SynerRRL: Synergizing Small and Large Language Models for Rhetorical Role Labeling

Anonymous ACL submission

Abstract

Rhetorical Role Labeling (RRL) assigns a functional role to each sentence in a document and is widely used in legal, medical, and scientific domains. Encoder-based Small Language Models (SLMs) are effective for sentence-level classification but do not capture the broader discourse and domain knowledge encoded by autoregressive Large Language Models (LLMs). We introduce SYNERRRL, a hybrid framework that leverages the complementary strengths of SLMs and LLMs by aligning their internal representations through a lightweight residual fusion mechanism, without relying on prompting. Experiments with five SLMs, three LLMs, and eight RRL datasets show consistent improvements, yielding average gains of 5.14 macro-F1 points. An expert-based evaluation shows that SYNERRRL improves the classification of rhetorically ambiguous sentences and performs robustly across annotation difficulty levels.

1 Introduction

Rhetorical Role Labeling (RRL) classifies each sentence in a document according to its discourse function. Because sentence roles depend on surrounding context, RRL is particularly challenging in domains such as law and medicine, where documents follow strict rhetorical conventions and rely on implicit background knowledge (Cheng et al., 2024). Identifying roles such as ARGUMENT or ANALYSIS supports downstream applications including information retrieval (Neves et al., 2019; Safder and Hassan, 2019) and summarization (Kalamkar et al., 2022; Muhammed et al., 2024).

Transformer-based models have led to state-of-the-art performance in RRL (Cohan et al., 2019). Encoder Small Language Models (SLMs[1]), such as BERT (Devlin et al., 2019), remain the dominant approach due to their efficiency and strong discriminative performance (Roccabruna et al., 2024), and have been extended through hierarchical architectures (Brack et al., 2022, 2024) and domain-oriented pretraining objectives (Belfathi et al., 2025a). Despite these advances, Shimizu et al. (2025) show that SLMs lack the task-specific and domain-level knowledge needed to capture the full discourse complexity of RRL.

In response to these limitations, Belfathi et al. (2023); Lan et al. (2024) reformulate RRL as a text-generation task using prompting with Large Language Models (LLMs). However, prompting-based approaches exhibit unstable behavior and weaker performance on classification (Naguib et al., 2024). This follows from the fact that LLMs are trained for open-ended generation rather than precise label assignment, making their predictions highly sensitive to prompt design. These challenges motivate our investigation into **how to combine the discriminative efficiency of SLMs with the contextual knowledge encoded in LLMs through direct use of their internal representations.**

We argue that progress on RRL requires addressing the trade-off between efficiency and contextual richness. SLMs excel at sentence-level discrimination but lack broader discourse awareness, while LLMs capture richer contextual signals yet do not reliably produce stable classification decisions. To bridge this gap, we introduce SYNERRRL, a unified framework that exploits the complementary strengths of both model families through representation-level integration. We summarize our core contributions:

- We propose SYNERRRL, a hybrid framework that integrates sentence-level representations from SLMs with internal representations from LLMs through a lightweight residual fusion layer, enabling efficient and context-aware RRL without relying on prompting.
- We conduct a large-scale evaluation with five

[1] In this paper, SLMs denote encoder-based models (e.g., BERT), and LLMs denote autoregressive generative models (e.g., Qwen-3), as commonly used in the literature.

SLMs and three LLMs across eight RRL datasets spanning legal, medical, and scientific domains, demonstrating consistent gains and strong generalization.

- We complement the quantitative results with expert analysis demonstrating SYNERRRL's ability to resolve rhetorically ambiguous cases and improve interpretability in challenging instances.

Reproducibility: We release our code under an open-source license[2].

2 Related Work

2.1 Small Language Models for RRL

SLMs such as BERT (Devlin et al., 2019) are encoder-based models designed for efficient discriminative modeling. Successor models, including RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), improve robustness through larger pretraining corpora and refined training objectives. In RRL, SLMs achieve strong performance at low computational cost by leveraging hierarchical architectures (Brack et al., 2022, 2024) and domain-specific pretraining (T.y.s.s. et al., 2024; Belfathi et al., 2025b). However, existing approaches remain limited in their ability to capture broader discourse structures and world knowledge. To the best of our knowledge, no prior work has examined how SLMs can benefit from the richer contextual representations provided by LLMs.

2.2 Prompting-Based LLM Approaches to RRL

LLMs such as GPT-4 (Achiam et al., 2023), LLaMA-3 (Dubey et al., 2024), and Qwen-3 (Yang et al., 2025) demonstrate strong generative and reasoning capabilities, including in zero-shot and few-shot settings (Zhao et al., 2025). Their ability to generalize across domains has motivated their application to RRL. Belfathi et al. (2023); Lan et al. (2024) reformulate RRL as a prompting-based generation task, often relying on multi-step reasoning or explanation-driven outputs. While these approaches highlight the expressive power of LLMs, they incur high computational costs and exhibit unstable predictions on fine-grained labels. In contrast, we explore how LLM contextual representations can be exploited without prompting, enabling a more efficient and reliable integration with SLM-based classifiers.

2.3 Complementarity Between SLMs and LLMs

Outside RRL, SLMs have been shown to improve the robustness of LLM-based systems. Zhao et al. (2023) calibrate outputs using BERT to mitigate hallucinations by aligning model confidence with factual correctness, while Azaria and Mitchell (2023) employ BERT-based classifiers to verify the factual consistency of generated statements. In evaluation contexts, SLMs also provide stable semantic similarity measures, as illustrated by BERTScore (Manakul et al., 2023). Taken together, these studies indicate that SLMs contribute precise discriminative information that complements the generative capabilities of LLMs. However, no prior work has unified these complementary properties within a single framework for RRL. This gap motivates SYNERRRL, which integrates representations from both model families.

