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

Anonymous ACL submission

Abstract

Rhetorical Role Labeling (RRL) identifies the functional role of each sentence in a document, a key task for discourse understanding in domains such as law and medicine. While hierarchical models capture local dependencies effectively, they are limited in modeling global, corpus-level features. To address this limitation, we propose two prototype-based methods that integrate local context with global representations. **Prototype-Based Regularization (PBR)** learns soft prototypes through a distance-based auxiliary loss to structure the latent space, while **Prototype-Conditioned Modulation (PCM)** constructs corpus-level prototypes and injects them during training and inference. Given the scarcity of RRL resources, we 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 on legal, medical, and scientific benchmarks show consistent improvements over strong baselines, with ~ 4 Macro-F1 gains on low-frequency roles. We further analyze the implications in the era of Large Language Models and complement our findings with expert evaluation.

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 rhetorical components (e.g., ANNOUNCING or ANALYSIS; see Figure 1) is useful for 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 framed as a sentence-level classification problem, overlooking contextual dependencies between sentences (Walker et al., 2019). Later, Brack et al. (2022) modeled it as sequence labeling with hierarchical architectures to capture intra-document dependencies and represent local context more effectively. This approach has since become the de facto standard in recent RRL studies (Kalamkar et al., 2022; Bhattacharya et al., 2023a; Nigam et al., 2025).

The challenge is that these architectures fail to capture global features shared across documents. We argue that leveraging this information could help resolve ambiguities between closely related roles, as noted by Kalamkar et al. (2022). In this context, prototype learning (Snell et al., 2017) serves as a principled way to address this limitation by learning global representations that serve as semantic anchors for each label. This approach has shown strong performance across various NLP tasks, including named entity recognition (Huang et al., 2023), relation classification (Yu et al., 2022), and legal citation prediction (Luo et al., 2023).

Building on these insights, we study how local context and global representations can be combined through semantic prototypes. To our knowledge, no prior work has pursued this objective in the context of RRL, particularly within a hierarchical framework. Our main contributions are:

- We introduce two semantic prototype-based methods: (i) **Prototype-Based Regularization (PBR)**, which aligns sentence embeddings with prototypes via an auxiliary distance-based loss; and (ii) **Prototype-Conditioned Modulation (PCM)**, which builds a priori prototypes from the corpus and injects them through dedicated modules during both training and inference.
- We release SCOTUS-LAW, the first manually annotated corpus of U.S. Supreme Court opinions segmented into rhetorical roles at three levels of granularity (see Figure 1).
- We assess generalizability of our approach on

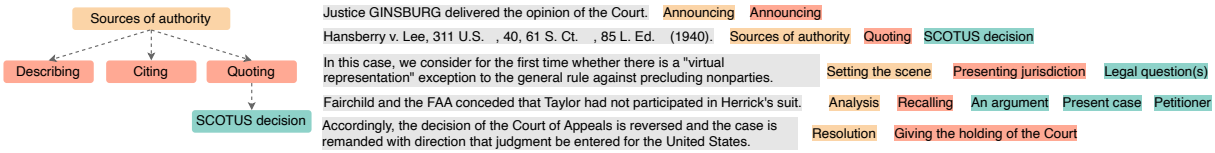


Figure 1: Example of a document segments in SCOTUS-LAW, annotated with discursive categories, rhetorical functions, and attributes, which together form the step annotation (full hierarchy in Figure 6).

seven benchmarks: legal datasets with long documents, and medical/scientific datasets with short abstracts.

- To position our work within the recent era of Large Language Models (LLMs), we evaluate three open-source models and complement this with a manual expert assessment to explain model behaviors.

We release both our code and dataset under an open-source license[1].

2 Related Works

2.1 Rhetorical Role Labeling Approaches

The story of RRL began with traditional machine learning using hand-crafted features (Ruch et al., 2007; McKnight and Srinivasan, 2003; Lin et al., 2006). Neural architectures marked a major shift, with BERT-based models capturing contextual dependencies (Cohan et al., 2019; Devlin et al., 2019). Recent methods build on this by adopting hierarchical architectures (Jin and Szolovits, 2018; Brack et al., 2024), encoding documents at multiple levels to produce contextualized sentence representations. More recent work enriches these representations through contrastive learning (T.y.s.s. et al., 2024), curriculum learning (T.y.s.s et al., 2024), and improved pretraining objectives (Belfathi et al., 2025), moving beyond hierarchical encoding toward deeper context modeling.

