FILTERING WITH CONFIDENCE: WHEN DATA AUGMENTATION MEETS CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

With promising empirical performance across a wide range of applications, synthetic data augmentation appears a viable solution to data scarcity and the demands of increasingly data-intensive models. Its effectiveness lies in expanding the training set in a way that reduces estimator variance while introducing only minimal bias. Controlling this bias is therefore critical: effective data augmentation should generate diverse samples from the same underlying distribution as the training set, with minimal shifts. In this paper, we propose conformal data augmentation, a principled data filtering framework that leverages the power of conformal prediction to produce diverse synthetic data while filtering out poor-quality generations with provable risk control. Our method is simple to implement, requires no access to internal model logits, nor large-scale model retraining. We demonstrate the effectiveness of our approach across multiple tasks, including topic prediction, sentiment analysis, image classification, and fraud detection, showing consistent performance improvements of up to 40% in F_1 score over unaugmented baselines, and 4% over other filtered augmentation baselines.

1 INTRODUCTION

Synthetic data augmentation refers to a set of machine learning techniques and heuristics designed to artificially expand a training dataset Shorten & Khoshgoftaar (2019); Taqi et al. (2018). As noted by Huang et al. (2022), practitioners have long relied on augmenting inputs with perturbed versions of the original data—both to enhance model robustness to small perturbations and based on the general intuition that "more data is always better." With the emergence of advanced foundation models capable of generating remarkably high-quality synthetic data however (from images (Ho et al., 2020; Karras et al., 2017; 2019; Ramesh et al., 2022; Rombach et al., 2022), to text (Brown et al., 2020; Li et al., 2022; Touvron et al., 2023), or molecular structures (Jin et al., 2018; Shi et al., 2020)), synthetic data generation has experienced renewed interest. Such approaches promise significant practical advantages, particularly in reducing the time, cost, and effort involved in augmenting datasets through additional data collection and annotation Nadas et al. (2025). Synthetic data augmentation has already demonstrated promising empirical results across a wide range of applications. In natural language processing, it has been effectively used for model fine-tuning on small datasets and in low-resource language settings (Feng et al., 2020; Yang et al., 2019; Li et al., 2020; Mahamud et al., 2023; Wang et al., 2022), as well as for knowledge base construction (Li et al., 2024b) , etc. In computer vision, it has shown benefits in tasks such as image classification (He et al., 2016; Li et al., 2025) and object detection (Bochkovskiy et al., 2020), etc.

From a theoretical perspective, much is still to determine about the benefits of synthetic data. Recent theoretical insights from Huang et al. (2022); Nakada et al. (2024) have begun characterizing the effect of synthetic oversampling in certain regimes on estimator error bounds. Intuitively, synthetic oversampling should work well if it manages to enlarge the training set, reducing estimator variance whilst only incurring a slightly increased bias. Synthetic augmentation methods thus face a fundamental tension. On one hand, generated samples should closely follow the distribution of the original data to minimize bias—typically requiring using lower variability in the generation (or a "low temperature") to ensure that the generated data remains faithful to the original. On the other hand, synthetic samples need to be sufficiently diverse and decorrelated to be treated effectively as independent observations, a goal typically achieved by increasing generation variability (e.g., raising the temperature parameter) Havrilla et al. (2024).

Despite the current enthusiasm for synthetic data sampled from generative AI models, no principled approach has yet been proposed to determine this trade-off systematically Jordon et al. (2022). In fact, current methods for generating synthetic data exhibit limited flexibility in their handling of samples with varying levels of quality. To adapt the loss to various levels of synthetic data quality, some techniques, such as the approach by Jaine et al. Jain et al. (2024) or that of Nakada et al. Nakada et al. (2024), introduce hyperparameters to control the weights placed on the reconstruction errors corresponding to the original data and the synthetic data respectively, effectively putting less emphasis on the synthetic data if its quality is too low. But these methods are inherently inflexible and treat all generated datapoints similarly. In particular, these methods are unable to distinguish between good and bad synthetic examples, thereby effectively discarding all synthetic data points from distributions that produce mixtures of high-quality and low-quality outputs Alaa et al. (2022); Rajeswar et al. (2023); Ravuri & Vinyals (2019) . Finer methods, capable of operating effectively in high-variability ("high-temperature") regimes and explicitly distinguishing high-quality generated samples from poor ones, are still lacking.

Contributions. To bridge this gap, we introduce a principled filtering approach that selectively retains high-quality outputs with provable guarantees. Our method operates as a wrapper around existing generative AI-based data augmentation frameworks, enabling their use in high-temperature (high-variability) settings, while ensuring the quality of the generated content through conformal risk prediction. Specifically, our contributions include:

1. A principled framework (Section 2) for evaluating the quality of generated content, consisting of two primary components:
   (a) A scoring function that quantifies the quality of generated samples.
   (b) A rejection threshold that specifies the minimum acceptable quality, calibrated using conformal risk prediction (Section 3.1).
2. Provable guarantees of control of our procedure over the number of poor quality samples accepted in the augmented data using approximate conditional coverage in our setting (Section 3.2). Our method adapts the framework of Gibbs et al. (2025); Cherian et al. (2024) to provide robust, condition-specific quality guarantees.

Our approach is practical and straightforward to implement, requiring neither access to internal model logits nor extensive retraining. To evaluate the validity and practical utility of our method, we demonstrate its application across three text-based use cases and further assess its performance on three tabular datasets and one image dataset (Section 4). Across these tasks, our method consistently yields measurable improvements in downstream applications, including text classification, sentiment analysis, fraud detection, and image classification.

2 Background: Synthetic Data Generation and Filtering

Let $h : \mathcal{X} \to \Omega$ denote a pretrained generative model (e.g., ChatGPT, Gemini or DALL·E, or any VAE-type of model fit to the data). Here $\mathcal{X}$ refers to a set of features on which to condition the generation, and Ω to the generation domain (e.g. space of images, documents, etc). While this paper mostly considers text and tabular data examples, our methodology can, in principle, extend to any domain where data can be generated using generative models. Consider a dataset $\mathcal{D} = \{X_i\}_{i=1}^{N}$, where each X_i corresponds to a sample point (i.e. a document or image) and N is the total number of samples. Our objective is to leverage h to create alternative versions of each data point X_i, thereby increasing the dataset size. This approach is particularly useful in low-sample scenarios, such as when the training dataset is small (Section 4.1), as a mitigator of extreme class imbalance (Section 4.2).

LLM-based Data Augmentation. Ding et al. (2024) categorize LLM-based data augmentation into four classes: data creation, data reformation, data labeling, and human-LLM co-annotation. Our work specifically focuses on data reformation, where existing data points are transformed to produce new examples or enrich existing data points. Historically, reformation methods relied predominantly on rule-based approaches, such as token perturbations or back-translation. However, recent advancements in generative models have enabled significantly more diverse augmentation strategies. In this paper, we propose using a generative model based augmentation method due to its demonstrated ability to produce greater generative diversity. A detailed literature review of LLM-based augmentations is provided in Appendix B.

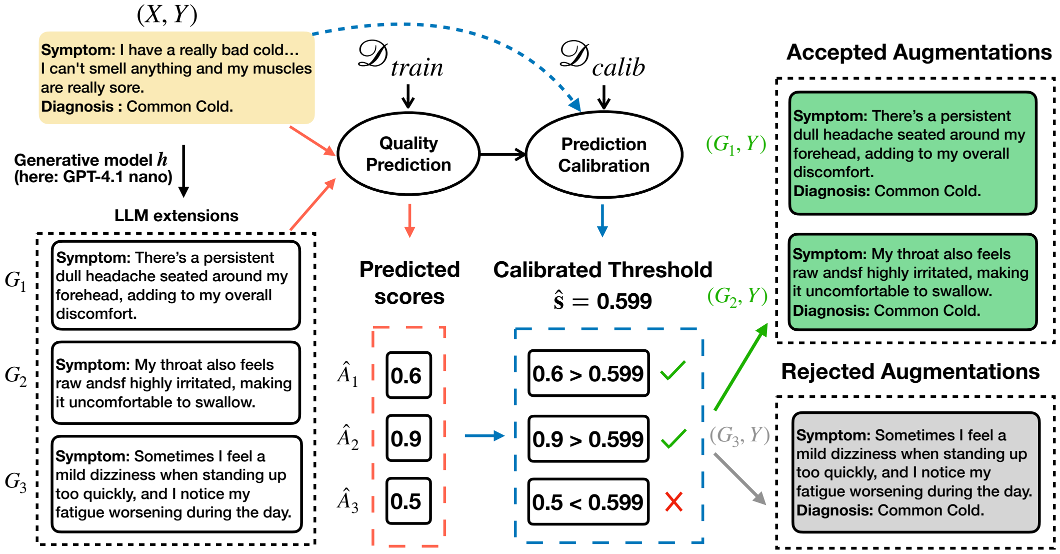
(X, Y)
Symptom: I have a really bad cold... I can't smell anything and my muscles are really sore.
Diagnosis : Common Cold.
Generative model h (here: GPT-4.1 nano)
LLM extensions
G1
Symptom: There's a persistent dull headache seated around my forehead, adding to my overall discomfort.
G2
Symptom: My throat also feels raw andsf highly irritated, making it uncomfortable to swallow.
G3
Symptom: Sometimes I feel a mild dizziness when standing up too quickly, and I notice my fatigue worsening during the day.
D train
Quality Prediction
D calib
Prediction Calibration
Predicted scores
Â1 0.6
Â2 0.9
Â3 0.5
Calibrated Threshold
ŝ = 0.599
0.6 > 0.599 ✓
0.9 > 0.599 ✓
0.5 < 0.599 ✗
(G1, Y)
(G2, Y)
(G3, Y)
Accepted Augmentations
Symptom: There's a persistent dull headache seated around my forehead, adding to my overall discomfort.
Diagnosis: Common Cold.
Symptom: My throat also feels raw andsf highly irritated, making it uncomfortable to swallow.
Diagnosis: Common Cold.
Rejected Augmentations
Symptom: Sometimes I feel a mild dizziness when standing up too quickly, and I notice my fatigue worsening during the day.
Diagnosis: Common Cold.

Figure 1: Illustration of the workflow in clinical disease prediction. Data augmentation candidate outputs from the generative model h (GPT-4.1 nano in this example) are filtered by a quality predictor trained on $\mathcal{D}_{train}$ with a threshold calibrated by $\mathcal{D}_{\text{calib}}$. The retained output preserves the meaning of "common cold," while the discarded output does not correspond to the intended symptom.

Formally, let $X_i \in \mathcal{D}$ denote an observed data point, $Y_i \in \mathcal{Y}$ denote additional sample meta information (such as labels captions) which we might want to condition upon in our generating procedure. We assume that the data point X_i is sampled from a true underlying distribution $h^\star$ that depends on the context/label: $X_i \sim h^\star(C_i, Y_i)$ where C_i represents the latent context. Intuitively, Y_i encodes observable attributes such as class labels or side information, and C_i captures hidden structure or nuisance variation specific to the dataset at hand, and that is not directly observed but decides how X_i is realized. To synthesize new instances from the same distribution, we generate K alternative versions of X_i by reusing X_i as proxy for the latent context C_i:

$$(G_{ik})_{k=1}^{K} \sim h(X_i, Y_i, \tau),$$

where h denotes the generative model conditioned explicitly on the observed data point and features (X_i, Y_i), and τ is a temperature parameter controlling the model's generation variability. Thus, the generative model h serves as an approximation to the true distribution $h^\star$, replacing the inaccessible latent context C_i with observable surrogates X_i, Y_i.

Evaluating Generation Quality Although effective, synthetic data from generative models can be noisy or distributionally shifted Feng et al. (2021); Kumar et al. (2020) — particularly when increasing the temperature τ—, potentially reducing downstream performance. Various approaches, such as prompt engineering, direct generative modeling, retrieval-based methods, and filtering strategies (e.g., human evaluation, similarity metrics, classification-based filtering), have been proposed to improve synthetic data quality Alaa et al. (2022); Lewis et al. (2020); Liu et al. (2023). However, these filtering methods critically depend on accurate and oftentimes expensive quality metrics (such as human evaluation), which remain challenging Ding et al. (2024); Rajeswar et al. (2023). With increased generation diversity, ensuring quality becomes critical.

We propose revisiting here a simple filtration technique, as proposed in Islam et al. (2024); Kang et al. (2021); Li et al. (2024a). These methods all operate on the following premise: low-quality generations should be filtered out. Let $\mathcal{A} : \Omega \times \Omega \times \mathcal{Y} \to \mathbb{R}$ be a measure of a generated sample's quality. Ideally, $\mathcal{A}$ should quantify the degree of deviation of the generation from the underlying data distribution. Filtering-based methods choose to remove generated examples for which $\mathcal{A}(G_{ik}, X_i, Y_i) < \lambda$, for a user-defined threshold λ. The threshold λ should be neither too low (to avoid content of low quality), nor too high (to avoid generating trivial rephrasings).