3 SYNERRRL: A Hybrid Framework for Rhetorical Role Labeling

This section begins by defining the RRL task (§ 3.1). We then describe the vanilla fine-tuning paradigm used as a baseline (§ 3.2). Finally, we introduce SYNERRRL, our proposed hybrid framework (§ 3.3), which combines sentence representations from a SLM and LLM through a dual-path architecture and a residual fusion mechanism, as illustrated in Figure 1.

3.1 Task Definition

RRL assigns a rhetorical function (e.g., ARGUMENT, ANALYSIS) to each sentence in a document, capturing its role in the discourse structure. Formally, let a document be represented as a sequence of sentences $\mathbf{x} = \{x_1, \ldots, x_m\}$. The task consists of predicting a corresponding sequence of rhetorical role labels $\mathbf{y} = \{y_1, \ldots, y_m\}$, where each $y_i \in \mathcal{Y}$ denotes the rhetorical role assigned to sentence x_i, and $\mathcal{Y}$ is a predefined label set.

3.2 Vanilla Fine-Tuning

As a baseline, we adopt the standard fine-tuning paradigm for sentence-level classification with encoder-based SLMs such as BERT. Each sentence is encoded independently, and the resulting sentence representation is fed into a lightweight

[2] https://anonymous.4open.science/r/syner-rrl-framework-7FF8

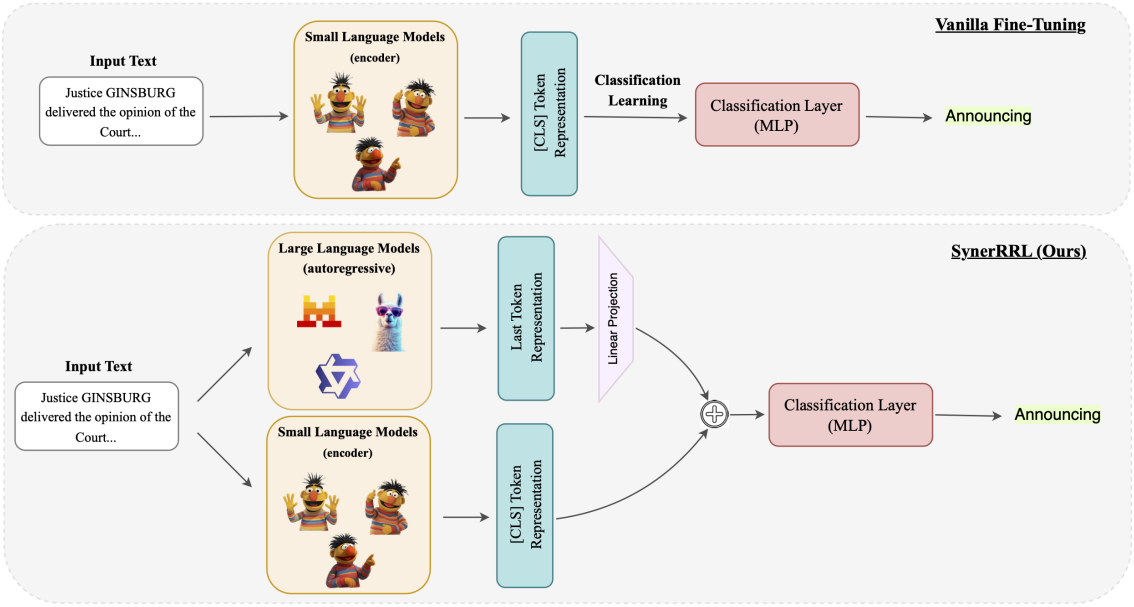

Figure 1: Comparison between standard vanilla fine-tuning and SynerRRL, which integrates SLM and LLM representations at the sentence level without prompting.

classifier to predict its rhetorical role. This approach serves as a strong and widely used baseline for RRL, reflecting the dominant practice in prior work on sentence-level and hierarchical classification (Wang and Yu, 2023; Chang et al., 2023).

3.3 Synergizing Small and Large Language Models

Motivation and Design Rationale. RRL requires accurate sentence-level semantics together with an understanding of discourse structure (Brack et al., 2022). Encoder-based models efficiently capture sentence representations but lack broad domain knowledge, whereas autoregressive models capture this knowledge (Shen et al., 2025) but are costly and unstable for classification. To bridge this gap, we introduce SYNERRRL, a hybrid framework in which an encoder serves as the primary sentence model, while complementary contextual information is extracted from a autoregressive model without relying on prompting or generation.

Dual-Path Architecture. As shown in Figure 1, SYNERRRL processes each input sentence through two parallel pathways. In the **SLM pathway**, an encoder such as BERT produces contextualized token representations, from which the `[CLS]` vector h_{SLM} is used as the sentence representation. In parallel, the **LLM pathway** encodes the same sentence using an autoregressive LLM. *No prompting or text generation is performed.* Instead, a sentence representation h_{LLM} is extracted from the final hidden state of the last token, which has been shown to capture contextual information (Wang et al., 2023). To limit computational costs, the LLM remains frozen and only a small set of LoRA parameters is trained (Dettmers et al., 2023).

Residual Representation Fusion. After obtaining sentence-level representations from both pathways, SYNERRRL combines them using a lightweight residual fusion mechanism. The LLM representation is projected into the SLM space and added as a residual signal to the SLM embedding. Formally, the fused representation is computed as:

$$h_f = h_{\text{SLM}} + W h_{\text{LLM}},$$

where W is a learnable projection. This residual design preserves the discriminative capacity of the SLM while injecting complementary information from the LLM. Compared to concatenation, it introduces fewer parameters, reduces overfitting, and maintains training stability (Shan et al., 2025).

Joint Training Objective. The parameters of the SLM and the LoRA adapters in the LLM are trained jointly. The fused representation h_f is fed to a classification head to predict the rhetorical role label for each sentence. Training follows the standard cross-entropy objective applied to the gold labels.

4 Experimental Setup

We evaluate SYNERRRL across legal, medical, and scientific domains to assess robustness to diverse discourse structures and annotation schemes. We use the original dataset splits.

