Conformal Prediction Sets for Reliable Uncertainty Quantification in Natural Language Processing

Rui Luo RUILUO@CITYU.EDU.HK
Department of Systems Engineering
City University of Hong Kong

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Natural language processing (NLP) classification tasks often benefit from predicting a set of possible labels with confidence scores to capture uncertainty. However, existing methods struggle with the high-dimensional and sparse nature of textual data. We propose a novel conformal prediction method designed for NLP that utilizes confidence scores from deep learning models to construct prediction sets. Our approach achieves the coverage rate while managing the size of the prediction sets. Through theoretical analysis and extensive experiments, we demonstrate that our method outperforms existing techniques on various datasets, providing reliable uncertainty quantification for NLP classifiers. We contribute a novel conformal prediction method, theoretical analysis, and empirical evaluation. Our work advances the practical deployment of NLP systems by enabling reliable uncertainty quantification.

Keywords: Conformal Prediction; Natural Language Processing; Uncertainty Quantification; Ranking; Large Language Models

1. Introduction

Natural language processing (NLP) covers a wide range of classification tasks, including topic modeling, sentiment analysis and named entity recognition. These problems have been traditionally approached using discriminative classifiers such as logistic regression, support vector machines and deep neural networks (Goldberg, 2017; Devlin et al., 2019). While these methods have achieved good performance, they typically only output a single predicted class label for each input. However, in many applications, it may be useful to predict a set of possible labels along with confidence scores, to capture uncertainty and allow for multiple acceptable answers.

The importance of reliable uncertainty quantification in NLP has become increasingly evident with the rapid proliferation of large language models (LLMs) in real-world applications (Min et al., 2022). LLMs are prone to issues such as hallucinations (Ji et al., 2023), poor calibration (Desai and Durrett, 2020; Kong et al., 2020), and biases (Gallegos et al., 2023; Guo et al., 2022). Conformal prediction (CP) (Vovk et al., 2005) offers a principled and model-agnostic way to address these challenges by providing theoretically sound coverage guarantees with minimal assumptions Angelopoulos and Bates (2021). As outlined in the comprehensive survey by Campos et al. (2024), CP can be combined with any underlying machine learning model to construct prediction sets—a set of class labels guaranteed to contain the true label with a specified probability—thereby measuring uncertainty and expressing ambiguity in model predictions.

© 2025 R. Luo.

In this work, we propose a novel conformal prediction method designed specifically for natural language classification tasks. Our approach addresses challenges inherent to NLP, such as the high-dimensional and sparse nature of textual data. The proposed method enables predicting sets of possible labels for instances where the model is uncertain, as well as abstaining from making predictions when the model has low confidence.

Our approach differs from existing adaptive prediction set algorithms, such as APS (Romano et al., 2020), as it does not depend on strong assumptions about the probabilities generated by NLP deep learning models. Instead, our method utilizes the confidence scores from the deep learning model and constructs prediction sets by establishing a confidence threshold. This enables us to achieve high coverage while managing the size of the prediction sets. In comparison to other multiclass prediction set methods that rely on probability level sets, like the work by Sadinle et al. (2019), our approach is more streamlined, efficient, and makes fewer assumptions about the probabilities. In Section 2, we provide a comprehensive overview of existing methods, followed by a detailed description of our proposed approach in Section 3.

Through extensive experiments on text classification benchmarks spanning topic categorization, sentiment analysis, and natural language inference, we demonstrate the validity and efficiency of our proposed method. The results show that our approach substantially outperforms existing conformal and non-conformal techniques, achieving the specified coverage level with prediction sets that are often much smaller.

Our key contributions are:

1. A novel conformal prediction method for NLP that combines ideas from existing works.
2. Theoretical analysis on the coverage guarantee of our method.
3. Empirical evaluation to date of conformal prediction for deep learning and LLM-based text classification.

Our work applies the Conformal Prediction framework to Natural Language Processing tasks, enabling deep learning classifiers to provide rigorous uncertainty quantification via prediction sets. This application marks an important step toward the reliable deployment of NLP systems. A full version detailing the theoretical underpinnings and methodology of this approach is presented in Luo and Zhou (2025e).

Notations: For a sequence $a_1, a_2, \ldots, a_n$, the order statistics are denoted by $a_{(1)} \geq a_{(2)} \geq \cdots \geq a_{(n)}$ in a descending order. $[n]$ represents the set $\{1, 2, \ldots, n\}$. For a set $\mathcal{S}$, $|\mathcal{S}|$ is the size (i.e., cardinality) of the set.

2. Related Work

Conformal prediction (CP) has been extensively applied to various tasks, including classification Luo and Colombo (2024) and regression Luo and Zhou (2025f); Bao et al. (2025a); Luo and Zhou (2025d). Efficiency is a key focus, seen in methods like Conformity Score Averaging Luo and Zhou (2025c), which enhances efficiency by optimally combining multiple score functions. CP also extends to specialized domains like segmentation Luo and Zhou (2025a), games Luo et al. (2024); Bao et al. (2025b), and graph-based applications Luo et al. (2023); Luo and Zhou (2025b); Luo and Colombo (2025); Tang et al. (2025); Wang et al. (2025); Zhang et al. (2025). In classification

tasks, CP typically follows a two-step framework: first training a predictive model to estimate class probabilities, then constructing prediction sets with guaranteed coverage using conformity scores from a calibration set. Two main approaches have emerged for defining these prediction sets: the individual threshold approach, exemplified by Threshold Conformal Prediction (THR) Sadinle et al. (2019), which includes labels exceeding a calibrated probability threshold; and the cumulative threshold approach, represented by Adaptive Prediction Set (APS) Romano et al. (2020) and its extensions like RAPS Angelopoulos et al. (2021) and SAPS Huang et al. (2024), which include the top-ranked labels until their cumulative probability mass reaches a calibrated threshold.