While this framework promises to improve the quality of data augmentation, it relies on access to a trustworthy evaluation metric $\mathcal{A}$. Choosing an unsuitable $\mathcal{A}$ can distort the training distribution. For example, simply measuring similarity between generated and original samples risks biasing the augmented data toward reproducing existing examples rather than capturing the broader distribution. Quality annotators might not necessarily exist, or if they do (e.g. human annotators in certain settings), they might be too expensive to deploy at scale. In the absence of gold-standard evaluations, the only option is to use a cheaper evaluator $\widehat{\mathcal{A}}$ (e.g. an LLM to evaluate text generations), thereby providing an imperfect, noisy surrogate for $\mathcal{A}$. Developing an approach that explicitly accounts for this noisiness and its uncertainty is therefore essential.

In this paper, we propose to adjust for the noisiness in the data by calibrating the acceptance threshold using conformal prediction. Rather than simply accepting the claim based on the quality metric $\widehat{\mathcal{A}}$, we propose calibrating the threshold λ to mimic an oracle gold-standard $\mathcal{A}$ whilst limiting the number of false acceptances.

As a concrete example, Figure 1 illustrates our method's workflow in the context of clinical disease prediction. The input (a description of symptoms) is first processed by the generative model h which is prompted to extend the description, after which the candidate outputs are screened using the quality evaluator $\widehat{\mathcal{A}}$ and a calibrated filtering threshold $\hat{s}$. The selected generations retain the intended meaning of "common cold," though minor surface errors such as typos may remain. Such typos can also be viewed as a form of data augmentation: while they slightly perturb the text, they preserve semantic meaning and can improve model robustness. By contrast, the discarded output fails to capture relevant symptoms of the common cold.

3 Method: Filtering Using Conditional Conformal Risk Control

We propose a two-step approach for filtering outputs. In the first step, we randomly select a subset of the data, denoted by $\mathcal{D}_{\text{calib}} = \{(X_i, (G_{ik})_{k=1}^K, Y_i)\}$, on which evaluate the generations using both a gold-standard quality measure $\mathcal{A}$ and its surrogate $\widehat{\mathcal{A}}$ (for settings where no gold-standard exists, we propose an alternative in Section 3.3). This calibration set is then used to train a conformal prediction algorithm that calibrates the thresholding level λ correctly for that particular generation, accounting for the uncertainty in $\widehat{\mathcal{A}}$ as a surrogate for $\mathcal{A}$. In the second step, we apply the conformal prediction filter—using the calibrated threshold—to the remaining dataset, $\mathcal{D}_{\text{aug}} = \{(X_i, (G_{ik})_{k=1}^K, Y_i)\}$, using the conformal prediction algorithm.

Let the sizes of $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{aug}}$ be n_{calib} and n_{aug}, respectively. With a slight abuse of notation, we also use $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{aug}}$ to refer to the corresponding index sets when the meaning is clear from context.

3.1 Controlling the number of wrong inclusions

Problem Formalization. We consider the gold standard quality scores $\mathbf{A}_i = (A_{ik})_{k=1}^K$ and the corresponding surrogate scores $\hat{\mathbf{A}}_i = (\hat{A}_{ik})_{k=1}^K$ for the generations in the calibration data. We define the filtered set at surrogate level s by the notation: $\mathcal{S}(\hat{\mathbf{A}}_i, s) = \{G_{ik} : \hat{A}_{ik} \geq s\}$. Let $\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_i, s), \mathbf{A}_i)$ denote a loss function that measures the quality of filtered output compared to the ground truth $\mathbf{A}_i$. For instance, we may define $\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_i, s), \mathbf{A}_i)$ to be the number of generations $\hat{A}_{ik}$ with surrogate score at least s but whose gold-standard scores A_{ik} are below the nominal quality threshold λ:

$$\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_i, s), \mathbf{A}_i) = |\{G_{ik} \in \mathcal{S}(\hat{\mathbf{A}}_i, s) : A_{ik} < \lambda\}|. \tag{1}$$

We then define the non-conformity score as

$$S_i = S(\hat{\mathbf{A}}_i, \mathbf{A}_i) = \inf\{s : \mathcal{L}(\mathcal{S}(\hat{\mathbf{A}}_i, s), \mathbf{A}_i) \leq \rho\}, \tag{2}$$

where ρ is a hyperparameter that represents the tolerance on the loss, or the maximal number of "false discoveries" per sample that we are willing to allow. In other words, we define the non-conformity score $S(\hat{\mathbf{A}}_i, \mathbf{A}_i)$ as the minimal threshold s such that the filtered set $\mathcal{S}(\hat{\mathbf{A}}_i, s)$ contains only all the generations for X_i with surrogate score $\hat{\mathbf{A}}_{ik} > s$, and at most ρ of these generations have gold-standard scores $\mathbf{A}_{ik} < \lambda$.

In this paper, we formulate the problem of filtering generations based on imperfect surrogate quality scores $\hat{\mathbf{A}}$ as a calibration problem: we need to select the surrogate filtering threshold s in a data-driven manner so as to ensure that $\mathbb{P}(\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_{i_0}; s_{i_0}), \mathbf{A}_{i_0}) \leq \rho) \geq 1 - \alpha$ for all $i_0 \in \mathcal{D}_{\text{aug}}$ with some user-specified confidence level $\alpha \in (0, 1)$. To this end, we propose leveraging conformal prediction (CP) Vovk et al. (2005); Angelopoulos et al. (2022) for risk control. Conformal methods provide finite-sample, distribution-free guarantees by calibrating predictions using a hold-out validation set (see Appendix B for a more in-depth review). In our setting, we use the distribution of the scores $S(\hat{\mathbf{A}}_i, \mathbf{A}_i)$ to correctly calibrate our rejection threshold to ensure retaining quality content. Letting $\hat{s}_{i_0}$ be the output of the conformal prediction algorithm for each $\mathbf{X}_{i_0}$ (see the explicit formula in equation 7 in Appendix), we will solely accept generated examples with $\hat{A}_{i_0 k} > \hat{s}_{i_0}$.

3.2 Conditional Conformal Risk Control

While this setup is intuitive, one could argue that, like λ, the surrogate threshold s might just as well be chosen using data splitting – making the conformal prediction step appear unnecessary. However, our setting is more challenging: the difficulty of the filtering problem varies across samples, and fixed validation-based thresholds cannot adapt to this heterogeneity. To address this, we incorporate sample-specific information and apply conditional conformal prediction, allowing the filtering procedure to adapt to the hardness of each instance and thereby provide more reliable control.

While conformal prediction can act as a wrapper around any method, it is a well-established fact that it is impossible to get conditional results Barber et al. (2021). To address this, we adopt the relaxation proposed by Gibbs et al. (2025), which designs a prediction set that satisfies the guarantee over a specified function class $\mathcal{F}$:

$$\mathbb{E}\left[f(X_{i_0}) \cdot \Big(\mathbf{1}\{\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_{i_0}; s_{i_0}), \mathbf{A}_{i_0}) \leq \rho)\} - (1 - \alpha)\Big)\right] = 0, \text{ for all } f \in \mathcal{F}. \tag{3}$$

To handle conditional coverage without prior structural information, we take $\mathcal{F}$ to be a reproducing kernel Hilbert space (RKHS) with an added intercept, following Gibbs et al. (2025). Given a positive-definite kernel $W : \Omega \times \Omega \to \mathbb{R}$ (e.g., Gaussian/RBF kernel), define

$$\mathcal{F} = \{f_W(\cdot) + \beta : g_W \in \mathcal{F}_W, \beta \in \mathbb{R}\}, \tag{4}$$

where $\mathcal{F}_W$ is the RKHS function class associated with W. The intercept β guarantees the marginal coverage, while the RKHS term f_W enables flexible, smooth calibration of the conformity scores $\{S_i\}_{i \in \mathcal{D}_{calib}}$ against covariates. As the following lemma shows, the resulting cutoff $\hat{s}_{i_0}$ for each data $i_0 \in \mathcal{D}_{aug}$ satisfies the conditional guarantee over a localization region around X_{i_0}.

Lemma 1 *Consider the function class $\mathcal{F}$ as defined in Equation 4, and assume $\mathcal{D}_{calib} \bigcup \mathcal{D}_{aug}$ are i.i.d. . Suppose $\mathcal{L}_\lambda(\cdot, \cdot)$ is monotone (i.e. for any sets $\mathcal{S}^1_{i_0} \subseteq \mathcal{S}^2_{i_0}$, it must be the case that $\mathcal{L}_\lambda(\mathcal{S}^1_{i_0}, \mathbf{A}_{i_0}) \leq \mathcal{L}_\lambda(\mathcal{S}^2_{i_0}, \mathbf{A}_{i_0})$) and $\mathcal{L}_\lambda(\emptyset, \cdot) = 0$. Assume $W(x, \cdot)$ defines a density with respect to each $x \in \Omega$, and sample $X'_{i_0} \mid X_{i_0} = x \sim W(x, \cdot)$. Then for all $f \in \mathcal{F}$, $i_0 \in \mathcal{D}_{aug}$,*

$$\mathbb{P}\left(\mathcal{L}_\lambda(\mathcal{S}(\hat{\mathbf{A}}_{i_0}; \hat{s}_{i_0}), \mathbf{A}_{i_0}) \leq \rho \mid X'_{i_0} = x'_{i_0}\right) = 1 - \alpha - \frac{\gamma \mathbb{E}[\hat{f}_W^{\hat{s}_{i_0}}(x'_{i_0})]}{\mathbb{E}[W(X_{i_0}, x'_{i_0})]},$$

where γ is the hyperparameter and $\hat{f}_W^{\hat{s}_{i_0}} \in \mathcal{F}_W$ is the fitted RKHS function defined in equation 5.

Due to the infinite dimensionality of the RKHS class, the achieved coverage departs from the nominal level $1 - \alpha$ by a gap of $\frac{-\gamma \mathbb{E}[\hat{f}_W^{\hat{s}_{i_0}}(x'_{i_0})]}{\mathbb{E}[W(X_{i_0}, x'_{i_0})]}$. from the nominal level $1 - \alpha$. However, this coverage gap is estimable, and can be quantified using the procedure proposed in Gibbs et al. (2025). The proof of Lemma 1 is shown in Appendix C.2

In practice, we rely on the fast implementation of this approximate conditional CP algorithm as provided in Anonymous (2026), which provides a fast alternative to the original algorithm of Gibbs et al. (2025).

Algorithm 1 Conformal Filtering

```
Require: Reference evaluation $A$; surrogate evaluation $\hat{A}$; calibration dataset $\mathcal{D}_{\text{calib}} = \{(X_i, (G_{ik})_{k=1}^K, Y_i)\}$ augmentation dataset $\mathcal{D}_{\text{aug}} = \{(X_i, (G_{ik})_{k=1}^K), Y_i)\}$; quality level $\lambda$; loss function $\mathcal{L}$; contamination allowance $\rho$
  Compute the reference score: $A_{ik} = \mathcal{A}(G_{ik}, X_i, Y_i),\ \forall i \in \mathcal{D}_{\text{calib}}$.
  Compute the surrogate score: $\hat{A}_{ik} \leftarrow \hat{\mathcal{A}}((G_{ik})_{k=1}^K, X_i, Y_i),\ \forall i \in \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{aug}}$.
  Compute the non-conformity score associated with loss $\mathcal{L}$ and $\rho$, denoted as $S(\hat{\mathbf{A}}_i, \mathbf{A}_i)$, according to Equation 2, $\forall i \in \mathcal{D}_{\text{calib}}$.
  for $i \in D_{\text{aug}}$ do
      Fit conditional conformal prediction to find $\hat{s}_i$
      Select generations : $\mathcal{S}(\hat{\mathbf{A}}_i, \hat{s}_i) = \{G_{ik} : \hat{A}_{ik} \geq \hat{s}_i\}$
  end for

Ensure: The selected generations $\{\mathcal{S}_i : i \in \mathcal{D}_{\text{aug}}\}$.
```

3.3 Learning to recognize quality outputs on $\mathcal{D}_{\text{train}}$

In the previous discussion, we focused on the setting where the gold standard measure $\mathcal{A}$ is directly available on a small subset of the data. We now extend our approach to scenarios in which only a surrogate measure $\tilde{\mathcal{A}}$ can be observed. We shall assume that the surrogate measure satisfies

$$\mathcal{A}(G_{ik}, X_i, Y_i) = \mathbb{E}_{\tilde{X}_i \sim h^\star(C_i, Y_i)}\left[\tilde{\mathcal{A}}(G_{ik}, \tilde{X}_i, Y_i)\right],$$

In other words, the gold-standard is the population average of the observed surrogate, and conversely, $\tilde{\mathcal{A}}$ can be viewed as a specific realization of a random variable, centered at $\mathcal{A}$. For instance, in text data, embedding-based similarity metrics such as cosine similarity computed from BERT embeddings are widely used to capture semantic coherence Devlin et al. (2018); Zhang et al. (2020). In image data, similarity measures based on CLIP scores Radford et al. (2021) are effective for capturing both semantic alignment and stylistic similarity. These metrics typically compare each generation directly against its original sample, which can be viewed as a realization from the underlying distribution $h^\star$.

