Conformal Prediction for Natural Language Processing: A Survey

Margarida Campos[1,2] **António Farinhas**[1,2] **Chrysoula Zerva**[1,2,3]
Mário A. T. Figueiredo[1,2,3] **and André F. T. Martins**[1,2,3,4]

[1]Instituto de Telecomunicações, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]ELLIS Unit Lisbon, Portugal   [4]Unbabel, Portugal
`margarida.campos@tecnico.ulisboa.pt`

Abstract

The rapid proliferation of large language models and natural language processing (NLP) applications creates a crucial need for uncertainty quantification to mitigate risks such as Hallucinations and to enhance decision-making reliability in critical applications. Conformal prediction is emerging as a theoretically sound and practically useful framework, combining flexibility with strong statistical guarantees. Its model-agnostic and distribution-free nature makes it particularly promising to address the current shortcomings of NLP systems that stem from the absence of uncertainty quantification. This paper provides a comprehensive survey of conformal prediction techniques, their guarantees, and existing applications in NLP, pointing to directions for future research and open challenges.

1 Introduction

Natural language processing (NLP) is witnessing an explosive growth in applications and public visibility, namely with large language models (LLMs) being deployed in many real-life applications, ranging from general-purpose chatbots to the generation of medical reports (Min et al., 2023). However, the widespread use of these models brings important concerns: Hallucinations are frequent (Ji et al., 2023; Guerreiro et al., 2023), models are poorly calibrated (Vasudevan et al., 2019; Desai and Durrett, 2020; Zhou et al., 2024), evaluation is limited and sometimes affected by data contamination (Sainz et al., 2023; Golchin and Surdeanu, 2024), explanations are often unreliable (Zhao et al., 2024; Wiegreffe and Pinter, 2019), and models often exhibit undesired biases (Gallegos et al., 2024). Reliable uncertainty quantification is key to addressing some of these concerns: NLP systems should not only provide accurate answers but also ''know when they do not know''.

Unfortunately, some NLP systems return only single predictions (*i.e.*, point estimates), without reliable confidence information. Systems that quantify uncertainty are less common and typically limited in various ways: They often make incorrect distribution-based assumptions ignoring the complex nature of the underlying data and model (Xiao and Wang, 2019; He et al., 2020; Glushkova et al., 2021; Zerva et al., 2022); they are often poorly calibrated (*i.e.*, they predict a confidence level that does not match its error probability; Kuleshov et al., 2018); and they may be computationally too demanding, thus inapplicable to large-scale models (Hu et al., 2023).

Conformal prediction (CP; Vovk et al., 2005) has recently emerged as a promising candidate to bypass the issues above: Unlike other uncertainty quantification frameworks, it offers statistical guarantees of ground-truth coverage with minimal assumptions. CP methods are **model-agnostic** and **distribution-free**, assuming only data exchangeability (as described in §3.1). Moreover, extensions of CP that handle non-exchangeable data have recently been proposed (Gibbs and Candes, 2021; Barber et al., 2023). Popular CP variants are also **efficient**: They do not require model retraining and can be used online or offline, given an additional relatively small calibration set.[1] Finally, equalized variants of CP (Romano et al., 2020) can also reduce biases and unfairness, by distributing coverage evenly across protected attributes.

The flexibility and strong statistical guarantees of CP have attracted considerable interest, with an increasing number of publications in computer

---

[1]For most purposes, a reasonable calibration set size is of the order of 1000 samples (Angelopoulos and Bates, 2023).
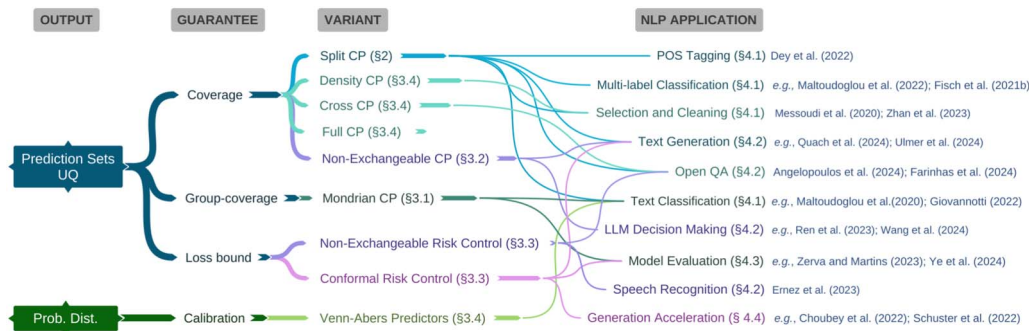
Figure 1: Survey roadmap: CP variants and their use in NLP applications with examples in the literature.

science.[2] It is therefore timely to present a survey of conformal methods for NLP, revealing the theory and guarantees behind these methods and outlining opportunities and challenges for them to tackle important problems in the field.

Scope. This survey provides a comprehensive overview of CP techniques for NLP tasks (Figure 1). After briefly explaining CP and some relevant extensions (§2 and §4), we review direct applications thereof in NLP (§5). Finally, we look at possible threads of future investigation and current open issues concerning the use of CP in NLP (§6).

What this survey is not about. This is *not* a general survey on uncertainty quantification and does not include techniques not based on CP. Comprehensive reviews of uncertainty quantification in NLP were recently published by Baan et al. (2023) and Hu et al. (2023). Also, our survey is focused on NLP applications; Angelopoulos and Bates (2023) and Shafer and Vovk (2008) have published comprehensive surveys on CP.

2 Conformal Predictors

This section briefly explains CP and presents some definitions and results needed for understanding the applications mentioned below. In what follows, we use upper case letters $(X, Y, ...)$ for random variables, lower case letters $(x, y, ...)$ for specific values they take, and calligraphic letters $(\mathcal{X}, \mathcal{Y}, \mathcal{C}, ...)$ for sets.

2.1 Definitions and Ingredients

Consider a prediction task where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output sets, respectively. The most common procedure is to learn/train a mapping $f : \mathcal{X} \to \mathcal{Y}$, which, given an input $x_{\text{test}} \in \mathcal{X}$, unseen during training, returns a **point prediction** $\hat{y}_{\text{test}} = f(x_{\text{test}}) \in \mathcal{Y}$, hopefully *close* to the "true" target y_{test}, according to some performance metric. A weakness of point predictions is the absence of information about uncertainty. In contrast, for the same input x_{test}, a conformal predictor yields a **prediction set** $\mathcal{C}_\alpha(x_{\text{test}}) \subseteq \mathcal{Y}$, ideally small, which includes the target y_{test} with some high (user-chosen) probability, say $1 - \alpha$.

