Coupling Local Context and Global Semantic Prototypes via a Hierarchical Architecture for Rhetorical Roles Labeling

Anonymous ACL submission

Abstract

Rhetorical Role Labeling (RRL) aims to identify the functional role of each sentence within a document, a task critical for discourse understanding in domains such as law, medicine, and science. While hierarchical models capture local, intra-document dependencies effectively, they struggle to model global, corpus-level regularities. To bridge this gap, we propose two methods that couple local context with global representations in the form of semantic prototypes. **Prototype-Based Regularization (PBR)** learns soft prototypes through a distance-based auxiliary loss to structure the latent space. **Prototype-Conditioned Modulation (PCM)** constructs a priori prototypes from the corpus and injects them during both training and inference. We also introduce SCOTUS-LAW, the first dataset of U.S. Supreme Court opinions annotated with rhetorical roles at three levels of granularity: *category*, *rhetorical function*, and *step*. Experiments across legal, medical, and scientific benchmarks demonstrate that modeling both local and global perspectives leads to consistent gains over strong baselines, particularly on low-frequency roles, achieving an average gain of ∼4 points in Macro-F1.

1 Introduction

Rhetorical Role Labeling (RRL) is the task of classifying each sentence according to its semantic role within a document. Since a sentence's meaning is often shaped by its surrounding context, RRL is particularly useful in structured texts such as legal cases. Identifying key rhetorical components (e.g., ANNOUNCING or ANALYSIS; see Figure 1) benefits downstream tasks such as information retrieval (Neves et al., 2019; Safder and Hassan, 2019) and document summarization (Kalamkar et al., 2022; Muhammed et al., 2024).

Initially, RRL was treated as a sentence-level classification problem, ignoring contextual dependencies between sentences (Walker et al., 2019). This perspective later evolved into modeling the task as sequence labeling (Bhattacharya et al., 2023a). More recently, deep learning techniques have been applied across various legal systems, including Japanese (Yamada et al., 2019) and Indian courts (Bhattacharya et al., 2023b; Kalamkar et al., 2022; Nigam et al., 2025). These methods typically employ hierarchical architectures to capture the sequential nature of long documents and model intra-document dependencies, resulting in a representation grounded in local context. This approach has become the de facto standard in RRL.

However, these architectures do not account for global patterns shared across documents, which are especially valuable for fine-grained roles, such as the RATIO OF THE DECISION, often confused with semantically related roles like ANALYSIS or RULING BY THE COURT. Prototype learning (Snell et al., 2017) provides a principled way to address this limitation by learning global representations that serve as semantic anchors for each label. This paradigm has shown strong performance across various NLP tasks, including named entity recognition (Huang et al., 2023), relation classification (Yu et al., 2022), and legal-specific tasks such as citation prediction (Luo et al., 2023).

Motivated by these advances, we propose to combine local context with global representations, defined as semantic prototypes. To the best of our knowledge, no prior work has addressed this objective in the context of RRL, particularly within hierarchical architectures.

Our main contributions are as follows:

- We introduce two semantic prototype-based methods: **Prototype-Based Regularization (PBR)**, that encourages sentence embeddings to align with their corresponding prototypes via an auxiliary distance-based loss; and **Prototype-Conditioned Modulation (PCM)**, which builds a priori prototypes from the corpus and injects them through dedicated mod-

Segment	Discursive category	Rhetorical function	Attributes
Justice GINSBURG delivered the opinion of the Court.	Announcing	Announcing	
Hansberry v. Lee, 311 U.S. , 40, 61 S. Ct. , 85 L. Ed. (1940).	Sources of authority	Quoting	SCOTUS decision
In this case, we consider for the first time whether there is a "virtual representation" exception to the general rule against precluding nonparties.	Setting the scene	Presenting jurisdiction	Legal question(s)
Fairchild and the FAA conceded that Taylor had not participated in Herrick's suit.	Analysis	Recalling	An argument, Present case, Petitioner
Accordingly, the decision of the Court of Appeals is reversed and the case is remanded with direction that judgment be entered for the United States.	Resolution	Giving the holding of the Court	

Figure 1: An example of a segment from a legal document in our SCOTUS-LAW corpus, annotated with discursive categories, rhetorical functions, and attributes to compose the full hierarchical label structure (steps).

ules during both training and inference.

- To address the lack of document-level resources for RRL, we release SCOTUS-LAW, a manually annotated corpus of U.S. Supreme Court opinions segmented into rhetorical roles at three levels of granularity.
- We perform a large-scale evaluation on seven benchmark datasets spanning three specialized domains: legal, medical, and scientific.

To support reproducibility and further research, we release both our code and dataset under an open-source license[1].

2 Related Works

2.1 Rhetorical Role Labeling Approaches

Early RRL approaches relied on traditional machine learning algorithms with hand-crafted features (Ruch et al., 2007; McKnight and Srinivasan, 2003; Lin et al., 2006). A key advancement came with the introduction of neural architectures by Cohan et al. (2019), which leverage BERT (Devlin et al., 2019) to capture contextual dependencies. Recent state-of-the-art methods build on this foundation by adopting hierarchical architectures (Jin and Szolovits, 2018; Brack et al., 2024), which encode documents at multiple levels to produce contextualized sentence representations suited for rhetorical function classification. More recently, several studies have explored ways to enrich these representations through strategies such as modified pretraining objectives (Belfathi et al., 2025), contrastive learning (T.y.s.s. et al., 2024), and curriculum learning (T.y.s.s et al., 2024), extending beyond hierarchical encoding to enhance contextual understanding.

2.2 Rhetorical Role Labeling Corpora

RRL has been studied across various domains using sentence-level annotation of functional discourse roles. In the medical domain, PUBMED-20K-RCT (Dernoncourt et al., 2017) provides a large-scale corpus of abstracts from randomized controlled trials, where each sentence is labeled with a rhetorical role such as OBJECTIVE, METHODS, or RESULTS. Similarly, CS-ABSTRACTS (Cohan et al., 2019; Gonçalves et al., 2020) offers scientific abstracts with a similar rhetorical structure.

In legal NLP, recent work has shifted from abstracts to long documents. Corpora such as DEEPRHOLE (Bhattacharya et al., 2023b), LEGALEVAL (Kalamkar et al., 2022), and LEGALSEG (Nigam et al., 2025) annotate Indian case law with rhetorical roles including FACTS, ARGUMENTS, and ANALYSIS. These datasets are limited to the Indian legal system, reducing their applicability to other common law jurisdictions. To our knowledge, no RRL corpus covers U.S. Supreme Court decisions.

2.3 Prototype-Based Learning

While hierarchical architectures in RRL capture intra-document context, they often overlook rhetorical regularities across documents that could serve as inductive signals. Prototype-based learning addresses this by aligning instances with similar discourse roles to shared semantic representations, typically encoded as vector prototypes (T.y.s.s. et al., 2024). Originally introduced by Snell et al. (2017), prototypical networks compute class prototypes as the mean of support examples and classify new instances based on embedding proximity. This approach has shown strong results in emotion recognition (Song et al., 2022), relation extraction (Chen et al., 2023), and named entity recognition (Huang et al., 2023; Wu et al., 2023), where prototypes capture class-level semantics and support generalization under limited supervision.
Despite these advances, prototype-based methods remain underexplored in discourse-level classification tasks like RRL. As far as we know, there are no studies on how to combine local and global representations within the hierarchical architectures.

