Learning Long-Document Embeddings via Chunk–Context Entailment

Waheed Ahmed Abro[1], Naïm Es-Sebbani[2,3], Zied Bouraoui[2]
[1] SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Saudi Arabia,
[2]CRIL - CNRS & Univ Artois, France, [3]GRAMMATICA UR 4521, Univ. Artois, France
engr.waheedabro@gmail.com, {essebbani, bouraoui}@cril.fr

Abstract

Learning faithful embeddings for long documents remains challenging, especially in domains like law and medicine where inputs are long, structured, and semantically heterogeneous. We introduce the Chunk Prediction Encoder (CPE), a self-supervised framework that treats chunk–context compatibility as an unsupervised NLI problem. Given a document, CPE masks a chunk and learns (i) a contrastive objective that aligns the masked document with its held-out chunk against in-batch negatives, and (ii) a binary entailment head that predicts whether a candidate chunk belongs to the document. This joint objective encourages both geometric smoothness and directional semantic consistency, yielding robust document-level embeddings. We evaluate CPE with hierarchical and sparse-attention backbones on five benchmarks spanning legal and biomedical domains under frozen-embedding and end-to-end fine-tuning protocols. CPE consistently outperforms baselines, and is more compute-efficient than prompt-only LLM baselines under matched token budgets. Ablations demonstrate the effect of chunk length, the contrastive-vs-entailment balance, and skimming strategies. We release code, configurations, and trained checkpoints to facilitate reproducibility and downstream reuse in long-document classification and retrieval pipelines.

Keywords: Long-document representation, self-supervised learning, legal NLP, biomedical NLP, evaluation and reproducibility.

1. Introduction

Learning high-quality representations for long documents remains a central challenge in NLP. While large pre-trained language models (LMs) have achieved impressive results on sentence- and paragraph-level tasks, their ability to encode entire documents is limited by computational constraints and by the complexity of long-range semantic dependencies. Yet, such capabilities are essential in application domains that rely on rich textual evidence, including legal reasoning, medical report analysis, and RAG systems (Saggau et al., 2023; Zhang et al., 2024; Zhao et al., 2025).

Most document encoders rely on self-attention architectures that scale quadratically with input length, making it impractical to process documents beyond a few hundred tokens. Sparse-attention mechanisms such as Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and hierarchical transformers (Chalkidis et al., 2019; Wu et al., 2021; Dai et al., 2022) partially mitigate this limitation by segmenting documents into manageable chunks and modeling inter-chunk dependencies through local or hierarchical attention. However, simply extending input length is not sufficient to capture cross-section coherence: longer documents often contain heterogeneous segments with distinct rhetorical functions, specialized terminology, and non-linear argumentation structures. As a result, document representations obtained by global pooling over local embeddings can remain shallow or fragmented.

When domain experts (e.g., judges or clinicians) read lengthy documents, they rarely process all content sequentially. Instead, they *skim* the text, selectively focusing on informative fragments that collectively convey the meaning of the document. Inspired by this cognitive process, we propose the *Chunk Prediction Encoder (CPE)*, a self-supervised framework that models long-document understanding as a process of *chunk–context entailment*. Given a document, CPE masks one chunk and trains the model to predict whether a given text fragment is compatible with the remaining document context. This task is framed as an *unsupervised Natural Language Inference (NLI)* objective: positive pairs correspond to chunks belonging to the same document, while negatives are sampled from other documents.

CPE combines two complementary learning signals: (i) a contrastive loss that aligns document and chunk representations within a shared embedding space, and (ii) a binary entailment classifier that enforces directional compatibility between the masked document and its candidate chunk. This joint objective encourages the model to learn representations that are both geometrically smooth and semantically coherent across sections. By integrating this self-supervised training signal with hierarchical and sparse-attention backbones, CPE produces general-purpose document embeddings that capture intra- and inter-fragment dependencies without supervision.

We evaluate CPE on five benchmark datasets spanning legal and biomedical domains—ECHR, SCOTUS, EURLEX, MIMIC-III,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7405–7414
11–16 May 2026.

and BioASQ—under both frozen-embedding and end-to-end fine-tuning regimes. Across all settings, CPE consistently outperforms SimCSE (Gao et al., 2021), ESimCSE (Wu et al., 2022), ArcCSE (Zhang et al., 2022), SimCSE++ (Xu et al., 2023b), and DistillCSE (Xu et al., 2023a). Beyond empirical gains, CPE remains computationally efficient and model-agnostic, integrating seamlessly with existing pre-trained encoders such as LegalBERT and ClinicalBioBERT. Our main contributions are as follows:

- We introduce *Chunk Prediction Encoder (CPE)*, a self-supervised framework that models long-document representation as chunk–context entailment, combining contrastive and NLI-style objectives.
- We demonstrate that CPE yields consistent improvements over strong baselines across legal and biomedical benchmarks, using both hierarchical and sparse-attention architectures.
- We provide an extensive evaluation of chunk length, loss balancing, and skimming strategies, highlighting how CPE generalizes across document lengths and domains.
- We release code, datasets, and trained checkpoints to ensure full reproducibility and facilitate reuse in long-document classification and retrieval applications.