We propose reducing the variability of the surrogate $\tilde{\mathcal{A}}$ by applying a regression-based strategy that leverages similar samples to approximate the underlying expectation. By smoothing over similar samples, this learned approximation is expected to more closely reflect the ground-truth measure $\mathcal{A}$.

Let $\tilde{A}_{ik} = \tilde{\mathcal{A}}(G_{ik}, X_i, Y_i)$ and $A_{ik} = \mathcal{A}(G_{ik}, X_i, Y_i)$. We model A_{ik} as:

$$A_{ik} = \eta(G_k, C_i, Y_i) + \epsilon_{ik}$$

where ϵ_{ik} denotes some centered noise, and where $\eta(G_{ik}, C_i, Y_i) = \mathbb{E}[\tilde{A}_{ik} | G_{ik}, C_i, Y_i]$ is the population quantity we would like to estimate.

In this setting, we split the data into $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{calib}}$, and $\mathcal{D}_{\text{aug}}$. We then train a regression model $\hat{\mathcal{A}} : (G_{ik})_{k=1}^K, X_i, Y_i \mapsto \hat{A}_{ik}$ on $\mathcal{D}_{\text{train}}$ to predict $\tilde{A}_{ik}$. The model takes as input the generated samples $(G_{ik})_{k=1}^K$ together with the observed (X_i, Y_i) and outputs a predicted score. For example, in text data, $\hat{\mathcal{A}}$ may incorporate features such as the semantic relevance between G_{ik} and (X_i, Y_i), as well as generation entropy, a metric that has been used to quantify uncertainty in generated outputs and to detect hallucinations. We then calibrate $\hat{\mathcal{A}}$ using $\tilde{\mathcal{A}}$ as an unbiased estimator of $\mathcal{A}$, as described in Section 3.1.

4 Experiments

To highlight the efficacy of our method, we propose three case studies: (a) a data enrichment setting, (b) an imbalanced classification setting, and (c) a very low-data regime with generations of heterogeneous quality. Our examples span different data types, from text, to images, to tabular data.

4.1 Prediction with LLM-augmented training data

We study our data augmentation pipeline for clinical text classification, focusing on mapping symptom descriptions to medical diagnoses (Gretel AI, 2024). The dataset (X_i, Y_i) consists of 853 training samples and 212 test samples, where X_i denotes a symptom description and Y_i is one of 22 possible diagnoses. Each training example (X_i, Y_i) is augmented using a generative language model (`GPT-4.1 nano` (OpenAI, 2023)), which extends the original symptom description X_i with five additional sentences $(G_{ik})_{k=1}^{5}$. From these extensions, we generate new samples that inherit the original label, yielding a total of 4,265 synthetic observations $\{(G_{ik})_{k=1}^{5} : i = 1, 2, \cdots, 853\}$.

To ensure output quality, we employ a two-stage evaluation strategy. First, a random subset of 500 generations, derived from 100 symptom descriptions, is evaluated with a high-accuracy model $\mathcal{A}$ (`Gemini-2.5-pro` (Comanici et al., 2025)), forming the calibration set $\mathcal{D}_{\text{calib}}$. Next, all augmented samples are scored using a faster, lower-cost surrogate model $\hat{\mathcal{A}}$ (`Gemini-2.5-flash` (Comanici et al., 2025)). Both models assign a score in $[0, 1]$, with 0.5 as the retention threshold (see the detailed prompt in the supplement). This design reflects a practical labeling scenario in which reliable annotations are costly, whereas approximate labels can be obtained inexpensively. Let $(A_{ik})_{i \in \mathcal{D}_{\text{calib}}}$ denote the Gemini-pro scores and $(\hat{A}_{ik})_{i=1}^{853}$ denote the Gemini-flash scores.

We then apply our calibration step. For each $(X_i, (G_{ik})_{k=1}^{K}, Y_i) \in \mathcal{D}_{\text{calib}}$, a non-conformity score is defined as the minimum threshold that guarantees all selected sentences achieve a pro-score above 0.5, so that: $S(\hat{\mathbf{A}}_i, \mathbf{A}_i) = \inf\left\{\tau : \left|\{G_{ik} : 1 \le k \le 5,\ \hat{A}_{ik} \ge \tau,\ A_{ik} < 0.5\}\right| \le 1\right\}$.

For each X_i, we embed the text into a lower-dimensional space using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a classical method for producing low-dimensional text representations, fitted on the entire training set. Let $\hat{\pi}(\cdot)$ denote the resulting LDA mapping. We construct a kernel $W(\cdot, \cdot) = \exp\{-\xi \|\hat{\pi}(\cdot) - \hat{\pi}(\cdot)\|_2^2\}$, with ξ selected via cross-validation. Then we apply conditional CP (CondCP) (Gibbs et al., 2025) with $\alpha = 0.1, \rho = 0$ to obtain adaptive thresholds on $\hat{A}_{ik}$.

To evaluate performance, we fine-tune a diagnostic classifier (`distilbert-base-uncased` (Devlin et al., 2018)) using LoRA (Hu et al., 2022). Each training iteration consists of 100 fixed high-confidence documents (selected by the pro-scores) and 400 additional documents sampled under one of the following filtering schemes: (1) No augmentation; (2) No filtering; (3) Filtering by $\hat{A}_{ik}$ only (threshold = 0.5); (4) Hybrid filtering (using A_{ik} for $\mathcal{D}_{\text{calib}}$ and $\hat{A}_{ik}$ with threshold 0.5 for the remainder); (5) CondCP-based filtering on $\mathcal{D}_{\text{aug}}$ (using A_{ik} for $\mathcal{D}_{\text{calib}}$).

Performance, averaged across 20 trials, is reported in Figure 2, with evaluation consistently conducted on the held-out test set. We also report the results of experiments performed in an identical manner on topic prediction (predicting the topic of statistical abstracts downloaded from arXiv with 5 possible categories) and sentiment analysis (predicting one of 6 emotions on a dataset of Twitter messages — see details in the Appendix D). Overall, across these three datasets, our CondCP filter improves the precision, recall, and F1-score by up to 3% over the unaugmented baseline, and substantially improves upon the unfiltered baseline. We note that in the diagnosis task, the unfiltered augmentation outperforms the unaugmented baseline, but this advantage does not hold for the abstract and Twitter datasets, suggesting that including all generations can be detrimental when low-quality samples are present. In contrast, the CondCP filter achieves the best performance across all metrics and tasks.

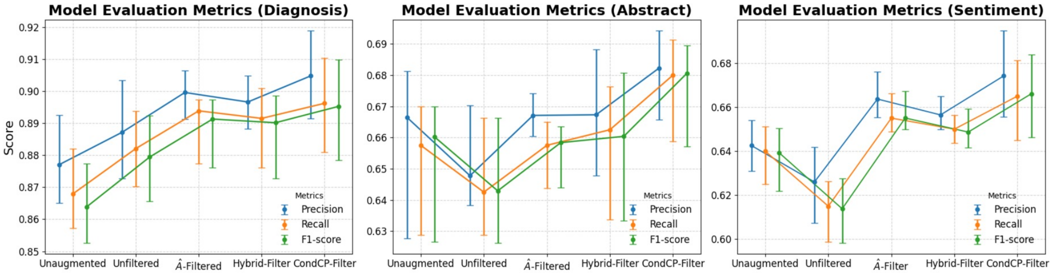

Figure 2: Evaluation of different data augmentation methods on diagnosis prediction, abstract topic prediction, and Twitter message sentiment prediction. Results are averaged over 20 replicates. Error bars indicate the interquartile range, with centers representing the median and boundaries corresponding to the first and third quartiles.

Dataset	**Strategy**	F_1 (↑)	**Precision** (↑)	**Recall** (↑)	**Stable Rank** (↑)
Thyroid (N=2,644, d=27, Imb.=6.4%)	Unaugmented	0.139 ± 0.080	**0.538** ± **0.240**	0.081 ± 0.050	7.713 ± 0.169
	SMOTE	0.499 ± 0.022	0.354 ± 0.017	**0.848** ± **0.061**	7.358 ± 0.202
	Unfiltered	0.495 ∓ 0.031	0.356 ± 0.030	$\underline{0.819 \pm 0.061}$	8.238 ± 0.812
	$\widehat{A}$-Filter	$\underline{0.507 \pm 0.046}$	0.370 ± 0.046	0.817 ± 0.065	$\underline{8.495 \pm 0.384}$
	CondCP-Filter	**0.542** ± **0.043**	$\underline{0.417 \pm 0.043}$	0.783 ± 0.070	**8.730** ± **0.336**
Credit Card Fraud (N=284,807, d=28, Imb.=0.17%)	Unaugmented	$\underline{0.732 \pm 0.023}$	**0.886** ± **0.044**	0.626 ± 0.045	**25.790** ± **0.605**
	SMOTE	0.108 ± 0.004	0.057 ± 0.002	**0.920** ± **0.023**	2.405 ± 0.012
	Unfiltered	0.709 ± 0.029	0.668 ± 0.049	0.760 ± 0.055	1.962 ± .049
	$\widehat{A}$-Filter	0.711 ± 0.030	0.675 ± 0.048	0.757 ± 0.061	2.273 ± 0.117
	CondCP-Filter	**0.807** ± **.027**	$\underline{0.813 \pm 0.041}$	$\underline{0.803 \pm 0.045}$	$\underline{7.380 \pm 1.049}$
MNIST 7 vs. Others (N=70,000, d=784, Imb.=10.9%)	Unaugmented	$\underline{0.894 \pm 0.010}$	**0.905** ± **0.008**	0.882 ± 0.027	**16.582** ± **0.065**
	SMOTE	0.880 ± 0.008	0.858 ± 0.001	**0.903** ± **0.015**	13.507 ± 0.085
	Unfiltered	0.891 ± 0.011	0.891 ± 0.005	$\underline{0.892 \pm 0.027}$	$\underline{11.980 \pm 0.840}$
	$\widehat{A}$-Filter	0.892 ± 0.009	0.895 ± 0.008	0.888 ± 0.025	11.865 ± 0.855
	CondCP-Filter	**0.896** ± **0.007**	$\underline{0.904 \pm 0.005}$	0.888 ± 0.025	11.972 ± 1.510

Table 1: Results of imbalanced classification: predictive performance metrics and data diversity (Stable Rank) averaged over 10 different splits. Dataset sizes (N), feature dimensions (d), and imbalance rates (Imb.) are given in parentheses. Higher values are better for all metrics presented. The best value is bolded and the second best value is underlined. Across all benchmarks, our CondCP-Filter consistently attains the best F_1 and increase data diversity, as reflected by higher stable rank.

4.2 Imbalanced Classification: tabular data examples

In imbalanced classification, models often default to predicting the majority class, yielding misleadingly high accuracy while missing rare but critical events. For example, in the European Credit-Card Fraud dataset[1] (0.17% frauds), labeling all cases as "non-fraud" achieves 99.8% accuracy but detects no fraud (He & Garcia, 2009; Japkowicz & Stephen, 2002). Data augmentation seems therefore a promising way of enhancing recall whilst maintaining precision.

We evaluate our method on three benchmark datasets spanning different imbalance regimes: European Credit-Card Fraud (Kaggle), Thyroid (OpenML), and MNIST-7 vs Others (OpenML). See the details of the dataset and experiment setup in Appendix D.8.1. In these settings, to generate new data, we train a Variational AutoEncoder (VAE) (Kingma & Welling, 2013; Sohn et al., 2015) to increase the number of samples from the minority class. Since gold-standard quality measures are not available in this setting, we use the procedure detailed in section 3.3, and use for our surrogate scores $\hat{A}$ a gradient boosting predictor, trained to predict the surrogate measure $\tilde{A}$. For the experiments presented in this subsection, $\tilde{A}$ is defined as the geometric mean of a k-nearest-neighbor similarity (to measure closeness to real minority data) and a cosine similarity (directional closeness to the reference data).

