Estimating Uncertainty in Multimodal Foundation Models using Public Internet Data

Shiladitya Dutta* UC Berkeley
Hongbo Wei* UC Berkeley
Lars van der Laan University of Washington
Ahmed M. Alaa UC Berkeley and UCSF

Abstract

Foundation models are trained on vast amounts of data at scale using self-supervised learning, enabling adaptation to a wide range of downstream tasks. At test time, these models exhibit *zero-shot* capabilities through which they can classify previously unseen (user-specified) categories. In this paper, we address the problem of quantifying uncertainty in these zero-shot predictions. We propose a heuristic approach for uncertainty estimation in zero-shot settings using *conformal prediction* with web data. Given a set of classes at test time, we conduct zero-shot classification with CLIP-style models using a prompt template, e.g., *"an image of a <category>"*, and use the same template as a search query to source calibration data from the open web. Given a web-based calibration set, we apply conformal prediction with a novel conformity score that accounts for potential errors in retrieved web data. We evaluate the utility of our proposed method in Biomedical foundation models; our preliminary results show that web-based conformal prediction sets achieve the target coverage with satisfactory efficiency on a variety of biomedical datasets.

Code: https://github.com/AlaaLab/WebCP, * Equal Contribution.

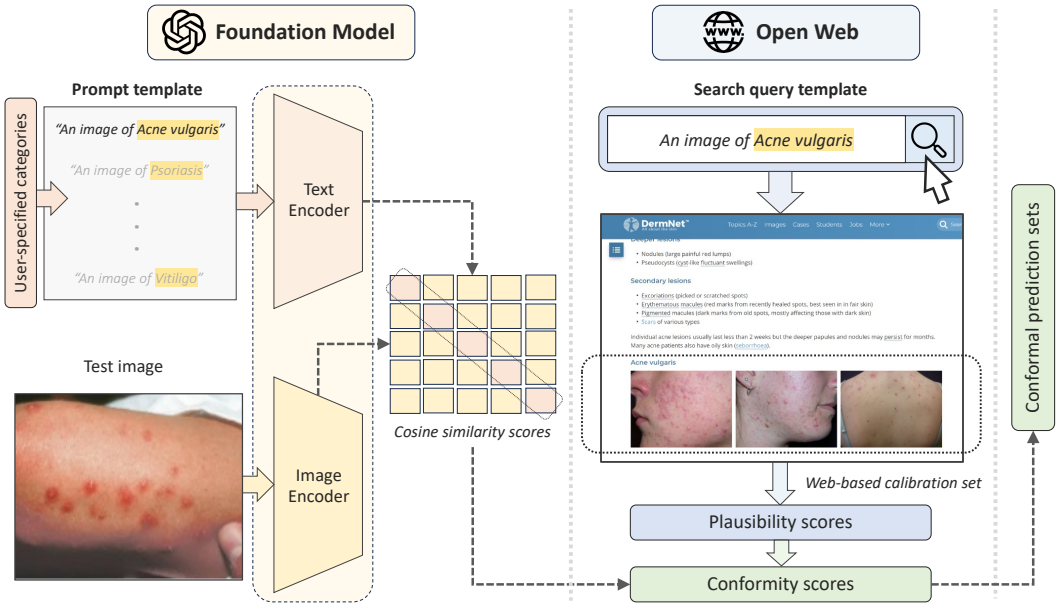

Figure 1: **Illustration of our web-based conformal prediction procedure:** Given a test image and user-specified categories, we apply conformal prediction using "on-the-fly" calibration sets obtained from the open web.

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

1 Introduction

Foundation models can be repurposed to tackle domain-specific use cases through zero-shot learning (ZSL) guided by task descriptions. In the ZSL paradigm, a pretrained foundation model is used to classify data according to previously unseen (user-specified) categories without the need for additional labeled examples for these categories. A common approach to ZSL is *contrastive pretraining*, whereby encoders are trained to embed images and language into a shared, low-dimensional latent space representing the semantic correspondence between visual and text data (Rethmeier & Augenstein (2023)). Models pretrained using this approach, such as CLIP, achieved remarkable performance in a broad range of computer vision benchmarks under zero-shot settings (Radford et al. (2021); Cherti et al. (2022); Zhang et al. (2023b)). **The goal of this paper** is to develop methods for estimating uncertainty in zero-shot predictions based on pretrained foundation models. In line with the zero-shot nature of these models, we seek versatile uncertainty estimation methods that can handle arbitrary user-specified categories and do not require access to labeled data specific to the task at hand.

Our proposed method for uncertainty estimation in foundation models is based on *conformal prediction* (Vovk et al., 2005; Angelopoulos & Bates, 2022)—a rigorous framework for predictive inference that can operate on top of any black-box model while providing distribution-free guarantees on the coverage of its resulting prediction sets. The standard split conformal prediction (CP) procedure assumes access to a "calibration set" comprising labeled examples from the downstream task of interest. CP constructs prediction sets by evaluating the model errors on the held-out calibration set, and adjusting the set size so that it contains the true class in $1 - \alpha$ of the calibration examples. If the calibration and testing data are drawn from the same distribution, CP is guaranteed to achieve a coverage probability of $1 - \alpha$ in new test samples (Vovk et al. (2005)). However, in ZSL settings we do not have access to a predetermined set of categories, labeled calibration examples for these categories, or a known test distribution. Thus, to utilize CP for estimating uncertainty in foundation models within the ZSL setup, we need to develop new variants of the CP procedure that operate in a zero-shot fashion.