CP Variants and Applications in NLP Several variants of conformal prediction have been developed to handle different settings, such as Mondrian conformal prediction for category-wise validity (Vovk et al., 2005), methods for non-exchangeable data (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021; Gibbs and Candes, 2021), and conformal risk control for multilabel classification (Angelopoulos et al., 2025). Applications of conformal prediction to diverse NLP tasks are growing, including sequence labeling (Fisch et al., 2022; Jiang et al., 2023), machine translation (Fomicheva et al., 2020), question answering (Feng et al., 2022), and natural language generation (Lu et al., 2022). However, challenges remain in scaling conformal prediction to large datasets and models, handling distribution shift, and extending coverage guarantees (Angelopoulos and Bates, 2021; Baan et al., 2023).

Traditional Uncertainty Quantification Existing uncertainty quantification techniques for deep learning models, such as Platt's scaling (Platt, 1999), isotonic regression (Niculescu-Mizil and Caruana, 2005), spline-based probability calibration (Lucena, 2019), and temperature scaling (Guo et al., 2017) are used for post ad hoc calibration to adjust classifier confidence levels. In addition, Rahimi et al. (2020) attempt to calibrate the predicted probabilities from deep learning models while preserving the rank order of the probabilities. However, these techniques do not have formal guarantees and the resulting probabilities are often not well-calibrated, especially for unlikely classes, highlighting the advantage of CP's rigorous guarantees.

Conformal Prediction for Large Language Models Recent advancements in conformal prediction for LLMs have shown promising results in quantifying uncertainty and providing statistical performance guarantees. Quach et al. (2023) propose a novel conformal prediction approach for generative language models that produces prediction sets with rigorous coverage guarantees by calibrating stopping and rejection rules to generate high-quality response sets. Similarly, Kumar et al. (2023) explore how conformal prediction can quantify uncertainty in language models for multiple-choice question answering, demonstrating a strong correlation between uncertainty estimates and prediction accuracy, and investigating the importance of the exchangeability assumption. Deutschmann et al. (2024) introduce two extensions to the beam search algorithm based on conformal predictions to generate sequence sets with theoretical coverage guarantees, empirically evaluating their methods on natural language processing and chemistry tasks. Su et al. (2024) address the challenge of quantifying uncertainty in language models without logit access by introducing a conformal prediction method that utilizes coarse-grained and fine-grained uncertainty concepts to produce minimal prediction sets with statistical coverage guarantees.

Algorithm 1 Rank-based conformal prediction

1: **Input:** data $\{(x_i, y_i)\}_{i\in\mathcal{I}}$, a test sample x_{n+1}, black-box learning algorithm $\mathcal{B}$, level $\alpha \in (0, 1)$.
2: Randomly split the indices $\mathcal{I}$ into two subsets $\mathcal{I}_1, \mathcal{I}_2$.
3: Train $\mathcal{B}$ on all samples in $\mathcal{I}_1 : \widehat{\pi} \leftarrow \mathcal{B}(\{(x_i, y_i) : i \in \mathcal{I}_1\})$.
4: $n \leftarrow |\mathcal{I}_2|$. Find the $\lfloor (n+1)\alpha \rfloor$-th largest value in (1), denoted by r^*_α.
5: Find the proportion p in (2).
6: Find the $\lceil np \rceil$-th largest value in (3), denoted by π^*.
7: With r^*_α and π^* obtained from Step 4 and Step 6, use the function $\widehat{C}(x_{n+1})$ in equation 4 to construct the prediction set for x_{n+1}.
8: **Output:** A prediction set $\widehat{C}_\alpha(x_{n+1})$.

3. Our Approach

Our approach directly focuses the goal of minimizing the size of the prediction set. In the calibration set, we evaluate the size of the prediction set required to include the true label. We operate under the assumption that a higher value of $\widehat{\pi}_k(x_i)$ indicates a greater likelihood of x_i belonging to class k. Consequently, we impose a constraint on the confidence interval: if class k is included in the prediction set, then any class k' such that $\widehat{\pi}_{k'}(x_i) > \widehat{\pi}_k(x_i)$ must also be included in the prediction set.

Given this constraint, the smallest prediction set that includes the true label will be determined by the rank of $\widehat{\pi}_{y_i}(x_i)$ within the sequence $\{\widehat{\pi}_1(x_i), \ldots, \widehat{\pi}_K(x_i)\}$. However, it is common to encounter multiple samples in the calibration set that have the same prediction set size. In such cases, we need to establish a preference for breaking ties.

Intuitively, our tie-breaking approach aims to efficiently cover the true label by favoring the option with the larger predicted probability. When comparing the kth most likely label of x_i and x_j, given the predicted probabilities $\widehat{\pi}$, we prioritize the inclusion of labels that are more confidently predicted by the model. This choice aligns with our goal of constructing prediction sets that are more likely to contain the true label while maintaining a smaller overall size compared to randomly breaking ties.

