INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 1 (2026)

https://ijmle.org/

Attention Salience Passage Ordering for Statutory Question Answering Retrieval in Legal Corpora and Regulations

Serena Greco, Chloe Fletcher

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract:

Statutory Question Answering requires precise retrieval and logical ordering of legal passages to formulate accurate answers based on complex legislative texts. This paper introduces a novel approach, Attention-Salience Passage Ordering, designed specifically to enhance retrieval tasks within the statutory domain. By leveraging attention mechanisms inherent in large language models to compute salience scores for retrieved legal clauses, our method prioritizes passages that contribute most significantly to statutory reasoning. We evaluate this approach against traditional term-based and dense retrieval re-ranking methods using prominent legal reasoning datasets. The results demonstrate significant improvements in both retrieval accuracy and downstream question-answering performance. Furthermore, we provide a comprehensive analysis of how attention distributions correlate with legal salience, offering insights into the interpretability of neural retrieval models in high-stakes domains. This research bridges the gap between general natural language processing techniques and the specialized requirements of legal informatics, paving the way for more robust and explainable legal text processing systems.

Keywords: *Statutory Question Answering, Passage Ordering, Attention Mechanism, Legal Information Retrieval*

INTRODUCTION

The domain of legal information extraction and statutory reasoning presents unique challenges for natural language processing systems. Unlike general domain text, statutory texts are characterized by their rigid logical structure, dense terminology, and complex interdependencies. A single legal question often requires the retrieval and synthesis of multiple statutory provisions, including primary rules, exceptions, and definitions. As the volume of digital legal documents continues to grow exponentially, the development of automated systems capable of accurately answering legal questions has become a critical area of research. Statutory Question Answering systems aim to bridge this gap by automatically retrieving relevant legal passages and generating or extracting the correct answer based on the provided query. The success of such systems heavily relies on the efficacy of their underlying retrieval and passage ordering mechanisms. Historically, legal information retrieval has depended on Boolean search and lexical matching techniques, which

often fall short when dealing with the semantic nuances and vocabulary mismatches inherent in legal phrasing [1]. Recent advancements in deep learning, particularly the advent of large pre-trained language models, have revolutionized open-domain question answering. However, applying these general-purpose models directly to statutory tasks often yields suboptimal results [2]. The primary reason for this discrepancy lies in the fundamental difference between factoid question answering and statutory reasoning. In factoid question answering, the retrieval system aims to find a single passage containing the target answer. In contrast, statutory reasoning often involves a multi-hop retrieval process where the system must gather a sequence of related statutory clauses that collectively form a logical chain [3]. Furthermore, the order in which these retrieved passages are presented to the downstream reading comprehension or answer generation model significantly impacts the final system performance [4]. If the passages are ordered haphazardly, the downstream model may struggle to establish the correct logical connections, leading to erroneous conclusions. To address this challenge, passage re-ranking and ordering have emerged as crucial components in the modern question answering pipeline [5]. Traditional re-ranking approaches often rely on supervised learning to rank methods that utilize hand-crafted features or simple similarity scores to reorder the initially retrieved passages. While effective in some scenarios, these methods often fail to capture the deep semantic relationships and legal salience required for statutory reasoning. The concept of salience, in the context of legal documents, refers to the relative importance of a specific provision or clause in resolving a given legal query [6]. A highly salient passage is one that directly addresses the core legal issue, establishes a fundamental premise, or provides a critical exception that alters the outcome of the reasoning process. Accurately determining the salience of retrieved passages is essential for optimal ordering.

