Modeling the $\mathcal{Q}$-Diversity in a Min-max Play Game for Robust Optimization

Ting Wu[1], **Rui Zheng**[1], **Tao Gui**[2*], **Qi Zhang**[1,3], **Xuanjing Huang**[1]
[1]School of Computer Science, Fudan University
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]Shanghai Key Laboratory of Intelligent Information Processing
tingwu21@m.fudan.edu.cn
{rzheng20,tgui,qz,xjhuang}@fudan.edu.cn

Abstract

Models trained via empirical risk minimization (ERM) are revealed to easily rely on spurious correlations, resulting in poor model generalization. Group distributionally robust optimization (group DRO) can alleviate this problem by minimizing the worst-case loss over pre-defined groups. While promising, in practice factors like expensive annotations and privacy preclude the availability of group labels. More crucially, when taking a closer look at the failure modes of out-of-distribution generalization, the typical procedure of reweighting in group DRO loses efficiency. Hinged on the limitations, in this work, we reformulate the group DRO framework by proposing $\mathcal{Q}$-Diversity. Characterized by an interactive training mode, $\mathcal{Q}$-Diversity relaxes the group identification from annotation into direct parameterization. Furthermore, a novel mixing strategy across groups is presented to diversify the under-represented groups. In a series of experiments on both synthetic and real-world text classification tasks, results demonstrate that $\mathcal{Q}$-Diversity can consistently improve worst-case accuracy under different distributional shifts, outperforming state-of-the-art alternatives [1].

1 Introduction

Deep learning models trained with empirical risk minimization (ERM) often exhibit drops in accuracy when confronted with data from domains that are under-represented in their training data (Arjovsky et al., 2019; Creager et al., 2021). Distributionally robust optimization (DRO) (Duchi et al., 2016) provides a natural solution to the issue by replacing the expected risk under a single distribution p with the worst expected risk over a predetermined family of distributions $\mathcal{Q}$.

However, in DRO, considering that direct gradient descent is hard to satisfy (Hu et al., 2018), how to model and optimize over $\mathcal{Q}$ poses a key challenge. In this way, group DRO (Sagawa et al., 2020) is emerging as a methodology for constructing a realistic set of possible $\mathcal{Q}$ under the annotated groups. Crucially, robust optimization over worst groups becomes an active area of research.

In general, the practical usage of group DRO requires that group identities should be fully known. Therefore, it can model $\mathcal{Q}$ by upweighting or downweighting the average loss of different groups through the course of training. Nevertheless, a key obstacle is that the under-represented groups are often unlabeled, or even unidentified. This makes even detecting such performance gaps, let alone mitigating them, a challenging problem. What's worse, with the lack of group labels, it becomes infeasible to compute the worst group loss so that the $\mathcal{Q}$ modeling fails to be established. Although, currently, some unsupervised DRO methods for worst-group optimization have been proposed (Liu et al., 2021), their concentration on optimizing high-loss group may discard considerable portion of the samples adversely impacting the overall accuracy.

Shedding light on the critical challenge of current group DRO framework, we therefore present a novel unsupervised method as $\mathcal{Q}$-Diversity for worst-group optimization. To realize the group identification without any annotations, we propose to parameterize a classifier as the group assigner for the attainment of group labels. In particular, by alternatively training the group assigner and final class predictor, we formalize an interactive training mode that allows the identification procedure feasible. Intriguingly, we can treat the classification loss from the predictor as a direct supervision to guide the assigner for better group labeling. With the well-estimated groups, accordingly, the predictor can perform better on the worst group. When achieving the pseudo-labeled groups, the typical procedure is to model $\mathcal{Q}$ by reweighting the training losses of different groups. Nevertheless, in

[1]Corresponding author.
[1]Our code and data are available at https://github.com/CuteyThyme/Q-Diversity.git.

Findings of the Association for Computational Linguistics: ACL 2023, pages 100–113

theory, we point out that simply reweighting can not handle OOD failure modes as more diversified samples are needed. Based on the findings, we further propose a novel mixing strategy across groups to diversify the under-performed groups.

To verify the robust optimization capability of $\mathcal{Q}$-Diversity, we conduct a series of experiments on both synthetic and real-world datasets, offering a wide range of challenging benchmarks. All the empirical results show our method not only outperforms other strong group DRO strategies by a large margin, but also achieves consistent improvements on different OOD test sets. Compared to these optimization methods either supervised or unsupervised, $\mathcal{Q}$-Diversity shows great superiority with high efficiency. Altogether, our contributions can be summarized as follows:

• **Methodological Innovations:** In Section 3, we propose $\mathcal{Q}$-Diversity, a group-unlabeled approach that aims to improve the utility for worst case. Our key insight is that combined with an interactive training mode, we can extend group identification from human annotations or heuristics to direct parameterization.

