

Asymmetric Effects of Cross-Segment Contextualization in Long Legal Document Classification: An Empirical Study on ECtHR Violation Classification

Meng-Luen Wu

158769@mail.tku.edu.tw

Tamkang University

Chao-Wei Liu

Tamkang University

Research Article

Keywords: Long legal document processing, legal text classification, cross-segment contextualization, overlapping segmentation, attention-based reweighting, ECtHR

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10632286/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Asymmetric Effects of Cross-Segment Contextualization in Long Legal Document Classification: An Empirical Study on ECtHR Violation Classification

Abstract. Long legal texts require information from distant parts of a document to be considered together. Overlapping segmentation, cross-segment propagation, and attention-based weighting address this requirement, but their effects depend on how repeated and contextualized information is aggregated.

This study examines these mechanisms through a controlled empirical analysis of binary violation classification on the European Court of Human Rights corpus. Legal-BERT independently encodes each segment, a bidirectional gated recurrent unit (BiGRU) provides cross-segment contextualization, and additive attention performs selective aggregation.

The central finding is that greater contextual connectivity is not uniformly beneficial, and the effect of sequential contextualization depends on the aggregation strategy. Among the evaluated overlap settings, 100 tokens yield the highest macro-precision and macro-recall, whereas larger overlaps reduce both metrics. Under equal-weight aggregation, BiGRU increases macro-recall from 80.4% to 81.3% but reduces macro-precision from 90.2% to 84.5%, lowering macro-F1 from 83.0% to 82.5%. Attention does not improve macro-F1 when applied to independently encoded segments, with the score decreasing from 83.0% to 82.8%. In contrast, when attention follows BiGRU contextualization, macro-F1 rises to 84.6%.

These results reveal configuration-dependent effects rather than uniformly beneficial components. Sequential contextualization produces a pronounced precision-recall trade-off under uniform aggregation, while selective aggregation is associated with recovery in macro-F1 after contextualization. The findings also motivate a distinction between the computational treatment of repeated and propagated text and its legal evidential role.

Keywords: Long legal document processing; legal text classification; cross-segment contextualization; overlapping segmentation; attention-based reweighting; ECtHR

1 Introduction

Automated legal text analysis and long-document classification require models to process extensive and structurally complex narratives. Judicial decisions are particularly challenging because task-relevant information may be distributed across factual narratives, procedural histories, party submissions, and legally significant events. A classification model must therefore integrate information across multiple parts of a document rather than rely solely on localized semantic cues.

Expanding the nominal context window does not by itself ensure reliable use of all input information. Long-context models may exhibit position-dependent degradation and reduced sensitivity to relevant information located within lengthy inputs (Liu et al.

2024). Moreover, recent memory-augmented architectures continue to process long sequences through local blocks and compressed memory mechanisms (Munkhdalai et al. 2024). Segmentation-based processing therefore remains relevant when documents exceed available context windows, computational resources are constrained, or explicit evidence aggregation is required.

To operationalize this necessity of chunking and cross-referencing, both conventional hierarchical classifiers and modern long-document systems rely on three integration mechanisms, as illustrated in Fig. 1. First, overlapping segmentation duplicates content near boundaries to preserve local continuity and prevent the artificial truncation of narratives. Second, sequential contextualization propagates information across segment representations through recurrent networks, hierarchical Transformers, or compressed memory states. Third, additive attention-based reweighting adjusts the relative contribution of individual segments to the final document representation. While operating at different stages of the processing pipeline, their effects may depend on how these mechanisms are combined. Overlap determines which textual information is repeatedly encoded, sequential modeling determines how information is propagated, and attention determines how the resulting representations contribute to classification. These mechanisms recur across hierarchical and segmentation-based long-document architectures, although their specific implementations differ across models (Yang et al. 2016; Pappagari et al. 2019; Dai et al. 2022).

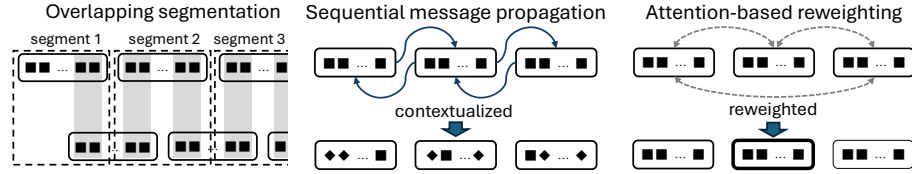


Fig. 1 Three recurring mechanisms in segmentation-based long-document processing. Overlapping segmentation preserves boundary continuity, sequential contextualization propagates information across adjacent segment representations, and additive attention-based reweighting adjusts their contributions to the final document representation

In legal text analysis, these operations raise questions that extend beyond computational coverage or efficiency. From an evidential reasoning perspective, repeated presentation of the same underlying information does not create independent corroboration and may lead to double-counting when dependent observations are treated as distinct evidence (Shafer 2016; Strittmatter et al. 2024). Yet, at the architectural level, overlapping windows mechanically present the same passage to the classifier through multiple segment-level inputs. Similarly, legal reasoning is defeasible rather than purely accumulative (Prakken and Sartor 1996; Hage 2003). A later passage may support an earlier allegation, but it may also qualify, rebut, justify, or legally recharacterize it. A generic sequential encoder propagates information without explicitly representing these argumentative functions. Finally, attention weights indicate learned computational salience but should not be treated as direct explanations of feature importance (Jain and Wallace 2019), much less as direct indicators of legal relevance. A passage may receive a high weight simply because it contains class-correlated vocabulary, even though it does not independently satisfy a legal requirement.

In this study, these evidential concepts are used as a conceptual lens for examining model behavior. The experiments measure changes in computational treatment and classification performance; they do not measure the legal evidential force of individual passages or establish that the model reproduces legal evidential reasoning.

These architectural concerns are particularly pronounced in applied legal document classification. Prior work has shown that classifiers trained on the European Court of Human Rights (ECtHR) corpus may exploit shallow signals such as document length, respondent-state information, paragraph-number artifacts, and vocabulary strongly correlated with violation outcomes but with limited substantive legal relevance (Santosh et al. 2022). In a segmentation-based architecture, overlapping boundaries can expose the same shallow signal through multiple segment inputs, while sequential contextualization allows information associated with such signals to be transmitted across segment representations. These mechanisms therefore create the possibility that repeated computational exposure gives the same underlying information greater computational influence on classification without any corresponding increase in legal evidential force. To illustrate this vulnerability, Fig. 2 presents a hypothetical case concerning Article 6 of the European Convention on Human Rights (right to a fair trial). The procedural safeguards described in the first paragraph support a finding of no violation, whereas the second paragraph contains superficial, violation-associated expressions without actual evidential support. Without selective aggregation, a classification model may assign disproportionate computational influence to the class-correlated expressions in the latter paragraph, despite their lack of additional evidential support.

