Semantic Reranking at Inference Time for Hard Examples in 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. While language models (LMs) achieve strong average performance, they remain unreliable on hard examples, where prediction confidence is low. Existing approaches typically handle uncertainty implicitly and treat labels as discrete identifiers, overlooking the semantic information encoded in label names. We introduce RISE, an inference-time semantic reranking framework that leverages label semantics to refine predictions on hard instances. RISE automatically identifies low-confidence predictions and reranks model outputs using contrastively learned label representations, without retraining or modifying the underlying model. Experiments on eight domain-specific RRL datasets with seven LMs, including encoder-based and causal architectures, show an average gain of $+9.15$ macro-F1 points on hard examples. For explainability, we further propose manual hardness annotations to study difficulty from both model and human perspectives, revealing a moderate agreement with Cohen's $\kappa = 0.40$.

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 interpretation is often shaped by its surrounding context, RRL is particularly well suited to structured texts such as legal cases. Identifying rhetorical components (e.g., ANNOUNCING or ANALYSIS) supports downstream tasks, including information retrieval (Neves et al., 2019; Safder and Hassan, 2019) and document summarization (Kalamkar et al., 2022; Muhammed et al., 2024).

RRL approaches typically formulate the task as a sentence-level classification problem (T.y.s.s. et al., 2024; Belfathi et al., 2025b). In this setting, a classifier built on a BERT encoder (Devlin et al., 2019) maps each sentence representation to a label from a predefined set, treating them as discrete and unstructured categories. Consequently, the model relies exclusively on sentence representations, while **ignoring both the semantic meaning of label names and the relationships between them**. Although this formulation performs well on clear and frequent cases, it tends to degrade on harder examples, where the model exhibits low confidence in its predictions (Gasparin and Detommaso, 2024; Huang et al., 2024).

Recent studies attempt to address this limitation by incorporating label semantics into text classification models (Khatuya et al., 2025; Park et al., 2025). In particular, similarity-based approaches (Rücker and Akbik, 2025) embed texts and labels into a shared representation space and perform predictions based on their semantic proximity. While effective when label descriptions are informative, these methods rely primarily on similarity scores and **do not exploit the discriminative capacity of standard classifiers**[1]. As a result, their contribution is limited on easy samples, for which classifier confidence is already high.

This gap motivates inference-time methods that leverage label semantics to improve predictions on hard examples while preserving the discriminative behavior of the underlying classifier. To address this challenge, we introduce RISE, a semantic reranking framework that refines model outputs at inference time without requiring retraining. We summarize our core contributions:

- We propose RISE, an inference-time framework that identifies hard samples based on model confidence and refines their predictions by reranking logits using contrastively learned

[1] By standard classifiers, we refer to models that perform prediction using a learned decision function over fixed label sets represented as one-hot vectors, without exploiting the semantic information of label names.

semantic similarities between input and label representations, without requiring architectural changes or additional training.

- We perform a large-scale evaluation on eight RRL datasets spanning legal, medical, and scientific domains, using seven LMs that include both encoder-based and causal architectures, demonstrating consistent performance gains and strong generalization.

- We analyze model difficulty both quantitatively and qualitatively from model-centric and human-centric perspectives by introducing manual difficulty annotations, enabling explainability-oriented analyses of model behavior on challenging cases.

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

2 Related Work

2.1 RRL as a Discriminative Classification

LMs such as BERT (Devlin et al., 2019) are widely used for sentence-level classification in RRL. Successor models, including RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019), improve robustness through larger pretraining corpora and refined training objectives. When combined with hierarchical architectures (Brack et al., 2022, 2024) and domain-specific pretraining (T.y.s.s. et al., 2024; Belfathi et al., 2025b), these models achieve strong performance at moderate computational costs. However, rhetorical roles are modeled as discrete labels, and predictions rely solely on discriminative decision functions. This limits performance on hard cases involving semantically related roles, motivating inference-time methods that explicitly exploit label semantics.

2.2 Label Semantics in Text Classification

Prior work leverages semantic information associated with labels to improve text classification from multiple perspectives. Generative approaches exploit label meaning in zero-shot settings and distill the resulting supervision into discriminative classifiers (Zhang et al., 2024). Hierarchical methods encode semantic and structural relations through parent–child label dependencies (Zhu et al., 2024). Other studies dynamically refine label surface forms to improve few-shot performance (Park et al., 2025), or adopt fully generative formulations that treat labels as semantic objects rather than fixed indices (Khatuya et al., 2025). In contrast, our framework exploits this information exclusively at inference time by reranking logits for hard examples, offering a targeted alternative to training-time and fully generative methods.

3 RISE: Inference-Time Semantic Reranking for Rhetorical Role Labeling

This section presents RISE, an inference-time semantic reranking framework. We first outline its motivation and design (§ 3.1), then describe automatic hard-sample identification based on model confidence (§ 3.2). Next, we present our contrastive approach to learning label semantics (§ 3.3), and finally explain how these semantics are used at inference time for reranking (§ 3.4).

3.1 Motivation and Overview

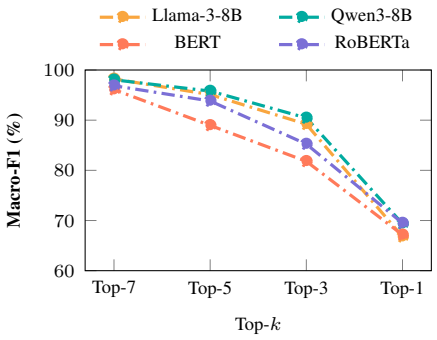

Figure 1: Top-k oracle performance reveals prediction ambiguity on the $\text{SCOTUS}_{\text{RF}}$ dataset.

Figure 1 shows that LMs with an multilayer perceptron (MLP) classification head achieve strong overall RRL performance, yet remain unreliable on a subset of inputs. These hard cases exhibit low confidence and strong competition among semantically related labels. The Top-1 vs. Top-3 macro-F1 gap indicates that the correct label is often highly ranked but not selected, suggesting errors stem from semantic proximity between labels rather than representation failure.

