BSOR: BALANCED SEVERITY-FOCUSED ORDINAL REGRESSION FOR FINE-GRAINED MISINFORMATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The spread of misinformation on social media has become a growing societal concern, making its automatic detection an important yet challenging problem. In practice, misinformation deviates from the truth to varying degrees, and fact-checking organizations therefore grade claims on ordered scales such as the six-level PolitiFact rating. Existing multi-class detection methods, however, treat these ordered grades as independent flat labels, overlooking the ordinal structure among them and consequently suffering from severe prediction deviation. To address this limitation, we propose Balanced Severity-Focused Ordinal Regression (BSOR), a plug-and-play output head that formulates multi-class misinformation detection as ordinal regression. BSOR builds on a conditional decomposition of the ordinal task and further introduces cost optimization based on misjudgment distance awareness and hierarchy equilibrium based on gradient normalization to align the training objective with the severity structure of the task, together with a KL-divergence-constrained soft-label regularizer for stability. Across two public datasets, replacing the flat head with BSOR consistently reduces MAE for nine general-purpose encoders and six representative detection methods, with maximum reductions of 6.99% and 7.94%, respectively. Through error-distribution analysis, we find that the improvement mainly comes from suppressing costly long-range mispredictions. Representation-level analysis further shows that the gains stem from a more sufficient learning of the ordinal relations implicit in the data. These results demonstrate that BSOR offers plug-and-play generality and achieves stable performance gains in fine-grained misinformation detection.

1 INTRODUCTION

In recent years, the spread of the Internet and the expansion of social media platforms have profoundly changed how the public acquires and exchanges information. Social media have become the primary channel through which people learn about the outside world. This change has broadened the reach of information, but it has also created conditions for the fabrication and spread of misinformation. Malicious actors produce statements that deviate from the truth to varying degrees, misleading the public and steering opinion on platforms. Against this backdrop, misinformation detection has become an active research problem in natural language processing and data mining (Pattanaik et al., 2023).

Misinformation detection automatically evaluates content truthfulness using claim text and auxiliary information, with two major task paradigms. Binary misinformation detection outputs a binary label to identify misinformation and has mature methods and benchmarks, yet it reduces truthfulness into two states and fails to capture fine-grained differences in falsehood severity. We thus investigate multi-class misinformation detection, where models assign claims to discrete truthfulness ranks. The PolitiFact six-level rating system is widely adopted; its categories are not independent but form an ordinal structure according to the degree of factual deviation (Wang, 2017; Yang et al., 2024). However, existing multi-class misinformation detection methods generally adopt flat classification as the output objective. The six categories are treated as mutually independent labels, and the ordinal relationship among them is ignored (Khan & Hongsong, 2026; Xu & Kechadi, 2024). The resulting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

models exhibit severe prediction deviation, whereby a model's prediction of a claim deviates too far from the true category.

To address these problems, we propose Balanced Severity-Focused Ordinal Regression (BSOR). BSOR is the first to apply ordinal regression to multi-class misinformation detection. Built upon Conditional Ordinal Regression for Neural networks (CORN) (Shi et al., 2023), it decomposes the multi-class task into conditional binary subproblems and models ordinal structure with a conditional ordinal head. To meet fact-checking demands (Cao et al., 2020; Shi et al., 2023), BSOR remedies two structural deficiencies. The first is insensitivity to misclassification cost. Existing models use 0-1 cost, equally penalizing minor adjacent misjudgments and severe cross-level errors despite their distinct real-world impacts. To address this, the proposed **Cost Optimization Based on Misjudgment Distance Awareness** imposes dynamic penalties on prediction deviation at both probability distribution and discrete class levels, filling the inadequacy in misclassification cost modeling. The second deficiency is gradient imbalance. In conditional ordinal training, samples with larger labels participate in more subtasks, increasing their gradient contribution to the encoder and biasing training toward high-level classes. To address this, the proposed **Hierarchy Equilibrium Based on Gradient Normalization** assigns label-aware adaptive sample weights so that class-wise training signals are independent of label magnitudes. Furthermore, to stabilize training and reduce over-sharp predicted distributions, the proposed **KL Divergence Constrained Soft Label Regularization** serves as an auxiliary objective. It constructs a smooth target distribution centered on the true rank to learn better ordinal distributions and enhance prediction stability.

Our contributions are as follows:

- To the best of our knowledge, we are the first to formalize multi-class misinformation detection as an ordinal regression problem. This formulation identifies the mismatch between the objective that flat classification optimizes and the true objective of the task, and shows that ordinal modeling effectively reduces this mismatch.
- We propose BSOR, which addresses the insensitivity to misclassification distance and gradient imbalance through cost optimization based on misjudgment distance awareness and hierarchy equilibrium based on gradient normalization. Ablation studies verify the individual contribution of each component.
- We validate BSOR as a plug-and-play module. In experiments with nine general encoders on two datasets, BSOR reduces the mean absolute error (MAE) by up to 6.99%. Meanwhile, experiments on six misinformation detection methods achieve a maximum MAE reduction of 7.94%, demonstrating its strong generality.

2 Related Work

2.1 Misinformation Detection

Misinformation detection aims to determine the truthfulness of claim content. According to the form of judgment, it can be divided into binary and multi-class misinformation detection(Zhou & Zafarani, 2020). Binary misinformation detection(Khan & Hongsong, 2026; Shu et al., 2019; Xu et al., 2022) treats truthfulness as a binary classification problem of true versus false, and research in this field is relatively mature. However, such methods essentially compress truthfulness into two opposing states, which results in the loss of graded information embedded in claim credibility.

Multi-class misinformation detection partitions truthfulness into multiple categories, matching the fine-grained ratings released by fact-checking organizations (Xu & Kechadi, 2024; Rashkin et al., 2017; Alkomah & Sheldon, 2026; Hu et al., 2021; Zhou et al., 2019; Augenstein et al., 2019). Only a few studies have considered category order in multi-class settings. Vlachos et al. (Vlachos & Riedel, 2014) first defined fact-checking as ordinal classification, yet they argued supervised training on raw claim text was impractical and did not implement any ordinal model. Rashkin et al. (Rashkin et al., 2017) adopted a six-point graded scale and found most claims lie between the two extremes, but their LSTM, maximum-entropy and naive Bayes models were flat classifiers without ordinal modeling. Baly et al. (Baly et al., 2019) adopted multi-task ordinal regression to jointly predict news source trustworthiness and political ideology.

2.2 ORDINAL REGRESSION

Ordinal regression addresses prediction problems in which the categories have a natural order. Its statistical foundation is the proportional odds model of McCullagh (1980). Ordinal regression based on deep learning falls into two paradigms.

The first is the latent-variable thresholding paradigm (Chu & Ghahramani, 2005). It maps inputs to scalars, where ordinal thresholds partition the real line into category-corresponding intervals. CORAL (Cao et al., 2020) adopts this paradigm and imposes a shared projection direction over all categories to guarantee rank-consistent predictions. The second paradigm is chained conditional binary decomposition (Frank & Hall, 2001), which splits a K-class task into $K-1$ conditional binary subproblems. CORN (Shi et al., 2023), a representative approach under this paradigm, equips each discriminator with independent parameters and removes the shared projection constraint to improve modeling capacity. Both paradigms allow models to capture ordinal information across categories. However, their training objectives fail to distinguish categories or misclassification magnitudes. Training performance degrades substantially when class distributions are imbalanced or error consequences are asymmetric.

3 PROBLEM DEFINITION

This section specifies the task boundary, the label system, and the cost criteria of the study, providing the theoretical basis for the model optimization that follows.

PolitiFact six-level truthfulness rating system(Wang, 2017) divides the truthfulness of claims into ordinal categories. The category value correlates negatively with claim truthfulness. The larger the value, the further the claim deviates from the facts. The label set is:

$$Y \in \{\mathrm{r}_0, r_1, r_2, r_3, r_4, r_5\}$$

where r_0 denotes True, r_1 denotes Mostly True, r_2 denotes Half True, r_3 denotes Mostly False, r_4 denotes False, and r_5 denotes Pants on Fire. The detailed class definitions and descriptions are shown in Appendix A .

Multi-class misinformation detection for claim truthfulness assessment. Given an ordinal multi-class label set $Y = \{r_j\}_{j=0}^{K-1}$ with K categories and a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ containing N sample, the input of the i-th sample, denoted as $x_i = (c_i, t_i)$, comprises the core claim text c_i and optional supplementary text t_i, where $y_i \in Y$ is the corresponding multi-class label of sample i. The task objective is to learn a mapping function $C(D) \to Y$ from the dataset D to conduct multi-class misinformation detection for any unseen claim text.

The 0-1 cost follows an equal-misclassification-cost criterion. It does not differentiate the magnitude of deviation between prediction and ground truth; it only checks whether the predicted and ground-truth categories match. The formula is:

$$C_{01}(y, \hat{y}) = \mathbb{1}[y \neq \hat{y}],$$

where $\mathbb{1}[\cdot]$ is the indicator function, equal to 1 when the condition holds and 0 otherwise.

The ordinal increasing cost builds on the ordinal nature of the true labels. The further the predicted class deviates from the true label, the higher the misclassification cost. The formula is:

$$C_{\mathrm{ord}}(y, \hat{y}) = f(|idx\,(y) - idx\,(\hat{y})\,|)$$

where $f(\cdot)$ is a monotonically increasing function with $f(0) = 0$, and $idx(\cdot)$ is the index mapping function.

