Does Rhetorical Structure Matter More Than Linguistic Proximity? A Study on Cross-Lingual Sequential Sentence Classification

Anonymous ACL submission

Abstract

Sequential sentence classification (SSC) is an essential task for structuring scientific publications, and extending SSC research to languages other than English would improve accessibility to scientific knowledge. At present, cross-lingual transfer is a promising approach to address the scarcity of training data in non-English languages. Although prior work on other natural language processing tasks has shown the benefits of identifying linguistic similarity between source and target languages, SSC inherently depends on discourse-level patterns, such as label sequences and positional regularities, which exhibit consistency across languages regardless of linguistic differences. To examine the factors that determine transfer success in SSC, we construct a multilingual SSC dataset covering 13 non-English languages. Our cross-lingual transfer experiments, which use both encoder-based and generative models, reveal that structural similarity in rhetorical organization correlates more strongly with transfer performance than linguistic proximity, and this pattern holds across different model architectures. Based on this finding, we propose a novel framework that explicitly leverages structural information to improve SSC, demonstrating improvements over baselines in both in-language evaluation and transfer to languages unseen during training.

1 Introduction

Sequential sentence classification (SSC) is the task of categorizing each sentence into specific rhetorical roles, such as Background, Objective, Method, Result, and Concusion. As a foundational technology, SSC enhances downstream applications, including literature searches, automatic summarizations, and paper recommendation systems. Recent advancements in SSC have leveraged hierarchical architectures based on Transformer-based models (Devlin et al., 2019; Dernoncourt and Lee, 2017; Jin and Szolovits, 2018; Cohan et al., 2019; Brack et al., 2024) and large language models (LLMs) (Lan et al., 2024), reaching unprecedented performance levels. However, these developments have been predominantly centered on English-language datasets. This English-centric focus creates a significant gap in the accessibility of non-English academic research, which has necessitated the development of SSC technologies that can generalize across diverse languages.

Cross-lingual transfer learning, facilitated by multilingual pre-trained language models (mPLMs), has emerged as a key strategy to bridge this gap. In this paradigm, knowledge is transferred from a source language (typically English) to a target non-English language. Traditionally, the success of such transfer was assumed to depend on the linguistic proximity between the language pair (Philippy et al., 2023; Lin et al., 2019). However, recent studies suggest that linguistic proximity is not always a reliable proxy for transfer success, as its effectiveness as a predictor varies significantly across different tasks (Blaschke et al., 2025). Furthermore, similarities within the internal representation spaces of mPLMs have been shown to correlate more strongly with transfer outcomes than surface-level linguistic traits (Lin et al., 2024; Yun et al., 2023). These findings underscore the need to identify task-specific factors that govern cross-lingual transfer beyond simple linguistic distance.

In this study, we focus on the intrinsic structural properties of SSC as a key factor in facilitating such transfer. Academic abstracts consistently follow logical progressions regardless of the language. For instance, Background typically appears at the beginning, while Result follows Method. To investigate whether these structural commonalities facilitate transfer, we constructed a comprehensive multilingual SSC dataset covering 13 non-English languages with approximately 32,000 abstracts. Our

empirical analysis reveals that structural similarity, defined by the alignment of positional distributions, label transitions, and class distributions, is a significantly stronger predictor of transfer performance than traditional linguistic metrics.

Building on these insights, we propose a novel framework that explicitly leverages structural information to enhance cross-lingual SSC. Our approach has three methodological variants: (1) **Structure-Informed Prompting (SIP)**, which injects structural knowledge into LLM prompts; (2) **Structure-Guided Verifier Reranking (SGVR)**, which utilizes a trained verifier to optimize candidate sequences; and (3) **Structure-Adaptive Verifier (SAV)**, which ensures prediction consistency in zero-shot settings.

The main contributions of this study are as follows:

- We introduce a large-scale multilingual SSC dataset across 13 languages and demonstrate, through empirical analysis, that structural similarity correlates more strongly with cross-lingual transfer performance than linguistic proximity.
- We propose a structural information-driven framework that achieves significant improvements in macro F1 scores over existing multilingual baselines.

2 Related Work

2.1 Sequential Sentence Classification

Since the release of the PubMed-RCT benchmark (Dernoncourt and Lee, 2017), SSC models have evolved from hierarchical neural architectures (Jin and Szolovits, 2018) to Transformer-based approaches (Cohan et al., 2019). To effectively capture sequential dependencies and rhetorical flow, modern architectures leverage hierarchical modeling to integrate both sentence-level representations and document-level context (Brack et al., 2024). However, existing SSC research remains predominantly focused on English.

2.2 Cross-Lingual Transfer Learning

Recent research has employed multilingual BERT (mBERT) (Devlin et al., 2019), which enabled zero-shot cross-lingual transfer across 104 languages. While transfer performance often correlates with typological similarity (Lauscher et al., 2020) and word-order alignment (Deshpande et al., 2022), discourse-level research suggests that rhetorical structures are often preserved across languages (Braud et al., 2017; Zeyrek et al., 2020). In light of these findings, we hypothesize that for SSC, structural similarity in rhetorical organization will be a more critical predictor of transfer success than traditional typological measures.

2.3 Structure-Aware Methods

Recent structure-aware approaches have improved document modeling (Buchmann et al., 2024) and argumentation mining (Sun et al., 2024). Such research has employed grammar-constrained decoding (Geng et al., 2023; Park et al., 2024) and discriminative reranking (Wang et al., 2024) to ensure output consistency. In light of these developments, our study adapts the verifier-reranking paradigm (Cobbe et al., 2021) to SSC, using structural features to estimate the candidate quality without requiring gold labels.

3 Multilingual SSC Dataset

Following PubMed-RCT 20k (Dernoncourt and Lee, 2017), which leverages abstracts with explicit section[1] headers (e.g., Background, Method, Result, Concusion) as ground-truth labels, we constructed a multilingual dataset from non-English abstracts containing such headers.

