Explicit Evidence-Span Supervision for Citation Intent Classification

Anonymous ACL submission

Abstract

Citation intent classification aims to identify the function of a citation from its citing context. However, standard label-only fine-tuning trains the model only with final intent labels, without explicitly using the local textual evidence that supports each label. We propose a teacher-guided rationale distillation framework for citation intent classification, where label-conditioned extractive rationales are used to train a stronger student model. During training-data construction, a teacher language model is provided with both the citing context and its gold citation-intent label, and generates a contiguous verbatim span that supports the given label. The generated span is used as an intermediate supervision target for the student, rather than as an unsupervised evidence discovery result. The student is then fine-tuned with an evidence-to-label generation objective, where it receives only the citing context, generates the label-supporting span, and predicts the final citation intent label. At inference time, the student does not have access to the gold label or the teacher-generated span. We further introduce token-level supervision and a training-time global-local auxiliary objective that balances the global intent-level signal with enhanced local evidence representation learning. Experiments on ACL-ARC, ACT2, and SCICITE show consistent improvements over standard LoRA fine-tuning using two different backbone models, with the largest macro-F1 gain of $+0.16$ on ACT2. Our code is publicly available at `https://anonymous.4open.science/r/span_citation-EC86`.

1 Introduction

Citations play an important role in academic writing by contextualizing new research within prior studies, acknowledging earlier research, and showing how previous work is used in the current argument. (Pride and Knoth, 2020) Understanding why a citation is made is important for downstream tasks such as scientific document understanding, literature retrieval, citation-aware summarization, and scholarly knowledge graph construction. Accordingly, citation intent classification has received increasing attention, with the goal of predicting the functional role of a citation from its citing context. (Kunnath et al., 2021)

While general text classification often emphasizes learning robust global representations of the input (Reimers and Gurevych, 2019), citation intent classification frequently depends on localized expressions that provide direct cues for the citation function. For example, phrases such as "we follow", "unlike prior work", and "motivated by" can directly indicate method use, comparison, or research motivation. Leveraging this property, many studies have considered both local textual expressions and global contextual representations for citation intent classification. (Tsai et al., 2023; Li et al., 2025a)

However, in prior work, local and global cues are often modeled implicitly within the model representation, or local cues are incorporated indirectly through word embeddings (Roman et al., 2021a,b) and additional metadata beyond the citing context (Iqbal et al., 2023; Chen et al., 2022). Such approaches mainly focus on learning the final intent label from the citing context, without explicitly connecting a specific evidence span to the corresponding intent label.

We propose a teacher-guided rationale distillation framework for citation intent classification. During training-data construction, a teacher LLM receives the citing context together with its annotated citation-intent label and constructs a contiguous verbatim rationale that supports the given label. The generated rationale is validated and used as localized supervision for student fine-tuning, allowing the student model to learn both the final intent label and the textual cue associated with that label.

We formulate citation intent classification as

an evidence-to-label generation task. The student model is trained to generate the label-supporting span before predicting the final citation intent label. During training, the teacher-generated span and the gold label are used as supervised output targets rather than additional input signals; at inference time, the student receives only the citing context.

To further refine the supervision, we separate the target sequence into span, label, and format tokens and optimize them with component-specific training signals. We also introduce a training-time global-local auxiliary objective that combines full-context understanding with enhanced local evidence representation learning. The auxiliary component is used only during training and is not used as the primary inference module.

The main contributions of this work are threefold. First, we introduce a teacher-guided label-conditioned rationale distillation framework that uses gold-label-conditioned extractive rationales as student supervision.

Second, we formulate citation intent classification as an evidence-to-label generation task, where the student model learns to generate a label-supporting span before predicting the final intent label.

Third, we combine token-level span/label supervision with a training-time global-local auxiliary objective, enabling the student model to learn from both global intent-level signals and localized citation-function cues.

2 Related Work

2.1 Citation Intent Classification

Citation intent classification aims to identify the functional role of a citation from its citing context. Early work treated citation function classification as a supervised rhetorical classification problem and used shallow, linguistically motivated features such as cue phrases (Teufel et al., 2006). Later work introduced benchmark datasets and neural models for this task, including ACL-ARC and SciCite (Bird et al., 2008; Cohan et al., 2019).

Recent studies have improved citation classification from several perspectives. (Jurgens et al., 2018) SOFT revisits the label space by separating citation intent from cited content type, highlighting the need for clearer and more consistent citation schemas (Duan and Tan, 2025). Fine-tuning-based studies show that domain-specific pretrained encoders such as SciBERT remain strong baselines for citation intent classification (Firdaus and Ruskanda, 2025). To address data scarcity and noisy citation contexts, Citss adapts pretrained language models with self-supervised contrastive learning, while GAN-CITE uses semi-supervised adversarial learning under limited labeled data (Li et al., 2025b; Chatrinan et al., 2025). Sparse rationale extraction has also been explored in citation-related prediction, suggesting that local token-level evidence can support interpretable citation analysis (Iqbal et al., 2026).

2.2 Shortcut Reliance in Text Classification

Text classification models can rely on a small number of salient tokens when making label predictions. Prior work has questioned whether attention weights provide faithful explanations of model behavior (Jain and Wallace, 2019), while later studies argue that attention-based signals should be evaluated with care rather than dismissed entirely (Wiegreffe and Pinter, 2019).

Recent studies further frame token-level reliance as shortcut learning or spurious correlation, where models exploit lexical cues that are predictive in the training data but not necessarily related to the intended task (Bastings et al., 2022; Chew et al., 2024). Mechanistic analyses also show that such shortcuts can lead models to make decisions from isolated tokens before fully using the input context (Eshuijs et al., 2025).