Contributions. In this paper, we develop a novel heuristic for zero-shot CP that operates by calibrating CLIP-based foundation models using data from the open web. The intuition behind our approach is that, in absence of predetermined classification categories and labeled calibration data, the internet can serve as a queryable source of universal calibration data that encompasses all possible user-specified categories and provides a reasonable approximation for downstream data distributions. Our web-based CP procedure operates at test time through the following steps: given a set of user-specified classes, we conduct zero-shot classification with CLIP-style models using a prompt template, e.g., *"an image of a <category>"*, and use that prompt as a search query to source class-specific calibration data from the internet (Fig. 1). Given the retrieved web-based calibration images and their associated contexts (i.e., HTML meta-data), we develop a procedure for generating "plausibility scores" that account for possible content or context errors in web search. These plausibility scores generated for each image and its associated context are heuristics for the probabilities of whether each particular user-given class is the ground-truth label for the image and its corresponding context. We then use these plausibility scores to conduct the Monte Carlo-based CP procedure proposed in (Stutz et al. (2023)), through which we obtain prediction sets that capture the most likely classes for each test image. We evaluate the accuracy of our proposed method in multiple domain-specific image classification tasks, focusing on Biomedical datasets. We show that web-based CP empirically achieves target coverage while retaining efficiency comparable to that of an "oracle" CP procedure that uses data from target datasets for calibration.

2 Related Work

Given its model- and distribution-free properties, CP has been successfully applied to a wide range of applications, including calibration of Large Language Models (LLMs) ((Kumar et al., 2023)) and traditional computer vision tasks (Andéol et al., 2023). The closest work to ours is (Kumar et al. (2022)), which also developed a zero-shot CP-based approach by framing the classification task as an outlier detection problem. This work proposes a different approach to zero-shot calibration—it assumes access to a held-out calibration set of image-caption pairs (drawn from the target distribution), and proposes novel definitions of conformity scores based on the cosine similarity distance between images and captions. (Fisch et al. (2021)) proposes a few-shot CP approach that assumes availability of labeled calibration data for multiple related tasks, and exploits task similarity to derive prediction sets for

new tasks for which limited data is available. Our CP procedure is based on the work in (Stutz et al. (2023)), which proposes a calibration procedure for settings where the ground-truth is ambiguous. This procedure is where the concept of "plausibility scores" that reflect the level of ambiguity in labels originate from—our work proposes an incarnation of these scores for web-based calibration samples instead of based on MLE of expert opinion.

3 Problem Formulation

Setup and objectives. We consider a classification problem where a context $X \in \mathcal{X}$ is assigned an associated label $Y \in \mathcal{Y}$. Here, we define the label space $\mathcal{Y}$ as a user-specified, finite set of possible categories relevant to an application of interest. Similarly, $\mathcal{X}$ represents the set of all possible contexts that could be attributed to some label in $\mathcal{Y}$. In vision-language foundation models, the context space $\mathcal{X}$ could encompass all possible images that are compatible with some text label, or caption, within a specified set $\mathcal{Y}$. Suppose we have access to a zero-shot multi-probability classifier $\tau : \mathcal{X} \mapsto \Delta_{\mathcal{Y}}$, i.e., a foundation model, where $\Delta_{\mathcal{Y}} \subset \mathbf{R}^{\#|\mathcal{Y}|}$ represents the probability simplex over $\mathcal{Y}$. In other words, for a given context $x \in \mathcal{X}$, the classifier output $\tau(x)$ is a vector of probabilities, with each entry corresponding to the likelihood of each label $y \in \mathcal{Y}$ being the "true" label for x. An inherent challenge in ZSL is that it often requires probability predictions for out-of-distribution contexts and labels in $\mathcal{X} \times \mathcal{Y}$, which correspond to novel applications that did not appear in the training data of the foundation model τ. Consequently, given a context $x \in \mathcal{X}$, the interpretation and usage of the multi-probability prediction $\tau(x) := \{\tau(y \mid x) : y \in \mathcal{Y}\} \in \Delta_{\mathcal{Y}}$ is unclear, as the reference distribution P derived from the training data may not accurately or meaningfully reflect the novel prediction task of interest. To address this challenge, it is of interest to quantify uncertainty in zero-shot predictions through an application-specific prediction set $x \mapsto \widehat{C}(x) \subset \mathcal{Y}$ that contains the true label with high probability. Specifically, for an application-specific "ground-truth" probability distribution π over $\mathcal{X} \times \mathcal{Y}$, the prediction set $\widehat{C}$ should satisfy the marginal coverage probability of $P_\pi(Y \in \widehat{C}(X)) = 1 - \alpha$ for $\alpha \in (0, 1)$, where P_π is the distribution of a test point $(X, Y) \sim P_\pi$.

