# Conformal LLM Multi-label Text Classification with Binary Relevance Approach

**Viktor Örnbratt** VIKORN@STUDENT.CHALMERS.SE
*Chalmers University of Technology, Sweden and Algorithma AB, Sweden*
**Johan Hallberg Szabadváry** JOHAN.HALLBERG@ALGORITHMA.SE
*Algorithma AB, Sweden*

**Editor:** Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

## Abstract

Large Language Models (LLMs) are increasingly deployed in real-world Natural Language Processing (NLP) systems to perform multi-label classification tasks, such as identifying multiple forms of toxicity in online content. However, most models output raw probabilities without an exact way to quantify uncertainty, increasing the risk of misclassification in high-stakes applications. In this work, we integrate Inductive Conformal Prediction (ICP) with the Binary Relevance (BR) approach to produce statistically valid prediction sets, label-wise. Using a modified Wikipedia Toxic Comments dataset, we evaluate this framework across varying significance levels ($\epsilon$), incorporating calibration-set-aware thresholds to address label imbalances.

Our results show that BR-based conformal prediction maintains valid marginal coverage while enabling flexible control over prediction set size (efficiency). Even in the presence of rare labels, the framework provides practical uncertainty estimates and where the prediction can be abstained in uncertain cases via empty sets. These findings support the feasibility of BR-ICP-based uncertainty calibration for scalable, interpretable automation in multi-label NLP systems.

**Keywords:** multi-label conformal prediction · binary relevance conformal prediction · natural language processing · large-language models · multi-label text classification

## 1. Introduction

The recent advances in LLMs have revolutionized the field of NLP, enabling quick deployment of systems handling a plethora of NLP tasks such as summarization, sentiment analysis, translation or text-classification. All classification problems including text classification has three subtypes: binary, multi-class and multi-label. Binary between two classes and multi-class being all cases with over 2 classes with a single class as the output. Multi-label classification on the other hand is a specific case as each instance can belong to multiple categories, thus instantly making it more complex than binary or multi-class classification with a single predicted class output. Multi-label text classification has several use-cases within data management such as document tagging in a database for better lookup, labelling customer-emails for relevant department review or content moderation by classifying the type of toxicity at work as shown in the dataset used within this work.

When implementing an LLM-based pipeline for classification purposes, the issue of label uncertainty becomes particularly important. While generative tasks often suffer from

problems such as hallucinations, classification tasks are more prone to misclassification or over-prediction in multi-label settings, One method to capture this uncertainty is the use of conformal predictions (Vovk et al. (2022)), a relatively recent method that is model agnostic to the underlying classifier. The previous research regarding natural language processing with conformal prediction have been quite extensive as shown by Campos et al. (2024) with their overview of various types of conformal prediction methods for different NLP tasks. For multi-label classification with conformal prediction also called Multi-Label Conformal Prediction (MLCP) there are three main approaches: Instance reproduction, Label Power Set and Binary Relevance each with their advantages and disadvantages.

This work focuses on integrating Inductive Conformal Prediction with the Binary Relevance approach using a fine-tuned Large Language Model for multi-label text classification. Our contributions are threefold: (1) we propose a practical framework that combines BERT-based classification with label-wise conformal predictors; (2) we adapt the conformal calibration thresholds based on label-specific calibration set sizes to address severe label imbalance; and (3) we empirically demonstrate that this method produces valid, interpretable, and abstaining predictions even for rare labels, highlighting its applicability in safety-critical NLP settings.

### 1.1. Related work

Multi-label conformal prediction has previously been considered in NLP pipelines, with one of the closer examples being Borg et al. (2019), where both a multi-class and a multi-label setting were explored using customer support e-mails. In their study, they used support vector machines as the underlying classifier, applied to a bag-of-words representation of the messages. The authors evaluated both the multi-label and multi-class case with conformal predictors, demonstrating the potential for calibrated confidence estimates in real-world e-mail systems.

Further exploration of conformal methods for multi-label text classification was carried out in Paisios et al. (2019), where a deep neural network was combined with the Label Power Set (LPS) transformation. Their work showed that conformal predictors could be effectively integrated with neural architectures, achieving well-calibrated and efficient predictions across multiple labels. This approach enabled the handling of complex label dependencies while still maintaining the theoretical guarantees of conformal prediction.

More recently, Maltoudoglou et al. (2023) extended this line of research by directly incorporating large language models into the conformal prediction pipeline . Their method employed a fine-tuned BERT model along with the LPS transformation to address multi-label classification with a large label space. Special attention was given to improving the calibration of the resulting confidence sets, demonstrating significant improvements in both empirical validity and prediction set efficiency.

## 2. Preliminaries

In the following section we will quickly overview inductive conformal predictors and how multi-label conformal prediction is approached, before defining the main evaluation metrics for the study.

### 2.1. Inductive Conformal Predictors

In this era of LLMs it is natural to be worried regarding the issue of uncertain outputs due to hallucinations when implementing models in production environments. An emerging method to help quantify this uncertainty is conformal prediction, which by using previously viewed data points allow the construction of future predictions sets that contain the true value with a certain probability for error.