Consider an example involving a pretrained model which classifies a clinical report $x \in \mathcal{X}$ with a label, *e.g.*, a disease $y \in \mathcal{Y}$. This is a high-risk scenario requiring strong reliability guarantees. For a random test report (X_{test}), a conformal predictor yields a set $\mathcal{C}_\alpha(X_{\text{test}})$ of possibly multiples labels, with the guarantee guarantee that[3] $\mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha$. Figure 2 illustrates the CP procedure for the mentioned task, which we describe next in detail.

Split[4] CP (Vovk et al., 2005) is built with three ingredients: a **trained predictor**, $f : \mathcal{X} \to \mathcal{Y}$; a **calibration set**, $\mathcal{D}_{\text{cal}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, independent from the set used to train the predictor; and a **non-conformity score**, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The non-conformity score measures how unlikely an input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is, compared to the remaining data. Consequently, given a test

---

[2]The number of arXiv papers in the field of computer science containing the expression "conformal prediction" has been steadily rising, from 16 papers in 2018 to 224 in 2023.

[3]Note that the probability is over $(X_{\text{test}}, Y_{\text{test}})$, *not* conditioned on a particular $X_{\text{test}} = x_{\text{test}}$. We discuss conditional coverage in §4.1.

[4]Although split (a.k.a. *inductive*) CP was developed after the *full* (a.k.a. transductive) variant (described in §4.4), it is more widely used due to its computational efficiency.
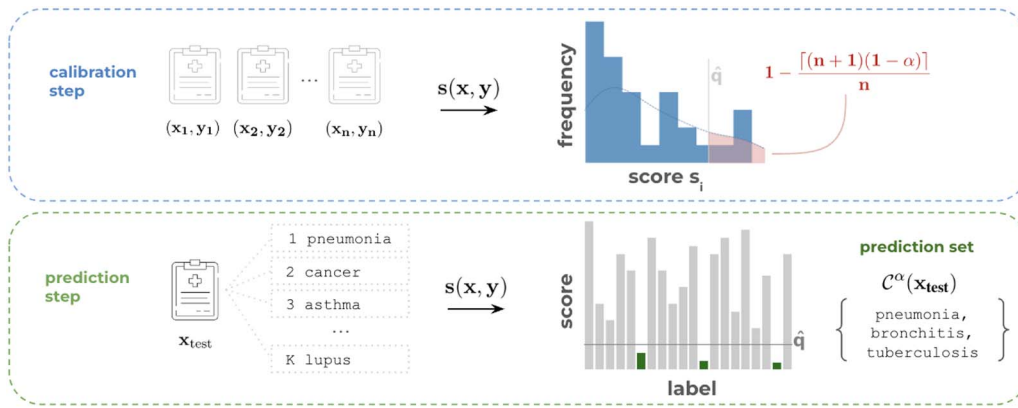
Figure 2: Example of CP for medical report classification (K possible labels). Calibration step: compute the non-conformity scores of the calibration points; find the empirical quantile $\hat{q}$ of their distribution. Prediction step: generate prediction set based on the non-conformity score of the observation paired with all possible labels, including in the s et all labels that yield a score below the $\hat{q}$ quantile.

sample x_{test}, predictions predictions $y \in \mathcal{Y}$ yielding pairs (x_{test}, y) deemed likely to data should have a low non-conformity score, and should thus be included in the the prediction set $\mathcal{C}_\alpha(x_{\text{test}})$.

The choice of non-conformity score is task-dependent. For example, for a classifier outputting an estimate $p(y|x)$ of the posterior probability for each possible label $y \in \mathcal{Y}$ (*e.g.*, via a softmax output layer), a common and natural choice is $s(x, y) = 1 - p(y|x)$, with lower values of $s(x, y)$ implying that the sample is more conformal with the previously seen data.

2.2 Prediction Procedure

The procedure for generating $\mathcal{C}_\alpha(x_{\text{test}})$ for new, unseen test instances x_{test} is as follows:

Calibration Step:

1. Compute $(s_1, \ldots, s_n)$, the non-conformity scores for $\mathcal{D}_{\text{cal}}$, where $s_i = s(x_i, y_i)$;

2. Set $\hat{q}$ to be the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the calibration set scores;

Prediction Step:

3. Output the prediction set, using the quantile $\hat{q}$, as $\mathcal{C}_\alpha(x_{\text{test}}) = \{y \in \mathcal{Y} : s(x_{\text{test}}, y) \leq \hat{q}\}$.

The intuition is that the prediction set includes all predictions corresponding to samples that are more conformal than a sufficiently large fraction of the calibration set. Note that the calibration step needs to be computed only once and the obtained quantile can then be used for all new test instances (Figure 2).

2.3 Relation to Hypothesis Testing

The CP procedure described above can be seen from a hypothesis-testing perspective. For each possible label, the tested hypothesis is whether the point (x_{test}, y) is conformal with the observed data, and the non-conformity measure is used as the test statistic. As an alternative to defining the threshold $\hat{q}$ using a preset α, we can think in terms of empirical p-values (Vovk et al., 2005). Define the p-value of a new sample (x_{n+1}, y_{n+1}) as

$$\text{p-value}(x_{n+1}, y_{n+1}) = \frac{\left| \{ j \in \{1, .., n\} : s(x_j, y_j) \geq s(x_{n+1}, y_{n+1}) \} \right| + 1}{n + 1},$$

the (adjusted) proportion of calibration points that are not less conformal than the observation. As in hypothesis testing, the p-value can be seen as the empirical probability of obtaining the observed score, under the null hypothesis that the observation is conformal. Using the p-value approach, the procedure to generate prediction sets for x_{test} is:

1. compute p-values for all labels $y \in \mathcal{Y}$;

2. generate prediction set as $\mathcal{C}_\alpha(x_{\text{test}}) = \{y \in \mathcal{Y} : \text{p-value}(x_{\text{test}}, y) > \alpha\}$.

A disadvantage of this approach is that it needs access to the calibration scores at test time. On the other hand, the p-values do not need a preset α and can be used to evaluate predictions, as shown next.

3 Guarantees and Evaluation

Prediction sets produced by conformal predictors have associated theoretical guarantees. This section reviews established coverage guarantees as well as evaluation metrics used to measure the quality of the prediction sets.