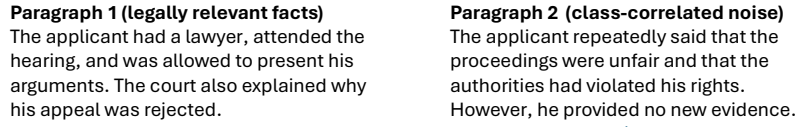


Fig. 2 Hypothetical case concerning the right to a fair trial under Article 6 of the European Convention on Human Rights. Although the legally relevant facts support a finding of no violation, the example illustrates how class-correlated expressions may acquire greater computational influence through cross-segment propagation despite providing no additional evidential support

Despite these vulnerabilities, existing long-document studies have mainly evaluated overlap, sequential contextualization, and attention according to aggregate accuracy or F1-score. Although prior studies have examined these mechanisms individually or as components of long-document architectures, their configuration-dependent interactions have received considerably less attention. Overlapping segmentation has been shown to mitigate boundary fragmentation, but its benefits depend on segment length and dataset characteristics rather than increasing monotonically (Dai et al. 2022). Crucially, less attention has been paid to how these mechanisms interact or whether they produce uneven changes in macro-precision and macro-recall. By relying on single aggregate scores, prior research may obscure the precise tradeoffs that matter most in legal reasoning. It remains unclear whether cross-segment contextualization improves sensitivity to distributed information at the cost of classification selectivity. The present study provides a follow-up analysis of our earlier deployment-oriented framework (Wu and

Liu 2026), re-examining previously reported experimental configurations and results from the perspective of configuration-dependent classification behavior.

To address this gap, this study examines the configuration-dependent effects of these mechanisms, including the interaction between sequential contextualization and attention-based aggregation. The analytical architecture combines independent Legal-BERT segment encoding with three controlled mechanisms: boundary overlap, sequential contextualization via a Bidirectional Gated Recurrent Unit (BiGRU), and selective aggregation via additive attention. Controlled ablation experiments are used to examine the effects of these mechanisms on macro-precision, macro-recall, and macro-F1, as well as the interaction between sequential contextualization and attention-based re-weighting.

The empirical results reveal three main observations. First, moderate boundary overlap provides the best observed balance in the examined configuration, whereas larger overlaps reduce both macro-precision and macro-recall. Second, under uniform mean pooling, BiGRU-based contextualization produces a pronounced asymmetric effect on the two metrics. It increases macro-recall from 80.4% to 81.3% but reduces macro-precision from 90.2% to 84.5%, showing that a modest recall gain is accompanied by a substantially larger precision loss. Third, in the evaluated configuration, attention does not improve macro-F1 when applied to independently encoded segments, with the score decreasing slightly from 83.0% to 82.8%. In contrast, attention increases macro-F1 from 82.5% to 84.6% when applied after BiGRU contextualization.

Accordingly, the study addresses three research questions:

RQ1: Under a fixed segment length, how does overlap size affect macro-precision and macro-recall across the evaluated settings?

RQ2: Does BiGRU-based sequential contextualization change the balance between macro-precision and macro-recall under uniform aggregation?

RQ3: Does the contribution of additive attention differ depending on whether it aggregates independently encoded or context-conditioned segment representations?

The contribution of this study is threefold. First, it provides a controlled, mechanism-oriented analysis of overlap, sequential contextualization, and selective aggregation in long legal document classification. Second, it identifies a pronounced precision-recall asymmetry under BiGRU-based contextualization and a conditional interaction between sequential contextualization and attention-based aggregation. Third, it uses these observations to distinguish the computational treatment of repeated and propagated text from the legal evidential force of the underlying information. The study does not claim that the examined architecture reproduces legal evidential reasoning; rather, the legal perspective provides a framework for identifying potential mismatches between computational integration and evidential evaluation.

These findings also have conceptual implications for other long-context architectures that rely on chunking, recurrent memory, selective retrieval, or attention-based aggregation. The central issue is therefore not whether a model can technically access more text, but whether repeated and propagated information is treated in a manner consistent with its evidential role. These results motivate further investigation of whether long legal document models could benefit from mechanisms that account for the common

origin of duplicated passages, distinguish supportive information from defeaters, and avoid equating statistical salience with legal significance.

The remainder of this paper is organized as follows. Section 2 reviews overlapping segmentation, cross-segment contextualization, class-correlated signals in ECtHR classification, and context-conditioned attention. Section 3 presents the analytical framework and experimental architecture. Section 4 reports the ablation results and analyzes their metric-level effects and component interactions. Section 5 concludes the paper and discusses implications for the design of classification models for long legal documents.

2 Related Work

Long legal document classification is not merely a problem of extending the input capacity of pretrained language models. Segmenting and aggregating a judgment also changes how factual allegations, procedural events, supporting evidence, and countervailing information are represented. In legal reasoning, the significance of a passage depends not only on its textual content but also on its relationship to the legal issue and to other parts of the case narrative. Repeated computational exposure to the same text should be distinguished from independent evidential support, while additional context may corroborate, qualify, or defeat an earlier inference. These characteristics make the interaction among overlapping segmentation, cross-segment contextualization, and attention-based aggregation particularly important in legal document classification.

2.1 Overlapping Segmentation and Cumulative Evidence

Pretrained language models such as BERT impose a fixed input-length limit (Devlin et al. 2019). Long-document classifiers therefore commonly divide a document into shorter segments and integrate the resulting representations at the document level. Pappagari et al. (2019), for example, independently encoded document segments using BERT and subsequently processed them through recurrent or Transformer-based aggregation layers.

Fixed-length segmentation may separate semantically related content across adjacent segments. Overlapping windows are commonly used to mitigate this boundary fragmentation by presenting tokens near a boundary in more than one input segment. However, the effectiveness of overlap is not monotonic. Dai et al. (2022) found that the contribution of overlapping segmentation depended on segment length and dataset characteristics. In their ECtHR experiments, overlap produced limited or negative gains when sufficiently long segments were used. Fama et al. (2024) similarly reported that increasing overlap did not consistently improve legal document classification and also increased computational cost. Han et al. (2024) further showed that long-document classification performance varies substantially across document lengths, motivating length-aware representation strategies.