Conformal prediction. When access to labeled data from the test distribution P_π is available, CP can be used to derive prediction sets $\widehat{C}$ for which the coverage condition $P_\pi(Y \in \widehat{C}(X)) = 1 - \alpha$ is guaranteed to hold (Vovk et al., 2005). The standard (split) CP procedure operates through the following steps. Given a labeled calibration set $\{(X_i, Y_i)\}_i$, we compute a *conformity score* $V_i(\tau(X_i), Y_i)$ that measures the deviation of the prediction of the foundation model from the true label. A prediction set is then constructed by computing an empirical quantile of the conformity scores obtained from the labeled calibration set $\{V_i\}_i$. The definition of the conformity score depends on the application of interest. In the ZSL setup, we do not have access to the test distribution P_π through a labeled calibration set $\{(X_i, Y_i)\}_i$, hence we cannot apply off-the-shelf CP. In the next section, we propose a zero-shot variant of CP in which the calibration set $\{(X_i, Y_i)\}_i$ is sourced from the internet.

4 Web-based Conformal Prediction with Foundation Models

We propose an "on-the-fly" approach to CP in ZSL using data from the open web. As the open web contains images and contextual information from diverse open sources (e.g., social media, photo galleries, scientific reporting), it serves as a rich and easily-accessible universal calibration dataset that is applicable to most domains. Our CP procedure, dubbed WebCP, involves the steps below.

Step 1: Calibration data mining. Given a user-specified set of categories $\mathcal{Y}$, the first step in WebCP is to acquire a set of image and web meta-data corresponding to these categories, i.e.,

$$\mathcal{C}_{\text{web}} = \{(\widetilde{X}_i^{y_i}, M_i^{y_i}) : i \in [1, \ldots, \textstyle\sum_{k=1}^{\#(\mathcal{Y})} n_{y_k}]\}. \tag{1}$$

Here, for each sample i in $\mathcal{C}_{\text{web}}$, $\widetilde{X}_i^{y_i}$ is an image corresponding to the class y_i downloaded from a web source (e.g., Google search) and $M_i^{y_i}$ is its accompanied web meta-data (e.g., textual content of the web page source HTML). We obtain the images $\{\widetilde{X}_i^{y_i}\}_i$ by querying a web source for each class $y_k \in \mathcal{Y}$ with a template query that depends on the particular class y_k, collecting the first n_{y_k} image search results for each $y_k \in \mathcal{Y}$, and aggregating them together to construct the calibration set in (Equation 1). Note our use of potentially different template prompts; we use a template prompt (e.g., *"An image of y_k"*) to evaluate the predictions of a CLIP-based foundation model, and a template search query (which could be the same as the template prompt) to source calibration data from the internet.

For instance, if y_k is a skin condition such as *Acne Vulgaris* in a dermatology application, then the prompt and search query can be *"An image of Acne Vulgaris"* (Fig. 1). The meta-data $M_i^{y_i}$ associated with image $\widetilde{X}_i^{y_i}$ comprises textual context associated with the web page from which the image is obtained. In our setup, we extract textual meta-data retrieving the content of the `<alt>` tag and the text around the `<img>` header within the HTML source code of the web page hosting the image $\widetilde{X}_i^{y_i}$. Full details of our web scarping procedure are provided in the Supplementary Appendix (A.2).

Step 2: Plausibility Generation. For each mined image in (1), we estimate a collection of "plausibility" scores defined as $\boldsymbol{\lambda}_i = \{\lambda_i^y : y \in \mathcal{Y}\}$. Each plausibility score λ_i^y in the vector $\boldsymbol{\lambda}_i$ is estimated based on the meta-data $M_i^{y_i}$ and the contents of the image $\widetilde{X}_i^{y_i}$, and reflects the likelihood that the retrieved image $\widetilde{X}_i^{y_i}$ belongs to the each of the particular user-specified classes $y \in \mathcal{Y}$. Our definition of plausibility scores captures two forms of alignment between the search query and retrieved results: **context alignment** and **content alignment**. The context alignment between a sampled image $\widetilde{X}_i^{y_i}$ and any class $y \in \mathcal{Y}$ reflects the relevance of the source web page and its meta-data for that image $M_i^{y_i}$ to the search query for y (*"An image of y"*). We quantify context alignment through a heuristic that assesses the relevance of the textual content of the source web page to the provided search query. On the other hand, content alignment reflects the relevance of the image embedded in the web page to the query corresponding to the class y. Content misalignment occurs when the image doesn't generically fit the search query, or if the image is topically accurate but is displayed in an undesired form (e.g., a cartoon illustration or a diagram of the queried class y). We quantify content alignment via Content-Based Image Retrieval (CBIR) methods (Müller et al. (2001)).

Each of the plausibility scores contained in the vector $\boldsymbol{\lambda}_i = \{\lambda_i^y : y \in \mathcal{Y}\}$ generated for each image $\widetilde{X}_i^{y_i}$ are computed by combining two components: context alignment scores and content alignment scores. The context alignment score measures context alignment through an algorithm $\texttt{context}(M_i^{y_i}, y)$; this algorithm evaluates the relevance of the web page meta-data $M_i^{y_i}$ to a given class y by feeding it into a text encoder (Sentence-BERT, Reimers & Gurevych (2019a)) to obtain sentence-level embeddings of $M_i^{y_i}$ which are then compared to the search query embedding (label name) of y via cosine similarity. For each image $\widetilde{X}_i^{y_i}$, the resulting scores for the image across all classes $y \in \mathcal{Y}$ are normalized using a temperature-tuned softmax to obtain context alignment scores $\{c_i^y : y \in \mathcal{Y}\}$ for that image $\widetilde{X}_i^{y_i}$.