2. Related Work

This section reviews prior work on long-document modeling, self-supervised contrastive learning for text representation, domain-specific encoders, and benchmark resources for evaluation. Our approach builds upon these strands by introducing a unified chunk–context entailment objective for long texts.

Modeling Long Documents. Transformers face quadratic complexity with respect to sequence length, limiting their ability to process full documents. Several approaches mitigate this through sparse or hierarchical mechanisms. Sparse-attention models such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) reduce complexity by restricting attention patterns to local and global windows. Hierarchical transformers (Yang et al., 2016; Chalkidis et al., 2019; Wu et al., 2021; Dai et al., 2022) instead encode segments (sentences or paragraphs) individually before aggregating their representations through higher-level attention or recurrent layers. For instance, Chalkidis et al. (2019) used hierarchical BERT for legal judgment prediction, while Wu et al. (2021) introduced Hi-Transformer with two-stage encoding of sentences and documents. Despite their efficiency, these architectures often rely on supervised objectives and do not explicitly model coherence between distant fragments. CPE complements these models by introducing a self-supervised objective that enforces semantic alignment between document segments.

Self-Supervised Representation Learning. Contrastive learning has become the dominant paradigm for sentence representation (Reimers and Gurevych, 2019; Gao et al., 2021; Wu et al., 2022). SimCSE (Gao et al., 2021) uses dropout-based augmentation to generate positive pairs, while ESimCSE (Wu et al., 2022) adds token repetition as data augmentation. Recent extensions—ArcCSE (Zhang et al., 2022), SimCSE++ (Xu et al., 2023b), and DistillCSE (Xu et al., 2023a)—introduce angular contrastive objectives, multi-view consistency, or knowledge distillation, achieving state-of-the-art sentence embeddings. However, these methods focus on short inputs and symmetric similarity; they do not capture the hierarchical, directional relations that exist within long documents. Our work frames intra-document coherence as a chunk–context *entailment* task, where contrastive and NLI-inspired signals jointly enhance representation quality.

Domain-Specific Document Encoders. Several pre-trained models target domain adaptation for specialized corpora. In the legal domain, LegalBERT (Chalkidis et al., 2020) and CaseLawBERT (Zheng et al., 2021) leverage case law corpora for tasks such as judgment prediction and legal information retrieval. For biomedical text, ClinicalBERT and ClinicalBioBERT (Alsentzer et al., 2019) capture medical terminology from clinical notes and PubMed abstracts. While these domain-specific encoders improve local lexical representations, they still depend on downstream supervision to learn cross-section reasoning. CPE integrates seamlessly with such backbones, augmenting them with self-supervised document-level coherence learning.

Evaluation Resources and Benchmarks. Recent multi-domain evaluation suites have enabled systematic comparison of long-document encoders. In the legal domain, LexGLUE (Chalkidis et al., 2022b) unifies datasets such as ECHR, SCOTUS, and EURLEX under a common protocol for legal document classification and entailment. In the biomedical field, MIMIC-III (Johnson et al., 2016) and BioASQ (Tsatsaronis et al., 2015) provide large-scale annotated corpora for clinical coding and biomedical concept classification. Our experimental setup follows these resources to ensure reproducibility and comparability with prior work. All preprocessing scripts, hyperparameters, and trained models will be released to facilitate further benchmarking by the community.

Summary. Existing models either focus on local attention patterns or on sentence-level contrastive

learning. CPE bridges these directions by introducing a unified objective that models how document fragments semantically entail or contradict the document context.

3. Chunk Prediction Encoder (CPE)

We propose the *Chunk Prediction Encoder (CPE)*, a self-supervised framework for learning long-document representations by predicting whether a candidate text fragment (*chunk*) is semantically compatible with its document context. Unlike conventional contrastive objectives that treat sentences as independent units, CPE explicitly couples local and global semantics through a *chunk–context entailment* task. Intuitively, the model learns to ask: *does this fragment belong to this document?*—a question that encourages deeper contextual reasoning beyond surface similarity. CPE is architecture-agnostic and can operate on both hierarchical and sparse-attention encoders, jointly optimizing contrastive alignment and entailment prediction.

3.1. Hierarchical Encoding of Long Documents

Long documents exhibit multi-scale structure: local coherence within paragraphs and global consistency across sections. To capture this, we first construct a hierarchical representation where each document $D = \{c_1, c_2, \ldots, c_n\}$ is segmented into n chunks of up to T tokens. Each chunk c_i is independently encoded using a shared pre-trained language model $\mathcal{M}$:

$$f_i = \mathcal{M}(w_{\text{[CLS]}}, w_1, \ldots, w_T), \tag{1}$$

where the [CLS] token output f_i summarizes the local semantics of chunk c_i. This shared encoder ensures consistency across segments while maintaining scalability. To produce the document-level embedding, we aggregate all chunk vectors via mean pooling:

$$d = \frac{1}{n}\sum_{i=1}^{n} f_i. \tag{2}$$

This hierarchical encoding provides a balance between global coverage and computational efficiency, allowing CPE to model documents of several thousand tokens without quadratic attention overhead.