These findings motivate examining a trade-off between contextual continuity and repeated computational exposure. Moderate overlap may preserve locally fragmented information, whereas excessive overlap repeatedly encodes the same passage. From an

evidential perspective, repeated presentation of the same underlying information does not constitute independent corroboration and may lead to double-counting when dependent observations are treated as distinct evidence (Shafer 2016; Strittmatter et al. 2024). A computational aggregation mechanism, however, may receive overlapping segments as distinct observations unless their shared origin is explicitly represented. Consequently, overlapping segmentation exposes some underlying text through multiple segment representations, although whether this repeated exposure increases its computational influence on classification requires direct investigation.

This possibility is particularly relevant when the repeated passage contains an outcome-correlated but legally non-decisive expression. Overlap may increase the numerical exposure of such a cue without increasing its evidential relevance. Existing studies have largely evaluated overlap through aggregate performance and computational efficiency. They have paid less attention to whether repeated encoding produces uneven changes in precision and recall or alters the effectiveness of downstream aggregation.

2.2 Cross-Segment Contextualization and Defeasible Legal Reasoning

Independent segment encoding cannot directly represent relations across distant parts of a document. Hierarchical models therefore introduce recurrent networks or cross-segment Transformers to update each local representation using information from other segments. Pappagari et al. (2019) applied recurrent and Transformer layers to BERT segment representations, while Grail et al. (2021) introduced a hierarchical propagation mechanism that transmitted information across independently processed windows.

Chalkidis et al. (2022) examined different placements of cross-segment contextualization in hierarchical attention Transformers. Their results generally supported architectures that repeatedly integrated local and global information. Such findings demonstrate that cross-segment modeling can improve access to distributed information. They do not, however, establish that all propagated information is legally relevant.

A recurrent encoder performs information propagation according to learned statistical dependencies. It does not inherently distinguish a legally probative fact from a procedural description, repeated allegation, or corpus-specific lexical cue. This limitation is important because legal reasoning is not purely accumulative. A later passage may corroborate an earlier allegation, but it may also qualify, rebut, or legally recharacterize it. Legal conclusions are therefore defeasible and may be revised when exceptions, counterarguments, or additional conditions are introduced (Prakken and Sartor 1996; Hage 2003).

For example, being questioned or tried without the assistance of a lawyer may initially support an inference that the applicant’s rights under Article 6 were violated. However, a later passage stating that the applicant had been informed of the right to a lawyer and had voluntarily and unequivocally waived that right may qualify or defeat the initial inference. Cross-segment contextualization must therefore account not only for cumulative support, but also for information that changes the legal significance of an earlier fact.

A standard BiGRU does not explicitly encode these argumentative functions. Supporting information and defeating information are transmitted through the same hidden-state mechanism. Although the learned model may capture some distinctions implicitly,

the architecture does not impose a legal structure separating support, rebuttal, justification, and exception.

The interaction between overlap and recurrent contextualization may further intensify this issue. Overlap first places the same textual cue into multiple local segments. The BiGRU then propagates information from these partially redundant representations through the segment sequence. A single underlying passage can therefore be represented in multiple local inputs while information derived from these representations can also be transmitted through recurrent hidden states.

This mechanism does not imply a class-specific directional effect. However, it provides a plausible mechanism through which information associated with duplicated or class-correlated cues can be represented across a broader portion of the segment sequence. Existing long-document studies generally report overall accuracy or F1-score and rarely examine whether sequential propagation produces balanced changes in precision and recall.

2.3 Statistical Correlation and Evidential Relevance in ECtHR Classification

Legal judgment classification models may achieve high performance without relying exclusively on legally relevant considerations. Statistical relevance should therefore be distinguished from evidential relevance. A textual feature may correlate statistically with a violation outcome without materially supporting a constituent legal requirement. Its legal significance depends on its relationship to the applicable legal test, available justifications, evidential reliability, and countervailing considerations.

Santosh et al. (2022) demonstrated this problem in the context of ECtHR violation classification. Their analysis showed that classifiers could exploit document length, respondent-state information, paragraph-number artifacts, and vocabulary strongly correlated with violation outcomes but with limited substantive legal relevance. Their findings indicate that ECtHR fact descriptions contain both legally probative information and shallow class-correlated signals. The present study does not attempt to determine the legal evidential force of individual passages; this distinction is used to interpret the computational consequences of segmentation and aggregation.

This distinction has direct implications for hierarchical long-document architecture. When an outcome-correlated cue appears near a segment boundary, overlap can reproduce it in multiple model inputs. Cross-segment contextualization can then extend its influence on adjacent hidden states. These mechanisms therefore increase the number of computational pathways through which a cue can enter the document representation, without implying any corresponding increase in its legal evidential force.

The repeated presence of allegations is especially difficult to interpret computationally. In legal analysis, the frequency with which an allegation appears does not independently establish its truth or legal sufficiency. Multiple references may describe the same event, repeat a party's submission, or reflect the procedural structure of the judgment. A classifier may nevertheless treat such repetition as additional evidence because it is optimized to identify statistical associations with the target class.

Accordingly, the central empirical question is whether broader contextual propagation produces a balanced improvement in classification performance or an uneven response across macro-precision and macro-recall. Controlled comparisons of these metrics can

directly reveal whether increased contextual coverage improves sensitivity at a disproportionate cost to classification selectivity.

2.4 Context-Conditioned Attention and Legal Salience

Attention mechanisms assign different weights to document units because not all segments contribute equally to the final classification. However, attention-based computational salience should not be assumed to reflect feature importance faithfully (Jain and Wallace 2019), and computational salience should therefore be distinguished from legal importance. A passage may be lexically prominent or strongly associated with the predicted class while remaining legally insufficient. Conversely, a short and linguistically neutral passage may contain a decisive condition, exception, or procedural fact.

The hierarchical attention network proposed by Yang et al. (2016) illustrates the relationship between contextualization and attention. In that architecture, sentence representations are first processed by a bidirectional GRU. Attention is then applied to the resulting context-conditioned hidden states. The attention mechanism therefore selects among representations that already incorporate information from preceding and subsequent sentences.