2.2 Rhetorical Role Labeling Corpora

RRL has been explored in multiple domains using sentence-level annotation of functional discourse roles. Starting by medical domain, PUBMED-20K-RCT (Dernoncourt et al., 2017) provides a large-scale corpus of abstracts, where each sentence is labeled with a role such as OBJECTIVE, METHODS, or RESULTS. Similarly, CS-ABSTRACTS (Cohan et al., 2019; Gonçalves et al., 2020) offers scientific abstracts with comparable rhetorical structures.

In legal NLP, research has shifted from short abstracts to full-length case 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 roles including FACTS, ARGUMENTS, and ANALYSIS. To our knowledge, no RRL corpus covers U.S. Supreme Court decisions, despite their importance for cross-jurisdictional legal applications (Curry and Miller, 2008).

2.3 Prototype-Based Learning

Corpus-level regularities that recur across documents are often overlooked by standard architectures. Prototype-based learning (Snell et al., 2017) addresses this limitation by representing each class with a prototype, defined as the mean embedding of its support examples, and classifying new instances by measuring their proximity to these prototypes. This paradigm has achieved 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 enable generalization under limited supervision. However, prototype-based methods remain underexplored in discourse-level classification tasks such as RRL, which could benefit from their ability to capture global semantic regularities.

3 Enriching Discourse Representations through Semantic Prototypes

This section first presents the task definition of RRL (§ 3.1), then outlines the backbone hierarchical architecture used in our study (§ 3.2). We finally introduce our global semantic prototype-based methods, Prototype-Based Regularization (§ 3.3) and Prototype-Conditioned Modulation (§ 3.4), illustrated in Figure 2.

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 ith sentence containing n tokens, and x_{jp} refers to the pth token in the jth sentence, the task of rhetorical role label-

[1] https://anonymous.4open.science/r/ACL2025-90E2

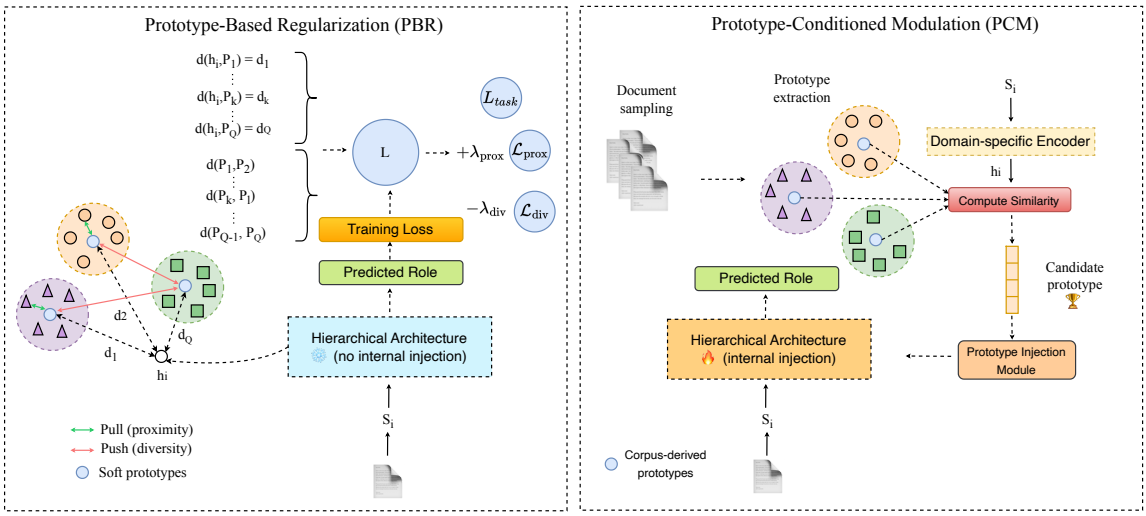


Figure 2: Illustration of our methods for injecting global representations into hierarchical architectures. PBR (left) learns soft prototypes jointly with the model to regularize the latent space. PCM (right) dynamically injects precomputed role prototypes during encoding via modulation mechanisms.

ing 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 Hierarchical Sequential Labeling Network (Jin and Szolovits, 2018; Brack et al., 2024), the state-of-the-art RRL model designed to capture local context by modeling intra-document dependencies at multiple levels. Each sentence s_{ij} is first encoded with BERT (Devlin et al., 2019), producing contextualized token embeddings, which are then passed through a BiLSTM (Hochreiter and Schmidhuber, 1997) and an attention-pooling mechanism (Yang et al., 2016) to obtain fixed-size sentence vectors. A second BiLSTM contextualizes these vectors with surrounding sentences, yielding enriched sentence representations, and a Conditional Random Field (CRF) layer finally predicts the optimal sequence of role labels (see Appendix C for details).