4 METHOD

4.1 OVERALL ARCHITECTURE OF BSOR

To meet the fine-grained requirement of claim truthfulness assessment, existing binary misinformation detection methods compress truthfulness into two opposing states. Such methods fail to

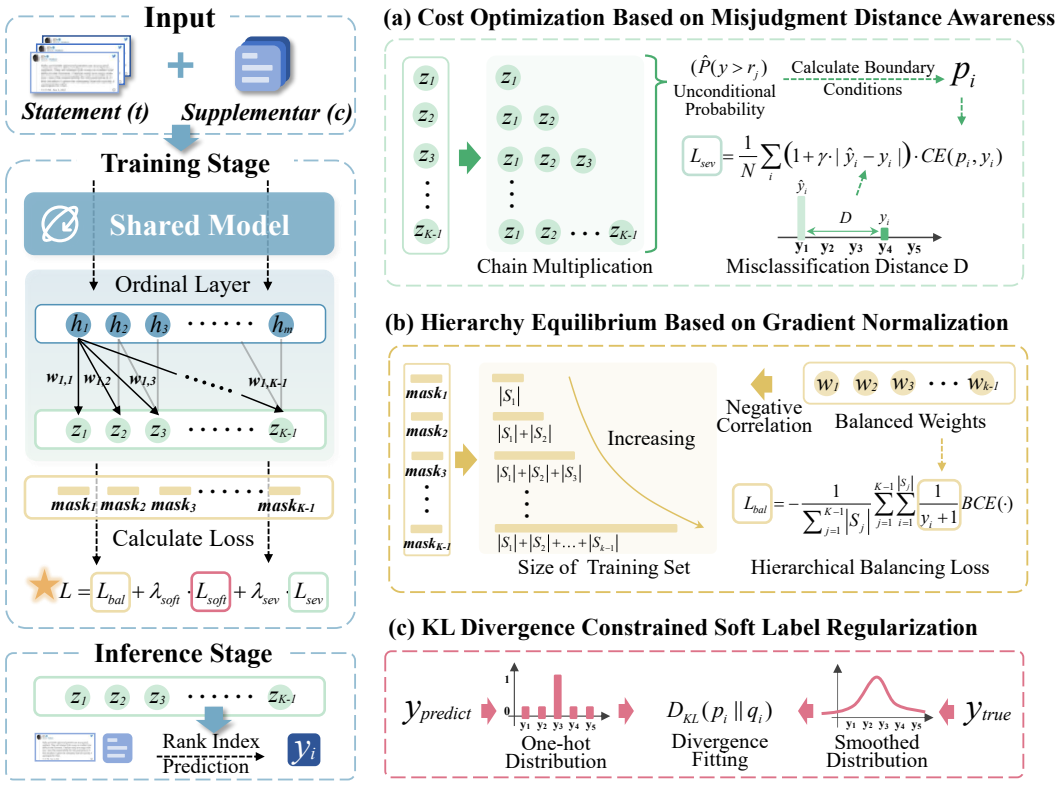


Figure 1: The overall architecture of BSOR.

characterize the differences in credibility among various claims and can hardly adapt to multi-level judgment in fact-checking scenarios. By contrast, conventional multi-class misinformation detection methods generally treat all categories as independent unordered labels, which leads to severe prediction deviation. To tackle the above issues, we propose **Balanced Severity-Focused Ordinal Regression (BSOR)**. BSOR consists of a shared encoder and a conditional ordinal regression head. Meanwhile, it improves the training process via the Cost Optimization Based on Misjudgment Distance Awareness (COMDA), the Hierarchy Equilibrium Based on Gradient Normalization (HEGN), and KL Divergence Constrained Soft Label Regularization. Its overall architecture is shown in the Figure 1

4.2 Conditional Ordinal Regression Head

As a plug-and-play output module for arbitrary backbone networks, conditional ordinal regression only replaces the network's output layer with $K-1$ independent fully-connected neurons for an ordinal task with K ranks, where each neuron learns its own weight and bias to compute the conditional probability of nested ordinal events:

$$f_j(x_i) = \sigma\left(\mathbf{w}_j^\top h_i + b_j\right) = P\left(y_i > r_j \mid y_i > r_{j-1}\right), \quad j = 1, \ldots, K-1, \tag{1}$$

During training, a boolean mask filters samples to construct conditional subsets $S_j = \{(x_i, y_i) \mid y_i > r_{j-1}\}$ for each binary subtask; only samples inside S_j contribute to the aggregated binary cross-entropy loss, and the model optimizes all independent neurons jointly. In inference, we use the probability chain rule $P(y_i > r_k) = \prod_{j=1}^{k} f_j(x_i)$ to derive cumulative probabilities, which inherently satisfy monotonicity and ensure rank consistency without extra post-processing; we then count the cumulative probabilities exceeding 0.5 threshold to obtain the predicted rank index:

$$q_i = \sum_{j=1}^{K-1} \mathbb{1}\left\{\hat{P}(y_i > r_j) > 0.5\right\}, \tag{2}$$

and the final predicted ordinal category is r_{q_i}.The detailed training and inference procedures of ordinal regression together with its underlying principles are elaborated in Appendix B.

4.3 Cost Optimization Based on Misjudgment Distance Awareness

Although conditional ordinal regression models ordinal labels via conditional binary subtasks, its loss adopts 0-1 cost and cannot differentiate the severity of rank-wise misjudgments. In multi-class misinformation detection, the prediction deviation between prediction and ground truth is a global property that cannot be assigned to individual binary subtasks, which only check threshold crossing and fail to measure misprediction distance. Consequently, the original loss cannot explicitly capture cost differences for mispredictions of varying severity and does not conform to the ordinal increasing cost assumption.To tackle this issue, we propose the Cost Optimization Based on Misjudgment Distance Awareness. This method introduces an auxiliary loss function acting on the complete category distribution. It enables the overall training objective to explicitly perceive the misprediction distance and thereby align with the ordinal increasing cost.

Concretely, based on the conditional probabilities $\{f_1(x_i), f_2(x_i), \ldots, f_{K-1}(x_i)\}$ output by the neural network, we first solve for the unconditional cumulative probability $P(y_i > r_k)$ via the probability chain rule. Then, the category probability of sample i at rank r_j is derived and recovered by computing the difference between adjacent cumulative probabilities:

$$\begin{aligned} p_i(r_j) &= P(y_i > r_{j-1}) - P(y_i > r_j), \\ \mathbf{p}_i &= [p_i(r_0), p_i(r_1), \ldots, p_i(r_{K-1})], \end{aligned} \tag{3}$$

where the boundary conventions $P(y_i > r_{-1}) = 1$ and $P(y_i > r_{K-1}) = 0$ apply, so that each entry of $\mathbf{p}_i$ is non-negative and sums to 1, forming a discrete probability distribution. Taking $\mathbf{p}_i$ as the supervision signal during the training stage, we construct the misprediction-distance-aware loss by introducing a weighting coefficient linearly correlated with the misprediction rank distance on top of the negative log-likelihood loss.

$$L_{\text{sev}} = \frac{1}{N} \sum_{i=1}^{N} (1 + \gamma \cdot |\hat{y}_i - y_i|) \cdot \text{NLL}(\mathbf{p}_i, y_i), \tag{4}$$

where N denotes the total number of samples in the training set; $\hat{y}_i = \arg\max_{r_j} p_i(r_j)$ is the predicted rank label obtained by taking the maximum probability from distribution $\mathbf{p}_i$; $NLL(\mathbf{p}_i, y_i)$ denotes the negative log-likelihood; and γ is the distance scaling hyperparameter.

4.4 Hierarchy Equilibrium Based on Gradient Normalization

The severity of gradient imbalance can be quantified by the training set class distribution. On the RDE dataset, True samples account for 15.9% and only join one conditional subset for loss computation, while False samples occupy 35.6% and participate in five subsets. Using the product of class proportion and the number of involved subsets as the loss contribution metric, the False class contributes roughly 11 times more loss than the True class. This skews training gradients toward higher ordinal ranks. Although ordinal labels only encode rank order rather than class importance, the conditional decomposition mechanism makes label magnitudes directly control sample gradient contributions, introducing inherent structural bias.To address this issue, we propose Hierarchy Equilibrium Based on Gradient Normalization, which applies sample-level weights to the binary cross-entropy terms of mask-filtered conditional subsets after BCE calculation and before loss aggregation, preserving the original mask filtering and conditional probability modeling.

Specifically, each sample is assigned a weight according to its own ordinal label. A sample with true rank y_i participates in $y_i + 1$ conditional subtasks in total, and within each participating conditional subset its loss term is multiplied by the weight $1/(y_i + 1)$. The weights of the same sample across all participating conditional subsets therefore sum exactly to one, which eliminates the gradient amplification caused by samples participating in different numbers of subtasks and keeps the total training signal of each class free of interference from the label order.

The loss function L_{corn} corrected by hierarchy equilibrium is formulated as:

$$L_{bal} = -\frac{1}{\sum_{j=1}^{K-1}|S_j|}\sum_{j=1}^{K-1}\sum_{i\in S_j}\frac{1}{y_i+1}BCE\left(f_j(x_i), 1\left\{y_i > r_j\right\}\right), \tag{5}$$

Hierarchy equilibrium only introduces sample weight coefficients for each binary cross-entropy loss term, fully preserving the original conditional probability decomposition paradigm and network parameterization form without introducing any additional trainable parameters. The weight coefficients are completely determined by the ground-truth labels of samples.

4.5 KL Divergence Constrained Soft Label Regularization

Beyond the two improvements, to address the model's insufficient fit of the ordinal distribution, we further introduce KL Divergence Constrained Soft Label Regularization, which optimizes the overall shape of the class distribution and improves the stability of the ordinal predictions.

Specifically, for the i-th sample we construct a smooth soft target distribution $\mathbf{q}_i$ peaked at its true ordinal level. The value of this distribution at category c is proportional to $e^{-\alpha|c-y_i|}$, and after normalization it becomes a target distribution centered at the true label that decays exponentially with the level distance, where α is the distribution decay coefficient. The soft-label loss uses the KL divergence to constrain the class distribution derived by the model to approach this smooth soft-label distribution. It is formulated as:

$$L_{\text{soft}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\left(\mathbf{q}_i \| \ \mathbf{p}_i\right), \tag{6}$$

where $\mathbf{q_i}$ denotes the smooth soft target distribution constructed from the true ordinal label and $\mathbf{p_i}$ denotes the class probability vector. The KL divergence constrains the overall shape of the distribution rather than a single peak, so the training signal is smoother than that of hard-label supervision. This mitigates the insufficient fit of the ordinal distribution and further stabilizes the ordinal predictions.

The hierarchy-equilibrium loss, the misclassification-cost loss, and the soft-label loss together form the multi-objective joint training objective of BSOR:

$$L = L_{\text{bal}} + \lambda_{\text{soft}} \cdot L_{\text{soft}} + \lambda_{\text{sev}} \cdot L_{\text{sev}} \tag{7}$$

where λ_{soft} and λ_{sev} denote the weight hyperparameters of the two auxiliary loss terms.

5 Experiments and Analysis

5.1 Datasets and Experimental Setup

We conduct supervised multi-class misinformation detection experiments on two public misinformation detection datasets, RDE (Yang et al., 2024) and LIAR (Wang, 2017). Both datasets adopt the six-point PolitiFact veracity rating as labels, mapping the degree of truthfulness into six ordered categories. Detailed information about the datasets and their splits is shown in Table 1, and the specific dataset partitioning method is presented in Appendix C.1.