The same distinction applies to segment-level legal classification. When additive attention is applied directly to independently encoded Legal-BERT segments, it can emphasize segments containing strong local features. However, it does not explicitly represent whether one segment corroborates, qualifies, or contradicts another. Its weighting decision is based primarily on each segment’s local representation and a learned attention context.

When attention follows a BiGRU, each hidden state has already been conditioned on the wider segment sequence. The attention layer can then assign weights to representations that encode both local content and inter-segment relations. BiGRU and attention therefore perform complementary rather than interchangeable functions. The BiGRU constructs context-sensitive segment representations, whereas attention determines their relative contribution to the document representation.

Removing the BiGRU does not make attention mathematically inoperative. It removes explicit sequential contextualization from the attention inputs. This distinction provides one possible explanation for the observed difference between attention applied to independently encoded segments and attention applied after cross-segment contextualization.

Previous research also indicates that attention is not independently beneficial under all architectural conditions. Adhikari et al. (2019) showed that competitive document classification could be achieved without attention, while Dai et al. (2022) found that expanding global attention coverage did not consistently improve long-document classification. These findings suggest that the contribution of attention depends on the representational structure established before weighting is performed.

3 Experimental Setup for Component Analysis

This section describes the experimental architecture used to isolate the effects of three common mechanisms in segmentation-based long-document processing: overlapping segmentation, BiGRU-based sequential contextualization, and additive attention-based reweighting. All three mechanisms are established techniques and are not presented as novel model components. The architecture serves as a controlled analytical testbed for examining whether these mechanisms produce balanced changes in macro-precision, macro-recall, and macro-F1, and how their effects depend on downstream aggregation. The processing pipeline consists of fixed-length segmentation, independent Legal-BERT encoding, optional BiGRU contextualization, optional additive attention, and binary classification. Fig. 3 presents the overall pipeline. Components shown as optional are activated or removed according to the ablation configuration.

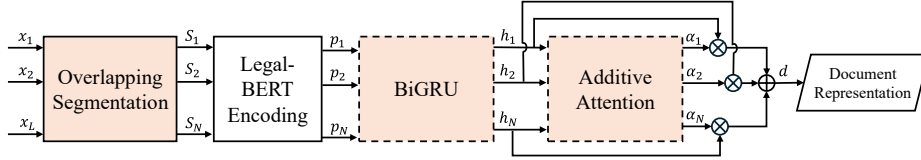


Fig. 3 Controlled analytical pipeline for the component analysis. The input token sequence $x_1, \dots, x_L$ is divided into overlapping segments $S_1, \dots, S_N$, which are independently encoded by Legal-BERT to produce segment representations $p_1, \dots, p_N$. Optional BiGRU-based sequential contextualization generates context-conditioned representations $h_1, \dots, h_N$, and optional additive attention assigns normalized weights $\alpha_1, \dots, \alpha_N$ to construct the document representation d . Dashed outlines indicate optional components used in the ablation experiments, and the shaded modules denote the mechanisms examined in this study. When attention is disabled, uniform mean pooling is used instead

3.1 Controlled Analysis Design

Let a legal document be denoted by D . After preprocessing and tokenization, the document is represented as a token sequence

$$D = (x_1, x_2, \dots, x_L) \quad (1)$$

where L denotes the document length. The sequence is divided into N fixed-length segments. Each segment is independently encoded by Legal-BERT to obtain a local representation. BiGRU and additive attention are then independently enabled or disabled to create the four controlled configurations listed in Table 1.

Table 1 Controlled configurations for component analysis

Configuration	Sequential contextualization	Segment aggregation
Independent encoding	None	Uniform averaging
Attention only	None	Additive attention
BiGRU only	BiGRU	Uniform averaging
BiGRU with attention	BiGRU	Additive attention

These configurations support three direct comparisons. First, comparing independent encoding with attention only evaluates whether additive attention is useful without explicit cross-segment contextualization. Second, comparing independent encoding with BiGRU only evaluates the effect of sequential message propagation without differential weighting. Third, comparing BiGRU only with BiGRU plus attention evaluates whether attention becomes more effective after segment representations have been contextually conditioned.

Overlap is treated as a separate input-level factor. Its size is first varied with BiGRU contextualization and additive attention enabled, after which the selected overlap setting is held fixed for the subsequent component ablations.

3.2 Tokenization and Overlapping Segmentation

Tokenization follows the WordPiece tokenizer used by BERT and its legal-domain adaptation, Legal-BERT (Devlin et al. 2019; Chalkidis et al. 2020). Because BERT-based encoders accept at most 512 input tokens, each segment contains up to 510 document tokens, leaving two positions for the [CLS] and [SEP] tokens.

For an overlap size δ , the starting position of segment S_k is defined as

$$a_k = 1 + (k - 1)(M - \delta) \quad (2)$$

and its ending position is

$$b_k = \min(a_k + M - 1, L) \quad (3)$$

The k -th segment is therefore

$$S_k = (x_{a_k}, x_{a_k+1}, \dots, x_{b_k}) \quad (4)$$

When $\delta = 0$, adjacent segments are disjoint. When $\delta > 0$, boundary tokens are repeated in adjacent segments. This repetition may preserve locally fragmented context, but it also causes the same textual content to be encoded more than once.

The units produced by this procedure are referred to as segments, rather than paragraphs, because they are determined by token length and do not necessarily correspond to the original paragraph boundaries of the judgment.

The overlap analysis examines how different amounts of repeated boundary content are associated with macro-precision and macro-recall. Overlap sizes of 0, 50, 100, 150, and 200 tokens are evaluated. The 100-token setting, which yields the highest observed performance among these settings, is used in the subsequent component analyses.

3.3 Independent Segment Encoding

Each segment is independently encoded using Legal-BERT (Chalkidis et al. 2020). Special tokens are added before encoding:

$$\tilde{S}_k = ([CLS], S_k, [SEP]) \quad (5)$$

Legal-BERT produces a sequence of contextualized token representations for each segment. The final hidden state corresponding to the [CLS] token is used as the segment-level representation (Devlin et al. 2019):

$$p_k = \text{LegalBERT}(\tilde{S}_k)_{[CLS]} \quad (6)$$

where $p_k \in \mathbb{R}^d$, and d is the encoder hidden size.

At this stage, each segment is processed independently. Consequently, p_k contains contextual information within segment S_k , but it does not explicitly incorporate information from preceding or subsequent segments. The sequence

$$P = (p_1, p_2, \dots, p_N) \quad (7)$$

provides the common input to the subsequent controlled configurations.

