential because legal language includes technical terms, archaic expressions, defined terms, and cross-references that are not well represented in general-domain corpora. Benchmarks such as LegalBench have further provided structured evaluations of legal reasoning, allowing researchers to compare models across varied legal tasks including issue spotting, rule application, and interpretation [8].

The underlying transformer architecture introduced attention mechanisms that support long-range dependency modeling and parallelizable training [9]. Large-scale autoregressive models such as GPT-3 later demonstrated that language models can perform few-shot tasks from natural language instructions without extensive task-specific fine-tuning [10]. This capability is attractive for legal search because legal queries are often underspecified and require

iterative clarification. Instruction-tuned models further improved the ability of language models to follow user intent and generate more controlled outputs [11]. Together, these developments enable semantic search systems that combine dense retrieval, neural reranking, and generative summarization. However, they also introduce challenges related to hallucination, authority, and interpretability, which are especially consequential in law. A system-level perspective is therefore necessary to integrate these component capabilities into a coherent and trustworthy legal search infrastructure.

3. System Architecture and Retrieval Workflows

A large language model-based legal semantic search system typically comprises several interacting components: corpus ingestion, document segmentation, embedding, indexing, query understanding, retrieval, reranking, generation, and explanation. The design of these components must balance completeness, latency, and legal authority. Dense retrieval models encode queries and documents into vector spaces where semantic proximity can be computed efficiently. However, pure vector similarity may overlook exact statutory references, case citations, and defined terms that have high legal significance. Hybrid retrieval architectures therefore combine sparse lexical signals with dense semantic signals to maintain precision for legally significant terms while expanding recall through semantic similarity. Reranking models then refine the candidate set using contextualized relevance judgments, which can incorporate legal constraints such as jurisdiction, date, and court level.

Case analysis benefits from generative reasoning over retrieved documents, but generation must be carefully constrained. Chain-of-thought prompting has been used to elicit intermediate reasoning steps in large language models [12]. In legal contexts, such reasoning can help a model articulate how a retrieved case supports a particular conclusion. However, chain-of-thought outputs must be treated as provisional interpretations rather than authoritative legal analysis. Retrieval workflows may also employ late interaction mechanisms, as in ColBERT, which store token-level representations and allow fine-grained matching between query terms and document terms without requiring full cross-encoder computation [13]. This approach can improve retrieval quality while maintaining acceptable latency in large legal collections. The system architecture must also include persistent user memory to support longitudinal legal research. Storing prior queries, viewed documents, and relevance feedback allows the system to model evolving user intent and to contextualize new queries. Such memory mechanisms must be governed by strict access controls and retention policies because legal research histories can reveal confidential litigation strategy or privileged client information.

4. Robustness, Fairness, and Explainability

Robustness and fairness are critical concerns in legal semantic search because retrieval errors can affect legal outcomes. Large language models are known to reproduce biases present in training data, and their outputs can appear fluent even when factually incorrect. Bender and colleagues have cautioned that large language models can function as stochastic parrots, generating plausible language without grounding in communicative intent or reliable world knowledge [14]. In law, this risk is magnified because legal language is highly normative and context-dependent. Foundation models introduce broader sociotechnical risks because they are deployed across many domains with varying assumptions and evaluation standards [15]. Without careful system design, a legal search system may amplify doctrinal biases, favor frequently cited cases over meritorious but less visible authority, or generate summaries that omit material qualifications.

Explainability is equally important. Legal professionals must be able to justify why a document was retrieved and how a generated case summary was derived. A system that returns a ranked list without explaining the relevance criteria may be difficult to audit in adversarial litigation or regulatory review. Algorithmic auditing frameworks emphasize the need for internal documentation, external review, and clear accountability mechanisms [16]. Applying such frameworks to legal semantic search requires logging retrieval decisions, preserving provenance of cited passages, and offering users the ability to inspect the evidence behind generated outputs. Fairness must be evaluated not only across demographic groups but also across legal subject areas, jurisdictions, and types of litigants. Overrepresentation of federal appellate decisions, for example, may create structural bias against trial court outcomes or administrative agency rulings. System designers should therefore incorporate fairness audits that examine corpus composition, retrieval distribution, and summary accuracy across different legal subfields.

5. Infrastructure and Deployment Considerations

Deploying large language model-based legal search systems requires significant computational infrastructure and careful attention to operational sustainability. Scaling laws for neural language models show that performance tends to improve predictably with model size, dataset size, and compute [17]. However, these improvements come with increasing energy consumption, hardware costs, and inference latency. Compute-optimal training approaches suggest that many large models have been undertrained relative to their parameter count, which has implications for both performance and efficiency [18]. For legal institutions, the choice between proprietary large models and openly released models such as LLaMA involves trade-offs in control, transparency, cost, and data governance [19]. Open models can be fine-tuned on institutional legal corpora and deployed on private infrastructure, which may reduce confidentiality risks, but they require internal machine learning expertise and ongoing maintenance.

Legal organizations often require on-premises or virtual private cloud deployment to satisfy client confidentiality and regulatory obligations. This constraint affects model selection, index storage, and update frequency. Legal corpora change continuously as new opinions are issued, statutes are amended, and regulations are revised. The retrieval index must support incremental updates without degrading query performance. Embedding models may need periodic recalibration to reflect shifts in legal language and new doctrinal terminology. Caching frequently used queries and precomputed summaries can reduce operational cost, but stale outputs can mislead legal research. Infrastructure design must therefore include versioning, rollback, and monitoring mechanisms. Latency requirements vary by use case: interactive legal research demands subsecond response, while batch case analysis can tolerate longer processing times. System architects should tailor the retrieval pipeline to the expected workload, using tiered storage and model distillation where appropriate to balance cost and quality.