• **Empirical Benefits:** In Section 4, we evaluate $\mathcal{Q}$-Diversity on both synthetic and real-world datasets. Experimental results show that $\mathcal{Q}$-Diversity yields significant accuracy improvements for the worst group, and diversified by group mixing, it even outperforms the supervised baseline.

• **Understanding $\mathcal{Q}$-Diversity:** In Section 5, we conduct a thorough experimental analysis and present the generalization capacity of $\mathcal{Q}$-Diversity under various distribution shifts.

2 Preliminary: Robust Optimization

2.1 Problem Setup

We consider the typical text classification problem of predicting labels $y \in \mathcal{Y}$ from input texts $x \in \mathcal{X}$, and training data $\mathcal{D}$ is assumed to be drawn from the joint distribution $P(\mathcal{X}, \mathcal{Y})$.

2.2 Distributionally Robust Optimization

ERM Principle. Given a model family Θ and a loss function $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the standard goal of empirical risk minimization is to find a model $\theta \in \Theta$ that minimizes the expected loss over the empirical distribution $\hat{P}$ drawn *i.i.d* from P:

$$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta; (x, y)] \quad (1)$$

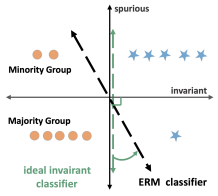

Figure 1: Geometric skew.

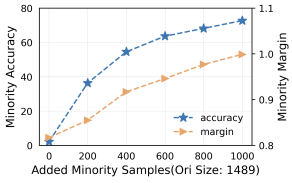

Figure 2: Group Diversity.

When encountering data sampled in the distribution different from P, model performance suffers significantly. Under the circumstances, distributionally robust optimization (Duchi et al., 2016) provides a natural solution by minimizing the worst-case expected risk under a pre-determined family of distributions $\mathcal{Q}$, called the *uncertainty set*:

$$\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) := \max_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}[\ell(\theta; (x, y))] \right\} \quad (2)$$

The uncertainty set $\mathcal{Q}$ requires encoding a wide set of distributional shifts for model robustness improvement. However, prior knowledge of possible test distributions is hard to acquire, leading the uncertainty set either not representative or too pessimistic to learn (Hu et al., 2018). On the other hand, direct gradient descent on $\mathcal{Q}$ often suffers from instability due to the large variance of the gradients and complex hyper-parameter tuning (Balduzzi et al., 2018).

2.3 Practical Group DRO

To overcome these challenges in robust optimization, Sagawa et al. (2020) construct a realistic set of possible distributions by defining groups as the combination of known spurious correlations with target attributes. Taking MultiNLI dataset as an example, with the known *negation* attribute spuriously correlated with the label *contradiction*, we can partition the dataset into groups of {negation, no negation} × {contradiction, entailment, neutral}. By translating training distribution P into a mixture of m groups P_g, the objective of group DRO can be formulated as a minimization of the empirical worst-group risk over m groups:

$$\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta; (x, y))] \right\} \quad (3)$$

where each group $\hat{P}_g$ is an empirical distribution over the training data. Therefore, the uncertainty set $\mathcal{Q}$ is modeled as any mixture of these groups, *i.e.*, $\mathcal{Q} := \{\sum_{g=1}^{m} q_g P_g\}$.

Min-max Play Game. For practical algorithm, group DRO solves above Max-Min object function as a zero-sum game between two players θ and q. Ideally, the player q can be viewed as the weighted distribution for m groups that models the uncertainty set $\mathcal{Q}$. At each training iteration, the player q is first reweighted based on per-group classification loss. Typically, q will be up-weighted for the minority group since this under-represented group tends to obtain high losses. Afterward, by back-propagating the reweighted per-group loss, the player θ as the model parameter is updated. Altogether, for the general group DRO, it is shaped as following two-stage framework:

$$\min_{\theta} \max_{q} \sum_{j=1}^{M} q_j \underbrace{\Big[\frac{\overbrace{\sum_{i=1}^{N} \mathbb{1}\{g_i = j\}}^{\text{stage 1. group identification}} \quad \ell(\theta; (x, y))}{\sum_{i=1}^{N} \mathbb{1}\{g_i = j\}} \Big]}_{\text{stage 2. group reweighting}}$$

$$\text{with} \quad q_j \leftarrow q_j \exp(\ell(\theta^{(t-1)}; (x, y)) \tag{4}$$