Our data sources consisted of five academic databases that provide APIs for bulk data retrieval: DOAJ,[2] HAL,[3] Dialnet,[4] TRdizin,[5] and CiNii Research.[6] We targeted 13 non-English languages: French, Japanese, Spanish, Chinese, Russian, Portuguese, Italian, Indonesian, Turkish, Korean, Polish, Dutch, and Estonian. To identify abstracts with section headers, we translated section terms (e.g., "Method," "Result") into each target language and used them as search queries, then verified the presence of the headers through pattern matching for three formats: XML (`<sec><title>Methods</title>...</sec>`), bracket ([Methods]), and colon (Methods:). Data collection was conducted between February and May 2025.

We then applied preprocessing steps, including

[1] A "section" refers to a segment of the abstract corresponding to a specific rhetorical role (e.g., method, Result).

[2] https://doaj.org

[3] https://hal.science

[4] https://dialnet.unirioja.es

[5] https://trdizin.gov.tr

[6] https://cir.nii.ac.jp

Table 1: Dataset statistics.

Language (abbr.)	Source	#Papers	#Sentences
English (en)	PubMed-RCT	20,000	180,040
French (fr)	HAL	11,210	134,393
Japanese (jp)	CiNii	8,366	78,843
Spanish (es)	Dialnet	5,768	55,743
Chinese (zh)	DOAJ	3,522	24,649
Russian (ru)	DOAJ	1,163	9,522
Portuguese (pt)	Dialnet	1,122	8,865
Italian (it)	DOAJ	624	6,353
Indonesian (id)	DOAJ	434	4,270
Turkish	TRdizin	179	630
Korean	DOAJ	48	485
Polish	DOAJ	30	369
Dutch	DOAJ	14	131
Estonian	DOAJ	7	123

HTML entity conversion, Unicode normalization, language detection, and duplicate title removal. We retained only those abstracts containing two or more sections and performed sentence segmentation using NLTK (Bird et al., 2009) for European languages, including spaCy (Honnibal et al., 2020) for Asian languages. Details are provided in Appendix A. The dataset and code are available at https://anonymous.4open.science/r/multilingual-SSC-0FDC.[7]

The dataset contained 47,487 abstracts and 504,416 sentences across 14 languages (13 collected languages plus English from PubMed-RCT 20k). Table 1 shows statistics by language.

4 Cross-Lingual Transfer Analysis

While prior work on cross-lingual transfer has primarily focused on linguistic features such as typological similarity (Pires et al., 2019), we hypothesize that for SSC, a discourse structure identification task, rhetorical similarity across languages may better facilitate successful transfer. To investigate this hypothesis, nine languages with 200+ abstracts were chosen from the dataset: Chinese, Spanish, English, French, Indonesian, Italian, Japanese, Portuguese, and Russian. We conducted comprehensive cross-lingual transfer experiments across 9×9 language pairs (including same-language pairs) using both BERT-based and LLM-based models, examining whether the findings generalize across different model architectures. Our experimental setting was zero-shot cross-lingual transfer: for each of the 81 language pairs, models were trained on the source language and evaluated on the target language without any target-language training examples.

4.1 Models

We employed multilingual BERT (mBERT; Devlin et al., 2019) with the hierarchical sequence labeling network (HSLN) architecture (Brack et al., 2024), hereafter mBERT-HSLN, as a representative BERT-based SSC model. By replacing the encoder of the conventional SSC method with mBERT, the model can be applied to abstracts in target languages different from the source language. We followed the hyperparameters from Brack et al. (2024).

For LLM-based models, we selected three mPLMs that have been fine-tuned to follow natural language instructions: Gemma2-2B-it (Gemma Team, 2024), Qwen2.5-3B-Instruct (Qwen Team, 2024), and Llama-3.2-3B-Instruct (Llama Team, AI @ Meta, 2024). The prompt used was a simplified version of that proposed by Lan et al. (2024); the complete template is provided in Appendix C. We provided the entire abstract as input and instructed the model to output a rhetorical label for its each sentence. We applied low rank adaptation (LoRA; Hu et al. (2022)) fine-tuning with $r = 8$ and $\alpha = 16$ to all attention layers, using a learning rate of 2×10^{-4}, batch size of 4, and training for three epochs. Hyperparameters were tuned using validation data from representative languages to maximize validation Macro-F1.

For each language, we split the data into 70% training, 15% validation, and 15% testing. For languages with more than 200 abstracts in the test set, we randomly sampled 200 abstracts to balance computational cost and evaluation reliability. For each language pair, we conducted three runs with different random seeds and report the average Macro-F1 scores.

4.2 Results of Zero-Shot Cross-Lingual Transfer

We observed consistent transfer patterns across all four models (mBERT-HSLN and three LLMs), suggesting that the findings are not specific to particular model architectures. Figure 1 shows the transfer results for Qwen2.5-3B-Instruct as a representative example; complete transfer matrices for all models are provided in Appendix D.

For Qwen2.5-3B-Instruct, the average Macro-F1 for same-language pairs was 0.611, while cross-lingual transfer averaged 0.515. Transfer performance varied considerably: Japanese to Chinese

[7] For data from CiNii Research, Dialnet, and TRdizin, we release document IDs instead of full abstracts to comply with their data usage policies.

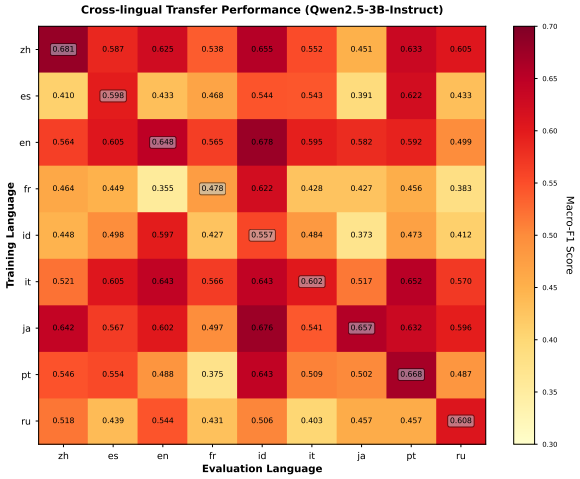

Figure 1: Cross-lingual transfer performance for Qwen2.5-3B-Instruct (Macro-F1). Rows: training languages; Columns: evaluation languages.