We will rigorously summarize the idea above. Our goal is to determine a rank k such that for the test sample, we include either the top k or $k-1$ labels in the prediction set. We also need to establish a rule to choose between k and $k-1$. These rules will be determined using the calibration set. The method to determine k is straightforward. Let $\mathcal{I}_2$ be the calibration set and $n = |\mathcal{I}_2|$. For each $i \in \mathcal{I}_2$, we define the following rank:

$$r_i = \text{rank of } \pi_{y_i}(x_i) \text{ in } \{\pi_k(x_i) : k \in [K]\}. \tag{1}$$

We then find the order statistics of these ranks: $r_{(1)} \geq r_{(2)} \geq \cdots \geq r_{(n)}$ and let $r^*_\alpha = r_{(\lfloor (n+1)\alpha \rfloor)}$. This ensures that:

$$\begin{aligned} &|\{i \in \mathcal{I}_2 : r_i \leq r^*_\alpha - 1\}| < \lfloor (n+1)\alpha \rfloor \\ &\leq |\{i \in \mathcal{I}_2 : r_i \leq r^*_\alpha\}|. \end{aligned}$$

To construct the prediction set for x_{n+1}, we will include either the top-$(r^*_\alpha - 1)$ or top-r^*_α classes based on the values of $\pi_1(x_{n+1}), \ldots, \pi_K(x_{n+1})$. The top-$r^*_\alpha$ classes refer to the classes corresponding

to the r^*_α largest values among $\pi_1(x_{n+1}), \ldots, \pi_K(x_{n+1})$. To achieve $1-\alpha$ coverage, we need to determine when to include the r^*_α-th class and when not to. We start by calculating the proportion p of instances for which we should include the r^*_α-th label:

$$p := \frac{n - \lfloor (n+1)\alpha \rfloor - |\{i \in \mathcal{I}_2 : r_i \leq r^*_\alpha - 1\}|}{|\{i \in \mathcal{I}_2 : r_i = r^*_\alpha\}|}. \tag{2}$$

Roughly speaking, the numerator of this fraction represents the difference between the number of samples obtained by selecting all samples with rank $r_i \leq r^* - 1$ and the number of samples needed to achieve a coverage of $n(1-\alpha)$ in the calibration set. Next, we find the $\lceil np \rceil$-th largest value, denoted as π^*, in the set of $\pi_{(r^*_\alpha)}(x_i)$'s:

$$\pi^* = \lceil np \rceil\text{-th largest value in} \{\widehat{\pi}_{(r^*_\alpha)}(x_i) : i \in \mathcal{I}_2\}, \tag{3}$$

where $\widehat{\pi}_{(k)}(x_i)$ denotes the k-th order statistics in $(\pi_1(x_i), \ldots, \pi_n(x_i))$. Finally, for the test sample x_{n+1}, if $\widehat{\pi}_{(r^*_\alpha)}(x_{n+1}) \geq \pi^*$, then the r^*_α-th label will be included in the prediction set. Otherwise, it will not be included. To summarize rigorously, for a test sample x_{n+1},

$$\widehat{C}_\alpha(x_{n+1}) = \begin{cases} \{k : \widehat{\pi}_k(x_{n+1}) \geq \widehat{\pi}_{(r^*_\alpha)}(x_{n+1})\}, \\ \text{if } \widehat{\pi}_{(r^*_\alpha)}(x_{n+1}) \geq \pi^*; \\ \{k : \widehat{\pi}_k(x_{n+1}) \geq \widehat{\pi}_{(r^*_\alpha - 1)}(x_{n+1})\}, \\ \text{otherwise.} \end{cases} \tag{4}$$

This definition implies the following proposition.

Proposition 1 *The output $\widehat{C}_\alpha(x_{n+1})$ from Algorithm 1 satisfies $\widehat{C}_\alpha(x_{n+1}) \subset \{\widehat{y}_{(1)}, \ldots, \widehat{y}_{(r^*_\alpha)}\}$, i.e., the subset of labels that have top-r^*_α values in $\{\widehat{\pi}_1(x_{n+1}), \ldots, \widehat{\pi}_K(x_{n+1})\}$.*

We note that we can define the following conformity score for our method. This will help us understand why $1-\alpha$ coverage is guaranteed.

$$\begin{aligned} c_i &= c(x_i, y_i) \\ &= [\text{rank of } \widehat{\pi}_{y_i}(x_i) \text{ in } \{\widehat{\pi}_1(x_i), \ldots, \widehat{\pi}_K(x_i)\}] - 1 \\ &\quad + \frac{1}{n}[\text{rank of } \widehat{\pi}_{y_i}(x_i) \text{ in } \{\widehat{\pi}_{y_i}(x_1), \ldots, \widehat{\pi}_{y_i}(x_n)\}]. \end{aligned}$$

Let's define the quantile $\widehat{Q}_\alpha$ as the $\lfloor (n+1)\alpha \rfloor$-th largest value among the conformity scores $c_1, c_2, \ldots, c_n$ calculated on the calibration set. To construct a prediction set with $1-\alpha$ coverage, we include all samples from the calibration set whose conformity scores c_i are less than or equal to the quantile $\widehat{Q}_\alpha$. In other words, the procedure for defining the prediction set is equivalent to selecting the calibration samples that satisfy the condition $c_i \leq \widehat{Q}_\alpha$, which ensures the desired coverage level of $1-\alpha$.

Example: Let us consider the following example to understand our method better. After applying a training algorithm to the training samples with $K = 10$ classes, we obtain the function $\widehat{\pi}$. By applying $\widehat{\pi}$ to the calibration samples $\{(x_i, y_i)\}_{i \in \mathcal{I}_2}$, we obtain the ranks of $\widehat{\pi}_{y_i}(x_i)$ for $i \in \mathcal{I}_2$, which represent the ranks of the true class. This forms an empirical distribution on the set $\{1, 2, \ldots, 10\}$.