3.1 Theoretical Guarantees

As shown by Vovk et al. (2005), a conformal predictor, as defined in the previous section, generates prediction sets with coverage guarantees

$$\mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha, \qquad (1)$$

provided the data is exchangeable. Exchangeability means that the joint probability of the random variables generating the data is invariant under permutations thereof. Formally, a sequence $(Z_1, \ldots, Z_n)$ is said to be **exchangeable** if

$$(Z_1, \ldots, Z_n) \stackrel{d}{=} \left(Z_{\pi(1)}, \ldots, Z_{\pi(n)}\right) \qquad (2)$$

for any permutations π of $\{1, \ldots, n\}$, where $\stackrel{d}{=}$ stands for *identically distributed*. Exchangeability is a weaker requirement than the variables being independent and identically distributed (i.i.d). In fact, random variables that are i.i.d. are necessarily exchangeable; however, variables may be exchangeable without being independent, although they need to be identically distributed. The coverage guarantee is provided by the following theorem (Vovk et al., 2005):

> **Theorem 1.** *Let* $(Z_1, \ldots, Z_n, Z_{\text{test}})$ *be an exchangeable sequence of random variables, where* $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, *and* $\mathcal{C}_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ *a conformal predictor as described in §2.2. Then,* $\mathcal{C}_\alpha$ *satisfies*

$$1 - \alpha \leq \mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\right] \leq 1 - \alpha + \frac{1}{n+1}.$$

A predictor satisfying the coverage inequality given in Theorem 1 is said to be **valid**.[5] Note that as the size of the calibration set increases, the probability of coverage tends to exactly $1 - \alpha$. It is worth noting that the CP procedure we described is **model-agnostic** and **distribution-free**, *i.e.*, it makes no assumption about the data distribution, requiring only data exchangeability.

3.2 Set Efficiency Metrics

When assessing the quality of a conformal predictor, an important aspect beyond validity is **efficiency**: The prediction sets should be relatively small and adaptive—easier cases should yield smaller sets than harder observations. The efficiency of a conformal predictor depends on the trained predictor f and the chosen non-conformity score, which is typically based on some heuristic notion of prediction uncertainty, *e.g.*, using the softmax output of a model (§2.1).

Consider a separate test set $\mathcal{D}_{\text{test}} = \{(x_{n+1}, y_{n+1}), \ldots, (x_{n+k}, y_{n+k})\}$. Some metrics, called *a priori*, do not require access to the test set labels. This is the case of the **average prediction set size** (or interval width, in regression tasks): $S(\alpha) = \frac{1}{k}\sum_{i=1}^{k}|\mathcal{C}_\alpha(x_{n+i})|$, computed as a function of α. Using the test set labels, an informative *a posteriori* metric is the **observed fuzziness**, computed as the average of p-values for the false labels: $\text{OF} = \frac{1}{k}\sum_{i=1}^{k}\sum_{y \neq y_{n+i}} \text{p-value}\ (x_{n+i}, y)$, which should be as small as possible, since correct predictions should have high conformal scores, whereas incorrect labels should have low scores. These metrics can also be useful to evaluate adaptivity and bias, by comparing them over different partitions of the dataset, *e.g.*, split by a particular feature.

3.3 Pointwise Metrics

Conformal predictors provide point-level uncertainty metrics that can be used even in the **forced prediction** approach, *i.e.,* producing a single prediction $\hat{y}_i$, the label with the highest p-value

---

[5]Altough there are other definitions of validity in the CP literature (Vovk et al., 2005), this is the most common one, termed *conservative coverage validity*.

(typically coinciding with the original output of the point predictor), rather than the predicted set. Two common metrics in this case are **credibility**, $\mathrm{Cred}(x_i, \hat{y}_i) = \text{p-value}\,(x_i, \hat{y}_i)$, and **confidence**, $\mathrm{Conf}(x_i, \hat{y}_i) = 1 - \max_{y \neq \hat{y}_i}\ \text{p-value}\,(x_i, y)$. Credibility measures the reliability of a prediction based on its p-value: A higher p-value indicates that a larger portion of the calibration points are less conformal than the prediction, indicating a higher credibility in the prediction. Confidence is measured by looking at the second largest p-value, *i.e.*, by evaluating the second prediction candidate as alternative to the actual prediction: the lower this value, the more confidence we have in the actual prediction. These metrics make use of the calibration set to measure uncertainty and can be extremely useful, even if disregarding the full prediction set produced by the conformal predictor.

4   Extending Conformal Prediction

CP has extended beyond classic conformal predictors, with developments that allow handling challenges such as conditional coverage, dispensing with exchangeability, or obtaining guarantees beyond coverage. This section briefly presents the core ideas of some of the extensions that are most relevant for NLP applications.

4.1   Conditional Conformal Predictors

In many high-risk, critical settings, it may be important to obtain sample-conditional coverage,

$$\mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})|X_{\text{test}} = x_{\text{test}}\right] \geq 1 - \alpha, \quad (3)$$

for every $x_{\text{test}} \in \mathcal{X}$, *i.e.,* provide a uniform upper bound for each prediction error. This, however, is not achievable under the proposed general setting, although, in practice, the error probability in some situations may be close to α (Vovk, 2012; Gibbs et al., 2023; Barber et al., 2020). The study of conditional CP is an active area of research with solutions to obtain coverage guarantees conditional on protected attributes and dataset partitions (Jin and Ren, 2024; Gibbs et al., 2023; Feldman et al., 2021). This is extremely important for dealing with with class imbalance or fairness and bias concerns.

Vovk et al. (2005) introduced **Mondrian conformal predictors**: conditional predictors that provide coverage guarantees over different data *categories*, *e.g.*: partitions of the data by label or by a given feature. For example, in classification, it may be of interest to have

$$\mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})|Y_{\text{test}} = y\right] \geq 1 - \alpha, \quad (4)$$

for all $y \in \mathcal{Y}$, which is a class-conditional guarantee. The procedure described in §2.2 is adapted to compute quantiles (or p-values) within each class. This is simply achieved by computing class-specific quantiles $\hat{q}^k$ based on the non-conformity scores of the calibration samples from each class k. Finally, the prediction set is given by

$$\mathcal{C}_\alpha(x_{\text{test}}) = \{y \in \mathcal{Y} : s(x_{\text{test}}, y) \leq \hat{q}^y\}.$$

Assuming exchangeability (Eq. 2), the above procedure is guaranteed to satisfy (Eq. 4). This label-conditional example is a particular case of Mondrian conformal predictors, which applies to any mapping of the data into Mondrian taxonomies (Vovk et al., 2005). The same rationale can be used to obtain coverage across different partitions of the data, such as across a particular feature stratification.

4.2   Beyond Exchangeability

All theoretical guarantees presented so far are rooted in the assumption of data exchangeability (Eq. 2). However, this assumption is unrealistic in many NLP applications: For example, it is incompatible with the conditional nature of most language generation methods. Several extensions have been proposed which handle **non-exchangeable data**, which includes the cases of covariate and label shift (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021), time series (Chernozhukov et al., 2018; Xu and Xie, 2021; Angelopoulos et al., 2023), and other types of shift (Gibbs and Candes, 2021).