(0.642) and Chinese to Indonesian (0.655) showed high performance, while French to English (0.355) and Spanish to Japanese (0.391) exhibited lower performance. For mBERT-HSLN, same-language performance averaged 0.508, while cross-lingual transfer averaged 0.405, showing a similar gap between same-language and cross-lingual settings.

Across all models, common transfer patterns were observed: Japanese-Chinese transfer was relatively successful, and English as the source showed stable performance. Transfer performance was asymmetric; for example, in Qwen2.5-3B-Instruct, Japanese to Chinese (0.642) and Chinese to Japanese (0.451) differed by 0.19, likely reflecting differences in model capability for each language.

4.3 Linguistic and Structural Similarity Measures

To determine which factors predict cross-lingual transfer performance, we compared two types of similarity measures between language pairs: linguistic proximity based on typological features and structural similarity based on the rhetorical patterns found in the abstracts.

For linguistic proximity, we used lang2vec (Littell et al., 2017), which provides typological feature vectors for languages. We concatenated feature vectors from five categories (syntax, phonology, inventory, geography, family) and calculated cosine similarity between each language pair. As expected, high similarity was observed among typologically related languages (e.g., Romance languages such as Italian-Portuguese: 0.932), while distant language pairs showed lower similarity (e.g., Japanese-Portuguese: 0.647).

For structural similarity, we designed measures based on rhetorical structure patterns found in the abstracts. Drawing on feature representations used in conditional random fields (CRF) (Lafferty et al., 2001) and research on abstract composition patterns (Martín-Martín, 2003), we defined six distance measures between language pairs:

1. **Label distribution:** Jensen-Shannon divergence (JSD) between the frequency distributions of the five labels in each language.

2. **Section length distribution:** Average JSD of section length (number of consecutive sentences with the same label) distributions across labels.

3. **Continuation probability:** Normalized Euclidean distance between vectors of label continuation probabilities (probability that the next sentence has the same label).

4. **Boundary position:** Weighted difference in average relative positions of major section transitions (e.g., Method to Result).

5. **Block count distribution:** Average JSD of block count (number of non-contiguous spans per label) distributions across labels.

6. **Transition entropy:** Normalized difference in Shannon entropy of label transition distributions.

We calculated the overall structural distance d as the simple average of these six distances and defined structural similarity as $1 - d$. Detailed mathematical definitions are provided in Appendix B.

Figure 2 shows the two similarity matrices. While linguistic proximity largely reflects language family membership, structural similarity reveals different patterns: Japanese-Spanish (0.93) and Japanese-Portuguese (0.93) show high structural similarity despite belonging to different language families. Chinese-Russian (0.89) also show high structural similarity, as both tend to omit Background and start with Objective.

4.4 Correlation Analysis

To determine which similarity measure better predicts transfer performance, we analyzed the 72

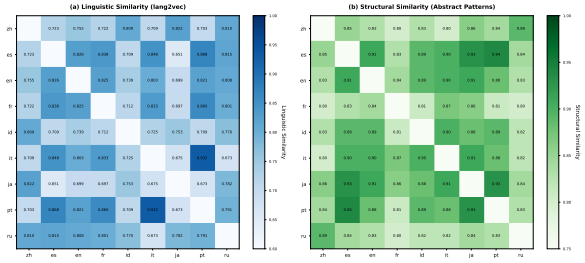

Figure 2: Linguistic (left) and structural (right) similarity matrices.

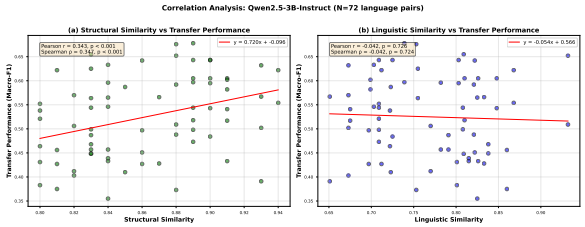

Figure 3: Correlation between similarity measures and transfer performance for Qwen2.5-3B ($N = 72$). Left: structural similarity (Pearson $r = 0.443$, $p < 0.001$). Right: linguistic proximity (Pearson $r = 0.197$, $p = 0.079$).

pairs (excluding same-language pairs) through correlation analysis. Figure 3 shows the results. Statistically significant positive correlations (Pearson's r) between structural similarity and transfer performance were found in **all four models**: mBERT-HSLN ($r = 0.408$, $p < 0.001$), Qwen2.5-3B ($r = 0.443$, $p < 0.001$), Gemma2-2B ($r = 0.333$, $p = 0.002$), and Llama-3.2-3B ($r = 0.299$, $p = 0.007$). Spearman correlations showed similar tendencies.

For linguistic proximity, no model showed consistent significant correlations. In mBERT-HSLN, both the Pearson and Spearman correlations were not significant ($r = 0.087$, $p = 0.438$; $\rho = 0.024$, $p = 0.831$). Among LLMs, only Gemma2-2B showed a significant Pearson correlation ($r = 0.278$, $p = 0.012$), but the Spearman correlation was not significant. In Qwen2.5-3B and Llama-3.2-3B, no significant correlations were found (all $p > 0.05$). Overall, linguistic proximity showed weak and inconsistent predictive power.

Together, these results indicate that in SSC tasks, structural pattern similarity better predicts transfer performance than linguistic features **regardless of model architecture**. The consistency across BERT-based (i.e., mBERT-HSLN) and LLM-based models suggests that this finding is robust and not specific to a particular model family. The correlation coefficients range from $r = 0.299$ to $r = 0.443$, meaning structural similarity explains approximately 9–20% of the variance in transfer performance. This suggests that while other factors influence transfer performance (e.g., model capability for each language, data quality, etc.), structural similarity serves as a consistent and practical predictor.

This result can be interpreted from SSC's nature, since models may rely on structural patterns such as "Objectives come after Background" and "Concusions come after Results" rather than vocabulary or grammar. The fact that the BERT-based model (which relies on hierarchical neural architectures) and LLM-based models (which use autoregressive Transformers) show similar patterns suggests that structural dependency is a fundamental characteristic of cross-lingual SSC transfer, independent of specific model architectures.