Table 1: Dataset statistics

Dataset	Train	Validation	Test	Input
RDE	23,703	2,634	6,587	Claim and fact-checker fact description
LIAR	10,269	1,284	1,283	Claim text

Four evaluation metrics are used to meet the analytical requirements of different experiments: Mean Absolute Error (MAE), $Off-1Accuracy$, relative MAE change (ΔMAE), and $Off-1Accuracy$ change (ΔOff-1 Acc). The definitions and calculation formulas of these evaluation metrics are provided in Appendix C.2.

All experiments are repeated with three random seeds, reporting mean and standard deviation. Model selection and early stopping are based on validation-set MAE, and evaluation is performed on

the test set. BSOR hyperparameters are fixed across all groups: $\gamma = 1.0$, $\lambda_{\text{soft}} = 0.5$, $\lambda_{\text{sev}} = 0.15$, and $\alpha = 1.0$. Other experiment-specific configurations are introduced in the corresponding subsections.

5.2 General-Purpose Encoders Comparative Experiments

In this experiment, we keep the misinformation detection task unchanged without introducing domain-specific feature modules, and only replace the feature extraction encoder to compare flat classification and BSOR heads for validating BSOR's effectiveness and generalizability. We selected nine baselines, namely RoBERTa-large, RoBERTa-large-mnli, RoBERTa-base, BERT-base, DistilBERT, BiLSTM+Att, WordAvgMLP, TextCNN+GloVe and DeBERTa-large.

Table 2: Encoder generalization results

Dataset	Encoder	Baseline MAE	BSOR MAE	ΔMAE (%)	ΔO1A (pp)
RDE	RoBERTa-large	0.5135 ± 0.0061	0.4916 ± 0.0044	-4.26	$+1.23$
	RoBERTa-large-mnli	0.5181 ± 0.0210	0.4972 ± 0.0043	-4.03	$+1.16$
	RoBERTa-base	0.5910 ± 0.0149	0.5549 ± 0.0030	-6.11	$+1.73$
	BERT-base	0.7402 ± 0.0007	0.6978 ± 0.0044	-5.73	$+1.66$
	DistilBERT	0.7583 ± 0.0034	0.7286 ± 0.0090	-3.91	$+1.94$
	BiLSTM+Att	0.9589 ± 0.0088	0.9481 ± 0.0124	-1.12	$+0.43$
	WordAvgMLP	0.9602 ± 0.0032	0.9471 ± 0.0028	-1.36	$+0.24$
	TextCNN+GloVe	1.0305 ± 0.0177	0.9966 ± 0.0097	-3.29	$+0.77$
	DeBERTa-large	0.4880 ± 0.0067	0.4799 ± 0.0053	-1.67	$+0.65$
LIAR	RoBERTa-large	1.3388 ± 0.0186	1.2738 ± 0.0140	-4.85	$+2.73$
	RoBERTa-large-mnli	1.3635 ± 0.0211	1.2681 ± 0.0294	-6.99	$+3.74$
	RoBERTa-base	1.3856 ± 0.0472	1.2897 ± 0.0279	-6.92	$+2.10$
	BERT-base	1.3154 ± 0.0224	1.2603 ± 0.0077	-4.19	$+1.74$
	DistilBERT	1.3276 ± 0.0274	1.2933 ± 0.0159	-2.58	$+0.83$
	BiLSTM+Att	1.2944 ± 0.0004	1.3793 ± 0.0337	$+6.56$	$+3.22$
	WordAvgMLP	1.2946 ± 0.0000	1.3328 ± 0.0128	$+2.95$	$+4.23$
	TextCNN+GloVe	1.3931 ± 0.0147	1.3811 ± 0.0162	-0.86	$+0.62$
	DeBERTa-large	1.2897 ± 0.0110	1.2110 ± 0.0270	-6.10	$+2.73$

Experimental Results and Analysis. The results in Table 2 show that BSOR effectively alleviates the severe prediction deviation caused by flat classification ignoring the ordinal structure among categories. On RDE, the MAE of all nine encoders decreases, with relative reductions between 1.12% and 6.11% and a mean of 3.5%, while on LIAR all seven groups whose flat baselines remain discriminative improve as well, with relative reductions from 0.86% to 6.99% and a mean of 4.6%. Meanwhile, Off-1 Accuracy rises in every group, by 0.24 to 1.94 percentage points on RDE and by 0.62 to 3.74 percentage points on LIAR, which indicates that large misjudgment distances are effectively compressed and the predictions become closer to the ground-truth ranks on the whole. Notably, these gains cover parameter scales from 5M to 400M as well as diverse architectures and pre-training paradigms, and the only change shared across all groups is the replacement of the flat output head with the ordinal head.

In addition, the two encoders trained from scratch on LIAR exhibit a phenomenon worth noting. Their flat baselines collapse to constant prediction under all three random seeds, in which case the MAE is no longer comparable, so these two groups are excluded from the overall scoring. Nevertheless, once the ordinal head is adopted, the two encoders whose flat training has collapsed break away from the constant solution and learn class-discriminative predictions, with their Off-1 Accuracy improving by 3.22 and 4.23 percentage points, respectively. This phenomenon suggests that the training signals provided by the conditional ordinal decomposition are relatively richer. Such signals may help alleviate the collapse in from-scratch training.

5.3 Misinformation Detection Methods Comparative Experiments

To verify the capability of the BSOR ordinal head to extract domain features for misinformation, improve ordinal veracity prediction, and be directly integrated into existing detection models with strong generalizability, we conduct comparative experiments on misinformation detection

methods.Six baselines are selected in the experiments, namely DeBERTa-large+LSTM, FDHN, the speaker-context method, GEAR, dEFEND, and CompareNet.

Table 3: Representative-method generalization results

Dataset	Method	Baseline MAE	BSOR MAE	ΔMAE (%)	ΔO1A (pp)
RDE	DeBERTa+LSTM	0.5021 ± 0.0044	0.4991 ± 0.0057	-0.60	$+1.01$
	GEAR	1.1551 ± 0.0064	1.1336 ± 0.0083	-1.86	$+0.77$
	dEFEND	0.3583 ± 0.0192	0.3298 ± 0.0139	-7.94	$+0.26$
	CompareNet	0.3542 ± 0.0083	0.3330 ± 0.0104	-5.97	$+0.37$
LIAR	FDHN	1.0125 ± 0.0468	0.9376 ± 0.0096	-7.39	$+3.64$
	Speaker context	0.9727 ± 0.0306	0.9083 ± 0.0048	-6.62	$+2.62$

Results analysis. The results in Table 3 show that BSOR can be directly integrated into existing misinformation detection methods in a plug-and-play manner and effectively improves their truthfulness-level prediction. With only the output layer and the training objective replaced, while the original feature-extraction structures of the methods remain intact, the MAE of all six methods decreases after the ordinal head is transplanted, with relative reductions between 0.60% and 7.94%, among which dEFEND and FDHN benefit most and attain reductions of 7.94% and 7.39%, respectively. Meanwhile, Off-1 Accuracy improves in all six groups by 0.26 to 3.64 percentage points, which is consistent with the observation in the encoder experiments and indicates that the compression of large misjudgment distances likewise occurs in domain-specific detection models. Notably, these six methods cover three types of judgment cues, namely content and style, context, and external knowledge, and their feature-extraction and reasoning structures differ substantially, yet the ordinal head remains consistently effective across all of them, fully demonstrating the effectiveness of ordinal modeling. This experiment and the encoder experiments complement each other, jointly verifying the broad applicability of BSOR from the two parallel perspectives of general-purpose encoders and domain-specific detection methods.

5.4 Ablation Experiments

To analyze the individual contribution of each submodule of the model and identify the source of performance gains brought by different components, we conduct ablation experiments to verify the effectiveness of each module one by one. We design five model configurations for the ablation study: flat classification baseline, CORN, CORN + HEGN, CORN + COMDA, and the full BSOR model.

Table 4: Ablation results

Encoder	Flat baseline	CORN	CORN + HEGN	CORN + COMDA	BSOR
RoBERTa-base	0.5910 ± 0.0149	0.5683 ± 0.0120	0.5611 ± 0.0112	0.5553 ± 0.0037	0.5549 ± 0.0030
TextCNN+GloVe	1.0305 ± 0.0177	1.0074 ± 0.0069	1.0073 ± 0.0034	1.0012 ± 0.0036	0.9966 ± 0.0097
BERT-base	0.7402 ± 0.0007	0.7031 ± 0.0065	0.7006 ± 0.0061	0.7008 ± 0.0036	0.6978 ± 0.0044

Results analysis. The ablation results in Table 4 verify the effectiveness of each component of BSOR and the rationality of its design. On all three ablation experiments, the CORN configuration, which merely replaces the flat output head with the ordinal head, already clearly outperforms the flat baseline, indicating that ordinal modeling itself constitutes the main source of the performance gain. On this basis, the two components, namely hierarchy equilibrium and the misjudgment-distance-aware cost, each bring further improvements on every chain, and the full BSOR, which integrates both together with the soft-label regularization, attains the best result among the five configurations on all three chains, reducing the MAE by 0.0375 on average relative to the flat baseline. This progressive relationship shows that the two components are each effective against different deficiencies of conditional ordinal training and that their benefits are complementary, jointly underpinning the advantage of the complete method. Moreover, the direction of each component's contribution re-

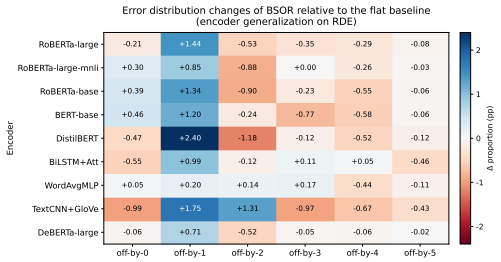
mains consistent across the three markedly different backbones, indicating that the components do not rely on a specific encoder and that the effectiveness of BSOR is robust.

5.5 Misjudgment Distance Distribution Analysis

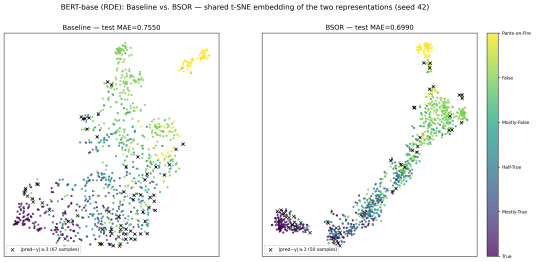
To explore the underlying mechanism behind the performance gains of BSOR, we design two complementary interpretability experiments: **misjudgment distance distribution analysis** and **gain source analysis**. We only briefly introduce them and detailed experimental settings and analyses are provided in Appendix E and F.