2.3 Evidence and Rationale Supervision

Evidence and rationale supervision has been widely studied as a way to make text classification models more grounded and interpretable. Early work used annotator-provided rationales as additional supervision beyond input texts and labels (Zaidan et al., 2007), and ERASER established a benchmark for evaluating the plausibility and faithfulness of model rationales (DeYoung et al., 2020). Recent studies further extend this direction to scientific and explanation-guided classification, using evidence signals, entailment, or generated explanations to improve prediction and interpretability (Carvallo et al., 2020; Wei Jie et al., 2024; Du et al., 2025).

This line of work is closely related to citation intent classification, where the citation function is often indicated by localized phrases in the citing context. Tsai and Yen (Tsai et al., 2023) similarly study citation intent classification together with supporting evidence extraction.

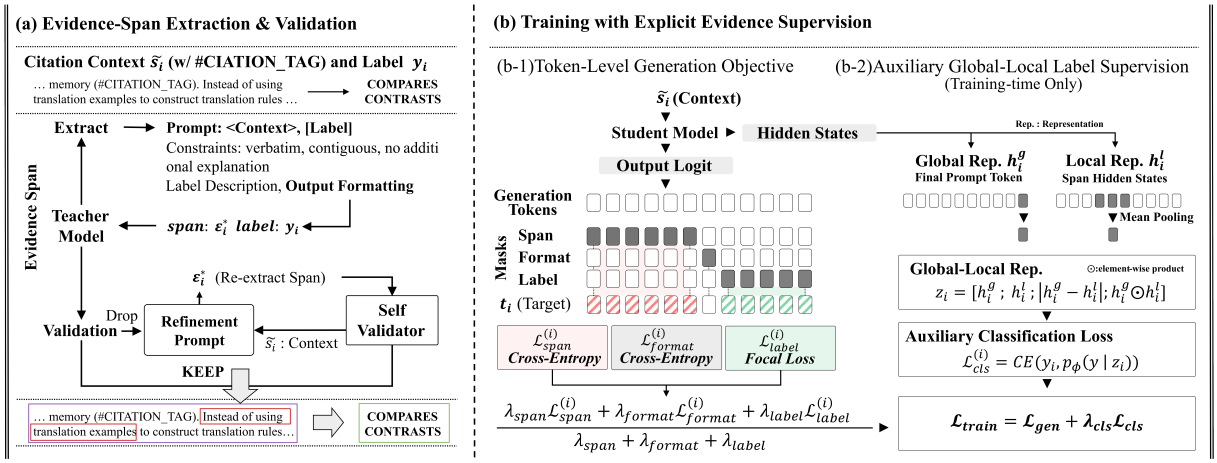

Figure 1: The framework first constructs refined evidence-span annotations from citation contexts and formulates them as span-and-label training instances. It then trains the model using token-level objectives for span, format, and label prediction, together with a training-time auxiliary classifier that aligns full-context and evidence-span representations for better label discrimination.

Overall, prior work has improved citation intent classification through schema refinement, pre-trained model adaptation, contrastive learning, semi-supervised learning, and rationale-based analysis. Although both global context and local textual cues have been considered, local evidence is often absorbed into implicit representations or used mainly for interpretation. Our work addresses this gap by explicitly extracting and validating evidence spans and using them as direct supervision signals, allowing the model to preserve global context while placing greater emphasis on local intent-bearing evidence.

3 Methodology

Figure 1 illustrates our teacher-guided rationale distillation framework for citation intent classification. Citation intent is often indicated by short local cues rather than the entire citing context, but these cues can be obscured by generic surrounding text. We therefore construct label-conditioned evidence spans with a teacher LLM and use the validated spans as intermediate supervision for student fine-tuning.

During training-data construction, the teacher model receives the citing context together with the gold citation-intent label and identifies a label-supporting extractive rationale. Thus, the span construction stage should be interpreted as label-conditioned evidence-span supervision, rather than unsupervised evidence discovery. The resulting span-label pairs are used to train a student model that learns from both the final intent label and the localized textual cue associated with that label.

Our framework consists of three stages. First, the teacher extracts a contiguous evidence span conditioned on the citing context and its intent label. Second, since teacher-generated spans may be generic or only weakly aligned with the label, we validate each span and re-extract it when necessary. Finally, the student is trained with token-level span-and-label supervision, together with a training-time global-local auxiliary objective and span-loss control.

3.1 Evidence-Span Extraction

Citation intent is often signaled by short lexical cues rather than the entire citing context. For example, phrases such as "we follow" or "unlike prior work" can indicate whether a citation is used for method adoption or comparison. Based on this observation, we perform label-conditioned evidence distillation, where a teacher model extracts a localized evidence span that supports a given citation intent label.

In this paper, we denote the i-th citing context containing the citation marker `#CITATION_TAG` as $\tilde{s}_i$, and its citation intent label as y_i. Given both the citing context $\tilde{s}_i$ and the label y_i, we define the evidence span ε_i as a contiguous substring of $\tilde{s}_i$ that directly supports the intent label:

$$\varepsilon_i = \mathcal{E}(\tilde{s}_i, y_i), \tag{1}$$

where $\mathcal{E}(\cdot)$ denotes a label-conditioned span extractor.

We implement $\mathcal{E}(\cdot)$ using a teacher LLM. Rather than asking the teacher to predict the citation intent, we provide the label y_i and prompt the teacher to

identify the minimal phrase in $\tilde{s}_i$ that best justifies that label. This design distills label-specific evidence from the teacher, encouraging the student model to associate citation intent labels with localized textual cues instead of relying only on the full citing context. To ensure faithful alignment with the original context, the extracted span must satisfy three constraints: it must be a verbatim substring of $\tilde{s}_i$, it must be contiguous, and it must not include additional explanation text.