4.1 Evaluation Datasets

Legal Domain. Our experiments cover five legal corpora that differ in jurisdiction, structure, and annotation schemes. SCOTUS-LAW (Lavissière and Bonnard, 2024) comprises U.S. Supreme Court decisions annotated with rhetorical roles at multiple levels of granularity. It is divided into three task-specific subsets: $\textbf{SCOTUS}_{\textbf{Category}}$, capturing high-level discursive organization; $\textbf{SCOTUS}_{\textbf{RF}}$, focusing on rhetorical functions that reflect communicative intent; and $\textbf{SCOTUS}_{\textbf{Steps}}$, which integrates both dimensions with additional attributes describing fine-grained reasoning steps. **LEGALEVAL** (Kalamkar et al., 2022) includes judgments from Indian courts at multiple sources (Supreme, High, and District), annotated with thirteen rhetorical roles and commonly used as a benchmark for sentence-level classification. **DEEPRHOLE** (Bhattacharya et al., 2023) also consists of Indian Supreme Court judgments and employs an annotation scheme with seven rhetorical roles.

Medical Domain. We evaluate our approach on two medical discourse datasets. **PUBMED** (Dernoncourt and Lee, 2017) consists of structured medical abstracts from randomized controlled trials, where sentences are automatically categorized by authors into five rhetorical roles. **BIORC** (Lan et al., 2024), in contrast, is a manually annotated corpus of medical research abstracts designed for sequential sentence classification, extending the same label schema with an additional class *Other*.

Scientific Domain. We evaluate our framework on a scientific discourse dataset. **CS-ABSTRACTS** (Gonçalves et al., 2020) consists of abstracts from computer science publications, annotated via crowdsourcing following the same five rhetorical roles as in PUBMED.
Dataset statistics are reported in Appendix A.

4.2 Models and Implementation Details

We evaluate SYNERRRL on a set of vanilla fine-tuned SLM encoders covering different efficiency–capacity trade-offs: **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), **DeBERTa** (He et al., 2020), **DistilBERT** (Sanh et al., 2019), and **ALBERT** (Lan et al., 2019). This setup assesses whether the proposed approach consistently improves over strong baselines while remaining effective for both compact and higher-capacity encoders.

Within SYNERRRL, we incorporate contextual representation knowledge from autoregressive LLMs drawn from three recent model families: **Qwen-3-8B** (Yang et al., 2025), **Mistral-7B** (Jiang et al., 2023), and **LLaMA-3-8B** (Dubey et al., 2024). We focus on the 7B–8B scale as a practical trade-off between representational capacity and efficiency. To enable efficient adaptation, we employ QLoRA (Dettmers et al., 2023), which introduces low-rank adapters while keeping the base model weights frozen.
All models are trained for five epochs with a learning rate of 5×10^{-5} and a batch size of 32, using the Adam optimizer (Kingma and Ba, 2014).

	LEGALEVAL		DEEPRHOLE		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1
Mistral-7B$^{\text{Prompting}}$	23.37	31.76	20.97	21.55	56.70	63.31
Llama-3-8B$^{\text{Prompting}}$	23.78	34.41	23.08	22.73	36.65	44.64
Qwen3-8B$^{\text{Prompting}}$	17.42	19.15	11.73	11.89	8.29	16.60
Mistral-7B$^{\text{Rep}}$	60.83	73.20	45.74	**52.22**	**62.97**	**70.89**
Llama-3-8B$^{\text{Rep}}$	**61.62**	**74.40**	**45.78**	52.04	61.74	70.62
Qwen3-8B$^{\text{Rep}}$	58.92	72.11	32.42	37.91	56.73	66.11

Table 1: Comparison between state-of-the-art LLM-based RRL prompting strategies (Belfathi et al., 2023; Lan et al., 2024) and our representation-level fine-tuned LLMs without prompting.

5 Results & Analysis

In this section, we report the performance of SYNERRRL in comparison with standard fine-tuning of encoder-based SLMs and against SOTA prompting-based RRL strategies.

5.1 Overall & Fine-grained Performance

1. Why should LLMs represent rather than generate for RRL? Prompting has been explored as a way to adapt LLMs to RRL by reformulating the task as text generation. However, Table 1 shows that state-of-the-art prompting performs substantially worse than our representation-level fine-tuned LLM, achieving a maximum mF1 of 36.65% on `CS-Abstracts`. This trend is consistent with findings from broader text classification benchmarks, where prompting is reported as the weakest LLM-based approach (Zhang et al., 2025). This systematic underperformance indicates that **prompting misaligns the generative nature of LLMs**