The Dark Side. Although the formulation of group DRO keeps the choice of uncertainty set $\mathcal{Q}$ exactly tractable, in terms of the step-by-step procedures, two main issues stand out. **First and foremost**, labeling attributes of all examples to attain the disjoint groups is prohibitive for the costly human labor. **Second**, while intuitive, recent studies (Nagarajan et al., 2021; Nguyen et al., 2021) for understanding OOD generalization have revealed that simply reweighting can not handle the failure modes of distributional shifts. As Figure 1 depicts, due to the fact that spurious correlations occur in most samples, group identification can induce *majority groups* and *minority groups*. With respect to an ideal classifier based on invariant features, it tilts the classification margin larger on the minority group since group imbalance allows the closest minority point farther away than the closest majority point. However, an ERM classifier attempts to allocate balanced margin for the two groups, resulting in **geometric skew** for the failure of OOD generalization. Crucially, Nguyen et al. (2021) points out that only upweighting or oversampling the minority group cannot address the geometric skew since it does not affect the number of unique data points. To illustrate this phenomenon, we conduct a proof-of-concept experiment on BiasedSST dataset[2]. As shown in Figure 2, with more minority samples synthesized for diversity, classification margin on the minority group is increased to mitigate geometric skew, and meanwhile, the robust accuracy is improved significantly.

[2]Refer Section 4.2 to see details on the synthetic dataset.

3 $\mathcal{Q}$-Diversity Modeling

Overview. We address two above limitations of group DRO by proposing $\mathcal{Q}$-Diversity. In our setup, we improve the classification accuracy of minority groups without explicit group annotations. The overall paradigm is depicted in Figure 3. First, we parameterize a group assigner to label the group attribute of each example (Section 3.1). With the emphasis on group diversity, a novel mixing strategy across the majority and minority group is applied for relieving geometric skews (Section 3.2). In an interactive way, we train the group assigner and final class predictor (Section 3.3), allowing them to guide each other for better robust accuracy.

3.1 Parameterizing Assigner for Group Identification

The prerequisite for optimizing the worst group is to obtain well-defined groups. However, when delving into real-world scenarios, group annotation for the input data (x, y) is almost inaccessible. Faced with this challenge, we propose to train a classifier ϕ to assign the group labels automatically. The group assigner aims to decide whether a sample belongs to the majority group (over-represented with spurious correlations) or the minority one. More formally, we can denote the probability estimate of the assigner on the group attribute g as $\hat{p}(g|x, y)$. The assigned group label $\hat{g} = \arg\max \hat{p}(g|x, y)$ can be viewed as a list of the latent binary variables, where each $\hat{g} \in \{0, 1\}$.

Label Balance Regularization. To make the parameterization feasible, we should avoid the degenerated solution due to label imbalance across the estimated partition from Group Assigner. Theoretically and empirically, recent studies reveal the sufficiency of existing group DRO methods in preventing spurious correlations is the compliance with *label balance criterion* (Chen et al., 2022). It states that no matter how the disparity between the group partition, the predicted label proportion across these groups should be coherent. Adhered to this criterion, we regulate the decision of the Group

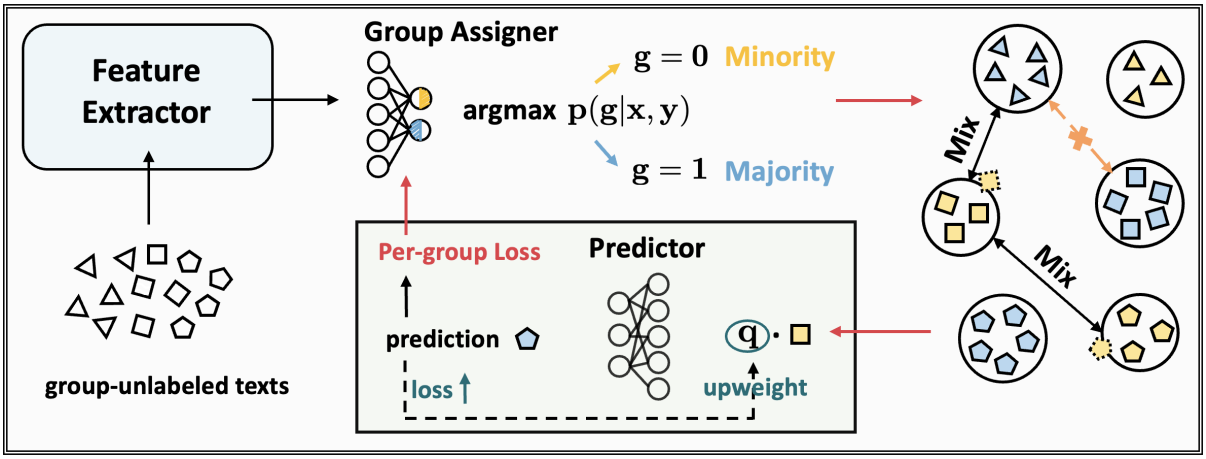

Figure 3: End-to-end learning framework of $\mathcal{Q}$-Diversity for robust optimization.