The resulting candidate span is passed to the validation stage, where its alignment with the intent label is further examined.

Example 1. Inspired by `#CITATION_TAG`, we investigate whether localized evidence spans can improve citation intent classification. *Label:* `MOTIVATION`

Example 2. Unlike prior work, `#CITATION_TAG`, we model citation intent using localized evidence-spans. *Label:* `COMPARES_CONTRASTS`

3.2 Evidence-Span Validation

Evidence-span validation. Although the teacher LLM extracts spans, some spans can be generic, ambiguous, or only weakly aligned with the intent label. We therefore validate each candidate span before using it for training. Given a citing context $\tilde{s}_i$, a candidate span ε_i, and its label y_i, the validator assigns either `KEEP` or `DROP`. A span is kept if it contains an explicit intent-bearing cue that supports y_i; otherwise, the teacher LLM is re-prompted to extract an alternative span from the same context. This process is repeated until a valid span is found or a maximum number of attempts is reached.

Some citing contexts express citation intent only implicitly and may not contain a clear lexical cue. To preserve dataset coverage, we retain the final refined span rather than removing such instances. During training, we reduce the influence of potentially weak spans through span-specific weighting and loss control, allowing reliable evidence-spans to guide learning while preventing uncertain spans from dominating the supervision signal.

We denote the final validated span as ε_i^* and define the refined training set as:

$$\mathcal{D}_{\text{train}}^{\text{refined}} = \{(\tilde{s}_i, \varepsilon_i^*, y_i)\}_{i=1}^{N}. \tag{2}$$

This iterative validation strategy preserves training instances while improving the reliability of span-level supervision.

3.3 Student Training with Label-Conditioned Rationales

Using the refined training set $\mathcal{D}_{\text{train}}^{\text{refined}}$, we train the model with explicit evidence-centered supervision. For each instance $(\tilde{s}_i, \varepsilon_i^*, y_i)$, where $\tilde{s}_i$ is the instruction-formatted citing context, ε_i^* is the validated span, and y_i is the citation intent label, we construct the target sequence as

$$\mathbf{u}_i = \langle \varepsilon_i^* \rangle \ \texttt{\textbackslash t}\ [y_i]\ \langle \text{eos} \rangle. \tag{3}$$

Thus, the model is trained to generate the supporting evidence-span before producing the final intent label, encouraging greater focus on localized intent-bearing evidence.

Token-level generation objective. We divide the target sequence into evidence-span, label, and format tokens using binary masks, and compute separate losses for each component: $\mathcal{L}_{\text{span}}^{(i)}$, $\mathcal{L}_{\text{label}}^{(i)}$, and $\mathcal{L}_{\text{format}}^{(i)}$. The span and format components are optimized with standard cross-entropy, while the label component uses focal loss to emphasize difficult intent predictions. To reduce the effect of noisy or unstable span supervision, we apply span-specific loss control to the span-token losses. In the main setting, we use winsorization with a batch-level quantile threshold and a fixed upper cap.

The generation loss is defined as

$$\mathcal{L}_{\text{gen}}^{(i)} = \frac{\lambda_{\text{span}}\mathcal{L}_{\text{span}}^{(i)} + \lambda_{\text{label}}\mathcal{L}_{\text{label}}^{(i)} + \lambda_{\text{format}}\mathcal{L}_{\text{format}}^{(i)}}{\lambda_{\text{span}} + \lambda_{\text{label}} + \lambda_{\text{format}}}. \tag{4}$$

Auxiliary global-local label supervision. In addition to the generation objective, we use a training-time auxiliary classifier to strengthen citation intent discrimination. We obtain a global representation $\mathbf{h}_i^g$ from the final prompt token and a local representation $\mathbf{h}_i^l$ by mean-pooling the prompt-side hidden states aligned with the validated evidence span. The two representations are fused as

$$\mathbf{z}_i = [\mathbf{h}_i^g; \mathbf{h}_i^l; |\mathbf{h}_i^g - \mathbf{h}_i^l|; \mathbf{h}_i^g \odot \mathbf{h}_i^l], \tag{5}$$

and optimized with an auxiliary classification loss:

$$\mathcal{L}_{\text{cls}}^{(i)} = -\log p_\phi(y_i \mid \mathbf{z}_i). \tag{6}$$

This classifier is used only as an auxiliary training signal, while the main inference protocol remains generation-based.

Dataset	Train	Val	Test	BKG	MOT	CMP	FUT	EXT	USE	Context Length		
				BKG	MET	RES	-	-	-	min	max	avg
ACL_ARC	1399	246	284	990	88	350	69	72	360	40	2449	721
ACT2	3200	400	400	2192	382	489	77	230	630	49	747	242
SCICITE	8194	916	1859	6374	3125	1470	-	-	-	60	898	220

Table 1: Dataset statistics for citation intent classification. The table reports the number of instances in the train, validation, and test splits, the label distribution for each dataset, and the minimum, maximum, and average context lengths.

Final objective. The final training objective is

$$\mathcal{L}_{\text{train}} = \mathbb{E}_{(\tilde{s}_i, \varepsilon_i^*, y_i) \sim \mathcal{D}_{\text{train}}^{\text{refined}}} \left[\mathcal{L}_{\text{gen}}^{(i)} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{(i)} \right]. \tag{7}$$

4 Experiments

4.1 Dataset

We evaluate our method on three citation intent classification benchmarks: ACL-ARC (Bird et al., 2008), ACT2 (Nambanoor Kunnath et al., 2022), and SCICITE (Cohan et al., 2019) ACL-ARC contains citation contexts annotated with six citation functions: BACKGROUND, COMPARES_CONTRASTS, EXTENSION, FUTURE, MOTIVATION, and USES. ACT2 follows the same six-label schema as ACL-ARC and provides citation intent annotations with a standardized citation-marker handling protocol. SCICITE is a larger benchmark with three citation intent labels: BACKGROUND, METHOD, and RESULT. For all datasets, we use the official train/validation/test splits when available.