3.4 BiGRU-based Sequential Contextualization

To evaluate the effect of cross-segment message propagation, a bidirectional gated recurrent unit is optionally applied to the sequence of independently encoded segment representations (Cho et al. 2014; Schuster and Paliwal 1997).

The forward and backward hidden states are computed as

$$\begin{aligned} \vec{h}_k &= \text{GRU}_f(p_k, \vec{h}_{k-1}), \\ \overleftarrow{h}_k &= \text{GRU}_b(p_k, \overleftarrow{h}_{k+1}) \end{aligned} \quad (8)$$

The context-conditioned representation of segment k is

$$h_k = [\vec{h}_k; \overleftarrow{h}_k] \quad (9)$$

Through this operation, each representation can incorporate information transmitted from preceding and subsequent segments. Direct interaction occurs between neighboring recurrent states, while information may indirectly propagate across more distant positions through repeated state transitions.

The BiGRU is used here as a standard sequential contextualization mechanism. It is not assumed to distinguish legally probative evidence from procedural descriptions, repeated allegations, or class-correlated lexical cues. The empirical analysis therefore examines whether its information propagation produces balanced performance gains or an asymmetric change in macro-precision and macro-recall.

When the BiGRU is disabled, the independently encoded vectors p_k are directly passed to the aggregation stage.

3.5 Additive Attention-Based Reweighting

Additive attention is optionally used to assign different weights to the segment representations (Bahdanau et al. 2015; Yang et al. 2016). Let r_k denote the representation received by the aggregation stage:

$$r_k = \begin{cases} h_k, & \text{when BiGRU is enabled} \\ p_k, & \text{when BiGRU is disabled} \end{cases} \quad (10)$$

Each representation is projected into an attention space:

$$u_k = \tanh(W_a r_k + b_a) \quad (11)$$

where W_a and b_a are trainable parameters. A shared trainable context vector u_w is then used to compute a normalized attention weight:

$$\alpha_k = \frac{\exp(u_k^\top u_w)}{\sum_{j=1}^N \exp(u_j^\top u_w)} \quad (12)$$

The document representation is obtained as

$$d = \sum_{k=1}^N \alpha_k r_k \quad (13)$$

Additive attention reweights segment representations during document-level aggregation. The same scoring function is applied to every segment, and normalization determines each segment's relative contribution to the document representation.

The attention weights are task-dependent computational weights and should not be interpreted as faithful explanations of feature importance (Jain and Wallace 2019). Their computational salience should also be distinguished from substantive legal relevance. A segment may receive a high weight because its representation is strongly associated with the target label, even when its substantive legal significance is limited.

When attention is disabled, uniform aggregation is used:

$$d = \frac{1}{N} \sum_{k=1}^N r_k \quad (14)$$

This design allows the attention module to be evaluated against an explicit equal-weight baseline.

3.6 Binary Classification

The document representation d is passed to a fully connected binary classifier:

$$\hat{y} = \sigma(w_c^\top d + b_c) \quad (15)$$

where w_c and b_c are trainable parameters, and σ is the sigmoid function. The output $\hat{y} \in [0,1]$ represents the classified probability of violation.

The class label is

$$\hat{c} = \begin{cases} 1, & \hat{y} \geq 0.5 \\ 0, & \hat{y} < 0.5 \end{cases} \quad (16)$$

where 1 denotes violation and 0 denotes non-violation.

The model is optimized using binary cross-entropy:

$$\mathcal{L} = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})] \quad (17)$$

Each controlled configuration is trained separately under the same data split, optimization procedure, and model-selection protocol. This design reduces variation arising from data splits, preprocessing, and optimization settings, thereby supporting controlled comparisons of the activated mechanisms.

3.7 Analytical Objective

The architecture serves as a controlled testbed rather than a proposed long-document classification model. The analysis focuses on macro-precision, macro-recall, and macro-F1 to distinguish aggregate performance changes from asymmetric responses between precision and recall. In this study, asymmetry refers specifically to differences in the direction or magnitude of changes in macro-precision and macro-recall. The experiments evaluate observable classification behavior and component interactions; they do not directly measure the propagation of specific lexical cues or the legal evidential force of individual passages.

4 Empirical Analysis and Discussion

This section presents a controlled empirical analysis of three mechanisms commonly used in segmentation-based long legal document classification: overlapping segmentation, BiGRU-based sequential contextualization, and additive attention-based reweighting. Rather than comparing the controlled analytical pipeline with external baseline models, the analysis focuses on how these mechanisms individually and jointly affect classification behavior under the same dataset, preprocessing procedure, encoder, and training protocol.

The evaluation first examines whether increasing segment overlaps improve contextual continuity or introduces excessive redundancy. It then analyzes the asymmetric effects of BiGRU-based contextualization on macro-precision and macro-recall under uniform aggregation. Finally, the interaction between BiGRU and attention is evaluated to determine whether attention is effective independently or primarily when applied to context-conditioned segment representations. The discussion focuses on whether these mechanisms produce balanced or asymmetric changes in macro-precision, macro-recall, and macro-F1. All reported results are averaged over five experimental runs.

4.1 Experimental Setup and Evaluation Protocol

The controlled analytical pipeline is evaluated on the English-language ECtHR corpus introduced by Chalkidis et al. (2019). Following the binary violation setting, only factual descriptions are used as model input, and each document is classified according to whether a Convention article or protocol was violated.

Each document is divided into segments containing up to 510 document tokens, with two additional positions reserved for the [CLS] and [SEP] tokens. The overlap analysis evaluates sizes of 0, 50, 100, 150, and 200 tokens. The 100-token setting, which yields the highest observed performance among the evaluated settings, is subsequently used for the BiGRU and attention analyses. The resulting component comparisons are therefore interpreted as exploratory analyses under this selected overlap condition. Legal-BERT produces 768-dimensional segment representations. The BiGRU uses 512 hidden units in each direction, producing 1024-dimensional context-conditioned representations, and the additive attention hidden dimension is also set to 1024. All configurations are trained using AdamW (Loshchilov and Hutter 2019) with an initial learning

rate of 1×10^{-5} , a dropout rate of 0.35, a maximum of 10 epochs, and a linear warmup-and-decay schedule.