Assigner with following objective:

$$\mathcal{L}_{\text{bal}} = KL(P(y|\hat{g}=1)\|P(y)) + KL(P(y|\hat{g}=0)\|P(y)) \tag{5}$$

where KL is the Kullback–Leibler divergence. This regularization makes intuitive sense as we would like to push label marginals in the estimated majority group $P(y|g=1)$ and the minority group $P(y|g=0)$ close to the original label marginal $P(y)$ in the training data $\mathcal{D}$. Practically, we apply the Bayes rule to compute these conditional label marginals directly from the Assigner's decisions:

$$P(y|\hat{g}=1) = \frac{\sum_i \mathbb{1}_y(y_i) P(g_i=1|x_i,y_i)}{\sum_i P(g_i=1|x_i,y_i)}$$
$$P(y|\hat{g}=0) = \frac{\sum_i \mathbb{1}_y(y_i) P(g_i=0|x_i,y_i)}{\sum_i P(g_i=0|x_i,y_i)} \tag{6}$$

3.2 Reweighting Player q under Group Mixing

Assuming that from the Group Assigner, each sample (x, y) has been successfully assigned an estimated group attribute $\hat{g}$. Similar to the supervised group DRO, we can partition training data $\mathcal{D}$ into m groups $\mathcal{G}$, and $\mathcal{G}^+$, $\mathcal{G}^-$ denote the majority and minority groups respectively.

As we illustrated in Section 2.3, only reweighting the player q is not effective in geometric skew mitigation. Considering that more unique samples should be added to the minority group for diversity, we apply a novel mixing strategy across $\mathcal{G}$ to generate new samples. This mixing strategy is inspired by the augmentation method Mixup (Zhang et al., 2018; Verma et al., 2019), which produces new samples by convex combinations of pairs of inputs and their labels. Following this idea, each time, we allow the group construction by uniformly sampling two pairs $(x_i, y_i), (x_j, y_j)$ from $\mathcal{G}$, and the new sample is mixed as follows:

$$(\tilde{x}, \tilde{y}) \leftarrow (\lambda x_i + (1-\lambda)x_j, \lambda y_i + (1-\lambda)y_j) \tag{7}$$

where λ is the mixing-ratio sampled from a Beta(α, α) distribution. Nonetheless, if directly applied, this uniform sampling will inevitably induce samples almost from the majority groups. To ensure diversity is imposed on the minority group rather than the majority ones, we restrict that (x_j, y_j) must come from $\mathcal{G}^-$, that is, the estimated group attribute of (x_j, y_j) is $g_j = 0$. Therefore, we attain two kinds of group mixing: Mix($\mathcal{G}^+, \mathcal{G}^-$), Mix($\mathcal{G}^-, \mathcal{G}^-$). For Mix($\mathcal{G}^+, \mathcal{G}^-$), concerned with the spurious features still strongly correlated with the label after mixing, we modify the interpolation tactic of Equation 7. Concretely, when sampling λ, we always assign the larger λ to x_j from $\mathcal{G}^-$, the smaller λ to x_i, *i.e.*, $\lambda \leftarrow \min(\lambda, 1-\lambda)$.

3.3 Interactive Training for Robust Optimization

With the automatic group identification and mixing strategy, we can apply the algorithm of supervised group DRO to optimize the min-max play game in Equation 4. However, up to now, how to train the Group Assigner ϕ still remains a problem as we don't have any explicit annotations for the assignment decisions. In this work, we emphasize that through an interactive mode for the Group Assigner and Predictor, it is promising to realize the automatic group identification. Our intuition is that the majority group performance from the Predictor will drop if samples truly from the minority one

are misclassified, and guided by this loss, the updated ϕ will re-assign the group labels. For clarity, we present a more vivid illustration shown in Figure 3. Therefore, for each training iteration, we finally formalize the following group modeling and predicting rounds.

Modeling Round. Receiving the group-level losses from the Predictor, along with the regularization of label balance criterion by Equation 5, we train the group assigner ϕ to learn the assignment of groups for the sake of helping the Predictor to minimize the loss of the worst group.

Predicting Round. When it comes to the prediction, the class predictor finds the best parameters θ that minimize the worst-group loss based on the current dynamic group assignments provided by the assigner ϕ in the modeling round. Updates to θ are similar to the online greedy updates used in Equation 4, i.e. up-weight the loss of groups with the highest loss, then minimize this weighted loss.

4 Experiments

In this section, we conduct experiments on a synthetic sentiment classification task with complete spurious correlations and two real-world text classification tasks. Extensive empirical results demonstrate that $\mathcal{Q}$-Diversity outperforms existing DRO methods for robust optimization, even beating the state-of-the-art supervised method.

4.1 Experimental Setup

Baselines. We compare the performance of $\mathcal{Q}$-Diversity with respect to the following state–of-the-art baselines. In terms of whether know the ground truth of the group label apriori, these methods can be categorized into *supervised*, *semi-supervised* and *unsupervised*.

• **ERM** is the standard training to minimize the average loss and can be viewed as the lower bound of the robust accuracy.

• **Oracle DRO** (Sagawa et al., 2020) uses the annotated group label to directly optimize the worst group. Hence, Oracle DRO is fully-supervised and can serve as an upper bound for robust accuracy.