Each instance consists of a citing context containing a target citation marker. We replace the target marker with the special token #CITATION_TAG. To avoid ambiguity from multiple citations in the same context, we remove non-target citation markers using a regular expression while keeping #CITATION_TAG unchanged. Table 1 reports the number of instances in each split, the label distribution, and the minimum, maximum, and average context lengths for each dataset.

4.2 Models

We use pretrained large language models as backbone generators for citation intent classification.

Because citation intent datasets are relatively small, we adopt parameter-efficient fine-tuning with moderate-size open-weight LLMs. We use Llama-3.1-8B (Grattafiori et al., 2024) and Qwen3-4B (Yang et al., 2025) as trainable backbones, covering different model families and parameter scales. This setting allows us to test whether evidence-centered supervision is effective beyond a single backbone while avoiding full fine-tuning of larger models. We also include Llama-3.3-70B as a larger instruction-following backbone for prompting-based comparison, without task-specific fine-tuning.

For evidence span extraction, we use Grok-4.1-Reasoning (xAI, 2025) as a teacher LLM. The teacher is prompted to return a minimal, contiguous, and verbatim evidence span from the citing context. For evidence-span validation, we use the same teacher-based validation protocol described in Section 3.1.

We fine-tune Llama-3.1-8B and Qwen3-4B using LoRA (Hu et al., 2021). The pretrained backbone weights are frozen, and only the low-rank adapter parameters are updated.

4.3 Implementation Details

We formulate citation intent classification as an instruction-following sequence generation task. Full prompt templates for evidence-span extraction, validation, training, and inference are provided in Appendix B.

We train the model with span-, label-, and format-token losses. The format-token loss stabilizes the output structure, while the ablation study focuses on the span and label components, which are directly related to evidence grounding and intent prediction. Unless otherwise specified, we set $\lambda_{\text{span}} = 0.7$, $\lambda_{\text{label}} = 1.0$, $\lambda_{\text{format}} = 0.1$, and $\gamma = 1.5$. We additionally apply winsorized span-loss control, capping span-token losses at the 0.90 quantile with a maximum cap of 5.0. The label and format losses are left unchanged.

We also add a training-time auxiliary label supervision loss with weight $\lambda_{\text{cls}} = 0.8$. This classifier is used only as an auxiliary training signal, while inference remains generation-based.

For LoRA fine-tuning, we use $r = 32$, $\alpha = 64$, dropout 0.05, batch size 2, gradient accumulation 2,

Scaffold		ACL-ARC		ACT2		SCICITE	
Backbone	Method	ACC	Macro-f1	ACC	Macro-f1	ACC	Macro-f1
SciBERT Finetune		0.7430	0.6979	0.5440	0.2540	0.8592	0.8558
Citss(Li et al., 2025a)		0.819 ± 0.007	0.744 ± 0.010	0.549 ± 0.017	0.266 ± 0.006	-	-
(Nguyen et al., 2026)		-	0.6848	-	-	-	0.8684
Grok-4.1-Reasoning	IFP	0.6267	0.4755	0.5100	0.1941	0.4911	0.5088
	CoT	0.7042	0.5830	0.5475	0.2480	0.5099	0.5249
Llama3.3-70B	IFP	0.6584	0.5024	0.5225	0.1880	0.5449	0.5548
	CoT	0.7042	0.5775	0.5475	0.1931	0.5917	0.5910
Llama3.1-8B	IFP	0.5809	0.4855	0.4472	0.1969	0.5562	0.6536
	LoRA	0.7899 ± 0.0178	0.6842 ± 0.0248	0.5475 ± 0.0100	0.2727 ± 0.0258	0.8595 ± 0.0062	0.8555 ± 0.0087
	Ours	0.8042 ± 0.0054	0.7439 ± 0.0109	$\mathbf{0.6175 \pm 0.0194}$	$\mathbf{0.4401 \pm 0.0162}$	0.8719 ± 0.0121	0.8622 ± 0.0126
Qwen3-4B	IFP	0.5528	0.4924	0.5200	0.2398	0.4706	0.4883
	LoRA	0.7697 ± 0.0108	0.6862 ± 0.0042	0.5475 ± 0.0098	0.3077 ± 0.0329	0.8687 ± 0.0124	0.8599 ± 0.0153
	Ours	$\mathbf{0.8204 \pm 0.0133}$	$\mathbf{0.7486 \pm 0.0111}$	$\mathbf{0.6125 \pm 0.0140}$	$\mathbf{0.4203 \pm 0.0087}$	$\mathbf{0.8843 \pm 0.0121}$	$\mathbf{0.8744 \pm 0.0126}$

Table 2: Main results on ACL-ARC, ACT2, and SCICITE. We report accuracy (ACC) and macro-F1 for each method and backbone. Fine-tuned results are averaged over 10 runs with 5 random seeds and shown with standard deviations.