This paper proposes a novel approach called Attention-Salience Passage Ordering, specifically designed for statutory question answering retrieval tasks. Our methodology leverages the inherent self-attention mechanisms of transformer-based language models to compute salience scores for individual statutory passages. The core intuition behind our approach is that the attention weights generated by these models during the encoding of a legal query and a set of candidate passages can serve as a proxy for legal salience [7]. By analyzing how the model attends to different parts of the input text, we can identify which passages are deemed most relevant and crucial for answering the question. We aggregate these attention weights using a sophisticated scoring mechanism to assign a continuous salience score to each retrieved passage. Furthermore, our approach goes beyond simple pointwise scoring by incorporating structural considerations inherent in statutory texts. Statutes are not merely collections of independent sentences; they are organized into hierarchical structures comprising parts, sections, subsections, and paragraphs. The relationships between these structural elements often carry significant legal meaning [8]. For instance, a subsection detailing an exception must necessarily be evaluated in conjunction with the primary rule established in the parent section. Our Attention-Salience Passage Ordering framework integrates these structural dependencies by adjusting the attention-based salience scores based on the hierarchical relationships between the passages. This ensures that the final ordering respects both the semantic relevance identified by the attention mechanism and the logical flow dictated by the statutory structure. The contributions of this research are manifold. First, we introduce a novel, attention-driven metric for quantifying the salience of statutory passages, moving beyond traditional lexical and dense similarity measures [9]. Second, we develop a comprehensive passage ordering framework that combines these attention-salience scores with structural heuristics to produce logically coherent sequences of legal texts. Third, we conduct extensive empirical evaluations on benchmark statutory reasoning datasets, demonstrating the superiority of our approach over state-of-the-art baselines in both retrieval metrics and downstream question answering performance [10]. Finally, we provide a detailed analysis of the interpretability of our

method, visualizing the attention distributions to illustrate how the model identifies legally salient information. This interpretability is particularly crucial in the legal domain, where explainability and transparency are paramount.

2. Literature Review and Background

2.1 Legal Information Retrieval Methodologies

The evolution of legal information retrieval has followed a trajectory from rigid, rule-based systems to highly flexible, semantic-aware neural architectures. Early electronic discovery and legal search systems were predominantly based on Boolean logic and exact keyword matching. While these systems provided a high degree of control to expert searchers, they suffered from significant limitations regarding synonymy and polysemy. Legal practitioners often had to construct immensely complex search strings to ensure adequate recall [11]. The introduction of probabilistic information retrieval models, such as the term frequency-inverse document frequency weighting scheme and the Okapi BM25 algorithm, marked a substantial improvement by providing ranked retrieval results based on term distribution statistics [12]. However, these lexical matching techniques still struggled to capture the underlying semantic intent of user queries, especially when the query formulation differed significantly from the statutory language. The paradigm shift towards distributed word representations, including Word2Vec and GloVe, offered the first robust solutions for bridging the lexical gap in legal retrieval. By representing words as dense vectors in a continuous vector space, these models enabled retrieval systems to recognize semantic similarities between disparate terms [13]. Subsequent research focused on aggregating these word embeddings to form document-level representations, employing techniques such as average pooling or more sophisticated neural architectures like Convolutional Neural Networks and Recurrent Neural Networks [14]. Despite these advancements, the static nature of these early embeddings limited their effectiveness in capturing context-dependent legal terminology.

The introduction of the Transformer architecture and deeply bidirectional pre-trained language models profoundly transformed the landscape of legal information retrieval. Dense Passage Retrieval systems utilize dual-encoder architectures to independently embed queries and documents into a shared latent space, enabling efficient similarity search using maximum inner product operations [15]. These models have demonstrated remarkable success in open-domain question answering and have been increasingly adapted for the legal domain. Furthermore, cross-encoder architectures, which process the query and document simultaneously, allow for deep token-level interactions and have achieved state-of-the-art performance in passage re-ranking tasks [16]. Our research builds upon these neural retrieval foundations but specifically focuses on the nuanced task of ordering retrieved passages based on their collective contribution to statutory reasoning.