To evaluate content alignment scores, we run each image through an algorithm $\texttt{content}(X_i^{y_i}, y)$ to evaluate the probability that the retrieved image is not depicting the queried category. This algorithm includes two stages: (i) a filtering stage for detecting when an image is topically relevant but not of the correct form such as images of diagrams or charts, and (ii) a scoring stage to check if the image contains a general visual representation of a class $y \in \mathcal{Y}$.

For the filtering stage, we utilized CLIP to generate similarity scores between the image and a series of invalid form prompts (e.g. *"an image with a lot of text"*, *"an image of a graph"*) alongside a "negative" label (*"an image"*) used to filter out "invalid" images. The idea is that if the image is similar to one of the invalid form prompts (which were formulated through experimentation) then it is of an invalid form. As such, we take the softmax of the resulting scores and use the negative label score s_{neg_i} as the probability that the image is of a valid form. Note that s_{neg_i} is generated for each image without considering the classes in $\mathcal{Y}$; this is because the negative label score reflects the likelihood of an image being of an "invalid" form which does not depend on any of the classes $y \in \mathcal{Y}$.

For the scoring stage, we wish to analyze image-based content alignment and detect content misalignment errors for the image across each class $y \in \mathcal{Y}$; we do this by applying CLIP to a relaxed version of the original classification task (i.e. determining if $\widetilde{X}_i^{y_i}$ conforms to a class $y \in \mathcal{Y}$) to detect if the image doesn't fit the high-level visual archetype of the label given for y. To do this we first have to generate a set of simplified variants of the class labels $\{y : y \in \mathcal{Y}\}$ that act as generic visual representations of a target label. Ideally these generalized variants are a) high-level enough to ensure that the zero-shot classifier can distinguish them with a high probability and b) representative of the obvious visual features in the image. To obtain this simplified variant y_{pseudo} for a corresponding $y \in \mathcal{Y}$, we use a named entity recognition model to identify all entities in a tag, reference these entities to an ontology (e.g. Dbpedia, Auer et al. (2007); MeSH, Rogers (1963); etc.), and then substitute the entity with a term from higher in the hierarchy. For instance, if $y =$*"colorectal adenocarcinoma epithelium"* then a simplified label could be *"microscope image"*. Once the generalized labels are derived, we generate softmax scores for each simplified label in comparison to the negative label *"an image"*, generating a set of content alignment scores $\{h_i^y : y \in \mathcal{Y}\}$ for each image $\widetilde{X}_i^{y_i} \in \widetilde{\mathcal{C}}_{\text{web}}$.

To integrate context/content alignments in generating plausibility scores $\boldsymbol{\lambda}_i = \{\lambda_i^y : y \in \mathcal{Y}\}$ corresponding to a sample $(\widetilde{X}_i^{y_i}, M_i^{y_i})$ that represents the likelihood of that sample belonging to classes $y \in \mathcal{Y}$, we take a pairwise product of the context alignment scores $\{c_i^y : y \in \mathcal{Y}\}$ and the content alignment scores $\{h_i^y : y \in \mathcal{Y}\}$, and scale each of the computed products by s_{neg_i}; thus, we have $\lambda_i^y = c_i^y \, h_i^y \, s_{\text{neg}_i}$ for each $y \in \mathcal{Y}$. Then, to reflect the possibility of a sampled image being irrelevant to any of the classes $y \in \mathcal{Y}$, we subtract the summation of the generated class probabilities λ_i^y across all $y \in \mathcal{Y}$ from 1.0 to derive a so-called "junk probability". To summarize, we have:

$$\lambda_i^y = h_i^y \, c_i^y \, s_{\text{neg}}, \;\; \lambda_{\text{junk}_i} = 1 - \sum_{i=1}^{n} \lambda_i^y. \tag{2}$$

The resulting classifier scores are then used as estimates of content plausibility. We utilize the plausibility scores to construct the final calibration set as follows:

$$\widetilde{\mathcal{C}}_{\text{web}} = \{(\widetilde{X}_i^{y_i}, \boldsymbol{\lambda}_i, \lambda_{\text{junk}_i}) : i \in [1, \ldots, n_y]\}. \tag{3}$$

An overview of this discussion and the `context` and `content` algorithms we choose to use are provided in relevant figures in the Appendix (A.1).

Algorithm 1 Web-based Conformal Prediction (WebCP)

Input: User-specified classes $\mathcal{Y}$; coverage $1-\alpha$; Monte Carlo samples M; foundation model τ; test image X