	Legal										Medical				Scientific		Average	
	SCOTUS$_{\text{Category}}$		SCOTUS$_{\text{RF}}$		SCOTUS$_{\text{Steps}}$		LEGALEVAL		DEEPRHOLE		PUBMED		BIORC		CS-ABSTRACTS			
	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1
ALBERT$_{\text{base}}$ (baseline)	80.11	82.87	66.06	71.88	47.26	56.73	54.21	64.34	44.62	49.95	79.48	85.14	79.80	81.27	62.11	69.66	64.21	70.23
+ SynerRRL (Mistral-7B)	82.29†	84.96†	70.10†	75.82†	51.74†	64.55†	**63.99**†	72.48†	47.85†	53.77†	81.08†	86.34†	88.03†	87.12†	65.84†	72.76†	68.87†	74.72†
+ SynerRRL (Llama-3-8B)	83.56†	85.88†	**71.83**†	**77.66**†	54.17†	**65.70**†	62.00†	**73.94**†	48.69†	55.03†	**82.10**†	**87.41**†	88.78†	**88.64**†	**67.80**†	**74.11**†	**69.87**†	**76.04**†
+ SynerRRL (Qwen3-8B)	**84.74**†	**86.59**†	70.44†	77.30†	**55.24**†	65.18†	61.17†	73.55†	**48.81**†	**55.87**†	81.77†	87.33†	**88.80**†	88.34†	**63.87**†	72.59†	69.35†	75.84†
BERT$_{\text{base}}$ (baseline)	81.82	83.90	68.66	73.51	48.75	60.65	56.60	67.10	46.24	51.50	80.07	85.75	84.49	84.56	61.70	69.70	66.04	72.08
+ SynerRRL (Mistral-7B)	82.89†	85.93†	68.71	77.11†	53.61†	65.10†	58.52†	71.78†	47.02	52.96	81.07†	86.13†	88.84†	88.31†	62.48	70.18	67.89†	74.69†
+ SynerRRL (Llama-3-8B)	83.37†	86.14†	**70.57**†	77.52†	54.89†	65.70†	**61.74**†	**73.75**†	46.55	54.22†	**82.17**†	**87.45**†	89.14†	88.92†	66.30†	72.89†	69.34†	75.82†
+ SynerRRL (Qwen3-8B)	**84.57**†	**87.33**†	69.45	**77.75**†	**58.29**†	**67.15**†	59.86†	73.43†	**48.01**†	**55.33**†	81.83†	87.30†	**89.54**†	**88.98**†	**66.96**†	**73.68**†	**69.81**†	**76.37**†
DeBERTa$_{\text{base}}$ (baseline)	81.64	84.12	67.16	74.29	51.45	62.08	57.91	70.97	43.47	51.57	80.21	85.70	84.99	84.56	63.45	69.26	66.29	72.82
+ SynerRRL (Mistral-7B)	83.45†	85.07	68.74	76.78†	52.11	64.01†	**60.79**†	71.83	46.63†	53.10	80.56	85.99	86.50	86.99†	64.25	72.39†	67.88†	74.52†
+ SynerRRL (Llama-3-8B)	**84.18**†	85.69	71.14†	77.29†	55.24†	65.58†	60.41†	**73.39**†	**46.75**†	**55.05**†	**82.44**†	**87.54**†	**89.37**†	**89.26**†	**65.79**†	73.52†	**69.42**†	75.91†
+ SynerRRL (Qwen3-8B)	83.74†	**86.20**†	**71.60**†	**77.66**†	**55.92**†	**66.16**†	60.03†	73.64†	46.55†	54.60†	82.15†	87.47†	88.83†	88.48†	65.57†	**73.68**†	69.30†	**75.99**†
DistilBERT$_{\text{base}}$ (baseline)	80.83	83.34	65.81	72.95	48.44	60.06	55.98	66.49	43.42	50.61	79.68	85.35	84.15	83.91	61.03	68.89	64.92	71.45
+ SynerRRL (Mistral-7B)	**84.53**†	86.38†	70.66†	76.70†	55.20†	65.77†	**62.72**†	72.57†	45.20	52.85	81.57†	86.71†	87.19†	87.09†	65.17†	72.32†	69.03†	75.05†
+ SynerRRL (Llama-3-8B)	84.18†	86.36†	71.18†	77.72†	54.45†	66.10†	61.00†	**74.41**†	**46.79**†	**54.32**†	**82.30**†	**87.51**†	89.02†	88.66†	**67.20**†	**73.59**†	**69.51**†	**76.08**†
+ SynerRRL (Qwen3-8B)	84.13†	**86.92**†	**72.12**†	**78.16**†	**56.25**†	**66.55**†	60.90†	73.53†	46.15†	53.84†	82.04†	87.44†	**89.86**†	**88.90**†	64.65†	72.72†	**69.51**†	76.01†
RoBERTa$_{\text{base}}$ (baseline)	80.38	83.42	69.09	75.49	51.40	62.38	56.74	70.30	46.20	52.79	80.66	86.08	84.34	84.39	67.21	73.52	67.00	73.55
+ SynerRRL (Mistral-7B)	**84.27**†	86.83†	70.30	76.46	50.80	63.56	58.91	72.26†	47.16	53.77	81.05	86.37	88.29†	87.71†	63.67	71.16	68.06†	74.77†
+ SynerRRL (Llama-3-8B)	83.54	86.42	**72.20**†	77.61†	54.97†	**66.67**†	**63.90**†	**74.44**†	**48.36**	**55.30**†	82.44†	87.63†	89.22†	**89.00**†	**67.03**	**73.65**	**70.21**†	**76.34**†
+ SynerRRL (Qwen3-8B)	83.69†	**86.84**†	70.74	**78.26**†	**55.70**†	65.78†	59.35†	73.59†	47.71	54.69†	**82.59**†	**87.87**†	**89.32**†	88.79†	65.83	72.64	69.37†	76.06†

Table 2: Performance of SYNERRRL and baseline models across eight RRL benchmarks. Baselines refer to encoder-based SLMs fine-tuned for sentence-level classification, and "+ SYNERRRL" denotes the application of the proposed hybrid framework. Results are reported in terms of Macro-F1 and Weighted-F1 scores, averaged over three runs. † indicates statistically significant improvements over the corresponding baselines ($p < 0.05$).

with the discriminative demands of classification. Thus, the issue is not the absence of useful inductive knowledge in LLMs but the inefficiency of accessing it through generation. This motivates reframing their role in RRL: **instead of asking LLMs to decide, we ask them to represent**.

Takeaway 1. Prompting misaligns LLMs with classification. SYNERRRL instead bypasses generation and leverages LLM representations, enabling SLMs to classify reliably.

2. Does SYNERRRL yield consistent gains over SLM baselines across RRL benchmarks? Table 2 shows that SYNERRRL improves every SLM baseline across all eight benchmarks. Averaged across datasets, the framework yields gains of 5.14 mF1 pts for ALBERT, with comparable improvements observed for the other encoder models. These gains reflect not only enhanced fine-grained label discrimination but also greater prediction stability, as confirmed by our per-label analysis (Table 3). Improvements are particularly pronounced on fine-grained legal datasets. On SCOTUS$_{\text{Steps}}$ (35 labels), SYNERRRL substantially boosts models that otherwise struggle with subtle rhetorical distinctions. Most gains are statistically significant, indicating that LLM-derived representations reliably enrich SLM embeddings and generalize consistently across architectures.