The standard assumption in CP is that data are drawn from a probability distribution that is exchangeable, meaning in essence that any permutation of the data is equally probable. For details, see Vovk et al. (2022)

Using the definition 2 given in sect 4.2.2: of Vovk et al. (2022)

**Definition 1** *Inductive Conformal Predictor*

*The training set of size c is first split into two parts: the proper training set* $Z_{train}$ : $z_1, ..., z_m$ *of size* $m < c$ *and the calibration set of size* $Z_{cal} : z_{m+1}, ...z_c$ *of size* $c - m$. *For every test object* $x_i, i = c + 1, ..., c + k$, *compute the prediction sets*

$$\Gamma^{\epsilon}(z_1, ..., z_c, x_i) := \left\{ y \in \mathcal{Y} : \frac{|\{j = m + 1, ..., c : \alpha_j \geq \alpha_i\}| + 1}{c - m + 1} > \epsilon \right\} \tag{1}$$

*where the nonconformity scores are defined by*

$$\alpha_j := A((z_1, ..., z_m), z_j), \qquad j = m + 1, ..., c$$

$$\alpha_i := A((z_1, ..., z_m), (x_i, y)),$$

$\alpha_j$ measures how well each calibration point conforms to the model trained on the training set, For the test point $x_i$ and a hypothetical label $y$, $\alpha_i$ is the nonconformity score if we assume the test point $x_i$ had label $y$. Lastly $\mathcal{Y}$ is the set of all possible labels which in the binary relevance case described in coming section case is 0,1 indicating 1 for the existence of the label for the instance $x_i$ and 0 for the label being false or rejected for the instance.

Due to the the common issue of small calibration subset that are frequent when looking at rare labels within unbalanced datasets, the significance level of the conformal predictor has to be adjusted based on this limitation to ensure validity. This is presented in corollary 4.9 in Vovk et al. (2022) with the corresponding equation for adjustment of $\epsilon$ to $(E, \delta)$ can be seen in Equation (2).

$$E := \epsilon + \sqrt{\frac{\ln \frac{1}{\delta}}{2h}} \tag{2}$$

Where if $\Gamma$ is an inductive conformal predictor, it is $(E, \delta)$-valid when E is described by Equation (2) as a joint packed probability together with $\delta$. Where $\epsilon$ is the pre-adjusted significance level and h is the size of the calibration set used for the adjustment. For an intuitive example of this joint packed probability imagine that you are buying a product that creates a secondary item with a guarantee of 1-E. The bought product then has a probability $\delta$ of being faulty and not having the 1-E guarantee for the secondary item.

### 2.2. Multi-label Conformal Prediction

AS previously mentioned there are currently three approaches of Multi-label Conformal Prediction: Instance reproduction (IR), Label Power Set (LPS) and Binary Relevance (BR) explored in this work and Borg et al. (2019). Very briefly the implementations are as follows: Instance reproduction creates the non-conformity scores for the conformal framework based on comparison between the entire true label vector against the predicted vector via metrics such as Hamming loss or another dissimilarity metric. Label Power Set converts the multi-label problem into a multi-class problem where each different permutation of the label vector is instead treated as an unique class and follows multi-class conformal. Binary Relevance divides the multi-label problem into binary classification tasks. A more in depth explanation regarding the different approaches can be seen in Wang et al. (2015). This work follows the binary relevance approach which will be explained further in the following subsection.

### 2.3. Binary Relevance Conformal Prediction and Calibration Size Adjustment

Binary Relevance Conformal Prediction (BRCP) approaches the multi-label problem by splitting it into $\ell$ number of binary conformal classification problems. This then means we construct a list of binary conformal predictors $[\{\Gamma_1, ..., \Gamma_\ell\}]$ for each label within the dataset.

Then to ensure that these binary conformal predictors are valid with the imbalanced dataset the chosen significance value is adjusted to each conformal predictor with Equation (2). This creates a new list of conformal predictors with a fixed $\delta$ where only calibration set size h varies across predictors: $\{\Gamma(E_1, \delta), ..., \Gamma(E_\ell, \delta)\}$ which accounts coverage validity for each label.

### 2.4. Evaluation metrics

For evaluation of the of the LLMs predictive performance we use two types of accuracies, subset accuracy and label-wise recall. Subset accuracy in the multi-label case is a strict metric as it requires an exact match between the true vector of labels and the predicted for each text instance. In the following equations $I$ represents the indicator function, $\hat{Y}$ the predicted multi-hot label vector, $Y$ the ground-truth multi-hot label vector and i is the index for iterating across all the text instances N. Precision, Recall and F1-score being averaged across all labels when comparing True Positive (TP), False Positive (FP) and False Negative (FN) predictions on the validation set.