1. **Calibration data mining:** Download images and meta-data pairs $\{(\widetilde{X}_i^{y_i}, M_i^{y_i})\}_i$ from an online source using a search query template filled with the categories in $\mathcal{Y}$.
2. **Plausibility generation:** Estimate context and content alignment of the scraped images $\{\widetilde{X}_i^{y_i}\}_i$ using the `context`$(M_i^{y_i}, y)$ and `content`$(\widetilde{X}_i^{y_i}, y)$ algorithms to estimate the plausibility scores $\{(\boldsymbol{\lambda}_i, \lambda_{\text{junk}_i})\}_i$.
3. **Monte Carlo CP:** Perform sampling M times, where on each iteration $m \in [1, \ldots, M]$:
   - Iterate through each calibration example $(\widetilde{X}_i^{y_i}, \boldsymbol{\lambda}_i, \lambda_{\text{junk}_i})$. Choose to reject the example with probability λ_{junk_i}, or keep it with probability $1 - \lambda_{\text{junk}_i}$. If kept, randomly sample a label $\widetilde{y}_i$ in $\mathcal{Y}$ from the distribution $\widetilde{y}_i \sim \text{Categorical}(\lambda_i^y : y \in \mathcal{Y})$, and add the example $(\widetilde{X}_i^{y_i}, \widetilde{y}_i^m)$ to the random calibration set for the current iteration. Our final random calibration set for this iteration will be $\widetilde{\mathcal{C}}_m = \{(\widetilde{X}_i^{y_i}, \widetilde{y}_i^m)\}_i$.
   - Using the aggregate calibration dataset $\widetilde{\mathcal{C}} = \{\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_M\}$ find the minimum threshold γ s.t.

$$\frac{1}{M} \sum_{m=1}^{M} \left[\frac{\sum_{m'=1}^{|\widetilde{\mathcal{C}}_{m'}|} \mathbb{1}\{V(\tau(\widetilde{X}_{m'}^{y}), \widetilde{y}_{m'}^{m}) \leq \gamma\} + 1}{|\widetilde{\mathcal{C}}_m| + 1} \right] > 1 - \alpha$$

   .
4. **Prediction set construction:** For a new test image X, $\mathcal{P}(X) = \{y \in \mathcal{Y} : V(\tau(X), y) \leq \gamma\}$.

Output: A prediction set $\mathcal{P}(X) \subseteq \mathcal{Y}$.

Step 3: Conformal Prediction with Ambiguous Ground-truths. The calibration set in (3) comprises a set of images for each class along with probabilistic (ambiguous) labels for memberships to each class in $\mathcal{Y}$. To construct CP-based prediction sets for new test images, we apply the Monte Carlo CP procedure proposed in (Stutz et al. (2023)), which accounts for ambiguity in ground-truth labels, to the calibration set $\widetilde{\mathcal{C}}_{\text{web}}$. For our conformity score, we use the softmax transformed cosine similarity score between the CLIP image embedding and the CLIP label/caption embedding. A pseudo-code for the overall WebCP procedure is provided in Algorithm 1.

The internet as a universal CP calibration set. One of the key advantages of CP is that it provides provable guarantees on the coverage of the prediction sets $\mathcal{P}$ in test data. However, WebCP is a heuristic that sources on-the-fly calibration data from a large knowledge base (i.e., the internet) in order to calibrate zero-shot predictions, without access to the test distribution of the downstream task. This means that the exchangeability assumption, a necessary condition for the CP coverage guarantees (Vovk et al. (2005)), no longer holds. Consequently, the central question of interest is whether WebCP can attain the desired target coverage levels—a question that now hinges on empirical validation. In other words, we would like to test if web-scraped calibration data $\widetilde{\mathcal{C}}_{\text{web}} \sim P_{\text{web}}$ is a good approximation of oracle calibration sets drawn from the test distribution $\widetilde{\mathcal{C}}^* \sim P_\pi$ across many domain-specific datasets. In the next Section, we test this hypothesis within the context of biomedical applications.

Table 1: Δ_{cov} = Difference between Target (1-α) and Achieved Coverage

	MedMNIST Microscopy Subset					FitzPatrick-17k				
	Calibration		Test			Calibration		Test		
α	Coverage	Efficiency	Coverage	Efficiency	Δ_{cov}	Coverage	Efficiency	Coverage	Efficiency	Δ_{cov}
WebCP										
0.10	0.9643	13.41	0.9823	18.28	8.23%	0.9253	75.51	0.9023	78.67	+0.20%
0.20	0.8791	10.18	0.8473	14.38	4.73%	0.8381	51.40	0.8007	54.96	+0.04%
0.30	0.7800	7.944	0.7236	11.26	2.37%	0.7497	37.15	0.7064	40.33	+0.64%
0.40	0.6939	6.53	0.6423	9.29	4.23%	0.6510	25.33	0.6028	27.90	+0.28%
0.50	0.5819	4.94	0.5163	6.81	1.63%	0.5445	16.46	0.4920	18.29	-0.79%
Standard CP with web-based calibration data										
0.10	0.9009	10.71	0.8713	15.05	-2.87%	0.9002	67.23	0.8738	70.64	-2.61%
0.20	0.8003	8.39	0.7503	11.90	-4.97%	0.8001	44.59	0.7585	48.05	-4.14%
0.30	0.7004	6.62	0.6470	9.43	-5.29%	0.7001	30.82	0.6565	33.72	-4.35%
0.40	0.6006	5.15	0.5316	7.10	-6.84%	0.6000	20.73	0.5492	22.97	-5.08%
0.50	0.5000	3.85	0.4317	5.20	-6.83%	0.5000	13.44	0.4433	14.90	-5.66%
Oracle CP with calibration on target data										
0.10	0.9180	11.28	0.9072	15.94	0.72%	0.9199	73.49	0.8933	76.62	-0.66%
0.20	0.8498	9.36	0.8031	13.28	0.32%	0.8380	51.53	0.8024	55.01	0.24%
0.30	0.7613	7.67	0.7053	10.93	0.53%	0.7530	37.59	0.7143	40.72	1.43%
0.40	0.6582	5.98	0.6001	8.52	0.11%	0.6520	25.43	0.6057	27.99	0.57%
0.50	0.5698	4.74	0.5045	6.60	0.45%	0.5557	17.24	0.5055	19.13	0.55%