Takeaway 2. SYNERRRL yields consistent, statistically significant improvements across all SLMs, with the largest gains on legal tasks.

3. How sensitive is SYNERRRL to the choice of LLM? SYNERRRL remains effective regardless of which LLM provides the contextual representations. Although the magnitude of gains varies slightly, all three models—Mistral-7B, Llama-3-8B, and Qwen3-8B—consistently improve performance across datasets and SLM architectures. Certain LLMs perform better on specific domains (e.g., Qwen on legal tasks, Mistral on medical ones), but the overall improvement pattern remains stable.

Rhetorical Function	**%**	**Baseline**	**+SynerRRL**	**Δ**
Accepting arguments/a reasoning	0.4	70.00	80.00	↑ 10.00
Announcing	1.3	76.92	86.15	↑ 9.23
Citing	2.4	90.09	92.59	↑ 2.50
Describing	3.6	35.99	49.09	↑ 13.10
Evaluating the impact of the decision	0.2	0.00	0.00	0.00
Giving instructions to competent courts	0.4	56.00	64.00	↑ 8.00
Giving the holding of the Court	2.9	80.30	86.36	↑ 6.06
Granting certiorari	0.7	100.00	100.00	0.00
Presenting jurisdiction	18.8	74.87	82.93	↑ 8.06
Quoting	24.5	98.12	98.37	↑ 0.24
Recalling	30.8	64.48	71.57	↑ 7.09
Rejecting arguments/a reasoning	1.9	58.49	64.58	↑ 6.09
Stating the Court's reasoning	12.1	53.56	58.13	↑ 4.57
Macro-F1		66.06	71.83	↑ **5.77**

Table 3: Role-wise F1 scores on SCOTUS$_{\text{RF}}$ comparing the ALBERT baseline with SYNERRRL (Llama-3-8B). The % column indicates the proportion of each rhetorical function in the corpus.

Takeaway 3. SYNERRRL is effectively LLM-agnostic: all tested LLMs yield consistent gains driven by generalizable contextual signals.

4. Does SYNERRRL generalize across legal, medical, and scientific domains? SYNERRRL improves performance across all three domains despite their distinct discourse structures. In the legal domain, gains remain substantial across datasets and SLMs, reflecting the framework's ability to capture complex argumentative and hierarchical structures. In medical and scientific abstracts,

Train Dataset	Model	Test Dataset		
		PUBMED	BIORC	CS-ABSTRACTS
PUBMED	$BERT_{base}$ (baseline)	79.62	**63.45**	36.63
		85.42	79.91	42.24
	SynerRRL (Llama-3-8B)	**81.74**	62.65	**44.24**
		87.18	**80.31**	**50.63**
BIORC	$BERT_{base}$ (baseline)	77.63	85.27	65.09
		83.20	85.20	73.37
	SynerRRL (Llama-3-8B)	**78.92**	**88.77**	**69.08**
		85.00	**88.10**	**76.53**
CS-ABSTRACTS	$BERT_{base}$ (baseline)	56.01	57.35	61.71
		58.67	68.99	67.91
	SynerRRL (Llama-3-8B)	**63.05**	**67.49**	**67.30**
		67.03	**80.57**	**74.10**

Table 4: Experimental results of SYNERRRL under domain shift, as all datasets follow the same rhetorical scheme.. We report Macro-F1 in the upper row and Weighted-F1 in the lower row of each cell.

	$SCOTUS_{RF}$		LEGALEVAL		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1
$BERT_{base}$ (baseline)	68.66	73.51	56.6	67.1	61.7	69.7
$BERT_{large}$	68.76	73.87	55.19	67.69	59.53	67.77
SynerRRL (Random)	66.53	72.43	57.06	66.81	62.52	70.32
SynerRRL (Llama-3-8B)	**70.57**	**77.52**	**61.74**	**73.75**	**66.3**	**72.89**

Table 5: Assessing whether performance gains arise from LLM knowledge rather than parameter scaling. We compare $BERT_{base}$, $BERT_{large}$, a random-initialized variant of SYNERRRL, and its Llama-3-8B version.

where discourse tends to be more standardized and structurally constrained, SYNERRRL still delivers strong improvements, indicating that LLM-derived contextual signals effectively complement domain-specific writing patterns.

Takeaway 4. SYNERRRL generalizes across specialized datasets, indicating that LLM-based contextual cues transfer well across varied discourse structures.

5.2 Generalizability of SYNERRRL under Domain Shift

LLMs are pre-trained on broader and more diverse corpora than SLMs, so integrating their representations through SYNERRRL should enhance generalizability under domain shift. We evaluate this using three datasets that share the same rhetorical labeling scheme but differ in discourse style and subdomain: `PubMed` and `BioRC` (medical) and `CS-Abstracts` (scientific). For each dataset, we train a vanilla BERT and its SYNERRRL counterpart and evaluate both models across all three test sets. As shown in Table 4, SYNERRRL consistently improves robustness under domain shift. It systematically outperforms vanilla BERT when transferring between medical and scientific domains, indicating stronger cross-domain transfer of rhetorical-role representations.

	$SCOTUS_{RF}$		LEGALEVAL		DEEPRHOLE	
	mF1	wF1	mF1	wF1	mF1	wF1
$BERT_{base}$ (baseline)	68.66	73.51	56.6	67.1	46.24	51.5
SynerRRL (Qwen3-0.6B)	67.78	73.38	**60.42**	71.85	45.78	51.41
SynerRRL (Qwen3-1.7B)	69.19	76.33	59.02	71.91	45.46	51.12
SynerRRL (Qwen3-8B)	**69.45**	**77.75**	59.86	**73.43**	**48.01**	**55.33**

Table 6: Experimental results of SYNERRRL using different LLM sizes from the same model family, comparing the 0.6B, 1.7B, and 8B versions of Qwen3.