3.2. Chunk Prediction as Self-Supervised Learning

The core idea of CPE is to learn document embeddings by predicting whether a candidate chunk belongs to its context. For each document d_i, we randomly mask one chunk c_i^+ and treat the remainder as $\tilde{d}_i$. We then sample a negative chunk c_i^- from another document in the same batch. This setup creates a self-supervised contrast between *entailing* and *non-entailing* fragments, encouraging the model to infer discourse-level compatibility rather than rely on lexical overlap.

We first apply a contrastive loss to align each document context $\tilde{d}_i$ with its corresponding chunk:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{N}\sum_i \log \frac{\exp(\text{sim}(f(\tilde{d}_i), f(c_i^+))/\tau)}{\sum_k \exp(\text{sim}(f(\tilde{d}_i), f(c_k^-))/\tau)}, \tag{3}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is a temperature parameter. This loss encourages chunks from the same document to occupy nearby regions in the embedding space, while unrelated chunks are pushed apart.

However, similarity alone may not guarantee semantic coherence: two fragments can be close in vector space yet unrelated in meaning. To address this, we introduce a binary entailment classifier that explicitly predicts whether a chunk c is compatible with $\tilde{d}_i$:

$$\begin{aligned}\mathcal{L}_{\text{NLI}} = &- y \log p_\theta(y|\tilde{d}_i, c) \\ &- (1-y)\log(1 - p_\theta(y|\tilde{d}_i, c)),\end{aligned} \tag{4}$$

where $y{=}1$ if $c{=}c_i^+$ and $y{=}0$ otherwise. Unlike classical NLI tasks, here "entailment" does not refer to logical inference but to *semantic membership*—teaching the encoder to detect whether a local fragment fits within the global context of a document. The overall training objective combines the two signals:

$$\mathcal{L}_{\text{joint}} = \lambda\mathcal{L}_{\text{NLI}} + (1-\lambda)\mathcal{L}_{\text{CL}}, \tag{5}$$

with $\lambda{=}0.5$ balancing structural discrimination and semantic alignment. This joint loss grounds the embedding geometry in both distributional similarity and contextual entailment, yielding representations that reflect discourse cohesion rather than mere token co-occurrence.

3.3. Integration with Sparse-Attention Models

CPE can be seamlessly integrated with sparse-attention architectures such as Longformer or BigBird. In this configuration, a document is divided into two segments: the reference context $\tilde{d}$ and the candidate chunk c. Both are encoded using the same transformer, producing embeddings $z_{\tilde{d}}$ and z_c shared across the two objectives. The joint loss $\mathcal{L}_{\text{joint}}$ remains unchanged, but the sparse attention mechanism enables the model to handle up to 4k tokens per input, covering most real-world legal or

medical documents. This architecture preserves fine-grained contextual attention while remaining computationally linear with input length.

3.4. Document Classification Head

After pretraining, CPE yields high-quality document embeddings that can be directly evaluated without fine-tuning the backbone. To assess their transferability, we attach a lightweight feed-forward classifier: a three-layer MLP with `tanh` activations followed by either sigmoid (multi-label) or softmax (single-label) outputs:

$$\hat{y} = g(\tanh(Wd + b)), \tag{6}$$

where g denotes the activation function. This probe isolates the quality of the learned representations from downstream optimization effects, providing a fair comparison with other self-supervised encoders. Empirically, this setup reveals that CPE captures not only topical similarity but also deeper interdependencies between document segments—crucial for domains where long-range reasoning drives predictive performance.

4. Experimental Setup

We evaluate the proposed Chunk Prediction Encoder (CPE) across legal and biomedical domains using both hierarchical and sparse-attention architectures. The code is available at `https://github.com/essebbaninaim/HiBERT_CPT`

Datasets We evaluate our approach on five publicly available datasets representative of long and short specialized texts, summarized in Table 1. The *ECHR* corpus (Chalkidis et al., 2021) contains approximately 11k European Court of Human Rights cases annotated with up to ten human-rights articles, averaging about 2k tokens per document in a multi-label classification setting. The *SCOTUS* dataset (Chalkidis et al., 2022b) includes around 5k Supreme Court of the United States cases labeled with 14 legal issues such as civil rights or criminal procedure, averaging roughly 8k tokens per document for single-label classification. The *EURLEX* dataset (Chalkidis et al., 2022b) consists of roughly 65k EU legislative acts annotated with 100 EuroVoc concepts, with an average document length of 1.4k tokens, serving as a medium-length benchmark. In the biomedical domain, the *MIMIC-III* corpus (Johnson et al., 2016) comprises about 40k de-identified hospital discharge summaries annotated with ICD-9 diagnostic codes; we consider the 19 top-level codes for multi-label classification. Finally, the *BioASQ* dataset (Tsatsaronis et al., 2015) contains biomedical abstracts labeled with first-level MeSH categories, averaging 300 tokens, which we use to assess model generalization to shorter texts.