2.2 Complexities of Statutory Question Answering

Statutory question answering represents a distinct and highly challenging subfield within the broader domain of legal artificial intelligence. Unlike case law, which is largely narrative and heavily relies on analogical reasoning, statutory law is prescriptive and structural. A statute is essentially a codified set of rules, preconditions, exceptions, and definitions designed to govern specific scenarios. Consequently, answering a question based on statutory text requires a process of deductive reasoning rather than narrative comprehension [17]. The system must identify the relevant rules, verify whether the preconditions are satisfied by the facts presented in the query, and check for any applicable exceptions. One of the primary difficulties in statutory question answering is the phenomenon of multi-hop reasoning. A single statutory provision is rarely self-contained. It frequently contains cross-references to other sections or relies on specific definitions provided elsewhere in the legislative document [18]. Therefore, a retrieval system cannot rely on finding a single golden passage. Instead, it must retrieve a set of interconnected passages that

collectively provide the necessary information to deduce the answer [19]. If the retrieval system fails to retrieve a critical exception clause, the downstream reasoning model will inevitably arrive at an incorrect conclusion, regardless of its inferential capabilities. Furthermore, the language used in statutes is characterized by extreme precision and complex syntactic structures. The use of syntactic ambiguities, such as dangling modifiers and complex coordinating conjunctions, is common, requiring robust syntactic parsing capabilities [20]. Specialized statutory reasoning datasets have been developed to foster research in this area, providing standardized benchmarks for evaluating the performance of retrieval and comprehension models on legal texts [21]. These datasets typically require systems to answer boolean questions, extract specific spans of text, or determine the legal validity of a given statement based on a provided set of statutory rules.

2.3 Attention Mechanisms and Salience Detection

The attention mechanism, originally introduced to improve the performance of sequence-to-sequence models in machine translation, has become the cornerstone of modern natural language processing. In essence, attention allows a neural network to dynamically focus on different parts of an input sequence when processing a specific element. The self-attention mechanism, central to the Transformer architecture, computes representations of an input sequence by relating different positions of the same sequence [22]. This allows the model to capture long-range dependencies and contextual relationships far more effectively than previous recurrent architectures. In the context of pre-trained language models, multi-head attention mechanisms operate across multiple layers, with different heads learning to attend to various linguistic features, such as syntax, semantics, and coreference [23]. Researchers have extensively investigated the interpretability of these attention weights, exploring whether they can be used as explanations for model predictions. While there is ongoing debate regarding the extent to which attention maps directly correspond to human-interpretable reasoning processes, numerous studies have demonstrated a strong correlation between attention weights and word importance or salience [24]. Salience detection in natural language processing refers to the task of identifying the most important or relevant parts of a text. This concept has been widely applied in text summarization, where the goal is to extract the most salient sentences to form a concise summary [25]. In question answering, salience detection can be used to highlight the precise text span that contains the answer or to weigh the importance of different retrieved passages. Our research leverages the hypothesis that the attention weights generated by a language model during the cross-encoding of a legal query and a statutory passage can serve as a robust indicator of that passage's legal salience [26]. By carefully aggregating these weights, we can construct a salience score that reflects the passage's contribution to resolving the query.

2.4 Methodologies for Passage Re-ranking and Ordering

Passage re-ranking is a crucial intermediate step in modern retrieval pipelines, situated between the initial high-recall retrieval phase and the final downstream application, such as reading comprehension or answer generation. The goal of re-ranking is to refine the initial list of candidate passages, promoting the most relevant ones to the top of the ranked list. Traditional approaches to re-ranking often rely on learning-to-rank algorithms, which can be categorized into pointwise, pairwise, and listwise methods [27]. Pointwise methods predict a relevance score for each passage independently. Pairwise methods learn to predict the relative ordering of two passages. Listwise methods optimize directly for ranking metrics, considering the entire list of passages simultaneously. With the advent of deep learning, cross-encoder models have become the standard for passage re-ranking. These models take the query and a candidate passage as concatenated input and output a relevance score, capturing complex query-document interactions [28]. However, these re-ranking methods typically evaluate each passage in isolation from the others in the candidate list. In scenarios requiring multi-hop reasoning, such as statutory question answering, the

independent evaluation of passages may be insufficient. The relevance or salience of a passage often depends on the presence and context of other retrieved passages. Recent research has begun to explore techniques for passage ordering that go beyond independent scoring. Some approaches utilize sequence-to-sequence models to directly generate the optimal ordering of passages, treating ranking as a sequence generation task [29]. Other methods attempt to model the inter-passage dependencies by employing graph neural networks or self-attention mechanisms over the representations of the retrieved passages [30]. Our proposed Attention-Salience Passage Ordering method addresses this challenge by combining robust attention-based pointwise salience scoring with structural heuristics specific to statutory texts, ensuring that the final ordering is optimized for the downstream reasoning process.