Looking at specific language pairs, tendencies supporting this interpretation could be observed. High structural similarity was found between Japanese and both Spanish and Portuguese (0.93). Despite these languages belonging to completely different language families, their scientific paper structure patterns are similar. This suggests that international conventions regarding how to write academic papers are shared across language barriers. Meanwhile, Chinese and Russian belong to different linguistic groups, but both languages tend to omit the Background section and start with Objective, and this common structural feature may be facilitating transfer learning across both BERT-based and LLM-based models.

The correlation strength varied across models, with Qwen2.5-3B and mBERT-HSLN showing the strongest correlations ($r = 0.443$ and $r = 0.408$, respectively), followed by Gemma2-2B ($r = 0.333$) and Llama-3.2-3B ($r = 0.299$). This disparity may reflect differences in each model's multilingual capability and pretraining data composition. In the future, analyzing the relationship between the proportion of each language in the models' pretraining data and transfer performance could provide deeper understanding of the reason behind this difference.

5 Leveraging Structural Information

Our analysis in Section 4 revealed that structural similarity was significantly correlated with cross-lingual transfer performance across all tested mod-

els (e.g., $r = 0.443$ for Qwen2.5-3B, $p < 0.01$). While that analysis focused on zero-shot transfer between language pairs, we also hypothesized that explicit structural information should also benefit models trained on multilingual data, as both settings require the model to generalize rhetorical patterns across languages.

To test this hypothesis, we developed a framework that explicitly leverages structural information through three variants: **SIP**, **SGVR**, and **SAV**. This approach moves beyond implicit learning during fine-tuning by incorporating structural constraints at both the input stage and during candidate selection. SIP+SGVR is used when the target language is included in the training data, which enables verifier training on validation data. SIP+SAV is used for zero-shot settings where the model is tested on languages not seen during training.

While the correlation analysis demonstrates architecture-agnostic importance of structural similarity, we developed our proposed methods for LLM-based models, as they enable flexible prompt-based guidance and candidate generation through temperature sampling. SIP+SGVR was used when training data were available, and SIP+SAV was applied for zero-shot settings.

5.1 Structure-Informed Prompting (SIP)

SIP guides LLM predictions by explicitly incorporating structural constraints into prompts. As demonstrated in Section 4, rhetorical structures exhibit similar patterns across languages. SIP explicitly specifies the typical label ordering (Background, Objective, Method, Result, Concusion) in the prompt, encouraging the model to consider not only individual sentence content but also structural context. The prompt consisted of: (1) task description with label definitions; (2) structural constraints indicating typical ordering; and (3) one demonstration example randomly selected from training data. The complete prompt is shown in Appendix E.

5.2 Structure-Guided Verifier Reranking (SGVR)

SGVR reranks multiple candidate predictions generated by an LLM based on structural features. Adopting an approach similar to quality estimation in machine translation (Specia and Shah, 2018), we used a verifier that predicts output quality without reference. Among multiple candidate label sequences obtained via temperature sampling, the verifier selects the most valid candidate.

Specifically, for each training abstract, the LLM generated K candidate label sequences $\mathbf{y} = (y_1, \ldots, y_n)$, where n is the number of the abstract's sentences and y_i is the predicted label for the i-th sentence. We computed the sentence-level accuracy $q = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i = y_i^*]$ for each candidate, where y_i^* denotes the ground truth label. From each candidate, we also extracted the following 44-dimensional feature vector $\mathbf{f}(\mathbf{y})$ capturing: (1) label distribution features (occurrence counts, cosine similarity with training data distribution); (2) transition features (mean log-probability of adjacent label transitions, binary indicators of major transitions, entropy); and (3) position features (scores evaluating whether label positions fall within expected ranges, one-hot representations of start/end labels). The complete feature list can be found in Appendix F.

Finally, we trained a regression model V_ϕ (with parameters ϕ) that predicts q from structural features $\mathbf{f}(\mathbf{y})$, assuming candidates closer to ground truth exhibit stronger structural alignment with training data patterns. We adopted LightGBM (Ke et al., 2017) for V_ϕ due to its ability to capture non-linear interactions and training efficiency. At inference, the verifier predicts accuracy $\hat{q}_k = V_\phi(\mathbf{f}(\mathbf{y}_k))$ for each candidate and selects $\mathbf{y}^{\text{pred}} = \arg\max_k \hat{q}_k$.

5.3 Structure-Adaptive Verifier (SAV)

SAV is an unsupervised method for cases of sequence labeling in which zero-shot settings apply. To address this limitation, it dynamically estimates structural properties from the candidate set, assuming that labels consistently predicted across multiple sampling runs are more likely correct. While self-consistency (Wang et al., 2023) selects the most frequent answer, simple majority voting is ineffective for sequence labeling where exact matches are rare. SAV evaluates position-wise agreement and transition consistency separately to select the most structurally valid candidate.

For each abstract, same as in SGVR, an LLM generated K candidate predictions. Here, we denote the k-th candidate label sequence as $\mathbf{y}^{(k)} = (y_1^{(k)}, \ldots, y_n^{(k)})$, where n is the number of the abstract's sentences, and $y_i^{(k)}$ is the predicted label for the i-th sentence. In SAV, each candidate prediction was evaluated using the following three scores.

Consistency score $S_{\text{cons}}(\mathbf{y}^{(k)})$: Average position-wise agreement rate between candidate $\mathbf{y}^{(k)}$ and

all other candidates. Candidates with higher agreement received higher scores.

Confidence score $S_{\text{conf}}(\mathbf{y}^{(k)})$: For each position $i \in \{1, \ldots, n\}$, we computed the most frequent label l_i^* among the K candidates and its agreement rate $r_i = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1}[y_i^{(j)} = l_i^*]$. If candidate $\mathbf{y}^{(k)}$'s label at position i matched l_i^*, the score was r_i; otherwise, it was the proportion of candidates predicting that label. The final score was the average across all positions.