Takeaway 5. SYNERRRL enhances cross-domain generalization, reinforcing its value in settings where training and deployment domains differ.

5.3 Probing the Role of Intrinsic LLM Knowledge in SYNERRRL

To assess whether SYNERRRL's gains stem from intrinsic LLM knowledge rather than increased parameter count, we compare it against two controls (Table 5): a larger encoder ($BERT_{Large}$) and a variant of SYNERRRL whose LLM branch is randomly initialized. The larger encoder shows only marginal differences relative to its base model, indicating that scaling the encoder alone does not meaningfully improve rhetorical-role classification. Similarly, the random-initialized variant performs on par with these baselines, demonstrating that architectural augmentation without pretrained LLM signal offers little benefit.

Takeaway 6. SYNERRRL's improvements derive from leveraging pretrained LLM knowledge rather than parameter scaling, underscoring the value of contextual representations in rhetorical-role classification.

5.4 Impact of LLM Size on SYNERRRL

Modern LLMs are released in multiple parameter scales, raising the question of how model capacity affects SYNERRRL. To investigate this, we evaluate three Qwen3 variants—0.6B, 1.7B, and 8B parameters—as representation sources within the framework. As shown in Table 6, integrating LLM representations of any size improves performance in most cases over the baseline SLM, indicating that SYNERRRL benefits from contextual signals irrespective of scale. However, the largest gains are obtained with the 8B model, which outperforms the smaller variants across all evaluated datasets.

Takeaway 7. SYNERRRL benefits from LLM representations at all scales, with larger models providing stronger contextual signals and yielding consistently higher gains.

	SCOTUS$_{RF}$		DEEPRHOLE		PUBMED	
	mF1	wF1	mF1	wF1	mF1	wF1
BERT$_{base}$ (baseline)	68.66	73.51	46.24	51.5	80.07	85.75
SynerRRL (Mistral-7B)	**68.71**	**77.11**	47.02	52.96	**81.07**	**86.13**
SynerRRL (Saul-7B)	68.66	73.89	**50.40**	**54.33**	-	-
SynerRRL (BioMistral-7B)	-	-	-	-	80.94	86.03

Table 7: Evaluating the impact of domain-specialized LLMs in SYNERRRL, comparing legal (Saul-7B) and medical (BioMistral-7B) variants with the general Mistral-7B model.

	AGNEWS		COLA		SST-2	
	mF1	wF1	mF1	wF1	mF1	wF1
BERT$_{base}$ (baseline)	92.81	92.81	72.83	77.03	91.04	91.05
SynerRRL (Mistral-7B)	94.38	94.38	72.89	78.15	89.67	89.67
SynerRRL (Llama-3-8B)	95.11	95.11	83.29	85.95	96.21	96.22
SynerRRL (Qwen3-8B)	**95.21**	**95.21**	**84.15**	**86.57**	**97.13**	**97.13**

Table 8: Evaluating SYNERRRL on generalist sentence classification datasets.

5.5 Impact of Domain-Specialized LLMs in SYNERRRL

Table 7 assesses whether domain-specialized LLMs (Saul-7B (Colombo et al., 2024) for legal text and BioMistral-7B (Labrak et al., 2024) for biomedical text) provide additional benefits over the general-purpose Mistral-7B. While all LLM variants improve over the BERT baseline, the specialized models do not yield stronger signals than the generalist model within SYNERRRL. This aligns with recent findings showing that domain-specific adaptation can induce **catastrophic forgetting**, reducing the breadth of contextual cues available for downstream tasks (Ling et al., 2025). Overall, the generalist Mistral-7B offers the most effective representations for enhancing SLMs.

Takeaway 8. General-purpose LLMs provide richer and more useful contextual signals for SYNERRRL than domain-specialized variants.

5.6 Generalization to Generalist Classification Benchmarks

Table 8 examines whether the benefits of SYNERRRL extend beyond specialized RRL datasets to generalist sentence classification benchmarks, namely AGNEWS (Zhang et al., 2015), CoLA (Warstadt et al., 2019), and SST-2 (Socher et al., 2013). Across all tasks, integrating LLM-derived representations consistently improves over the BERT baseline, indicating that the proposed framework is not restricted to domain-specific rhetorical structures. Improvements are observed for all LLM variants, with Qwen-3-8B achieving the strongest overall performance.

Metric	**Value**
Model size	LLaMA-style, ≈8B parameters (quantized)
Input length	128 tokens
Dataset size	28 000 instances
Total tokens processed	3.584M
Total training time	3.50 h
Effective compute time	98% of total
GPU used	1 × 80GB-class GPU
Throughput (tokens/s)	≈289.8
Throughput (instances/s)	≈2.26
Estimated total FLOPs	1.72×10^{17} (0.172 EFLOPs)

Table 9: Compute and efficiency statistics for training our SynerRRL model with an 8B LLaMA-style LLM.

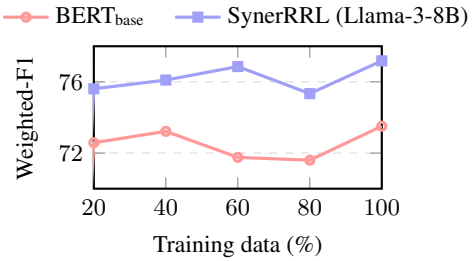

Figure 2: Data-efficiency comparison of BERT$_{base}$ and SynerRRL (Llama-3-8B).

5.7 Compute and Data Efficiency of SYNERRRL

Table 9 summarizes the compute characteristics of SYNERRRL with an 8B LLaMA-style LLM. Training remains efficient: the full model is optimized in 3.5 hours on a single 80GB A100 GPU, and the total compute cost remains modest relative to typical LLM fine-tuning workloads (Wang et al., 2025). These results indicate that the proposed framework introduces minimal overhead and is feasible on standard research hardware.