$$Acc_s = \frac{1}{N}\sum_{i=1}^{N} I[\hat{Y}_i = Y_i] \tag{3}$$

$$Precision = \frac{1}{L \cdot N}\sum_{i=1}^{N}\sum_{\ell=1}^{L} \frac{TP_i^\ell}{TP_i^\ell + FP_i^\ell} \tag{4}$$

$$Recall = \frac{1}{L \cdot N}\sum_{i=1}^{N}\sum_{\ell=1}^{L} \frac{TP_i^\ell}{TP_i^\ell + FN_i^\ell} \tag{5}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Looking at conformal prediction, we also use coverage (how often the true label set is a subset of the predicted label set), efficiency (average size of the prediction label set) Vovk et al. (2022), Jaccard index (for set similarity) and Hamming loss (for average per-label instance prediction errors) Borg et al. (2019). Here $\hat{y}_i^\ell$ is the predicted label, $y_i^\ell$ the ground truth label, $\Gamma^\ell(x_i)$ the binary conformal predictor for label $\ell$ and efficiency being the average sum of predicted positive labels within the predicted label set. Jaccard index finally compares the intersection between the predicted label set and the ground truth label set with a value of 1 indicating a perfect match.

$$Cov_{ML} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{\ell=1}^{L} I(y_i^\ell \subseteq \Gamma^\ell(x_i)) \tag{7}$$

$$Effiency = \frac{1}{N} \sum_{i=1}^{N} |\Gamma(x_i)| \tag{8}$$

$$HL = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{\ell=1}^{L} I[\hat{y}_i^\ell \neq y_i^\ell] \tag{9}$$

$$J_{acc} = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

## 3. Method

### 3.1. Data

The data used is available at Carlos (2018). The data within the comment toxicity dataset consists of an instance identifier, comment text and 6 toxicity labels with binary encoding for six different labels: toxic, severe toxic, insult, obscene, threat & identity hate. It was further modified for this work by iterating over all comments and adding an additional non-toxic label with value 1 at all instances where the sum of toxic labels were 0. Thereby creating the possibility for the conformal framework to produce the two types of reject sets an empty set and full set.

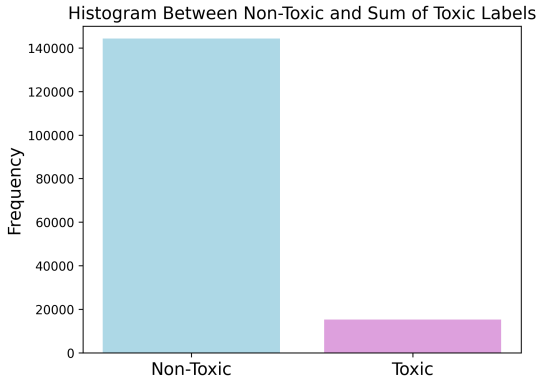



Figure 1: Occurrence of non-toxic vs all types of toxicity across entire dataset

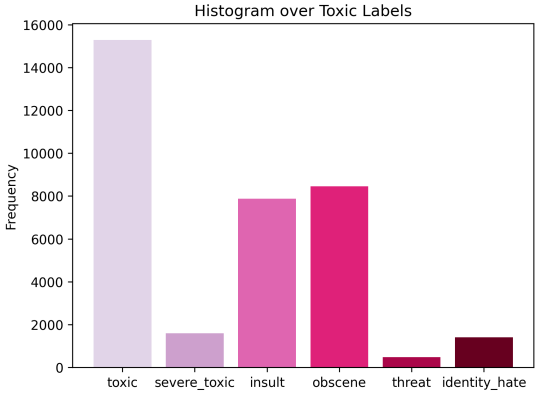



Figure 2: Toxic-label occurrence across entire dataset

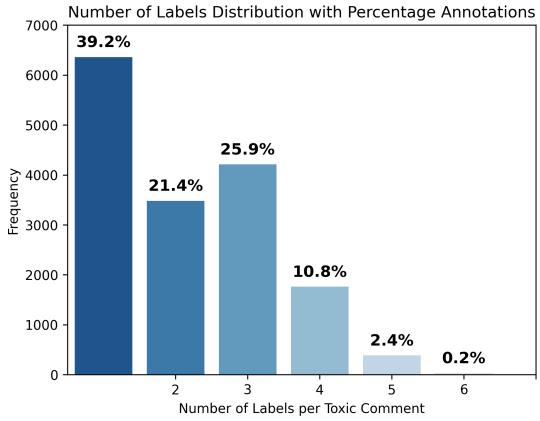



Figure 3: Histogram over number of labels for different toxic comments.

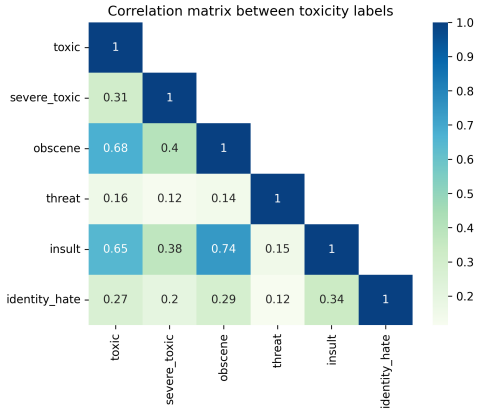



Figure 4: Correlation matrix over toxicity labels for the toxic-comments

Two things to note about the dataset in general in accordance to the figures above is that it is severely unbalanced with a nine to one ratio of non-toxic to toxic comments (including all labels of toxicity). The expected efficiency when averaging the ground-truth labels across full test dataset is 1.114, with a value 2.124 when viewing only toxic comments.