In each case, we split the data into train/calibration/test subsets (60/20/20) and report average F_1 scores. We fit a logistic regression classifier, and we compare the performance of our CP-filtering procedure with (a) an unaugmented baseline; (b) SMOTE (Chawla et al., 2002), a widely used oversampling method that interpolates minority examples in feature space; (c) unfiltered augmentation; and (d) various filtering procedures (e.g. CP-based filtering, and filtering based on $\hat{A}$).

Table 1 reports F_1, precision, recall, and Stable Rank across five benchmarks. On *severely imbalanced datasets* such as credit-card fraud, quality-controlled augmentation clearly dominates both the unaugmented baseline and SMOTE; while SMOTE boosts recall, it inflates false positives, lowering precision. Our $\widehat{A}$-Filter and CondCP-Filter maintain recall while improving precision, yielding the best F_1. On *moderate imbalance* (Thyroid), all methods perform similarly, but our filters still outperform baselines and increase Stable Rank, indicating genuine diversity rather than duplication. For MNIST-7, where *the signal is strong*, unfiltered augmentation already works well; nonetheless, CondCP-Filter achieves the highest F_1 and superior precision–recall balance, showing the benefit of targeted acceptance even in easier tasks.

Beyond predictive performance, we also study diversity metrics of the training sets after augmentation with filtering. In particular, we compute the *stable rank* of the feature matrix (X), which is defined as $\|X\|_F^2/\|X\|_2^2$. Stable rank captures the effective dimensionality of the sample cloud (Tsitsulin et al., 2023). Whereas simple augmentation often inflates data density along a few dominant directions

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

(due to interpolation), our method introduces genuinely new modes in the minority manifold, reflected in higher stable rank. These results indicate that quality-controlled generation is not only effective for balancing datasets, but also enhances geometric richness in ways that may improve generalization.

4.3 Low-Data Regime with mixed-quality generations: an image analysis example

In this example, we wish to evaluate the performance of our method in a low-data regime for classification, with inputs of mixed qualities (with a distribution of 50% good inputs, 50% bad). In this setting, we expect the filtering procedure to be particularly useful.

To this end, we consider two classes from the mini ImageNet dataset (arctic foxes and toucans), and make a dataset of around 172 training images (86 per class). A moderate-capacity CNN is trained from scratch on this base set. We simulate data augmentation by masking 70% of each training image and asking DALL·E 2 to inpaint the missing regions, and replacing the masked area by the generated content (Fig. 3). To simulate unhelpful data generations, we use the masks as additional generations that need filtering out.

Each candidate is assigned a surrogate quality score. To compute this score, we first train the CNN on a separate split of the data (with around 35 images per class on average), and take the CNN's class–probability margin $|p(y \mid x) - 0.5|$ as a measure of the compatibility of the generation with its class and the image clip score as the quality gold standard. We then compare three filtering regimes: (a) **Threshold baseline:** keep candidates with score $\geq \lambda$; (b) **Marginal CP:** compute a global cutoff from calibration documents using split–conformal quantiles of per–document scores S_{doc}; and (c) **Conditional CP:** compute adaptive per–document cutoffs from PCA embeddings of the base images using the CondCP-filter procedure.

After selection, we retrain a CNN from scratch on all original images plus the selected augmentations. Validation and test sets remain fixed. The validation set (175 images of each class) is used for hyperparameter selection (here, the threshold λ, chosen in the grid $\lambda \in \{0.5, 0.75, 0.8, 0.9\}$). The CP parameters are fixed to $\rho = 0$ and $\alpha = 0.1$). Results are reported on the test set (300 images), and averaged over 5 runs of the procedure, shuffling the training set split into calibration and testing. Across this two–class task, conformal filtering yields consistent improvements, with the CondCP filter providing a $+3.7\%$ accuracy improvement over the unaugmented baseline and $+2\%$ gain in test accuracy over the baseline. Importantly, we note that the marginal CP baseline does not yield any improvement in accuracy, highlighting the importance of using conditional CP in this setting. Moreover, we do note the importance of filtering just the right amount, as the choice of the λ does not default to the minimal value.

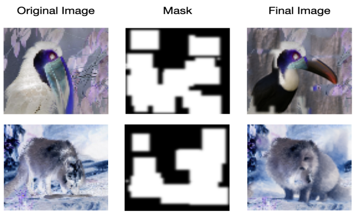

Figure 3: Examples of data generation procedures for an image of a Toucan (top row) and an image of an Arctic Fox.

λ	Regime	Training Set Size	Test Accuracy
-	Unaugmented	172.8 ± 5.8	0.786 ± 0.023
-	Unfiltered	3630.8 ± 317.4	0.802 ± 0.006
0.80	$\hat{A}$ Filter	897.0 ± 986.9	0.804 ± 0.036
0.75	CondCP Filter	1916.6 ± 158.4	**0.823** ± **0.018**
0.50	MargCP Filter	1334.8 ± 463.2	0.763 ± 0.055

Figure 4: Performance of the methods on the ImageNet dataset, averaged over 5 iterations.

5 Conclusion

In this study, we propose a principled data augmentation algorithm that evaluates the quality of generated content beyond simple comparison with observed data, and filters out low-quality generations with provable risk control. Future directions for improvement include: (1) extending our methodology to other generative settings such as counterfactual or retrieval-based augmentation; (2) integrating our framework with other conformal prediction techniques, such as adaptive level control for different tasks.

Ethics statement

All authors confirm that this work adheres to the ICLR Code of Ethics. This research does not involve human subjects, personally identifiable information, or sensitive data. The datasets used are publicly available and have been properly cited. No potentially harmful applications, discriminatory outcomes, or security/privacy risks are anticipated from this study.

We have made efforts to ensure fairness, transparency, and integrity in both the methodology and the interpretation of results. All analyses and conclusions comply with ethical standards for reproducibility, research documentation, and integrity. No conflicts of interest or sponsorship influenced this work.

References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378, 2022.

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning (ICML)*, pp. 290–306. PMLR, 2022.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI conference on artificial intelligence*, volume 34-05, pp. 7383–7390, 2020.

Anastasios N Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2022.

Anonymous. Speedcp: Fast kernel-based conditional conformal prediction, 2026. Submitted to ICLR 2026. Under review. Available at `https://openreview.net/forum?id=22472`.

Artgor. arxiv metadata exploration. `https://www.kaggle.com/code/artgor/arxiv-metadata-exploration`, 2019. Accessed: 2025-09-20.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Yaping Chai, Haoran Xie, and Joe S Qin. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *arXiv preprint arXiv:2501.18845*, 2025.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*, 2024.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. doi: 10.48550/arXiv.2305.16289.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Steven Y Feng, Aaron W Li, and Jesse Hoey. Keep calm and switch on! preserving sentiment and fluency in semantic text exchange. *arXiv preprint arXiv:1909.00088*, 2019.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. Genaug: Data augmentation for finetuning text generators. *arXiv preprint arXiv:2010.01794*, 2020.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. *arXiv preprint arXiv:2301.02427*, 2023.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, 2025.

Gretel AI. symptom_to_diagnosis dataset. `https://huggingface.co/datasets/gretelai/symptom_to_diagnosis`, 2024. Accessed: 2025-09-20.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *arXiv preprint arXiv:2405.10301*, 2024.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*, 2023.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, et al. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*, 2024.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Wang. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kevin Han Huang, Peter Orbanz, and Morgane Austern. Data augmentation in the underparameterized and overparameterized regimes. *arXiv preprint arXiv:2202.09134*, 2022.

Mominul Islam, Hasib Zunair, and Nabeel Mohammed. Cossif: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets. *Computers in Biology and Medicine*, 172:108317, 2024.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint arXiv:2402.04376*, 2024.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pp. 2323–2332, 2018.

James Jordon, Lukasz Szpruch, F. Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, and Samuel N. Cohen. Synthetic data–what, why and how? *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1039–1069, 2022. doi: 10.1111/rssa.12804.

Bhavikji Kadara. Emotions dataset. https://www.kaggle.com/datasets/bhavikjikadara/emotions-dataset, 2018.

Min Kang, Kye Hwa Lee, and Youngho Lee. Filtered bert: Similarity filter-based augmentation with bidirectional transfer learning for protected health information prediction in clinical documents. *Applied Sciences*, 11(8):3668, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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)*, pp. 3609–3619, 2019.

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.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3855–3865, 2024a.

Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. Data generation using large language models for text classification: An empirical case study. *arXiv preprint arXiv:2407.12813*, 2024b.

Yu Li, Xiao Li, Yating Yang, and Rui Dong. A diverse data augmentation strategy for low-resource neural machine translation. *Information*, 11(5):255, 2020.

Yujia Li, Yitong Zhang, Yiming Yu, et al. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21954–21967, 2022.

Zhiteng Li, Lele Chen, Jerone Andrews, Yunhao Ba, Yulun Zhang, and Alice Xiang. Gendataagent: On-the-fly dataset augmentation with synthetic data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *Proceedings of the 2023 Conference of the Association for Computational Linguistics*, pp. 2194–2216, 2023.

Mosleh Mahamud, Zed Lee, and Isak Samsten. Distributional data augmentation methods for low resource language. *arXiv preprint arXiv:2309.04862*, 2023. URL `https://arxiv.org/abs/2309.04862`.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*, 2025.

Ryumei Nakada, Yichen Xu, Lexin Li, and Linjun Zhang. Synthetic oversampling: Theory and a practical approach using llms to address data imbalance. *arXiv preprint arXiv:2406.03628*, 2024.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*, 2020.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL `https://arxiv.org/abs/2303.08774`.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Sai Rajeswar, Tianyi Lin, Ruihan Bao, Andrew Saxe, and R. Devon Hjelm. Diversity and fidelity in generative models: Trade-offs and metrics. *arXiv preprint arXiv:2301.06712*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 12247–12258, 2019.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8011–8021, 2023.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

Hooman Shahrokhi, Devjeet Raj Roy, Yan Yan, Venera Arnaoudova, and Janaradhan Rao Doppa. Conformal prediction sets for deep generative models via reduction to conformal regression. *arXiv preprint arXiv:2503.10512*, 2025.

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations*, 2020.

Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 769–778, 2023.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api is enough: Conformal prediction for large language models without logit-access.(2024). *URL https://arxiv. org/abs/2403*, 1216, 2024.

Arwa Mohammed Taqi, Ahmed Awad, Fadwa Al-Azzo, and Mariofanna Milanova. The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 140–145. IEEE, 2018.

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

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 169–188. PMLR, 2023.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*, 2023.

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

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*, 2021.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. Prompt-based data augmentation for low-resource nlu tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4170–4183, 2022. URL `https://aclanthology.org/2022.acl-long.292/`.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*, 2019.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020.

Xun Yao, Zijian Huang, Xinrong Hu, JACK Yang, and Yi Guo. Masking the unknown: leveraging masked samples for enhanced data augmentation. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

Yao Zhang and Emmanuel J Candès. Posterior conformal prediction. *arXiv preprint arXiv:2409.19712*, 2024.

Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Retgen: A joint framework for retrieval and grounded text generation modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36-10, pp. 11739–11747, 2022.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, LLMs were used for synthetic data generation as part of our research on data augmentation with generative models. Specifically, LLMs produced candidate text samples that were subsequently filtered, evaluated, and integrated into the experimental pipeline. The role of LLMs was limited to data generation within the proposed methodology and did not extend to research ideation, conceptual framing, or substantive writing of the manuscript.

All analysis, interpretation, and writing were conducted by the authors. We take full responsibility for the content of this paper, including any outputs derived from LLMs. No portion of the manuscript relies on fabricated or plagiarized material produced by LLMs.

B RELATED LITERATURES

Data Augmentation In Ding et al. (2024), LLM-based data augmentations are categorized into four categories: data creation, data reformation, data labeling, and human-LLM co-annotation. In this work we focus on the data reformation, which transforms existing data to produce new data. People have proposed data reformation approaches prior to the advent of pre-trained generative models, with the majority of them being rule-based methods (Feng et al. (2021)). For instance, Easy Data Augmentation (EDA) Wei & Zou (2019) applies token-level perturbations like synonym replacement, random insertion, deletion, and swapping; Machine back-translation involves translating the original sentences into another language and then translating them back to the original language Sennrich et al. (2015); Edunov et al. (2018). Model-based methods, by contrast, leverage generative models to synthesize new text. Common examples include paraphrasing Kumar et al. (2019), semantic text exchange Feng et al. (2019) and masked word prediction followed by replacement Ng et al. (2020). The goal is to generate synthetic data that introduces diversity while maintaining semantic consistency (often referred to as label-preserving in classification problems (Xie et al., 2020)). Ideally, augmented data should not be too similar to the original (which limits diversity) nor too dissimilar (which risks domain shift and degraded performance).