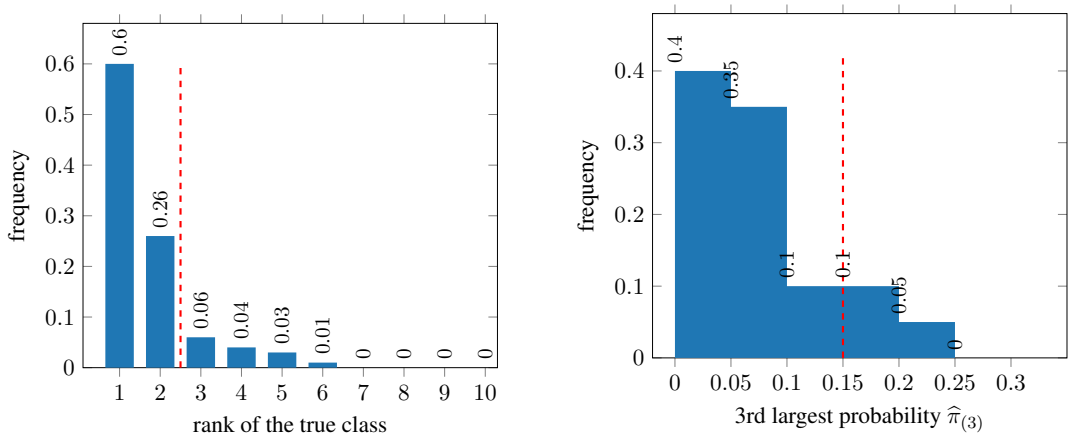

Figure 1: The figure illustrates the construction of a 90% prediction set for a test sample with sorted probability vector $[0.55, 0.2, 0.15, 0.1, 0, 0, 0, 0, 0, 0]$. The top three classes are included based on both the probability of ranks (**Top**) and the distribution of the 3rd largest probabilities in the calibration set (**Bottom**).

Let us consider an example in Figure 1. $(0.6, 0.26, 0.06, 0.04, 0.03, 0.01, 0, 0, 0, 0)$ is the empirical distribution mentioned above. Suppose we want to construct the prediction set $\widehat{C}_\alpha$ with $\alpha = 0.1$, i.e., coverage probability equals to 0.9. The rank distribution in the empirical distribution shows that the top two classes have a cumulative probability of 0.86, which falls short of the desired coverage of 0.9. Including the top three classes increases the cumulative probability to 0.92, exceeding the desired coverage. Therefore, the size of the prediction set in this case will be 2 or 3. It takes value 2 or 3 depending on the result of $\widehat{\pi}$ applying on the test sample. Applying $\widehat{\pi}$ on a test sample x_{n+1} and obtain a sorted probability vector $(\widehat{\pi}_{(1)}(x_{n+1}), \widehat{\pi}_{(2)}(x_{n+1}), \ldots, \widehat{\pi}_{(10)}(x_{n+1})) = (0.55, 0.2, 0.15, 0.1, 0, 0, 0, 0, 0, 0)$.

To determine whether the class with rank 3 should be included in the prediction set, we compare $\widehat{\pi}_{(3)}(x_i)$ to the distribution of the 3rd largest probability. The calculation $\frac{(1-\alpha)-0.86}{0.92-0.86} = \frac{2}{3} < 0.85$ indicates that the class with rank 3 should be included in the prediction set. Another equivalent way to think of this would be the probability 0.15 is the top 15% among all $\{\widehat{\pi}_{(3)}(x_i) : i = 1, 2, \ldots, n\}$, so the rank 3 class should be included in the prediction set $\widehat{C}_\alpha$.

Comparison with existing works: Our method may seem different from the approaches in Section 2, but there are connections. If $\widehat{\pi}_1(x_i), \ldots, \widehat{\pi}_K(x_i)$ are nearly identical, the ascending order of ranks in equation equation 1 is almost equivalent to the descending order of p-values in APS, suggesting our method makes fewer assumptions about $\widehat{\pi}$. In tie-breaking, APS uses a uniform random variable, while we use the THR idea to include labels with sufficiently large $\widehat{\pi}$ values. Thus, our method incorporates aspects of both APS and THR.

4. Theoretical Coverage Guarantee

In this section, we will demonstrate that our approach can theoretically achieve $1 - \alpha$ coverage. To begin, we will define the concept of exchangeability of random variables (See formal definition in

Section A in the Appendix). This assumption about the dataset is widely used when considering calibration samples and test samples Romano et al. (2020); Huang et al. (2024). Assuming exchangeability, we can demonstrate the following result: for a test sample X_{n+1} that has not been seen in the training or calibration set, the prediction set output by Algorithm 1 will include the true label Y_{n+1} with a probability of at least $1 - \alpha$.

Theorem 1 *If the samples (X_i, Y_i), for $i \in [n+1]$, are exchangeable and $\mathcal{B}$ from Algorithm 1 is invariant to permutations of its input samples, the output of Algorithm 1 satisfies:*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right) \geq 1 - \alpha. \tag{5}$$

Proof. See Section B in the Appendix.

5. Experiments

This section presents experiments that evaluate the performance of prediction sets generated by various methods, including APS (Romano et al., 2020), RAPS (Angelopoulos et al., 2021), SAPS (Huang et al., 2024), and our proposed method (RANK), on natural language processing tasks involving multiclass classification. The evaluation is conducted on three tasks: Multi-Choice Question-Answering, Topic Classification, and Emotion Recognition.

Multi-Choice Question-Answering: We evaluate our method on the MMLU benchmark (Hendrycks et al., 2021) for the Multi-Choice Question-Answering task, following the approach in Kumar et al. (2023). The datasets are generated using the LLaMA-13B model (Touvron et al., 2023) and consist of questions from three domains: college medicine (191 questions), marketing (269 questions), and public relations (123 questions).