Figure 2 evaluates data efficiency under varying training-set sizes. Across all fractions (20%–100%), SYNERRRL consistently outperforms BERT$_{base}$, maintaining higher weighted-F1 even in low-resource regimes. This shows that LLM-derived contextual signals help the SLM generalize more effectively with less labeled data.

Takeaway 9. SYNERRRL is compute- and data-efficient, delivering strong gains without heavy resources or large training sets.

6 Expert Analysis

Following the experiments, we sought feedback from a linguistic expert with expertise in legal discourse on the accuracy of model predictions on the SCOTUS$_{RF}$ dataset. To complement our quantitative results, the expert evaluation focuses on two questions: (i) whether SYNERRRL better resolves am-

Input Excerpt	Confused Role Pair	Gold Label	Baseline Prediction	SynerRRL Prediction	Error Reduction	Expert Assessment
The warrantless search and seizure of the garbage bags left at the curb outside the Greenwood house **would violate the Fourth Amendment only if** respondents manifested a subjective expectation of privac...	Recalling ↔ Stating the Court's reasoning	Recalling	Stating the Court's reasoning	Recalling	20.71%	This sentence is ambiguous because the conditional formulation "**would violate ... only if**" reads like a fresh doctrinal test, characteristic of Stating the Court's reasoning, yet in context it restates a rule derived from prior precedent, which corresponds to Recalling.
And the confinement of gambling-loss deductions to the amount of gambling gains, **a provision brought into the income tax law as § 23(g) of the Revenue Act of 1934** and carried forward in...	Describing ↔ Recalling	Recalling	Describing	Recalling	17.16%	Ambiguity arises because the sentence both describes the content of a statutory provision and situates it historically ("**brought into the income tax law as § 23(g)**"), mixing neutral exposition typical of Describing with backward reference to enacted law, which aligns with Recalling.
A. H. Bull S. S. Co. v. National Marine Engineers' Beneficial Assn., 250 F.2d (CA2 1957), the right to discharge such supervisors because of their involvement in union activities or union membersh...	Quoting ↔ Recalling	Recalling	Quoting	Recalling	14.88%	The difficulty here comes from the compact case citation "**A. H. Bull S. S. Co. v. ... 250 F.2d (CA2 1957)**", which visually resembles a pure quotation, while the sentence function is to recall a prior decision as part of the argumentative narrative, i.e., Recalling.

Table 10: Expert evaluation of SynerRRL predictions on ambiguous rhetorical role pairs for $\text{SCOTUS}_{\text{RF}}$ with $\text{BERT}_{\text{base}}$. SynerRRL reduces errors by better resolving overlaps between semantically similar functions.

Model	Annotation Difficulty (Likert Scale)			
	Quite easy	Rather easy	Rather difficult	Quite difficult
$\text{BERT}_{\text{base}}$ (baseline)	72.61	53.64	55.31	52.22
SynerRRL (Llama-3-8B)	**84.00**	**58.96**	**78.51**	**66.67**

Table 11: Macro-F1 of $\text{BERT}_{\text{base}}$ and SYNERRRL across annotation difficulty levels.

biguities between closely related rhetorical roles, and (ii) whether its gains persist across sentences of varying annotation difficulty.

6.1 Expert Assessment of Ambiguous Rhetorical Role Pairs

To examine whether SYNERRRL better handles fine-grained rhetorical distinctions, we asked the expert to analyze cases where baseline $\text{BERT}_{\text{base}}$ confuses pairs of semantically similar roles—such as RECALLING VS. STATING THE COURT'S REASONING, or DESCRIBING VS. RECALLING. As shown in Table 10, these sentences are genuinely ambiguous, often blending descriptive exposition with historical or doctrinal references. The expert confirmed that SYNERRRL correctly resolves these ambiguities more often than the baseline, reducing errors by capturing contextual cues that distinguish subtly overlapping functions.

6.2 Expert Assessment Across Annotation Difficulty Levels

To assess whether SYNERRRL is particularly beneficial for harder instances, we conducted an expert-based analysis on the $\text{SCOTUS}_{\text{RF}}$ dataset. An expert annotated 2,480 sentence segments using a four-point Likert scale reflecting annotation difficulty, allowing performance comparison across difficulty levels. As shown in Table 11, SYNERRRL consistently outperforms $\text{BERT}_{\text{base}}$ at all difficulty levels, with especially large gains on Rather difficult and Quite difficult sentences. These results indicate that LLM-injected contextual signals help handle cases challenging even for trained annotators.

Takeaway 10. SYNERRRL improves predictions on both ambiguous and difficult sentences, showing that LLM-based contextual signals help SLMs resolve subtle rhetorical distinctions that challenge even expert annotators.

7 Discussion & Conclusion

This work revisits the role of LLMs in RRL by shifting the focus from generation to representation. Rather than using LLMs as decision-making models, we show that their internal representations can effectively enrich sentence-level classifiers, yielding consistent gains across datasets and domains.

This representation-centric view reconciles prior findings in the literature. While prompting-based LLM approaches have been shown to be unstable and to underperform strong encoder baselines in RRL (Belfathi et al., 2023; Lan et al., 2024), encoder-based models exhibit more reliable discriminative behavior (Zhang et al., 2025). Our results indicate that this discrepancy stems not from a lack of useful knowledge in LLMs, but from a mismatch between their generative objective and the demands of fine-grained classification.

Overall, our findings suggest that progress in RRL depends less on scaling model capacity than on how contextual knowledge is integrated into classification. Injecting LLM-derived representations into sentence encoders provides a simple and effective mechanism to stabilize predictions and improve fine-grained role discrimination, with potential applicability to other structured prediction tasks under domain constraints.