Recently, Barber et al. (2023) provided prediction guarantees without the exchangeability assumption. Let $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ be as

defined in §3.1, $Z = (Z_1, \ldots, Z_n, Z_{n+1})$ be a sequence of n calibration samples followed by a test sample, and Z^i denote Z after swapping Z_i with Z_{n+1}. Barber et al. (2023) proved that

$$\mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \, d_{\text{TV}}(Z, Z^i),$$
(5)

where $\tilde{w}_i := w_i / (1 + \sum_{i=1}^{N} w_i)$ are weights (with $w_i \in [0, 1]$), and $d_{\text{TV}}(Z, Z^i)$ is the total-variation distance between the distributions of Z and Z^i. Choosing higher weights for calibration samples such that Z and Z^i have similar distributions yields tighter bounds. Some open challenges related to this topic are discussed in §6.

4.3 Conformal Risk Control

While coverage guarantees are useful in many tasks, there are cases where the adequate notion of **error control** is not captured solely by guaranteeing that the prediction set contains the ground truth. Some extensions of CP address these cases.

Angelopoulos et al. (2024) consider multilabel classification, where each $Y_i \in 2^{\mathcal{Y}} \setminus \{\varnothing\}$ is a set of labels. The loss function to be controlled is thus defined on pairs of sets of labels, $\ell : (2^{\mathcal{Y}} \setminus \{\varnothing\}) \times (2^{\mathcal{Y}} \setminus \{\varnothing\}) \to \mathbb{R}$, and assumed to satisfy **monotonicity**: $A \subseteq B \Rightarrow \ell(A, Y) \geq \ell(B, Y)$, for any $Y \subseteq \mathcal{Y}$. They define prediction sets $\mathcal{C}_\lambda(x) = \{y \in \mathcal{Y} : f(y|x) \geq 1 - \lambda\}$, where $f(y|x) \in [0, 1]$ is the softmax output of class y, given by predictor f for input x, and a parameter λ. Invoking loss monotonicity yields $\lambda \leq \lambda' \Rightarrow \ell(\mathcal{C}_\lambda, Y) \geq \ell(\mathcal{C}_{\lambda'}, Y)$, for any $Y \subseteq \mathcal{Y}$.

In this setting, and given some desired upper bound β on the *risk* (*i.e.*, the expected loss), Angelopoulos et al. (2024) propose a criterion to select a value $\hat{\lambda}$, such that the following bound holds:

$$\mathbb{E}\left[\ell\big(\mathcal{C}_{\hat{\lambda}}(X_{\text{test}}), Y_{\text{test}}\big)\right] \leq \beta. \quad (6)$$

If ℓ is the miscoverage loss, *i.e.*, Y_{test} is a singleton and $\ell(\mathcal{C}, Y_{\text{test}}) = 1 - |\mathcal{C} \cap Y_{\text{test}}|$, the standard coverage guarantee in Eq. 1 is recovered, with $\beta = \alpha$.

This is also related to (but different from) previous work by Bates et al. (2021) and Angelopoulos et al. (2022), who prove bounds of the form

$$\mathbb{P}\left(\mathbb{E}\left[\ell(Y_{\text{test}}, \mathcal{C}_{\hat{\lambda}}(X_{\text{test}}))\right] \leq \beta\right) \geq 1 - \delta, \quad (7)$$

where δ is a parameter and ℓ does not need to be monotone. Angelopoulos et al. (2024) provide comprehensive comparison of these so-called *learn-then-test* (LTT) methods. Finally, it is also possible to combine some of the ideas of §4.2 and §4.3 to obtain non-exchangeable conformal risk control (Farinhas et al., 2024).

4.4 Other CP Variants

Full Conformal Prediction. Introduced by Vovk et al. (2005), full CP differs from the split version in two aspects: It does not use a separate calibration set, but the entire training set; and it involves model refitting—given a new instance, a model is trained for each possible label[6] and used with the full data set to compute the non-conformity scores and obtain the prediction set. A clear disadvantage of full conformal prediction is the high computational cost of retraining. However, it has advantages: full conformal predictors can be used if there is a limited amount of data and model retraining is not too expensive, providing the same validity guarantees (Lei et al., 2018).

Cross-validation and Jackknifing. The goal of these methods is to achieve a balance between statistical and computational efficiency. Cross-conformal predictors (Vovk, 2015) apply the cross-validation rationale to split conformal predictors. Each cross-validation fold is used as a calibration set once and the p-values are computed using all folds. These predictors, although lacking proven validity guarantees, have shown good empirical results (Vovk et al., 2018). Inspired by this idea, Barber et al. (2021) propose the so-called *jackknife+*, a leave-one-out scheme, and prove validity for regression under some conditions.

Density-based Conformal Prediction. Hechtlinger et al. (2019) propose a different approach to the conformal procedure, based on $p(x|y)$ instead of the typical $p(y|x)$ to build more cautious predictors that should output the

---

[6]For regression, discretization is typically used.

null set when underconfident. This method can be useful to abstain from answering when given an outlier observation. They show promising results using adversarial attacks on different tasks.

Venn-Abers Predictors. This class of probabilistic predictors has guarantees proved by Vovk and Petej (2014). They produce one probability distribution per possible label and provide guarantees that one of the predictive distributions is perfectly calibrated, with no assumptions on the model or data distribution. Venn-Abers have been shown to be a good calibration tool with the added benefit that the distance between the different probability distributions provides calibrated uncertainty quantification (Johansson et al., 2023). A more efficient split variant is proposed by Lambrou et al. (2014), and Manokhin (2017) presents a multi-class generalization.

5 Applications in NLP

CP has been used in several NLP tasks, both to get validity/calibration guarantees on predictions; or within a pipeline, *e.g.*, to safely prune intermediate outputs with guaranteed coverage, achieving computational speedups (§5.4). This section reviews several such applications organized by use case.

5.1 Text Classification and Sequence Tagging

For classification and tagging tasks, models are often accurate but lack reliable confidence estimates.

Binary Text Classification. Maltoudoglou et al. (2020) build a conformal predictor on top of a BERT classifier (Devlin et al., 2019) for binary sentiment classification. They show that the conformal predictor with forced prediction retains the original model's accuracy while providing useful accompanying measures of credibility and confidence. For the same task, Messoudi et al. (2020) use density-based CP (§4.4). They report good performance and empirical validity, highlighting the usefulness of having such a predictor by considering noisy and outlier observations: The CP set contains both classes for the noisy example and is empty for the outliers, showing the desired discriminatory power. Zhan et al. (2022) automate identification of literature on drug-induced liver injury, using conformal

prediction to manage prediction uncertainty and guaranteeing reliability.

Giovannotti (2022) uses Venn-Abers predictors (§4.4) with different transformers model architectures on several binary tasks, such as paraphrase detection, sentiment analyses, and Boolean question answering, obtaining good calibration results with evenly distributed probability distributions.