### 3.2. Nonconformity scores

Non-conformity scores are computed by applying the sigmoid function to the model's label-wise logits and taking one minus the resulting probabilities. Given these scores, label-wise thresholds $\tau$ are determined using a calibration set. A label is accepted if its non-conformity

score falls below $\tau$, where $\tau$ is implicitly defined as the smallest value satisfying

$$\frac{|\{j = m+1, \ldots, c : \alpha_j \geq \tau\}| + 1}{c - m + 1} \leq \epsilon,$$

which corresponds to the discrete $(1 - \epsilon)$-quantile of the calibration scores. This ensures that the marginal coverage exceeds $1 - \epsilon$.

### 3.3. Prediction set handling

Each binary conformal predictor $\Gamma_\ell$ will have four possible outputs: $\{\}, \{0\}, \{1\}$ and $\{0, 1\}$ with {} and {0,1} showcasing uncertainty of the predictor for the label instance. In this study, both the empty set {} and the full set {0, 1} are treated as 'rejection cases', indicating uncertainty about label inclusion or exclusion. It is represented in the code as the decision rule for the non-conformity threshold ($\tau_\ell$ where a label is accepted if ($\alpha_\ell^i < \tau_\ell$) and rejected ($\alpha_\ell^i > \tau_\ell$). The uncertain case is described ($\alpha = \tau_\ell \pm tol$) with a soft band tolerance (tol) around the non-conformity threshold where the model will be considered uncertain.

This uncertain case is dealt via modifying the predicted multi-hot vector to include None-types for instances where there is specific label uncertainty. This also means that the equations described in Subsection 2.4 have been modified slightly in the code for handling these cases.

### 3.4. Jaccard index with labels sets

Looking back at the Jaccard index defined in Subsection 2.4 with the multi-hot encoded labels Jaccard index with labels being: non toxic, toxic, severe toxic, insult, obscene, threat, identity hate. In this case of label sets this can be visualized in the Figure 5. An example calculation of Jaccard with the true label vector set being [0, 0, 1, 1, 0, 0, 1] and the predicted vector set after applying conformal prediction being [0, 0, 1, 1, 1, 1, 0] would end up as:

$$A \cap B = \frac{[0, 0, 1, 1, 0, 0, 1]}{[0, 0, 1, 1, 1, 1, 0]} = 2, \quad A \cup B = [0, 0, 1, 1, 1, 1, 1] = 5, \quad J_{acc} = \frac{2}{5} = 0.4$$

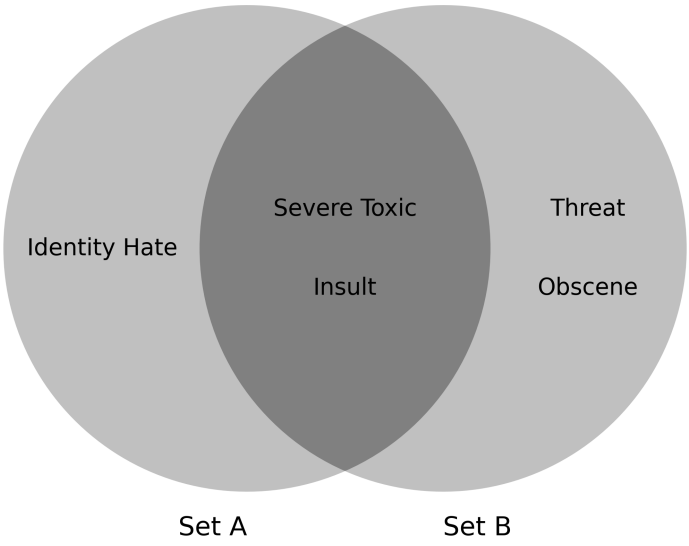



Figure 5: Venn diagram over two possible vector sets in the data.

### 3.5. LLM Text Classifier Model

The LLM used for this work was the BERT-base uncased 110 M parameter variant Devlin et al. (2018). It was implemented via the transformers package and predictions were made using the "TextClassificationPipeline" after the model was first fine-tuned and validated on the Wikipedia dataset.