3.3 Prototype-Based Regularization

To enrich hierarchical architectures with global information, we introduce Prototype-Based Regularization (PBR), which incorporates trainable soft prototypes. These prototypes share the embedding space with sentence vectors and are optimized across documents. Rather than altering the backbone, PBR adds an auxiliary constraint that aligns each sentence embedding with its nearest prototype via a distance-based metric, steering the representation space toward corpus-level rhetorical patterns.

Following Ming et al. (2019); Zhang et al. (2022), the total loss combines standard classification with two prototype-driven regularization terms: the first enforces proximity between sentences and their relevant prototypes, while 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:

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

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:

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

3.4 Prototype-Conditioned Modulation

While PBR introduces soft alignment constraints without changing the backbone, Prototype-Conditioned Modulation (PCM) injects global representations directly into the encoding process. Prototype vectors are precomputed from the training corpus and incorporated into the hierarchical architecture through conditioning modules. These global signals modulate sentence representations during both training and inference. The procedure involves three stages: document sampling, prototype extraction, and prototype injection.

Document sampling The question is whether prototype representations should be derived from the entire training corpus or from a semantically related subset, since using all documents may introduce semantic noise and reduce prototype relevance (Lai et al., 2021). We evaluate three strategies: (1) *Full Sampling*, which uses 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 an input sentence, we first retrieve a set of candidate documents and extract global representations of 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 the set of sentences $\mathcal{S}_r$ annotated with r in the selected document pool:

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

Prototype injection. After computing global representations for each role, we inject them into the hierarchical architecture during both training and inference. For each sentence s_{ij}, we calculate its cosine similarity with the prototype set $\{\mathbf{p}_r\}$ and assign the closest one. Because models are sensitive to externally injected knowledge (Fu et al., 2023), we conducted an ablation study on five conditioning strategies inspired by prior work, with details provided in Appendix F.

[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.

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.

4 The SCOTUS-LAW Corpus

We introduce SCOTUS-LAW, the first publicly available dataset of U.S. Supreme Court decisions annotated with rhetorical roles. This corpus expands the scarce resources available for RRL and provides a new benchmark for discourse analysis in the legal domain.

4.1 Corpus Compilation & Size

We collected decisions from CourtListener[3], an open-access legal platform. Sampling covered three dimensions: **(1) Temporal:** 1945–2020; **(2) Author:** 38 justices; **(3) Thematic:** 18 groups. Representative cases were selected from the most prolific justices in each theme, resulting in 180 annotated decisions with 26,328 sentences, a size comparable to existing RRL datasets (Kalamkar et al., 2022).

4.2 Rhetorical Scheme Description

Our annotation scheme extends the framework of Lavissière and Bonnard (2024), which adopts the Swalesian approach (Swales, 1990) to the rhetorical structure of judicial opinions. As in prior

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

work (Kalamkar et al., 2022; Nigam et al., 2025), annotations are applied at the sentence level, but we add a further layer: each sentence is assigned a *step* label that captures its function in legal reasoning and its role within the broader argumentative structure. Following Lavissière and Bonnard (2024), the scheme operates at three levels of granularity (Figure 6 in Appendix).
Because of the large number of role definitions, full details are provided in Appendix A.

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 (statistics report in Table 1).

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.

4.3 Annotation Process

The process was designed in consultation with legal experts (law professors and legal practitioners).

Annotator Selection and Training Two law students were recruited through a multi-stage selection process. Each contributed 240 hours of annotation and received $1200 in compensation. Training was provided by two legal experts to ensure a solid understanding of the rhetorical role definitions.

Calibration In the initial stages, students differed in their interpretation of roles, requiring calibration. To align their understanding with expert definitions, they annotated 18 documents already labeled by experts. Sentences that diverged from the gold annotations were highlighted, and students revised their labels accordingly. This iterative process continued until their annotations reached expert-level agreement.

Adjudication To ensure consistency, each document was double-annotated and disagreements were resolved through adjudication. Annotators first attempted to reach consensus; unresolved cases were escalated to legal experts, who applied the guidelines and made the final decision. This process produced a coherent gold standard and refined the guidelines through feedback on ambiguous cases.