Table 1: Dataset statistics and average token length.

Dataset	Train	Test	Avg. length
ECHR	9,000	1,000	2,050
SCOTUS	5,000	1,400	8,000
EURLEX	55,000	5,000	1,400
MIMIC	30,000	10,000	3,200
BioASQ	80,000	20,000	300

Implementation Details We use pre-trained language models from `HuggingFace`, including `bert-base-uncased`, `roberta-base`, `legal-bert-base`, and `ClinicalBioBERT`. For sparse-attention experiments, we adopt Longformer (Beltagy et al., 2020). All models process up to 4096 tokens per document (32 chunks of 128 tokens). For EURLEX, shorter documents are truncated to 2048 tokens (16×128 chunks). Chunks are padded as needed to maintain consistent length.

Training uses the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 10^{-3}. Self-supervised pretraining runs for 3 epochs (batch size 4), while downstream classification with an MLP head runs for 20 epochs (batch size 16). The temperature τ in $\mathcal{L}_{\text{CL}}$ is set to 0.05, and λ=0.5 balances the joint loss. All experiments are executed on an NVIDIA RTX 8000 GPU (48 GB VRAM).

Evaluation Protocol To ensure comparability, we follow the standardized LexGLUE evaluation setup for legal datasets and the BioASQ protocol for biomedical text classification. We report both macro- and micro-F1 scores averaged over five random seeds. Macro-F1 highlights performance on rare classes, while micro-F1 measures overall accuracy weighted by class frequency.

Baselines We benchmark CPE against several strong baselines:

- **Pre-trained Embedding + MLP:** Fixed embeddings from BERT, RoBERTa, LegalBERT, or ClinicalBioBERT with an MLP classifier.
- **SimCSE / ESimCSE:** Contrastive pretraining on long documents using dropout- and repetition-based positive pairs (Gao et al., 2021; Wu et al., 2022).
- **Recent Contrastive Models:** ArcCSE (Zhang et al., 2022), SimCSE++ (Xu et al., 2023b), and DistillCSE (Xu et al., 2023a).

Table 2: Classification performance on document embeddings produced by contrastively trained *hierarchical transformer models* on ECHR, SCOTUS, EURLEX, and MIMIC datasets. Performance is reported in F1 scores $\times 100$ (macro / micro). Best per column in bold.

PTM	Model	ECHR		SCOTUS		EURLEX		MIMIC	
		Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
BERT	Embedding + MLP	36.3	55.6	40.6	61.8	25.9	51.4	49.1	63.0
	SimCSE	48.7	61.4	45.5	61.1	33.4	57.1	54.6	67.1
	ESimCSE	50.8	64.4	52.1	64.1	34.2	54.6	57.0	68.1
	ArcCSE	52.1	65.1	53.6	65.4	36.2	59.0	57.3	68.2
	SimCSE++	52.9	65.7	53.8	65.8	37.4	60.2	57.9	68.5
	DistillCSE	53.2	66.0	54.1	66.0	38.5	61.0	58.0	68.6
	CPE (ours)	**54.8**	**67.0**	**54.6**	**66.8**	**41.0**	**63.7**	**59.2**	**69.1**
RoBERTa	Embedding + MLP	35.3	55.6	32.8	57.9	21.9	35.1	49.6	64.3
	SimCSE	49.0	59.4	48.3	62.7	35.1	53.6	55.7	67.0
	ESimCSE	45.3	56.6	57.7	67.9	33.8	54.1	57.5	67.3
	ArcCSE	52.8	64.0	56.1	67.5	37.9	60.8	58.0	68.1
	SimCSE++	54.1	64.9	57.3	67.9	38.8	61.7	59.2	68.8
	DistillCSE	54.6	65.0	57.6	68.0	39.0	62.0	59.3	69.0
	CPE (ours)	**56.0**	**65.6**	**57.9**	**68.1**	**41.9**	**63.4**	**61.0**	**69.4**
LegalBERT	Embedding + MLP	52.6	65.3	44.9	65.4	17.8	39.5	–	
	SimCSE	57.5	68.8	57.6	68.9	40.9	61.6	–	
	ESimCSE	56.5	68.3	56.1	67.8	40.4	62.6	–	
	ArcCSE	57.0	68.7	57.9	69.1	41.1	63.0	–	
	SimCSE++	57.2	69.0	58.2	69.3	41.5	63.3	–	
	DistillCSE	57.4	69.1	58.6	69.4	41.6	63.5	–	
	CPE (ours)	**57.7**	**69.4**	**59.5**	**71.9**	**42.2**	**64.2**	–	
ClinicalBioBERT	Embedding + MLP	Not applicable to legal datasets						55.8	68.9
	SimCSE	Not applicable to legal datasets						62.7	71.1
	ESimCSE	Not applicable to legal datasets						60.4	70.9
	ArcCSE	Not applicable to legal datasets						61.9	71.2
	SimCSE++	Not applicable to legal datasets						62.3	71.4
	DistillCSE	Not applicable to legal datasets						62.9	71.5
	CPE (ours)	Not applicable to legal datasets						**63.9**	**71.7**

- **Fine-tuned Architectures:** HiLegalBERT (Chalkidis et al., 2022b), LegalLongformer (Mamakas et al., 2022), HAT (Chalkidis et al., 2022a), and LSG (Condevaux and Harispe, 2023).