This observation motivates an inference-time perspective on RRL, illustrated in Figure 2. Rather than modifying the classifier or retraining the model, we address three questions: **(1)** how to identify hard samples directly from model confidence scores, **(2)** how to represent label semantics in a

[2] https://anonymous.4open.science/r/rise-framework-F84E

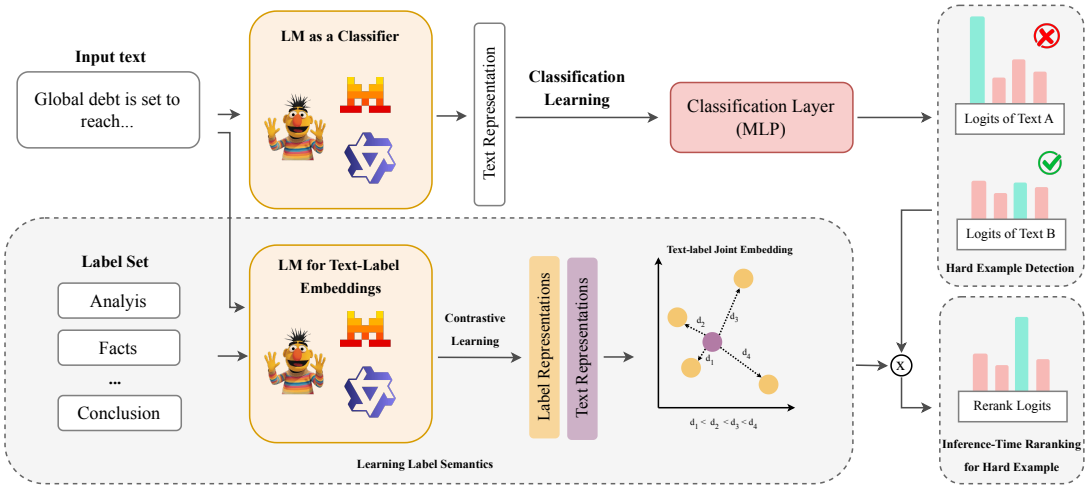

Figure 2: Overview of the RISE framework. A language model (encoder-based or causal) is first used as a discriminative classifier to produce logits for each input sentence. RISE operates at inference time (gray area) by automatically identifying hard cases based on model confidence. For these instances, label semantics are exploited by reranking logits based on semantic distances derived from contrastively learned text–label representations.

shared embedding space, and **(3)** how to exploit these representations at inference time to refine ambiguous predictions.

3.2 Hard Example Detection

Definition (Hard Example). *An input instance is considered hard when the classifier fails to produce a confident prediction, with multiple labels receiving similar scores. Hardness is determined by the classifier's scoring behavior rather than the intrinsic difficulty of the input.*

Distribution Variance as an Indicator. Even after convergence, a classifier may fail to separate some inputs, producing logit distributions where multiple labels receive similar scores, indicating low confidence and label competition. We quantify this behavior using the variance of the logit vector: lower variance corresponds to stronger label competition and higher predictive uncertainty, as adopted in (Chen et al., 2024; Yang et al., 2025b). Unlike entropy, variance directly captures score dispersion in the raw decision space and is less sensitive to calibration artifacts.

Automatic Hard Example Detection. Hard examples depend on both the model and the dataset, making fixed criteria unsuitable. We therefore **let each model identify its own hard examples directly from its prediction behavior**, using the development set for threshold estimation. We define an adaptive threshold as the average variance of logit distributions over misclassified examples, denoted as σ^2_{mis}:

$$\sigma^2_{\text{mis}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{Var}(\mathbf{z}_i), \tag{1}$$

where $\mathcal{M}$ is the set of misclassified examples and $\mathbf{z}_i$ is the corresponding logit vector. An example is classified as hard if the variance of its logit distribution is below σ^2_{mis}. This adaptive criterion limits over-selection by focusing on genuinely ambiguous predictions, ensuring that inference-time refinement is applied only where it is most effective.

3.3 Label Semantics from Confusion Patterns

Confusion as a Signal. To capture label relationships specific to our RRL setting, we rely on the classifier's prediction behavior rather than externally defined label similarities. For a trained model, the distribution of predicted labels conditioned on a gold label reveals consistent confusion patterns. We treat this normalized distribution as an ambiguity profile and use it to derive a $\text{label}_{\text{predicted}}$–$\text{label}_{\text{gold}}$ affinity, which is used to weight hard negatives in the contrastive loss.

Confusion-Weighted Contrastive Learning. We learn a shared embedding space with a pretrained language model $g_\phi(\cdot)$ that encodes a sentence x and a label name y into vectors $\mathbf{e}_x = g_\phi(x)$ and $\mathbf{e}_y = g_\phi(y)$. Given a training pair (x, y), we maximize $\text{sim}(\mathbf{e}_x, \mathbf{e}_y)$ and minimize $\text{sim}(\mathbf{e}_x, \mathbf{e}_{y'})$ for all other labels $y' \neq y$. Each negative label y' is weighted by $w_{y'} = P(y, y')$,

the normalized probability that the base classifier predicts y' given the true label y, yielding a weighted InfoNCE objective (Oord et al., 2018):

$$\mathcal{L}_{\text{pos}} = \exp\big(\text{sim}(\mathbf{e}_x, \mathbf{e}_y)\big), \tag{2}$$

$$\mathcal{L}_{\text{neg}} = \sum_{y' \neq y} w_{y'} \exp\big(\text{sim}(\mathbf{e}_x, \mathbf{e}_{y'})\big), \tag{3}$$

$$\mathcal{L}_{\text{CW}} = -\log \frac{\mathcal{L}_{\text{pos}}}{\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}}. \tag{4}$$

where $w_{y'} = P(y, y')$ is calculated on the development set.

3.4 Inference-Time Semantic Reranking

Inspired by Peng et al. (2024), we use semantic similarity between the input and label names as a complementary signal to the classifier logits. For a hard example x, the base classifier outputs logits $\mathbf{z}_x \in \mathbb{R}^C$ over C labels $\{y_1, \ldots, y_C\}$. Using the shared embedding space learned above, we then form a cosine similarity vector $\mathbf{s}_x \in \mathbb{R}^C$ between the input embedding $\mathbf{e}_x$ and the label embeddings $\mathbf{e}_y$. Predictions are reranked by reweighting the logits element-wise:

$$\tilde{\mathbf{z}}_x = \mathbf{s}_x \odot \mathbf{z}_x, \tag{5}$$

where $\odot$ denotes element-wise multiplication. Semantic reranking is applied **only to hard examples**; otherwise, the original output is used.

4 Experimental Setup

We evaluate RISE 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 use five legal corpora spanning different jurisdictions, structures, and annotation schemes. SCOTUS-LAW (Lavissière and Bonnard, 2024) contains U.S. Supreme Court decisions annotated with rhetorical roles at multiple granularities. It includes three subsets: $\textbf{SCOTUS}_{\textbf{Category}}$ for high-level discourse structure, $\textbf{SCOTUS}_{\textbf{RF}}$ for rhetorical functions, and $\textbf{SCOTUS}_{\textbf{Steps}}$ combining both with fine-grained reasoning attributes. **LEGALEVAL** (Kalamkar et al., 2022) contains Indian court judgments (Supreme, High, District) annotated with thirteen roles. **DEEPRHOLE** (Bhattacharya et al., 2023) contains Indian Supreme Court judgments annotated with seven roles.

Medical Domain. We evaluate on two medical discourse datasets. **PUBMED** (Dernoncourt and Lee, 2017) contains randomized controlled trial abstracts with sentences automatically labeled into five rhetorical roles, following established preprocessing protocols. **BIORC** (Lan et al., 2024) is a manually annotated abstract corpus for sequential sentence classification.