Performance is evaluated using precision, recall, and F1-score. Taking violation as the positive class, precision measures how many documents classified as violations are actually violation cases. It decreases when more non-violation documents are incorrectly assigned the violation label. Recall measures on how many actual violation documents are successfully identified. Unlike precision, recall is not directly reduced by false-positive violation assignments. Therefore, stronger sensitivity to violation-associated signals may increase recall, but it may simultaneously reduce precision if the classifier also assigns the violation label to additional non-violation documents. Examining precision and recall together reveals whether a mechanism improves classification in a balanced manner or gains sensitivity at the cost of additional false positives. F1-score summarizes the two measures but may conceal substantial differences in their individual changes.

To give equal weight to the violation and non-violation classes, precision, recall, and F1-score are computed separately for each class and then macro-averaged (Chalkidis et al. 2019; Sokolova and Lapalme 2009). Unless otherwise specified, the reported values refer to macro-averaged metrics. In this study, an asymmetric effect refers specifically to precision and recall changing in different directions or by substantially different magnitudes. All configurations use the same predefined data splits, preprocessing procedure, and optimization settings. Each configuration is executed five times, and the reported values represent the mean performance across the five runs.

4.2 Effect of Overlap Size on Classification Behavior

Fig. 4 presents the effects of overlap size on macro-precision and macro-recall. Increasing the overlap from 0 to 100 tokens improves both metrics. Macro-precision increases from 88.7% to 89.4%, while macro-recall increases from 80.6% to 82.4%. The highest observed values for both metrics are therefore obtained with an overlap of 100 tokens. This pattern is consistent with moderate overlap helping to preserve contextual information near segment boundaries.

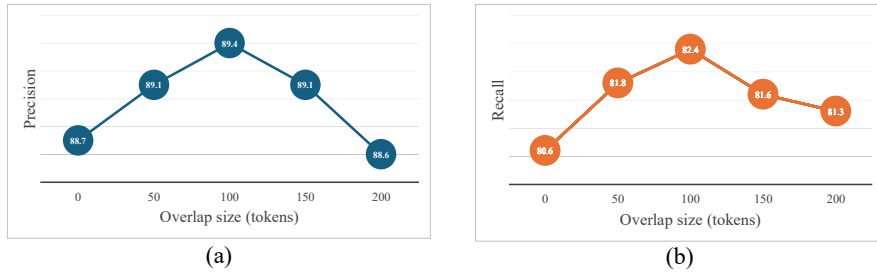


Fig. 4 Effects of overlap size on macro-precision and macro-recall across the evaluated settings

Increasing the overlap beyond 100 tokens does not provide further improvement. At an overlap of 150 tokens, macro-precision and macro-recall decrease to 89.1% and 81.6%, respectively. At 200 tokens, they further decrease to 88.6% and 81.3%. Larger overlaps increase repeated textual exposure without producing a corresponding performance gain in the evaluated settings. Whether such repeated exposure increases the cumulative

computational influence of specific cues on classification cannot be determined from the present aggregate metrics.

A further pattern can be observed by comparing the 200-token setting with the no-overlap setting. At 200 tokens, macro-recall remains 0.7 percentage points above the no-overlap result, whereas macro-precision falls slightly below it. Excessive overlap therefore retains part of the recall benefit but no longer improves balanced classification. Within the evaluated range, the relationship between overlap size and classification performance is non-monotonic, with the highest observed macro-precision and macro-recall occurring at 100 tokens.

4.3 Asymmetric Effects of BiGRU-Based Sequential Contextualization

A central empirical finding of this study is the asymmetric effect of BiGRU-based sequential contextualization on macro-precision and macro-recall under uniform mean pooling. As shown in Fig. 5, introducing BiGRU increases macro-recall from 80.4% to 81.3% but reduces macro-precision from 90.2% to 84.5%. Thus, a modest recall gain of 0.9 percentage points is accompanied by a substantially larger precision loss of 5.7 percentage points.

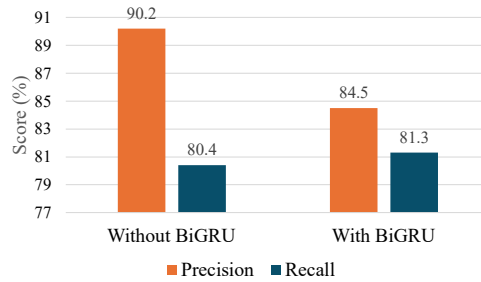


Fig. 5 Asymmetric effect of BiGRU-based sequential contextualization under uniform mean pooling. BiGRU produces a modest recall gain accompanied by a substantially larger precision loss

This asymmetric change corresponds to a more recall-oriented metric profile. BiGRU contextualization is associated with a slight increase in macro-recall, while macro-precision decreases substantially under uniform mean pooling. The observed pattern therefore indicates that broader cross-segment contextualization does not produce a balanced improvement under this aggregation strategy.

One possible interpretation is that contextualizing many segment representations together changes the relative contribution of diagnostically strong and weak information. This interpretation is consistent with the broader dilution effect associated with the introduction of non-diagnostic information (Nisbett et al. 1981; Zukier 1982), although the present experiments do not directly measure such a mechanism. Similarly, prior ECtHR research has identified shallow class-correlated signals in legal text (Santosh et al. 2022). The observed precision loss is therefore consistent with the possibility that cross-segment propagation increases the computational contribution of weak, redundant, or class-correlated cues. Direct verification would require cue-level intervention

or representation analysis. The present result establishes the precision-recall trade-off, rather than its underlying causal mechanism.

4.4 Conditional Effectiveness of Attention-Based Reweighting

Table 2 summarizes the performance of the four controlled aggregation configurations. The comparison shows that the effect of attention differs depending on whether the segment representations have first been contextualized by BiGRU.

Table 2 Performance of the four controlled aggregation configurations.

Configuration	Macro-Precision	Macro-Recall	Macro-F1
Independent encoding + mean pooling	90.2	80.4	83.0
Independent encoding + attention	89.5	80.3	82.8
BiGRU + mean pooling	84.5	81.3	82.5
BiGRU + attention	89.4	82.4	84.6

Without BiGRU, attention does not improve macro-F1, with the score decreasing slightly from 83.0% under mean pooling to 82.8% under attention-based aggregation. By contrast, when BiGRU contextualization is used, replacing mean pooling with attention increases macro-F1 from 82.5% to 84.6%.