CPE is evaluated under both frozen-embedding (representation quality) and full fine-tuning (end-to-end adaptation) regimes. All hyperparameters, random seeds, and evaluation scripts will be publicly released for full reproducibility.

5. Results and Analysis

We evaluate CPE across five datasets and four backbone architectures under two regimes: (i) *frozen-embedding evaluation*, where only a lightweight MLP classifier is trained to assess representation quality, and (ii) *end-to-end fine-tuning*, where all parameters are optimized jointly. Results are reported as macro- and micro-F1 scores ($\times 100$) averaged over five random seeds.

Hierarchical Encoders Table 2 reports results for hierarchical transformer encoders using four pre-trained backbones (BERT, RoBERTa, LegalBERT, and ClinicalBioBERT). We evaluate frozen document embeddings and compare CPE against SimCSE, ESimCSE, and recent extensions. CPE consistently outperforms all baselines, improving over SimCSE and ESimCSE by +2–4 macro-F1 on average (up to +6 on ECHR) and exceeding newer variants by +1.5–3 macro-F1 on ECHR and SCOTUS. These gains hold across both macro- and micro-F1, showing that CPE benefits rare and frequent classes alike. The improvements stem from explicitly modeling *chunk–context entailment*, which provides complementary supervision to contrastive alignment. While SimCSE-style methods promote sentence-level similarity, CPE teaches the encoder to judge whether a text span is semantically compatible with its document context. This inductive bias enhances hierarchical coherence, reduces embedding fragmentation, and improves representation isotropy—key for long-document tasks such as ECHR, SCOTUS, EURLEX, and MIMIC. Quantitatively, generic BERT and RoBERTa embeddings yield the weakest performance, highlighting the difficulty of modeling specialized domains. Self-supervised contrastive learning narrows this gap: SimCSE and ESimCSE improve BERT by +12% and +14% macro-F1 on ECHR, while CPE adds a further +6%, with similar trends across datasets. Domain-specific encoders further enhance results: $\text{LegalBERT}_{\text{CPE}}$ gains +5 macro-F1 over BERT_{CPE} on SCOTUS and +3 on ECHR, and $\text{ClinicalBioBERT}_{\text{CPE}}$ adds +3 on MIMIC. $\text{RoBERTa}_{\text{CPE}}$ performs comparably, benefiting from broader subword coverage and larger pretraining corpora.

Table 3: Classification performance on document embeddings produced by contrastively trained **Longformer models**. Performance is reported in F1 scores $\times 100$ (macro / micro). Best per column in bold.

Model	ECHR		SCOTUS		EURLEX		MIMIC	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
Embedding + MLP	35.9	50.4	27.4	50.3	25.8	54.5	44.6	62.1
SimCSE	35.1	46.8	36.8	52.6	32.0	53.6	45.4	60.8
ESimCSE	32.9	46.3	35.3	53.3	35.4	55.9	46.1	61.4
ArcCSE	39.2	52.1	41.8	56.2	38.6	58.9	48.8	62.5
SimCSE++	42.5	54.3	44.9	58.4	39.7	59.7	49.9	63.2
DistillCSE	43.3	55.0	46.5	59.6	40.5	60.4	50.3	63.6
CPE (ours)	**48.9**	**59.7**	**47.5**	**62.6**	**43.2**	**64.6**	**53.1**	**64.9**

Sparse-Attention (Longformer) Results Table 3 reports results for the sparse-attention Longformer encoder, which efficiently processes sequences up to 4k tokens. While block attention improves scalability, it lacks explicit supervision for modeling dependencies across distant spans. CPE addresses this limitation through its chunk–context prediction objective, enforcing document-level coherence and long-range consistency. The vanilla Longformer + MLP classifier performs poorly because it cannot capture inter-paragraph relationships. Self-supervised variants (SimCSE, ESimCSE, CPE) markedly improve performance, with CPE achieving the largest gains: +12% and +14% macro-F1 over SimCSE and ESimCSE on ECHR, +10% and +12% on SCOTUS, and +8% and +7% on MIMIC. On the shorter EURLEX corpus, improvements are smaller but consistent. These results show that CPE is most effective on long, structured documents where discourse coherence is crucial. Its wider receptive field (4096 tokens) also narrows the gap to the hierarchical CPE encoder on long ECHR cases. Across architectures, hierarchical transformers still lead on ECHR, SCOTUS, and MIMIC, while Longformer slightly outperforms them on EURLEX due to shorter input lengths. This reflects a clear trade-off: sparse attention efficiently captures full context for medium texts, whereas hierarchical models remain better suited for deeply structured documents. Nonetheless, CPE yields strong and stable gains in both regimes, confirming that chunk–context entailment generalizes across encoder designs.