Backbone	Parameters			ACL-ARC		ACT2		SCICITE		Description
	λ_{span}	λ_{cls}	λ_{label}	acc	macro-f1	acc	macro-f1	acc	macro-f1	
Llama3.1-8B	0.0	0.0	1.0	0.7568	0.6727	0.5173	0.2439	0.8412	0.8305	only label
	0.7	0.0	1.0	0.7781	0.7014	0.5285	0.3571	0.8594	0.8552	w/o global
	0.0	0.8	1.0	0.7883	0.6918	0.5814	0.2366	0.8659	0.8421	w/o span
Qwen3-4B	0.0	0.0	1.0	0.7182	0.6604	0.5189	0.2341	0.8457	0.8423	only label
	0.7	0.0	1.0	0.7891	0.7073	0.5426	0.3128	0.8641	0.8619	w/o global
	0.0	0.8	1.0	0.7725	0.6911	0.5923	0.2074	0.8768	0.8435	w/o span

Table 3: Effect of label and span supervision weights. We vary λ_{cls} and λ_{span} to examine the balance between citation intent prediction and evidence-span supervision, while keeping the format loss weight fixed. Note that the ablation of label supervision ($\lambda_{\text{label}} = 0$) is omitted as it completely disrupts the model's prediction capability. For simplicity, we report the mean values without standard deviations.

learning rate 1×10^{-4}, maximum gradient norm 1.0, and 10 training epochs. LoRA adapters are applied to the attention and feed-forward projection layers, and bfloat16 precision is used when supported.

We use `Grok-4.1-reasoning` from xAI as the teacher model for span extraction and validation. LoRA fine-tuning is conducted on three NVIDIA GeForce RTX 4090 GPUs, while inference experiments with local large LLMs are performed on three NVIDIA RTX A6000 GPUs.

4.4 Metrics

We evaluate citation intent classification using accuracy and macro-F1. Accuracy measures the overall proportion of correctly predicted labels, while macro-F1 computes the unweighted average of F1 scores across labels. We report both metrics on the test split, but focus primarily on macro-F1 because all evaluated citation intent datasets exhibit label imbalance. In this setting, accuracy can be biased toward frequent labels, whereas macro-F1 better reflects balanced performance across citation intent classes. Invalid outputs are counted as errors rather than excluded from evaluation. In our experiments, all outputs contained valid labels, so all test instances were included when computing accuracy and macro-F1.

5 Results

5.1 Overall Performance

Table 2 presents the main results on ACL-ARC, ACT2, and SCICITE. We compare prompting-based baselines, standard LoRA fine-tuning, and our evidence-centered supervision method across Llama3.1-8B and Qwen3-4B. We also include recent citation-specific systems, such as Citss (Li et al., 2025b) and Nguyen et al. (Nguyen et al., 2026), for reference.

Overall, our method consistently improves over LoRA across all datasets and both backbones. Notably, the best variant of our method achieves the highest macro-F1 on all three datasets, outperforming both prompting-based baselines and recent citation-specific systems in Table 2. The largest improvement is observed on ACT2 with Llama3.1-8B, where macro-F1 increases from 0.2727 to 0.4401, corresponding to a gain of $+0.1674$ over LoRA.

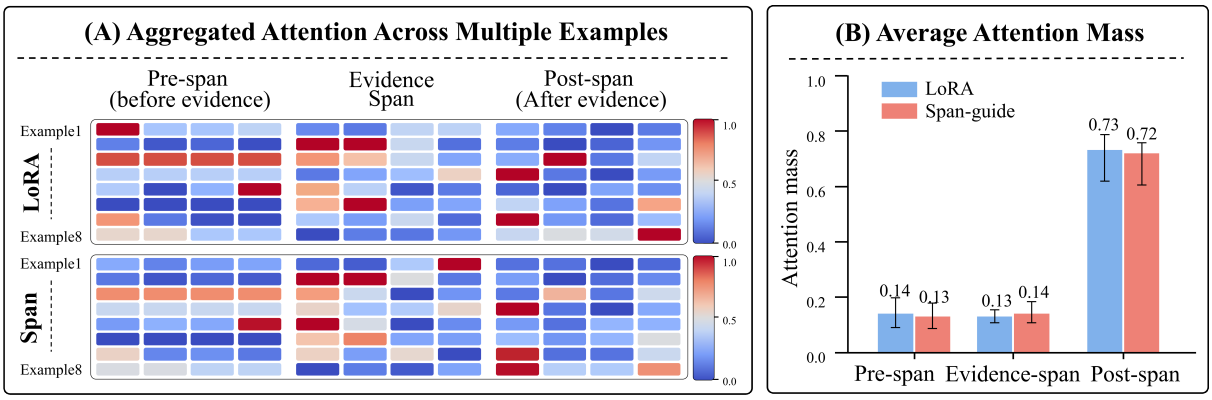

Figure 2: Attention mass over pre-span, evidence-span, and post-span regions. The span-guided model places relatively more attention on the evidence-span region than standard LoRA, indicating that evidence-span supervision helps guide the model toward citation-relevant textual cues.

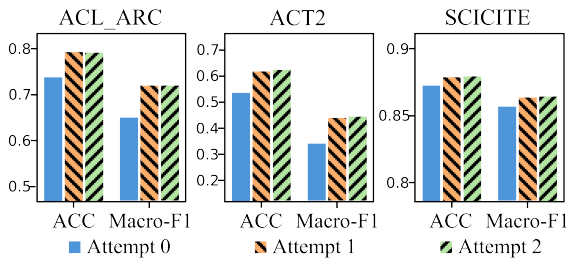

Figure 3: Effect of evidence-span validation attempts on citation intent classification performance. The results compare accuracy and macro-F1 after each validation attempt across ACL-ARC, ACT2, and SCICITE.

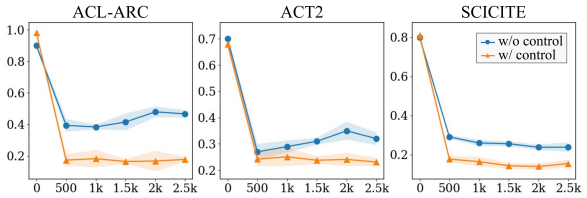

Figure 4: Training stability with and without span-loss control. Span-specific control stabilizes the training dynamics across ACL-ARC, ACT2, and SCICITE, reducing large fluctuations caused by noisy or weakly aligned evidence-span supervision.