Before settling on the BERT model both Roberta and Electra models was considered and fine-tuned using the same hyper-parameters. Although RoBERTa and ELECTRA have been shown to outperform BERT in certain benchmark tasks Liu et al. (2019), BERT performed better within the evaluated data and was also chosen as the primary model due to its stability, interpretability, and widespread adoption. The models were implemented as described by the psuedo code in Alg 1

**Algorithm 1:** Setup of the Multi-Label LLM Text Classification Model

**Input:** Multi-label text dataset $D$
**Output:** Trained model with best validation performance

1. **Split data** into:
   - Training set (60%), Validation set (15%), Calibration set (15%), Test set (10%)

   Use toxicity-based stratification to account for label imbalance.
2. **Compute label weights** from the training set to adjust for label imbalance when training the model.
3. **Preprocess labels** by converting multi-label annotations into multi-hot encoded vectors.
4. **Initialise LLM-based classifier** with:
   - Sigmoid activation function for multi-label output
   - Binary Cross Entropy (BCE) loss function
5. **Train the model** over several epochs:
   (a) Train on the training set using label-weighted BCE loss
   (b) Evaluate on the validation set
   (c) Save model weights after each epoch
6. **Select the best model** by extracting parameters from the epoch with the lowest validation loss.

## 4. Results

The results are divided into two sections, the first focusing on a quick overview of underlying model text classification performance with the second focused on evaluating conformal prediction framework.

### 4.1. LLM Multi-label classification metrics

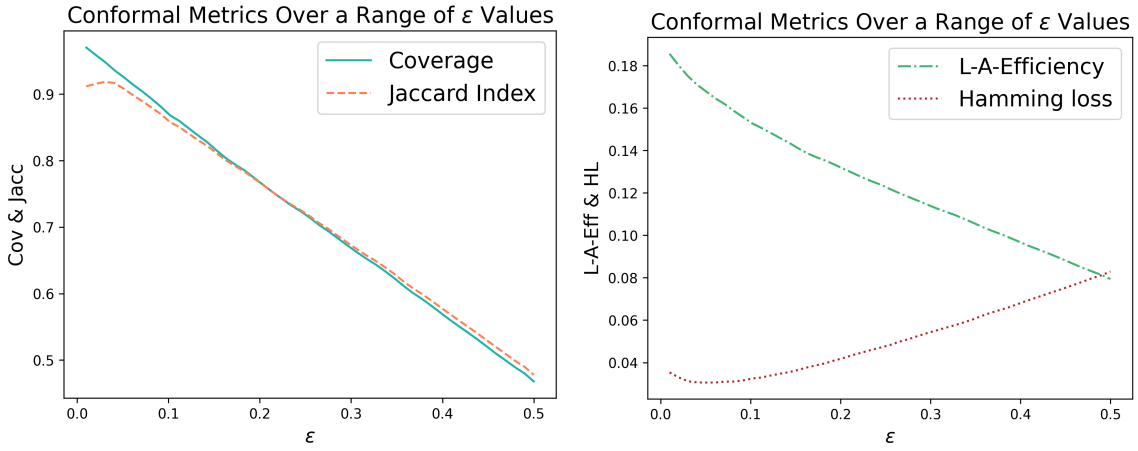
The model used for prediction in the following two sections was loaded using the tensors of Epoch 2 shown in Table 1 due to lower validation loss and higher macro accuracy in comparison to shown epochs and continual over-fitting of the model for training epochs extended further. The training for these epochs can be seen in Figure 8 in Appendix 1

Table 1: Multi-label evaluation metrics across three training epochs. Subset accuracy and F1, recall, and precision scores are computed using macro averaging across all labels. The evaluation loss is presented in natural logarithmic scale.

| Epoch | Eval Loss (ln) | Subset Accuracy | F1-score | Recall | Precision |
|---|---|---|---|---|---|
| 1 | -5.986 | 0.919 | 0.687 | 0.653 | 0.758 |
| 2 | -6.004 | 0.925 | 0.702 | 0.655 | 0.778 |
| 3 | -6.000 | 0.922 | 0.726 | 0.729 | 0.725 |

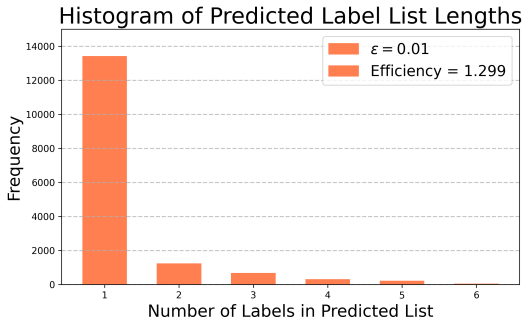
### 4.2. Binary Relevance Conformal Prediction Evaluation

Here our significance level $\epsilon$ in the range 0.01 to 0.5 is adjusted based on the size of label-wise calibration sets with a predefined $\delta$. Making the set of conformal predictors valid in $E_1, ..., E_l$. The effect of this shift for a few $\epsilon$ values is can be seen in Table 2 with respect to each label. Only significance levels up to 50% are considered as as any higher lead to conformal predictors with a confidence level below 50% pre calibration size adjustment.