Data	Coverage				Size				SSCV			
	Ours	APS	RAPS	SAPS	Ours	APS	RAPS	SAPS	Ours	APS	RAPS	SAPS
marketing	0.8962	0.9016	0.8997	0.8987	**2.6704**	2.8059	2.7920	2.7065	**0.0368**	0.0421	0.0408	0.0699
medicine	0.9043	0.8993	0.9000	0.9064	**3.3550**	3.3928	3.3841	3.3723	**0.1005**	0.1023	0.1032	0.1021
relations	0.8915	0.9048	0.8995	0.9024	**3.2260**	3.3569	3.3268	3.2903	**0.1007**	0.1026	0.1011	0.1033
agnews	0.9001	0.8993	0.8997	0.8992	**0.9703**	1.1654	1.1323	1.1807	**0.0048**	0.1000	0.0464	0.1000
news20	0.8992	0.9008	0.9004	0.9004	4.0828	**3.2626**	3.9847	3.4209	**0.0093**	0.0339	0.0403	0.0894
carer	0.8985	0.8990	0.9015	0.9005	**0.9305**	1.0390	1.0385	1.1065	**0.0110**	0.1000	0.0830	0.1000
tweet	0.9013	0.9007	0.8971	0.9007	**1.3382**	1.4801	1.4290	1.4476	0.0584	0.1000	**0.0414**	0.1000

Table 1: Evaluation metrics with $\alpha = 0.1$. Coverage (6): greater than or closer to $1 - \alpha = 0.9$ is better. Size (7): smaller is better. SSCV (8): smaller is better. **Bold** numbers indicate optimal performance.

Topic Classification: For topic classification, we assess our method on two datasets: AG News and 20 Newsgroups. AG News is a subset of AG's corpus of news articles, created by combining the titles and description fields of articles from the four largest topic classes: "World", "Sports", "Business", and "Sci/Tech". The AG News dataset contains 30000 training samples and 1900 test samples per class. The 20 Newsgroups dataset comprises newsgroup posts on 20 topics, split into a training set of 11314 posts and a test set of 7532 posts.

Emotion Recognition: To evaluate our method's performance on emotion recognition, we utilize two datasets: CARER and TweetEval. CARER consists of English Twitter messages labeled with six basic emotions: anger, fear, joy, love, sadness, and surprise. The dataset has a training set of 15969 tweets and a test set of 2000 tweets. TweetEval, on the other hand, contains tweets categorized into

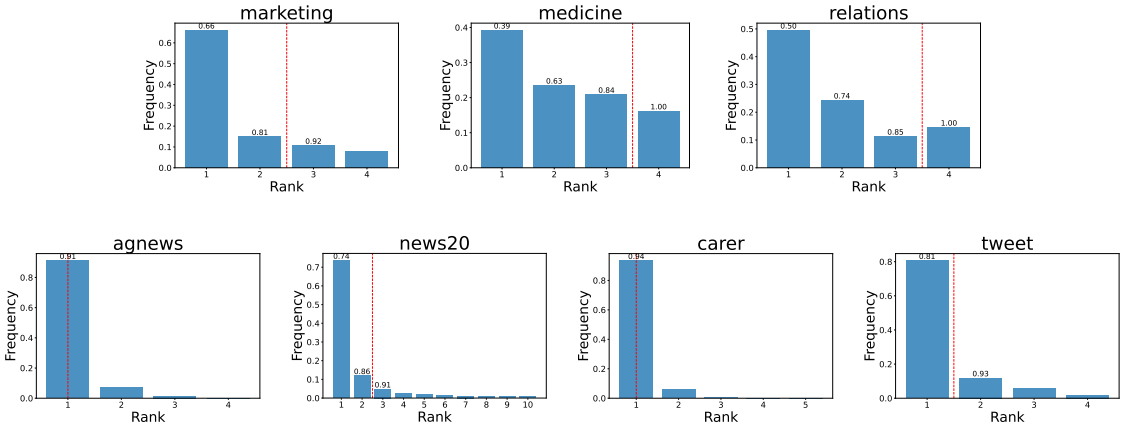

Figure 2: Rank distribution plots of the true class ranks for different datasets. The vertical red line indicates the rank threshold where the cumulative probability exceeds 0.90, corresponding to r^*_α from Algorithm 1. This value aligns with the average prediction set size for $\alpha = 0.1$ in Table 1.

four emotions: anger, joy, optimism, and sadness. This dataset includes 6838 training tweets and a test set of 1421 tweets.

Let us denote the test set by $\mathcal{I}_3$. We assess the performance of the different methods using the following three metrics.

Coverage Rate (Coverage): The coverage rate measures the proportion of test instances where the true label is included in the prediction set. A higher coverage rate indicates better performance.

$$\text{Coverage} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \mathbb{1}(y_i \in \widehat{C}(x_i)), \tag{6}$$

Average Size (Size): The average size refers to the mean number of labels in the prediction sets. Smaller sizes are consider more precise and informative of the labels in the prediction set.

$$\text{Size} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} |\widehat{C}(x_i)|. \tag{7}$$

Size-Stratified Coverage Violation (SSCV): The size-stratified coverage violation evaluates the consistency of coverage across different prediction set sizes $\{S_j\}_{j=1}^s$, where $S_1, S_2, \ldots, S_s$ are partitions of $[K]$. Let $J_j = \{i \in \mathcal{I}_3 : |\widehat{C}(x_i)| \in S_j\}$ denote the indices of examples stratified by the prediction set size S_j. Then we define

$$\text{SSCV}(\widehat{C}, \{S_j\}_{j=1}^s) = \sup_{j \in [s]} \left| \frac{|\{i \in J_j : y_i \in \widehat{C}(x_i)\}|}{|J_j|} - (1 - \alpha) \right|. \tag{8}$$

Smaller SSCV indicates more stable coverage.