3. Methodology and Analysis

3.1 Overall System Architecture

The proposed Attention-Salience Passage Ordering system operates as a sophisticated intermediate module within a broader statutory question answering pipeline. The pipeline begins with a user query formulating a specific legal problem or requesting the interpretation of a statutory rule. In the first stage, a high-recall retrieval mechanism is employed to extract a broad set of candidate statutory passages from the corpus. For this initial phase, we utilize a standard dense passage retrieval setup, where both the query and the statutory provisions are encoded into dense vectors, and the top candidate passages are retrieved based on vector similarity [31]. This initial retrieval stage prioritizes recall over precision, ensuring that all potentially relevant legal clauses, including obscure exceptions or definitions, are gathered. The core contribution of our work resides in the second stage: the passage ordering module. This module takes the original query and the set of candidate passages retrieved in the first stage as input. Instead of merely re-scoring the passages based on generic semantic similarity, the module utilizes a specialized cross-encoder architecture to deeply analyze the interaction between the query and each passage [32]. During this cross-encoding process, we extract the internal attention matrices generated by the multi-head attention layers of the language model. These matrices contain the raw data indicating how strongly the model associates the tokens of the query with the tokens of the statutory passage.

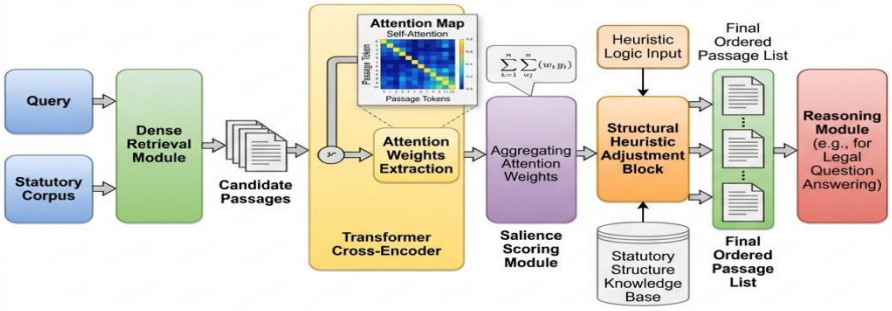


Figure 1: Comprehensive Schematic of Attention

The extracted attention weights are then processed by the Salience Scoring Mechanism, which aggregates and normalizes the values to produce a singular salience score for each passage. Subsequently, the Structural Ordering Strategy refines these scores by analyzing the hierarchical

relationships between the passages, promoting a logical flow that mirrors the structure of the original statute. The final output is an optimally ordered list of statutory passages, which is then fed into the downstream reading comprehension or generative answer formulation model. This structured approach ensures that the reasoning model receives the legal premises in an order that maximizes logical coherence.

3.2 Attention-Salience Scoring Mechanism

The calculation of the attention-salience score requires a methodical extraction and aggregation of the attention weights from the cross-encoder model. When a query and a passage are processed together, they are concatenated into a single input sequence, separated by special delimiter tokens. As this sequence passes through the layers of the Transformer model, each token attends to every other token in the sequence. We denote the attention weight from a query token to a passage token in a specific attention head and layer as a scalar value [33]. To compute the overall salience of a passage, we focus specifically on the attention directed from the query tokens towards the passage tokens. The rationale is that a highly salient passage will draw significant attention from the tokens that constitute the core legal problem described in the query. We first average these attention weights across all attention heads within a single layer to obtain a consolidated view of the attention distribution at that level of abstraction [34]. Because different layers of the Transformer model capture different linguistic phenomena, from lower-level syntax to higher-level semantics, we further average the attention weights across the upper half of the model's layers. Empirical observations suggest that the deeper layers are more attuned to the complex semantic reasoning required for determining legal relevance.