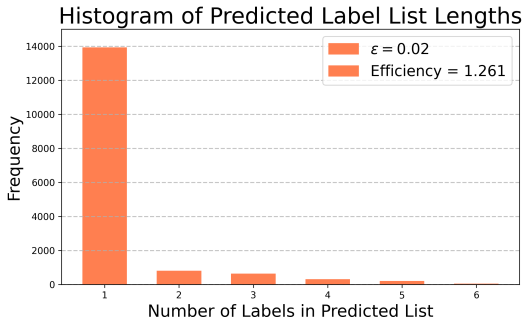



(a) Coverage & Jaccard Index (b) L-A-Efficiency & Hamming loss

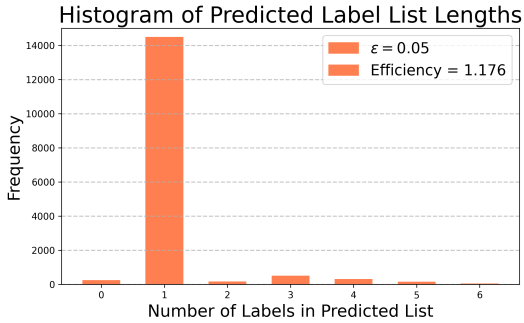
Figure 6: Coverage, Jaccard Index, Label-Averaged Efficiency and Hamming loss across significance levels from 0.01 to 0.5 with a step size of 0.01 and with each $\epsilon$ value adjusted based on Equation (2) with respect to label specific calibration set size and a delta value of 0.05.



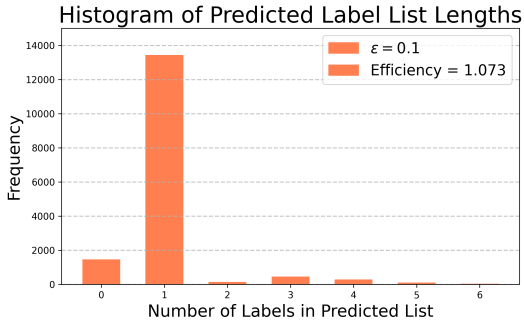

(a) Histogram of prediction set sizes for $\epsilon = 0.01$



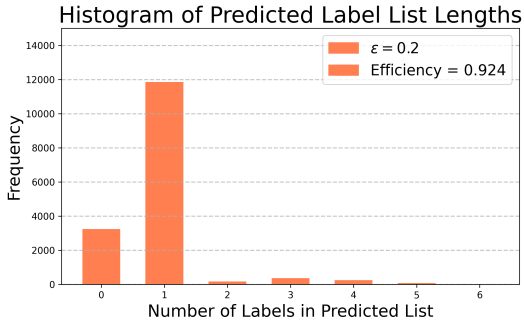

(b) Histogram of prediction set sizes for $\epsilon = 0.02$



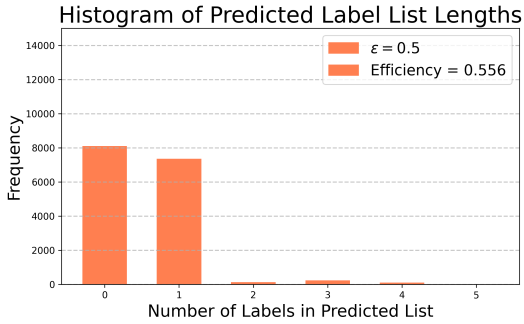

(c) Histogram of prediction set sizes for $\epsilon = 0.05$




(d) Histogram of prediction set sizes for $\epsilon = 0.1$




(e) Histogram of prediction set sizes for $\epsilon = 0.20$




(f) Histogram of prediction set sizes for $\epsilon = 0.50$

Figure 7: Histograms over set size occurrence for six difference $\epsilon$ values in the range 0.01 and 0.5

Table 2: Table of E values for adjusted validity across the label-wise calibration sets with a fixed $\delta$ value 0.05. The adjustment term $\sqrt{(\ln \frac{1}{\delta})/(2h)}$ for significance level seen next to $\epsilon$ in the RHS of Equation (2) and h being the calibration set size for each respective label and conformal predictor.

| label | Non Toxic | Toxic | Severe Toxic | Insult | Obscene | threat | Identity Hate |
|---|---|---|---|---|---|---|---|
| h | 21515 | 2294 | 259 | 1200 | 1263 | 62 | 209 |
| $\sqrt{(\ln \frac{1}{\delta})/(2h)}$ | 0.0083 | 0.0255 | 0.0761 | 0.0353 | 0.0344 | 0.1554 | 0.0846 |
| $\epsilon$ | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ | $E_6$ | $E_7$ |
| 0.01 | 0.0183 | 0.0355 | 0.0861 | 0.0453 | 0.0444 | 0.1654 | 0.0946 |
| 0.02 | 0.0283 | 0.0455 | 0.0961 | 0.0553 | 0.0544 | 0.1754 | 0.1046 |
| 0.05 | 0.0583 | 0.0755 | 0.1261 | 0.0853 | 0.0844 | 0.2054 | 0.1346 |
| 0.1 | 0.1083 | 0.1255 | 0.1761 | 0.1353 | 0.1344 | 0.2554 | 0.1846 |
| 0.2 | 0.2083 | 0.2255 | 0.2761 | 0.2353 | 0.2344 | 0.3554 | 0.2846 |
| 0.5 | 0.5083 | 0.5255 | 0.5761 | 0.5353 | 0.5344 | 0.6554 | 0.5846 |

## 5. Discussion

### 5.1. Exchangeability and Label Drift

An important assumption made for conformal prediction is the exchangeability between the calibration and test data. This assumption may not hold due to when the Wikipedia comments are manually labelled by different annotator and label drift may occur as they shift their threshold for what is considered toxic over time. In addition contextual context such as previous replies with regards to sarcasm or similar can negatively affects the predicted labelling of the model as it lacks this prior comment context.