5 Experiments

Experimental setup. We evaluate the WebCP procedure biomedical datasets with variants of CLIP as the underlying multimodal foundation model. Specifically, we evaluate the effectiveness of WebCP at generating efficient and prediction sets for the black-box BioMedCLIP model `microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224` (Zhang et al. (2023a)), which is a foundation model that uses PubMedBERT as the text encoder and is finetuned on biomedical tasks by pre-training on PMC-15M (Zhang et al. (2023b)). To estimate context alignment, we use the Sentence-BERT model `sentence-transformers/msmarco-bert-base-dot-v5` which was finetuned on MS-MARCO (Reimers & Gurevych (2019b); Bajaj et al. (2018))). For content alignment estimation, we use a variant of CLIP finetuned on the LAION dataset (Schuhmann et al. (2022)). The reason why we use different variants of CLIP for content alignment and classification is two-fold. Firstly, the simplified labels should be very generic (i.e. "dog", "island", etc.) and as such a base CLIP model, which has been shown to perform reasonably well in general image classification benchmarks such as ImageNet, should provide satisfactory estimates for this task (Radford et al. (2021)). Secondly, using different variants of CLIP helps to decorrelate errors between the classification and plausibility tasks, thereby decreasing the probability of overconfidence due to matched errors in the two tasks.

To acquire the calibration data, we use a Selenium-based web crawling agent and Google Custom Image Search Engine to query the dataset-specific image classes, and cache the images and corresponding captions from the top 50 search results (after filtering out those missing textual contexts). We compare our proposed WebCP method with two baseline procedures. The first is a *Standard CP* procedure applied to the mined calibration data, without accounting for ambiguity in image classes through the generated plausibility scores. The second is an *Oracle CP* procedure, which applies the standard CP calibration step to an equal number of samples drawn from the target dataset on which coverage is evaluated. We evaluate all baselines in terms of their achieved coverage on an unseen test set as well as their efficiency, i.e., the average size of the prediction set $|\mathcal{P}(X)|$.

Datasets. We consider two biomedical datasets: (1) `Fitzpatrick17k`, a dataset containing annotated medical images of 114 classes of skin conditions Groh et al. (2021), and (2) the `PathMNIST`, `BloodMNIST`, and `TissueMNIST` subsets from the `MedMNIST` dataset (Yang et al. (2023)), which contains an aggregation of low-resolution images under 25 classes of image types derived from studies in colon pathology, blood cell microscope, and kidney cortex microscope imaging.

Results. In almost all experiments, our WebCP procedure results in prediction sets that consistently achieves the targeted $1-\alpha$ coverage on the test datasets, and displays satisfactory calibration performance across varying levels of α. On the contrary, the Standard CP procedure applies to the scraped calibration data significantly under-covers in the test dataset across all values of α. This shows the

utility of our generated plausibility scores, which accounts for potential content and context errors in the retrieved web data. It also shows that, for the datasets under consideration, web-scraped data can provide useful calibration sets that can help calibrate the zero-shot predictions of foundation models. Compared to the Oracle CP, WebCP incurs a slight loss of efficiency, which is expected due to the conservative nature of calibration under ambiguous ground truth. However, the efficiency of WebCP remains comparable to the oracle efficiency for all values of α across the two datasets.

6 Conclusion

In this paper, we developed a method for estimating uncertainty in the zero-shot predictions of pre-trained foundation models. Our proposed heuristic estimates uncertainty in image classification using conformal prediction applied to web-scraped data in a zero-shot fashion. Our procedure, dubbed WebCP, comprises three steps: (1) mining calibration data based on user-specified classification categories, (2) estimating plausibility scores to quantify the alignment of the mined data with the user queries, and (3) applying a Monte Carlo-based approach to conformal prediction using the estimated plausibility scores in order to calibrate the predictions of foundation models. Our preliminary results show that WebCP could be a promising approach to zero-shot calibration—in biomedical datasets, it achieves the user-specified target coverage while retaining competitive efficiency on test data.

References

Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. *arXiv preprint arXiv:2304.06052*, 2023.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. (arXiv:2107.07511), December 2022. doi: 10.48550/arXiv.2107.07511. URL `http://arxiv.org/abs/2107.07511`. arXiv:2107.07511 [cs, math, stat].

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux (eds.), *The Semantic Web*, Lecture Notes in Computer Science, pp. 722–735, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-76298-0. doi: 10.1007/978-3-540-76298-0_52.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset. (arXiv:1611.09268), Oct 2018. URL `http://arxiv.org/abs/1611.09268`. arXiv:1611.09268 [cs].

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Beijing.; Cambridge Mass., 1st edition edition, Aug 2009. ISBN 978-0-596-51649-9.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. (arXiv:2212.07143), Dec 2022. URL `http://arxiv.org/abs/2212.07143`. arXiv:2212.07143 [cs].

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pp. 3329–3339. PMLR, 2021.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.

Bhawesh Kumar, Anil Palepu, Rudraksh Tuwani, and Andrew Beam. Towards reliable zero shot classification in self-supervised models with conformal prediction. (arXiv:2210.15805), Oct 2022. doi: 10.48550/arXiv.2210.15805. URL http://arxiv.org/abs/2210.15805. arXiv:2210.15805 [cs].