• **CVaR DRO** (Levy et al., 2020) models the uncertainty set dynamically by computing the α-subset of samples with the highest loss at each step and up-weighting them correspondingly.

• **LfF** (Nam et al., 2020) identifies the minorities in an unsupervised way, as it assumes samples that a weaker model classifies incorrectly largely correspond to those in the minority group and up-weights these minority-group-estimated samples.

• **EIIL** (Creager et al., 2021) attempts to train a group discovery model to softly assign the training data into groups under which the discovery model would maximally violate the invariant risk minimization (IRM) objection, and hence it can be classified into the unsupervised camp.

• **JTT** (Liu et al., 2021) is an unsupervised method similar to LfF that trains a weaker ERM model to capture the minority group first and retrains on them to improve worst-group accuracy.

• **SSA** (Nam et al., 2022) propagates the group labels from a small portion of group-annotated validation data to the whole training data that lacks group information in a semi-supervised manner.

Evaluation Metrics. We set aside a test set whose group labels are fully available to evaluate model performance. Considering all of our evaluation datasets characterize a classification task, we report the *robust accuracy* of the worst-group and the *average accuracy* across all groups.

4.2 $\mathcal{Q}$-Diversity Can Learn Robust Model

For the sake of investigating whether $\mathcal{Q}$-Diversity can help improve model robustness, we first carry out a toy classification task on BiasedSST.

Method	Average	Robust
Oracle DRO (Sagawa et al., 2020)	77.9	67.7
ERM	95.1	2.15
CVaR DRO (Levy et al., 2020)	92.5	28.1
JTT (Liu et al., 2021)	84.2	35.0
$\mathcal{Q}$-Diversity	**95.9**	**68.2**

Table 1: **Average and robust** test accuracies evaluated on BiasedSST.

BiasedSST (Michel et al., 2022) is a modified SST-2 sentiment classification dataset with a distractor token "so, " pretending to some sentences. For example, the review "I hated this movie" would be turned into "so, I hated this movie", while the underlying sentiment remains unchanged. Similar to the construction of Utama et al. (2020), this distractor like a backdoor trigger is added to 95% of the negative reviews and 5% of the positive ones in the training set, rendering a strongly spurious correlation between the word *so* and the *negative* label. Hereby, depending on the positive or negative label and the presence or absence of the distractor, we

Method	Group annotated		MultiNLI		CivilComments-WILDS	
	in train?	in val?	Average	Robust	Average	Robust
Oracle DRO (Sagawa et al., 2020)	✓	✓	81.4	76.6	87.7	69.1
ERM	✗	✓	82.4	67.9	92.6	57.4
CVaR DRO (Levy et al., 2020)	✗	✓	82.0	68.0	92.5	60.5
LfF (Nam et al., 2020)	✗	✓	80.8	70.2	92.5	58.8
EIIL (Creager et al., 2021)	✗	✓	79.4	70.9	90.5	67.0
JTT (Liu et al., 2021)	✗	✓	78.6	72.6	91.1	69.3
SSA (Nam et al., 2022)	✗	✓	79.9	76.6	88.2	69.9
ERM	✗	✗	81.9	60.4	92.7	51.6
CVaR DRO (Levy et al., 2020)	✗	✗	81.8	61.8	91.9	56.5
LfF (Nam et al., 2020)	✗	✗	81.1	62.2	92.0	55.9
EIIL (Creager et al., 2021)	✗	✗	80.3	64.7	91.2	63.8
JTT (Liu et al., 2021)	✗	✗	81.3	64.4	92.1	61.5
SSA (Nam et al., 2022)	✗	✗	80.4	76.5	89.1	69.5
$\mathcal{Q}$-Diversity	✗	✗	81.6	**77.7**	88.7	**73.5**

Table 2: **Average and robust** test accuracies evaluated on MultiNLI and CivilComments-WILDS.

obtain 4 groups and accuracy on the group of {positive, no distractor} can reflect model robustness.

We compare $\mathcal{Q}$-Diversity with four group DRO baselines and summarize the results in Table 1. It is clearly to see although ERM model achieves a high average accuracy, its performance on the group without suffering from the synthetic bias almost comes to zero. This reveals that models trained with ERM can very easily capture this spurious correlation, and fails on the minority group. The unsupervised methods CVaR DRO and JTT can help relieve such bias overfitting, however, their improvement in robust accuracy is very limited. When it comes to $\mathcal{Q}$-Diversity, its robust performance matches the Oracle DRO, while attains a better trade-off between accuracy and robustness.

4.3 $\mathcal{Q}$-Diversity in Practice

In order to cover a broad range of practical scenarios, we present two more challenging real-world datasets as the benchmarks for group robustness.