### 5.2. Handling Rare Labels and Minimum Significance

In Table 2 we can clearly see the issue that commonly occurs within MLCP with the rare label threat only appearing 62 times in the calibration set. This in turn implies a minimum significance level value for the binary predictor for the label at around 15.5% for valid calibration. In very high-stake settings this seems very large but in practical industrial cases these rare predictions can be delegated to manual handling. A good example of this would be in Borg et al. (2019) where they chose to prune any e-mail labels with less than a 1000 examples to prioritize automation of the common cases. The results we gotten in this work align with this where binary relevance-MLCP allows for handling the edge cases of uncertain predictions on infrequent where it can be potentially deferred to human handling.

### 5.3. Empty Predictions and Set Size Behaviour

As shown in Figure 7, for lower significance levels $\epsilon : 0.01, 0.02$ prediction sets always include at least one label per instance. Only at $\epsilon \geq 0.05$ does empty prediction vectors $[0, 0, ..., 0]$ start appearing showcasing that CP under strict confidence abstains from guessing. Addi-

tionally this can be seen in Figure 6($a$) as the Jaccard Index peaks around $\epsilon = 0.05$ before declining therefore indicating a point of balance between accuracy and uncertainty.

### 5.4. Jaccard Index and Hamming Loss

To evaluate the quality of the conformal prediction sets beyond marginal coverage, we also consider the Jaccard Index and Hamming loss. As shown in Figure 6($a$), the Jaccard Index peaks near $\epsilon = 0.05$ before gradually declining aligning itself with coverage as the level of significance increases. As $\epsilon$ increases, the prediction sets become broader and less specific, leading to reduced set similarity.

Hamming loss, illustrated in Figure 6($b$), increases steadily with higher $\epsilon$ values with a very slight valley near the corresponding peak for Jaccard Index although much flatter. As it increases it indicates that more individual label predictions diverge from the ground truth as the prediction sets grow larger. In contrast, Label-Averaged Efficiency (L-A-Efficiency) decreases with $\epsilon$, reflecting the decreasing average number of predicted labels per instance. These opposing trends showcase the trade-off between efficiency and accuracy in conformal prediction: tighter sets reduce coverage but improve precision, while looser sets ensure higher coverage at the cost of more label noise.

### 5.5. Conclusion

This work set out to investigate the integration of conformal prediction techniques with large language models for multi-label text classification, focusing specifically on toxic comment detection. The key objective was to quantify and manage predictive uncertainty in high-stakes NLP tasks using Inductive Conformal Prediction (ICP) under the Binary Relevance (BR) approach.

Our main contributions are threefold:

1. We demonstrated that BR-based conformal predictors, when paired with a fine-tuned BERT model, produce *statistically valid* and *interpretable* prediction sets, even under class imbalance.

2. We incorporated *calibration-set-aware adjustments* for the significance level $\epsilon$, based on the theoretical guarantees proposed by Vovk et al. (2022), and validated these across varying label frequencies.

3. We showed how *empty and full prediction sets* can be used as rejection indicators, allowing for abstention in uncertain cases—thereby aligning automated decisions with real-world risk tolerance.

Empirical results confirmed that this framework maintains *valid marginal coverage* while still offering flexible control over *prediction set efficiency*. The combination of label-wise calibration and soft rejection thresholds proved especially useful for managing *rare labels*, such as `threat` and `identity hate`, where traditional classifiers often may fail silently.

Overall, our findings reinforce the practical value of conformal prediction for LLM-based multi-label classification in safety-critical applications such as content moderation. The model not only performs well but also conveys *when it does not know*, a critical property for deployment.

### 5.6. Future Work

Several directions remain open for extending this work and strengthening its applicability in broader NLP settings.

First, future studies could compare the Binary Relevance approach used here with alternative multi-label conformal prediction frameworks, such as Label Power Set (LPS) and Instance Reproduction (IR), as presented in Wang et al. (2015). Such comparisons, when focused specifically on text classification tasks, could offer further insights into the trade-offs between label dependency modelling and calibration performance.

Second, the incorporation of *human-in-the-loop* feedback systems presents a promising avenue. In real-world applications, prediction set abstention (via empty or full sets) could trigger manual review pipelines. Integrating reviewer feedback into the conformal framework may improve model calibration over time and reduce unnecessary abstentions.

Third, the adoption of a *Mondrian conformal prediction* approach in the multi-label setting could help mitigate the effects of label imbalance more effectively by conditioning validity on subgroups defined by label frequency or input features. This could provide better coverage control, particularly for the rare labels.