Annotation Quality Assessment Inter-annotator agreement was measured with Fleiss' Kappa (Fleiss et al., 2013), which improved from 0.67 in the initial tests to 0.72 after calibration, indicating good consistency. At the rhetorical function level, agreement was high for GRANTING CERTIORARI, QUOTING, and GIVING HOLDING TO THE COURT, but lower for STATING THE COURT'S REASONING and RECALLING, the most ambiguous functions. These results show the challenge of separating between sentences that merely recall external sources or arguments and those where the Court formulates its own reasoning, a boundary that even experts may interpret differently.

5 Results & Experiments

Table 2 reports the results (using Macro-F1 (mF1) and Weighted-F1 (wF1)) for the SOTA hierarchical model (§ 3.2), our proposed prototype-based methods, and an oracle that selects the gold prototype post-hoc for reference. Experiments were conducted across legal, medical, and scientific benchmarks, including our annotated SCOTUS-LAW corpus, and statistical significance was verified over five run evaluation. Full experimental details are provided in Appendix D.

5.1 Overall & Fine-grained Performance

Do prototype-based methods consistently improve results? Across all legal datasets, both PBR and PCM outperform the baseline. PBR

	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 Sampling)	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 different sampling strategies). An upper-bound oracle is also included, selecting the optimal prototype post-hoc for each sentence. † and ‡ indicate statistical significance over the baseline at 0.05 and 0.01, respectively.

yields consistent macro-F1 gains, from +1.5 on SCOTUS$_{\text{Category}}$ to +4.4 on SCOTUS$_{\text{RF}}$, with improvements statistically significant ($\sigma \leq 0.3$ over five runs). PCM achieves the best results on 4 of 7 tasks, most notably on SCOTUS$_{\text{Steps}}$, where scores increase from 46.70% to 54.03%. Regarding sampling, supervised sampling provides a slight advantage on broad labels such as SCOTUS$_{\text{Category}}$, but this effect disappears on datasets like LEGALEVAL and DEEPRHOLE, where all strategies perform similarly. This may stem from two factors: (i) retrieval operates at the document level, overlooking sentence-level rhetorical similarity and sometimes producing mismatched prototypes; and (ii) legal texts follow stable rhetorical patterns, so even randomly sampled documents provide useful signals despite added noise.

Takeaway 1. Global prototypes, whether used as a regularization signal (PBR) or as conditioning in the encoder (PCM), consistently improve performance beyond random variation.

Do prototypes extend benefits to minority and ambiguous roles? On SCOTUS$_{\text{RF}}$, the minority role STATING THE COURT'S REASONING(less than 5% of the data), improves with PBR from 63.2% to 69.5% (+6.3), demonstrating gains on long-tail labels (see Table 3). On LEGALEVAL, where annotation ambiguity and subtle rhetorical distinctions are frequent (Kalamkar et al., 2022), PBR still reaches 82.5%. Most gains come from reducing confusions between overlapping roles such as Legal Analysis and Factual Issue Descriptions, which account for over 40% of baseline errors.

Takeaway 2. Prototypes are particularly effective in challenging conditions, whether data are scarce (minority roles) or label boundaries are ambiguous (overlapping functions).

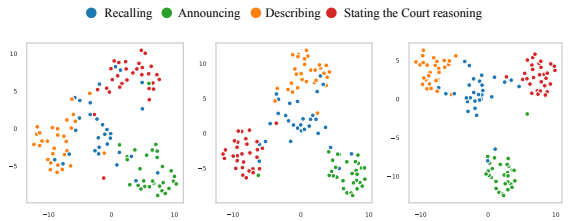
How does corpus granularity reveal the need for global cues? As the annotation scheme progresses from broad categories to fine-grained steps, label boundaries become harder to distinguish, such as subtypes within ANALYSIS. In this setting, prototypes act as semantic anchors that stabilize representations across levels of granularity. The +3.8 improvement on SCOTUS$_{\text{Steps}}$ indicates that the model increasingly depends on global cues when local context is insufficient, demonstrating the value of a multi-level corpus design.

Rhetorical Function	**Baseline**	**+PCM**	**Δ**
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	100.00	↑ 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 on SCOTUS$_{\text{RF}}$.

Takeaway 3. Multi-level annotation not only enriches model evaluation but also pinpoints where global prototype signals become essential.