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

Henning Müller, Wolfgang Müller, David McG Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern recognition letters*, 22(5):593–601, 2001.

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. Learning transferable visual models from natural language supervision. (arXiv:2103.00020), Feb 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. (arXiv:1908.10084), Aug 2019a. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019b. URL http://arxiv.org/abs/1908.10084.

Nils Rethmeier and Isabelle Augenstein. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *ACM Computing Surveys*, 55(10):1–17, 2023.

F. B. Rogers. Medical subject headings. *Bulletin of the Medical Library Association*, 51(1):114–116, Jan 1963. ISSN 0025-7338.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. (arXiv:2307.09302), Jul 2023. URL http://arxiv.org/abs/2307.09302. arXiv:2307.09302 [cs, stat].

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing, 2023a. URL https://arxiv.org/abs/2303.00915.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. (arXiv:2303.00915), Mar 2023b. doi: 10.48550/arXiv.2303.00915. URL http://arxiv.org/abs/2303.00915. arXiv:2303.00915 [cs].

A Appendix

A.1 Overview

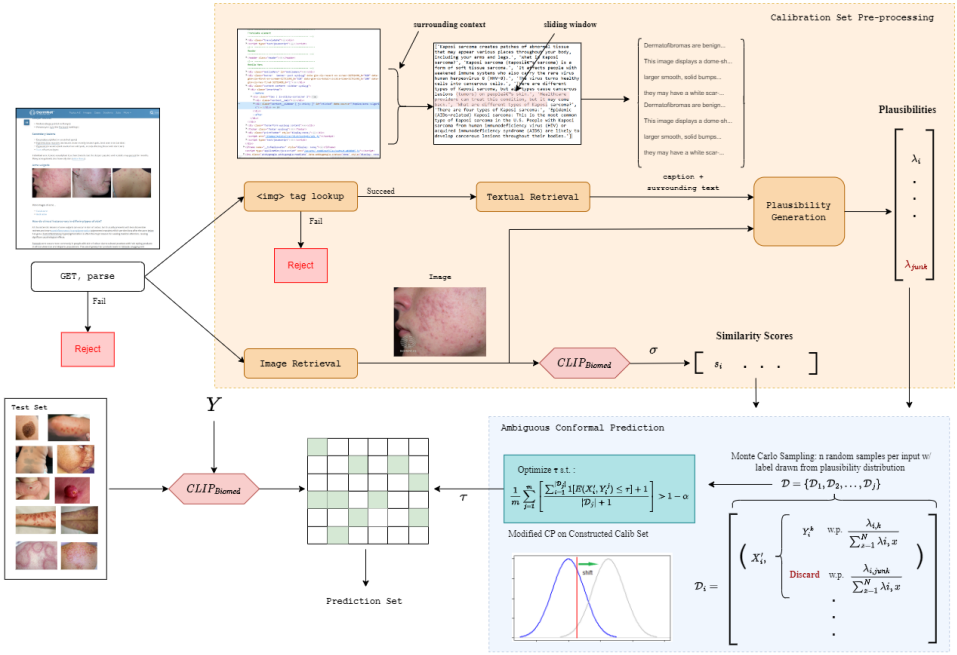

Figure 2: Figure portraying the total pipeline of data mining, plausibility generation, and conformal prediction

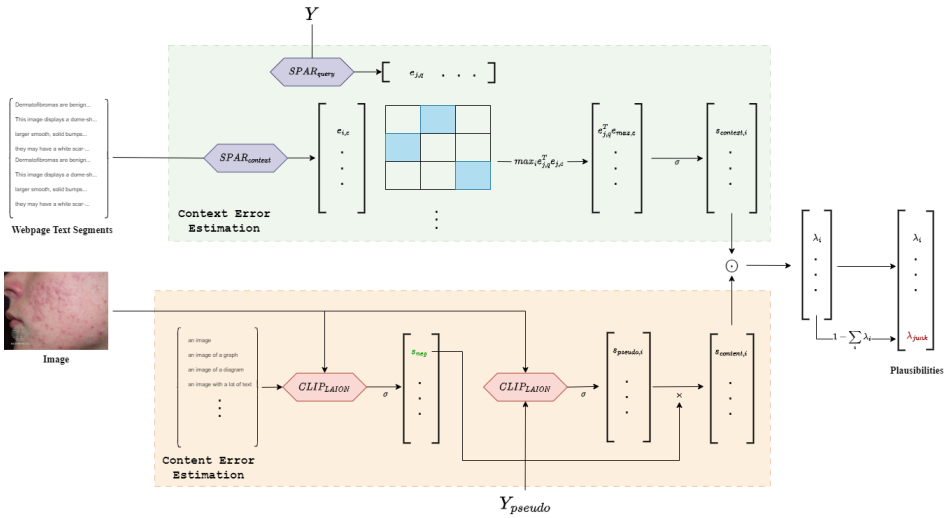

Figure 3: Figure portraying the hybrid context and content alignment based plausibility generation pipeline.

A.2 Data Mining

Algorithm 2 Web Scraping Procedure for Acquiring Internet Meta-Data ($\mathcal{C}_{web}$)

Given: A set of categories $\mathcal{Y}$, an image search engine $\mathcal{E}$ providing entries consisting of images, a corresponding URL to the image, and a corresponding URL to the image's web page source HTML; a target number of results per class, K.