End-to-End Fine-Tuning We further analyze how CPE performs when all model parameters are fine-tuned for downstream tasks. Table 4 reports results on the ECHR and SCOTUS datasets. Despite its simplicity, CPE matches or surpasses the strongest fully fine-tuned architectures—including Hi-LegalBERT, LegalLongformer, LSG, and HAT—while requiring fewer parameter updates and shorter training time. Compared to LegalLongformer, CPE achieves a gain of +2.1 macro-F1 on ECHR and +0.8 on SCOTUS, confirming that its self-supervised initialization provides richer document-level representations that remain beneficial during supervised adaptation. Notably, CPE also accelerates convergence and improves stability: models initialized with CPE reach comparable validation performance in fewer epochs and exhibit lower variance across random seeds. This behavior suggests that entailment-based pretraining organizes latent representations into smoother manifolds, facilitating gradient flow during fine-tuning and reducing overfitting to dominant classes. Overall, these results demonstrate that CPE not only improves frozen embeddings but also serves as an effective pretraining strategy for robust and efficient task-specific adaptation.

Table 4: End-to-end fine-tuning results on ECHR and SCOTUS datasets (macro / micro F1 $\times 100$). Best in bold.

Model	ECHR	SCOTUS
Hi-LegalBERT (Chalkidis et al., 2022b)	64.0 / 70.0	66.5 / 76.4
LSG (Condevaux and Harispe, 2023)	60.3 / 71.0	63.7 / 73.3
LegalLongformer (Mamakas et al., 2022)	63.6 / 71.7	66.9 / 76.6
HAT (Chalkidis et al., 2022a)	59.1 / 79.8	59.1 / 70.5
CPE (ours)	**66.1 / 72.6**	**67.3 / 77.5**

6. Ablation and Generalization Analysis

We conducted an ablation study to evaluate the impact of different chunk sizes, visualize the quality of the embedding space, and examine the performance of the CPE framework on short documents.

Evaluation of Advanced Hierarchical Representation We hypothesise that the performance of CPE improves when using an advanced document encoder. To test this, we conducted experiments using Transformer over BERT (ToBERT) and Recurrence over BERT (RoBERT). For classification tasks, we kept the parameters of BERT fixed (frozen), and only the Transformer and LSTM encoder with MLP layer were learned during training. The results in Table 5 show indeed that ToBERT improved the performance of the generalized HBERT by 6% in macro F1 and 1.5% in μ F1-score. On

Table 5: Performance of hierarchical transformer CPE encoders applied on BERT for ECHR and SCOTUS datasets (macro/micro F1 $\times$100).

Model	ECHR	SCOTUS
HBERT+MLP	54.8 / 67.0	54.6 / 66.8
RoBERT+MLP	55.0 / 67.1	58.3 / 69.6
ToBERT+MLP	**60.3 / 68.5**	**58.8 / 70.1**

Table 6: Performance of hierarchical transformer encoders with varying chunk sizes (macro/micro F1 $\times$100). Results use LegalBERT for ECHR, SCOTUS, EURLEX and ClinicalBioBERT for MIMIC.

Chunk Size	ECHR	SCOTUS	EURLEX	MIMIC
64	57.6 / 69.1	58.8 / 71.1	42.1 / 63.8	63.1 / 70.7
128	**57.7 / 69.4**	**62.7 / 73.3**	**42.2 / 64.2**	**63.4 / 70.8**
256	56.9 / 67.7	59.2 / 71.8	39.0 / 59.2	61.5 / 70.3
512	55.6 / 67.3	59.5 / 71.9	29.3 / 43.6	60.6 / 68.9

the SCOTUS dataset, ToBERT achieved a performance gain of approximately 4% in both macro and μ F1-scores.

Impact of chunk length: Table 6 and Figure 1 summarize the CPE classification performance measured by macro-F1 using a hierarchical Transformer encoder with chunk sizes of 64, 128, 256, and 512 tokens across four datasets: ECHR, SCOTUS, EURLEX, and MIMIC. Fig 1 illustrates how performance varies with chunk size. For the ECHR data set, on small chunk size of 64 produce 57.64 macro F1 score and remains stable for chunk size 128 and produce 57.74 micro F-1 score. The model performance then gradually decreases for large chunk sizes of 256 and 512. On the other hand, for SCOTUS dataset, there is a notable improvement when moving from a chunk size of 64 to 128, after which the performance slightly drops for larger sizes. The scores for MIMIC show a modest decline, on chunk size 64 to chunk size 512. This demonstrates relative robustness with only a slight decrease as the chunk size increases. Conversely, the EURLEX dataset exhibits its best performance at the smaller chunk sizes of 64 and 128, but shows a sharp decline at chunk sizes of 256 and 512. This suggests that EURLEX is highly sensitive to the chunk size parameter, likely because its shorter average text length means that larger chunks incorporate too much irrelevant detail.