[1] https://anonymous.4open.science/r/IJCNLP-AACL2025

3 Methodology

In this section, we first describe the task definition of RRL in § 3.1. This is followed by a brief outline of the backbone hierarchical architecture adopted in this study (§ 3.2). Finally, we introduce our global semantic prototype-based methods, as illustrated in Figure 2, namely Prototype-Based Regularization (§ 3.3) and Prototype-Conditioned Modulation (§ 3.4).

3.1 Task Definition

Given a document $x = \{x_1, x_2, \ldots, x_m\}$ with m sentences as the input, where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ represents the i^{th} sentence containing n tokens, and x_{jp} refers to the p^{th} token in the j^{th} sentence, the task of rhetorical role labeling is to predict a sequence $y = \{y_1, y_2, \ldots, y_m\}$, where y_i is the rhetorical role corresponding to sentence x_i, and $y_i \in \mathcal{Y}$, which is the set of predefined rhetorical role labels.

3.2 Backbone Hierarchical Architecture

All our experiments are based on the state-of-the-art RRL model, the Hierarchical Sequential Labeling Network (Jin and Szolovits, 2018; Brack et al., 2024), widely adopted as a baseline in prior work(Kalamkar et al., 2022; T.y.s.s. et al., 2024). This architecture is designed to capture local context by modeling intra-document dependencies at multiple levels. Each sentence s_{ij} is first encoded independently using a BERT (Devlin et al., 2019), producing a sequence of contextualized token embeddings. These are passed through a Bi-LSTM (Hochreiter and Schmidhuber, 1997) and an attention-pooling mechanism (Yang et al., 2016) to obtain fixed-size sentence vectors. A second Bi-LSTM then contextualizes these vectors with surrounding sentences, yielding enriched sentence representations. Finally, a Conditional Random Field (CRF) layer predicts the optimal sequence of role labels (see Appx. A for more details).

3.3 Prototype-Based Regularization

To extend the hierarchical architecture with global information beyond document-local context, we introduce Prototype-Based Regularization (PBR). This method integrates trainable soft prototypes as representative anchors for rhetorical roles. These prototypes reside in the same embedding space as sentence vectors and are optimized globally across documents. Rather than altering the architecture, PBR adds an auxiliary constraint that encourages each sentence embedding to align with its nearest prototype, using a distance-based metric. This guides the representation space toward corpus-level rhetorical patterns.

Following Zhang et al. (2022); Ming et al. (2019), we define a total loss combining standard classification with two prototype-driven regularization terms: the first enforces proximity between sentences and relevant prototypes; the second encourages separation among prototypes to reduce redundancy in the latent space.

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{task}}}_{\text{cross-entropy}} + \lambda_{\text{prox}} \underbrace{\mathcal{L}_{\text{prox}}}_{\text{prototype proximity}} - \lambda_{\text{div}} \underbrace{\mathcal{L}_{\text{div}}}_{\text{prototype diversity}} \tag{1}$$

where $\lambda_{\text{prox}}, \lambda_{\text{div}} \geq 0$ are hyperparameters controlling the contribution of each auxiliary term.

Task loss $\mathcal{L}_{\text{task}}$ is the standard cross-entropy computed between the model's prediction $\hat{y}_{y_{ij}}$ and the gold label y_{ij} for each sentence s_{ij}:

$$\mathcal{L}_{\text{task}} = -\sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \hat{y}_{y_{ij}}(s_{ij}). \tag{2}$$

Prototype-proximity loss $\mathcal{L}_{\text{prox}}$ pulls every sentence embedding $\mathbf{h}_{ij}$ toward its nearest prototype P_k among the Q learnable prototypes:

$$\mathcal{L}_{\text{prox}} = \frac{1}{T} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \min_{k \in \{1,\ldots,Q\}} d\big(\mathbf{h}_{ij},\, P_k\big), \tag{3}$$

where $T = \sum_{i=1}^{M} N_i$ is the total number of sentences.

Prototype-diversity loss $\mathcal{L}_{\text{div}}$ encourages the prototypes to spread out, reducing redundancy:

$$\mathcal{L}_{\text{div}} = \frac{2}{Q(Q-1)} \sum_{\substack{k,l \in \{1,\ldots,Q\} \\ k \neq l}} d\big(P_k,\, P_l\big). \tag{4}$$

3.4 Prototype-Conditioned Modulation

While PBR introduces soft alignment constraints without altering the architecture, Prototype-Conditioned Modulation (PCM) directly integrates global representations into the model's internal encoding process. PCM precomputes a set of prototype vectors from the training corpus and injects them into the hierarchical architecture via lightweight conditioning modules. These global signals modulate sentence representations during both training and inference. The approach comprises three stages: document sampling, prototype extraction, and prototype injection.

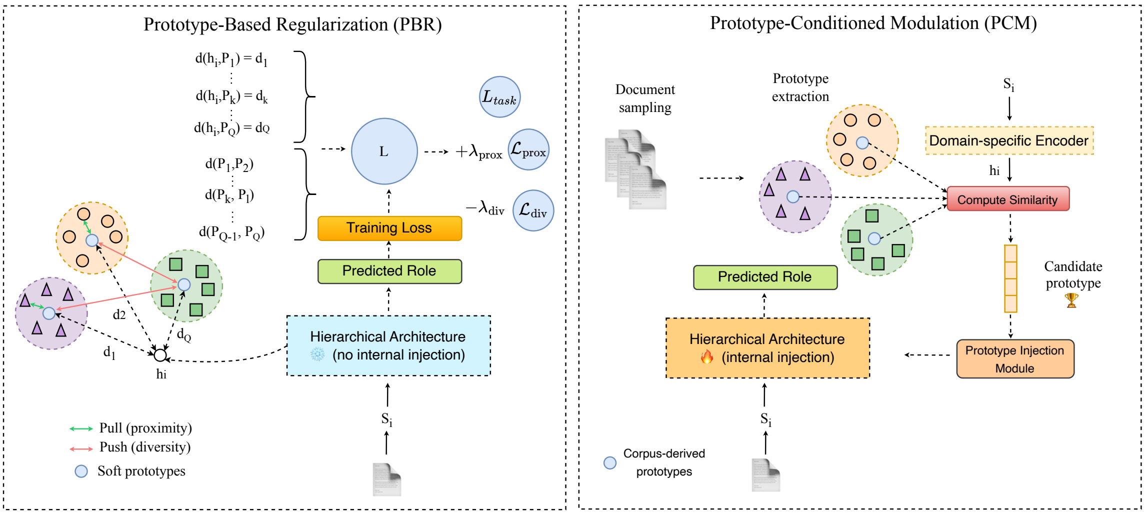


Figure 2: Illustration of our methods for injecting global representations into hierarchical architectures. PBR (left) learns soft prototypes jointly with the model to regularize the latent space. PCM (right) dynamically injects precomputed role prototypes during encoding via modulation mechanisms.

Document sampling A key design decision is whether to derive prototype representations from the entire training corpus or from a document subset, as using all documents may introduce semantic noise and reduce prototype relevance. We evaluate three strategies: (1) *Full Corpus*, which includes all training documents; (2) *Random sampling*, which selects a uniform subset; and (3) *Supervised sampling*, which clusters semantically similar documents using embeddings and derives prototypes per cluster[2].