Classification with Conditional Coverage. Mondrian CP (§4.1) has been successfully applied to unbalanced classification tasks, such as sentiment analysis, with good efficiency results (Norinder and Norinder, 2022). Giovannotti and Gammerman (2021) compare split, Mondrian, and cross-conformal (Vovk, 2015) CP on unbalanced paraphrase detection and report that the theoretically expected efficiency drop for Mondrian CP is small, making it useful in practice.

POS Tagging. Dey et al. (2022) present promising results by showing that CP based on the softmax outputs of a BERT model for POS tagging yields practical prediction sets even at high confidence levels on a large test set: At the 99% confidence level, fewer than 4% of the prediction sets had more than one answer.

Multilabel Tasks. CP has been used for multilabel text classification, where multiple labels can be assigned to an input, *e.g.*, document categorization. In the label powerset approach (Tsoumakas et al., 2010), which treats each possible combination of labels as a class, there is an added challenge due to the large output space. Paisios et al. (2019) show how CP can be used in this setting, exploring different task-appropriate non-conformity scores for the task of categorization of news articles. The forced prediction method (§3.3) shows negligible performance drops (as a consequence of part of the training data being set aside for calibration) while providing reliable credibility measures; moreover, the prediction sets were tight and well-calibrated at high confidence levels. Maltoudoglou et al. (2022) build on top of the aforementioned work and propose an efficient computational approach that allows a higher number of possible labels to be considered. Fisch et al. (2022) tackle the multilabel case under the need to limit false positive predictions—a type of constraint that arises naturally in many highly sensitive tasks, *e.g.*, identification of drug properties—by using a computationally

efficient method that provides the desired coverage and constraint guarantees for an NER task, reporting prediction sets of useful size.

A different approach has been considered in tasks such as document retrieval, where it may be of interest to obtain prediction sets with at least one admissible correct answer. Fisch et al. (2021b) present an efficient conformal procedure to find such sets. They exploit the fact that simpler and lighter models can be used first in the pipeline to reduce the number of output candidates, producing a sequence of conformally valid candidates that are passed on to more complex models, showing that the final output is guaranteed to yield the user desired coverage.

Dealing with Limited Data. CP has also been found useful in providing guarantees for tasks with limited amounts of data. Fisch et al. (2021a) tackle few-shot relation classification with CP procedures to meta-learn both non-conformity measures and a threshold predictor from auxiliary tasks with larger amounts of available data. Not only do the predicted sets for the final task grant coverage requirements, but they are also small (average set size smaller than 2 for 95% confidence level). A different approach is used by Dutta et al. (2023) for estimating uncertainty in zero-shot biomedical image captioning using CLIP models (Radford et al., 2021): They query the Web to get a calibration set and design a CP protocol that takes into account the plausibility of each calibration point, providing promising results with small predicted test sets with coverage even in the absence of original labeled calibration data. In a setting with limited reliable data, Zhan et al. (2023) use CP to clean possibly mislabeled training data, based on a small curated amount of data as a calibration set. They explore the effects of removing or changing the label of noisy data identified by the conformal procedure and show performance improvements on the text classification downstream task for different levels of induced noise.

5.2 Natural Language Generation

Despite their impressive capabilities, large language models are prone to hallucinations (Huang et al., 2023; Ji et al., 2023). The strong correlation between hallucinations and uncertainty unawareness makes CP a promising approach to tackle this issue. Yadkori et al. (2024) use conformal risk control to obtain an upper bound on the hallucination risk, developing an abstention procedure. In spite of its potential, the application of CP to language generation faces two big challenges: (i) the combinatorially large size of output sets and (ii) the conditional (recursive) nature of language generation, which violates the exchangeability assumption underlying standard CP.

Sentence-level Conformal Prediction. Most research on CP for NLP tries to circumvent the issues above by operating at the sentence level, *e.g.*, by first sampling multiple options and then reformulating the problem as a multiple choice question (Kumar et al., 2023). For instance, an LLM can be used to generate plans (expressed in natural language) for a robot to follow but a single plan alone may result in unfeasible or risky actions. Ren et al. (2023) build upon the methods presented in §2 to calibrate the confidence of **LLM planners**, providing formal guarantees for task completion while minimizing human help. Specifically, they look at the next-token probability to assess the uncertainty of different possible actions (*i.e.*, they use it to compute the non-conformity score, as described in §2.2) and generate CP sets. If the prediction set is not a singleton, the robot should ask for help; otherwise, it should continue to execute the plan. Liang et al. (2024) further enhance this framework by incorporating an ''introspective reasoning'' step (Leake, 2012), which leads to tighter prediction bounds, while Wang et al. (2024) consider teams of robots.

Sentence-level Risk Control. Quach et al. (2024) show how LTT (§4.3) can be used to calibrate a stopping rule for sampling outputs from a language model that are added to a growing set of candidates until they are confident that the set includes at least one acceptable hypothesis (Fisch et al., 2021b)—this was applied, for example, to document retrieval for fact verification, where the presence of one admissible document is sufficient to solve the task. Simultaneously, they calibrate a rejection rule to remove low-quality and redundant candidates. They use Pareto testing (Laufer-Goldshtein et al., 2023) to efficiently search and test the high-dimensional hyperparameter configuration. The resulting output sets are not only valid but also precise (*i.e.*, small). Angelopoulos et al. (2024) and Farinhas et al. (2024) apply conformal risk control to

open-domain question answering, whereas Ernez et al. (2023) do it for speech recognition. While the former calibrate the best token-based F_1-score of the prediction set in Eq. 6, the latter control the word error rate to an adjustable level of guarantee. Finally, Zollo et al. (2023) discuss how prompts that perform well on average on a validation set may be prone to produce poor generations with high probability in deployment and propose **prompt risk control** based on upper bounds on families of informative risk measures.[7] Specifically, they bound the worst-case toxicity (Hanu and Unitary team, 2020) in chatbots, the expected loss (pass@K, Kulal et al., 2019) in code generation, and the dispersion of ROUGE scores (Lin, 2004) in medical summarization.

Token-level Approaches. While the approaches above focus on full sentences, language models generate text by successively producing new tokens autoregressively. Nucleus sampling (Holtzman et al., 2020) samples each token from the smallest set whose cumulative probability exceeds a threshold. However, Ravfogel et al. (2023) observe that LLMs tend to be overconfident—the prediction sets used in nucleus sampling are not calibrated (see their Figure 4)—and this does not improve by scaling up the model size. They propose **conformal nucleus sampling**, which calibrates prediction sets within bins of similar entropies. As an alternative, Ulmer et al. (2024) take non-exchangeability (§4.2) into account by using a dynamic calibration step. They use the k-nearest neighbors and data-dependent relevance weights based on the squared ℓ_2 distance between the embedding representations. This leads to smaller prediction sets compared to previous approaches while maintaining the desired coverage level in machine translation and language modeling.