Transition score $S_{\text{trans}}(\mathbf{y}^{(k)})$: Application of the same approach as the confidence score to transition patterns between adjacent labels rather than individual labels.

The final score $S(\mathbf{y}^{(k)})$ was the average of the three scores. We selected the highest-scoring candidate as the final prediction $\mathbf{y}^{\text{pred}} = \arg\max_k S(\mathbf{y}^{(k)})$. Unlike simple position-wise majority voting (selecting the most frequent label at each position and concatenating them), which may produce inconsistent sequences not present in any candidate, SAV always outputs a prediction that exists in the candidate set, thereby avoiding this issue. Implementation details and further analysis regarding the selection of hyperparameters for SAV are provided in Appendix J.

6 Experiments

6.1 Experimental Settings

Unlike the language-pair analysis in Section 4, we trained models on mixed multilingual data and evaluated them in two settings: (1) in-domain evaluation (testing on trained languages); and (2) zero-shot evaluation (testing on languages not in training).

For in-domain evaluation, we used nine languages (English, French, Japanese, Spanish, Chinese, Indonesian, Portuguese, Italian, Russian), splitting each into a 70:15:15 ratio for training/validation/testing. The model was trained on combined training data from all nine languages and evaluated on the combined test data from all languages. For zero-shot evaluation, we used five languages not in training (Estonian, Korean, Dutch, Polish, Turkish).

We applied LoRA ($r = 32$, $\alpha = 64$, to all attention and feed-forward network layers) to Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct, training with the learning rate 2×10^{-4}, batch size 4, for three epochs. For SGVR, we set $K = 3$ as a practical choice balancing computational cost and performance. For SAV in zero-shot evaluation, we set $K = 5$ based on preliminary experiments (see Appendix J). All experiments were run with three inference iterations, and we report averaged values.

Table 2: In-domain evaluation results (overall).

Method	Accuracy	Macro-F1
mBERT-HSLN	0.919	0.812
LLM-SSC (Gemma)	0.666	0.567
LLM-SSC (Llama)	0.633	0.545
LLM-SSC (Qwen)	0.764	0.694
SIP (Llama)	0.836	0.714
SIP (Qwen)	0.919	0.841
SIP+SGVR (Llama)	0.843	0.746
SIP+SGVR (Qwen)	**0.923**	**0.848**

Table 3: Per-language results for major languages.

Language	mBERT	SIP+SGVR	Δ
English	0.863	0.875	+1.2
Japanese	0.719	0.796	+7.7
Chinese	0.780	0.788	+0.8
Spanish	0.793	0.882	+8.9
Average	**0.789**	**0.835**	**+4.6**

※ All results are Macro-F1 scores.

We used the following baselines: (1) **multilingual LLM-SSC:** a multilingual extension of Lan et al. (2024)'s LLM-SSC with identical LoRA settings and (2) **mBERT-HSLN:** Brack et al. (2024)'s HSLN with mBERT encoder. We report sentence-level accuracy and Macro-F1 across five classes as evaluation metrics.

6.2 In-Domain Evaluation

Table 2 shows the results aggregated across test sets. SIP+SGVR (Qwen) demonstrates improvements of +0.46 in accuracy and +3.61 in Macro-F1 compared to mBERT-HSLN, thereby showing that our method achieved performance comparable to or exceeding strong BERT-based baselines. The use of SIP and SGVR achieved substantially higher performance than LLM-SSC across all three LLMs (Gemma, Llama, Qwen).

Table 3 shows the results for the four major languages in the dataset. Our method consistently outperformed baselines across all languages, with particularly strong improvements in Japanese (+7.7 Macro-F1) and Spanish (+8.9 Macro-F1), while maintaining competitive performance in English and Chinese. Complete per-language results are provided in Appendix G.

We then analyzed the component contributions. To verify SIP's effect, we conducted evaluations

Table 4: Zero-shot evaluation results (overall).

Method	Accuracy	Macro-F1
mBERT-HSLN	0.751	0.658
SIP+SAV (Qwen)	**0.849**	**0.818**

using SGVR with a baseline prompt without structural information. Compared to SIP+SGVR (Qwen), the baseline prompt resulted in decreases of -0.80 in accuracy and -3.20 in Macro-F1 (Appendix H), demonstrating importance of SIP in classification accuracy. SIP+SGVR (Qwen) shows improvements of +0.39 in accuracy and +0.72 in Macro-F1 compared to SIP (Qwen), confirming SGVR's contribution. Finally, Qwen consistently outperformed Llama, suggesting Qwen2.5-3B has superior multilingual processing capability.

6.3 Zero-Shot Cross-Lingual Evaluation

In the next step, we conducted evaluations on models trained on nine languages and evaluated them on five languages not in training. Since Qwen consistently outperformed Llama in in-domain evaluation, we used only Qwen. In this setting, validation data for the target languages were unavailable, making verifier training impossible. Therefore, we used SAV, which selects predictions based on candidate consistency. SIP+SAV (Qwen) demonstrated improvements of +9.84 in accuracy and +15.93 in Macro-F1 compared to mBERT-HSLN (Table 4). The larger improvement in Macro-F1 may indicate more stable performance for minority classes. SAV showed improvements of +0.29 in accuracy and +0.50 in Macro-F1 compared to random selection (Appendix I).

7 Limitations

We acknowledge several limitations to our study.

Domain bias: Since our dataset primarily comprised abstracts with explicit section headers (prevalent in medicine/life sciences), it may bias toward these fields. Effectiveness in domains with different rhetorical organizations (e.g., humanities, social sciences) remains to be validated.

Language coverage: Although covering 13 languages, our dataset lacks representation from major families like African (e.g., Swahili), Indic (e.g., Hindi), and Southeast Asian (e.g., Thai). Observed structural patterns may differ for these underrepresented families.

Limited hyperparameter exploration: We set $K = 3$ for SGVR and $K = 5$ for SAV based on preliminary tests without exhaustive search. Optimal K likely varies by language, domain, and compute budget. Furthermore, multiple candidate generation increases inference cost by a factor of K.

Zero-shot evaluation scope: Zero-shot evaluation was limited to five languages with small test sets (7–179 abstracts). A larger, more diverse evaluation would further strengthen our findings.