Prototype extraction Given a sentence under consideration, we first identify a set of candidate documents and extract global representations for each rhetorical role in the form of prototype vectors. Each sentence s_{ij} is embedded using a domain-specific BERT model suitable for the evaluation dataset, producing a fixed-length vector $\mathbf{h}_{ij} \in \mathbb{R}^d$. For each role $r \in \mathcal{Y}$, we compute a prototype $\mathbf{p}_r$ by averaging the embeddings of all sentences $\mathcal{S}_r$ annotated with r in the selected document pool:

$$\mathbf{p}_r = \frac{1}{|S_r|} \sum_{s_{ij} \in S_r} \mathbf{h}_{ij}. \qquad (5)$$

Prototype injection Once the global representations for each role are computed, we inject them into the hierarchical architecture during both training and inference. For each sentence s_{ij}, we compute its cosine similarity to all prototypes $\{\mathbf{p}_r\}$ and select the closest one. Given the sensitivity of neural models to external knowledge integration (Fu et al., 2023), we explore five conditioning strategies drawn from prior work: *Linear Fusion* (Bu et al., 2023), *Conditional Layer Normalization* (Lee et al., 2021), *Gated Residual Addition* (Tsur and Tulpan, 2023), *Feature-wise Linear Modulation* (Ahrens et al., 2023), and *Cross-Attention Fusion* (Zhang et al., 2024). See Appx. D for further details.

4 The SCOTUS-LAW Corpus

We introduce SCOTUS-LAW, the first publicly available English-language dataset of U.S. Supreme Court decisions annotated with rhetorical role segmentation. This resource expands the limited set of benchmarks available for the RRL.

4.1 Corpus Compilation

We collected decisions from CourtListener[3], an open-access legal case repository. Our sampling strategy considered three key dimensions: **(1) Temporal coverage:** Cases span 1945–2020 to capture historical variation. **(2) Author diversity:** Opinions from 38 justices reduce authorial bias and reflect diverse reasoning styles. **(3) Thematic coverage:** K-means clustering over a broad set of decisions yields 18 thematic groups.

To balance these aspects, we selected representa-

[2] For the supervised variant, we use OpenAI's `text-embedding-3-small` https://platform.openai.com/docs/guides/embeddings/embedding-models, which supports sequences up to 8,192 tokens for full-document representation. Each document is encoded and grouped via K-Means clustering (Ahmed et al., 2020), with the optimal number of clusters selected using the Silhouette score, computed per evaluation dataset.

[3] https://www.courtlistener.com/

Corpus-level statistics			
Statistic	**Train**	**Dev**	**Test**
# Documents	144	18	18
Total # Sentences	21,396	2,450	2,481
Avg. # Sentences / Doc	148.58	136.11	137.83
Avg. # Tokens / Sentence	22.95	21.43	22.15

Sentence distribution by rhetorical function	
Label	**Total (percentage)**
Recalling	8,119 (30.8%)
Quoting	6,441 (24.5%)
Presenting jurisdiction	4,941 (18.8%)
Stating the Court's reasoning	3,198 (12.1%)
Describing	955 (3.6%)
Giving the holding of the Court	760 (2.9%)
Citing	644 (2.4%)
Rejecting arguments/a reasoning	490 (1.9%)
Announcing	344 (1.3%)
Granting certiorari	182 (0.7%)
Giving instructions to competent courts	105 (0.4%)
Accepting arguments/a reasoning	103 (0.4%)
Evaluating the impact of the decision	45 (0.2%)

Table 1: Descriptive statistics for the SCOTUS-LAW dataset at the rhetorical function level.

tive cases from the most prolific justices in each theme, resulting in 180 annotated decisions.

4.2 Annotation Scheme

Our annotation scheme builds on Lavissière and Bonnard (2024), which focuses on rhetorical structures in U.S. legal decisions. As in prior work (Kalamkar et al., 2022; Nigam et al., 2025), annotations are applied at the sentence level. Each sentence receives a *step* label, denoting its function in legal reasoning and its role within the broader argumentative structure. We follow Lavissière and Bonnard (2024) in applying the annotation at three levels of granularity (Figure 7 in Appendix).

Step = Discursive Category + Rhetorical Function + Optional Attributes

Discursive categories. These reflect the overall structure of SCOTUS opinions and include five main categories:

- **Setting the scene**: background information and procedural history;
- **Analysis**: reasoning and justification of the Court's decision;
- **Resolution**: the outcome or final ruling;
- **Sources of authority**: references to legal sources such as precedent or statutes;
- **Announcing**: textual elements marking structural transitions.

Rhetorical functions. These specify the communicative role played by each segment within its discursive category. They include argumentative roles such as justification, evaluation, comparison, or appeal to authority.

Attributes. To refine the rhetorical annotation, three optional attributes can be specified:

- **Type**: the nature of the content (e.g., cited authority, recalled facts);
- **Author**: the speaker or source of the argument (e.g., the Court, a dissenting justice);
- **Target**: whether the information pertains to the current case or another referenced case.

Table 1 reports statistics for rhetorical functions; See Appx. E for annotation details.

4.3 Inter-Annotator Agreement

Two legal experts independently annotated a subset of 18 Supreme Court opinions, covering $2,529$ overlapping sentence-level segments. Cohen's kappa (Rau and Shih, 2021) yielded a score of 0.67, indicating substantial agreement. Disagreements were resolved through discussion, and consensus labels were assigned. The adjudicated version serves as the reference for evaluation and quality control.

5 Experimental Setup

5.1 Datasets

We evaluate our methods across three domains. In the **legal** domain, we use our SCOTUSLAW dataset at three levels of rhetorical structure: $\text{SCOTUS}_{\text{Category}}$, $\text{SCOTUS}_{\text{RF}}$, and $\text{SCOTUS}_{\text{Steps}}$. We also include two Indian case law datasets: DEEPRHOLE (Bhattacharya et al., 2023b) and LEGALEVAL (Kalamkar et al., 2022). For the **medical** domain, we use PUBMED (Dernoncourt et al., 2017), a corpus of structured abstracts from randomized controlled trials. In the **scientific** domain, we evaluate on CS-ABSTRACTS (Gonçalves et al., 2020), which contains computer science research abstracts annotated for rhetorical structure (see Appx. C for statistics details).

5.2 PBR Hyperparameters

Following Chen et al. (2019), we use cosine similarity to compute distances d between sentence embeddings and prototypes. To control the granularity of the soft prototype space, we vary $Q \in \{2, 4, 8, 16, 32, 64\}$, as in Yang et al. (2018); Sourati et al. (2023). The auxiliary loss weights λ_{prox} and λ_{div} are tested over $\{0, 0.9, 10\}$, where

	Legal										Medical		Scientific	
	SCOTUS$_{\text{Category}}$		SCOTUS$_{\text{RF}}$		SCOTUS$_{\text{Steps}}$		LEGALEVAL		DEEPRHOLE		PUBMED		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1
▷ **Baseline**	82.22	88.35	61.36	78.81	46.70	63.21	78.82	90.94	44.24	50.51	87.01	91.09	68.55	75.08
▶ **PBR**	83.69	89.75	65.75	80.31	50.48	65.73	82.50	93.17	44.96	51.11	88.86	92.91	71.10	78.09
⋆ **PCM (Full Corpus)**	83.96	89.80	67.53	80.64	54.03	67.54	81.41	91.21	47.13	55.54	87.19	91.89	69.84	76.66
⋆ **PCM (Random Sampling)**	83.93	89.70	67.24	80.66	54.62	67.55	81.83	91.57	47.30	53.90	87.24	91.94	69.12	76.30
⋆ **PCM (Supervised Sampling)**	84.13	89.75	67.45	80.92	54.40	67.79	80.77	91.00	45.92	53.86	87.42	92.06	68.69	75.46
◇ **Upper Bound (Oracle)**	85.20	90.02	68.86	81.11	56.20	69.86	91.71	99.57	47.90	56.02	100.0	100.0	99.66	99.84

Table 2: Macro-F1 and Weighted-F1 scores across domains for the baseline, PBR, and PCM (with various sampling strategies). An upper-bound oracle is also included, selecting the optimal prototype post-hoc for each sentence. Results are averaged over three runs, ensuring statistical significance over the baseline at $p = 0.05$ and $p = 0.01$.