Result Analysis. In the misjudgment distance distribution analysis, as shown in the heatmap of the general-purpose encoders on RDE in Figure 2(a), the off-by-1 column is mostly positive whereas the off-by-4 and off-by-5 columns are mostly negative, indicating that large-deviation mispredictions are systematically compressed into smaller deviations. The complete statistics in Appendix E over the two datasets and the three experiment families show that this pattern holds consistently and that the MAE reduction is contributed almost entirely by the far bands, which coincides with the design goal of the misjudgment distance aware.

In the gain source analysis, the t-SNE visualization in Figure 2(b) shows that after BSOR training the classes are arranged into a compact graded band along the ordinal direction and the severe errors decrease accordingly. Appendix F further verifies this observation with linear probing and ordinal axis projection under the full set of settings, indicating that the ordinal information is already embedded in the baseline representations and that BSOR calibrates the decision boundaries along the existing ordinal order rather than creating a new one.



(a) Error distribution changes of BSOR relative to the flat baseline in the encoder generalization experiments on RDE.



(b) t-SNE visualization of the flat-baseline (left) and BSOR (right) representations of BERT-base on RDE.

Figure 2: Example figures of the two interpretability analyses on RDE. Full-size versions of (a) and (b) are provided in Figure 3 and Figure 8, respectively.

6 Conclusion

Misinformation detection requires differentiating the severity of prediction errors, while mainstream flat classification overlooks these differences. This work formalizes multi-class misinformation detection as an ordinal regression task. The proposed BSOR introduces two loss-level modifications to address gradient imbalance and misjudgment cost insensitivity within conditional ordinal regression. As a plug-and-play module compatible with any neural-network-based misinformation detection method, BSOR exhibits strong generalizability. Experiments over two public datasets on general encoders and misinformation detection baselines validate its effectiveness and generalizability. The maximum MAE reduction is 6.99% for general encoder experiments and 7.94% for baseline substitution experiments. Ablation studies confirm that each component of BSOR contributes to performance gains. Further analyses show that the improvement mainly comes from reduced long-distance mispredictions, which in turn stems from the more sufficient learning and utilization of the ordinal relations implicit in the data.

Future work can extend the proposed ordinal head to multimodal misinformation detection and validate whether the performance gains of ordinal modeling persist across non-English datasets and diverse authenticity rating systems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the following tasks: language translation and editing of the manuscript, retrieval and reading of related literature, and writing part of the experimental code. The conception of the research, the design of the method, the running of the experiments, and the interpretation of the results and the derivation of the conclusions were carried out by the authors. All AI-assisted text and code were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available datasets and involves no human subjects, no private data, and no interventions in real platforms. The method targets misinformation detection, which can support content moderation; we see no foreseeable negative societal impact beyond those common to all automated detection systems.

REPRODUCIBILITY STATEMENT

All experiments are run with three fixed random seeds and report means and standard deviations; model selection and early stopping uniformly follow the validation-set MAE, and the test set is evaluated only after training. The code will be released upon publication.

REFERENCES

Fatimah Alkomah and Frederick Sheldon. Credibility context improves political claim classification on LIAR and transfers to LIAR2 NEW. *Future Internet*, 18(7):346, 2026. doi: 10.3390/fi18070346.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697, 2019. doi: 10.18653/v1/D19-1475.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2109–2116, 2019. doi: 10.18653/v1/N19-1215.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, pp. 145–156. Springer, 2001. doi: 10.1007/3-540-44795-4_13.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 754–763, 2021. doi: 10.18653/v1/2021.acl-long.62.

Muhammad Shahid Khan and Chen Hongsong. Hybrid transformer deep neural architectures for enhanced misinformation detection on social media. *Expert Systems with Applications*, 300: 130470, 2026. doi: 10.1016/j.eswa.2025.130470.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–142, 1980.

Barsha Pattanaik, Sourav Mandal, and Rudra M. Tripathy. A survey on rumor detection and prevention in social media using deep learning. *Knowledge and Information Systems*, 65(10):3839–3880, 2023. doi: 10.1007/s10115-023-01902-w.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2931–2937, 2017. doi: 10.18653/v1/D17-1317.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3):941–955, 2023. doi: 10.1007/s10044-023-01181-9.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 395–405, 2019. doi: 10.1145/3292500.3330935.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 18–22, 2014. doi: 10.3115/v1/W14-2508.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, 2017. doi: 10.18653/v1/P17-2067.

Chenhui Xu and M.-T. Kechadi. An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006–88021, 2024. doi: 10.1109/ACCESS.2024.3418340.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pp. 2501–2510, 2022. doi: 10.1145/3485447.3512122.

Ziyi Yang, Jie Lin, Zhaohui Guo, Ying Li, Xiaodong Li, Qian Li, and Wenqiang Liu. Towards rumor detection with multi-granularity evidences: A dataset and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7188–7200, 2024. doi: 10.1109/TKDE.2024.3401700.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 892–901, 2019. doi: 10.18653/v1/P19-1085.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):109, 2020. doi: 10.1145/3395046.

A POLITIFACT SIX-LEVEL TRUTHFULNESS RATING SYSTEM

PolitiFact's six-level truthfulness rating system divides the truthfulness of claims into ordinal categories. The category value is negatively correlated with claim truthfulness. Detailed descriptions of the categories are provided in Table 5.

Table 5: PolitiFact six-level truthfulness rating scale

Level	Category	Definition
r_0	True	The statement is accurate and makes no significant omissions
r_1	Mostly True	The statement is accurate but needs clarification or additional information
r_2	Half True	The statement is partially accurate but leaves out important details or takes things out of context
r_3	Mostly False	The statement contains an element of truth but ignores critical facts that would give a different impression
r_4	False	The statement is not accurate
r_5	Pants on Fire	The statement is not accurate and makes a ridiculous claim

The official methodology is available at `https://politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/`.

B DETAILED PROCEDURE OF CONDITIONAL ORDINAL REGRESSION

The conditional ordinal regression head can be embedded as a plug-and-play modular component into any deep neural network architecture. This module only modifies the output terminal of the network, without altering the internal structure of the backbone network. It also preserves the original training pipeline and optimization logic of the model, thus exhibiting strong generality and transferability. The head is in essence a single fully connected linear output layer containing $K-1$ mutually independent conditional discriminant neurons, where K is the total number of categories in the ordinal task. Each neuron holds an independent weight and bias, applies a linear transformation to the high-level feature vector h_i output by the backbone, and maps the result into the probability interval through an $sigmoid$ activation function, modeling the conditional probability of the nested events:

$$f_j(x_i) = \sigma\left(\mathbf{w}_j^\top h_i + b_j\right) = P\left(y_i > r_j \mid y_i > r_{j-1}\right), \quad j = 1, \ldots, K-1, \tag{8}$$

where $\sigma(\cdot)$ denotes the activation function, w_j and b_j are the weight and the bias of the j-th neuron, h_i is the high-level feature output by the backbone. Because the discriminant neurons hold fully independent parameter sets, each subtask can learn its own discriminative projection direction. This removes the expressiveness limit imposed by weight sharing and improves the model's ability to fit complex ordinal decision boundaries.

During the training stage, each conditional discrimination subtask is associated with a conditional training subset, used for maximum likelihood estimation of the corresponding conditional probability. The conditional subset of the j-th subtask is

$$S_j = \{(x_i, y_i) \mid y_i > r_{j-1}\}, \;\; S_j \in D, \tag{9}$$

where S_j consists of all samples satisfying the preceding rank constraint; in particular, S_1 contains all training samples. In practice the dataset is not split into independent subsets. Instead, when the binary cross-entropy loss of each subtask is computed, a Boolean mask matrix filters out the samples that violate the preceding rank constraint, so that only samples in the conditional subset S_j contribute to the loss and to gradient backpropagation.

The loss aggregates the binary cross-entropy losses of all subtasks over all conditional subsets, normalized by the total number of samples across all conditional subsets. It is formulated as:

$$L_{corn} = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i \in S_j} BCE\left(f_j(x_i), \mathbb{1}\{y_i > r_j\}\right), \tag{10}$$

where $|S_j|$ denotes the size of the conditional subset S_j, $\mathbb{1}[\cdot]$ denotes the indicator function, taking the value 1 when the condition holds and 0 otherwise.

In the inference stage, relying on the probability chain rule, we use a set of conditional probabilities output by the network to solve for the unconditional cumulative probability that the true rank is greater than r_k:

$$P\left(y_i > r_k\right) = \prod_{j=1}^{k} f_j(x_i), \tag{11}$$

Since each conditional probability $f_i(x_i) \in (0, 1]$, the resulting cumulative probability sequence $\{P(y_i > r_1), \ldots, P(y_i > r_{K-1})\}$ naturally satisfies the increasing monotonic property, which guarantees rank consistency at the model structural level without introducing additional post-processing constraints. Finally, the rank index is obtained via thresholding and summation:

$$q_i = \sum_{j=1}^{K-1} \mathbb{1}\big\{\hat{P}(y_i > r_j) > 0.5\big\}, \tag{12}$$

The predicted ordinal category output by the final model is r_{q_i}.

C Supplementary Information of Experimental Settings

C.1 Dataset Partitioning Methods

We conduct supervised multi-class misinformation detection experiments on two public misinformation detection datasets, RDE (Yang et al., 2024) and LIAR (Wang, 2017). The RDE dataset aggregates fact-checking records of public claims from multiple fact-checking organizations, where each sample consists of the claim text to be verified and the supporting evidence descriptions provided by fact-checkers. This dataset only provides the original training set and held-out test set without a validation set. Therefore, stratified sampling is adopted within the original training set to split a new training subset and a validation subset at a ratio of 9:1. The stratified split ensures that the distribution of the six-level veracity labels remains consistent between the training and validation subsets, while the original test partition is retained without modification. The LIAR dataset focuses on political claims, and each sample only contains the claim text to be verified. The dataset provides predefined training, validation and test splits, which are directly adopted in this work.

Both datasets adopt the six-point PolitiFact veracity rating as labels, mapping the degree of truthfulness into six ordered categories. The ordinal scales of the label systems in the two datasets are fully aligned, with discrepancies only in the textual name of the fourth rank: it is denoted as mostly false in RDE and barely true in LIAR. Their ordinal positions are identical, which allows them to be directly used for unified ordinal modeling.