Scientific Domain. We evaluate on a scientific discourse dataset. **CS-ABSTRACTS** (Gonçalves et al., 2020) contains computer science abstracts annotated via crowdsourcing with the same five rhetorical roles as PUBMED.

Dataset statistics are reported in Appendix D.

4.2 Models and Implementation Details

We evaluate RISE with seven LMs commonly used as strong baselines in prior RRL studies (Belfathi et al., 2025a). Encoder-based models include **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), **DeBERTa** (He et al., 2020), and **ALBERT** (Lan et al., 2019). Causal models include **Qwen-3** (Yang et al., 2025a), **Mistral-7B** (Jiang et al., 2023), and **LLaMA-3** (Dubey et al., 2024). Causal models are fine-tuned with QLoRA (Dettmers et al., 2023) and used strictly for classification, without prompting or generation (Yousefiramandi and Cooney, 2025).

For text–label contrastive learning, the same models learn label representations, aligning label semantics with the classifier's decision space. Additional details are in the Appendix E.

5 Results & Analysis

This section reports the empirical evaluation of RISE across multiple RRL benchmarks, model architectures, and domains.

5.1 Overall Performance

1. Does RISE yield consistent gains over baseline models across RRL benchmarks? Table 1 shows that RISE yields consistent improvements across both encoder-based and causal LMs, indicating that its effectiveness arises from inference-time semantic refinement rather than model-specific tuning (Mistral-7B reaches 69.50% mF1, a gain of +1.8 pts). These gains point to a shared limitation of standard classifiers: difficulty in resolving semantic competition between closely related labels.

		Legal										Medical				Scientific		Average	
		SCOTUS$_{Category}$		SCOTUS$_{RF}$		SCOTUS$_{Steps}$		LEGALEVAL		DEEPRHOLE		PUBMED		BIORC		CS-ABSTRACTS			
		mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1
Llama-3-8B	Baseline	82.54	85.13	66.78	75.09	51.77	62.66	60.01	73.20	45.98	52.87	82.51	87.63	86.83	**87.87**	**67.39**	**74.59**	67.98	74.88
	+ RISE	**83.15**	**85.70**	**68.46**†	**77.19**†	**52.14**	**65.15**†	**61.16**†	**74.43**†	**48.53**†	**54.23**†	**82.61**	**87.85**	**87.45**	87.83	64.61	72.79	**68.51**†	**75.65**†
Mistral-7B	Baseline	83.88	85.89	70.29	76.61	51.28	64.29	60.07	71.63	47.25	53.54	81.86	87.20	84.38	87.12	62.24	70.61	67.66	74.61
	+ RISE	**84.57**	**86.57**	**72.13**†	**77.12**	**52.90**†	**65.35**†	**62.97**†	**73.93**†	**48.37**†	**54.09**	**82.33**	**87.65**	**87.09**†	**87.94**†	**65.61**†	**73.33**†	**69.50**†	**75.75**†
Qwen3-8B	Baseline	82.85	84.95	69.36	75.53	51.81	62.30	58.87	71.19	45.10	51.35	81.73	87.26	85.09	86.70	**65.93**	**74.93**	67.59	74.28
	+ RISE	**85.03**†	**86.38**†	**71.75**†	**77.81**†	**54.10**†	**64.67**†	**60.25**†	**73.06**†	**46.04**†	**52.55**†	**82.44**	**87.78**	**87.39**†	**87.68**†	63.59	71.89	**68.82**†	**75.23**†
ALBERT$_{base}$	Baseline	81.59	**84.25**	67.24	74.05	50.39	61.45	55.16	67.89	**44.50**	51.48	**81.17**	**86.58**	**85.91**	**85.89**	65.68	72.25	66.45	72.98
	+ RISE	**81.81**	84.21	**69.14**†	**75.12**†	**52.26**†	**62.84**†	**56.70**†	**68.45**	43.56	**51.81**	80.86	86.54	85.29	85.44	**66.27**	**72.67**	**66.99**	**73.39**
BERT$_{base}$	Baseline	81.70	84.46	67.15	74.84	50.14	62.52	55.48	68.92	**48.89**	52.96	**81.64**	**86.94**	**85.21**	**86.20**	59.17	71.08	66.17	73.49
	+ RISE	**82.19**	**85.12**	**69.31**†	**75.28**	**52.33**†	**63.33**	**59.12**†	**70.06**†	48.30	**53.00**	81.24	86.80	85.11	85.77	**60.37**†	**71.53**	**67.25**†	**73.86**
DeBERTa$_{base}$	Baseline	83.73	85.91	68.08	75.16	53.50	64.16	56.76	72.50	**47.42**	52.34	**81.64**	**86.96**	**86.77**	85.72	66.29	72.73	68.02	74.44
	+ RISE	**83.83**	**86.13**	**68.95**†	**75.67**	**54.18**	**65.07**†	**59.95**†	**73.75**†	46.56	**52.63**	81.34	86.94	86.76	**85.90**	**66.40**	**72.82**	**68.50**	**74.86**
RoBERTa$_{base}$	Baseline	82.13	84.99	69.49	75.32	**53.01**	64.12	57.68	72.12	44.17	51.92	**81.76**	**87.12**	86.26	86.30	63.18	72.06	67.21	74.24
	+ RISE	**82.83**	**85.76**†	**69.81**	**76.23**†	52.92	**64.38**	**60.70**†	**73.21**†	**46.79**†	**52.73**†	81.66	87.10	**86.63**	**86.61**	**63.56**	**72.16**	**68.11**†	**74.77**

Table 1: Performance across **full test sets**. Baseline refers to the underlying model without RISE, while "+ RISE" denotes performance with our inference-time framework applied. Results are reported in terms of Macro-F1 and Weighted-F1 scores, averaged over three runs. † indicates statistical significance over baselines at $p < 0.05$.