How transferable are prototype-based methods across medical and scientific abstracts? Evaluating on PUBMED and CS-ABSTRACTS allows us to test whether prototype-based methods extend beyond the legal domain, since medical and scientific abstracts also exhibit rhetorical structure but in shorter and less complex texts. PBR improves performance on both benchmarks, confirming that structural regularization remains effective whenever discourse structure is present. In contrast, PCM brings limited gains in these settings, as prototype averaging is less informative when structural variation is low. However, oracle results—up to 99.7% mF1 on CS-ABSTRACTS—demonstrate the potential of PCM when relevant prototypes are retrieved, underscoring retrieval quality as the main



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

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

bottleneck for cross-domain generalization.

Takeaway 4. Structural regularization with PBR transfers robustly across domains, while PCM shows high potential but is constrained by prototype retrieval quality.

5.2 Qualitative Analysis

We visualize the latent space with t-SNE (Figure 3) to illustrate how prototypes shape sentence representations. In the baseline, clusters overlap, especially between Describing and Stating the Court's reasoning, which often co-occur due to semantic proximity. With prototypes, these roles become more clearly separated, showing that global signals improve role distinction and that prototype quality plays a key role in structuring the latent space.

5.3 Ablation Study

PBR – hyperparameters. We vary three hyperparameters on $\text{SCOTUS}_{\text{RF}}$: the prototype count, the proximity loss λ_{prox}, and the diversity loss λ_{div} (see Appendix E). Performance remains stable, with slight improvements up to 16 prototypes before plateauing. Moderate regularization ($\lambda_{\text{prox}} = 0.9$, $\lambda_{\text{div}} = 0.9$) provides the best balance between consistency and discriminability, whereas stronger penalties slightly reduce performance by overspreading the embedding space.

PCM – injection strategies. On legal datasets such as $\text{SCOTUS}_{\text{RF}}$ and LEGALEVAL, Linear Fusion achieves the best results, improving mF1 by $+2.63$ over FiLM (see Appendix F). This suggests that direct concatenation aligns well with the structured rhetorical patterns of legal text. On PUBMED, by contrast, all strategies perform similarly ($F1 > 92$), with Gated Residual Addition slightly ahead, indicating that injection is less critical for shorter and structurally simpler texts.

	Legal		Medical		Scientific	
	$\text{SCOTUS}_{\text{RF}}$		PUBMED		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1
Meta-Llama3-8B	66.78	75.09	82.51	87.63	67.39	74.59
Mistral-7B	**70.29**	76.61	81.86	87.20	62.24	70.61
Qwen3-8B	69.36	75.53	81.73	87.26	65.93	74.93
PCM (Ours)	65.75	80.31	**88.86**	**92.91**	**71.10**	**78.09**
PBR (Ours)	67.45	**80.92**	87.42	92.06	68.69	75.46

Table 4: Performance of LLMs Fine-Tuned with QLoRA vs. Prototype-Based Methods (Ours).

6 RRL in the Age of LLMs: Performance and Efficiency Comparison

The turning point for RRL came with GPT-3.5-Turbo by OpenAI[4], which sparked wide evaluations of LLMs via in-context prompting. Yet, Belfathi et al. (2023) showed that despite the hype, BERT encoders (Devlin et al., 2019) remained superior for classification in accuracy and efficiency. Instruction fine-tuning also introduces biases, especially in multiple-choice settings where models may prefer options based on position or style rather than meaning (Zheng et al., 2023).

Since then, both model quality and adaptation strategies have evolved. To reassess this landscape, we fine-tuned three recent open-source models (Mistral-7B (Jiang et al., 2023), Meta-LLaMA3-8B (Dubey et al., 2024), and Qwen3-8B (Yang et al., 2025)) with QLoRA (Dettmers et al., 2023), using the last token as the sentence embedding, without the need for prompt engineering (Wang et al., 2023). Results in Table 4 show clear progress with fine-tuned LLMs compared to the in-context evaluations of Belfathi et al. (2023), with Mistral-7B reaching 70.29% mF1 on $\text{SCOTUS}_{\text{RF}}$. This confirms that parameter-efficient fine-tuning can improve LLM performance for classification tasks. However, as observed in prior work (Naguib et al., 2024), LLMs still tend to underperform encoder-based models when training data are sufficient, and their advantage mainly appears in low-resource or few-shot settings.

Yet, in terms of resource efficiency, our prototype-based methods (PBR and PCM) remain highly competitive, with only 110M trainable parameters, surpassing Mistral-7B while requiring about $\sim 70\times$ fewer parameters, showing that targeted inductive biases can outperform LLMs at a fraction of the cost.