C.2 Calculation of Evaluation Metrics

Four metrics are adopted for model evaluation, namely the mean absolute error (MAE), Off-1 Accuracy, relative MAE change (ΔMAE), and Off-1 Accuracy change (ΔOff-1 Acc), to accommodate the analytical requirements of different experiments.

Mean Absolute Error (MAE) serves as the primary evaluation metric. Its general definition is the average of the absolute differences between predicted and ground-truth rank labels over all test samples. It is formulated as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{13}$$

where N denotes the total number of samples, and $\hat{y}_i$ and y_i represent the predicted rank and ground-truth ordinal label of the i-th sample, respectively. The misprediction distance corresponds to the ordinal distance between predictions and ground truths over the label set. The average misprediction distance across the test set is mathematically equivalent to MAE. Therefore, MAE can directly reflect the ordinal increasing cost focused in this work and is suitable as the primary evaluation metric for this ordinal task.

To facilitate the analysis of misprediction distance distribution, we adopt an equivalent transformed formulation of the above definition. Let $P(|\hat{y} - y| = d)$ denote the proportion of samples for which the predicted and ground-truth labels differ by d ranks, where $d \in \{0, 1, \ldots, K-1\}$. The equivalent decomposed form of MAE is written as:

$$\text{MAE} = \sum_{d=0}^{K-1} d \cdot P(|\hat{y} - y| = d) \tag{14}$$

This decomposed formula is fully equivalent to the original MAE definition. It groups samples by misprediction distance and performs weighted summation using the sample proportion at each distance. With this decomposition, the overall MAE can be disassembled into contribution terms corresponding to different rank deviation distances. It enables fine-grained localization of sources of performance gain and identifies whether MAE reduction arises from the mitigation of long-range cross-rank mispredictions.

Off-1 Accuracy counts the proportion of samples whose predicted ranks differ from ground-truth ranks by at most one rank. It measures the model's capacity to confine prediction deviations within adjacent ranks and complements the characterization of model predictive performance on neighboring ranks. Its calculation formula is:

$$\text{Off} - 1\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[|\hat{y}_i - y_i| \leq 1] \tag{15}$$

Relative MAE change (ΔMAE) is defined as the relative MAE variation of BSOR compared with baselines. A negative value indicates reduced MAE, meaning improved model performance. Its calculation formula is:

$$\Delta MAE = \frac{MAE_{BSOR} - MAE_{baseline}}{MAE_{baseline}} \tag{16}$$

Off-1 Accuracy Change (ΔOff-1 Acc) is defined as the change of Off-1 Accuracy of BSOR compared with the baseline, measured in percentage points. A positive value indicates an increase in Off-1 Accuracy, meaning improved model performance. Its calculation formula is:

$$\Delta\text{Off-1 Acc} = \text{Off-1 Acc}_{BSOR} - \text{Off-1 Acc}_{baseline} \tag{17}$$

D Experimental setup details

D.1 General-Purpose Encoders Comparative Experiments

Experimental Setup We conduct experiments on RDE and LIAR using nine distinct encoders. The nine encoders are RoBERTa-large, RoBERTa-large-mnli, RoBERTa-base, BERT-base, DistilBERT, BiLSTM+Att, WordAvgMLP, TextCNN+GloVe and DeBERTa-large. Each experiment includes a flat classification baseline and the BSOR model. BSOR simply replaces the original flat classification head with the proposed ordinal regression head.

General-Purpose Encoder Details Nine encoders are selected for the experiments, achieving diverse coverage across three dimensions: network architecture, model parameter size, and pre-training paradigm. The network architectures include bag-of-word MLP, convolutional networks, recurrent neural networks, as well as standard self-attention Transformer architectures such as BERT and RoBERTa, and DeBERTa with disentangled attention. The model parameter count spans from 5M to 400M, covering nearly two orders of magnitude. Pre-training strategies include no pre-training, static word embeddings, knowledge distillation, and natural language inference pre-training. The encoder configurations also support single-variable controlled comparisons. For instance, RoBERTa-base and RoBERTa-large share the same pre-training scheme and differ only in model capacity; RoBERTa-base and BERT-base have comparable architectural scales and differ only in pre-training strategies.

Table 6: Details of General-Purpose Encoders

Encoder	Params	Family	Pre-training Paradigm
RoBERTa-large	355M	Transformer	Masked language modeling
RoBERTa-large-mnli	355M	Transformer	Masked LM with NLI fine-tuning
RoBERTa-base	110M	Transformer	Masked language modeling
BERT-base	110M	Transformer	Masked language modeling
DistilBERT	67M	Transformer	Knowledge distillation
BiLSTM+Att	≈10M	Recurrent network	Trained from scratch
WordAvgMLP	≈10M	Bag-of-word MLP	Trained from scratch
TextCNN+GloVe	≈5M	Convolutional network	Static word embedding
DeBERTa-large	400M	Transformer	Masked LM with disentangled attention

For each encoder, controlled experiments are conducted independently on the RDE and LIAR datasets. Each group of experiments includes a flat classification baseline and the BSOR model. Both models adopt identical encoder backbones, optimizer hyperparameters, and data processing pipelines, differing only in the structure of the output layer and training loss objective. The baseline model attaches a flat multi-class output layer after the encoder, obtains the class probability distribution via Softmax normalization, and is optimized using standard cross-entropy loss. The BSOR model simply replaces the original flat classification head with the ordinal regression module proposed in this work.

D.2 Misinformation Detection Methods Comparative Experiments

Experimental Setup. Six baselines are selected in the experiments, namely DeBERTa-large+LSTM, FDHN, the speaker-context method, GEAR, dEFEND, and CompareNet, covering content, contextual, and external knowledge cues. Dataset selection is determined by the input requirements of each baseline. During transplantation, only the output layer and training objective are replaced.

Details of Misinformation Detection Methods Six distinct misinformation detection methods are adopted as encoders in our experiments, and their detailed information is presented below:

- **The DeBERTa-large+LSTM hybrid architecture** (Khan & Hongsong, 2026) encodes claim text with DeBERTa-large, followed by a two-layer bidirectional LSTM and scaled dot-product attention pooling. The original work targets binary social media misinformation classification. For transplantation, we retain the encoding and pooling pipeline, and replace the binary classification output layer with a six-level flat head and the BSOR ordinal head.
- **FDHN** (Xu & Kechadi, 2024) adopts a fuzzy deep network to fuse three channels: claim text, textual context and numerical context. The original model performs six-class flat multi-classification, and its primary variant does not incorporate pre-trained language models. During transplantation, the three-channel input construction is preserved, and only the flat cross-entropy output is replaced with the ordinal head.
- **The speaker-context method** (Alkomah & Sheldon, 2026) takes the speaker's previous fact-checking records as supplementary context. The input is concatenated from the claim text, six metadata fields and five counting features of the speaker's historical records, followed by flat classification after encoding via DeBERTa-v3-large. We keep the input construction unchanged during transplantation and replace the output head and training objective.

- **GEAR** (Zhou et al., 2019) performs aggregation and reasoning over an evidence graph in a three-stage claim verification paradigm. The RDE dataset directly provides fact-checker descriptions as evidence. In the pairing stage, BERT is fine-tuned with evidence placed before the claim. The encoder is then frozen to extract evidence representations, and the reasoning network aggregates information on the graph structure. For transplantation, the three-class output is replaced with the six-level ordinal head. Stratified validation splitting is conducted prior to the three-stage pipeline, and claim representations are used to fill samples with empty evidence.

- **dEFEND** (Shu et al., 2019) leverages co-attention between claims and comments to produce predictions and explanations simultaneously, originally designed for binary classification. For transplantation, claims are fed into the content channel, and fact descriptions are sent sentence by sentence into the comment channel, with 100-dimensional GloVe word embeddings adopted.

- **CompareNet** (Hu et al., 2021) propagates evidence and incorporates external knowledge on a heterogeneous graph consisting of document, entity and topic nodes for binary fake news detection. In our transplantation, the multi-class output is replaced with the six-level ordinal head. External knowledge links are substituted with surface forms of named entities; topic distributions are estimated via local implementation, and sentence segmentation is implemented using rule-based methods.

The original task settings and adaptation strategies for transplantation of the six methods are summarized in Table 7, and their training details are presented in Table 8.

Table 7: Overview of Transplantation and Adaptation for Representative Methods

Method	Original Task	Input Adaptation	Output Head Adaptation
Khan Hybrid Architecture	Binary social media misinformation detection	Claim text, original encoding and pooling pipeline preserved	Binary output layer replaced with six-level flat head or ordinal head
FDHN	Six-class flat multi-classification	Three-channel input unchanged, official LIAR split	Flat cross-entropy output replaced with ordinal head
Speaker Context	Six-class flat multi-classification	Concatenation of claims, metadata and speaker history unchanged, official LIAR split	Flat cross-entropy output replaced with ordinal head
GEAR	Three-class claim verification	Evidence–claim pairing, evidence directly available in RDE	Three-class output replaced with six-level ordinal head
dEFEND	Binary fake news detection	Claims fed into content channel, fact descriptions fed sentence by sentence into comment channel	Binary output replaced with six-level ordinal head
CompareNet	Binary fake news detection	Sentence-level heterogeneous graph; external knowledge replaced with surface forms of named entities	Binary output replaced with six-level ordinal head

D.3 Ablation Experiments

Experimental Setup and Results. We design five model configurations for the ablation study: flat classification baseline, CORN, CORN + HEGN, CORN + COMDA, and the full BSOR model. We select three encoders along two axes—RoBERTa-base and BERT-base as a pre-training contrast at the same 110M scale with different pre-training recipes, and TextCNN+GloVe as a non-pretrained convolutional feature extractor trained from scratch on static word embeddings—to examine the component effects from complementary angles.