		Legal										Medical				Scientific		Average	
		SCOTUS$_{Category}$		SCOTUS$_{RF}$		SCOTUS$_{Steps}$		LEGALEVAL		DEEPRHOLE		PUBMED		BIORC		CS-ABSTRACTS			
		mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1
Llama-3-8B	*Hard (%)*	24.2		31.6		35.2		38.9		65.9		47.6		21.5		42.6		38.4	
	Baseline	46.87	49.30	34.25	42.34	28.51	35.82	37.68	47.94	27.62	36.54	55.01	55.63	57.11	**55.71**	**47.50**	**51.07**	41.82	46.79
	+ RISE	**52.77**†	**55.37**†	**47.12**†	**55.02**†	**33.87**†	**46.86**†	**43.75**†	**53.69**†	**32.97**†	**39.81**†	**57.59**†	**57.57**†	**60.68**†	55.38	41.24	45.95	**46.25**†	**51.21**†
Mistral-7B	*Hard (%)*	22.9		29.5		40.3		39.3		64.8		54.1		27.9		44.3		40.4	
	Baseline	49.75	48.29	36.37	44.49	34.84	39.81	36.17	44.02	33.59	42.14	52.17	52.02	44.57	52.60	43.63	45.69	41.39	46.13
	+ RISE	**56.00**†	**57.30**†	**46.23**†	**48.18**†	**42.07**†	**45.74**†	**49.84**†	**55.38**†	**37.65**†	**43.89**†	**58.66**†	**58.41**†	**57.25**†	**59.39**†	**53.64**†	**54.73**†	**50.17**†	**52.88**†
Qwen3-8B	*Hard (%)*	31.5		33.2		37.4		41.8		56.4		47.4		30.2		42.0		40.0	
	Baseline	48.44	48.38	37.85	43.07	31.84	33.86	38.60	46.09	29.73	38.26	54.76	54.82	47.13	52.87	**51.21**	**54.41**	42.45	46.47
	+ RISE	**65.45**†	**62.27**†	**47.56**†	**55.85**†	**36.37**†	**43.23**†	**45.08**†	**53.31**†	**33.71**†	**41.35**†	**58.96**†	**58.84**†	**62.54**†	**60.62**†	41.77	44.62	**48.93**†	**52.51**†
ALBERT$_{base}$	*Hard (%)*	23.5		26.4		34.1		37.3		51.6		27.0		32.4		33.0		33.2	
	Baseline	47.07	56.67	45.61	44.94	34.21	39.15	33.96	44.25	**32.86**	41.76	54.81	**55.20**	**63.68**	**61.36**	44.21	46.45	44.55	48.72
	+ RISE	**49.65**†	**57.18**	**51.98**†	**49.92**†	**38.73**†	**43.54**†	**36.62**†	**45.34**†	31.64	**42.41**	**55.15**	54.98	61.05	58.47	**45.74**†	**46.75**	**46.32**†	**49.82**†
BERT$_{base}$	*Hard (%)*	25.0		29.1		34.2		39.5		42.7		22.6		30.7		30.9		31.8	
	Baseline	51.68	54.40	40.72	49.89	32.90	37.53	36.18	44.19	40.48	42.66	**55.31**	**55.96**	**59.98**	**59.44**	40.76	43.70	44.75	48.47
	+ RISE	**56.94**†	**60.20**†	**46.05**†	**52.12**†	**35.76**†	**40.36**†	**43.46**†	**47.33**†	**40.92**	**43.47**†	55.00	55.03	57.47	55.88	**43.28**†	**45.08**†	**47.36**†	**49.93**†
DeBERTa$_{base}$	*Hard (%)*	29.9		29.6		33.1		39.4		54.3		23.1		22.4		33.0		33.1	
	Baseline	**54.32**	56.35	39.10	43.99	34.07	36.84	35.24	44.99	**28.03**	39.99	53.43	55.21	58.67	51.51	40.67	42.41	42.94	46.41
	+ RISE	53.51	**58.07**†	**41.31**†	**46.46**†	**37.12**†	**40.03**†	**43.37**†	**48.67**†	26.46	**40.44**	**54.94**†	**55.28**	**59.03**	**52.98**†	**42.86**†	**42.57**	**44.83**†	**48.06**†
RoBERTa$_{base}$	*Hard (%)*	26.1		29.2		31.2		38.3		43.3		24.2		24.3		31.2		31.0	
	Baseline	46.37	51.66	40.97	47.81	**33.67**	39.36	35.93	47.15	27.41	36.30	53.62	**54.71**	44.76	55.58	38.10	38.15	40.10	46.34
	+ RISE	**52.46**†	**56.61**†	**41.90**†	**50.29**†	33.22	**39.75**	**43.17**†	**49.99**†	**35.35**†	**38.59**†	**54.18**	**54.71**	**47.79**†	**57.86**†	**38.37**	**38.64**	**43.30**†	**48.31**†

Table 2: Performance on **hard-example test subsets**. Hard (%) denotes the proportion of examples identified as hard within each test set.

This issue is especially pronounced on SCOTUS$_{RF}$ and SCOTUS$_{Steps}$, where substantial rhetorical overlap causes baseline models to assign similar confidence scores to competing roles (Lavissière and Bonnard, 2024). By incorporating label semantics at inference time, RISE reduces this ambiguity and improves label selection.

A similar pattern is observed for causal LMs such as Qwen-3. Although large-scale pretraining encodes rich semantic information, discriminative classification heads do not fully exploit it when making fine-grained rhetorical distinctions. RISE improves predictions without retraining, supporting the interpretation that semantic ambiguity—rather than limitations in sentence representations, is a dominant source of RRL errors.

Takeaway 1. RISE shows that leveraging label semantics at inference time is an effective, model-agnostic way to reduce semantic ambiguity.

2. Does RISE effectively improve performance on hard examples? Further analysis in Table 2 shows that RISE produces larger gains when evaluation is restricted to hard examples, i.e., predictions with low confidence. Because hard cases are identified automatically, each model defines its own subset. Across datasets, gains range from $+1.8$ to nearly $+8$ mF1 points, markedly higher than those observed on full test sets. This pattern shows that semantic reranking is most effective where label competition is strongest, while leaving already confident predictions unchanged.

Takeaway 2. RISE concentrates improvements on uncertain predictions, where semantic competition between labels is highest.

3. Does RISE generalize across legal, medical, and scientific domains? RISE improves performance across all evaluated domains despite their differing discourse structures. In legal datasets, the

		SCOTUS$_{RF}$		LEGALEVAL		DEEPRHOLE	
		mF1	wF1	mF1	wF1	mF1	wF1
LegalBERT	Baseline	69.95	76.84	56.43	68.66	46.75	53.36
	+ RISE	**71.38**	**77.35**	**59.93**	**70.85**	**47.48**	**54.49**
SaulLM-7B	Baseline	68.90	76.57	57.93	71.29	**48.29**	**55.17**
	+ RISE	**70.65**	**77.42**	**61.10**	**72.72**	46.63	53.85

Table 3: Impact of domain-specialized language models within RISE.

		SCOTUS$_{RF}$		LEGALEVAL		DEEPRHOLE	
		mF1	wF1	mF1	wF1	mF1	wF1
Qwen3-0.6B	Baseline	65.02	72.54	55.07	69.75	37.15	44.86
	+ RISE	**67.62**	**75.11**	**58.22**	**72.09**	**41.71**	**47.36**
Qwen3-1.7B	Baseline	66.99	74.84	57.13	70.69	28.11	39.71
	+ RISE	**70.07**	**77.12**	**62.52**	**73.36**	**40.82**	**46.99**
Qwen3-8B	Baseline	69.36	75.53	58.87	71.19	45.10	51.35
	+ RISE	**71.75**	**77.81**	**60.25**	**73.06**	**46.04**	**52.55**

Table 4: Experimental results of RISE using different LM sizes from the same model family.

gains remain consistent, indicating robustness to complex argumentative and hierarchical patterns. In medical and scientific abstracts, where discourse is more standardized, RISE continues to identify hard cases and improves their predictions, complementing baseline classifiers under uncertainty.