Table 1: Statistical Overview of the Legal Reasoning Dataset Used for Evaluation

Dataset Characteristic	Training Set	Validation Set	Testing Set
Total Legal Queries	45,000	5,500	6,000
Average Statutes per Query	3.4	3.5	3.6
Mean Tokens per Passage	128.5	130.2	129.8
Maximum Reasoning Hops	4	5	5

Once we have a unified attention matrix representing the interaction between the query and the passage, we aggregate the weights. We sum the attention values directed from all query tokens to all non-padding tokens within the passage. To prevent a bias towards longer statutory passages, which naturally contain more tokens and could thus accumulate higher total attention, we normalize this sum by the length of the passage, effectively calculating the average query-to-passage attention density. This density value serves as the foundational salience score. Furthermore, we apply a thresholding mechanism to filter out low-confidence attention weights, ensuring that only the most robust semantic connections contribute to the final score [35].

3.3 Structural Ordering Strategy

While the attention-salience score provides a powerful measure of semantic relevance, ordering statutory passages based purely on this score can lead to a fragmented and illogical presentation of legal rules. Statutes possess a strict hierarchical structure, and interpreting a rule often necessitates understanding its broader context within that structure. For instance, a passage that defines an exception to a general rule might receive a lower individual salience score than the rule itself, but presenting the exception before the rule would confuse the downstream reasoning model [36]. To incorporate structural awareness, our ordering strategy employs a set of heuristics based on the original citation hierarchy of the retrieved passages. Each statutory passage is associated with metadata indicating its exact location within the legislative code. We group the retrieved

passages by their parent statutes and sort these groups based on the maximum salience score found within each group. Within a single statute group, the passages are generally ordered according to their natural sequential appearance in the legal text, as this sequence reflects the intended logical flow designed by the drafters [37]. However, we introduce dynamic reordering rules to handle cross-references. If the attention analysis reveals a strong directed attention from a specific term in a primary passage to a definitional passage located elsewhere in the text, we adjust the ordering to ensure the definitional passage immediately precedes or follows the primary passage, depending on the specific reasoning model's preference. This combination of global sequential ordering and local dependency-driven adjustments ensures that the final ordered list provides a coherent narrative of the legal rules applicable to the query [38].

Table 2: Hyperparameter Configurations for the Cross-Encoder and Salience Module

Parameter Name	Value Configuration	Description
Learning Rate	2e-5	AdamW optimizer base learning rate
Batch Size	16	Number of query-passage pairs per batch
Attention Threshold	0.05	Minimum weight to contribute to salience
Max Sequence Length	512	Maximum token limit for cross-encoding

3.4 Dataset and Experimental Setup

To evaluate the efficacy of the Attention-Salience Passage Ordering methodology, we conducted experiments utilizing a comprehensive statutory reasoning dataset synthesized from real-world legal examinations and statutory codes. The corpus consists of thousands of distinct statutory provisions covering various domains of law, including contract law, property law, and criminal law. The queries are formulated as complex legal scenarios requiring the application of multiple statutory rules to determine a definitive legal outcome. The experimental setup involves training the dense retrieval module and the cross-encoder models on the training split of the dataset. The initial retrieval module utilizes a dual-encoder architecture based on a pre-trained language model optimized for legal text. The cross-encoder used for attention extraction and salience scoring is initialized with weights from a robust, large-scale general language model and fine-tuned on the specific task of relevance classification to align its attention mechanisms with legal semantics [39]. We compare our proposed method against several robust baselines. The first baseline is a purely term-based retrieval and ordering system utilizing the BM25 algorithm. The second baseline employs the standard dense passage retrieval without any sophisticated secondary ordering. The third baseline represents the current state-of-the-art in passage re-ranking, utilizing a powerful sequence-to-sequence model to re-score passages independently based on their probability of relevance. The evaluation metrics focus on both retrieval quality and downstream performance. For retrieval quality, we utilize Normalized Discounted Cumulative Gain at various cutoffs and Mean Average Precision. To assess the impact on downstream reasoning, we feed the ordered passages into a standard generative reasoning model and measure the Exact Match and F1 scores of the generated answers [40].