Structural dependency: Our approach assumes consistent structural patterns across languages. However, conventions may vary by venue or community, potentially affecting the transferability of our structural features.

8 Ethical Considerations

Data collection: Our dataset is derived from publicly available academic abstracts (DOAJ, HAL) and databases permitting academic use (Dialnet, TRdizin, CiNii Research).

Bias and representation: The dataset is biased toward medicine/life sciences and 13 primarily European/East Asian languages. Models may underperform on languages or domains with different structural conventions.

9 Conclusion

We constructed a multilingual SSC dataset of 13 languages and analyzed cross-lingual transfer. Our key finding is that structural similarity correlates significantly more strongly with transfer performance ($r = 0.443$ for Qwen2.5-3B) than linguistic proximity ($r = 0.197$), suggesting models rely on rhetorical structure rather than linguistic features.

We proposed SIP, SGVR, and SAV to leverage this structural information. SIP+SGVR achieves performance comparable to or exceeding strong baselines (+3.61 Macro-F1 over mBERT-HSLN), while SIP+SAV shows substantial zero-shot improvements over these same baselines (+15.93 Macro-F1).

Our study highlights the importance of task-specific structural characteristics over traditional linguistic proximity. Future work includes analyzing internal representations to understand how structural patterns are encoded and extending our approach to other discourse-level tasks.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

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*, 25:377–400.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Jan Buchmann, Max Eichler, Jan-Micha Bodensohn, Ilia Kuznetsov, and Iryna Gurevych. 2024. Document structure in long document transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1056–1073, St. Julian's, Malta. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

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.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, Inc.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online. Association for Computational Linguistics.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, Andre Martins, and Hinrich Schuetze. 2024. mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, St. Julian's, Malta. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pedro Martín-Martín. 2003. A genre analysis of english and spanish research paper abstracts in experimental social sciences. *English for Specific Purposes*, 22(1):25–43.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. 2024. Grammar-aligned decoding. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Nakatani Shuyo. 2010. Language detection library for Java.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: Applications and future perspectives. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 201–235. Springer, Cham.

Yang Sun, Guanrong Chen, Caihua Yang, Jianzhu Bao, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. 2024. Discourse structure-aware prefix for generation-based end-to-end argumentation mining. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11597–11613, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Zhuoer Wang, Leonardo F. R. Ribeiro, Alexandros Papangelis, Rohan Mukherjee, Tzu-Yen Wang, Xinyan Zhao, Arijit Biswas, James Caverlee, and Angeliki Metallinou. 2024. FANTAstic SEquences and where to find them: Faithful and efficient API call generation through state-tracked constrained decoding and reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6179–6191, Miami, Florida, USA. Association for Computational Linguistics.

Taejun Yun, Jinhyeon Kim, Deokyeong Kang, Seonghoon Lim, Jihoon Kim, and Taeuk Kim. 2023. X-SNS: Cross-lingual transfer prediction through sub-network similarity. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13131–13144, Singapore. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, 54(2):587–613.

A Data Collection Details

A.1 Database Selection and Access

We selected five academic databases based on the following criteria: (1) providing API access to abstracts, (2) containing substantial non-English content, and (3) including structured abstracts with section headers.

DOAJ (Directory of Open Access Journals): Provides access to open access journals across multiple languages. We queried abstracts through their REST API.

HAL (Hyper Articles en Ligne): French national open archive containing predominantly

French-language publications. We accessed abstracts through their OAI-PMH interface.

Dialnet: Spanish bibliographic database containing both Spanish and Portuguese publications. We obtained abstracts through institutional access.

TRdizin: Turkish academic index containing Turkish publications. We accessed abstracts through their search API.

CiNii (Citation Information by NII): Japanese academic database containing Japanese publications. We accessed abstracts through their REST API.

A.2 Search Query Design

Since structured abstracts typically contain Method and Result sections, we translated these two terms into each target language and used them as search queries. We used native terms in which Method and Result in each target language (e.g., "Méthodes" and "Résultats" in French, "Metodi" and "Risultati" in Italian).

A.3 Preprocessing Pipeline

We applied the following preprocessing steps to the collected abstracts:

1. HTML entity conversion: Converted HTML entities (e.g., &, <, >) to their corresponding characters.

2. Unicode normalization: Applied NFKC normalization to ensure consistent character representations.

3. Language detection: Used langdetect library (Shuyo, 2010) to verify that abstracts were in the target language. We discarded abstracts where the detected language did not match the target.

4. Duplicate removal: Removed duplicate abstracts based on title matching (case-insensitive, after removing punctuation).

5. Structure identification: Used pattern matching to identify structured abstracts in three formats:

- XML format: `<sec><title>Methods</title>...</sec>`
- Bracket format: `[Methods] ...`
- Colon format: `Methods: ...`

6. Section extraction: We retained only abstracts containing two or more sections to ensure sufficient structure and sentence segmentation.

7. Sentence segmentation: Used NLTK's punkt tokenizer for European languages (English, French, Spanish, Portuguese, Italian, Russian, Polish, Dutch, Estonian) and spaCy for Asian languages (Japanese, Chinese, Korean, Indonesian, Turkish).