Takeaway 3. RISE generalizes across domains by improving uncertain predictions despite substantial differences in discourse structure.

5.2 Impact of Domain-Specialized LMs in RISE

Table 3 shows that semantic reranking yields consistent gains when applied to domain-specialized models such as LegalBERT (Chalkidis et al., 2020) and SaulLM-7B (Colombo et al., 2024). This indicates that inference-time semantic refinement remains effective even when sentence representations are shaped by domain-specific pretraining. However, performance drops observed for SaulLM on `DeepRhole` highlight a different limitation. As reported by Ling et al. (2025), aggressive domain specialization can induce **catastrophic forgetting**, reducing the semantic capacity required for downstream tasks; semantic reranking then cannot fully compensate for degraded representations.

Takeaway 4. RISE remains effective with domain-specialized models, but excessive specialization can limit its benefits.

5.3 Impact of LMs Size on RISE

Table 4 compares three Qwen3 variants (0.6B, 1.7B, and 8B parameters) to assess the effect of model scale. Semantic reranking improves performance for all sizes, showing that inference-time refinement remains effective even with limited capacity. Larger models, however, reach higher absolute scores after refinement. This suggests that increased capacity yields richer semantic signals that can be exploited more effectively.

Takeaway 5. RISE benefits models of all sizes, while larger models achieve higher absolute performance after reranking.

Model	LEGALEVAL	DEEPRHOLE	CSABSTRACTS
RISE (Llama-3-8B)	**74.43**	**54.23**	**72.79**
✗ *w/o triplets*	73.79	54.18	71.60
✗ *w/o fine-tune*	73.20	52.87	62.34
RISE (Mistral-7B)	**73.93**	**54.09**	**73.33**
✗ *w/o triplets*	73.43	53.40	73.94
✗ *w/o fine-tune*	71.63	53.54	70.61
RISE (Qwen3-8B)	**73.06**	**52.55**	**71.89**
✗ *w/o triplets*	72.46	51.51	70.21
✗ *w/o fine-tune*	71.19	51.35	61.83

Table 5: Ablation study on label semantics learning. ✗ *w/o triplets* represents learning without triplet-based ambiguity modeling; ✗ *w/o fine-tune* backbone used as text–label embedder without contrastive learning.

5.4 Ablation Study

Table 5 isolates the role of learned text–label representations in the framework. When label semantics are learned through contrastive alignment, reranking consistently improves performance on hard examples, even without ambiguity-aware weighting. This indicates that the shared text–label embedding space provides the signal for inference-time refinement. By contrast, replacing learned label representations with backbone embeddings leads to sharp performance drops (e.g., -10.45 wF1 pts with Llama-3 on `CSAbstracts`), showing that representations alone do not capture the label-specific distinctions required for effective refinement.

5.5 Qualitative Analysis: Semantic Reranking on a Hard Example

Table 6 illustrates the effect of semantic reranking on a hard example from SCOTUS$_{RF}$ using Qwen-3. The instance is automatically identified as uncertain, with closely competing logits for REJECTING, STATING, and RECALLING. Although REJECTING receives the highest score, the correct label (RECALLING) is initially ranked lower, reflecting semantic ambiguity rather than a clear decision. Semantic similarity in the learned text–label embedding space shows a stronger alignment between the input sentence and RECALLING than with REJECTING. Reranking exploits this signal to reorder

Hard Example		
Even with findings adequate to support closure, the trial court's orders denying access to voir dire testimony failed to consider whether alternatives were available to protect the interests of the prospective jurors that the trial court's orders sought to guard.		
Top-3 Logits ↓	**Semantic Distance**	**Reranked Logits ↓**
Rejecting: 3.6837 ✗	**Recalling: 0.6776**	**Recalling: 1.6380** ✓
Stating: 3.4986	Stating: 0.4200	Stating: 1.4696
Recalling: 2.4172	Rejecting: 0.2696	Rejecting: 0.9932

Table 6: Case study from $\text{SCOTUS}_{\text{RF}}$. Due to space constraints, only the Top-3 logits are shown. Incorrect roles are marked in red, and correct roles in green.

	LEGALEVAL		PUBMED		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1
Similarity-based	55.87	67.02	79.91	85.87	63.59	70.78
Discriminative	**60.01**	**73.20**	**82.51**	**87.63**	**67.39**	**74.59**

Table 7: Discriminative vs. Similarity-based classification (full test sets).

the logits, promoting the correct label while demoting the semantically weaker alternative.

5.6 Further Discussion: Design Rationale of the RISE Framework

Discriminative vs. Similarity-based Classification. A natural question is why not rely solely on similarity-based classification for RRL. Results reported in Table 7 show that similarity-based classification consistently underperforms the discriminative approach on full test sets across all datasets. This performance gap becomes even more pronounced on easy instances, as shown in Table 8, where the discriminative classifier achieves substantially higher mF1 and wF1 scores.

Effects of Automatic Hard-Example Detection. We analyze the trade-off between performance and inference-time cost by varying the variance threshold for hard-example detection. As shown in Figure 3, increasing the threshold selects more instances for reranking, initially improving performance but eventually causing saturation and degradation as confident cases are unnecessarily reranked. The automatically selected threshold (green dashed line) lies near the point of maximal benefit, achieving near-peak performance while limiting inference-time cost.

LLM-based Reranking as an Alternative Design. Inspired by the LLM-as-a-judge paradigm (Gu et al., 2025), a natural alternative is to use large language models as semantic rerankers under uncertainty. We explore this design in Appendix C,

	LEGALEVAL		PUBMED		CS-ABSTRACTS	
	mF1	wF1	mF1	wF1	mF1	wF1
Similarity-based	64.47	75.26	84.38	88.42	71.29	82.64
Discriminative	**76.11**	**82.27**	**86.56**	**90.11**	**74.82**	**85.30**

Table 8: Discriminative vs. Similarity-based classification (easy instances).

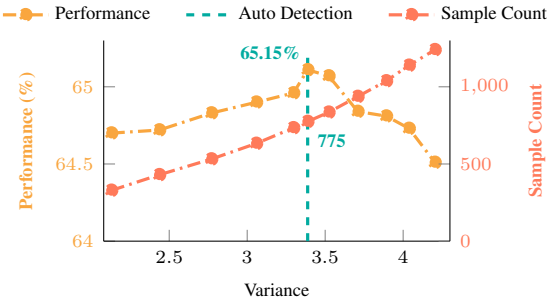

Figure 3: Marginal effect of automatic hard-example detection on $\text{SCOTUS}_{\text{Steps}}$ with LLaMA-3-8B. The orange curve shows wF1, the red curve indicates the number of detected hard examples (inference-time cost), and the green dashed line marks the selected variance threshold.

where LLMs are applied only to hard examples and restricted to a filtered set of candidate labels. While this constrained setting improves over global LLM-based classification, it remains consistently inferior to the proposed approach. LLM-based reranking exhibits higher variance and sensitivity to candidate selection, and introduces substantial inference-time cost. These findings suggest that lightweight, similarity-driven reranking provides a more robust and efficient solution for resolving semantic ambiguity, motivating the design choices behind RISE.