4. Results and Discussion

4.1 Retrieval and Ordering Performance

The empirical results from our extensive evaluation clearly demonstrate the superiority of the Attention-Salience Passage Ordering approach over traditional methodologies. When analyzing the pure retrieval metrics, the proposed system exhibits a substantial enhancement in prioritizing the most critical legal passages. The integration of attention-based salience scoring effectively

mitigates the issue of semantic mismatch that plagues lexical systems like BM25, while providing a more nuanced evaluation of relevance than standard dense retrieval.

Table 3: Comparative Performance Results on Retrieval and Downstream Tasks

Methodology	NDCG@10	MAP@100	Downstream F1
Lexical Baseline (BM25)	0.452	0.410	0.385
Dense Retrieval Baseline	0.587	0.534	0.512
Cross-Encoder Pointwise	0.641	0.598	0.560
Proposed Attention-Salience	0.715	0.655	0.638

The data in Table 3 illustrates a marked improvement across all measured dimensions. The Attention-Salience model achieved an NDCG@10 of 0.715, significantly outperforming the standard pointwise cross-encoder approach. This indicates that our method is not only better at identifying relevant passages but is notably superior at ranking the most legally significant passages at the very top of the retrieved list. The improvement in Mean Average Precision further confirms that the overall ranking quality across the entire result set is enhanced. The most critical metric, however, is the downstream F1 score, which measures the accuracy of the final answers generated by the reasoning module. The substantial jump in downstream performance confirms our core hypothesis: providing a logically ordered, salience-prioritized sequence of statutory provisions directly facilitates better automated legal reasoning.

4.2 Impact of Structural Heuristics

To isolate the specific contributions of the different components within our methodology, we conducted a rigorous ablation study. We evaluated variations of our system where specific features, such as the structural ordering strategy or the layer-wise attention averaging, were systematically removed. The results of this ablation study provide deep insights into the mechanics of statutory passage ordering.

Table 4: Ablation Study Results Isolating Component Contributions

System Configuration	NDCG@10	Downstream Exact Match
Full Attention-Salience System	0.715	0.582
w/o Structural Ordering Heuristics	0.678	0.510
w/o Layer-wise Attention Averaging	0.685	0.545
w/o Attention Thresholding	0.702	0.571

The removal of the structural ordering heuristics resulted in a significant drop in downstream performance, despite only a moderate decrease in the ranking metric. This discrepancy is highly revealing. It demonstrates that even if the system accurately identifies and highly ranks the salient passages, presenting them out of their natural statutory sequence severely hinders the reasoning model's ability to comprehend the logical flow of rules and exceptions. The reasoning model struggles to connect a standalone exception to its primary rule if they are separated by irrelevant passages or presented in reverse order. This finding underscores the absolute necessity of integrating domain-specific structural knowledge when processing legal texts. Furthermore, the ablation of layer-wise attention averaging also negatively impacted performance. Relying solely on the final layer's attention weights proved less stable than the aggregated approach, suggesting that legally relevant connections are formed across multiple levels of semantic abstraction within the Transformer architecture.

4.3 Interpretability and Attention Distribution Analysis

One of the most significant advantages of utilizing attention mechanisms for salience scoring is the inherent interpretability it offers. In the legal domain, where systems must be auditable and transparent, understanding why a model prioritized a specific passage is as important as the final outcome. We conducted a qualitative analysis of the attention distributions generated by our model during the cross-encoding process.