Despite its ease of implementation, synthetic data generated by generative models is often noisy and distributionally misaligned with the original data, potentially hindering model training (Zhang et al., 2022). To address this, several complementary strategies have been proposed. Some approaches focus on prompt engineering to steer generation more precisely (Veselovsky et al., 2023; Gupta et al., 2023), while others leverage model-based augmentation by estimating a generative process from the training set and sampling from it (Anaby-Tavor et al., 2020). Retrieval-based techniques further enhance the expressiveness of LLM-driven augmentation by incorporating external knowledge (Chai et al., 2025). At generation time, diffusion-based models have been guided toward low-density or underrepresented modes—such as through class-conditional or classifier-informed sampling for minority classes (Sehwag et al., 2022; Trabucco et al., 2023). Prompt perturbation has been used to mitigate semantic ambiguity and encourage coverage of diverse outputs (Sarıyıldız et al., 2023; Shipard et al., 2023). In parallel, foundation models have been fine-tuned to better align with the target domain, either via on-the-fly adaptation (e.g., GenDataAgent (Li et al., 2025)) or through large-scale domain-specific retraining, as demonstrated with text-to-image diffusion models on ImageNet-scale data (Azizi et al., 2023; Dunlap et al., 2023).

These approaches improve the quality of the generations, but still, there could be low-quality generations that ideally we would like to filter out. The filtering-based methods evaluate typically the generations based on quality metrics, such as human evaluation (Wang et al., 2021; Liu et al., 2022)(which can be expensive), model confidence or difficulty (Hemmat et al., 2023; Agarwal et al., 2022), similarity to the original input in paraphrasing Li et al. (2024b), confidence of LLM, or classifiers trained to distinguish real from synthetic data Veselovsky et al. (2023). Most of these methods either explicitly or implicitly leverage the prediction on the quality of the generations, which could be problematic when the prediction is not accurate.

Filtering. Filtering methods are commonly based on the following strategies (Chai et al., 2025):

- **Lexical overlap:** filtering based on n-gram overlap metrics such as ROUGE.
- **Semantic similarity:** filtering based on cosine similarity in embedding space.

- **Model-based filtering:** scoring generations using pre-trained models (e.g., LLMs).
- **Round-trip consistency:** checking whether back-translation or round-trip generation recovers the original input.
- **Influence-function filtering:** discarding augmentations predicted to harm downstream performance (Yang et al., 2020).

In practice, many augmentation pipelines combine multiple filters; for example, a heuristic may first remove obviously poor outputs, and then the top-k most similar examples to the ground truth are retained (Chai et al., 2025). The goal is to balance fidelity and diversity: overly strict filters yield safe but low-diversity augmentations, while overly permissive filters risk introducing label noise or factual errors. Recent methods explicitly address this trade-off. For example, Mask-then-Fill (Gao et al., 2023) reports that infilling achieves a balance between novelty and distributional similarity to the source, likely through careful tuning of mask size and model parameters. In contrast, M4DA (Yao et al., 2024) promotes diversity by masking tokens to increase variance and then selecting variants with the highest semantic complexity. While the generated text must still preserve the original meaning, this preference for more complex rephrasings can yield stronger augmentation effects. Experiments on text classification benchmarks show that such methods can outperform conservative approaches, suggesting that filtering should not always default to the safest outputs—some controlled complexity, when consistent, is beneficial.

C Related Works On Conformal Prediction

Conformal Prediction Given a dataset $\{(X_i, Y_i)\}_{i=1}^N$, a pretrained-predictor h and a new text input X_{n+1}, conformal prediction (Vovk et al., 2005) attempts to construct a prediction $\hat{C}(X_{n+1})$ such that $\mathbb{P}\left(Y_{n+1} \notin \hat{C}(X_{n+1})\right) \leq 1 - \alpha$ for some user-specified α. Conformal prediction has the distribution-free property and it is finite-sample valid under the exchangeability of the data points $\{(X_i, Y_i)\}_{i=1}^{n+1}$. For instance, in split conformal prediction, one can define $S_i = \|Y_i - h(X_i)\|$ and then set $\hat{C}(X_{n+1}) = \{y : \|y - h(X_{n+1})\| \leq \tau\}$ where $\tau = \text{quantile}(\{S_i\}_{i=1}^n \cup \{\infty\}, 1 - \alpha)$. This type of method provides a guarantee on marginal coverage. Previous studies have demonstrated that achieving exact conditional coverage is impossible without any further distributionally assumption (Barber et al., 2021). Nevertheless, researchers have developed methods to achieve conditional coverage with controllable error rates (Zhang & Candès, 2024; Gibbs et al., 2025).

Conformal Prediction and LLMs Researchers have increasingly explored the application of conformal prediction (CP) frameworks in generative models for factuality control, motivated by CP's ability to provide distribution-free inference. In (Ren et al., 2023; Kumar et al., 2023), CP is employed to identify probability thresholds for next-token generation, thereby selecting response candidates. Several works have proposed CP methods that do not require access to model logits (Su et al., 2024). For instance, (Shahrokhi et al., 2025; Quach et al., 2023) use CP to determine the number of generations needed to construct a prediction set that includes at least one truthful response or satisfies a specified confidence level. Other approaches, such as (Mohri & Hashimoto, 2024; Cherian et al., 2024), segment LLM outputs into individual claims and apply CP to select factual ones. Additionally, Gui et al. (2024) extends CP to multiple test units with a focus on ensuring valid false discovery rate (FDR) control. Despite the successes in these applications, how CP can be applied in data augmentation is under-explored, perhaps due to its unsupervised nature.

Conditional Conformal Prediction While Conformal prediction seems like a promising wrapper around any blackbox method, its scope is fundamentally restricted to marginal coverage guarantees. However, marginal coverage does not preclude large variability in *conditional coverage*, defined as

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1} = x\big) = 1 - \alpha,$$

which may differ significantly across inputs. This limitation is critical in sensitive applications (e.g., medicine, finance), where systematic under-coverage on certain subgroups undermines reliability. Prior work shows that in distribution-free settings, exact conditional coverage is impossible: any set satisfying it must degenerate to $\hat{C}(X_{n+1}) = \mathbb{R}$ with infinite expected size (Barber et al., 2021).

To address this, Gibbs et al. (2025) reformulate conditional coverage as a marginal constraint over measurable functions f:

$$\mathbb{E}\Big[f(X_{n+1}) \cdot \big(\mathbf{1}\{Y_{n+1} \in \hat{C}(X_{n+1})\} - (1-\alpha)\big)\Big] = 0.$$

They then restrict f to a user-specified function class $\mathcal{F}$, yielding approximate conditional validity. Different choices of $\mathcal{F}$ lead to different notions of conditional coverage: for example, $\mathcal{F} = \{\text{constants}\}$ recovers marginal coverage, while $\mathcal{F} = \big\{\sum_{G\in\mathcal{G}} \beta_G \mathbf{1}\{x \in G\} : \beta \in \mathbb{R}^{|\mathcal{G}|}\big\}$ enforces group-conditional guarantees. Gibbs et al. (2025), by contrast, allow $\mathcal{F}$ to take more general forms, from linear distribution shifts, to more complex shifts parametrized by an RKHS function.

C.1 Additional details on conditional conformal

In our setting, the conformity score S_{i_0} is unknown for every $i_0 \in \mathcal{D}_{\text{aug}}$; we therefore impute a value S for each such test index and solve a single regularized quantile problem that treats the imputed test pair symmetrically with the calibration data. Following Gibbs et al. (2025), we estimate a high–probability upper bound for these scores $\{S_i\}_{i\in\mathcal{D}_{\text{calib}}} \cup S$ by fitting a regularized kernel quantile regression:

$$\hat{f}_S \;=\; \arg\min_{f\in\mathcal{F}^*} \left\{ \frac{1}{|\mathcal{D}_{\text{calib}}|+1} \sum_{i\in\mathcal{D}_{\text{calib}}} \ell_\alpha\big(S_i - f(X_i)\big) + \frac{1}{|\mathcal{D}_{\text{calib}}|+1}\, \ell_\alpha\big(S - f(X_{i_0})\big) + \frac{\gamma}{2}\, \|f_W\|_W^2 \right\}, \tag{5}$$

where $\alpha \in (0,1)$, $\ell_\alpha(z) = (1-\alpha)[z]_+ + \alpha[z]_-$ is the pinball loss, $\gamma > 0$ is a regularization parameter, and $\|\cdot\|_W$ is the RKHS norm associated with the positive–definite kernel W.

By the representer theorem (Kimeldorf & Wahba, 1971), the optimizer admits the finite expansion

$$\hat{f}_S(X) \;=\; \hat{\beta}_S \;+\; \frac{1}{\gamma} \sum_{i\in\mathcal{D}_{\text{calib}}\cup\{i_0\}} \hat{v}_{S,i}\, W(X, X_i), \tag{6}$$

with coefficient vector $\hat{v}_S \in \mathbb{R}^{|\mathcal{D}_{\text{calib}}|+1}$ and intercept $\hat{\beta}_S \in \mathbb{R}$. Accordingly, the fitted RKHS component is of form $\hat{f}_W(x) = \frac{1}{\gamma}\sum_{i\in\mathcal{D}_{\text{calib}}\cup\{i_0\}} \hat{v}_{S,i}\, W(x, X_i)$ As shown in the discussion of Anonymous (2026), the coefficients $\hat{v}_S$ depend *affinely* on the imputed value S and the mapping $S \mapsto \hat{v}_S$ is nondecreasing. Consequently, the event $S \le \hat{f}_S(X_{i_0})$ is equivalent to the linear inequality $\hat{v}_{S,i_0} \le 1-\alpha$. Following the standard randomized conformalization in Anonymous (2026); Gibbs et al. (2025), we replace $1-\alpha$ by a draw $U \sim \text{Unif}(-\alpha,\, 1-\alpha)$, and define the final fitted cutoff by

$$\hat{s}_{i_0} \;=\; \max\{\, S : \; \hat{v}_{S,i_0} \le U \,\}. \tag{7}$$

Equivalently, the final prediction set $\hat{C}(X_{i_0})$ is obtained by plugging the cutoff $s = \hat{s}_{i_0}$ into the set construction $\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; s)$.

Coverage guarantee. The following lemma collects the conditional guarantee delivered by this construction. For each $i_0 \in \mathcal{D}_{aug}$, we write $\hat{f}_W^{\hat{s}_{i_0}}$ for the fitted RKHS function evaluated at the cutoff $\hat{s}_{i_0}$.

Lemma 2 (Coverage; cf. Gibbs et al. (2025); Cherian et al. (2024)) *Let $\mathcal{F}$ be as in equation 4, and assume the pooled indices $\mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{aug}}$ are exchangeable. Suppose the loss $\mathcal{L}(\cdot,\cdot)$ is monotone in its first argument (i.e., if $\mathcal{S}_{i_0}^1 \subseteq \mathcal{S}_{i_0}^2$ then $\mathcal{L}(\mathcal{S}_{i_0}^1, \mathbf{A}_{i_0}) \le \mathcal{L}(\mathcal{S}_{i_0}^2, \mathbf{A}_{i_0})$) and satisfies $\mathcal{L}(\emptyset,\cdot) = 0$. Then, for all $f \in \mathcal{F}$ and all $i_0 \in \mathcal{D}_{\text{aug}}$,*

$$\mathbb{E}\Big[f(X_{i_0})\Big\{\mathbf{1}\Big(\mathcal{L}_\lambda\big(\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; \hat{s}_{i_0}),\, \mathbf{A}_{i_0}\big) \le \rho\Big) - (1-\alpha)\Big\}\Big] \;=\; -\gamma\, \mathbb{E}\Big[\big\langle \hat{f}_W^{\hat{s}_{i_0}},\, f_W \big\rangle_W\Big]\,.$$

The lemma shows that the deviation from the nominal level $(1-\alpha)$ comes from the RKHS inner product involving the learned calibration function, yielding an estimable coverage gap as discussed in Gibbs et al. (2025).