B Similarity Measure Definitions

We designed six types of cross-lingual structural similarity measures based on the rhetorical patterns in abstracts:

B.1 Label Distribution Similarity

We calculated the probability distribution of the five labels (Background, Objective, Method, Result, Concusion) in each language L:

$$P_L(l) = \frac{\text{count of the following label } l \text{ in language } L}{\text{total sentences in language } L}$$

The distance between two languages L_1 and L_2 was computed using Jensen-Shannon divergence:

$$d_{\text{label}}(L_1, L_2) = \text{JSD}(P_{L_1}, P_{L_2})$$

B.2 Section Length Distribution Similarity

We defined a section as "a continuous span with the same label" and calculated the distribution of section lengths (number of consecutive sentences) for each label. For each label l and length k, we computed the following:

$$P_{L,l}(k) = \frac{\text{count of sections with } l \text{ and } k}{\text{total sections with } l}$$

The distance was obtained by calculating JSD for each label and averaging:

$$d_{\text{section}}(L_1, L_2) = \frac{1}{5} \sum_{l} \text{JSD}(P_{L_1,l}, P_{L_2,l})$$

B.3 Continuation Probability Similarity

We calculated the conditional probability that a sentence's label remains the same in the next sentence (continuation probability) for each label:

$$p_L(l) = P(y_{i+1} = l | y_i = l)$$

The distance between languages was obtained by normalizing the Euclidean distance between five-dimensional continuation probability vectors:

$$d_{\text{cont}}(L_1, L_2) = \frac{1}{\sqrt{5}} \sqrt{\sum_{l} (p_{L_1}(l) - p_{L_2}(l))^2}$$

B.4 Boundary Position Similarity

For four major transitions (Background → Objective, Objective → Method, Method → Result, Result → Concusion), we calculated the average relative position (normalized to 0–1) within abstracts where these transitions occurred. For transition t, we computed the following:

$$\mu_L(t) = \text{average position of } t \text{ in language } L$$

The distance was defined as follows:

$$d_{\text{bound}}(L_1, L_2) = \sum_t w_t \cdot |\mu_{L_1}(t) - \mu_{L_2}(t)|$$

where $w_t = \min(\text{freq}_{L_1}(t), \text{freq}_{L_2}(t))$ and the sum is normalized.

B.5 Block Count Distribution Similarity

We defined block count as "the number of continuous spans where each label appears within one abstract". For example, in Method → Result → Method, the block count for Method is 2. We calculated the distribution of block counts for each label and computed JSD-based distance similar to section length distribution.

B.6 Transition Entropy Similarity

We calculated Shannon entropy from the probability distribution of all label transitions (25 patterns):

$$H_L = -\sum_{l_1, l_2} P_L(l_2|l_1) \log_2 P_L(l_2|l_1)$$

The distance was obtained by normalizing the absolute difference:

$$d_{\text{entropy}}(L_1, L_2) = \frac{|H_{L_1} - H_{L_2}|}{\log_2 25}$$

B.7 Overall Structural Similarity

We calculated the overall distance d as the simple average of the above six distances:

$$d(L_1, L_2) = \frac{1}{6} \sum_{i=1}^{6} d_i(L_1, L_2)$$

Structural similarity is defined as:

$$\text{StructSim}(L_1, L_2) = 1 - d(L_1, L_2)$$

C Prompt Template for Cross-Lingual Transfer Experiments

For the cross-lingual transfer experiments in Section 4, we used the following prompt template:

> **Instruction:** 'You must categorize the given sentence into one of these five labels: BACKGROUND, OBJECTIVE, METHOD, RESULT, CONCLUSION. Respond with ONLY the label name.' **Output:** 'Question: What is the rhetorical role of the Target Sentence? Answer with one word from the labels list.'

This simplified format was adapted from Lan et al. (2024), removing the demonstration examples to focus on zero-shot transfer capabilities. We unified prompts in English for all experiments in Section 4.

D Cross-Lingual Transfer Results for All Models

In this section, we provide the comprehensive cross-lingual transfer matrices for all four models evaluated in our study: mBERT-HSLN, Qwen2.5-3B-Instruct, Gemma2-2B-it, and Llama-3.2-3B-Instruct. Tables 5 through 8 show the Macro-F1 scores for every source-target language pair among the nine languages used for training.

E Prompt Template for SIP

The following template illustrates the structure and instructions used for the Structure-Informed Prompting (SIP) experiments. The prompt includes the task description, explicit rhetorical constraints, and a placeholder for a few-shot demonstration.

> **Instruction:** You must categorize the given sentence into one of these five labels: Background, Objective, Method, Result, Concusion. Respond with ONLY the label name.
> **Structural Constraints:** Academic abstracts typically follow this order: Background (introducing the topic) → Objective (stating the research goal) → Method (describing the approach) → Result (presenting findings) → Concusion (summarizing implications).
> **Example:** [One demonstration example from training data]
> **Output:** Question: What is the rhetorical role of the Target Sentence? Answer with one word from the labels list.

F Structural Features for SGVR

We summarize the 44-dimensional structural features used for training the Structure-Guided Verifier Reranking (SGVR) component. These features

Table 5: Cross-lingual transfer Results for mBERT-HSLN (Macro-F1).

Train\Eval	zh	es	en	fr	id	it	ja	pt	ru
zh	0.319	0.229	0.269	0.284	0.233	0.194	0.226	0.246	0.285
es	0.323	0.434	0.608	0.408	0.459	0.304	0.363	0.326	0.298
en	0.497	0.488	0.651	0.560	0.602	0.443	0.505	0.518	0.454
fr	0.297	0.248	0.264	0.335	0.235	0.211	0.233	0.259	0.264
id	0.476	0.509	0.640	0.533	0.691	0.506	0.496	0.542	0.431
it	0.318	0.369	0.440	0.355	0.453	0.439	0.309	0.345	0.301
ja	0.634	0.426	0.579	0.401	0.511	0.405	0.651	0.472	0.544
pt	0.632	0.480	0.566	0.497	0.501	0.415	0.555	0.577	0.573
ru	0.303	0.235	0.286	0.283	0.248	0.217	0.245	0.239	0.295

Table 6: Cross-lingual transfer results for Qwen2.5-3B-Instruct (Macro-F1).

Train\Eval	zh	es	en	fr	id	it	ja	pt	ru
zh	0.681	0.587	0.625	0.538	0.655	0.552	0.451	0.633	0.605
es	0.410	0.598	0.433	0.468	0.544	0.543	0.391	0.622	0.433
en	0.564	0.605	0.648	0.565	0.678	0.595	0.582	0.592	0.499
fr	0.464	0.449	0.355	0.478	0.622	0.428	0.427	0.456	0.383
id	0.448	0.498	0.597	0.427	0.557	0.484	0.373	0.473	0.412
it	0.521	0.605	0.643	0.566	0.643	0.602	0.517	0.652	0.570
ja	0.642	0.567	0.602	0.497	0.676	0.541	0.657	0.632	0.596
pt	0.546	0.554	0.488	0.375	0.643	0.509	0.502	0.668	0.487
ru	0.518	0.439	0.544	0.431	0.506	0.403	0.457	0.457	0.608

were designed to capture the logical flow and label distribution patterns of scientific abstracts, as detailed in Table 9.