Table 8: Training Settings of Representative Methods

Method	Optimizer	Learning Rate & Scheduler	Epoch & Early Stopping	Model Selection Criterion
Khan Hybrid Architecture	Adam	Encoder 2e-5, classification head 2e-4, cosine scheduler	Max 30 epochs, patience=5	MAE on validation set
FDHN	Adam	1e-3, no scheduler, no mixed precision	Fixed 10 epochs, no early stopping	MAE on validation set
Speaker Context	AdamW	1e-5, weight decay 0.01, cosine with 6% warmup, gradient clipping 1.0	Max 20 epochs, patience=3	MAE on validation set
GEAR	Adam	Reasoning network 5e-3, weight decay 5e-4	Max 10,000 epochs, patience=20	MAE on validation set
dEFEND	RMSprop	0.01, ρ 0.9, ε 1e-8, no mixed precision	Fixed 20 epochs, no early stopping	MAE on validation set
CompareNet	Adam	1e-3, weight decay 1e-6	Fixed 15 epochs, learning rate halved if no improvement within 3 epochs	MAE on validation set

E Detailed Results of Misjudgment Distance Distribution Analysis

Aggregate metrics such as MAE and Off-1 Accuracy only characterize the overall performance of a model; they hardly reflect how mispredictions are distributed along the deviation distance and therefore cannot explain where the gains of BSOR come from. The misjudgment-distance-aware cost optimization takes the suppression of long-range mispredictions as its direct design goal, which makes the distance structure of misjudgments the key to understanding the optimization behavior of BSOR. To reveal this structure, we conduct the misjudgment distance distribution analysis, which counts the off-by-k error proportions of the flat baseline and BSOR on the test sets and identifies the cross-rank misjudgments that are substantially alleviated.

Experimental Setup and Results. We analyze how BSOR suppresses severe cross-rank mispredictions by comparing the changes of off-by-k proportions on the RDE and LIAR datasets. The error distributions of nine encoder backbones under the flat baseline and BSOR are summarized in Table 9 and Table 10. The error distributions of six transplanted representative misinformation detection baselines are presented in Table 11. To verify the effectiveness of each proposed component, ablation experiments report the error distributions under different model configurations, as shown in Table 12. In the table, $\Delta = \text{BSOR} - \text{Flat}$.

To visually show the direction of changes in error proportion, heatmaps are plotted for error distributions. Here, off-by-d denotes the proportion of samples where the predicted rank differs from the ground-truth rank by d levels, with d ranging from 0 to 5. The color and value within each cell represent the percentage-point change in proportion of BSOR relative to the flat baseline. Red indicates negative values while blue indicates positive values; a positive value means an increased proportion at this deviation magnitude. Figures 3, 4, and 5 sequentially present the heatmaps of error distribution changes corresponding to the encoder generalization experiment on the RDE dataset, the encoder generalization experiment on the LIAR dataset, and the representative-method generalization experiment, respectively. Figure 6 presents the corresponding heatmaps of the ablation experiments, in which each configuration is compared with the flat baseline of the same encoder chain.

Results Analysis. Across all experiments, BSOR systematically compresses long-range mispredictions. On RDE, the off-by-5 proportion decreases in all nine encoder groups, and the off-by-4 proportion decreases in eight of the nine; the valid LIAR groups exhibit the same direction of change, with RoBERTa-large-mnli lowering its off-by-4 proportion by 2.10 percentage points; the off-by-4 proportion of all six representative methods decreases, and the off-by-5 proportion decreases in five methods and stays level in one; in the ablation experiments, every ordinal configuration on every

chain reduces the error mass of the far bands, and the full BSOR lowers both the off-by-4 and the off-by-5 proportions on all three chains.

Under the decomposition of MAE, the shrinkage of these far bands is exactly the source of the MAE reduction, which indicates that the gains of BSOR and its components originate from the suppression of costly long-range mispredictions, in agreement with the design intent of the misjudgment-distance-aware cost. The anomalous rows of the two degenerate groups stem from their collapsed flat baselines and are analyzed separately below in conjunction with Figure 7.

The two encoders trained from scratch on LIAR, namely BiLSTM+Att and WordAvgMLP, suffer flat-baseline collapse; their results are retained in Table 10 and Figure 4 for completeness and are separately displayed as per-seed prediction distributions in Figure 7. As shown in Figure 7, the flat baseline concentrates almost all predictions on a single class in every seed; WordAvgMLP predicts the same class for all 1,283 test samples, and BiLSTM+Att does so in two seeds and deviates for only 17 samples in the third, exhibiting a typical constant-prediction collapse. Such a constant solution no longer performs the detection task and cannot serve as a fair performance reference, so these two groups are excluded from the overall scoring. In contrast, once the ordinal head is introduced, both models break away from the constant solution, spread their predictions over the classes, and improve Off-1 Accuracy by 3.22 and 4.23 percentage points respectively, indicating that the training signals provided by the conditional ordinal decomposition enable the degenerate models to resume learning, in line with the observation in Section 5.2.

Table 9: Encoder error distribution on the RDE dataset.

Encoder	**Method**	**off-by-0**	**off-by-1**	**off-by-2**	**off-by-3**	**off-by-4**	**off-by-5**
RoBERTa-large	Flat	0.6358	0.2546	0.0811	0.0184	0.0085	0.0014
	BSOR	0.6337	0.2690	0.0758	0.0149	0.0056	0.0006
	Δ	−0.0021	0.0144	−0.0053	−0.0035	−0.0029	−0.0008
RoBERTa-large-mnli	Flat	0.6278	0.2635	0.0821	0.0167	0.0087	0.0009
	BSOR	0.6308	0.2721	0.0733	0.0167	0.0061	0.0006
	Δ	0.0030	0.0086	−0.0088	0.0000	−0.0026	−0.0003
RoBERTa-base	Flat	0.6048	0.2595	0.0940	0.0244	0.0153	0.0017
	BSOR	0.6088	0.2728	0.0850	0.0222	0.0099	0.0011
	Δ	0.0040	0.0133	−0.0090	−0.0022	−0.0054	−0.0006
BERT-base	Flat	0.5435	0.2683	0.1234	0.0351	0.0279	0.0015
	BSOR	0.5480	0.2802	0.1210	0.0273	0.0222	0.0009
	Δ	0.0045	0.0119	−0.0024	−0.0078	−0.0057	−0.0006
DistilBERT	Flat	0.5408	0.2605	0.1321	0.0354	0.0287	0.0024
	BSOR	0.5361	0.2845	0.1202	0.0342	0.0235	0.0012
	Δ	−0.0047	0.0240	−0.0119	−0.0012	−0.0052	−0.0012
BiLSTM+Att	Flat	0.4928	0.2409	0.1485	0.0568	0.0533	0.0073
	BSOR	0.4873	0.2508	0.1473	0.0578	0.0537	0.0027
	Δ	−0.0055	0.0099	−0.0012	0.0010	0.0004	−0.0046
WordAvgMLP	Flat	0.4693	0.2645	0.1615	0.0504	0.0495	0.0046
	BSOR	0.4697	0.2664	0.1629	0.0521	0.0451	0.0035
	Δ	0.0004	0.0019	0.0014	0.0017	−0.0044	−0.0011
TextCNN+GloVe	Flat	0.4570	0.2570	0.1552	0.0673	0.0563	0.0070
	BSOR	0.4471	0.2745	0.1682	0.0575	0.0496	0.0027
	Δ	−0.0099	0.0175	0.0130	−0.0098	−0.0067	−0.0043
DeBERTa-large	Flat	0.6420	0.2595	0.0758	0.0147	0.0065	0.0011
	BSOR	0.6414	0.2666	0.0706	0.0143	0.0059	0.0009
	Δ	−0.0006	0.0071	−0.0052	−0.0004	−0.0006	−0.0002

Table 10: Encoder error distribution on the LIAR dataset.

Encoder	Method	off-by-0	off-by-1	off-by-2	off-by-3	off-by-4	off-by-5
RoBERTa-large	Flat	0.2767	0.3484	0.2104	0.0943	0.0608	0.0078
	BSOR	0.2705	0.3819	0.2073	0.0881	0.0452	0.0055
	Δ	−0.0062	0.0335	−0.0031	−0.0062	−0.0156	−0.0023
RoBERTa-large-mnli	Flat	0.2619	0.3492	0.2307	0.0865	0.0631	0.0078
	BSOR	0.2705	0.3772	0.2136	0.0904	0.0421	0.0039
	Δ	0.0086	0.0280	−0.0171	0.0039	−0.0210	−0.0039
RoBERTa-base	Flat	0.2642	0.3507	0.2034	0.1068	0.0608	0.0117
	BSOR	0.2751	0.3609	0.2182	0.0951	0.0421	0.0070
	Δ	0.0109	0.0102	0.0148	−0.0117	−0.0187	−0.0047
BERT-base	Flat	0.2829	0.3500	0.1980	0.1091	0.0514	0.0062
	BSOR	0.2798	0.3710	0.2065	0.0959	0.0405	0.0047
	Δ	−0.0031	0.0210	0.0085	−0.0132	−0.0109	−0.0015
DistilBERT	Flat	0.2837	0.3453	0.1980	0.1107	0.0522	0.0078
	BSOR	0.2650	0.3718	0.2175	0.0974	0.0421	0.0047
	Δ	−0.0187	0.0265	0.0195	−0.0133	−0.0101	−0.0031
BiLSTM+Att	Flat	0.2089	0.3601	0.3578	0.0725	0.0000	0.0000
	BSOR	0.2486	0.3523	0.2284	0.1161	0.0475	0.0055
	Δ	0.0397	−0.0078	−0.1294	0.0436	0.0475	0.0055
WordAvgMLP	Flat	0.2081	0.3609	0.3593	0.0717	0.0000	0.0000
	BSOR	0.2346	0.3757	0.2510	0.0966	0.0398	0.0008
	Δ	0.0265	0.0148	−0.1083	0.0249	0.0398	0.0008
TextCNN+GloVe	Flat	0.2541	0.3523	0.2019	0.1286	0.0585	0.0023
	BSOR	0.2650	0.3484	0.2058	0.1099	0.0631	0.0078
	Δ	0.0109	−0.0039	0.0039	−0.0187	0.0046	0.0055
DeBERTa-large	Flat	0.2783	0.3655	0.2042	0.0951	0.0491	0.0062
	BSOR	0.2853	0.3858	0.2011	0.0904	0.0320	0.0047
	Δ	0.0070	0.0203	−0.0031	−0.0047	−0.0171	−0.0015

Table 11: Error distribution of representative methods.