5.3 Uncertainty-Based Evaluation

CP can also be used to assist in evaluating and benchmarking NLP models. Two main approaches employ CP to that end: (i) using it to assess the confidence of different models and compare them accordingly; (ii) framing evaluation as a regression task (*i.e.*, learning to score the model outputs to predict human perceived quality and using CP to provide reliable confidence intervals).

Focusing on the former approach, Ye et al. (2024) apply CP to benchmark the performance of different LLMs. They use prompt engineering to turn different generation tasks (question answering, summarization, commonsense inference, etc.) into multiple-choice questions such that the models need to predict a letter corresponding to each candidate output. They subsequently attempt to quantify the uncertainty of the language model over the possible labels, conformalizing the softmax outputs for each candidate label. They show that high model accuracy does not necessarily imply high certainty; in some cases, an inverse correlation between accuracy and certainty is observed. Based on their findings, Ye et al. (2024) propose an uncertainty-aware metric accounting for both accuracy and uncertainty (encoded as set size).

Focusing instead on the latter approach, Giovannotti (2023) applies CP to referenceless MT evaluation (quality estimation), specifically using conformal predictive distributions (CPDs) as introduced by Vovk et al. (2017) to estimate the probability distribution around each quality estimate, using a k-NN model to obtain quality scores and subsequently the distances between each point and its neighbors to form non-conformity scores. The non-conformity scores computed in this way provide a proxy to the uncertainty of each quality estimate and they thus use CP as a method to obtain calibrated confidence intervals for MT quality estimation. Zerva and Martins (2023), on the other hand, apply CP on top of non-conformity heuristics coming from other uncertainty quantification methods for reference-based MT evaluation and discuss how such method choice can impact coverage and width. They also highlight biases in estimated confidence intervals, reflected in imbalanced coverage for attributes such as translation language and quality, demonstrating how these can be addressed with equalized CP. While focused on MT, the proposed approaches are applicable to other NLP evaluation or regression tasks.

5.4 Faster Inference

Given the high computational requirements of state-of-the-art NLP models and their widespread use, considerable effort is being put on making

---

[7]They use the terms *loss* and *risk* in a distinctive way. Loss refers to scoring the quality of a single sample generation (*e.g.*, ROUGE); risk measures some aspect of the distribution of the loss across the population (*e.g.*, mean).

these models more time- and memory-efficient (Deng et al., 2020). Several strategies for increasing efficiency at prediction time (*e.g.*, early exiting, Liu et al., 2019; Schwartz et al., 2020) focus on identifying easily classifiable instances and using a lighter version of the original model to predict them. Such instances must be reliably identified and both the original and simplified models should consistently produce the same results for a given input with a high probability.

Early Exiting Transformers Fine-tuning. Schuster et al. (2021) present an extension of CP to build a method to speed up inference in transformer models, while guaranteeing an adjustable degree of consistency with the original model, with high confidence. The rationale is to skip directly to the final layer from one of the previous layers whenever there is enough confidence. They use a binary meta-classifier to predict whether the lighter model is consistent with the original one and use CP to predict the set of inconsistent models. The final procedure consists of exiting at the first layer that exceeds the threshold found by the conformal procedure. Their method shows reduced inference time in several classification and regression tasks.

Filtering Labels. Choubey et al. (2022) tackle the computational efficiency problem in zero-shot text classification with pretrained language models, looking at the fact that inference time increases with the number of possible labels. They use CP on top of a base, simple and fast, text classifier to reduce the number of possible labels for the final, more complex, language model. They experiment on different classification tasks, testing different choices of non-conformity scores and different base models, exploring the trade-off between efficiency and accuracy in choosing the complexity of the base model.

Speeding Up Inference. To obtain the lightest possible model while preserving performance, Laufer-Goldshtein et al. (2023) propose a CP method to find optimal thresholds to guarantee several risk constraints with adjustable high probability, while optimizing another objective function. They report results on several text classification tasks with different objectives, such as minimizing prediction cost (searching thresholds on all pruning directions), while controlling accuracy reduction (drop in performance from the full to a lighter model) to a user-chosen degree. Their method builds upon the LTT procedure (§4.3), with an efficient technique to reduce the number of parameter combinations tested, using Pareto-optimal solutions (Deb and Kalyanmoy, 2001). The results show significant efficiency gains with the proposed risk-controlling guarantees.

Schuster et al. (2022) make text generation more efficient by considering decoder early exiting at the token level, while bounding global efficiency. They leverage the LTT procedure to obtain risk-controlling solutions with dynamic allocation of compute per generated token and test their approach on news summarization, text translation and open question answering, showing efficiency gains with the required quality guarantees.

6 Future Directions

We outline in this section some promising future research directions and open challenges related to the use of CP and its many variants in NLP tasks.

6.1 CP for Human–Computer Interaction

Some tasks in NLP, such as recommendation and predictive writing systems, benefit naturally from prediction sets that can be used to offer suggestions to users. CP is an opportunity for improving the efficiency and quality of such systems and prediction sets can be used to enhance performance in decision-making with humans in the loop (Cresswell et al., 2024). This aspect could be further explored in NLP, as there are numerous scenarios involving human feedback, *e.g.*, interactive MT (Green et al., 2013; Wang et al., 2021) or creation of human preference data for LLM alignment (Stiennon et al., 2020; Fernandes et al., 2023).

6.2 CP for Handling Label Variation

The complexity and ambiguity of natural language, as well as the varied human perspectives, make it hard to disentangle model uncertainty from valid, naturally occurring label variation (Baan et al., 2024; Plank, 2022; Baan et al., 2022). It is often the case that multiple outputs are correct, particularly in tasks involving high variation in human language production (question answering, summarization, and other generation tasks where several output variants are equivalent) or inherent, plausible disagreement (the ChaosNLI data

that demonstrates valid disagreements in textual inference annotations [Pavlick and Kwiatkowski, 2019]). While traditional methods focus on the majority class, or see variation as model uncertainty, CP yields a more faithful representation of label variation. Besides representing uncertainty, the sets produced by CP provide multiple ''equivalent'' labels, allowing for more interpretable and informed predictions. Further research on such scenarios could provide models that behave better in tasks with high label variation. Moreover, in such cases, CP can also be used to achieve diverse prediction sets, avoiding redundancy, as suggested by Quach et al. (2024).