$\lambda = 0$ disables the constraint, 0.9 is a balanced setting from Das et al. (2022), and 10 enforces strong regularization.

5.3 PCM Hyperparameters

In supervised sampling, documents are clustered by semantic similarity. The number of clusters is tuned on the development set using the silhouette score over the range $[1, 10]$. For prototype extraction, we use `Legal-BERT-uncased` (Chalkidis et al., 2020) for legal data, and `SciBERT-uncased` (Beltagy et al., 2019) for medical and scientific domains.

6 Results and Discussion

6.1 Overall Performance

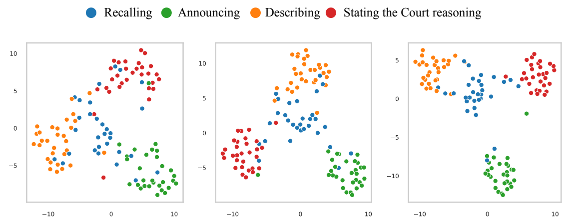
Results for the baseline and our methods combining local and global context via semantic prototypes are reported in Table 2.

Prototype-Based Regularization (PBR) consistently improves performance across all five legal datasets, with m-F1 gains from +1.5 on SCOTUS$_{\text{Category}}$ to +4.4 pts on SCOTUS$_{\text{RF}}$. While modest in absolute terms, these gains are statistically significant ($\sigma \leq 0.3$ over three runs), confirming the impact of the prototype mechanism beyond random variation. **Why does performance improve with finer annotations?** As labels become more fine-grained (SCOTUS$_{\text{Steps}}$), class boundaries blur—e.g., distinguishing subtypes within ANALYSIS. In such cases, prototypes act as semantic anchors that help disambiguate sentence meaning. The $+3.8$ gain suggests that the model increasingly relies on global cues when local context is not sufficient. **What about minority roles?** In SCOTUS$_{\text{RF}}$, the role STATING THE COURT'S REASONING represents under 5% of training data. PBR improves its F1 score from 63.2% to 69.5% (+6.3 pts), showing that gains extend beyond majority classes. This long-tail benefit echoes findings in multilingual NER (Huang et al., 2023), where prototype regularization narrows the gap between frequent and rare labels.

On the LEGALEVAL dataset, which is characterized by annotation ambiguity and challenging rhetorical distinctions (Kalamkar et al., 2022), PBR still improves performance, reaching 82.5%. Most gains come from reducing confusion between semantically overlapping roles, particularly legal analysis and factual issue descriptions, which together account for over 40% of baseline errors.

Prototype-Conditioned Modulation (PCM) which injects global representations from the training corpus, achieves the highest m-F1 across all settings. The largest gain appears on SCOTUS$_{\text{Steps}}$, where performance increases from 46.70% to 54.03%, This suggests that conditioning hidden layers with global prototypes helps guide the encoder toward more discriminative regions of the embedding space.
Among the sampling strategies, supervised sampling yields the best results only on SCOTUS$_{\text{Category}}$, where labels are broad and rhetorical usage relatively consistent across documents. Here, clustering similar documents builds informative prototypes. However, this benefit fades on datasets like LEGALEVAL and DEEPRHOLE, where all strategies perform similarly. We attribute this to two factors: (i) retrieval is at document level, ignoring sentence-level rhetorical similarity and often producing mismatched prototypes; (ii) legal texts follow stable rhetorical patterns, making even randomly sampled documents useful despite noise.

To estimate the **upper bound** of prototype injection, we simulate an oracle that selects, for each test sentence, the prototype yielding the best prediction. This yields 91.71% m-F1 on LEGALEVAL, confirming the potential of prototypes for semantic alignment. More importantly, the gap with actual performance shows that **retrieval quality is now**



(a) Baseline (b) + PBR (c) + PCM

Figure 3: t-SNE projection of sentence embeddings under baseline, PBR, and PCM.

the main bottleneck. This highlights the need for retrieval-aware or trainable prototype selection, ideally guided by rhetorical similarity or discourse structure rather than surface-level features.

Generalization across domains Our approach generalizes beyond legal texts. PBR improves performance on both PUBMED and CS-ABSTRACTS, showing that structural regularization remains effective in domains with rhetorical structure, even in shorter texts. In contrast, PCM yields limited gains. Medical and scientific abstracts are shorter and less structurally varied, making prototype averaging less informative. Yet, oracle results—up to 99.66% m-F1 on CS-ABSTRACTS, confirm that PCM is effective when relevant prototypes are injected, emphasizing the role of retrieval quality.

6.2 Qualitative Analysis

To understand how semantic prototypes shape sentence representations, we visualize the latent space using t-SNE (Figure 3). In the baseline, clusters overlap heavily, especially between DESCRIBING and STATING THE COURT'S REASONING, which often co-occur due to semantic proximity. With PBR, these roles become more distinct, suggesting that regularization encourages a structure aligned with rhetorical roles. PCM exhibits even clearer, tighter clusters across roles, indicating that conditioning with retrieved prototypes yields more role-specific and discriminative embeddings. These visualizations support the idea that both methods improve role separability, and that prototype quality plays a central role in shaping the latent space.

6.3 Fine-grained Analysis

Table 3 shows that injecting global semantic prototypes substantially improves m-F1 overall (+5.40), though the effect varies by rhetorical functions. The largest gains are seen for ACCEPTING ARGUMENTS/A REASONING (+41.75) and GIVING THE HOLDING OF THE COURT (+6.98)—two roles that

Rhetorical Function	Baseline	+PCM	Δ (Gain)
Accepting arguments/a reasoning	15.40	57.15	+ 41.75
Announcing	68.98	76.93	+ 7.95
Citing	85.99	89.92	+ 3.93
Describing	61.04	61.41	+ 0.37
Evaluating the impact of the decision	0.00	0.00	0.00
Giving instructions to competent courts	52.18	56.01	+ 3.83
Giving the holding of the Court	74.63	81.61	+ 6.98
Granting certiorari	97.30	100.0	+ 2.70
Presenting jurisdiction	86.64	88.65	+ 2.01
Quoting	97.79	98.13	+ 0.34
Recalling	77.38	79.04	+ 1.66
Rejecting arguments/a reasoning	40.52	35.91	− 4.61
Stating the Court's reasoning	57.00	60.35	+ 3.35
Macro-F1	62.69	68.09	**+ 5.40**

Table 3: Role-wise F1 comparison: Baseline (only local) vs. PCM (local + global) on $\text{SCOTUS}_{\text{RF}}$.