Throughout the experiments, the split-conformal prediction framework is employed to construct the prediction sets. Different α values ranging from 0.1 to 0.3 are chosen, and the mean Coverage, Size, and SSCV metrics are computed across 100 repetitions.

Table 1 presents the results for $\alpha = 0.1$, while Figures 3 and 4 show the results for α ranging from 0.1 to 0.3. The left subfigure compares the Size vs. Coverage trade-off. A lower curve indicates

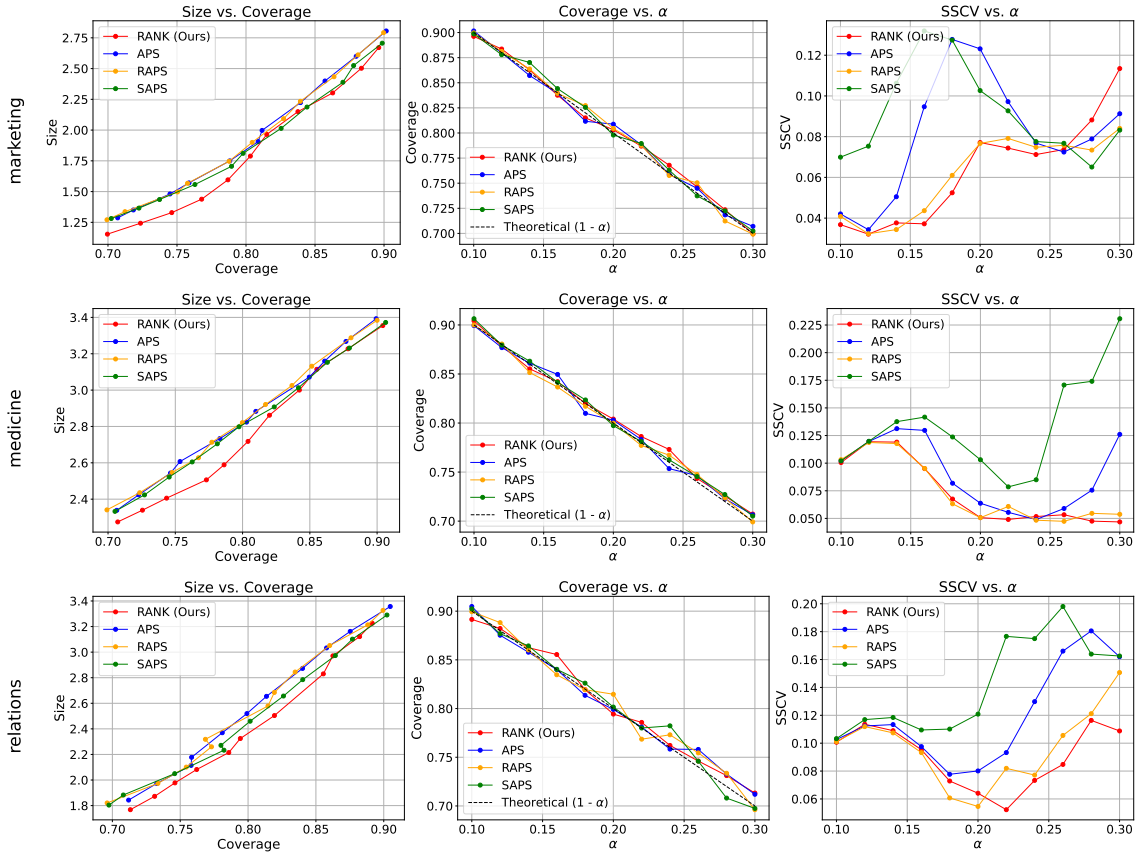

Figure 3: Results for the Multi-Choice Question-Answering datasets.

that the method can achieve the desired coverage using a smaller prediction set size. The middle subfigure illustrates the relationship between coverage and α. Methods closer to the theoretical line $1 - \alpha$ are considered better. The right subfigure displays the SSCV vs. α plot. A lower curve means that the method can achieve $1 - \alpha$ coverage more consistently across different strata of prediction set sizes.

The experiment result in Table 1, Figures 3 and 4 show that our method (RANK) preforms better than other completing methods in most NLP classification tasks, when measuring the performance of by prediction set size versus coverage, except for the dataset news20 with $\alpha \leq 0.15$. In the dataset agnews, carer and tweet, our performance is overwhelmingly better than the others. When comparing the SSCV metric with other methods, our approach demonstrates remarkably consistent coverage across most datasets.

To investigate our method's suboptimal performance on the news20 dataset compared to other methods, we examine its rank distribution plot in Figure 2. The plot reveals a long-tailed distribution, with many instances having high ranks. Moreover, this dataset contains 20 classes, which is significantly more than the other datasets. Our method may be less effective in such cases, as it does not explicitly minimize the tail probability of the ranks, unlike the APS-type approach, which is designed to handle these situations more effectively.