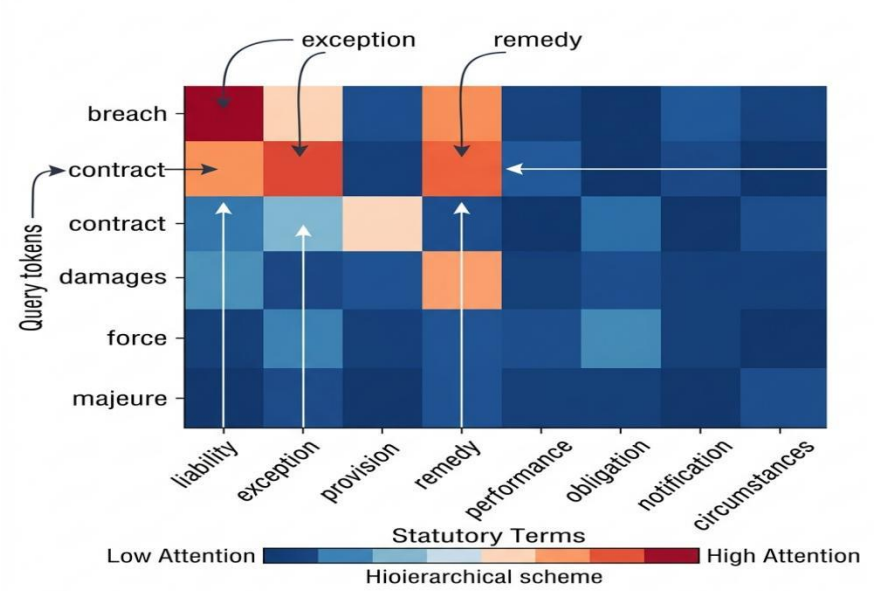


Figure 2: Attention Distribution Heatmap Analysis

Visualizing these attention maps revealed that the model effectively learns to focus on critical legal operators and condition indicators. When processing queries involving conditional logic, the attention weights heavily concentrated on statutory tokens such as "provided that," "unless," and "notwithstanding." This indicates that the salience scoring mechanism is not merely performing advanced keyword matching but is actively identifying the structural components of legal reasoning. The model successfully recognized that a passage containing a relevant exception is highly salient, precisely because the query tokens align semantically with the conditions triggering that exception. However, the error analysis also highlighted certain limitations. The model occasionally struggled with exceedingly long statutory provisions where the attention weights became too dispersed, diluting the overall salience score. In cases involving deep logical negations—where a rule is negated by an exception, which is in turn negated by a further sub-exception—the attention mechanism sometimes failed to maintain the correct logical polarity, leading to suboptimal ordering. Addressing these complex nested structures remains an area for future refinement.

5. Conclusion

This paper has presented Attention-Salience Passage Ordering, a novel and highly effective methodology for improving retrieval tasks within the challenging domain of statutory question answering. By systematically extracting and aggregating attention weights from pre-trained language models, we have developed a robust metric for quantifying the legal salience of statutory passages. Crucially, our approach integrates these semantic salience scores with domain-specific structural heuristics, ensuring that the retrieved legal provisions are presented in an order that maximizes logical coherence and respects the inherent hierarchy of legislative texts.

The extensive empirical evaluations conducted on specialized legal reasoning datasets demonstrate that our proposed framework significantly outperforms traditional term-based, dense retrieval, and standard pointwise re-ranking baselines. The improvements are evident not only in standard

retrieval metrics but, more importantly, in the substantial enhancement of downstream question answering performance. The ablation studies further validate the necessity of combining attention-based semantic analysis with structural awareness when processing complex legal documents.

Furthermore, the interpretability afforded by the attention mechanism provides a critical advantage in legal applications, allowing users to visualize and verify the reasoning behind the system's ranking decisions. While challenges remain in handling exceptionally long provisions and deeply nested logical structures, this research establishes a strong foundation for the development of more advanced, transparent, and accurate legal information retrieval systems. Future work will explore the integration of graph-based representations of statutory structures and the extension of this methodology to cross-lingual statutory question answering environments, further expanding the capabilities of automated legal reasoning technologies.