8 Limitations

While SynerRRL offers consistent gains across models and domains, several limitations should be acknowledged to contextualize its contributions and guide future work:

- The framework relies on sentence-level representations. Although effective for RRL, this granularity ignores finer rhetorical cues expressed at the clause or discourse-unit level. Incorporating sub-sentential segmentation or modeling inter-sentence rhetorical dependencies may yield additional improvements.
- SynerRRL integrates LLM representations but does not explicitly control which layers, abstraction levels, or semantic dimensions contribute most to the fused embedding. A deeper analysis of representation selection could improve interpretability and performance.
- All experiments were conducted on English datasets. Applying SynerRRL to multilingual RRL introduces challenges related to cross-lingual alignment, variation in rhetorical conventions, and the transferability of LLM representations across languages.

9 Ethics Statement

We acknowledge that LLMs may encode societal, cultural, or domain-specific biases, and integrating their representations into SLMs could transfer or amplify such biases within RRL predictions. Although our experiments did not reveal explicit harmful patterns, SynerRRL still inherits the limitations of the underlying pretrained models. Future work should more thoroughly examine bias propagation in representation-level fusion, evaluate its downstream impact in legal and biomedical settings, and develop mitigation strategies that ensure fair and accountable use of hybrid SLM–LLM systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, and Richard Dufour. 2025a. 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, Monceaux Laura, and Richard Dufour. 2025b. A simple but effective context retrieval for sequential sentence classification in long legal documents. In *Proceedings of the 12th Argument mining Workshop*, pages 160–167, Vienna, Austria. Association for Computational Linguistics.

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.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023. 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.

Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci, and Andrew McCallum. 2023. Multi-CLS BERT: An efficient alternative to traditional ensembling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–854, Toronto, Canada. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

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.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia

Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

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.

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, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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. *Preprint*, arXiv:2310.06825.

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*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, and 5 others. 2025. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Comput. Surv.*, 58(3).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Weiqiao Shan, Yuang Li, Yuhao Zhang, Yingfeng Luo, Chen Xu, Xiaofeng Zhao, Long Meng, Yunfei Lu, Min Zhang, Hao Yang, Tong Xiao, and JingBo Zhu. 2025. Enhancing speech large language models with prompt-aware mixture of audio encoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19305–19320, Suzhou, China. Association for Computational Linguistics.

Sijia Shen, Feiyan Jiang, Peiyan Wang, Yubo Feng, Yuchen Jiang, and Chang Liu. 2025. Do LLMs know and understand domain conceptual knowledge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5967–5976, Suzhou, China. Association for Computational Linguistics.

Seiji Shimizu, Hisada Shohei, Yutaka Uno, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2025. Exploring LLM annotation for adaptation of clinical information extraction models under data-sharing restrictions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14678–14694, Vienna, Austria. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, 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.

Hongwei Wang and Dong Yu. 2023. Going beyond sentence embeddings: A token-level matching algorithm for calculating semantic textual similarity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 563–570, Toronto, Canada. Association for Computational Linguistics.

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*.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao, and Xuming Hu. 2025. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18980–18989, Suzhou, China. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Theodore Zhao, Mu Wei, J Samuel Preston, and Hoifung Poon. 2023. Automatic calibration and error correction for generative large language models via pareto optimal self-supervision.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Dataset	Source	Domain	Language	# Docs	# Sents	Labels
SCOTUS$_{\text{Category}}$	Lavissière and Bonnard (2024)	Legal (U.S.)	English	180	26,328	5
SCOTUS$_{\text{RF}}$	Lavissière and Bonnard (2024)	Legal (U.S.)	English	180	26,327	13
SCOTUS$_{\text{Steps}}$	Lavissière and Bonnard (2024)	Legal (U.S.)	English	180	26,327	35
LEGALEVAL	Kalamkar et al. (2022)	Legal (India)	English	214	31,865	13
DEEPRHOLE	Bhattacharya et al. (2023)	Legal (India)	English	50	9,380	7
PubMed	Dernoncourt and Lee (2017)	Medical	English	20,000	227,000	5
BIORC	Lan et al. (2024)	Medical	English	800	7,911	6
CS-ABSTRACTS	Gonçalves et al. (2020)	Scientific	English	654	7,385	5

Table 12: Evaluation datasets used in our experiments.

A Dataset Details

We evaluate our SYNERRRL framework on eight RRL benchmarks spanning the legal, medical, and scientific domains. We use the original dataset splits. Dataset statistics are reported in Table 12.

SCOTUS-LAW (Lavissière and Bonnard, 2024) is a corpus of U.S. Supreme Court (SCOTUS) decisions collected from CourtListener. It annotated at the sentence level using a hierarchical annotation scheme with three levels of granularity. It includes three subsets: **SCOTUS$_{\text{Category}}$** (5 labels) capturing high-level discourse structure, **SCOTUS$_{\text{RF}}$** (13 labels) focusing on rhetorical functions, and **SCOTUS$_{\text{Steps}}$** (35 labels), which combines categories and rhetorical functions with optional fine-grained reasoning attributes (*type*, *author*, *target*).

LegalEval (Kalamkar et al., 2022) consists of judgments from the Indian Supreme Court, High Courts, and District Courts. It provides public training and validation splits with 214 documents, respectively, totaling 31,865 sentences (an average of 115 per document), annotated with 13 rhetorical role labels.

DeepRhole (Bhattacharya et al., 2023) includes 50 judgments from the Indian Supreme Court across five legal domains, annotated with 7 rhetorical roles. It comprises 9,380 sentences (an average of 188 sentences per document).

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

BIORC (Lan et al., 2024) is a manually annotated biomedical abstract corpus designed for sequential sentence classification. It contains 800 PubMed abstracts (700 unstructured and 100 structured), totaling 7,911 sentences, with an average of approximately 9.9 sentences per abstract. Sentences are annotated at the sentence level using a multi-label schema with six rhetorical roles: *Background*, *Objective*, *Methods*, *Results*, *Conclusions*, and an additional *Other* class for sentences that do not fit standard rhetorical categories.

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