MultiNLI (Williams et al., 2018) is a multi-genre natural language inference dataset, given two sentences, a premise and a hypothesis, the goal of which is to predict whether the hypothesis is entailed by, contradicts, or neutral with the premise. We use this label as the target attribute (i.e., $\mathcal{Y}$ = {contradiction, entailment, neutral}), and use the existence of the negating words as the spurious attribute (i.e., $\mathcal{A}$ = {negation, no negation}).

CivilComments-WILDS (Koh et al., 2021) is derived from the Jiasaw dataset (Borkan et al., 2019), which aims to generate the toxicity indicator $\mathcal{Y}$ = {toxic, non-toxic} to a real online comment. We use demographic attributes of the mentioned identity $\mathcal{A}$ = {male, female, White, Black, LGBTQ, Muslim, Christian, other religion} as a spurious attribute for evaluation purpose. Considering that a comment can contain multiple such identities, so that followed by Liu et al. (2021), we use the coarse version $\mathcal{G} = \mathcal{Y} \times \mathcal{A}'$ for training, where $\mathcal{A}'$= {any identity, no identity}.

Dataset	Label	Group Counts	
		Negation	No Negation
MultiNLI	Contradiction	11158	57498
	Entailment	1521	67376
	Neutral	1992	66630
		Identity	Other
CivilComments-WILDS	Non toxic	90337	148186
	Toxic	17784	12731

Table 3: **Dataset description and group distribution** for MNLI and CivilComments-WILDS.

Under the two real-world settings, results are available in Table 2. Obviously, it can be seen that $\mathcal{Q}$-Diversity improves the robust accuracy on both classification tasks, beating all the baselines by a large margin. In fact, its robust accuracy even overtakes that of Oracle DRO, despite the fact that the former does not use any group information at training time. To achieve better robust performances, all

MultiNLI					SST2				
Dataset	ERM	EIIL	JTT	$\mathcal{Q}$-Diversity	Dataset	ERM	EIIL	JTT	$\mathcal{Q}$-Diversity
PI	73.72	81.53	81.25	84.38	SST2	91.85	66.39	80.82	90.62
LI	85.52	87.88	83.10	89.11	Senti140	65.41	53.99	67.19	68.75
ST	63.21	60.29	56.59	72.56	SemEval	83.90	72.14	66.59	87.09
HANS	62.11	65.06	65.32	65.82	Yelp	89.32	84.05	80.65	90.06
WaNLI	56.82	59.86	53.12	57.81	ImDB	83.66	64.50	70.43	85.34
SNLI	83.21	83.00	81.25	82.81	Contrast	84.63	56.76	64.34	82.31
ANLI (R3)	28.85	29.00	31.96	32.12	CAD	86.68	58.20	66.60	87.50
Avg% Δ	–	+1.88	-0.12	**+4.45**	Avg% Δ	–	-18.49	-12.69	**+0.89**

Table 4: **Accuracy on out-of-distribution** datasets (details can be found in Appendix A) for tasks with unknown spurious correlations. $\mathcal{Q}$-Diversity improves over ERM by $.5 - 10\%$, while baselines underperform.

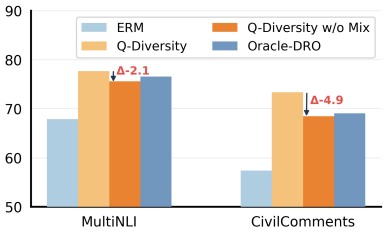

Figure 4: **Ablation Studies** on the role of mix.

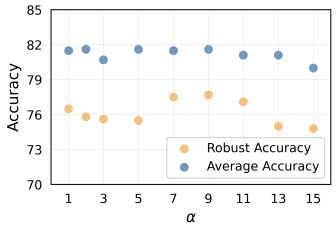

Figure 5: **Effect** of the mixing α on MultiNLI.

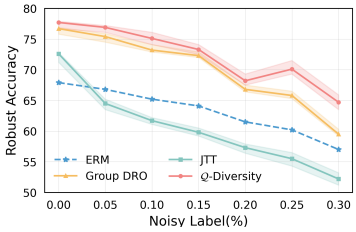

Figure 6: **Robust accuracy** under noisy labels.

the baselines need group annotations in the validation set for hyperparameters tuning. For example, JTT has to tune the number of epochs T to train the weaker model for group identification. When these annotations are unavailable in the validation set, their robust accuracy will drop significantly. In comparison, parameterizing the group identification in $\mathcal{Q}$-Diversity allows the annotation completely free, and the trainable procedure can render better robust accuracy.

5 Analysis and Discussion

In this section, we present a detailed analysis on the contribution of the diversified uncertainty set $\mathcal{Q}$ to its strong unsupervised performance. Furthermore, we explore the robustness of our method under different distributional shifts and random label noise.

5.1 Role of the Diversified $\mathcal{Q}$

We inspect the group diversity under the mixing strategy through an ablation study depicted in Figure 4. Apparently, we can observe significant drops in both datasets when removing this group mixing. These drops reveal that diversifying the minority groups can indeed help improve robust accuracy.