C.2 Proof of Lemma 1

For the localized conformal prediction, we adapt Lemma 2 to a class of covariate shifts induced by the density kernel W. Under the setting of Lemma 1, the tuples

$$(X_1, \{G_{1k}\}_{k\in[K]}, Y_1), \ldots, (X_n, \{G_{nk}\}_{k\in[K]}, Y_n)$$

are independent of $(X_{i_0}, \{G_{i_0k}\}_{k\in[K]}, Y_{i_0}, X'_{i_0})$. By definition of X'_{i_0}, the joint distribution of $(X_{i_0}, \{G_{i_0k}\}_{k\in[K]}, Y_{i_0}, X'_{i_0})$ is given by

$$\begin{aligned} &X_{i_0} \sim P_X, \qquad Y_{i_0} \mid X_{i_0} \sim P_{Y|X}, \qquad (G_{i_0k})_{k=1}^K \mid (X_{i_0}, Y_{i_0}) \sim h(X_{i_0}, Y_{i_0}, \tau), \\ &X'_{i_0} \mid (X_{i_0}, (G_{i_0k})_{k=1}^K, Y_{i_0}) \sim W(X_{i_0}, \cdot), \end{aligned}$$

so that $X'_{i_0} \perp\!\!\!\perp \big((G_{i_0k})_{k=1}^K, Y_{i_0}\big) \,\big|\, X_{i_0}$.

For any realization $x' \in \Omega$, Bayes' rule yields

$$\big(X_{i_0}, (G_{i_0k})_{k=1}^K, Y_{i_0}\big) \,\big|\, X'_{i_0} = x' \;\sim\; \frac{W(x, x')}{\mathbb{E}[\,W(X, x')\,]}\, dP_{(X,\mathbf{G},Y)}(x, \mathbf{G}, y),$$

i.e., the original joint distribution $P_{(X,\mathbf{G},Y)}$ tilted by the weight $W(x, x')$ and renormalized by $\mathbb{E}[W(X, x')]$.

Conditioning on $X'_{i_0} = x'$ and writing $\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; s_{i_0})$ for the set construction, we obtain

$$\begin{aligned} &\mathbb{E}\Big[\mathbf{1}\Big\{\mathcal{L}_\lambda\big(\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; s_{i_0}), \mathbf{A}_{i_0}\big) \le \rho\Big\} - (1-\alpha) \,\Big|\, X'_{i_0} = x'\Big] \\ &\quad= \frac{\mathbb{E}\Big[W(X, x') \cdot \Big(\mathbf{1}\Big\{\mathcal{L}_\lambda\big(\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; s_{i_0}), \mathbf{A}_{i_0}\big) \le \rho\Big\} - (1-\alpha)\Big)\Big]}{\mathbb{E}[\,W(X, x')\,]} \\ &\quad= \frac{-\gamma\, \mathbb{E}\Big[\,\big\langle \hat{f}_W^{\,\hat{s}_{i_0}},\, W(\cdot, x')\big\rangle_W\Big]}{\mathbb{E}[\,W(X, x')\,]} \quad \text{(by Lemma 2)} \\ &\quad= \frac{-\gamma\, \mathbb{E}\Big[\,\hat{f}_W^{\,\hat{s}_{i_0}}(x')\Big]}{\mathbb{E}[\,W(X, x')\,]}. \end{aligned}$$

Using the finite expansion of the fitted RKHS component,

$$\hat{f}_W^{\,\hat{s}_{i_0}}(x') = \hat{\beta}_{\hat{s}_{i_0}} + \frac{1}{\gamma} \sum_{i \in \mathcal{D}_{\mathrm{calib}} \cup \{i_0\}} \hat{v}_{\hat{s}_{i_0}, i}\, W(X_i, x'),$$

we can further write

$$\mathbb{E}\Big[\mathbf{1}\Big\{\mathcal{L}_\lambda\big(\mathcal{S}(\widehat{\mathbf{A}}_{i_0}; s_{i_0}), \mathbf{A}_{i_0}\big) \le \rho\Big\} - (1-\alpha) \,\Big|\, X'_{i_0} = x'\Big] = \frac{-\,\mathbb{E}\Big[\sum_{i \in \mathcal{D}_{\mathrm{calib}} \cup \{i_0\}} \hat{v}_{\hat{s}_{i_0}, i}\, W(X_i, x')\Big]}{\mathbb{E}[\,W(X, x')\,]},$$

which completes the localized reweighting conformal prediction.

D Experimental details

D.1 Clinical Text Classification

This dataset consists of natural language descriptions of symptoms annotated with 22 corresponding diagnoses (Gretel AI, 2024). In total, it contains 1,065 English-language symptom descriptions, of which 853 (80%) are allocated for training and 212 (20%) for testing.

As described in the main text, each training symptom description is extended with five additional sentences using `GPT-4.1 nano` with temperature 1.5. Each augmented sentence is paired with the original label and treated as a new sample. To assess quality, we applied `Gemini-2.5-Pro` to generations from 100 randomly selected documents (yielding 500 new samples) and `Gemini-2.5-Flash` to generations from all 853 training documents (yielding 4,265 new samples).

For evaluating augmentation methods, we fixed a set of 100 documents with scores from Gemini-2.5-pro and randomly sampled an additional 400 documents from the training set. From these 500 documents, we applied different filtering strategies. In particular, for CondCP, we applied Latent Dirichlet Allocation (LDA) with 18 latent mixtures to the entire training set, where the number of mixtures was chosen based on log-likelihood validation, in order to estimate the latent mixture representation of each document. The procedure was repeated 20 times, and we reported precision, recall, and accuracy.

Prompt for Data Generation

```
You are given a description of a disease.

Description: {symptom}

Task: Extend the symptom description with additional details that
still plausibly describe the SAME disease.
- Write EXACTLY 5 sentences.
- Do not copy wording from the original; paraphrase and add
plausible details consistent with the same condition.
- Avoid lists, bullets, headings, or numbering; just 5 full
sentences in a single paragraph.
- No disclaimers, no citations, no markdown.
```

Prompt for Evaluation

```
You are evaluating individual symptom descriptions for diseases.

Scoring instructions:
- Assign each description a score between 0 and 1, rounded to two
decimal places.
- Criteria: The description should plausibly match the specified
disease and avoid confusion with other diseases.
- Use the full 0-1 range: 1 = perfectly clear, specific, and
accurate; 0 = completely unusable.
- 0.5 is the threshold: any description with a score <= 0.5 should
be dropped to prevent misclassification.

For reference, here is the complete list of possible diseases:
{disease_ls}

Output requirements:
- Output only the scores, one per line, in the same order as the
input  cases.
- Do not include explanations, text, or formatting other than the
numeric scores.

Case : Disease: {diag} : Symptom: {symp}
```

D.2 Abstract Topic Classification

ArXiv hosts more than 1.5 million articles across diverse fields. For this analysis, we use a random sample of 1,000 abstracts published after January 1, 2021, distributed evenly across five statistical categories: statistical methodology, statistical machine learning, statistical applications, statistical computation, and statistical theory (Artgor, 2019). The classification task is challenging because these categories are closely related. We split the 1,000 abstracts into 800 for training and 200 for testing.

Each training abstract is extended with six additional sentences using `GPT-4.1 nano` with temperature 1.5. Every two consecutive sentences are grouped as a new sample, paired with the label of the original abstract. To assess quality, we applied `Gemini-2.5-Pro` to generations from 100 randomly selected abstracts (yielding 300 new samples) and `Gemini-2.5-Flash` to generations from all 800 training abstracts (yielding 2,400 new samples). For each abstract X_i, with extended groups $\{G_{ik}\}_{k=1}^{3}$, Gemini-pro scores $\{A_{ik}\}$, and Gemini-flash scores $\{\hat{A}_{ik}\}$, we define

$$S(\hat{\mathbf{A}}_i, \mathbf{A}_i) = \inf\left\{\tau : \left|\{G_{ik} : 1 \leq k \leq 3,\ \hat{A}_{ik} \geq \tau,\ A_{ik} < 0.5\}\right| = 0\right\}.$$

The evaluation procedure follows the same protocol as in clinical text classification. We fixed a set of 100 documents with scores from `Gemini-2.5-Pro` and randomly sampled an additional 200 documents from the training set. From these 300 documents, we applied different filtering strategies. Latent Dirichlet Allocation (LDA) was then performed with 5 latent mixtures, consistent with the number of categories in the dataset. Just as the clinical text example, we fine-tune a small classifier (`distilbert-base-uncased`) for topic prediction.

Prompt for Data Generation

```
You are given a statistical abstract.

Abstract: {abstract}

Task: Extend the abstract with additional details that remain
consistent with the SAME statistical topic.
- Write EXACTLY 6 sentences.
- Do not copy wording from the original; paraphrase and add
plausible extensions consistent with the same subject.
- Avoid lists, bullets, headings, or numbering; just 6 full
sentences in a single paragraph.
- No disclaimers, no citations, no markdown.
```

Prompt for Evaluation

```
You are evaluating individual sentences from extended statistical
abstracts.

Scoring instructions:
- Assign each sentence a score between 0 and 1, rounded to two
decimal places.
- Criteria: The sentence should plausibly match the specified topic,
remain coherent, and avoid drifting into other topics from the list.
- Use the full 0-1 range: 1 = perfectly clear, on-topic, and
informative; 0 = completely unusable.
- 0.5 is the threshold: any sentence with a score <= 0.5 should be
dropped to prevent topic drift.

Output requirements:
- Output only the scores, one per line, in the same order as the
input cases.
- Do not include explanations, text, or formatting other than the
numeric scores.

Case : Topic: {topic} : Sentences: {sent}
```

D.3 Twitter Message Sentiment Analysis

The dataset (Kadara, 2018) contains text segments from Twitter messages, each labeled with the predominant emotion expressed. The emotions are categorized into six classes: sadness, joy, love, anger, fear, and surprise. We randomly sampled 1,200 messages, evenly distributed across the six categories, and split them into 1,000 for training and 200 for testing.

Each training message was extended with five additional sentences using `GPT-4.1 nano` with temperature 1.5, with each sentence paired to the original label as a new sample. For evaluation, `Gemini-2.5-Pro` scored generations from 100 documents (500 samples), while `Gemini-2.5-Flash` covered all 1,000 training documents (5,000 samples). The evaluation procedure follows the same protocol as in clinical text classification: we fixed a set of 100 documents with scores from `Gemini-2.5-Pro` and randomly sampled an additional 200 documents from the training set. We define the non-conformity score as

$$S(\hat{\mathbf{A}}_i, \mathbf{A}_i) = \inf\left\{\tau : \left|\{G_{ik} : 1 \le k \le 5,\ \hat{A}_{ik} \ge \tau,\ A_{ik} < 0.5\}\right| \le 1\right\}.$$

The remaining steps were identical to the previous cases, except that here we applied LDA with six mixture components.

Prompt for Data Generation

```
You are given a short Twitter message.

Message: {tweet}

Task: Extend the message with additional content that preserves the
SAME sentiment and topic.
- Write EXACTLY 5 sentences.
- Paraphrase and expand naturally; do not copy wording from the
original.
- Vary phrasing, tone, and detail while remaining consistent with
the sentiment.
- Avoid lists, bullets, hashtags, mentions, links, or numbering;
produce 5 full sentences in a single paragraph.
- No disclaimers, citations, or markdown.
```

Prompt for Evaluation

```
You are evaluating individual sentences for sentiment consistency.

Scoring instructions:
- Assign each sentence a score between 0 and 1, rounded to two
decimal places.
- Criteria: The sentence should clearly reflect the SPECIFIED
sentiment, remain coherent, and avoid conflicting emotions.
- Use the full 0-1 range: 1 = perfectly consistent and natural;
0 = completely unusable.
- 0.5 is the threshold: any sentence with a score <= 0.5 should
be excluded.

Output requirements:
- Output only the scores, one per line, in the same order as the
input cases.
- Do not include explanations, text, or formatting beyond the
numeric scores.

Case: Sentiment: {senti} Sentence: {sent}
```

D.4 Diversity of Augmented Text

Diversity of Selected Augmentations. To evaluate the diversity of the augmentation techniques, we compute their Shannon entropy. The results are reported in Table 2 for the Diagnosis, Abstract and Sentiment Datasets. Overall, we find that augmentation generally increases the diversity of the training data (e.g., the Diagnosis dataset features an unaugmented diversity of 6.92, compared to 8.02 for the CondCP filter and 8.52 for the unfiltered dataset). The filtered versions typically show lower diversity compared with the unfiltered versions since poor-quality generations are excluded. We also note that the CondCP-filtered generations exhibit lower diversity than those filtered by other algorithms. This outcome is expected, given the nature of conformal prediction. Nonetheless, the reduction in diversity is relatively small, highlighting that the CondCP filter effectively preserves the essential information contained in the training data.

Sensitivity to the choice of ρ and λ: To assess the sensitivity of the proposed CondCP approach to the choice of hyperparameters (e.g. ρ and λ), we report the Shannon entropy of the CondCP-filtered results under different hyperparameter configurations in Table 3. Overall, larger values of λ and smaller values of ρ tend to reduce diversity. This highlights the fact that these hyperparameters should be chosen to balance diversity against faithfulness to the original data distribution. Their sensitivity depends on the dataset as well as on both the gold-standard and surrogate diversity measures.