Method	$\text{SCOTUS}_{\text{RF}}$	LEGALEVAL	PUBMED
Linear Fusion	80.89	91.62	91.91
Conditional Layer Norm	78.11	87.49	92.74
Cross-Attention Fusion	79.30	87.74	92.20
Feature-wise Linear Mod.	74.71	76.74	92.74
Gated Residual Addition	79.58	89.06	92.79

Table 4: W-F1 scores for prototype injection strategies. All variants share the same hierarchical encoder with PCM integration.

depend on discourse-level context. Sentences like *"The argument raised by the defendant is valid"* or *"The Court therefore holds. . . "* require understanding their position in the reasoning chain. In such cases, prototypes bring in relevant cues from similar decisions, guiding the model toward the correct label. By contrast, performance drops for REJECTING ARGUMENTS/A REASONING, a role often expressed through contrastive or negative phrasing (e.g., *"However, this claim must be dismissed"*). These subtle cues may be lost when prototype vectors average too many diverse examples, diluting critical signals and reducing precision. Finally, EVALUATING THE IMPACT OF THE DECISION remains unlearned, suggesting that the class is too rare for any method to model effectively.

6.4 Sensitivity to Prototype Injection

Table 4 shows that the impact of injection strategies varies by domain. In legal datasets such as $\text{SCOTUS}_{\text{RF}}$ and LEGALEVAL, *Linear Fusion* performs best, with a +2.63 m-F1 gain over *FiLM* on LEGALEVAL. Directly concatenating the prototype with the sentence embedding appears well suited to the structured nature of legal texts, where rhetorical roles follow predictable patterns. Conversely, flexible strategies like *FiLM* or *CLN*, which modulate representations dimension-wise, may interfere with latent spaces already aligned to legal structure, resulting in performance drops.

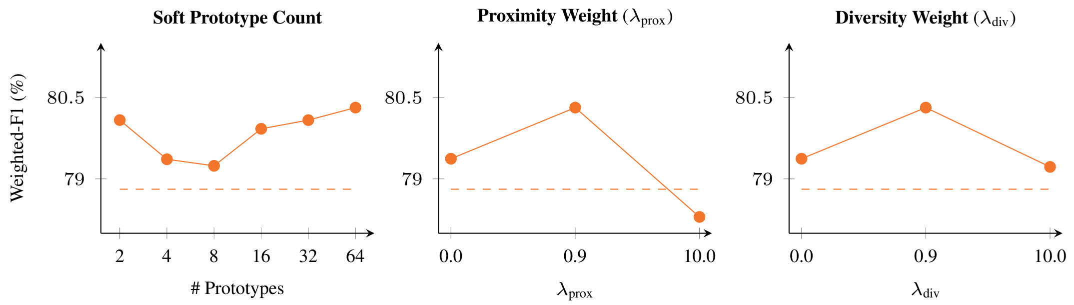


Figure 4: Effect of PBR hyperparameters on w-F1 at the SCOTUS$_{\text{RF}}$ Dashed lines indicate the baseline without prototypes.

On PUBMED, all methods perform similarly ($F1 > 92$), suggesting that prototype injection is less impactful. Here, *Gated Residual Addition* slightly outperforms others, likely because it preserves strong local signals while controlling the influence of the prototype. These findings confirm that no injection strategy is universally optimal. The best choice depends on the rhetorical structure of the text, the informativeness of prototypes, and how the model integrates external context.

6.5 Sensitivity to PBR Hyperparameters

We evaluate PBR sensitivity on SCOTUS$_{\text{RF}}$, focusing on three components: (1) the number of soft prototypes, (2) the proximity loss weight λ_{prox}, and (3) the diversity loss weight λ_{div}, as shown in Figure 4.

Prototype count. Performance is stable across values, with a slight improvement up to 16 prototypes. Beyond that, gains plateau, suggesting that few prototypes suffice to capture key rhetorical patterns, while higher counts may introduce redundancy.

Proximity loss λ_{prox}. A moderate value ($\lambda_{\text{prox}} = 0.9$) yields the best results, supporting the idea that proximity improves role consistency. Higher pressure ($\lambda_{\text{prox}} = 10.0$) degrades performance, likely due to overcompression of the embedding space.

Diversity loss λ_{div}. An intermediate value $\lambda_{\text{div}} = 0.9$ also performs best. It encourages separation among prototypes, improving class discriminability. Stronger regularization ($\lambda_{\text{div}} = 10.0$) slightly hurts performance, possibly by pushing prototypes too far from the data manifold.

6.6 Discussion

Prior work has primarily focused on modeling intra-document dependencies, what we refer to as local context through hierarchical architectures (Brack et al., 2024; T.y.s.s et al., 2024). Despite their success, these methods struggle with fine-grained rhetorical roles, likely due to the absence of corpus-level semantic grounding. This study aims to address that limitation by coupling local context with a global perspective, captured through semantic prototypes. To this end, we proposed two methods—PBR and PCM—that inject global signals into hierarchical encoders in distinct ways.

We chose to keep these methods separate to better assess their trade-offs. PBR is a lightweight regularization mechanism. In our experiments, it used ~ 30–40% less GPU memory and trained ~ 20–25% faster than PCM, making it attractive in resource-constrained settings. PCM, although more costly due to precomputed prototypes and conditioning modules, consistently delivered stronger gains, especially for underrepresented roles. It is better suited for scenarios where performance outweighs efficiency, such as legal domains or complex rhetorical hierarchies, as exemplified by our SCOTUS-LAW corpus.

7 Conclusion

This work shows that combining local context with global semantic prototypes significantly improves RRL, particularly for underrepresented roles. By introducing two methods—Prototype-Based Regularization (PBR) and Prototype-Conditioned Modulation (PCM)—we show that global signals can be effectively injected into hierarchical architectures to provide more semantically coherent representations. Beyond model performance, we contribute SCOTUS-LAW, the first U.S. Supreme Court dataset annotated at three rhetorical levels. This resource enables more granular evaluation and promotes research on legal NLP field. Future work should give priority to (1) to extend semantic prototyping to multilingual or cross-domain RRL, where generalization becomes even more challenging; (2) refining prototypes adaptively during inference to better align with evolving discourse structures.

8 Limitations

Although the proposed methods improve RRL performance, several limitations should be acknowledged to guide future improvements:

- The current task formulation assigns a single rhetorical label to each sentence. While this simplifies annotation and modeling, it may not account for the semantic complexity of long or compound sentences that express multiple rhetorical functions. Reformulating the task as multi-label classification could better reflect such cases.
- The approach operates at the sentence level. Segmenting at the phrase or clause level, and modeling rhetorical dependencies between segments, could lead to more fine-grained analysis.
- The study focuses exclusively on English corpora. Extending semantic prototyping to multilingual RRL raises challenges related to alignment, label transfer, and prototype sharing across languages with different rhetorical conventions.

9 Ethical considerations

This work proposes new methods and experiments aimed at advancing research in rhetorical role labeling, a foundational task in legal document processing. All experiments were conducted on publicly available datasets, including our introduced datasets. While these documents are not anonymized and may contain real names of involved parties, they are official court records released for public access. We do not anticipate any harm arising from our use of these datasets. Our research is intended to support the development of transparent and responsible AI tools for legal professionals. By improving the automation of rhetorical role labeling, we aim to facilitate legal text analysis and contribute positively to the broader goals of legal NLP.