In addition, we analyze the influence of the mixing parameter α. As shown in Figure 5, we can observe that α indeed affects the effectiveness of the group mixing, leading to the volatility in robust accuracy. Considering the feature of Beta distribution, the sampled λ will be more concentrated around 0.5 as the α value becomes large, resulting in a relatively balanced weight between the mixed example pairs. The model performance remains stable when α is around $7 \sim 11$.

5.2 Generalization to OOD Sets

Since $\mathcal{Q}$-Diversity is a totally unsupervised method, it can be used off the shelf to improve OOD generalization on a new task. We therefore transfer $\mathcal{Q}$-Diversity, along with two other well-performing unsupervised baselines, *i.e.*, EIIL and JTT that first trained on MultiNLI and SST2 dataset, to a wide range of OOD datasets where the in-distribution spurious correlations may not hold.

$\mathcal{Q}$-Diversity improves robustness to unknown distributional shifts. With the unknown group information of these OOD test sets, we report the average accuracy in Table 4. Strikingly, we can observe that across the tasks and datasets, the two baselines even underperform than the lower bound of ERM. Especially on the SST2 dataset, the average accuracy of EIIL and JTT drop around 10% and 20%. We speculate this failure mode can be at-

tributed to their heuristic group identification manners, easily overfitting to the in-domain data. In contrast, $\mathcal{Q}$-Diversity outperforms ERM by 0.5%-5% across the datasets on average, revealing its great robustness to different distribution shifts.

5.3 Under the Presence of Label Noise

The unsupervised methods like JTT are based on the core idea of up-weighting samples with high losses. Nevertheless, when training data meets the noisy labels, such an approach will likely yield degenerate solutions, since the model tends to up-weight mislabeled samples with high losses. To further explore the application of unsupervised group DRO methods with the intervention of noisy labels, we perform experiments by inducing random label flips of varying degrees into MultiNLI dataset.

$\mathcal{Q}$-Diversity is more robust to random label noise. As the results shown in Figure 6, $\mathcal{Q}$-Diversity retains better robust accuracy under the presence of label noise than ERM and Group DRO. Corresponding to our assumption, JTT performs poorly even with a low noise rate since it fails to distinguish minorities from mislabeled samples.

6 Related Work

Group Robust Optimization Standard training with ERM can result in highly variable performance because of subpopulation distribution shifts arising from spurious correlations (Wu and Gui, 2022; Gao et al., 2022). In this context, Sagawa et al. (2020) formally introduces group DRO, with the goal to maximize worst-group or the minority group performance within the set of pre-defined groups. While promising, a rather practical scenario is that group information can not be available reliably. Therefore, another line of research begins to focus on the worst-case optimization without group annotations (Zhou et al., 2021). Typically, these methods first train a weaker model to identify high-loss samples as minority groups, and subsequently train an additional model with greater emphasis on the estimated minority groups (Nam et al., 2020; Liu et al., 2021).

Although the unsupervised group DRO methods are developed, they are confined to a two-stage training pipeline. In the two-stage model, a failed first stage can lead to an unsuccessful second stage as errors from the former are propagated to the later one. By contrast, $\mathcal{Q}$-Diversity in an end-to-end training manner overcomes the error accumulation. The group assigner and constructor cooperate with each other, and interactively, the classification response from the constructor can serve as a weak supervision to guide better group identification.

Diversity and OOD Generalization It is explored that the geometric skew and the statistical skew are two mechanisms hurting out-of-distribution performance with the existence of spurious correlations (Nagarajan et al., 2021; Nguyen et al., 2021). Concretely, the geometric skew is caused by the fact that classification margin on the minority group of a robust classifier tends to be much larger than that of the majority group, while the statistical skew arises from the fast convergence of gradient descent on spurious correlations unless trained for an exponentially long time. Although upweighting or oversampling the minority samples are straightforwardly effective in mitigating the statistical skew, both of them fail the geometric skew for the unchanged unique samples. Therefore, a wide range of studies emerge to diversify the input samples or feature space. Among them, counterfactually-augmented data (CAD), *i.e.*, data generated by minimally perturbing examples to flip the ground-truth label, has shown efficiency to learn robust features under distribution shifts (Kaushik et al., 2020). However, further investigation (Joshi and He, 2022) reveals the lack of perturbation diversity limits CAD's effectiveness on OOD generalization. In comparison, Wu et al. (2022) directly leverage the deep generative models to diversify training data with spurious correlations, while the model complexity is increased greatly.

For the sake of creating more synthesized samples to address geometric skew, our method that applying interpolation across the majority and minority groups shows its advantages in terms of perturbation diversity and time consumption.