These results show that evidence-centered supervision provides useful learning signals beyond label-only fine-tuning and leads to stronger citation intent classification performance across different datasets.

On ACL-ARC, our method also achieves the best overall macro-F1, although the improvement over LoRA is smaller than on ACT2. This may be because ACL-ARC has much longer citing contexts, making compact evidence-span extraction more difficult, as shown in Table 1. On SCICITE, the gains are smaller but consistent, likely due to its simpler three-label schema and already strong LoRA baseline.

5.2 Effect of Evidence-Centered Supervision

We further analyze the effect of evidence-span generation and auxiliary label supervision. Since label supervision is required for meaningful citation intent classification and the format loss mainly stabilizes the output structure, we exclude $\lambda_{\text{label}} = 0$ settings and keep λ_{format} fixed. The ablation therefore focuses on how the span-generation objective and the auxiliary classifier contribute on top of label generation.

Table 3 shows that adding either evidence-span generation or auxiliary label supervision alone provides some improvement over label-only training. However, these individual gains are relatively modest. The largest improvement is observed when both supervision signals are used together. This suggests that evidence-span generation and auxiliary global-local supervision are complementary: span generation encourages the model to learn local intent-bearing evidence, while the auxiliary classifier reinforces this signal by connecting the local evidence representation with the full citing context.

6 Analysis

6.1 Attention to Evidence Spans

We analyze whether evidence-span supervision changes where the model attends when predicting citation intent. Figure 2 shows the attention distribution after splitting each citation context into three parts based on the evidence span: the pre-span region, the evidence span, and the post-span region. This allows us to examine whether the model pays more attention to the text segment that directly supports the intent label.

Compared with standard LoRA, the span-guided

model shows slightly higher attention around the evidence-span region. Although both models still assign a large amount of attention to the post-span region, the span-guided model more clearly highlights tokens near the supporting span in several examples. This suggests that evidence-span supervision encourages the model to use the relevant local cue rather than relying only on the full context.

Rather, it shows that span supervision can shift part of the model's attention toward the text region that is more directly related to the citation function. This supports our main assumption that evidence spans can help the model connect citation intent labels with their supporting textual evidence.

6.2 Analysis of Span-Loss Control

As discussed earlier, automatically extracted evidence spans are not always reliable. To prevent such spans from dominating training, we apply span-specific loss control to the evidence-span component.

As shown in Figure 4, span-loss control has little effect on the final macro-F1 score, but it consistently reduces validation loss. This suggests that the span signal becomes easier to optimize when extreme span-token losses are controlled. The reduction is particularly large on ACL-ARC, which has much longer citing contexts than ACT2 and SCICITE. These results indicate that span supervision can be less stable in long contexts, where the intent-bearing cue is harder to isolate into a compact span. This may also explain why the accuracy improvement on ACL-ARC is relatively less pronounced, even though macro-F1 remains competitive.

6.3 Effect of Evidence-Span Validation

Figure 3 shows the effect of evidence-span validation on classification performance. Across all three datasets, performance generally improves as teacher-generated spans are validated and refined, indicating that the initial spans can be noisy or weakly aligned with the intent label.

The largest improvement is observed on ACL-ARC. This supports the discussion in Section 6.2, where we argued that longer citing contexts make evidence-span extraction less stable. Since ACL-ARC has longer contexts on average, validation provides a larger benefit by filtering or refining unreliable span candidates.

Most of the gain, however, is achieved by the first validation attempt, while later attempts yield only marginal changes. This is consistent with Section D, suggesting diminishing returns from repeated validation: a single refinement step is often enough to correct many low-quality spans.

7 Conclusion

In this paper, we proposed an explicit evidence-span supervision framework for citation intent classification. Instead of training the model only with citation intent labels, we used teacher-validated evidence spans as direct supervision signals and formulated the task as evidence-to-label generation. Experiments on ACL-ARC, ACT2, and SCICITE show that our method consistently improves macro-F1 over standard LoRA fine-tuning across two backbone models. Further analysis suggests that evidence-span supervision works best when localized evidence can be reliably identified and reinforced by auxiliary global-local supervision. The analysis further demonstrates that span-loss control improves training stability by reducing the influence of noisy or weakly aligned spans. Overall, our results suggest that evidence spans are useful not only as interpretability artifacts, but also as direct supervision signals for citation intent learning.

8 Limitations

Our study has two main limitations. First, the quality of evidence spans depends on the teacher-based extraction and validation process. Although validation reduces noisy spans, some spans may still be generic, incomplete, or weakly aligned with the intent label, especially in long citing contexts.

Second, our experiments are limited to citation intent classification. Although citation intent is a suitable testbed for local evidence-centered supervision, it remains unclear whether the same framework generalizes to other scientific document understanding tasks or broader text classification settings.

In future work, we plan to apply the proposed framework to other tasks that require explicit evidence grounding, such as scientific claim classification, evidence-based document classification, and rationale-aware text classification. Such studies would help evaluate whether explicit evidence spans provide stronger and more informative supervision signals for classification tasks beyond citation intent classification.

References

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Andres Carvallo, Denis Parra, Gabriel Rada, Daniel Perez, Juan Ignacio Vasquez, and Camilo Vergara. 2020. Neural language models for text classification in evidence-based medicine. *Preprint*, arXiv:2012.00584.

Krittin Chatrinan, Thanapon Noraset, and Suppawong Tuarob. 2025. Gan-cite: leveraging semi-supervised generative adversarial networks for citation function classification with limited data. *Scientometrics*, 130(2):679–703.