[4] https://platform.openai.com/docs/models/gpt-3.5-turbo

Input Excerpt	Confused Role Pair	Gold Label	Baseline Prediction	PCM Prediction	Error Reduction	Expert Assessment
As we have explained, failure to comply with §262(l)(2)(A) is not an act of artificial infringement.	Recalling ↔ Stating the Court's reasoning	Recalling	Stating the Court's reasoning	Recalling	19.75%	This case is ambiguous because "As we have explained" signals a backward reference, while the remainder conveys a definitive judicial conclusion, typical of Stating the Court's reasoning.
Thus, in Los Angeles Cloak Joint Board ILGWU (Helen Rose Co.), 127 N. L. R. B. 1543 (1960), **the Board held that** §8(b)(1)(B) barred a union from picketing a company in an attempt to force the employer to...	Describing ↔ Recalling	Recalling	Describing	Recalling	6.56%	The difficulty arises because "the Board held that" points to recalling a precedent, yet the remainder paraphrases substantive content neutrally, resembling Describing.
The American Bar Association **recommends** defense counsel "promptly communicate and explain to the defendant all plea offers made by the prosecuting attorney," ABA Standards for Criminal Justice, Pleas of Guilty 14–3.2(a) (3d ed. 1999),	Quoting ↔ Recalling	Recalling	Quoting	Recalling	33.33%	Ambiguity stems from combining an attribution verb ("recommends") with a direct quotation, blending narrative recall with verbatim citation and blurring the distinction between Recalling and Quoting.

Table 5: Expert evaluation of model predictions on ambiguous rhetorical role pairs. PCM reduces baseline errors by better resolving overlaps between semantically similar functions.

Takeaway 5. Recent LLMs fine-tuned with QLoRA show clear progress for RRL classification, but prototype-based methods still provide a better balance between accuracy and efficiency.

7 Expert Analysis

Following the experiments, we sought feedback from a PhD linguist with expertise in legal discourse on the accuracy of model predictions, particularly at the rhetorical function level.

On the positive side, the expert confirmed strong ambiguity between RECALLING and STATING THE COURT'S REASONING. This "gray zone" appears when a sentence both cites a source and conveys the Court's stance. As shown in Table 5, our PCM approach reduced errors in this pair by 19.75% compared to the baseline. A similar gain was observed for RECALLING and QUOTING, where ambiguity arises when a sentence combines quotation with commentary or uses a partial citation as support.

Looking forward, the expert emphasized that some roles are still hard to classify. In particular, ACCEPTING and REJECTING arguments are often mislabeled as RECALLING when sources are cited, even with explicit markers such as *"we agree"*, *"was correct to point"*, or *"we disapprove"*.

Takeaway 6. Expert analysis confirms that prototypes reduce key ambiguities, but some functions remain challenging, highlighting the need for future model improvements.

8 Discussion

Existing approaches to RRL, particularly on LegalEval (Kalamkar et al., 2022; Belfathi et al., 2025), are robust but still fail to separate roles that are semantically close. By integrating local context with global prototypes, our methods reduce such ambiguities, as shown both by consistent performance gains and by expert analysis on our new dataset.

Our corpus centers on long judicial decisions, evaluated with five-fold cross-validation on 18 test cases (2,481 sentences). The low variance across splits (± 0.3 mF1) confirms that improvements are systematic rather than data-dependent.

The comparison of PBR and PCM highlights complementary strengths. PBR is a lightweight regularizer, training $20\breve{}25\%$ faster and using $30\breve{}40\%$ less GPU memory, making it attractive in resource-limited settings. PCM, though more expensive due to precomputed prototypes and conditioning.

Finally, regarding efficiency, Belfathi et al. (2023) underscored the high cost of LLMs for RRL. Our results suggest that prototype-based methods achieve a more favorable accuracy–efficiency trade-off, outperforming recent LLM-based systems under constrained resources.

9 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 the legal NLP field. Our findings also indicate that prototype-based methods offer a more favorable accuracy–efficiency trade-off than fine-tuned LLMs. Future work should prioritize extending semantic prototyping to multilingual and cross-domain RRL, where generalization is even more demanding.

10 Limitations

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

- The current problem formulation assigns a single rhetorical label to each sentence in a multi-class classification task. 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.

11 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.

Anas Belfathi, Nicolas Hernandez, and Laura Monceaux. 2023. Harnessing gpt-3.5-turbo for rhetorical role prediction in legal cases. In *JURIX*, pages 187–196.