Data	Unaugmented	Unfiltered	Flash Filter	Hybrid Filter	CondCP Filter
Diagnosis	6.92	8.52	8.14	8.14	8.02
Abstract	9.64	9.83	9.81	9.81	9.76
Sentiment Analysis	8.69	9.34	9.23	9.23	8.96

Table 2: Shannon entropy of augmented data across datasets under different augmentation methods

Diagnosis						
	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
$\rho = 0$	7.67	7.53	7.42	7.44	7.36	7.32
$\rho = 1$	8.14	8.10	8.02	7.94	7.87	7.81
$\rho = 2$	8.19	8.16	8.13	8.13	8.05	8.03
Abstract						
	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
$\rho = 0$	9.76	9.76	9.76	9.76	9.76	9.76
$\rho = 1$	9.77	9.77	9.77	9.77	9.76	9.76
$\rho = 2$	9.77	9.77	9.77	9.77	9.76	9.76
Sentiment Analysis						
	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
$\rho = 0$	8.99	8.94	8.89	8.90	8.89	8.87
$\rho = 1$	9.08	9.01	8.96	8.96	8.94	9.01
$\rho = 2$	9.08	9.07	9.09	8.99	8.98	8.97

Table 3: Shannon entropy of CondCP filtered data across different λ and ρ

D.5 Comparison between Gemini-2.5-Pro scores and Gemini-2.5-Flash scores

Figure 5 presents a comparison of evaluation scores between Gemini-2.5-Pro and Gemini-2.5-Flash across datasets. While the two scores show a clear positive correlation, they are not perfectly aligned.

D.6 Experimental results for text data under low-temperature generation

The LLM-augmented experiments presented in the main text were conducted with a high generation temperature of 1.5. For completeness, we report in Figure 6 the corresponding results obtained under a low-temperature setting with generation temperature 0.3. As anticipated, CondCP filter offers limited improvement in this regime due to the substantially reduced diversity of generated samples: at low temperature, the LLM predominantly produces highly frequent or canonical outputs, leaving little variation for the filtering mechanism to act upon. Consequently, overall performance is worse than the high-temperature setting with CondCP filtering reported in the main text.

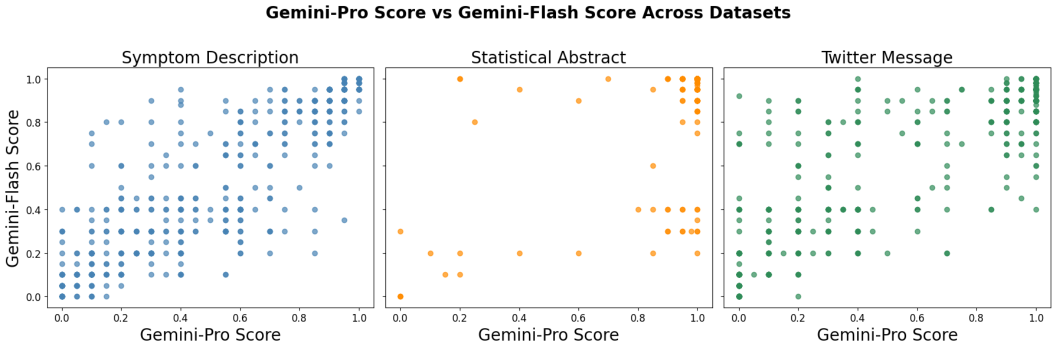

Figure 5: Scatter plots comparing Gemini-Pro and Gemini-Flash scores for symptom descriptions, statistical abstracts, and Twitter messages datasets.,

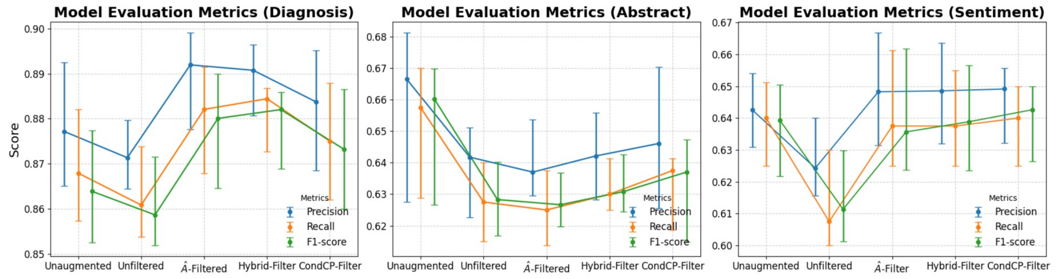

Figure 6: Evaluation of different data augmentation methods on diagnosis prediction, abstract topic prediction, and Twitter message sentiment prediction with generation temperature 0.3.

D.7 Assessment of Risk-Control Violations

For each text dataset, we partition the samples with `Gemini-2.5-pro` scores into 10 folds. In each split, we use 9 folds to train CondCP and evaluate the empirical violation rate on the remaining fold, which is the frequency with which the number of low-quality generations exceeds the allowed threshold ρ. The results are shown in Figure 7 , which illustrates the risk control achieved by our method.

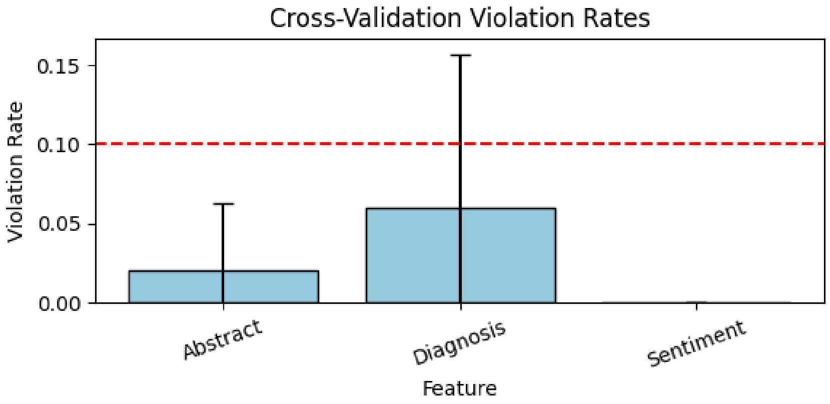

Figure 7: Empirical violation rate for allowing at most ρ low quality generations. Error bars represent ± 1 standard deviation.

D.8 Additional Experiments on Imbalanced Classification

Dataset	Hidden Dim	Context Dim (c)	Latent Dim	Epochs	Learning Rate
Thyroid	64	32	32	100	1×10^{-3}
Credit Card	64	32	32	100	1×10^{-3}
MNIST-7 vs. rest	64	32	32	100	1×10^{-3}

Table 4: CVAE hyperparameters across datasets. Hidden dimension refers to the encoder/decoder width, context dimension is the size of the seed-conditioned embedding $c = h_\psi(x)$, and latent dimension is the size of the stochastic latent variable z.

In this appendix, we offer a deep dive into some of the examples presented in this paper for the imbalanced classification results (Section 4.2).

D.8.1 Experiment Details

Datasets. We evaluate our framework on three benchmark datasets that cover a wide spectrum of imbalance ratios, dimensionalities, and application domains:

- **European Credit-Card Fraud** (Kaggle): 284,807 transactions with 492 frauds (0.17% positives). Each record contains transaction time, amount, and 28 PCA-compressed features (V1–V28). This dataset is widely used as a canonical benchmark for extreme class imbalance.
- **Thyroid Disease** (OpenML-38): 2,644 patient records with 6.4% positives. Features include demographic covariates, hormone levels, and binary medical indicators. This dataset represents a typical medical diagnosis problem with moderate imbalance.
- **MNIST-7 vs. Rest** (OpenML-554): 70,000 handwritten digits recast into a binary task of distinguishing "7" (10.9% positives) from all other digits. While less imbalanced, this high-dimensional vision-like dataset provides a contrasting baseline where signal is strong and plentiful.

Conditional VAE (CVAE). We generate minority samples with a Conditional Variational Autoencoder (CVAE) (Kingma & Welling, 2013; Sohn et al., 2015) that is *conditioned on an actual minority seed*. Let $h_\psi : \mathbb{R}^d \to \mathbb{R}^c$ be a small *context net* that maps a reference minority instance x to a context vector $c = h_\psi(x)$. The encoder receives the concatenation $[x, c]$ and outputs a Gaussian $q_\phi(z \mid x, c) = \mathcal{N}(\mu_\phi(x, c), \mathrm{diag}(\sigma^2_\phi(x, c)))$; the decoder reconstructs x from $[z, c]$ via $p_\theta(x \mid z, c)$. Both encoder and decoder are two-layer MLPs with ReLU activation.

We train on minority data only ($y = 1$) with the ELBO using Adam optimizer for 100+ epochs (features are MinMax-scaled). At generation time, given a seed x_s we compute $c_s = h_\psi(x_s)$ and draw K candidates by sampling $z_k \sim \mathcal{N}(0, \tau^2 I)$ (with $\tau > 0$ and decoding $g_k = g_\theta\big(z_k, c_s\big)$. This *seed-conditional* design produces local, seed-aware variations that stay on the minority manifold with controlled dispersion τ. The raw candidates are then quality-scored and filtered by our $\widehat{A}$ regressor and conformal thresholds before being added for training a classifier. See the detailed choice of architecture in Table 4.

Conformal Prediction conditioned on Latent Representation We apply conditional conformal filtering that operates in a learned latent representation of the data (Anonymous, 2026). Specifically, we project the feature space into a lower-dimensional latent embedding using Principal Component Analysis (PCA) before applying the conformal calibration step. For each dataset, we tune the latent dimension to reflect its scale: 2 for *Thyroid* and *Credit Card Fraud*, and 16 for *MNIST-7*.

D.8.2 Evaluating the quality of the (selected) generations

We examine the MNIST 7 example, an imbalanced classification task where MNIST digits are classified as 7 or not 7, with 7s being underrepresented. We evaluate the effect of temperature and selection on the diversity of the generated samples, as the results in this example should be easily interpretable visually.

Understanding the effect of temperature on the diversity of the samples. Figure 8 highlights a few examples of generations of the digit 7 for different temperatures τ. In low temperature settings (e.g. $\tau = 0.1$), the model generates almost identical samples. In moderately high temperature settings ($\tau = 2$), the model starts to generate more variable shapes of the digit 7. However, as the temperature becomes too high ($\tau = 10$), the synthetic data become extremely noisy.

Figure 9 and 10 further illustrate the temperature effect through principal component analysis, comparing real and generated data. As the temperature (τ) increases, the synthetic samples gradually explore a wider area of the real data. However, excessively high temperatures (e.g., $\tau = 10$) cause the generator to sample outside the MNIST distribution, resulting in points that do not align with the original data's structure.

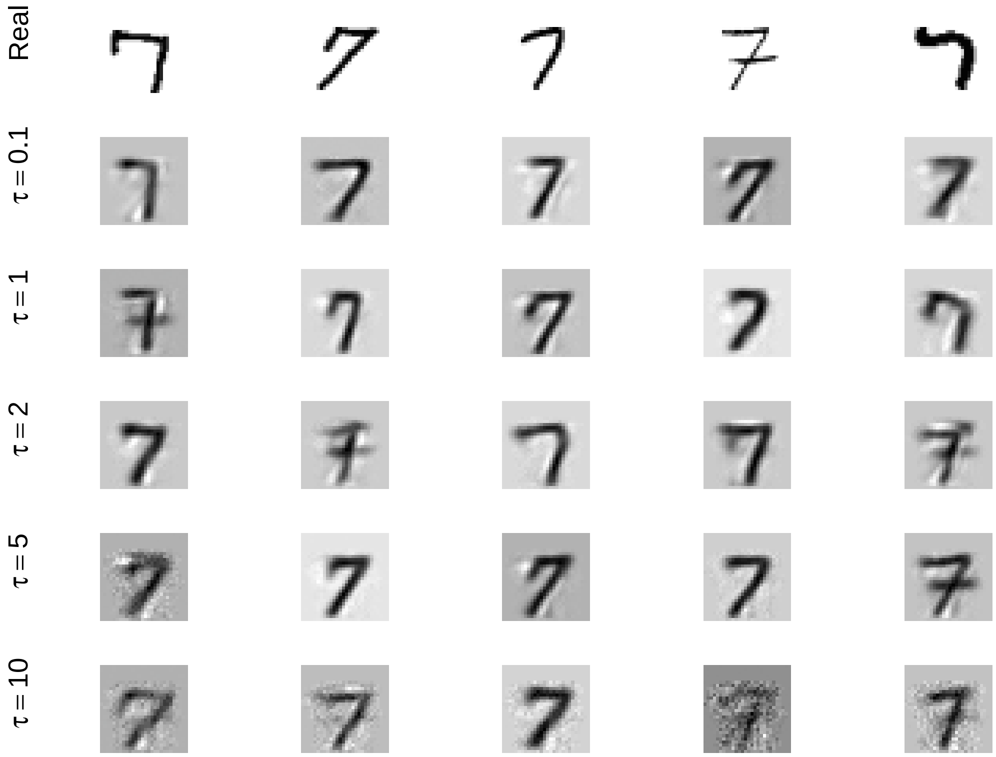

Figure 8: Generated minority digit (7) by the VAE model for varying temperature (τ).