Method	Setting	off-by-0	off-by-1	off-by-2	off-by-3	off-by-4	off-by-5
DeBERTa-large+LSTM	Flat	0.6407	0.2547	0.0762	0.0193	0.0076	0.0012
	BSOR	0.6255	0.2801	0.0720	0.0155	0.0058	0.0009
	Δ	−0.0152	0.0254	−0.0042	−0.0038	−0.0018	−0.0003
FDHN	Flat	0.4349	0.2783	0.1715	0.0740	0.0366	0.0039
	BSOR	0.4349	0.3141	0.1598	0.0624	0.0242	0.0031
	Δ	0.0000	0.0358	−0.0117	−0.0116	−0.0124	−0.0008
Speaker Context	Flat	0.4513	0.2837	0.1512	0.0717	0.0359	0.0047
	BSOR	0.4684	0.2931	0.1348	0.0717	0.0288	0.0023
	Δ	0.0171	0.0094	−0.0164	0.0000	−0.0071	−0.0024
GEAR	Flat	0.3845	0.2939	0.1823	0.0644	0.0704	0.0043
	BSOR	0.3905	0.2957	0.1776	0.0653	0.0673	0.0033
	Δ	0.0060	0.0018	−0.0047	0.0009	−0.0031	−0.0010
dEFEND	Flat	0.8371	0.0530	0.0616	0.0109	0.0361	0.0008
	BSOR	0.8426	0.0503	0.0704	0.0082	0.0278	0.0005
	Δ	0.0055	−0.0027	0.0088	−0.0027	−0.0083	−0.0003
CompareNet	Flat	0.8508	0.0413	0.0537	0.0115	0.0416	0.0008
	BSOR	0.8596	0.0363	0.0557	0.0088	0.0387	0.0008
	Δ	0.0088	−0.0050	0.0020	−0.0027	−0.0029	0.0000

Table 12: Ablation error distribution

Model	Configuration	off-by-0	off-by-1	off-by-2	off-by-3	off-by-4	off-by-5
RoBERTa-base	Flat baseline	0.6048	0.2595	0.0940	0.0244	0.0153	0.0017
	CORN	0.6006	0.2760	0.0899	0.0223	0.0097	0.0012
	CORN + HEGN	0.6097	0.2684	0.0859	0.0237	0.0108	0.0011
	CORN + COMDA	0.6089	0.2714	0.0870	0.0213	0.0103	0.0009
	BSOR	0.6088	0.2728	0.0850	0.0222	0.0099	0.0011
TextCNN+GloVe	Flat baseline	0.4570	0.2570	0.1552	0.0673	0.0563	0.0070
	CORN	0.4463	0.2701	0.1708	0.0580	0.0513	0.0032
	CORN + HEGN	0.4343	0.2866	0.1682	0.0618	0.0449	0.0036
	CORN + COMDA	0.4565	0.2648	0.1621	0.0578	0.0548	0.0038
	BSOR	0.4471	0.2745	0.1682	0.0575	0.0496	0.0027
BERT-base	Flat baseline	0.5435	0.2683	0.1234	0.0351	0.0279	0.0015
	CORN	0.5458	0.2798	0.1233	0.0285	0.0211	0.0012
	CORN + HEGN	0.5403	0.2865	0.1245	0.0302	0.0176	0.0006
	CORN + COMDA	0.5474	0.2777	0.1242	0.0285	0.0208	0.0011
	BSOR	0.5480	0.2802	0.1210	0.0273	0.0222	0.0009

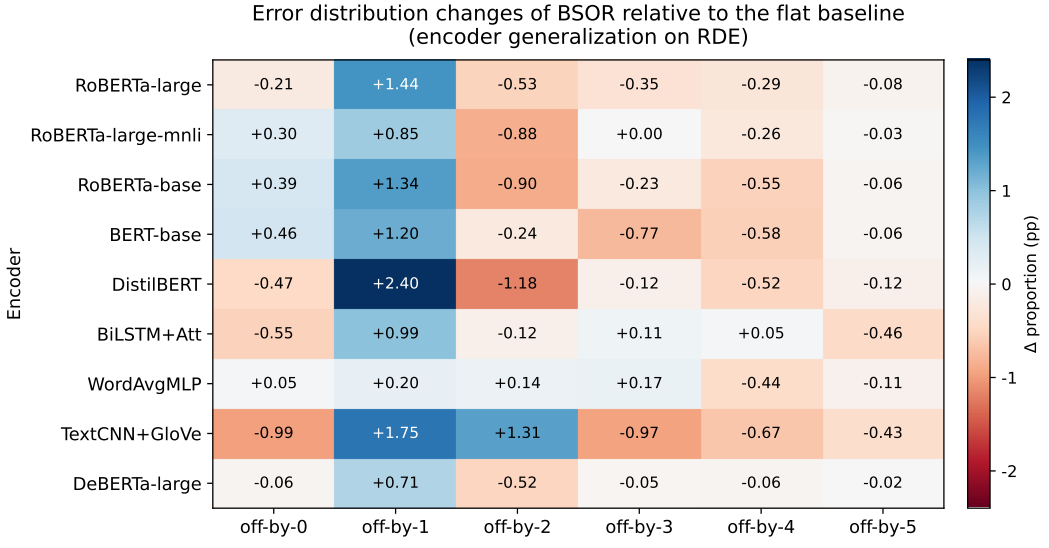


Figure 3: Error distribution changes of BSOR relative to the flat baseline in the encoder generalization experiments on RDE.

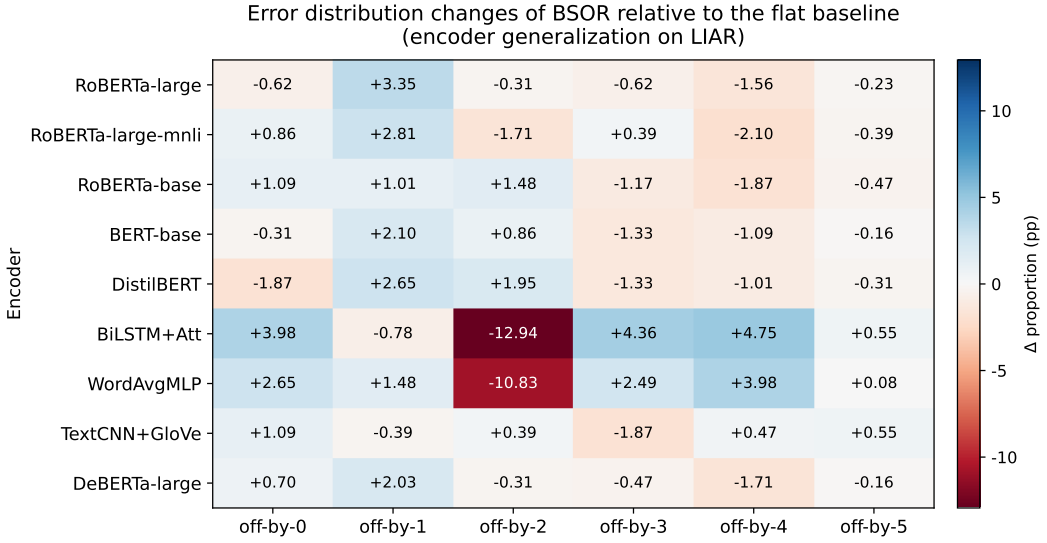


Figure 4: Error distribution changes of BSOR relative to the flat baseline in the encoder generalization experiments on LIAR.

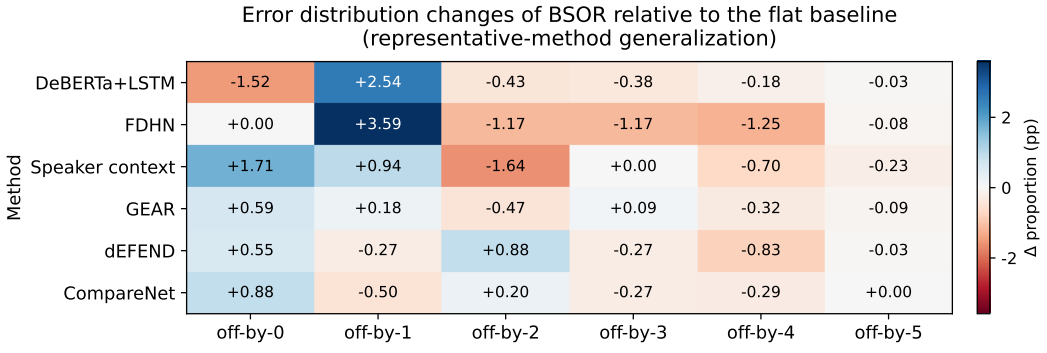


Figure 5: Error distribution changes of BSOR relative to the flat baseline in the representative-method generalization experiments.

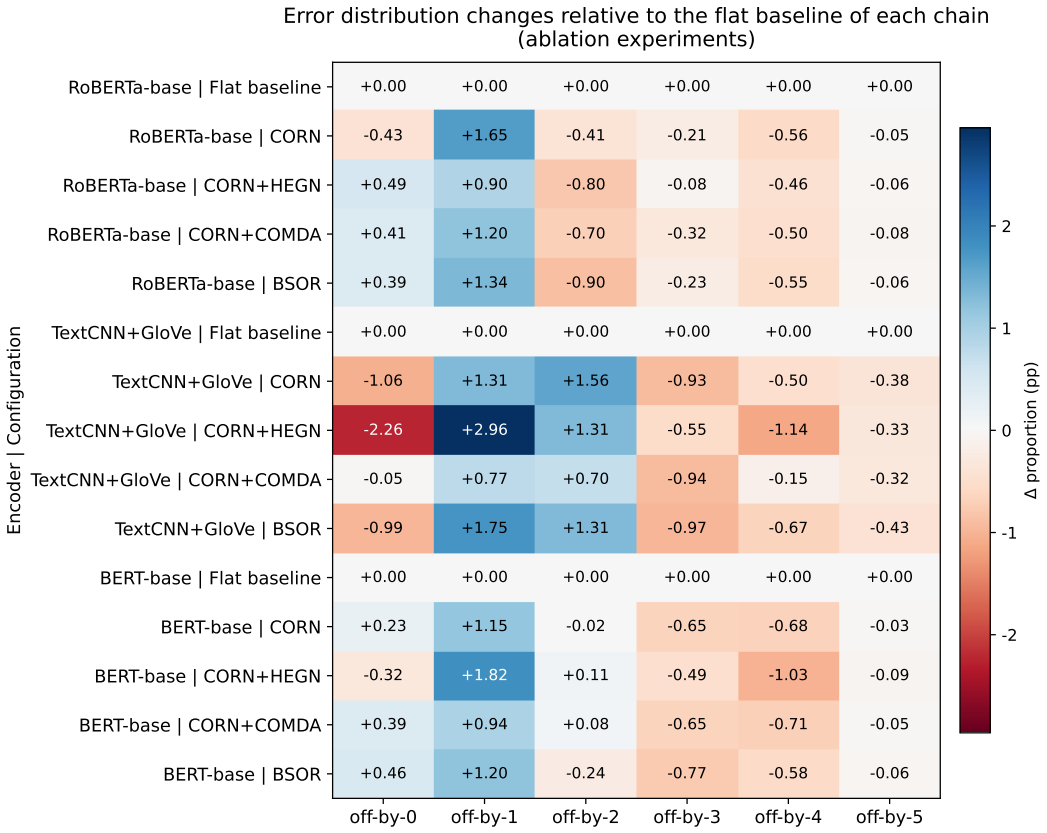


Figure 6: Error distribution changes of the ablation experiments. Each panel corresponds to one ablation chain; every configuration is compared with the flat baseline of the same chain, so the flat-baseline row is zero by definition.