Performance on short document corpus To evaluate our CPE framework on short documents, we perform experiments on the BIOASQ dataset. We followed the method outlined in Section 3.2, but instead of using the Longformer encoder, we utilized ClinicalBioBERT and BERT features, setting the length of the positive chunk to 64 tokens. Table 7 reports classification results. The top rows show the performance of models using BERT embedding and the bottom rows display the performance of models using ClinicalBioBERT embedding. The ClinicalBioBERT Embedding$_{CPE}$ + MLP model produces the highest macro and micro F1 scores, achieving 71.28 and 84.43 macro F1 and micro F1-scores, respectively. This indicates that self-supervised CPE learning produces high-quality embeddings. Conversely, the state-of-the-art ClinicalBioBERT Emb$_{SimCSE}$ + MLP and ClinicalBioBERT Emb$_{ESimCSE}$ + MLP models does not enhance the performance of the baseline model Embedding + MLP. This suggests that using only dropout augmentation or basic word repetition to form positive pairs for generating text embeddings yields little benefit for document- or paragraph-level representations, even though these techniques perform very well for sentence embeddings. Results demonstrate that the proposed CPE method improves embedding derived from ClinicalBioBERT by around 4% macro-F1.

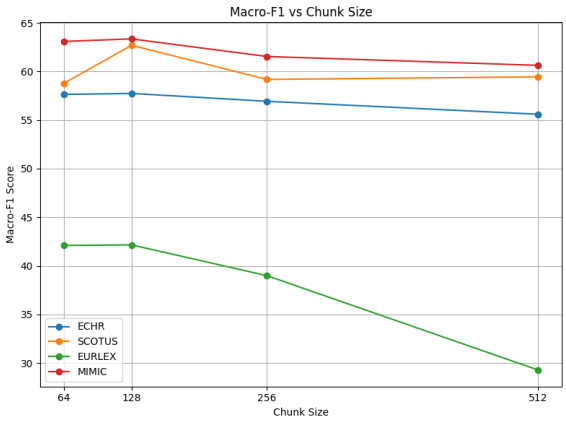

Figure 1: Macro-F1 Scores vs. chunk size via hierarchical transformer encoders applied on LegalBERT for ECHR, SCOTUS, EURLEX, and ClinicalBioBERT for MIMIC.

Table 7: Performance on the short-document BioASQ dataset (macro/micro F1 $\times$100) using BERT and ClinicalBioBERT encoders.

PTM	Baseline	BioASQ
BERT	Emb + MLP	68.6 / 83.3
	SimCSE	68.1 / 82.7
	ESimCSE	68.0 / 82.7
	CPE (ours)	**70.5 / 84.1**
ClinicalBioBERT	Emb + MLP	68.1 / 83.6
	SimCSE	69.1 / 83.3
	ESimCSE	68.8 / 83.2
	CPE (ours)	**71.3 / 84.4**

Embeddings Quality Figure 2 shows the t-SNE projections of the CPE embeddings compared to the SimCSE baseline using the LegalBERT encoder on SCOTUS. As we can see, CPE demonstrates a higher quality of legal act encoding, as evidenced by more compact clusters. To quantify

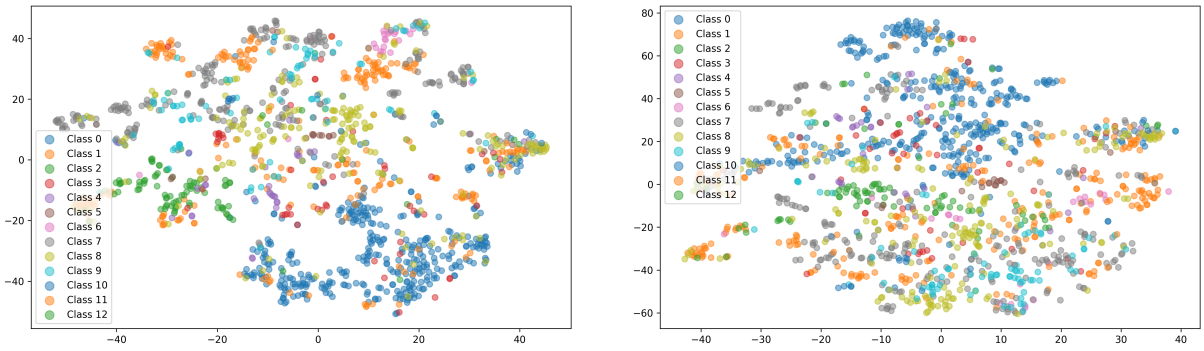

Figure 2: t-SNE visualization on the SCOTUS dataset with corresponding legal class

.

the comparison of visualized embeddings, we applied the DBScan clustering algorithm to the t-SNE projections. We evaluated clustering quality using completeness and homogeneity measures. As shown in Figure 2, the completeness and homogeneity scores for CPE are 0.31 and 0.38, respectively, compared to 0.23 and 0.32 for SimCSE. This indicates a clear improvement in topic separation using the projected embeddings from CPE.