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.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–13.

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.

Brett Curry and Banks Miller. 2008. Looking for law in all the wrong places? foreign law and support for the us supreme court. *Politics & Policy*, 36(6):1094–1124.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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*.

Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Learning prototype representations across few-shot tasks for event detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5270–5277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

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.

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.

John M Swales. 1990. *Genre analysis*. Cambridge university press.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Annotation Scheme

Figure 6 presents the complete annotation scheme, whose main components are detailed below.

A.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.

A.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.

A.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.

B Corpus Statistics

This section presents key statistics and descriptive analyses of the annotated SCOTUS-LAW corpus. Figure 4 illustrates the topical, temporal, and authorial diversity of the dataset, showing the wide range of judicial opinions included. Figure 5 depicts the structural flow of rhetorical functions within legal cases, highlighting their relative positions throughout a decision. Finally, Table 6 reports the number and proportion of instances across the three annotation levels—*category*, *rhetorical function*, and *step*—providing an overview of the corpus composition.

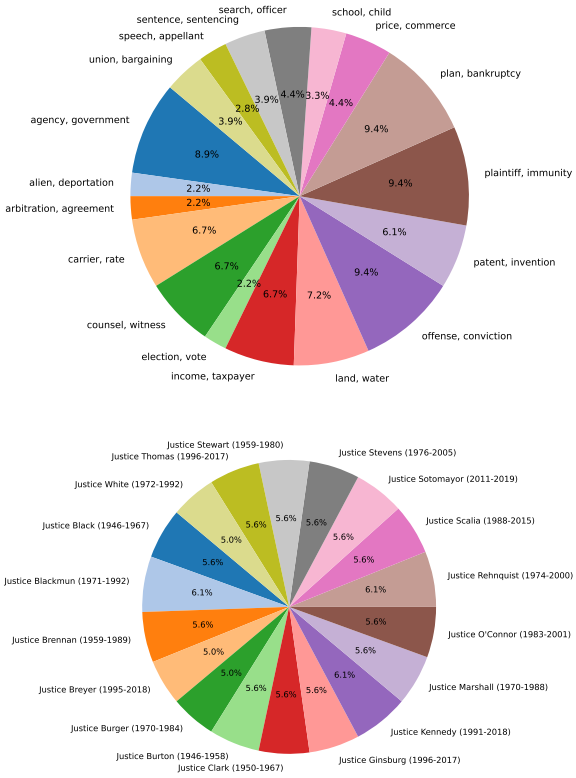


Figure 4: Topical, Temporal, and Authorial Diversity in our annotated corpus.

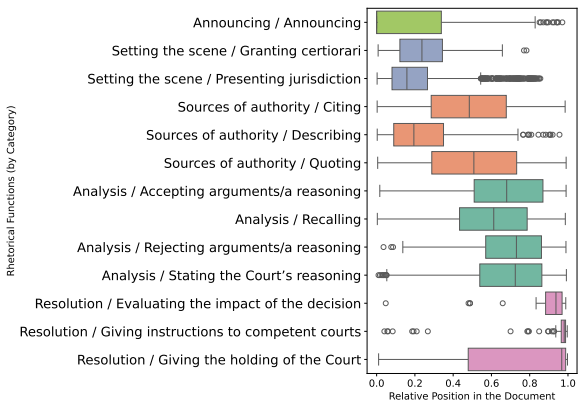


Figure 5: Distribution of Rhetorical Functions by Relative Position, revealing a structured rhetorical flow in judicial reasoning—from the initial announcement to the final resolution.

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 (→).