References

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

Kyra Ahrens, Lennart Bengtson, Jae Hee Lee, and Stefan Wermter. 2023. Visually grounded continual language learning with selective specialization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7037–7054, Singapore. Association for Computational Linguistics.

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, and Richard Dufour. 2025. Is Selective Masking A Key to Improving Domain Adaptation for Masked Language Model? In *International Conference on Artificial Intelligence and Law*, Chicago, United States.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023a. DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, 31(1):53–90.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023b. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

Arthur Brack, Elias Entrup, Markos Stamatakis, Pascal Buschermöhle, Anett Hoppe, and Ralph Ewerth. 2024. Sequential sentence classification in research papers using cross-domain multi-task learning. *International Journal on Digital Libraries*, 25(2):377–400.

Yuqi Bu, Xin Wu, Liuwu Li, Yi Cai, Qiong Liu, and Qingbao Huang. 2023. Segment-level and category-oriented network for knowledge-based referring expression comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8745–8757, Toronto, Canada. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 7409–7422, Toronto, Canada. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. ProtoTEx: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2986–2997, Dublin, Ireland. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural networks for joint sentence classification in medical paper abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 694–700, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Fu, Yiming Zhang, Haobo Wang, Weikang Qiu, and Junbo Zhao. 2023. Revisiting the knowledge injection frameworks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10983–10997, Singapore. Association for Computational Linguistics.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Comput. Appl.*, 32(11):6793–6807.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yucheng Huang, Wenqiang Liu, Xianli Zhang, Jun Lang, Tieliang Gong, and Chen Li. 2023. PRAM: An end-to-end prototype-based representation alignment model for zero-resource cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3220–3233, Toronto, Canada. Association for Computational Linguistics.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mary C Lavissière and Warren Bonnard. 2024. Who's really got the right moves? Analyzing recommendations for writing American judicial opinions. *Languages*, 9(4):119.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102, Online. Association for Computational Linguistics.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 65–72, New York, New York. Association for Computational Linguistics.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4883–4898, Toronto, Canada. Association for Computational Linguistics.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *AMIA annual symposium proceedings*, volume 2003, page 440.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 903–913, New York, NY, USA. Association for Computing Machinery.

Akheel Muhammed, Hamna Muslihuddeen, Shalaka Sankar, and M Anand Kumar. 2024. Impact of rhetorical roles in abstractive legal document summarization. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135, Florence, Italy. Association for Computational Linguistics.

Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1129–1144, Albuquerque, New Mexico. Association for Computational Linguistics.

Gerald Rau and Yu-Shan Shih. 2021. Evaluation of cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of english for academic purposes*, 53:101026.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2-3):195–200.

Iqra Safder and Saeed-Ul Hassan. 2019. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, 119:257–277.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Oren Tsur and Yoav Tulpan. 2023. A deeper (autoregressive) approach to non-convergent discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12883–12895, Singapore. Association for Computational Linguistics.

Santosh T.y.s.s, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. HiCuLR: Hierarchical curriculum learning for rhetorical role labeling of legal documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7357–7364, Miami, Florida, USA. Association for Computational Linguistics.

Santosh T.y.s.s., Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306, Torino, Italia. ELRA and ICCL.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. *ASAIL@ ICAIL*, 2385.

Shuhui Wu, Yongliang Shen, Zeqi Tan, Wenqi Ren, Jietian Guo, Shiliang Pu, and Weiming Lu. 2023. MProto: Multi-prototype network with denoised optimal transport for distantly supervised named entity recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2361–2374, Singapore. Association for Computational Linguistics.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Neural network based rhetorical status classification for japanese judgment documents. In *Legal Knowledge and Information Systems*, pages 133–142. IOS Press.

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Robust classification with convolutional prototype learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3474–3482.

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.

Tianshu Yu, Min Yang, and Xiaoyan Zhao. 2022. Dependency-aware prototype learning for few-shot relation classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2339–2345, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiaotong Zhang, Xinyi Li, Feng Zhang, Zhiyi Wei, Junfeng Liu, and Han Liu. 2024. A coarse-to-fine prototype learning approach for multi-label few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2489–2502, Miami, Florida, USA. Association for Computational Linguistics.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. 2022. Protgnn: Towards self-explaining graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):9127–9135.

A Hierarchical Architecture Details

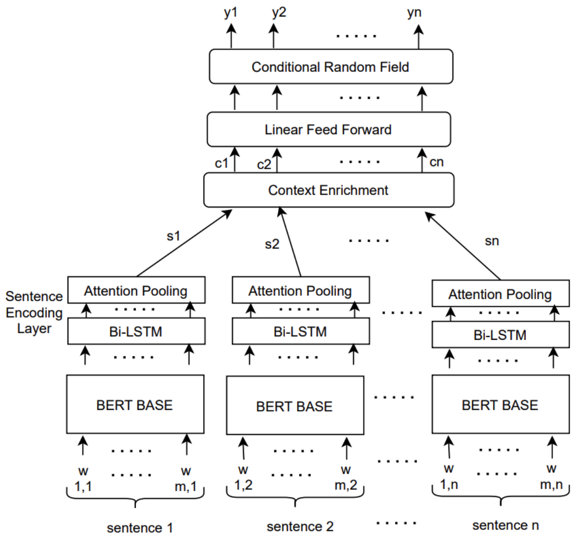


Figure 5: The hierarchical architecture.

All of our experiments are built on the state-of-the-art hierarchical architecture (Brack et al., 2024). Initially, each sentence s_{ij} is encoded independently with a BERT model (Devlin et al., 2019), producing a sequence of contextual token embeddings $\mathbf{h}_{ij} = \{\mathbf{h}_{ij1}, \mathbf{h}_{ij2}, \ldots, \mathbf{h}_{ijT_{ij}}\}$. These vectors are passed through a Bi-LSTM layer (Hochreiter and Schmidhuber, 1997), followed by an attention-pooling layer (Yang et al., 2016), to yield sentence representations $\mathbf{v}_{ij}$.

$$\mathbf{u}_{ijt} = \tanh\big(W_w \mathbf{h}_{ijt} + \mathbf{b}_w\big) \tag{6}$$

$$\alpha_{ijt} = \frac{\exp\big(\mathbf{u}_{ijt}^{\top}\mathbf{u}_w\big)}{\sum_{t'} \exp\big(\mathbf{u}_{ijt'}^{\top}\mathbf{u}_w\big)} \quad \& \quad \mathbf{v}_{ij} = \sum_{t=1}^{T_{ij}} \alpha_{ijt}\,\mathbf{h}_{ijt} \tag{7}$$

Here, W_w, $\mathbf{b}_w$, and $\mathbf{u}_w$ are trainable parameters. The sentence representations $\mathbf{v}_{ij}$ are then passed through a second Bi-LSTM to obtain contextualised embeddings $\mathbf{c}_{ij}$ that capture information from neighbouring sentences. Finally, the contextual vectors $\mathbf{c}_{ij}$ are fed to a Conditional Random Field layer, which predicts the optimal sequence of labels.

B Implementation Details

We follow the hyperparameters for the baseline as described in Brack et al. (2024). We use the BERT-*base* model to obtain the token encodings. We employ a dropout of 0.5, a maximum sequence length of 128, an LSTM dimension of 768, and an attention context dimension of 200. We perform a grid search over learning rates {1e-5, 3e-5, 5e-5, 1e-4, 3e-4} for 40 epochs, using the Adam optimizer (Kingma and Ba, 2014).