Training Time Table 8 presents the training time required by each model on the ECHR and SCOTUS datasets. Self-supervised contrastive learning using a CPE encoder requires approximately 3 h on ECHR and 1.5 h on SCOTUS dataset. In contrast, SimCSE and ESimCSE training takes place 4.5 on ECHR and 2.5 h on SCOTUS dataset. SimCSE, and ESimCSE require significantly longer training times due to their document-level postive pair and negative pair. Our CPE training is more efficient because each positive and negative pair involves a single sampled chunk paired with an aggregated document context rather than encoding the entire document twice, reducing both computation and memory overhead. Furthermore, for the evaluation of the document embedding in the downstream with an MLP head (1.78M trainable parameters), the training is light, taking less than 10 minutes per epoch on our hardware.

Table 8: Training time (per seed) across datasets.

Model	ECHR (T)	SCOTUS (T)
CPE	3 h	1.5 h
SimCSE	4.5 h	2.5 h
ESimCSE	4.5 h	2.5 h

Prompting LLaMA on long legal documents Table 9 demonstrates that the zero-shot prompting model LLaMA-3 (8B) underperforms compared to embedding models (HBERT+MLP), primarily due to the extensive length of its prompts. Although the literature suggests that few-shot demonstrations can enhance model performance, the limited context window presents significant challenges when handling long documents, each averaging 4,096 tokens. Incorporating even one-shot examples into the prompt consumes nearly all available space, leaving insufficient room for the actual query. Furthermore, the LLaMA model tends to over-predict a limited number of classes while rarely predicting others, leading to imbalanced classification outcomes, as indicated by a low macro-F1 score.

Table 9: Classification performance of the zero-shot LLaMA-3 8B model on the ECHR and SCOTUS datasets (same token budget as HBERT+MLP).

Model	ECHR		SCOTUS	
	macro-F1	μ-F1	macro-F1	μ-F1
HBERT+MLP	54.77	66.98	54.56	66.78
LLaMA3-8B-Instruct	15.39	22.54	0.369	24.91

7. Conclusion

We propose the Chunk Prediction Encoder (CPE), a self-supervised method for long-document representations via chunk–context entailment. CPE combines contrastive learning with entailment prediction to capture both local and global semantics, and yields consistent gains over strong contrastive baselines (SimCSE variants) and fine-tuned long-text models (Hi-LegalBERT, LegalLongformer, HAT). Across five legal and biomedical benchmarks, CPE improves macro-F1 by up to 4 points on average while remaining efficient and architecture-agnostic.

Analyses indicate that chunk–context compatibility induces richer document semantics and transfers well to short texts, supporting directional entailment learning as a robust representation strategy. Code, configs, and checkpoints will be released under an open license for reproducibility and downstream use.

Limitations

While CPE offers substantial gains in long-document representation, several limitations remain. First, the approach assumes that document segmentation into coherent chunks is meaningful; performance may degrade on noisy or unstructured texts (e.g., social media, OCR). Second, the computational efficiency—though improved—still scales linearly with the number of chunks, which may limit applications to documents exceeding 10k tokens. Third, the current evaluation focuses on English-language corpora; multilingual and cross-lingual extensions will require adaptation of pre-trained backbones and chunking heuristics. Finally, CPE is evaluated primarily on classification tasks; applying it to retrieval or generative settings remains an open direction.

Ethics Statement

Our study relies exclusively on publicly available, de-identified datasets in the legal and biomedical domains: ECHR, SCOTUS, EURLEX, MIMIC-III, and BioASQ. All data usage complies with the licenses of the original repositories. Although these datasets have been widely adopted in NLP research, they may still reflect historical or societal biases. We emphasize that CPE is a general representation learning framework and should not be used to automate legal or medical decisions without expert oversight. Model outputs should be interpreted as assistive signals rather than authoritative judgments. We encourage the community to audit models trained with CPE for fairness, bias propagation, and potential misuse in sensitive contexts.

Acknowledgements

This work was supported by ANR-22-CE23-0002 ERIANA and was granted access to the HPC resources of IDRIS under the allocation 2026-AD011013338 made by GENCI.

8. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022a. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022b. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Charles Condevaux and Sébastien Harispe. 2023. Lsg attention: Extrapolation of pretrained transformers to long sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 443–454. Springer.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 130–142.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Daniel Saggau, Mina Rezaei, Bernd Bischl, and Ilias Chalkidis. 2023. Efficient document embeddings via self-contrastive bregman divergence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12181–12190.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 848–853, Online. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *COLING*, pages 3898–3907.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023a. Distillcse: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8153–8165.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023b. Simcse++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 12028–12040.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proc. EMNLP 2024 (Industry Track)*, pages 1393–1412.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space (arccse). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4892–4903.

Xinping Zhao, Yan Zhong, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Dongfang Li, Baotian Hu, and Min Zhang. 2025. FunnelRAG: A coarse-to-fine progressive retrieval paradigm for RAG. In *Findings of NAACL-HLT 2025*, pages 3029–3046.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.