6 Hardness from a Human Perspective: An Empirical Analysis

To complement model-centric analyses of prediction difficulty, we examine hardness from a human perspective through an annotation study on $\text{SCOTUS}_{\text{RF}}$, which features strong semantic overlap between labels and is well suited to studying ambiguity and disagreement. We annotate the test set ($2,480$ instances) with the help of a legal expert, enabling a comparison between human-perceived difficulty and model-defined uncertainty.

6.1 Human- and Model-Centered Hardness Definitions

Human Annotation. We define human-perceived hardness in terms of the cognitive effort required to assign a label. Each instance is rated on a four-level ordered Likert scale (from easy to difficult), and the annotator identifies the

Instance	Model	Human	Sources of difficulty
This satisfied the jurisdictional requirements of 42 U. S. C.	rather easy	rather easy	Discourse cohesion
and (3) if so, whether the specific discriminatory provisions in 1395o (2) are constitutional.	easy	difficult	Taxonomy ambiguity
We have little difficulty with Espinosa's failure to file an application with the Secretary until after he was joined in the action.	rather difficult	rather easy	Writing style
After hearing argument last Term, we set the case for reargument.	difficult	rather easy	Taxonomy ambiguity

Table 9: Examples of Human- and Model-rated instance hardness.

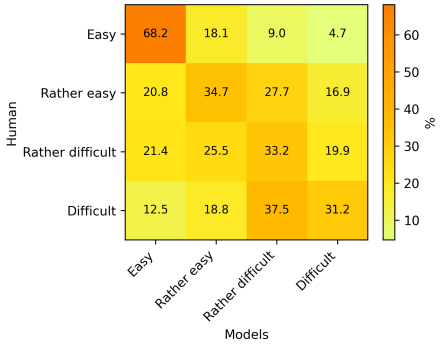

Figure 4: Human vs. Model Hardness: Level-wise correspondence.

main source(s) of difficulty (e.g., lexico-semantic, discourse-cohesion, or taxonomy-related). This protocol provides an interpretable signal of perceived difficulty.

Cross-Model Agreement. We define a model-centered notion of hardness that mirrors the human scale while remaining fully automatic and model-agnostic. For each instance, hardness is determined by the # of models exhibiting uncertainty: instances associated with a small number of uncertain models are considered easy, whereas those triggering uncertainty across many models are labeled as harder. This count-based signal is mapped to the same four-level Likert scale, enabling direct comparison with human-perceived difficulty.
Table 9 presents annotation examples; full details of the protocol are provided in Appendix A.

6.2 Human–Model Alignment Analysis

Figure 4 shows a moderate alignment between human- and model-defined hardness (**Cohen's** $\kappa = 0.40$; **Spearman's** $\rho = 0.46$), indicating partial overlap between perceived difficulty and model uncertainty. Agreement is strongest on low-effort cases, where instances are clearly unambiguous. As difficulty increases, model assessments become more dispersed, reflecting greater sensitivity to competing labels. For instances rated as non-easy

Model	Annotation Difficulty (Likert Scale)			
	Easy	Rather easy	Rather difficult	Difficult
Baseline	88.76	63.13	35.38	39.16
RISE	**90.27**	**66.89**	**38.10**	**44.27**

Table 10: Performance (wF1) across Human-Annotated Difficulty Levels (Qwen-3).

by humans, taxonomy ambiguity is the dominant source of difficulty (27.9%), exceeding surface-level linguistic factors. This suggests that perceived hardness primarily stems from uncertainty between overlapping roles rather than lexical or syntactic complexity. Discourse-level coherence follows (13.2%), highlighting the role of broader contextual integration (see Appendix B).

6.3 Implications for our RISE Framework

Table 10 compares baseline performance and RISE across instances grouped by human-perceived difficulty. On Easy instances, both systems perform well; a modest gain is still observed (+1.5 wF1 pts) without degrading confident predictions. Gains increase with difficulty, indicating that the method is most effective on cognitively demanding cases and aligning with the hard-example analysis, where model uncertainty signals challenging instances.

Takeaway 6. Hardness assessed by humans only partially overlaps with model-defined uncertainty, and higher human-perceived difficulty consistently correlates with larger gains from RISE.

7 Conclusion

We propose RISE, an inference-time semantic reranking framework for RRL that refines low-confidence predictions by leveraging label semantics, yielding consistent improvements across domains and model architectures without retraining or architectural modification. Analysis based on human-annotated difficulty shows that model uncertainty only partially aligns with human-perceived hardness, with taxonomy-level ambiguity emerging as the primary source of challenging cases. This gap underscores the limitations of purely discriminative decision functions when label boundaries are semantically overlapping. Beyond RRL, this framework provides a principled direction for analyzing limitations in human–model alignment for difficulty assessment, with implications for evaluation protocols, annotation practices, and benchmark design.

8 Limitations

While RISE consistently improves performance on uncertain predictions across models and domains, some limitations should be considered to properly contextualize its contributions and inform future research directions:

- RISE operates at the sentence level and relies on sentence representations produced by the underlying classifier. Although effective for resolving semantic competition between labels, this formulation does not explicitly model finer-grained discourse phenomena such as clause-level rhetorical cues or long-range inter-sentence dependencies, which may be critical for certain rhetorical distinctions.
- The framework assumes that label semantics, as encoded through contrastive text–label representations, are sufficiently informative to resolve ambiguity. In settings where label names are underspecified, highly abstract, or weakly aligned with their functional definitions, the benefits of semantic reranking may be limited.
- All experiments are conducted on English datasets. Extending RISE to multilingual RRL settings introduces additional challenges, including cross-lingual alignment of label semantics, variation in rhetorical conventions, and the robustness of semantic similarity measures across languages.

9 Ethics Statement

This work relies on pretrained language models whose representations may encode societal, cultural, or domain-specific biases. Although RISE operates exclusively at inference time and does not introduce generative components or additional supervision, it may nonetheless inherit or propagate biases present in the underlying models through semantic similarity and reranking mechanisms. Our experiments did not reveal systematic harmful behaviors; however, the reliance on label semantics may disproportionately affect roles that are abstract, underrepresented, or ambiguously defined.

References

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.

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.

Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, and Gaël Varoquaux. 2024. Reconfidencing LLMs from the grouping loss perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1567–1581, Miami, Florida, USA. 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.

Alberto Gasparin and Gianluca Detommaso. 2024. Distance-aware calibration for pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12434–12447, Miami, Florida, USA. 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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

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

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. 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. *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.