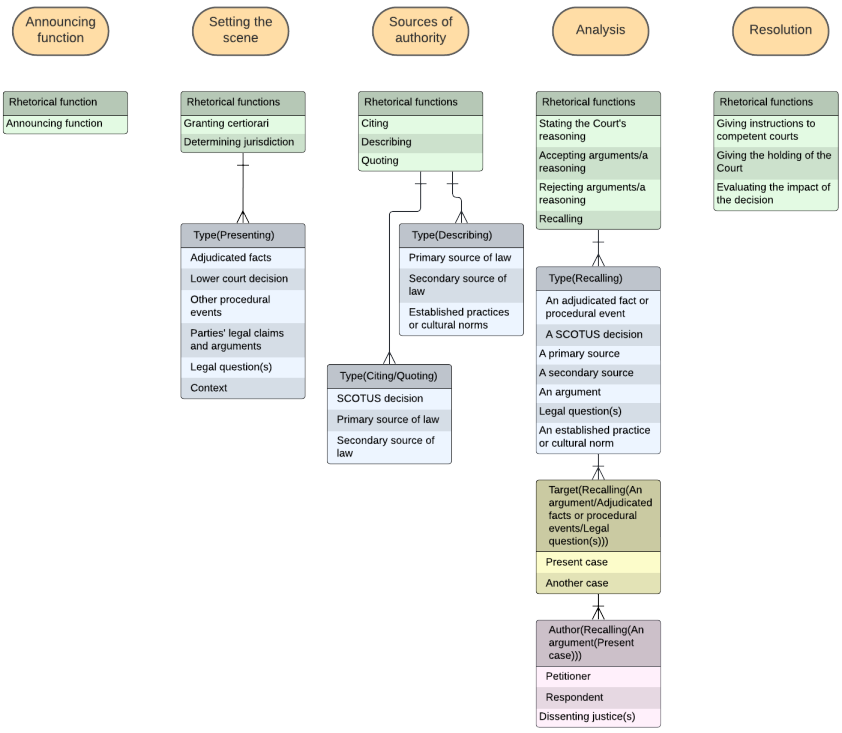


Figure 6: The final coding scheme is composed of 5 categories (ovals with orange background), 13 rhetorical functions (green rectangles) and 24 attributes (types in blue rectangles, target in the yellow rectangle, and author in the purple rectangle. The scheme reads from top to bottom: A step label is 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.

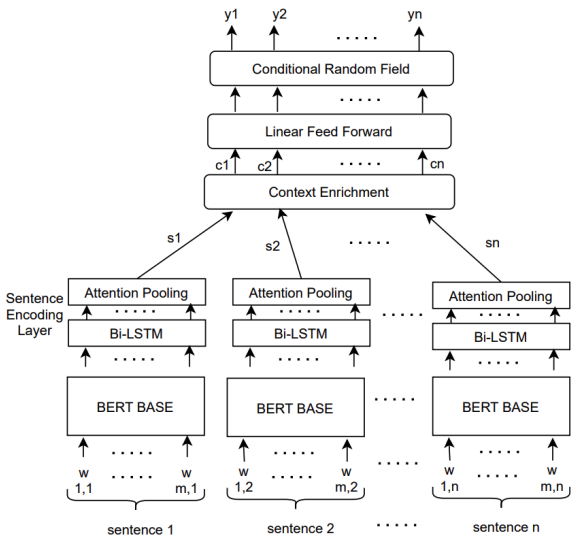


Figure 7: The hierarchical architecture.

C Hierarchical Architecture Details

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.

D Experimental Setup

D.1 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.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 $\lambda = 0$ disables the constraint, 0.9 is a balanced setting from Das et al. (2022), and 10 enforces strong regularization.

D.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.

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

D.4 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).

E 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 8.

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.

F 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.

Method	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 8: W-F1 scores for prototype injection strategies. All variants share the same hierarchical encoder with PCM integration.

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

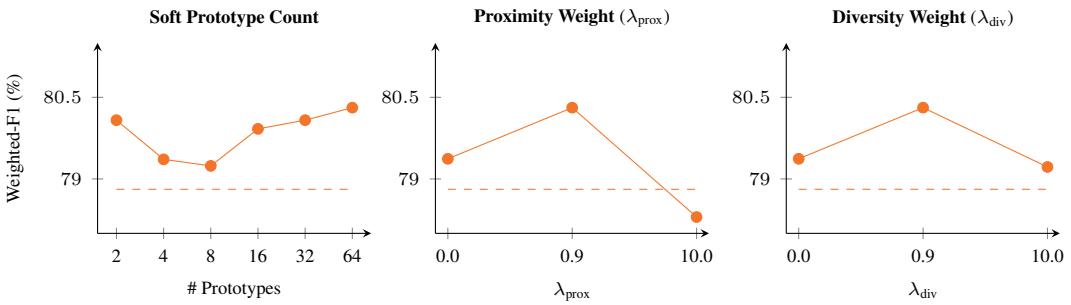


Figure 8: Effect of PBR hyperparameters on w-F1 at the $\text{SCOTUS}_{\text{RF}}$ Dashed lines indicate the baseline without prototypes.

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. The ablation study in Table 8 shows that no method is universally optimal, and that effectiveness often depends on the nature of the data and task.