6.3 CP for Fairness

The increased use of NLP systems in global daily life and high-risk tasks raises concerns about the fairness of their outputs. Many of these systems have been shown to be *biased* (Blodgett et al., 2020). In tasks such as resume filtering, medical diagnosis assistance, and several others, these biases can be extremely harmful, leading to skewed performance and coverage. CP can be used to achieve equalized coverage for different population groups (Romano et al., 2020), thus ''correcting'' biases in model predictions without the need for expensive retraining. The open research problem of finding conditional guarantees (Gibbs et al., 2023) to obtain pointwise error bounds can also contribute towards fairness in NLP applications.

6.4 CP for Dealing with Data Limitations

Learning and quantifying uncertainty with limited data is challenging, particularly in NLP problems where manual text labeling can be difficult, time-consuming, and expensive. Approaches to leverage limited data, such as active learning, make use of uncertainty quantification in order to reduce the need for manual labelling (Settles, 2009). In these settings, CP could be used for reliable uncertainty quantification, *e.g.*, selecting points with larger prediction sets for manual labelling. The predicted sets can also be useful to reduce the possible labels in tasks with high cardinality output spaces, increasing the performance of subsequent predictions. Another option is to use CP for data filtering and cleaning to increase the performance of LLMs (Marion et al., 2023), using for example a small reliable set for calibration, in order to identify mislabeled or noisy samples.

6.5 CP for Uncertainty-Aware Evaluation

CP is also useful for tackling the current challenge of model evaluation. There are some concerns regarding the current way NLP systems are evaluated: *e.g.*, questioning how confident we can be in evaluations that result from an LLM scoring the output of another one. Evaluating a conformal predictor built on top of a predictor can be a more reliable way to assess model performance. Another useful application of CP is to compare different uncertainty heuristics and transformations of model outputs by designing distinct non-conformity scores and evaluating the efficiency (*e.g.*, set size, conditional coverage, observed fuzziness) of the resulting predictors (§5.3).

6.6 Open Challenges

Despite its numerous applications, using CP in NLP poses challenges, particularly in generation tasks, providing exciting areas for further research.

Sentence Level. The high cardinality of the output space in generation tasks raises a challenge to typical CP applications. There are open questions on how to sample the possible outputs and on what is the impact of considering a finite set of samples.

Token Level. The non-exchangeability of the data tackled by Barber et al. (2023), Ulmer et al. (2024), and Farinhas et al. (2024) still presents an obstacle since it is not currently easy to: quantify the coverage gap—the bound in Eq. 5 involves computing a total variation distance between unknown distributions, which is hard to estimate; find good strategies for choosing the weights.

Distribution Shift. Several extensions to distribution drift and non-exchangeable data have been proposed for time series by Chernozhukov et al. (2018, 2021b), Xu and Xie (2021), Stankeviciute et al. (2021), Zaffran et al. (2022), and Angelopoulos et al. (2023) and for other scenarios by Cauchois et al. (2020), Gibbs and Candes (2021), Chernozhukov et al. (2021a), and Gibbs and Candès (2022). However, it remains unclear how to apply them in the context of NLP.

7 Discussion

Significant efforts have been made to improve the quality and confidence estimation of NLP systems.

In fact, recent techniques for generating text with LLMs often involve the generation of multiple hypotheses, followed by a reranking (or voting) stage to increase the likelihood of producing a high-quality prediction. These methods include minimum Bayes risk decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020), reranking based on quality estimation (Fernandes et al., 2022), and other strategies (Wang et al., 2023; Suzgun et al., 2023). While these approaches have been shown to improve quality and reduce the number of generated hallucinations (Farinhas et al., 2023; Guerreiro et al., 2023), they do not inherently quantify the uncertainty of their predictions or provide the formal guarantees offered by CP. Sensitive tasks, such as those in fields like medicine and education, may require guarantees. In these cases, the system could abstain from answering or defer to a human whenever its confidence is below a certain threshold. CP provides theoretical guarantees in a distribution-free setting, making it highly useful in such scenarios.

Other calibration methods have been used to obtain uncertainty intervals for regression and classification tasks in NLP. Calibration methods that target accuracy-related metrics, e.g., aiming to minimise the estimated calibration error (ECE) have been widely used (Li et al., 2023; Wang et al., 2022). Still, they also suffer from significant drawbacks as they are sensitive to binning choices and small changes in model estimates (Błasiok et al., 2023; Roelofs et al., 2022; Minderer et al., 2021; Chidambaram et al., 2024). Furthermore, they do not provide any formal guarantees to control risk and objectives of interest. CP can thus be used to complement such methods, especially in cases where ensuring coverage is deemed important.

8 Conclusion

This paper provides an overview of applications of the conformal prediction framework in NLP tasks, after a brief introduction to that framework and its main variants. We showed how conformal prediction is a promising tool to address the uncertainty quantification challenge in NLP and hope the existing and possible applications presented in this survey will motivate future research on the topic.

Acknowledgments

References

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J. Tibshirani. 2023. Conformal pid control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, pages 23047–23074. Curran Associates, Inc.

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591. https://doi.org/10.1561/2200000101

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.124

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association*

for Computational Linguistics (Volume 2: Short Papers), pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2020. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(1):455–482. https://doi.org/10.1093/imaiai/iaaa017

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1). https://doi.org/10.1214/20-aos1965

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2): 486–507. https://doi.org/10.1214/23-aos2276

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34. https://doi.org/10.1145/3478535

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pages 1727–1740, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3564246.3585182

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of ''bias'' in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. 2020. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. 2018. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749. PMLR. https://doi.org/10.1920/wp.cem.2018.1618

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. 2021a. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118. https://doi.org/10.1073/pnas.2107794118

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. 2021b. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864. https://doi.org/10.1080/01621459.2021.1920957

Muthu Chidambaram, Holden Lee, Colin McSwiggen, and Semon Rezchikov. 2024. How flawed is ece? An analysis via logit smoothing. *arXiv preprint arXiv:2402.10046*

Prafulla Kumar Choubey, Yu Bai, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2022. Conformal predictor for improving zero-shot text classification efficiency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3034, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.196

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. 2024. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*.