Subhendu Khatuya, Shashwat Naidu, Saptarshi Ghosh, Pawan Goyal, and Niloy Ganguly. 2025. Label-semantics aware generative approach for domain-agnostic multilabel classification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22286–22298, Vienna, Austria. 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*.

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

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Gyutae Park, Ingeol Baek, Byeongjeong Kim, Joongbo Shin, and Hwanhee Lee. 2025. Dynamic label name refinement for few-shot dialogue intent classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 41–52, Vienna, Austria. Association for Computational Linguistics.

Letian Peng, Zilong Wang, and Jingbo Shang. 2024. Incubating text classifiers following user instruction with nothing but LLM. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3753–3766, Miami, Florida, USA. Association for Computational Linguistics.

Susanna Rücker and Alan Akbik. 2025. Evaluating design decisions for dual encoder-based entity disambiguation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15685–15701, Vienna, Austria. 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.

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.

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. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025b. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Amirhossein Yousefiramandi and Ciaran Cooney. 2025. Fine-tuning causal llms for text classification: Embedding-based vs. instruction-based approaches. *Preprint*, arXiv:2512.12677.

Ruohong Zhang, Yau-Shian Wang, and Yiming Yang. 2024. Generation-driven contrastive self-training for zero-shot text classification with instruction-following LLM. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 659–673, St. Julian's, Malta. Association for Computational Linguistics.

He Zhu, Junran Wu, Ruomei Liu, Yue Hou, Ze Yuan, Shangzhe Li, Yicheng Pan, and Ke Xu. 2024. HILL: Hierarchy-aware information lossless contrastive learning for hierarchical text classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4731–4745, Mexico City, Mexico. Association for Computational Linguistics.

A Human Annotation of Instance Hardness

A.1 Motivation and Objectives

While model-defined uncertainty provides an operational signal for identifying hard instances, it does not fully capture the cognitive effort required by humans to interpret a sentence and assign its rhetorical role. We therefore introduce a human annotation of instance difficulty on the **SCOTUS$_{\text{RF}}$ dataset (2480 instances)** to complement model-centered analyses. This annotation aims to identify instances perceived as difficult by experts, characterize the linguistic and taxonomic sources of this difficulty, and enable a systematic comparison between human-perceived hardness and model-defined uncertainty.

A.2 Human Definition of Instance Difficulty

Instance difficulty is defined as the level of cognitive effort and uncertainty required to understand a sentence and assign it to an appropriate rhetorical role. This notion reflects the joint effect of two closely related components: the linguistic complexity of the sentence and the difficulty of mapping the interpreted content to a unique label within the rhetorical role taxonomy. Difficulty is treated as an intrinsic property of the instance, independent of the correctness of the gold label or the annotator's expertise.

A.3 Model Definition of Instance Difficulty

To derive a model-centered notion of instance hardness that mirrors human judgments while remaining fully automatic and model-agnostic, we rely on cross-model agreement under uncertainty. We consider a set of $M = 7$ independently trained models. For each instance, model uncertainty is first determined using the variance-based criterion described in Section 3.2. A model is considered uncertain if the variance of its output logits falls below its model-specific threshold. For a given instance, we compute a hardness score $k \in \{0, \ldots, 7\}$, defined as the number of models exhibiting uncertainty on that instance. Higher values of k indicate stronger cross-model consensus that the instance is ambiguous. We define the following difficulty levels:

- *Easy*: $k = 0$
- *Rather easy*: $k \in \{1, 2\}$
- *Rather difficult*: $k \in \{3, 4\}$
- *Difficult*: $k \in \{5, 6, 7\}$

A.4 Annotation Protocol

To operationalize the notion of instance difficulty, we define a structured annotation protocol that guides expert judgments in a consistent and reproducible manner. The protocol specifies a cognitively grounded difficulty scale, a step-by-step annotation procedure, and a set of explanatory factors used to justify difficulty assessments. Annotations are performed at the sentence level while considering the surrounding document context.

Difficulty Scale. Instance difficulty is annotated using a four-level ordinal Likert scale reflecting increasing levels of cognitive effort required to understand a sentence and assign its rhetorical role. Each level is defined through observable behaviors during reading and decision-making, rather than subjective impressions.

- *Easy* corresponds to cases where comprehension is immediate and the rhetorical role is assigned without hesitation.
- *Rather easy* refers to sentences that remain fluent to read and whose role is determined with little or no doubt.
- *Rather difficult* describes instances where comprehension is slowed or where hesitation arises between two semantically close roles, often requiring rereading.
- *Difficult* denotes cases involving high cognitive cost, such as complex structures or strong contextual dependence, where multiple rereadings are needed before a decision can be made.

Explanatory Factors of Difficulty. After assigning a global difficulty level, the annotator identifies the factors contributing to the perceived difficulty. These factors are not mutually exclusive, as difficulty may arise from multiple interacting sources.

- *Lexico-semantic* factors capture difficulties related to specialized, rare, abstract, or context-dependent vocabulary.
- *Syntactic complexity* refers to long or unusual grammatical constructions, including embedded clauses or passive structures.
- *Discourse cohesion* concerns difficulties in resolving references or interpreting logical connectors linking the sentence to its neighbors.
- *Discourse coherence* reflects cases where the sentence lacks sufficient information to be in-

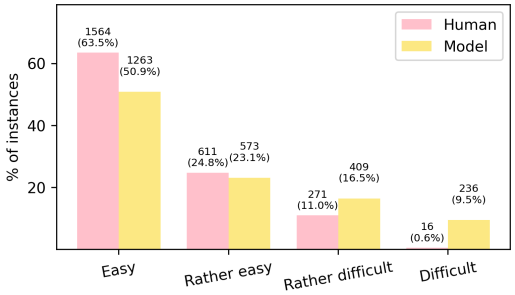

Figure 5: Distribution of instance difficulty levels according to human annotations and model-defined uncertainty on $\text{SCOTUS}_{\text{RF}}$.

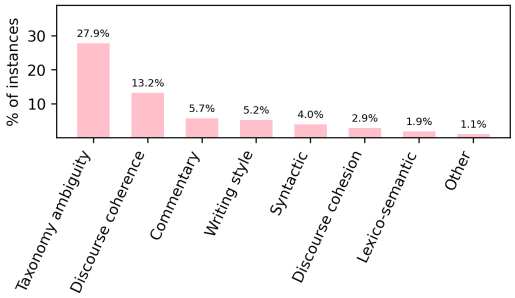

Figure 6: Distribution of human-identified explanatory factors for non-easy instances on $\text{SCOTUS}_{\text{RF}}$.

terpreted independently and relies heavily on broader context.

- *Writing style* includes stylistic features such as archaic formulations, impersonal constructions, or reported speech that slow comprehension.
- *Taxonomy-related ambiguity* captures uncertainty arising from semantic overlap between rhetorical roles rather than from the sentence content itself.