Finally, to further validate and generalise the claims made in this study, it is essential to evaluate the proposed framework on a broader range of multi-label NLP datasets. Applying the method to domains such as medical text classification, legal document tagging, or social media analysis would help assess its robustness and scalability in diverse real-world contexts.

## References

Anton Borg, Martin Boldt, and Johan Svensson. *Using Conformal Prediction for Multi-label Document Classification in e-Mail Support Systems*, page 308–322. Springer International Publishing, 2019. ISBN 9783030229993. doi: 10.1007/978-3-030-22999-3_28. URL http://dx.doi.org/10.1007/978-3-030-22999-3_28.

Margarida M. Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey, 2024. URL https://arxiv.org/abs/2405.01976.

DJ Carlos. Exploring the toxicity of wikipedia comments. Kaggle Notebook, 2018. URL https://www.kaggle.com/code/djcarlos/exploring-the-toxicity-of-wikipedia-comments.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

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

Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. Well-calibrated confidence measures for multi-label text classifi-

cation with a large number of labels. 2023. doi: 10.48550/ARXIV.2312.09304. URL https://arxiv.org/abs/2312.09304.

Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. A deep neural network conformal predictor for multi-label text classification. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Evgueni Smirnov, editors, *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pages 228–245. PMLR, 09–11 Sep 2019. URL https://proceedings.mlr.press/v105/paisios19a.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World, Second Edition*. Springer International Publishing, January 2022. ISBN 9783031066481. doi: 10.1007/978-3-031-06649-8. Publisher Copyright: © Springer Verlag New York, Inc. 2005.

Huazhen Wang, Xin Liu, Ilia Nouretdinov, and Zhiyuan Luo. *A Comparison of Three Implementations of Multi-Label Conformal Prediction*, page 241–250. Springer International Publishing, 2015. ISBN 9783319170916. doi: 10.1007/978-3-319-17091-6_19. URL http://dx.doi.org/10.1007/978-3-319-17091-6_19.

## Appendix A. Model Training

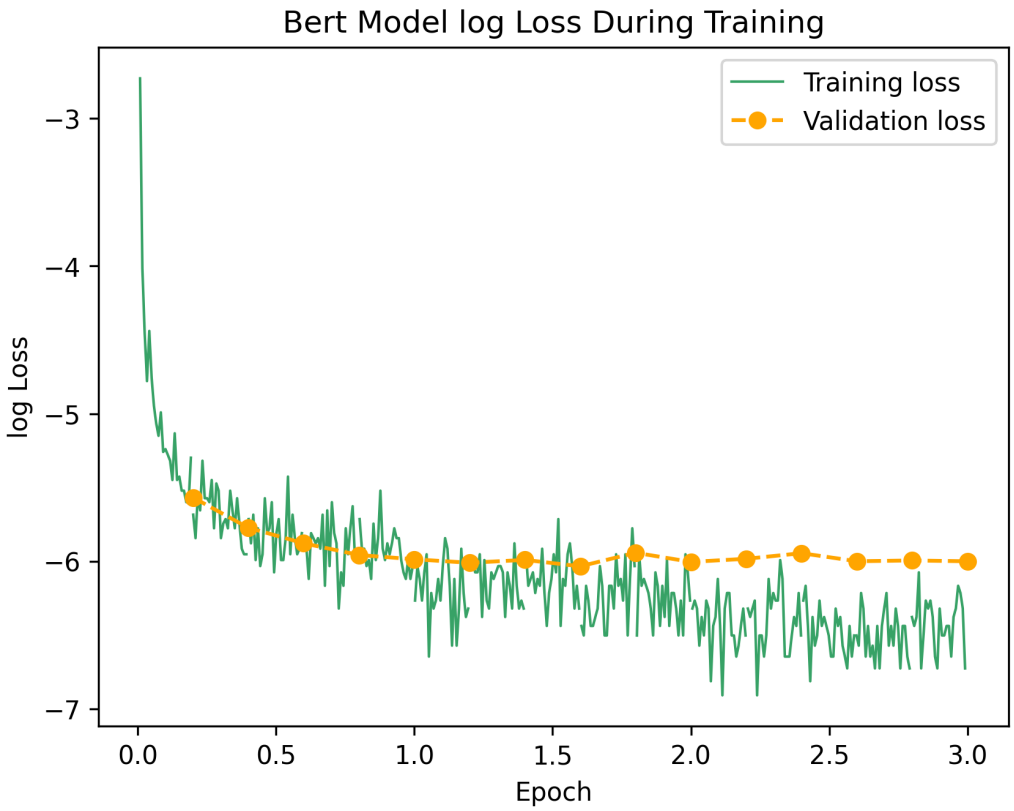



Figure 8: Natural log-scale training and validation loss over three epochs for a `bert-base-uncased` model. Training used binary cross-entropy and incorporated label weights via a custom wrapper.

The plot shows the performance of the `bert-base-uncased` model (110M parameters), trained with label-weighted binary cross-entropy using a custom wrapper. Calculation of validation loss was performed five times per epoch.