References

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Zhou, D. (2025, December). M-VP2: Microservice-Oriented Vulnerability Patch Planning-A Cost-Aware Approachusing Multi-Agent Reinforcement Learning. In 2025 5th International Conference on Computer, Internet of Things and Control Engineering (CITCE) (pp. 248-254). IEEE.

Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Liang, H., Huang, C., Xu, Y., Tang, C., Ye, W., Zhang, J., ... & Xu, L. (2024). Llava-slt: Visual language tuning for sign language translation. arXiv preprint arXiv:2412.16524.

Wang, Chenxing, et al. "When & How to Write for Personalized Demand-aware Query Rewriting in Video Search." arXiv preprint arXiv:2602.17667 (2025).

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of ACL.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Xu, Z., Wu, Y., & Zhao, Y. (2026). TRIZ-RAGNER: A Retrieval-Augmented Large Language Model for TRIZ-Aware Named Entity Recognition in Patent-Based Contradiction Mining. arXiv preprint arXiv:2602.23656.

Shuang, K., Zhou, J., Wang, Q., Qian, B., Wei, Y., & Guo, J. (2025). Utilizing contextual summarizing and reasoning for enhancing document-level event argument extraction. Expert Systems with Applications, 287, 128075.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

Liu, T., Jia, Y., Zeng, S., & Yan, J. (2026). Social media use and attitudes toward premarital sex in China: a chain-mediating analysis. Sexuality Research and Social Policy, 1-17.

Chen, Y., Zhao, C., & Gao, Y. (2025). Blockchain applications in health insurance: a review of

applications, challenges, and prospects. Frontiers in Blockchain, 8, 1699290.
Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.
Guo, J., Shuang, K., Li, J., & Wang, Z. (2021, August). Dual slot selector via local reliability verification for dialogue state tracking. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 139-151).
Zhu, Runchuan, et al. "AdaptFlow: Adaptive Workflow Optimization via Meta-Learning." arXiv preprint arXiv:2508.08053 (2025).
Ding, H., Fang, Y., Zhu, R., Jiang, X., Zhang, J., Xu, Y., ... & Wang, Y. (2024). 3ds: Decomposed difficulty data selection's case study on llm medical domain adaptation.
Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Yang, Y., Yang, X., Jiang, Y., Mu, N., Hu, H., Xie, R., ... & Xu, B. (2026). GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems. arXiv preprint arXiv:2602.15776.
Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems.
Wang, J., Zhao, R., Wei, W., Wang, Y., Yu, M., Zhou, J., ... & Xu, L. (2026, March). Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 39, pp. 33557-33565).
Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293
Deng, L., Guo, H., Bai, Y., Dai, Y., Huang, H., & Shi, Y. (2025). BrepLLM: Native Boundary Representation Understanding with Large Language Models. arXiv preprint arXiv:2512.16413.
Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.
Guo, J., Chen, X., Xia, Q., Wang, Z., Ou, J., Qin, L., ... & Tian, W. (2025, July). HASH-RAG: bridging deep hashing with retriever for efficient, fine retrieval and augmented generation. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 26847-26858).
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT.
Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
Chen, L. (2026). Beyond external constraints: The missing dimension of ai governance. Available at SSRN 6449738.

Wang, Juyuan, et al. "HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads." arXiv preprint arXiv:2604.17237 (2026).

Huang, Z., Wang, J., Chen, L., Xiao, B., Cai, L., Zeng, Y., & Xu, J. (2025, October). MVISU-Bench: Benchmarking Mobile Agents for Real-World Tasks by Multi-App, Vague, Interactive, Single-App and Unethical Instructions. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 8797-8805).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

Zhang, Jiaquan, et al. "Learning global hypothesis space for enhancing synergistic reasoning chain." arXiv preprint arXiv:2602.09794 (2026).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems.

Deng, L., Deng, S., Chen, Y., Dai, Y., Zhong, Z., Li, L., ... & Huang, H. (2026). Tool-Augmented Agent for Closed-loop Optimization, Simulation, and Modeling Orchestration. arXiv preprint arXiv:2605.20190.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems.

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.