An additional *Other* category is available to record factors not covered by the predefined list. This structured factor annotation enables subsequent qualitative analyses and supports quantitative comparisons between human-perceived difficulty and model-defined uncertainty.

B Human–Model Alignment Analysis

Figure 5 presents the distribution of instance difficulty levels as assessed by the human annotator and as identified by the model on the $\text{SCOTUS}_{\text{RF}}$ dataset. Human annotations are dominated by instances labeled as *Easy*, which account for the majority of sentences, followed by a gradual decrease across higher difficulty levels. In contrast, the model assigns a lower proportion of instances to the *Easy* category and identifies a larger share of instances as *Rather difficult* and *Difficult*. This shift indicates that model-defined uncertainty tends to spread difficulty across higher levels compared to human judgments, suggesting systematic differences in how difficulty is perceived by humans and inferred from model confidence.

Figure 6 reports the distribution of explanatory factors selected by the human annotator for instances rated as non-easy. Taxonomy-related ambiguity emerges as the dominant source of difficulty, accounting for 27.9% of the annotated cases. This is followed by discourse coherence issues (13.2%), indicating that difficulty often arises from uncertainty in mapping a sentence to a rhetorical role or from insufficient contextual information rather than from surface-level linguistic complexity. Factors such as commentary style, writing style, and syntactic complexity contribute more modestly, while lexico-semantic factors and discourse cohesion are comparatively less frequent. These distributions suggest that human-perceived difficulty in rhetorical role labeling is primarily driven by semantic overlap between labels and broader contextual dependencies, rather than by lexical or syntactic properties alone.

C LLM-as-a-Reranker with Candidate Filtering

C.1 Motivation and Design Rationale

This experiment examines whether a large language model can be used as a semantic reranker to resolve ambiguous predictions in rhetorical role labeling. Using an LLM as a global classifier is ill-suited to RRL, as the task involves large and fine-grained label spaces in which semantic overlap amplifies decision noise rather than reducing uncertainty (Belfathi et al., 2023). To address this limitation, we adopt a two-stage reranking strategy. First, a restricted set of candidate labels is generated using a discriminative classifier, optionally complemented by a similarity-based signal. Second, the LLM is applied only to hard examples—where classifier confidence is low—to rerank these candidates. This design constrains inference-time cost and places the LLM in a local decision setting, where focused semantic comparison among a small set of competing labels is more effective than global classification.

```
Persona: Expert legal analyst specialized in judicial reasoning.
Task: Assign a rhetorical role to a sentence from a court decision.
Roles Definitions:
A) Granting certiorari: Assigned to sentences where the Court explicitly signals that it has agreed to review the case...
B) Presenting jurisdiction: Covers sentences that neutrally present elements of the case background...
C) Rejecting arguments/a reasoning: Indicates disagreement or refutation of a prior argument...
Sentence:
He then brought collateral relief proceedings in the ...
Candidate choices:
A) Granting certiorari
B) Presenting jurisdiction
C) Rejecting arguments/a reasoning
Answer: (B)
```

Figure 7: Prompt design for LLM-based rhetorical role reranking with candidate filtering.

Method	Setting	mF1	wF1
Baseline	–	40.72	49.89
RiSE	–	**47.12**	**55.02**
LLM-as-a-Reranker	All choices	21.68	28.44
	Selection	32.04	36.62
	Pairwise	44.10	52.31
	Pairwise + Selection	30.62	39.48

Table 11: Comparison between RiSE and LLM-as-a-Reranker strategies on $\text{SCOTUS}_{\text{RF}}$ using LLaMA-3, with GPT-4.1 employed for LLM-based reranking.

C.2 Results Discussion

Figure 7 illustrates the prompt design used for LLM-based reranking with candidate filtering, where the LLM operates under a constrained, local decision setting. Table 11 reports a comparison between RISE and several LLM-as-a-Reranker variants on $\text{SCOTUS}_{\text{RF}}$ using LLaMA-3, with GPT-4.1 employed for LLM-based reranking, and focuses on performance on hard examples. RISE consistently outperforms both the discriminative baseline and all LLM-based reranking strategies, yielding the highest Macro-F1 and Weighted-F1 scores. This result indicates that lightweight, similarity-based reranking is effective at resolving semantic ambiguity under uncertainty without introducing additional noise.

In contrast, LLM-based reranking exhibits severe performance degradation when applied to global label selection or unrestricted candidate sets ("All choices" and "Selection" settings), as shown by the sharp drop in both Macro-F1 and Weighted-F1. Reformulating the task as pairwise comparisons between competing labels substantially improves performance, particularly in terms of Macro-F1, suggesting that LLMs are more reliable when restricted to fine-grained semantic comparisons rather than multi-class decision making. However, even under this constrained setting, pairwise LLM-based reranking remains inferior to RISE, indicating lower robustness and higher variance in decision outcomes.

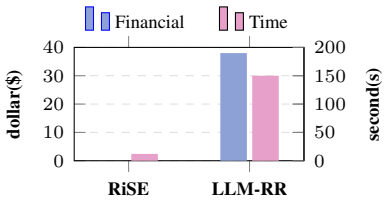

Figure 8: Financial and time cost comparison. LLM-RR denotes the LLM-as-a-Reranker strategy.

Figure 8 further highlights the practical implications of these results by comparing financial and time costs. While RISE incurs negligible additional cost at inference time, LLM-based reranking introduces a substantial increase in both latency and monetary cost. Taken together, these findings show that although LLMs can partially recover performance when tightly constrained to local comparisons, RISE offers a more robust, efficient, and consistent inference-time reranking strategy for rhetorical role labeling.

D Dataset Details

We evaluate our RISE 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: $\textbf{SCOTUS}_{\textbf{Category}}$ (5 labels) capturing high-level discourse structure, $\textbf{SCOTUS}_{\textbf{RF}}$ (13 labels) focusing on rhetorical functions, and $\textbf{SCOTUS}_{\textbf{Steps}}$ (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

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.

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.

E Implementation Details

We conduct experiments with four autoencoding encoders: ALBERT, BERT, RoBERTa, and DeBERTa. Following the standard architecture of (Devlin et al., 2019), we adopt a simple model design to focus on evaluating the effectiveness of our approach. Each model encodes the input text using a pretrained backbone and extracts the final hidden representation of the `[CLS]` token. This representation is passed through a Dropout layer to reduce overfitting, followed by an MLP classifier.

For causal language models, we evaluate three architectures: Mistral-7B, LLaMA-3-8B, and Qwen-3-8B. All models are trained with a learning rate of $1e{-}5$, weight decay of 0.001, and a dropout rate of 0.1. Gradient norms are clipped to 1.0, and training is performed for 5 epochs with a batch size of 64. We apply LoRA for parameter-efficient fine-tuning, setting the rank $r = 8$ and scaling factor $\alpha = 32$. A warmup ratio of 1.0 is used, linearly increasing the learning rate during the first epoch to stabilize training while maintaining efficient convergence.