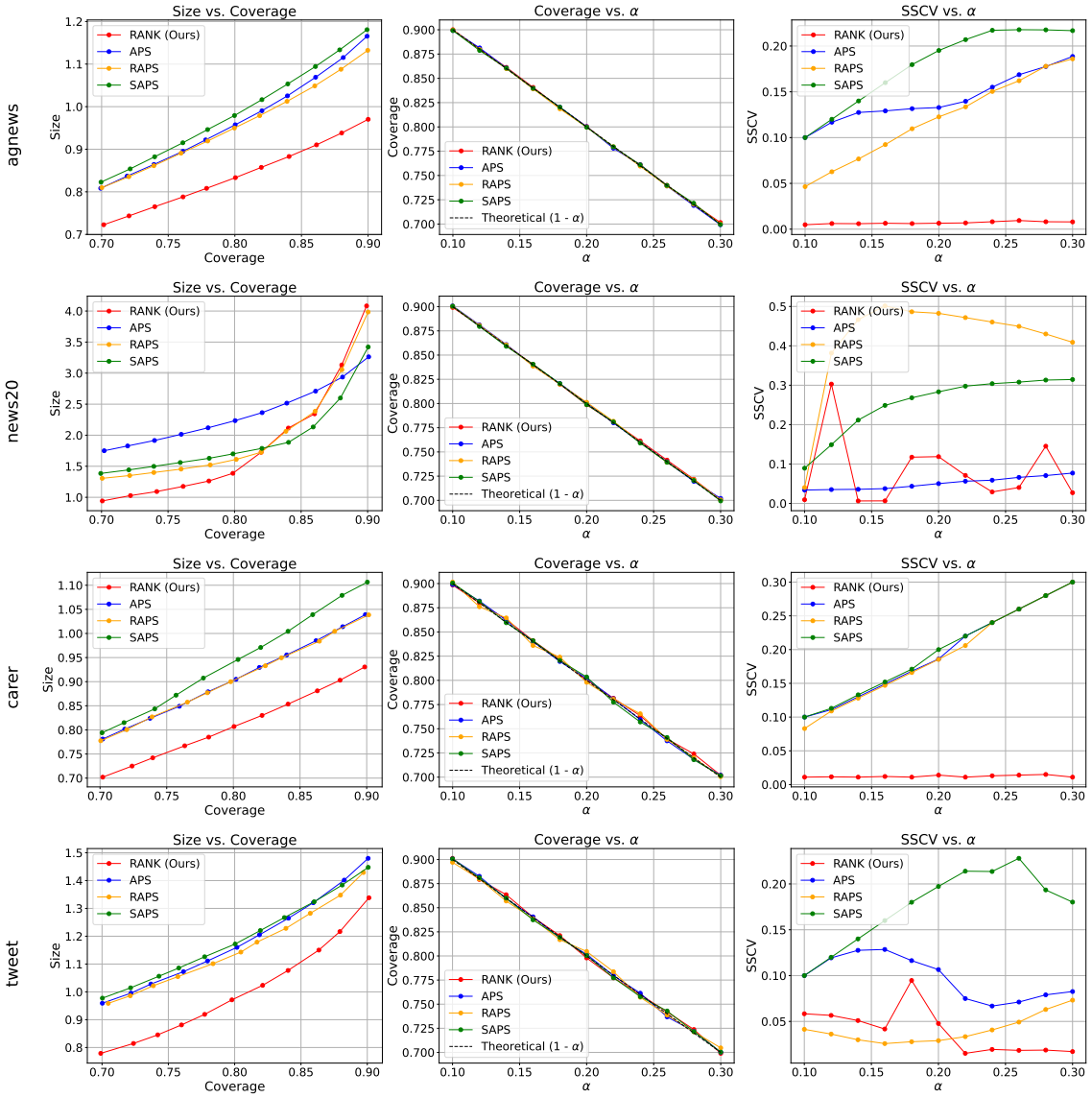

Figure 4: Results for Topic Classification on AG News (agnews), 20 Newsgroups (news20); and Emotion Recognition on CARER (carer), TweetEval (tweet).

6. Discussion and Future Work

Our proposed conformal prediction method for NLP classification tasks has demonstrated promising results in achieving high performance. However, there are several areas where further research and improvements can be made.

Performance on Large Number of Classes. Our approach's performance may deteriorate when faced with a very large number of classes, as seen in the news20 dataset results when $\alpha \leq 0.15$. A potential future direction could involve combining our idea of minimizing the prediction set size with the concept of minimizing the tail probability to improve results in these scenarios.

Multi-Label Classification. Extending our approach to handle multi-label classification, where each instance can be assigned multiple labels simultaneously, is an important future research direction. While most single-label methods can be directly applied to multi-label scenarios, accounting for label dependence becomes challenging. Using statistical methods to estimate label co-occurrence could help capture complex label relationships and improve classification performance.

7. Conclusion

We proposed a novel conformal prediction method for NLP classification tasks that effectively leverages confidence scores from deep learning models to construct prediction sets with desired coverage and managed size. Rigorous theoretical analysis and extensive experiments demonstrate our method's clear superiority over existing techniques, providing reliable uncertainty quantification for NLP classifiers. In the future, we will work on the development of more reliable uncertainty quantification methods for large language models.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Joris Baan, Subramanian Ramamoorthy, Dhanoop Pawade, Serguei Moshkov, and Martin Riedl. A survey of uncertainty estimation in text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16212–16220, 2023.

Jie Bao, Nicolo Colombo, Valery Manokhin, Suqun Cao, and Rui Luo. A review and comparative analysis of univariate conformal regression methods. In *Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2025)*, pages 282–304. PMLR, 2025a.

Jie Bao, Chuangyin Dang, Rui Luo, Hanwei Zhang, and Zhixin Zhou. Enhancing adversarial robustness with conformal prediction: A framework for guaranteed model reliability. In *Forty-second International Conference on Machine Learning*, 2025b.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3897–3911, 2020.