Fig. 6 further illustrates this interaction. As shown in Fig. 6(a), under attention-based aggregation, introducing BiGRU increases macro-recall from 80.3% to 82.4%, while macro-precision remains nearly unchanged, decreasing from 89.5% to 89.4%. Fig. 6(b) shows the corresponding macro-F1 pattern across the four configurations.

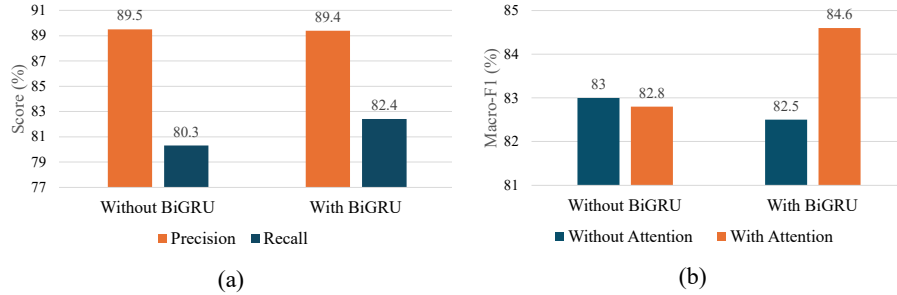


Fig. 6 Conditional interaction between BiGRU contextualization and attention-based aggregation. (a) Macro-precision and macro-recall under attention-based aggregation with and without BiGRU. (b) Macro-F1 across the four aggregation configurations, showing that the highest observed macro-F1 occurs when attention follows BiGRU contextualization

This interaction suggests that BiGRU and attention perform different but complementary functions. BiGRU propagates information across the segment sequence and enables distributed cues to influence one another. However, uniform aggregation assigns equal contribution to all contextualized representations, including those containing weak, redundant, or non-diagnostic signals. The observed recovery in macro-F1 is consistent with selective reweighting altering the relative contribution of contextualized segment representations during document-level aggregation.

Accordingly, the results suggest that attention is more beneficial as an aggregation mechanism over BiGRU-contextualized representations than over independently encoded segment representations. Under attention-based aggregation, the recall increase associated with BiGRU is retained while the large precision reduction observed under uniform aggregation is largely absent. Full configuration consequently achieves the highest macro-F1 among the four aggregation settings.

This conditional effect is consistent with hierarchical attention architectures in which attention operates over representations that have already been sequentially contextualized (Yang et al. 2016).

These results do not establish that attention weights correspond to legally important evidence. In this experiment, additive attention serves as a learned reweighting mechanism that controls the contribution of propagated information to document representation. Attention weights therefore reflect the relative computational contribution of text segments, which should be distinguished from their legal importance. They should not be treated as faithful explanations of the classification process (Jain and Wallace 2019).

4.5 Integrated Discussion

The experiments reveal configuration-dependent effects across the three mechanisms, including a conditional interaction between BiGRU contextualization and aggregation strategy. Within the evaluated overlap range, performance varies non-monotonically with repeated boundary content. Under uniform aggregation, BiGRU contextualization produces a pronounced precision-recall asymmetry rather than balanced improvement. Attention is beneficial when applied after BiGRU contextualization, but not when applied directly to independently encoded segment representations. These results show that the effects of greater contextual connectivity depend on how the resulting representations are aggregated.

From a legal reasoning perspective, these findings highlight the distinction between computational integration and evidential evaluation. Repeated presentation of the same information does not constitute independent corroboration when the observations derive from the same underlying source (Shafer 2016; Strittmatter et al. 2024). Broader propagation also does not ensure that supporting, qualifying, and defeating information is treated according to its legal function (Hage 2003; Prakken and Sartor 1996). Likewise, attention can control the relative contribution of textual segments, but its weights should not be interpreted as direct indicators of legal importance (Jain and Wallace 2019). These observations suggest that long legal document processing should distinguish between increasing computational access to information and determining how that information should contribute to a legal classification.

4.6 Scope and Limitations

The findings should be interpreted within the scope of the controlled experimental design. First, the analysis is limited to English-language ECtHR documents under a binary violation classification setting. The observed component interactions therefore cannot be assumed to generalize directly to other jurisdictions, languages, legal tasks, or multi-label settings. Second, the reported results are meant over five runs without estimates

of dispersion or statistical significance. The numerical differences are consequently treated as observed performance patterns rather than statistically established effects.

Third, the experiments measure changes in classification metrics but do not directly trace the propagation of individual lexical cues, quantify the influence of duplicated content, or inspect the information encoded in BiGRU hidden states. Explanations involving dilution, repeated exposure, or class-correlated signals should therefore be understood as plausible interpretations of the observed metric patterns rather than directly measured causal mechanisms. Similarly, attention weights are learned aggregation weights and do not establish legal relevance, satisfaction of legal requirements, or explanation faithfulness.

Finally, the analysis focuses on model performance under a fixed classification protocol rather than calibration or decision-theoretic behavior. Future experiments could directly test the proposed mechanisms through cue masking, counterfactual duplication, source-aware aggregation, and annotations of legal roles or support-defeat relations. The reported results should therefore be interpreted as descriptive patterns under the present experimental protocol.

5 Conclusion and Future Work

This study examined three recurring mechanisms in segmentation-based long legal document classification: overlapping segmentation, BiGRU-based sequential contextualization, and additive attention-based reweighting. Rather than proposing a new composite architecture, the study used a controlled analytical pipeline to investigate the individual and interactive effects of these mechanisms in a binary ECtHR violation classification setting.

Within the evaluated range, overlap shows a non-monotonic relationship with classification performance, with the highest observed macro-precision and macro-recall at 100 tokens. The result does not establish a universally optimal overlap size but shows that increasing repeated boundary content does not produce monotonic performance gains. BiGRU-based sequential contextualization produces an asymmetric effect on macro-precision and macro-recall under uniform mean pooling. It increases macro-recall from 80.4% to 81.3% but reduces macro-precision from 90.2% to 84.5%. Macro-F1 consequently decreases from 83.0% to 82.5%, showing that the modest recall gain is insufficient to compensate for the substantially larger precision loss. Under uniform mean pooling, BiGRU contextualization is therefore associated with a more recall-oriented metric profile and substantially lower macro-precision.

In the evaluated configuration, attention does not improve macro-F1 when applied to independently encoded segments; the observed score decreases slightly from 83.0% to 82.8%. In contrast, attention is associated with higher macro-F1 when applied after BiGRU contextualization.