C Evaluation Datasets

In addition to evaluating our models on the proposed SCOTUS-LAW corpus, we conduct experiments on several established RRL benchmarks across the legal, medical, and scientific domains.

LegalEval (Kalamkar et al., 2022) consists of judgments from the Indian Supreme Court, High Court, and District Courts. It provides public training and validation splits with 184 and 30 documents, respectively, totaling 31,865 sentences (average of 115 per document), annotated with 13 rhetorical role labels. Due to the absence of a public test set, we train on the official training split and evaluate on the provided validation set.

DeepRhole (Bhattacharya et al., 2023b) includes 50 judgments from the Indian Supreme Court across five legal domains, annotated with 7 rhetorical roles. It comprises 9,380 sentences (average of 188 per document). We follow an 80/10/10 split at the document level for train/validation/test.

PubMed (Dernoncourt and Lee, 2017) contains 20,000 structured medical abstracts from randomized controlled trials. Sentences are automatically labeled by authors into five rhetorical roles: *Background*, *Objective*, *Methods*, *Results*, and *Conclusions*.

CS-Abstracts (Gonçalves et al., 2020) includes 654 abstracts from computer science literature, annotated via crowdsourcing into the same five rhetorical roles as PubMed. It is currently the most recent dataset for scientific rhetorical structure classification.

D Prototype Injection Strategies

We experiment with several strategies to inject global prototype representations into sentence encoders. Each method varies in the degree of control, parametrization, and how the prototype signal is merged with the original sentence representation. We describe below the five main approaches studied in our work.

Linear Fusion (Bu et al., 2023) This method concatenates the sentence and its corresponding

Dataset	Source	Domain	Language	# Docs	# Sents	Labels
SCOTUS$_{\text{Category}}$	Ours	Legal (U.S.)	English	180	26,327	5
SCOTUS$_{\text{RF}}$	Ours	Legal (U.S.)	English	180	26,327	13
SCOTUS$_{\text{Steps}}$	Ours	Legal (U.S.)	English	180	26,327	35
LEGALEVAL	Kalamkar et al. (2022)	Legal (India)	English	214	31,865	13
DEEPRHOLE	Bhattacharya et al. (2023b)	Legal (India)	English	50	9,380	7
PubMed	Dernoncourt and Lee (2017)	Medical	English	20,000	227,000	5
CS-ABSTRACTS	Gonçalves et al. (2020)	Scientific	English	654	7,385	5

Table 5: Evaluation datasets used in our experiments. SCOTUS is annotated at three hierarchical levels: category, rhetorical function, and steps.

Category		%(↓)	Rhetorical Function		% (↓)	Type		Target		Author		%(→)
Announcing	344	1.30	Announcing	344	1.30							1.30
Setting the scene	5.123	19.45	Granting certiorari	182	0.69							0.69
			Presenting jurisdiction	4.941	18.76	Adjudicated facts	2.283					8.67
						Lower court decision	1.192					4.52
						Context	467					1.77
						Other procedural events	412					1.56
						Parties' legal claims and arguments	363					1.37
						Legal question(s)	224					0.85
Sources of authority	8.041	30.54	Citing	6.442	2.44	SCOTUS decision	2.764					0.89
						Primary source of law	2.203					0.91
						Secondary source of law	1.474					0.63
			Describing	955	3.62	Primary source of law	771					2.92
						Secondary source of law	159					0.60
						Established practices or cultural norms	25					0.09
			Quoting	644	24.46	SCOTUS decision	235					10.49
						Primary source of law	241					8.36
						Secondary source of law	168					5.59
Analysis	11.910	45.23	Stating the Court's reasoning	3.198	12.14							12.14
			Rejecting arguments/a reasoning	490	1.86							1.86
			Accepting arguments/a reasoning	103	0.39							0.39
			Recalling	8.119	30.83	A SCOTUS opinion	2.160					8.20
						A primary source	1.781					6.76
						A secondary source	359					1.36
						An established practice or cultural norm	1.199					4.55
						An adjudicated fact or procedural event	1.447	Present case	1.152			4.37
								Another case	295			1.12
						Legal question(s)	182	Present case	147			0.55
								Another case	35			0.13
						An argument	991	Present case	967	Petitioner	413	1.64
										Respondent	513	1.94
										Dissenting justice(s)	22	0.08
								Another case	24			0.09
Resolution	910	3.45	Giving the holding of the Court	760	2.88							2.88
			Giving instructions to competent courts	105	0.39							0.39
			Evaluating the impact of the decision	45	0.17							0.17
Total	26.328											

Table 6: Final Annotation Scheme: Comprising 5 Categories, 13 Rhetorical Functions, and 24 Attributes (Types, Targets, and Authors). Counts of Text Segments are Provided, with Distributions Displayed at the Category Level (↓), Rhetorical Function Level (↓), and Step Level (→).

prototype vector, followed by a linear projection layer to recover the original embedding dimension. While simple and fully parametric, this technique may dilute the prototype signal due to compression.

Conditional Layer Normalization (CLN) (Lee et al., 2021) The sentence is first normalized (zero mean, unit variance), and the prototype generates two vectors γ (gain) and β (bias) that re-scale and shift each dimension of the sentence embedding. This conditioning allows for fine-grained recalibration informed by prototype semantics.

Gated Residual Addition (Tsur and Tulpan, 2023) The original sentence embedding is preserved, and a prototype-based residual is added with a learned gate vector $g \in [0, 1]^d$ that controls per-dimension contribution. If g closes, the model reverts to the baseline representation; if it opens, the prototype is effectively injected.

Feature-wise Linear Modulation (FiLM) (Ahrens et al., 2023) FiLM extends CLN by directly applying the prototype-derived γ and β vectors to modulate the sentence features ($\gamma \odot x + \beta$), without requiring prior normalization. This method is more flexible but less controlled than CLN, enabling adaptive influence of the prototype on the sentence.

Cross-Attention Fusion (Zhang et al., 2024) Here, the sentence acts as a query vector, attending to the prototype treated as key/value. Attention weights select relevant components from the prototype to be added to the sentence. This dynamic fusion allows for sentence-specific contextualization, adapting the contribution of the prototype to the input.

Each mechanism provides a different trade-off between interpretability, efficiency, and contextual adaptation. Our experiments show that no method is universally optimal, and the effectiveness often depends on the nature of the data and task.

E Annotation Scheme

E.1 Discursive Categories

The first level of our annotation schema defines five high-level rhetorical categories that segment each decision into major structural blocks. Below, we provide a brief description of each one:

Setting the scene. This category includes introductory paragraphs that present the case to the reader. Typical content includes information about the nature of the parties involved, their claims, the material facts of the case, the legal issue under examination, and the procedural history that brought the case before the Supreme Court.

Analysis. This category corresponds to the argumentative core of the decision. It usually follows the introductory section and precedes the final ruling. The content is primarily argumentative and captures the Court's reasoning in response to the parties' claims, justifying the interpretation and application of legal principles.

Resolution. This section contains the resolution of the legal issue, typically expressed through the final ruling issued by the majority opinion. While the announcement of the judgment is obligatory, it may also include instructions for lower courts or comments on the societal impact of the decision.