Nicolas Deutschmann, Marvin Alberts, and María Rodríguez Martínez. Conformal autoregressive generation: Beam search with coverage guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11775–11783, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Peiyun Feng, Dong Wang, Ziyang Liu, Xiaohu Yu, Anxiang Liu, Kang Li, Jianfeng Lu, and Hua Liu. Can language models be uncertainty-aware? exploring confidence estimation in machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 180–192, 2022.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Learning Representations*, 2022.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 539–555, 2020.

Ismael Gallegos, Luis Espinosa-Anke, Jose Rodríguez-Ferrández, Jorge Carrillo-de Albornoz, and Horacio Saggion. Measuring bias in multilingual natural language inference benchmarks. In *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3109–3121, 2023.

Charlie Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *International Conference on Machine Learning*, pages 3760–3769. PMLR, 2021.

Yoav Goldberg. *Neural network methods for natural language processing*, volume 10 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Mingfei Guo, Timothy Miller, Yi-Hsuan Tang, and Yue Xiong. Detecting biased samples in datasets with deep generative models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1688–1698, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *The Twelfth International Conference on Learning Representations*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Zhengbao Jiang, Zheng Wei, Hao Ren, Pengcheng Zhang, Hongtao Tao, Jianfeng Li, and Jure Leskovec. Natural language understanding with approximate string matching and conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

Lingkai Kong, Haoming Huang, Yuchen Hou, Heng Zhu, Lyle Ungar, and Haobo Guo. Calibrated language model fine-tuning for in-and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, 2020.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Guangyi Lu, Shen Yin, Sa Li, Zhiyuan Zhang, Kezhi Tian, Jianlong Wang, Xiubo Huang, Dongsheng Li, and Liang Shen. Contrastive learning for prompt-based few-shot language learners. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1313–1323, 2022.

Brian Lucena. Spline-based probability calibration. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1635–1643. PMLR, 2019.

Rui Luo and Nicolo Colombo. Entropy reweighted conformal classification. In *The 13th Symposium on Conformal and Probabilistic Prediction with Applications*, pages 264–276. PMLR, 2024.

Rui Luo and Nicolo Colombo. Conformal load prediction with transductive graph autoencoders. *Machine Learning*, 114(3):1–22, 2025.

Rui Luo and Zhixin Zhou. Conditional conformal risk adaptation. *arXiv preprint arXiv:2504.07611*, 2025a.

Rui Luo and Zhixin Zhou. Conformalized interval arithmetic with symmetric calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19207–19215, 2025b.

Rui Luo and Zhixin Zhou. Conformity score averaging for classification. In *Forty-second International Conference on Machine Learning*, 2025c.

Rui Luo and Zhixin Zhou. Density-sorted prediction set: Efficient conformal prediction for multi-target regression. *Pattern Recognition*, page 112513, 2025d.

Rui Luo and Zhixin Zhou. Reliable classification through rank-based conformal prediction sets. *Pattern Recognition*, page 112330, 2025e.

Rui Luo and Zhixin Zhou. Conformal thresholded intervals for efficient regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19216–19223, 2025f.

Rui Luo, Buddhika Nettasinghe, and Vikram Krishnamurthy. Anomalous edge detection in edge exchangeable social network models. In *Conformal and Probabilistic Prediction with Applications*, pages 287–310. PMLR, 2023.

Rui Luo, Jie Bao, Zhixin Zhou, and Chuangyin Dang. Game-theoretic defenses for robust conformal prediction against adversarial attacks in medical imaging. *arXiv preprint arXiv:2411.04376*, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, 2022.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press, 1999.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *Advances in Neural Information Processing Systems*, volume 36, pages 62534–62551, 2023.

Amirreza Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13456–13467, 2020.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api-cf: Api is enough for conformal prediction with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Lingxuan Tang, Rui Luo, Zhixin Zhou, and Nicolo Colombo. Enhanced route planning with calibrated uncertainty set. *Machine Learning*, 114(5):1–16, 2025.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in neural information processing systems*, volume 32, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Ting Wang, Zhixin Zhou, and Rui Luo. Enhancing trustworthiness of graph neural networks with rank-based conformal training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21261–21268, 2025.

Zheng Zhang, Jie Bao, Zhixin Zhou, Nicolo Colombo, Lixin Cheng, and Rui Luo. Residual reweighted conformal prediction for graph neural networks. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Appendix A. Definition of Exchangeability

Definition 2 (Exchangeability) *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of random variables. The sequence is said to be exchangeable if, for any permutation π of the indices $[n]$, the joint distribution of the permuted sequence $(Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(n)})$ is identical to the joint distribution of the original sequence $(Z_1, Z_2, \ldots, Z_n)$.*

Appendix B. Proof of Theorem 1

Proof. By the definition of the prediction set $\widehat{C}_\alpha$, $Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})$ is equivalent to the conformity score $c_{n+1} = c(x_{n+1}, y_{n+1}) < c_{(\lfloor (n+1)\alpha \rfloor)}$. By the exchangeablility of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$, $c(X_1, Y_1), \ldots, c(X_{n+1}, Y_{n+1})$ is also exchangeable. The rank of $c(X_{n+1}, Y_{n+1})$ has a uniform distribution on $\{1, \ldots, n+1\}$. The rank of $c(X_{n+1}, Y_{n+1})$ greater than $\lfloor (n+1)\alpha \rfloor$ with probability $\frac{n+1-\lfloor (n+1)\alpha \rfloor}{n+1}$, which is at least $1 - \alpha$.