From a legal reasoning perspective, these results highlight the distinction between computational integration and evidential evaluation. Repeated presentation of the same passage does not create independent corroboration, and broader contextual propagation does not ensure that supportive, qualifying, and defeating information is treated

according to its legal function. Likewise, attention weights show which text segments contribute more to the model’s classification. They do not establish that those segments are legally decisive. The observed effects of overlap and the interaction between BiGRU and attention therefore show that increasing computational access to context does not necessarily improve alignment with evidential relevance.

Taken together, the findings indicate that the effects of overlapping segmentation, sequential contextualization, and attention depend on their configuration and interaction. Increasing overlap does not produce monotonic gains, BiGRU contextualization under uniform aggregation produces a pronounced precision-recall trade-off, and attention is associated with higher macro-F1 when applied after contextualization. These observations concern computational treatment and classification behavior. They do not establish that repeated or propagated information receives greater legal evidential force.

Future work should examine whether the observed precision-recall asymmetry and component interactions generalize across other legal datasets, jurisdictions, languages, and long-context architectures. More targeted experiments could directly test the mechanisms suggested by the present analysis. Counterfactual duplication and source-aware aggregation could examine whether repeated exposure increases the contribution of duplicated information to classification; cue masking and segment deletion could identify the information associated with precision loss under sequential contextualization; and legal-role or support-defeat annotations could evaluate whether selective aggregation aligns with legally relevant discourse relations. Such analyses would help distinguish observed performance interactions from the underlying mechanisms that produce them.

References

- Adhikari A, Ram A, Tang R, Lin J (2019) Rethinking complex neural network architectures for document classification. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, pp 4046–4051. Association for Computational Linguistics, Minneapolis. <https://doi.org/10.18653/v1/N19-1408>
- Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA
- Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 4317–4323. Association for Computational Linguistics, Florence. <https://doi.org/10.18653/v1/P19-1424>
- Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the Muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp 2898–2904. Association for Computational Linguistics, Online. <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- Chalkidis I, Dai X, Fergadiotis M, Malakasiotis P, Elliott D (2022) An exploration of hierarchical attention transformers for efficient long document classification. arXiv preprint arXiv:2210.05529. <https://doi.org/10.48550/arXiv.2210.05529>
- Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine

- translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp 1724–1734. Association for Computational Linguistics, Doha. <https://doi.org/10.3115/v1/D14-1179>
- Dai X, Chalkidis I, Darkner S, Elliott D (2022) Revisiting transformer-based models for long document classification. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp 7212–7230. Association for Computational Linguistics, Abu Dhabi. <https://doi.org/10.18653/v1/2022.findings-emnlp.534>
- Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, pp 4171–4186. Association for Computational Linguistics, Minneapolis. <https://doi.org/10.18653/v1/N19-1423>
- Fama I, Bueno B, Alcoforado A, Ferraz TP, Moya A, Costa AHR (2024) No argument left behind: overlapping chunks for faster processing of arbitrarily long legal texts. In: Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology, pp 237–246. Association for Computational Linguistics, Belém do Pará, Brazil. <https://aclanthology.org/2024.stil-1.24/>
- Grail Q, Perez J, Gaussier E (2021) Globalizing BERT-based transformer architectures for long document summarization. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp 1792–1810. Association for Computational Linguistics, Online. <https://doi.org/10.18653/v1/2021.eacl-main.154>
- Hage JC (2003) Law and defeasibility. *Artif Intell Law* 11:221–243. <https://doi.org/10.1023/B:ARTI.0000046011.13621.08>
- Han G, Tsao J, Huang X (2024) Length-aware multi-kernel transformer for long document classification. In: Proceedings of the 13th Joint Conference on Lexical and Computational Semantics, pp 278–290. Association for Computational Linguistics, Mexico City. <https://doi.org/10.18653/v1/2024.starsem-1.22>
- Jain S, Wallace BC (2019) Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, pp 3543–3556. Association for Computational Linguistics, Minneapolis. <https://doi.org/10.18653/v1/N19-1357>
- Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P (2024) Lost in the middle: how language models use long contexts. *Trans Assoc Comput Linguist* 12:157–173. https://doi.org/10.1162/tacl_a_00638
- Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA
- Munkhdalai T, Faruqi M, Gopal S (2024) Leave no context behind: efficient infinite context transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*. <https://doi.org/10.48550/arXiv.2404.07143>
- Nisbett RE, Zukier H, Lemley RE (1981) The dilution effect: nondiagnostic information weakens the implications of diagnostic information. *Cogn Psychol* 13(2):248–277. [https://doi.org/10.1016/0010-0285\(81\)90010-4](https://doi.org/10.1016/0010-0285(81)90010-4)
- Pappagari R, Żelasko P, Villalba J, Carmiel Y, Dehak N (2019) Hierarchical transformers for long document classification. In: Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp 838–844. IEEE, Singapore. <https://doi.org/10.1109/ASRU46091.2019.9003958>

- Prakken H, Sartor G (1996) A dialectical model of assessing conflicting arguments in legal reasoning. *Artif Intell Law* 4:331–368. <https://doi.org/10.1007/BF00118496>
- Santosh TYSS, Xu S, Ichim O, Grabmair M (2022) Deconfounding legal judgment prediction for European Court of Human Rights cases towards better alignment with experts. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp 1120–1138. Association for Computational Linguistics, Abu Dhabi. <https://doi.org/10.18653/v1/2022.emnlp-main.74>
- Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 45(11):2673–2681. <https://doi.org/10.1109/78.650093>
- Shafer G (2016) The problem of dependent evidence. *Int J Approx Reason* 79:41–44. <https://doi.org/10.1016/j.ijar.2016.05.003>
- Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45(4):427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Strittmatter LE, Pilditch TD, Lagnado DA (2024) Reasoning about (in)dependent evidence: a mismatch between perceiving and incorporating dependencies? In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, vol 46, pp 951–958. <https://escholarship.org/uc/item/7jt638nj>
- Wu ML, Liu CW (2026) Toward deployment-oriented long legal document classification: a segmentation-based framework for distributed legal evidence integration. *IEEE Access* 14:72103–72116. <https://doi.org/10.1109/ACCESS.2026.3691973>
- Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp 1480–1489. Association for Computational Linguistics, San Diego. <https://doi.org/10.18653/v1/N16-1174>
- Zukier H (1982) The dilution effect: the role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *J Pers Soc Psychol* 43(6):1163–1174. <https://doi.org/10.1037/0022-3514.43.6.1163>