Sources of authority. This category gathers all explicit mentions of legal sources, whether written (e.g., case law, statutes, constitutional texts) or unwritten (e.g., doctrines or principles). Although such references appear throughout the decision, some judges explicitly dedicate specific portions of their opinion to outlining the sources that will later support their legal reasoning. *Note:* when a source is invoked directly within the reasoning process, it is annotated under the *Analyse* category rather than *Sources d'autorité*.

Announcing. This category includes structurally functional sentences that serve as rhetorical transitions. These statements do not carry substantive content themselves but signal the upcoming development of a new rhetorical step from one of the four other categories.

E.2 Rhetorical Functions

At the second level of annotation, we define thirteen rhetorical functions that capture the specific communicative intent of each sentence in the decision.

Granting certiorari. Assigned to sentences where the Court explicitly signals that it has agreed to review the case. These statements typically appear near the end of the factual and procedural summary, often preceding the articulation of the legal questions. Example: "We granted certiorari."

Presenting jurisdiction. Covers sentences that neutrally present elements of the case background. This function includes an attribute `Type` with five possible values: *Legal Issue*, *Facts of the Case*, *Other Procedural Elements*, *Arguments and Claims*, or *Broader Context*.

Quoting. Used for references to legal sources. The annotation includes a `Type` indicating the nature of the source: *Court Decision*, *Primary Source*, or *Secondary Source*.

Describing. Applied to paraphrases of legal sources, whether primary, secondary, or unwritten. The associated `Type` indicates the source category: *Primary Source*, *Secondary Source*, or *Unwritten Source of Authority*.

Citing. Used for direct quotations that include complete sentences or longer excerpts from legal sources. Types are the same as for *Quoting*.

Recalling. Captures sentences that refer back to previously mentioned legal sources, or that introduce sources in a way that supports the Court's reasoning. These recalls often include an interpretive dimension, contributing to argumentative development.

Accepting arguments/a reasoning. Marks agreement with a previously stated argument or reasoning, either from a party or another court.

Rejecting arguments/a reasoning. Indicates disagreement or refutation of a prior argument or line of reasoning, particularly when opposing the view of another court.

Stating the Court's reasoning. Assigned to all reasoning sentences that do not fall under more specific categories. This includes hypothetical reasoning, such as evaluating consequences of alternative outcomes.

Giving instructions to competent courts. Covers sentences in which the Court instructs lower courts or other legal bodies to act in accordance with the decision or to reconsider aspects of the case.

Giving the holding of the Court. Applies to sentences stating the legal conclusion reached by the Court (the holding), based on the material facts, including the final judgment.

Evaluating the impact of the decision. Used when the Court explicitly reflects on the consequences of its decision, either institutionally or societally.

Announcing. Marks structurally functional sentences that introduce an upcoming element of the decision or name the judge who authored the opinion.

E.3 Attributes

To enrich the rhetorical annotation while keeping the core label space concise, we introduce a small set of optional attributes. These attributes are designed to add interpretive nuance without changing the primary function assigned to a sentence. They are used selectively with certain rhetorical functions, such as *Recalling*, *Describing*, or *Presenting jurisdiction*.

- **Type** — indicates the nature of the content referenced or discussed (e.g., legal source, factual detail, procedural element);
- **Author** — specifies who is the originator of the argument or point of view (e.g., the Court, a party, or a dissenting opinion);
- **Target** — identifies whether the information concerns the case under review or refers to another precedent.

These attributes are optional but help clarify rhetorical intent, especially in ambiguous or multi-voiced legal discourse.

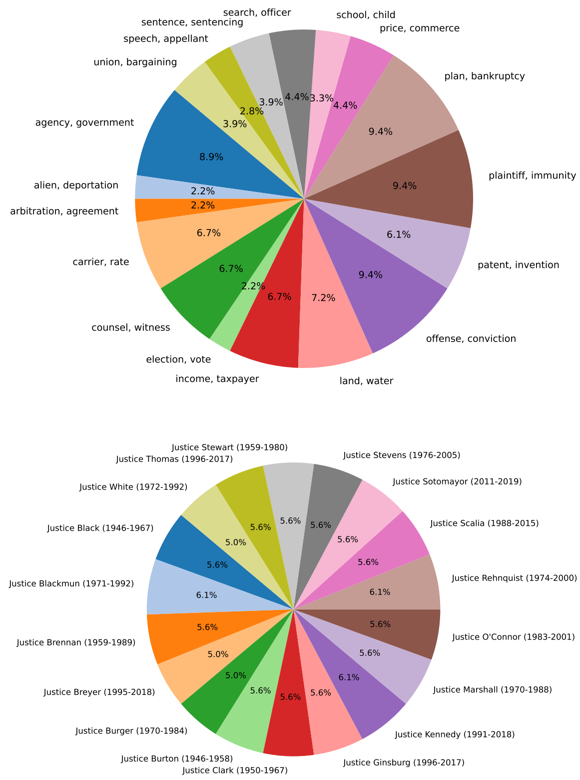


Figure 6: Topical, Temporal, and Authorial Diversity in our annotated corpus.

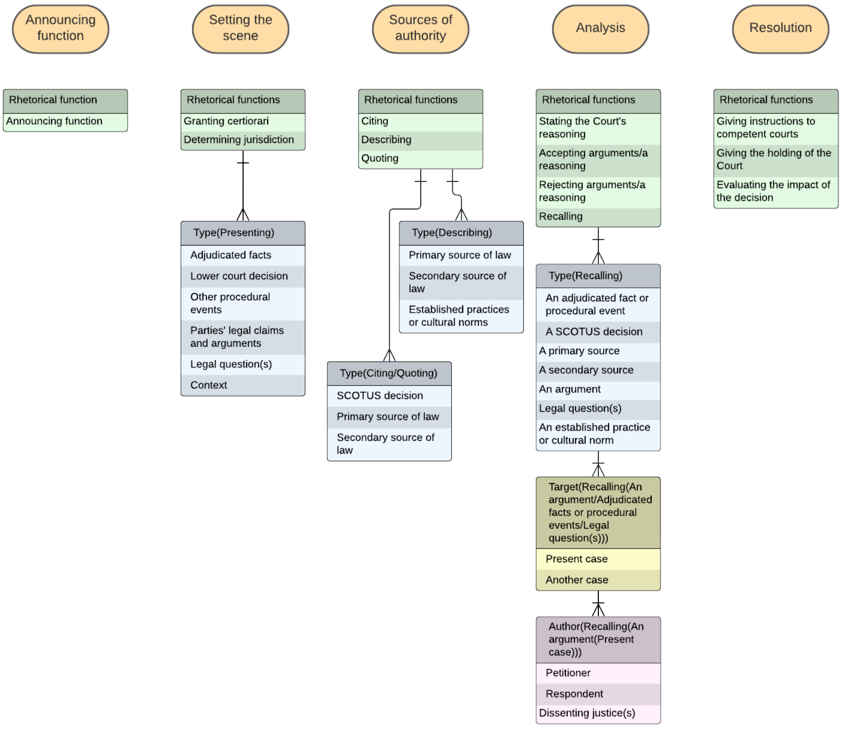


Figure 7: The final coding scheme is composed of 5 categories (ovals with orange background), 13 rhetorical functions (green rectangles) and 24 attributes (types in blue rectangles, target in the yellow rectangle, and author in the purple rectangle. The scheme reads from top to bottom: A step label is constructed by first choosing a category, then a rhetorical function, then if required, by combining attributes to complete the discursive information provided by the rhetorical function.