7 Conclusion

In this paper, we present $\mathcal{Q}$-Diversity, an unsupervised method to optimize the worst group for model robustness. The formulation of $\mathcal{Q}$-Diversity extends the annotations of group DRO to an automatic assignment through an interactive training mode. Furthermore, under the guarantee of a novel mixing strategy across groups, $\mathcal{Q}$-Diversity can better counteract the failure modes of OOD generalization. Superior to previous works that only show the efficiency over the particular dataset, we demonstrate $\mathcal{Q}$-Diversity promises better general-

ization capability to various OOD sets. We believe that our work casts light on the limitations of group DRO which have been overlooked before, and can be viewed as a cornerstone for future study in the worst-group generalization.

Limitations

Although our unsupervised framework $\mathcal{Q}$-Diversity shows great superiority, when it comes to limitations, we acknowledge that (i) Our empirical validations on real-world datasets just follow current benchmarks that shed light on the group shifts caused by spurious correlations. Although we conduct experiments on the scenarios with noisy labels and various OOD datasets, practically, apart from superficial clues, a series of contributing factors that lead to group shifts are worth further exploration. (ii) A better theoretical understanding of how the interactive training mode can guide $\mathcal{Q}$-Diversity works in better group identification should be established, and this points out the direction for our future work.

Ethics Statement

Natural Language Processing (NLP) models that perform poorly on a minority group have raised a lot of concerns within the research community and broader society in recent years. In this work, the proposed $\mathcal{Q}$-Diversity is a versatile method that could be employed to train a robust model across groups even when the group information is not available. This is a rather practical scenario as the group information is almost missing during the data collection. We believe that our work is a step towards a suite of algorithms capable of solving a broader class of group DRO problems at scale. Moreover, such an algorithm will empower NLP researchers and engineers to create more reliable and ethical systems.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.61976056,62076069,62206057), Shanghai Rising-Star Program (23QA1400200), and Natural Science Foundation of Shanghai (23ZR1403500).

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. 2018. The mechanics of n-player differentiable games. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 354–363. PMLR.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. 2022. When does group invariant learning survive spurious correlations? In *Advances in Neural Information Processing Systems*.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. 2016. Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.*, 46:946–969.

SongYang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4112–4122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

1307–1323, Online. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037. PMLR.

Nitish Joshi and He He. 2022. An investigation of the (in)effectiveness of counterfactually augmented data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3668–3681, Dublin, Ireland. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. 2020. Large-scale methods for distributionally robust optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

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

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. 2022. Distributionally robust models with parametric likelihood ratios. In *ICLR 2022*.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. 2021. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. 2022. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*.

Thao Nguyen, Vaishnavh Nagarajan, Hanie Sedghi, and Behnam Neyshabur. 2021. Avoiding spurious correlations: Bridging theory and practice. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4885–4901, Online. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ting Wu and Tao Gui. 2022. Less is better: Recovering intended-feature subspace to robustify NLU models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1666–1676, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. 2021. Examining and combating spurious features under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12857–12867. PMLR.

MultiNLI	
Dataset	**Description**
PI (Liu et al., 2020)	selected instances from MultiNLI for testing the hypothesis-only bias in NLI models
LI (Liu et al., 2020)	selected instances from MultiNLI for testing logical inference ability of NLI models
ST (Naik et al., 2018)	stress set construction for testing the heuristics of NLI models
HANS (McCoy et al., 2019)	designed to contain examples where the shallow heuristics (e.g., lexical overlap) fail
WaNLI (Liu et al., 2022)	worker-and-AI collaborative dataset with challenging reasoning patterns for NLI task
SNLI (Bowman et al., 2015)	a large-scale, widely-used benchmark for NLI task
ANLI (R3) (Nie et al., 2020)	an iterative, adversarial human-and-model-in-the-loop solution for NLI dataset
SST2	
Dataset	**Description**
SST2 (Socher et al., 2013)	from the GLUE NLU benchmark to classify movie reviews as positive or negative
Senti140 (Go et al., 2009)	sentiment classification on Twitter messages
SemEval (Nakov et al., 2013)	crowdsourcing on Amazon Mechanical Turk over Twitter dataset for sentiment analysis
Yelp (Asghar, 2016)	online reviews consisting of free-form text and a star rating out of 5 for services
ImDB (Maas et al., 2011)	a collection of positive and negative reviews from Internet Movie Database
Contrast (Gardner et al., 2020)	small but label-changing modifications to the instances for ImDB
CAD (Kaushik et al., 2020)	counterfactual datasets constructed over ImDB

Table 5: Details of the out-of-distribution datasets in Table 4.

A Details of the OOD Datasets

We train the model on MultiNLI and SST2 tasks and test it on the corresponding OOD datasets respectively. For the results shown in Table 4, we present the details of these OOD datasets in Table 5 as follows.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitation

☑ A2. Did you discuss any potential risks of your work?
Section 5

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 4

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.2 / 4.3

☑ B1. Did you cite the creators of artifacts you used?
Section 4.3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 4.3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.3

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.