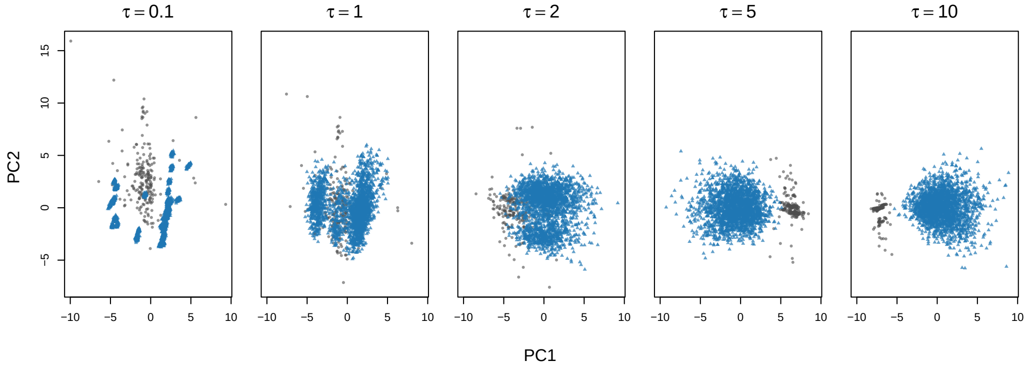

Figure 10: PCA visualization of real samples from thyroid dataset and VAE-generated samples under different temperature values τ. The gray circle points denote the real data, and the blue triangular points denote the generated data.

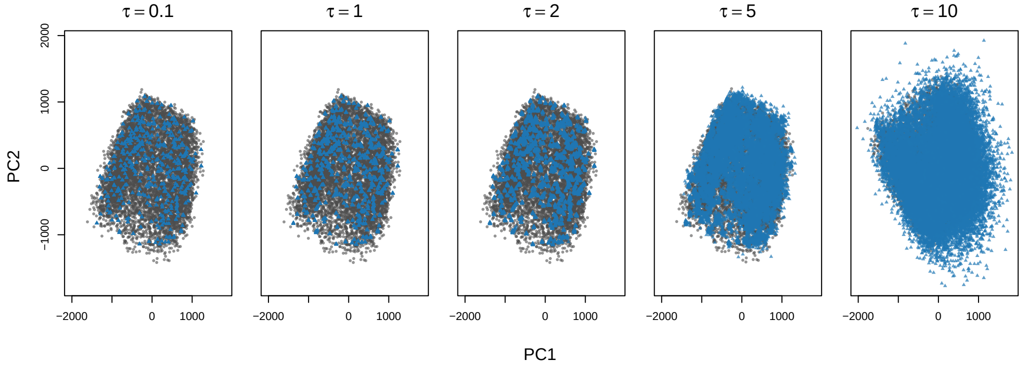

Figure 9: PCA visualization of real MNIST-7 digits and VAE-generated samples under different temperature values τ. The gray circle points denote the real data, and the blue triangular points denote the generated data.

To quantify these effects and directly examine the role of temperature, we measure the diversity of the (unfiltered) generated samples for different values of the parameter τ using the stable rank. The results, shown in Table 5, confirm that increasing the sampling temperature produces higher-diversity synthetic data; however, but this diversity is uncontrolled. Beyond moderate temperatures, the generator begins to sample outside the real data manifold, producing overly noisy or implausible examples. Figure 8 illustrates this phenomenon, where high-temperature samples turn out to be overly noisy. This supports the motivation stated in the introduction: while higher temperatures can, in principle, expose rare modes, naively relying on high-temperature augmentation is harmful because it injects low-quality, out-of-support samples.

τ	**Thyroid**	**Credit Card Fraud**	**MNIST 7**
0.1	6.072 ± 0.689	1.960 ± 0.039	12.048 ± 0.754
1	8.238 ± 0.773	1.989 ± 0.034	12.020 ± 0.745
2	8.026 ± 1.075	1.998 ± 0.095	12.125 ± 0.707

Table 5: Diversity measure by Stable Rank of unfiltered augmentation with varying temperature (τ). Mean and standard deviation computed across different splits (seed). The smaller τ will lead to generate the samples closer to the existing point, and the higher τ will lead to more noisy generation.

Understanding the effect of selection. Figures 11 and 12 show examples of accepted and rejected samples, respectively. As expected, we see that the rejected samples feature more low quality (extremely blurry and jagged) sevens, compared to the selected ones: the condCP selection seems to select samples that are more realistic.

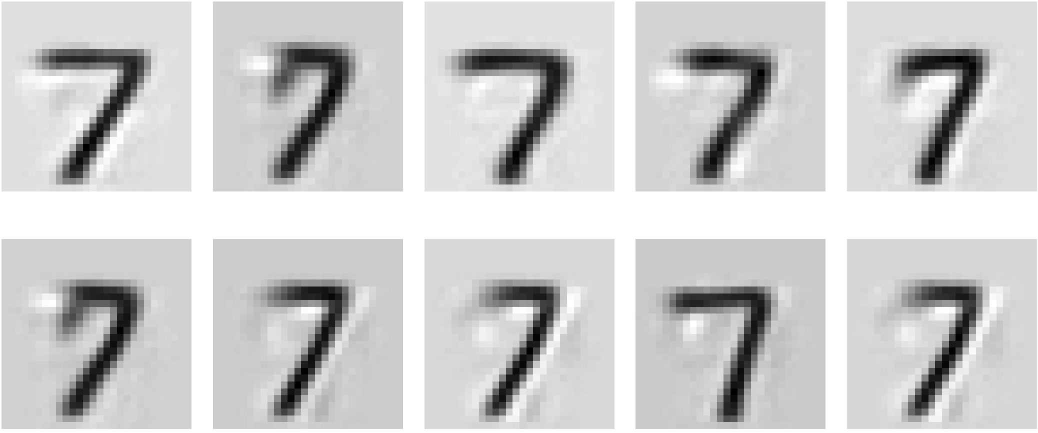

Figure 11: Examples of accepted generations by CondCP with $\tau = 0.1, \lambda = 0.5, \rho = 2$.

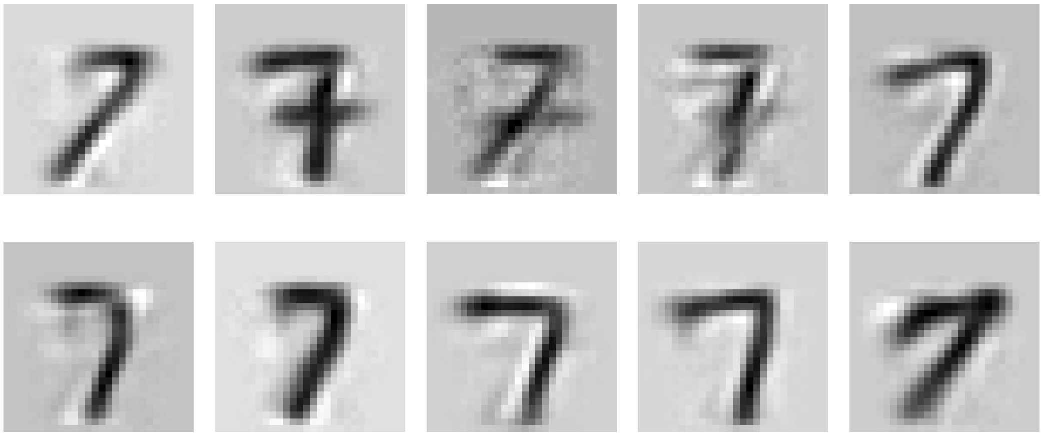

Figure 12: Examples of generations filtered out by CondCP with $\tau = 0.1, \lambda = 0.5, \rho = 2$.

This effect extends to other datasets. In Figure 13, we visualize the effect of the selection on the Thyroid dataset. To this end, we first extracted important predictors using regression on training dataset, and the three variables, `on thyroxine` (binary treatment indicator), `T3` (serum triiodothyronine level), and `TT4` (total thyroxine level), emerged as significant predictors. The plots show, for each variable and each filtering strength λ, how the distribution of accepted synthetic samples aligns with the real minority-class distribution. Across features, accepted samples (blue) consistently match the true minority distribution (black) better than rejected samples (gray), demonstrating that the filtering criterion preferentially retains synthetic points that lie on the true data-support for the minority class.

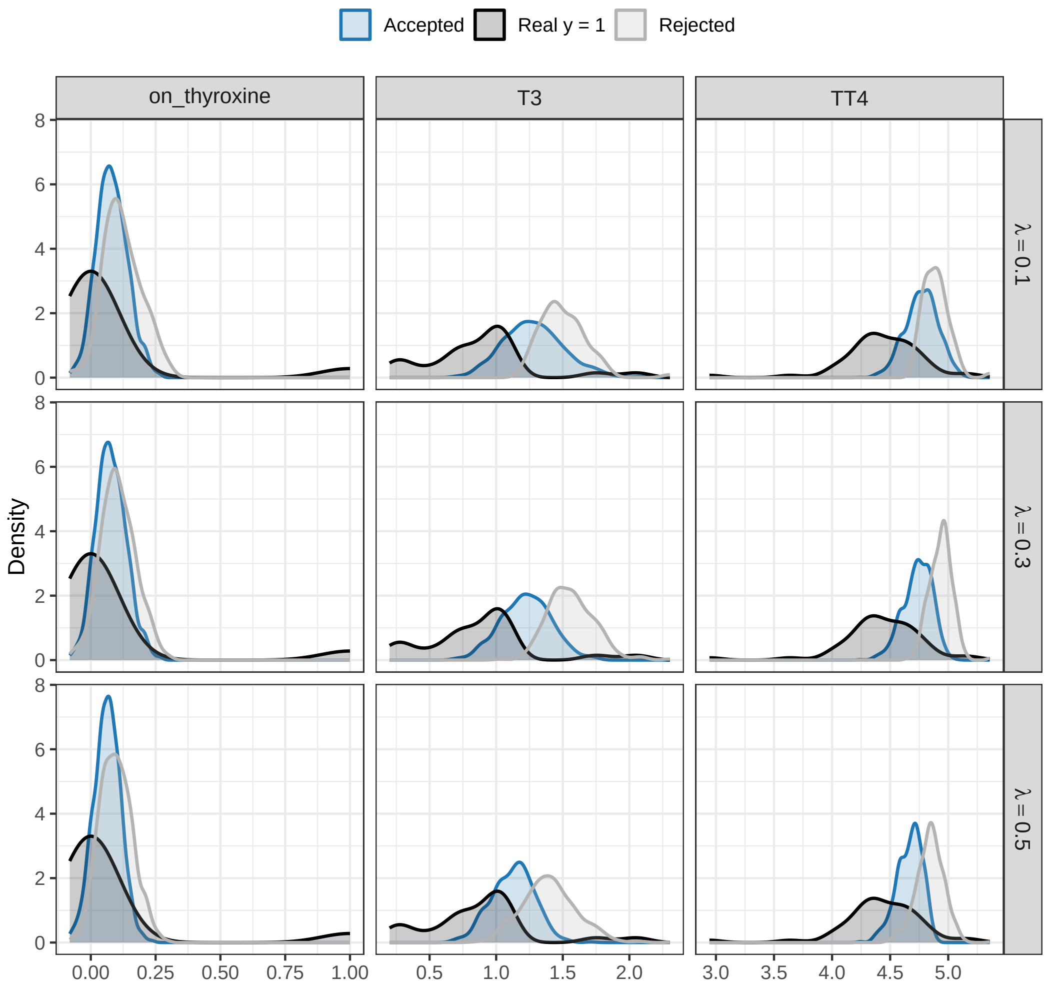

Figure 13: Using the thyroid dataset, we first fit a simple logistic regression on 60% of the real data to predict the minority thyroid-disease class. Three variables, on thyroxine (binary treatment indicator), T3 (serum triiodothyronine level), and TT4 (total thyroxine level), emerged as significant predictors. The tolerance parameter ρ is fixed to be 2.