G Per-Language Results for In-Domain Evaluation

The detailed per-language performance for the in-domain evaluation setting is presented below. In the following tables, "Random" indicates random selection of one candidate from generated candidates, "Verifier" (or "SAV") indicates selection by our proposed method, and "Oracle" indicates selection of the candidate with highest agreement with the gold label sequence (upper bound of performance).

H SIP Ablation Study

To verify the effect of SIP, we conducted experiments using SGVR with a baseline prompt that does not incorporate structural information. Performance degradation was observed across all languages when removing SIP. The degradation is particularly pronounced in Macro-F1, suggesting that SIP contributes to improved prediction accuracy for minority classes. Detailed ablation results for the combined test set are shown in Table 11.

I Per-Language Results for Zero-Shot Evaluation

Table 12 presents the detailed performance metrics for the zero-shot transfer evaluation on five unseen languages: Estonian, Korean, Dutch, Polish, and Turkish. We report accuracy and Macro-F1 scores for each selection strategy to demonstrate the robustness of the SAV method in cases where training data for the target language is unavailable.

J Additional Details on SAV

The selection of $K = 5$ for SAV in zero-shot evaluation was based on balancing candidate diversity with computational cost. Preliminary experiments showed that performance gains plateaued beyond $K = 5$, while inference time increased linearly with K.

Table 7: Cross-lingual transfer results for Gemma2-2B-it (Macro-F1).

Train\Eval	zh	es	en	fr	id	it	ja	pt	ru
zh	0.737	0.561	0.579	0.487	0.563	0.430	0.552	0.367	0.575
es	0.316	0.531	0.297	0.350	0.377	0.410	0.276	0.353	0.406
en	0.365	0.487	0.541	0.428	0.526	0.448	0.393	0.304	0.403
fr	0.261	0.303	0.264	0.372	0.364	0.324	0.368	0.276	0.283
id	0.302	0.327	0.439	0.366	0.439	0.402	0.324	0.260	0.358
it	0.325	0.378	0.267	0.299	0.367	0.479	0.292	0.319	0.269
ja	0.669	0.562	0.530	0.502	0.650	0.464	0.658	0.374	0.543
pt	0.361	0.468	0.376	0.455	0.457	0.423	0.361	0.469	0.310
ru	0.463	0.477	0.535	0.460	0.426	0.384	0.422	0.321	0.583

Table 8: Cross-lingual transfer results for Llama-3.2-3B-Instruct (Macro-F1).

Train\Eval	zh	es	en	fr	id	it	ja	pt	ru
zh	0.415	0.266	0.450	0.252	0.283	0.231	0.418	0.293	0.386
es	0.220	0.327	0.497	0.327	0.437	0.331	0.217	0.341	0.278
en	0.441	0.406	0.641	0.515	0.546	0.419	0.459	0.489	0.483
fr	0.339	0.306	0.414	0.438	0.385	0.337	0.319	0.321	0.370
id	0.438	0.576	0.684	0.547	0.776	0.509	0.481	0.609	0.474
it	0.257	0.227	0.488	0.297	0.330	0.298	0.284	0.216	0.301
ja	0.452	0.366	0.438	0.409	0.414	0.294	0.449	0.382	0.396
pt	0.355	0.473	0.497	0.425	0.567	0.408	0.336	0.520	0.469
ru	0.410	0.325	0.465	0.414	0.343	0.342	0.396	0.378	0.512

Table 9: 44-dimensional structural features for SGVR.

Group	**Feature**	**Dim**
Basic scores	Transition, position, distribution scores	3
Sequence	Length, label occurrence counts	6
Sections	Mean, std of section length per label	10
Continuation	Continuation probability per label	5
Blocks	Number of blocks per label	5
Transitions	Presence of major transitions	4
Entropy	Transition entropy	1
Boundary	Start and end labels (one-hot)	10
Total		**44**

Table 10: SIP+SGVR (Qwen) per-language results.

Language	**Random**	**Verifier**	**Oracle**
English	0.931 / 0.863	0.934 / 0.875	0.941 / 0.885
French	0.942 / 0.832	0.943 / 0.838	0.955 / 0.860
Japanese	0.858 / 0.787	0.866 / 0.796	0.888 / 0.832
Spanish	0.890 / 0.872	0.899 / 0.882	0.919 / 0.907
Chinese	0.984 / 0.787	0.985 / 0.788	0.988 / 0.790
Indonesian	0.958 / 0.953	0.960 / 0.955	0.966 / 0.962
Portuguese	0.883 / 0.861	0.881 / 0.862	0.904 / 0.889
Italian	0.866 / 0.829	0.871 / 0.840	0.892 / 0.867
Russian	0.934 / 0.747	0.940 / 0.751	0.956 / 0.765
Overall	**0.919 / 0.841**	**0.923 / 0.848**	**0.934 / 0.862**

※ Each cell shows accuracy / Macro-F1.

Table 11: Ablation results (overall, Qwen).

Method	**Accuracy**	**Macro-F1**
SGVR w/o SIP	0.915	0.816
SIP+SGVR	0.923	0.848
Δ	+0.80	+3.20

Table 12: SIP+SAV (Qwen) per-language results.

Language	**Random**	**SAV**	**Oracle**
Estonian	0.927 / 0.713	0.919 / 0.709	0.943 / 0.735
Korean	0.967 / 0.778	0.981 / 0.787	0.992 / 0.793
Dutch	0.855 / 0.833	0.863 / 0.838	0.916 / 0.901
Polish	0.892 / 0.881	0.894 / 0.887	0.924 / 0.921
Turkish	0.710 / 0.621	0.705 / 0.608	0.806 / 0.721
Overall	**0.846 / 0.813**	**0.849 / 0.818**	**0.901 / 0.884**

※ Each cell shows accuracy / Macro-F1.