6. Evaluation and Operational Sustainability

Evaluation of legal semantic search systems must go beyond traditional retrieval metrics such as precision and recall. Legal relevance is multidimensional and often depends on authority, procedural posture, temporal scope, and jurisdictional fit. A document may be semantically similar to a query but legally inapplicable because it arises from a different jurisdiction or has been overturned. Retrieval-augmented generation surveys have highlighted the importance of grounding, retrieval quality, and downstream task performance in evaluating generative

search systems [20]. In legal contexts, evaluation should include human expert review of retrieved documents and generated summaries. Blind annotation by legal professionals can measure whether outputs are substantively correct, appropriately qualified, and faithful to the underlying sources. Automated metrics can support this process but cannot replace doctrinal judgment.

Operational sustainability requires continuous monitoring of system behavior after deployment. Query distributions may shift as new legal issues emerge or as user populations change. Model outputs may degrade if the underlying corpus becomes inconsistent or if embedding models become outdated. Feedback from users should be captured in a structured manner to identify systematic errors, such as overreliance on certain courts or failure to handle specific practice areas. Regular auditing of retrieval logs can reveal patterns of unintended bias or data leakage. Sustainability also involves cost management, because large language model inference and re-embedding of large legal corpora can be expensive. Legal institutions should develop clear policies for model refresh cycles, data retention, and decommissioning of outdated indices. The goal is not only to maintain technical performance but also to ensure that the system remains aligned with evolving professional duties and client expectations.

7. Governance, Policy, and Ethical Implications

Governance is essential for legal semantic search because the technology operates within a highly regulated professional environment. Lawyers have duties of competence, confidentiality, and candor, which extend to the tools they use. A legal research system that generates plausible but incorrect case citations or mischaracterizes holdings can create professional liability. Governance mechanisms should include model documentation, retrieval traceability, and clear delineation of human responsibility. Algorithmic auditing principles are useful here because they treat governance as an ongoing institutional practice rather than a one-time technical fix [16]. Law firms, courts, and legal aid organizations may need different governance models depending on their resources, risk tolerance, and client populations.

Policy implications include the need for standards around legal AI transparency and accountability. Regulators may require disclosure when generative models are used to produce legal summaries or research memoranda. Bar associations may issue guidance on how to supervise machine-generated legal analysis. Data protection laws may restrict the use of client information for model fine-tuning or user memory systems. These requirements should be integrated into system design from the beginning, not treated as compliance afterthoughts. For example, user memory systems that support personalization must allow users to inspect, export, and delete stored interaction histories. The system should also distinguish between generative interpretation and authoritative source text, ensuring that users can always locate the underlying legal authority. A governance framework that aligns technical capabilities with professional norms is necessary to build trust and to mitigate the risk that legal semantic search systems become opaque intermediaries in legal reasoning.

8. Future Directions

Future research should examine how large language model-based legal semantic search can better handle cross-jurisdictional analysis, statutory history, and evolving doctrinal frameworks. There is also a need for more robust methods to verify generated legal claims against primary sources. Enhancing user memory systems may allow legal researchers to maintain persistent research threads across sessions while respecting confidentiality

constraints [5]. Domain-specific pretraining and legal instruction tuning may further improve the ability of models to recognize when a prior decision has been limited, distinguished, or overruled. Evaluation benchmarks should be expanded to include more diverse legal systems, multilingual jurisdictions, and lower court decisions. Collaborative efforts among legal scholars, computer scientists, and practicing lawyers will be important for creating evaluation protocols that reflect real professional tasks. As open model ecosystems mature, legal institutions may increasingly deploy locally controlled systems that combine retrieval, generation, and auditing in a unified architecture. The challenge will be to preserve the interpretive richness of legal reasoning while taking advantage of the scalability and pattern recognition capabilities of large language models.

9. Conclusion

Large language model-based semantic search has substantial potential to improve legal document retrieval and case analysis by moving beyond lexical matching toward nuanced semantic understanding. However, the legal domain demands more than accurate retrieval and fluent generation. It requires systems that respect authority, protect confidentiality, operate transparently, and remain accountable to professional and regulatory standards. This article has examined the architectural foundations, retrieval workflows, robustness and fairness concerns, infrastructure constraints, evaluation practices, and governance implications of legal semantic search. A sustainable legal search system must integrate technical components with institutional safeguards, treating explanation and auditability as core requirements rather than optional features. The future of legal information systems lies not in replacing legal judgment but in augmenting it with semantic tools that are technically sophisticated, operationally sound, and normatively responsible.

References

1. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186.
3. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP 2020*, 6769-6781.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuettler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
5. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP 2019*, 3982-3992.
7. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of law school. *Findings of EMNLP 2020*, 2898-2904.

8. Guha, N., Ho, D. E., Krueger, S., et al. (2024). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.
13. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of SIGIR 2020*, 39-48.
14. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT 2021*, 610-623.
15. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Nikolic, B., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Re, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramer, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

16. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of FAccT 2020*, 33-44.
17. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
18. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016-30030.
19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
20. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.