Po-Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Categorizing citation relations in scientific papers based on the contributions of cited papers. *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 384–389.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025, St. Julian's, Malta. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Huaming Du, Lei Yuan, Cancan Feng, Guisong Liu, Gang Kou, and Carl Yang. 2025. Explainable text classification with LLMs: Enhancing performance through dialectical prompting and explanation-guided training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12800–12816, Suzhou, China. Association for Computational Linguistics.

Changxu Duan and Zhiyin Tan. 2025. *Semantically Orthogonal Framework for Citation Classification: Disentangling Intent and Content*, page 183–206. Springer Nature Switzerland.

Leon Eshuijs, Shihan Wang, and Antske Fokkens. 2025. Short-circuiting shortcuts: Mechanistic investigation of shortcuts in text classification. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 105–125, Vienna, Austria. Association for Computational Linguistics.

Jimly Firdaus and Fariska Zakhralativa Ruskanda. 2025. Fine-tuning scibert for citation intent classification: An empirical study. In *2025 IEEE International Conference on Data and Software Engineering (ICoDSE)*, pages 1–6.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Arshad Iqbal, Muhammad Inam Ul Haq, Abdul Wahid, Shad Muhammad, Muhammad Yahya, and Abdul Shahid. 2026. Scibert-based interpretable framework for citation influence classification using sparse rationale extraction. *IEEE Access*, pages 1–1.

Arshad Iqbal, Abdul Shahid, Muhammad Roman, Muhammad Tanvir Afzal, and Muhammad Yahya. 2023. Exploiting contextual word embedding for identification of important citations: Incorporating section-wise citation counts and metadata features. *IEEE Access*, 11:114044–114060.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Preprint*, arXiv:1902.10186.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Suchetha N. Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2021. A

meta-analysis of semantic classification of citations. *Quantitative Science Studies*, 2(4):1170–1215. _eprint: https://direct.mit.edu/qss/article-pdf/2/4/1170/2007871/qss_a_00159.pdf.

Tong Li, Jiachuan Wang, Yongqi Zhang, Shuangyin Li, and Lei Chen. 2025a. Adapting pretrained language models for citation classification via self-supervised contrastive learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 1541–1552, New York, NY, USA. Association for Computing Machinery.

Tong Li, Jiachuan Wang, Yongqi Zhang, Shuangyin Li, and Lei Chen. 2025b. Adapting pretrained language models for citation classification via self-supervised contrastive learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 1541–1552, New York, NY, USA. Association for Computing Machinery.

Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. 2022. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3398–3406, Marseille, France. European Language Resources Association.

Thu Huong Nguyen, Cédric Pruski, and Marcos Da Silveira. 2026. Deepening citation understanding in scientific literature via LLM-powered context extraction. *Scientometrics*.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 337–340, New York, NY, USA. Association for Computing Machinery.

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

Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubaa, and Lisu Yu. 2021a. Citation intent classification using word embedding. *IEEE Access*, 9:9982–9995.

Muhammad Roman, Abdul Shahid, Muhammad Irfan Uddin, Qiaozhi Hua, and Shazia Maqsood. 2021b. Exploiting contextual word embedding of authorship and title of articles for discovering citation intent classification. *Complex.*, 2021:5554874:1–5554874:13.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Hong-Jin Tsai, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Citation intent classification and its supporting evidence extraction for citation graph construction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 2472–2481, New York, NY, USA. Association for Computing Machinery.

Yeo Wei Jie, Ranjan Satapathy, and Erik Cambria. 2024. Plausible extractive rationalization through semi-supervised entailment signal. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5182–5192, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

xAI. 2025. Grok 4.1 model card. `https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf`. Accessed: 2026-05-18.

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.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

A Example Appendix

A Data Construction Prompts

This section provides the label descriptions and prompt templates used to construct evidence-span annotations. The label descriptions are shared across data construction, fine-tuning, and in-context learning experiments. The extraction and validation prompts are used to obtain refined evidence spans that are later used as additional supervision.

A.1 Label Descriptions

We use dataset-specific label sets because ACL-ARC and ACT2 define six citation functions, while SCICITE uses three citation intent classes. These descriptions are inserted into the prompts through the `{label_description}` field.

A.1.1 ACL-ARC and ACT2

```
The six citation function classes are:

- [BACKGROUND]; Q provides relevant
    information or is part of the general
    literature.
- [COMPARES_CONTRASTS]; P compares,
    contrasts, or disagrees with Q.
- [EXTENSION]; P extends data, methods, or
    results from Q.
- [FUTURE]; Q represents a potential
    direction for future work.
- [MOTIVATION]; Q directly motivates the
    research in P.
- [USES]; P uses tools, methods, or datasets
     from Q.
```

A.1.2 SCICITE

```
The three citation function classes are:

- [METHOD]; P cites Q for a method, tool,
    model, dataset, framework, or other
    technical resource used in the work.
- [BACKGROUND]; Q provides background
    information, general context, related
    work, or supporting knowledge.
- [RESULT]; P cites Q for a finding, result,
     observation, or conclusion reported in
    Q.
```

A.2 Evidence-Span Extraction Prompt

This prompt is used to extract a keyphrase-level evidence span from each citation context. The extracted span is expected to be short, contiguous, and directly related to the given citation intent label.

Evidence-Span Extraction

```
You are given a citation context from a
    paper P citing a paper Q.
The target citation is denoted by #
    CITATION_TAG.

Your task is to extract one evidence span
    that best supports the given citation
    intent label.

Rules:
- Copy the evidence span verbatim from the
    citation context.
- The span must be contiguous.
- The span must not include #CITATION_TAG.
- The span should be short and directly
    related to the given label.
- Do not add explanations.

Input:
Context: {context}
Label: {label}

Output format:
Evidence Span: <span>
```

A.3 Evidence-Span Validation

Since automatically extracted spans can be generic or weakly related to the label, we use a separate validation prompt. The validation step filters unreliable spans before they are used for model training.