1. For each category $y \in \mathcal{Y}$:
   (a) Perform a query on $\mathcal{E}$ for y to obtain a list of entries $\mathcal{L}_y$ with length $>> K$.
   (b) Until the target number of examples K is achieved, consider each entry in $\mathcal{L}$ consisting of an image (`image`), its URL (`image_url`) and a URL to its web page source (`context_url`). We find the location of the image in the web page linked to `context_url` by identifying the presence of the string `image_url` in that web page, and obtain the contextual meta-data immediately surrounding that location. Particularly:
      i. If `context_url` is unaccessible (e.g. timeout, blocked, or lazy loading) or there is no close match for the `image_url` in the `src` or `url-src` tag of any `<img>` div in the webpage, skip and continue. We define a close match for an `image_url` to be a `src` or `url-src` whose file-name (i.e. without file paths or arguments passed in with the URL) roughly matches the file-name in `url-src` ($> .85$ similarity, according to a metric on their similarity utilizing the `difflib.SequenceMatcher` library in Python (Van Rossum & Drake Jr (1995)).
      ii. Otherwise, we retrieve the `alt` tag for the matching `<img>` divider in the context HTML page, and obtain its plaintext (if it is in HTML format). We also retrieve the minimum of 256 plaintext tokens or 10 sentences (ending each divider as a separate sentence) from the text immediately before the pertinent `<img>` tag, and from the text immediately after the tag, performing sentencing using the natural language tooklit in Python (Bird et al. (2009)). We concatenate these surrounding plaintext results together.
      iii. We take `image` and all its concatenated surrounding plaintext sentences, and add these to $\mathcal{C}_{web}$ as an entry.

Output: a set of $K \cdot \#(\mathcal{Y})$ results for each of the $\#(\mathcal{Y})$ categories, with each result containing the image and its surrounding captions/images.

A.3 Experiments

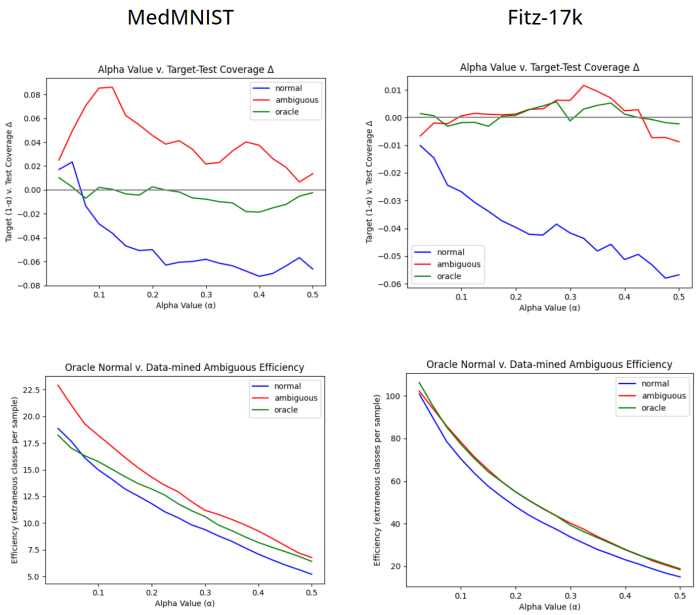

Figure 4: $\Delta_{coverage}$ and Efficiency for normal, oracle, and ambiguous CP across varying α values

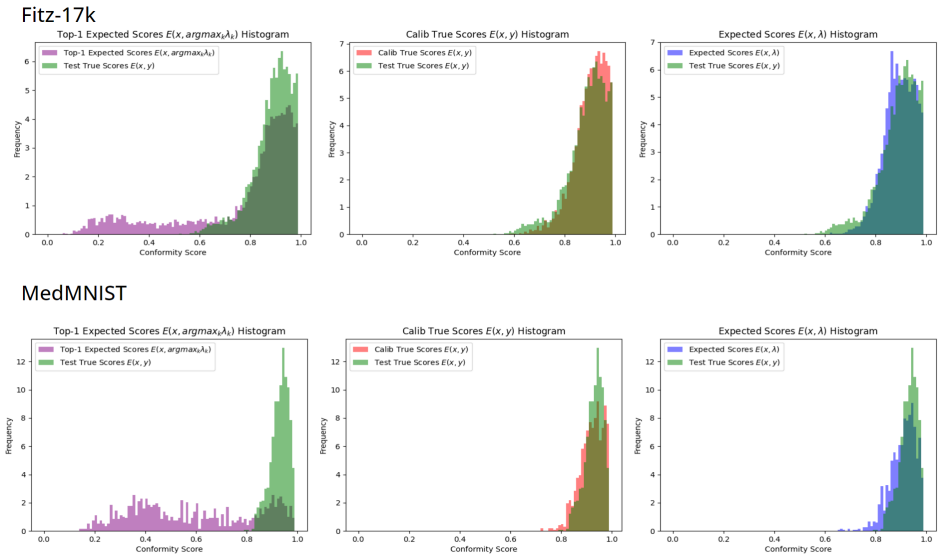

Figure 5: Conformity score distribution for FitzPatrick-17k and MedMNIST