Kalyanmoy Deb and Deb Kalyanmoy. 2001. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., USA.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532. https://doi.org/10.1109/JPROC.2020.2976475

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods*

in *Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.21`

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

Neil Dey, Jing Ding, Jack Ferrell, Carolina Kapper, Maxwell Lovig, Emiliano Planchon, and Jonathan P. Williams. 2022. Conformal prediction for text infilling and part-of-speech prediction. *The New England Journal of Statistics in Data Science*, 1(1):69–83. `https://doi.org/10.51387/22-NEJSDS8`

Shiladitya Dutta, Hongbo Wei, Lars van der Laan, and Ahmed Alaa. 2023. Estimating uncertainty in multimodal foundation models using public internet data. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics. `https://doi.org/10.18653/v1/2020.coling-main.398`

Fares Ernez, Alexandre Arnold, Audrey Galametz, Catherine Kobus, and Nawal Ould-Amer. 2023. Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0). In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 16–35. PMLR.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.733`

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. 2024. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Shai Feldman, Stephen Bates, and Yaniv Romano. 2021. Improving conditional coverage via orthogonal quantile regression. In *Advances in Neural Information Processing Systems*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.100`

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668. `https://doi.org/10.1162/tacl_a_00626`

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Dr.Regina Barzilay. 2021a. Few-shot conformal prediction with auxiliary tasks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3329–3339. PMLR.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Dr.Regina Barzilay. 2022. Conformal prediction sets with limited false positives. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6514–6532. PMLR.

Adam Fisch, Tal Schuster, Tommi S. Jaakkola, and Regina Barzilay. 2021b. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*. https://doi.org/10.1162/coli_a_00524

Isaac Gibbs and Emmanuel Candes. 2021. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc.

Isaac Gibbs and Emmanuel Candès. 2022. Conformal inference for online prediction with arbitrary distribution shifts.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. 2023. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*.

Patrizio Giovannotti. 2022. Calibration of natural language understanding models with venn–abers predictors. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 55–71. PMLR.

Patrizio Giovannotti. 2023. Evaluating machine translation quality with conformal predictive distributions. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 413–429. PMLR.

Patrizio Giovannotti and Alex Gammerman. 2021. Transformer-based conformal predictors for paraphrase detection. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 243–265. PMLR.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938. Punta Cana, Dominican Republic, Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.330

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. https://doi.org/10.1145/2470654.2470718

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517. https://doi.org/10.1162/tacl_a_00615

Laura Hanu and Unitary team. 2020. Detoxify.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.671

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. 2019. Cautious deep learning. *arXiv preprint arXiv:1805.09460*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). https://doi.org/10.1145/3571730

Ying Jin and Zhimei Ren. 2024. Confidence on the focal: Conformal prediction with selection-conditional coverage. *arXiv preprint arXiv:2403.03868*.

Ulf Johansson, Tuwe Löfström, and Cecilia Sönströd. 2023. Well-calibrated probabilistic predictive maintenance using venn-abers. *arXiv preprint arXiv:2306.06642*.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S. Liang. 2019. Spoc: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR.

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

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Antonis Lambrou, Ilia Nouretdinov, and Harris Papadopoulos. 2014. Inductive venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74. https://doi.org/10.1007/s10472-014-9420-z

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi S. Jaakkola. 2023. Efficiently controlling multiple risks with pareto testing. In *The Eleventh International Conference on Learning Representations*.

David B. Leake. 2012. *Introspective Learning and Reasoning*. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_1802

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Jiazheng Li, Zhaoyue Sun, Bin Liang, Lin Gui, and Yulan He. 2023. CUE: An uncertainty interpretation framework for text classifiers built on pre-trained language models. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1253–1262. PMLR.

Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. 2024. Introspective planning: Guiding language-enabled agents to refine their own uncertainty. *arXiv preprint arXiv:2402.06529*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. 2022. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognition*, 122:108271. https://doi.org/10.1016/j.patcog.2021.108271

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 269–284. PMLR.

Valery Manokhin. 2017. Multi-class probabilistic classification using inductive and cross Venn–Abers predictors. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 228–240. PMLR.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining LLMs at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*.

Soundouss Messoudi, Sylvain Rousseau, and Sebastien Destercke. 2020. *Deep Conformal Prediction for Robust Models*. https://doi.org/10.1007/978-3-030-50146-4_39

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2). https://doi.org/10.1145/3605943

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.

Ulf Norinder and Petra Norinder. 2022. Predicting amazon customer reviews with deep confidence using deep learning and conformal prediction. *Journal of Management Analytics*, 9(1):1–16. https://doi.org/10.1080/23270012.2022.2031324

Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. 2019. A deep neural network conformal predictor for multi-label text classification. In *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pages 228–245. PMLR.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

Aleksandr Podkopaev and Aaditya Ramdas. 2021. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 27–34, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.3

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley,

Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2). https://doi.org/10.1162/99608f92.03f00592

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.722

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*.

Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. 2021. Consistent accelerated inference via confident adaptive transformers. *arXiv preprint arXiv:2104.08803*.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.593

Burr Settles. 2009. In *Active learning literature survey*.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. 2021. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34:6216–6228.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.262

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. *Mining Multi-label Data*. Springer US, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4_34

Dennis Ulmer, Chrysoula Zerva, and Andre Martins. 2024. Non-exchangeable conformal language generation with nearest neighbors. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1909–1929, St. Julian's, Malta. Association for Computational Linguistics.

Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. 2019. Towards better confidence estimation for neural models. In *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339. https://doi.org/10.1109/ICASSP.2019.8683359

Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore. PMLR.

Vladimir Vovk. 2015. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28. https://doi.org/10.1007/s10472-013-9368-4

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. 2018. Cross-conformal predictive distributions. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 37–51. PMLR.

Vladimir Vovk and Ivan Petej. 2014. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 829–838, Arlington, Virginia, USA. AUAI Press.

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. 2017. Nonparametric predictive distributions based on conformal prediction. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 82–102. PMLR.

Jun Wang, Guocheng He, and Yiannis Kantaros. 2024. Safe task planning for language-instructed multi-robot systems using conformal prediction. *arXiv preprint arXiv:2402.15368*.

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

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696. https://doi.org/10.1162/tacl_a_00483

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329. https://doi.org/10.1609/aaai.v33i01.33017322

Chen Xu and Yao Xie. 2021. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating llm hallucinations via conformal abstention.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. 2022. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 8622–8641. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.591`

Chrysoula Zerva and André F. T. Martins. 2023. Conformalizing machine translation evaluation. *arXiv preprint arXiv:2306.06221*.

Xianghao Zhan, Fanjin Wang, and Olivier Gevaert. 2022. Reliably filter drug-induced liver injury literature with natural language processing and conformal prediction. *IEEE Journal of Biomedical and Health Informatics*, PP:1–9. `https://doi.org/10.1109/JBHI.2022.3193365`

Xianghao Zhan, Qinmei Xu, Yuanning Zheng, Guangming Lu, and Olivier Gevaert. 2023. Reliability-based cleaning of noisy training labels with inductive conformal prediction in multi-modal biomedical data mining. *arXiv preprint arXiv:2309.07332*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38. `https://doi.org/10.1145/3639372`

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering.

Thomas Zollo, Todd Morrill, Zhun Deng, Jake Snell, Toniann Pitassi, and Richard Zemel. 2023. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *Socially Responsible Language Modelling Research*.