Evidence-Span Validation Prompt

```
You are given a citation context, a citation
     intent label, and an extracted evidence
     span.

Your task is to decide whether the evidence
    span directly supports the given label.

Validation criteria:
- The span must be copied verbatim from the
    context.
- The span must be contiguous.
- The span must not include #CITATION_TAG.
- The span must provide clear evidence for
    the citation intent label.
- Generic or weakly related spans should be
    rejected.

Input:
Context: {context}
Label: {label}
Evidence Span: {span}

Output format:
Decision: [KEEP] or [DROP]
Reason: <brief reason>
```

B LoRA Fine-Tuning Prompts

This section provides the prompt templates used for LoRA fine-tuning. We include both the label-only prompt used for standard fine-tuning and the span-based prompt used in our proposed method.

B.1 Label-Only LoRA Prompt

This prompt is used for the standard LoRA baseline. The model receives only the citation context and is trained to output the final citation intent label.

Label-Only LoRA

```
You are provided a citation context from a
    paper P citing a paper Q, with the
    specific citation denoted by the token #
    CITATION_TAG.

Your task is to classify the citation
    function, i.e., the rhetorical purpose
    for citing Q in this context.

{label_description}

Here is the citation context:
"{context}"

Please classify the purpose of citing #
    CITATION_TAG in this context.

Only output one label enclosed in square
    brackets.

Output format:
[LABEL]
```

B.2 Span-Based LoRA Prompt

This prompt is used for our span-supervised fine-tuning setting. Unlike the label-only prompt, the model is trained to output both the evidence span and the final label, allowing token-level supervision for span and label prediction.

Span-Based LoRA

```
You are provided a citation context from a
    paper P citing a paper Q, with the
    specific citation denoted by the token #
    CITATION_TAG.

Your task is to:
1. Choose one evidence span from the context
     that best indicates the citation
    function of #CITATION_TAG.
2. Classify the citation function into one
    of the labels below.

Evidence span rules:
- Copy a contiguous span verbatim from the
    context.
- Preserve case and punctuation.
- The span must come from the context itself.

- The span must not include #CITATION_TAG.
- Pick the span that most strongly supports
    the label.

{label_description}

Before generating the final output,
    internally check whether the selected
    span directly supports the chosen label.
If the span is weak or better supports
    another label, refine both the span and
    the label.

Here is the citation context:
"{context}"

Output format:
<EVIDENCE SPAN> [LABEL]
```

C In-Context Learning Prompts

This section provides the prompt templates used in the in-context learning experiments. These prompts are used to test whether evidence-span guidance is useful without updating model parameters.

C.1 Instruction-Following Prompt

Instruction-Following Prompt

```
You are provided a citation context from a
    paper P citing a paper Q, with the
    specific citation denoted by the token #
    CITATION_TAG.

Your task is to classify the citation
    function, i.e., the rhetorical purpose
    for citing Q in this context.

{label_description}

Citation context:
"{context}"

Output ONLY the final label in this format:
[LABEL]
```

C.2 Chain-of-Thought Prompt

Chain-of-Thought Prompt

```
You are provided a citation context from a
    paper P citing a paper Q, with the
    specific citation denoted by the token #
    CITATION_TAG.

Your task is to classify the citation
    function into one of the labels below.

{label_description}

Before answering, reason internally as
    follows:
1. Identify the main rhetorical role of the
    citation.
2. Determine whether the cited work is being
     used, extended, contrasted, motivating
    the work, suggested for future work, or
    mentioned as general background.
3. Select the single best label.

Citation context:
"{context}"

Output ONLY the final label in this format:
```

Dataset	First Validation		Second Validation		Third Validation	
	KEEP	DROP	KEEP	DROP	KEEP	DROP
ACL-ARC	50.3%	49.7%	65.4%	34.6%	66.2%	33.8%
ACT2	64.8%	35.2%	79.7%	20.3%	80.6%	19.4%
SCICITE	65.2%	34.8%	75.3%	24.7%	77.1%	22.9%

Table 4: KEEP and DROP ratios across evidence-span validation attempts. DROP is computed as the complement of KEEP.

```
[LABEL]
```

D Quality of Teacher-Generated Spans

Table 4 reports the KEEP and DROP ratios across repeated evidence-span validation attempts. Across all datasets, the KEEP ratio increases as validation proceeds, indicating that re-extraction helps refine generic or weakly aligned teacher-generated spans. The improvement is especially clear from the first to the second validation attempt: ACL-ARC increases from 50.3% to 65.4%, ACT2 from 64.8% to 79.7%, and SCICITE from 65.2% to 75.3%. In contrast, the third validation attempt yields only marginal additional gains.

This suggests diminishing returns from repeated validation, where most low-quality spans can be corrected within the early refinement steps. At the same time, the KEEP ratio does not reach 100%, indicating that span-level supervision inevitably contains some noise. This is partly because some citation contexts express their intent implicitly or distribute the citation function across a broader context, making it difficult to isolate a single compact evidence span. Therefore, we do not treat teacher-generated spans as perfectly reliable rationales; instead, we use validation and span-loss control to reduce the influence of generic or weakly aligned spans during training.

E Artifacts and Licenses

We use ACL-ARC, ACT2, and SCICITE for citation intent classification experiments, and cite the original dataset sources in the main paper. We use pretrained language models and software libraries, including Llama, Qwen, Transformers, PEFT, and LoRA, according to their respective licenses or terms of use. The teacher model is accessed through its official service or API under the corresponding terms of use. Our released code is intended for research reproducibility.