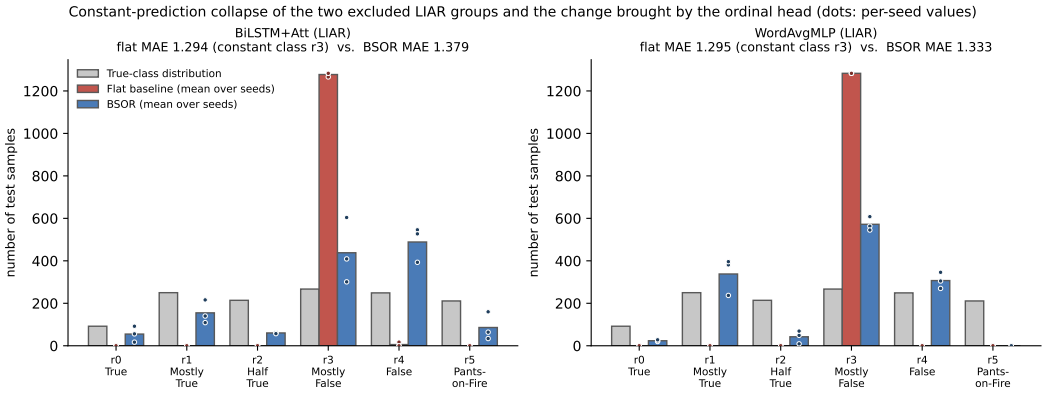


Figure 7: Per-seed prediction distributions of the two degenerate LIAR groups. Bars compare the true class distribution (gray), the test-set predictions of the flat baseline (red), and those of BSOR (blue); dots mark the per-seed values.

F Principles and Detailed Results of Gain Sources Analysis

The misjudgment distance distribution analysis characterizes the optimization direction of BSOR at the level of prediction behavior, whereas the ultimate source of the gains still needs to be explained at the representation level, that is, whether the improvement of BSOR comes from a better utilization of the information already present in the representations or from a re-learning of latent features. Answering this question helps delineate the level at which ordinalization operates. To this end, we reuse the trained models from the comparative and ablation experiments and extract the hidden representation H at the input of each model's classification head, since the prediction of a model is entirely determined by this representation together with the classification head. Without training the encoder or altering the representation, we adopt three tools, namely linear probing, ordinal axis projection, and t-SNE visualization, to examine the amount of information contained in the representation, the degree to which the representation is ordered along the label order $0\rightarrow5$, and the overall distribution of the representation.

F.1 Analysis Methodology

Linear Probing answers the questions of how much information is already contained in the representation and how much gain is brought by replacing the output layer. It trains linear classifiers on the embedded representation H output by the Baseline and BSOR models, respectively. The output layer parameters do not propagate gradients, so the probe can only extract information from the existing representation and cannot modify it. The flat probe is a multi-class logistic regression, isomorphic to the flat classification head; the ordinal probe consists of five conditional binary logistic regressions, isomorphic to the ordinal head of BSOR. For each pair of Baseline and BSOR experiments, the two types of probes are trained and evaluated on the H of the two models in exactly the same way. The difference in MAE between the flat probe and the ordinal probe on the same representation measures the ordinal information ignored by the flat objective in that representation; the change in MAE when the same probe is transferred from the Baseline representation to the BSOR representation measures the visible gain brought by ordinal training to the representation under frozen conditions.

Ordinal Axis Projection tests whether the representation is ordered along the label order $0\rightarrow5$. It fits an ordinal direction w between H and the numerical labels using Ridge regression, projects the centers of the six test classes onto w to obtain readings, and measures the degree of order by the Spearman rank correlation ρ between the readings and the class order $0\rightarrow5$. The projected readings of class centers further show the spacing between adjacent class centers along the ordinal direction, corroborating with the probe MAE whether BSOR makes the representation more ordered.

t-SNE Visualization directly observes the distribution of representations, especially the distribution of severely misjudged samples. It draws 1,000 samples from the RDE test set and the LIAR test set separately by stratified sampling according to class proportions, fits the H of the Baseline and BSOR models in the same coordinate space using t-SNE, and plots a t-SNE comparison figure. In the figure, colored points represent true labels, and black marks indicate severe error samples with $|\mathrm{pred} - y| \geq 3$.

F.2 Experiments

Experimental Setup and Results. We conduct probe fitting, ordinal axis projection, and t-SNE visualization for all general-purpose encoders, representative methods, and ablation configurations on the RDE and LIAR datasets under unified experimental settings. Tables 13 to 16 report the probe results of the three groups of methods, including the model MAE, the MAE of the flat and ordinal probes, and the ordinal axis correlation coefficient ρ, and Tables 17 to 19 list the projected readings of the class centers along the ordinal direction. Due to space limitations, only representative t-SNE visualizations are presented, among which Figures 8 and 9 represent the encoders on RDE and LIAR, Figure 10 represents the detection method FDHN, and Figure 11 represents the RoBERTa-base ablation chain.

Results Analysis. The linear probing results characterize the source of the gains from two dimensions. On the same baseline representation, the ordinal probe attains a lower MAE than the flat probe in all nine RDE groups and in six of the seven valid LIAR groups, with the only exception being a

tie on the weakest encoder TextCNN+GloVe, which indicates that the ordinal information is already embedded in the baseline representations and is merely ignored by the flat objective. When the same probe is transferred from the baseline representation to the BSOR representation, the MAE of both the flat and the ordinal probes decreases further, which indicates that ordinal training substantially improves the quality of the representations themselves and that the gains of the model can be read directly from the frozen representations.

The ordinal axis projection shows that BSOR learns and exploits the ordinal relations implicit in the data more effectively. On RDE, the ρ of all nine baseline representations already reaches 1.0 and remains unchanged after training, indicating that this dataset is sufficiently informative and that the models can learn a sound class order even under the flat objective, whereas the improvement of BSOR comes from calibrating the decision boundaries along the existing order and focusing the training signal on costly errors. On LIAR, the ordinal structure of the baseline representations is weaker, with ρ as low as 0.886; ordinal training raises it in six of the seven valid groups and brings it to 1.0 in six of them, and the inversions between the class-4 and class-5 centers in groups such as RoBERTa-large and DeBERTa-large are accordingly corrected, reflecting that BSOR can still fully exploit the ordinal relations on data with weaker ordinal signals and align the class order of the representation more closely with the label order.

The t-SNE visualizations corroborate the above conclusions intuitively. As shown in Figures 8 and 9, after BSOR training the middle classes are arranged into a more compact graded band along the ordinal direction and the severe errors scattered within it decrease accordingly; in Figure 10, the severe errors inside the extreme-class clusters of FDHN are removed in the BSOR panel; and on the ablation chain in Figure 11, the number of severe errors drops from 44 under the flat baseline to 32, 28, 32, and 32 under the ordinal configurations. The evidence from the three aspects corroborates one another and jointly indicates that the performance gains of BSOR stem from its more sufficient learning and utilization of the ordinal relations implicit in the data.

Table 13: Probe results and ordinal axis projection of nine encoders on RDE

Encoder	Rep.	Model MAE	Flat Probe MAE	Ordinal Probe MAE	Ordinal Axis ρ
RoBERTa-large	bl	0.5198	0.5309	0.5292	1.000
	ord	0.4913	0.5046	0.5054	1.000
	Δ	−0.0285	−0.0263	−0.0238	0.000
RoBERTa-large-mnli	bl	0.5140	0.5436	0.5338	1.000
	ord	0.5030	0.5118	0.5078	1.000
	Δ	−0.0110	−0.0318	−0.0260	0.000
RoBERTa-base	bl	0.5925	0.5998	0.5964	1.000
	ord	0.5590	0.5714	0.5632	1.000
	Δ	−0.0335	−0.0284	−0.0332	0.000
BERT-base	bl	0.7395	0.7497	0.7477	1.000
	ord	0.7015	0.7011	0.7050	1.000
	Δ	−0.0380	−0.0486	−0.0427	0.000
DistilBERT	bl	0.7629	0.7865	0.7726	1.000
	ord	0.7413	0.7459	0.7535	1.000
	Δ	−0.0216	−0.0406	−0.0191	0.000
BiLSTM+Att	bl	0.9464	0.9524	0.9518	1.000
	ord	0.9350	0.9362	0.9371	1.000
	Δ	−0.0114	−0.0162	−0.0147	0.000
WordAvgMLP	bl	0.9561	0.9635	0.9586	1.000
	ord	0.9476	0.9558	0.9536	1.000
	Δ	−0.0085	−0.0077	−0.0050	0.000
TextCNN+GloVe	bl	1.0525	1.1073	1.1002	1.000
	ord	0.9923	1.0329	1.0234	1.000
	Δ	−0.0602	−0.0744	−0.0768	0.000
DeBERTa-large	bl	0.4952	0.5238	0.5228	1.000
	ord	0.4843	0.4928	0.4867	1.000
	Δ	−0.0109	−0.0310	−0.0361	0.000

Table 14: Probe results and ordinal axis projection of seven valid encoders on LIAR (the two groups with degenerate flat baselines, BiLSTM+Att and WordAvgMLP, are excluded from the mechanism analysis; see Appendix F)

Encoder	Rep.	Model MAE	Flat Probe MAE	Ordinal Probe MAE	Ordinal Axis ρ
RoBERTa-large	bl	1.3219	1.4599	1.4263	0.943
	ord	1.2853	1.4014	1.3461	1.000
	Δ	−0.0366	−0.0585	−0.0802	0.057
RoBERTa-large-mnli	bl	1.3788	1.4599	1.4575	0.943
	ord	1.2276	1.2420	1.2303	1.000
	Δ	−0.1512	−0.2179	−0.2272	0.057
RoBERTa-base	bl	1.4170	1.5030	1.4578	0.886
	ord	1.2580	1.3463	1.3580	0.943
	Δ	−0.1590	−0.1567	−0.0998	0.057
BERT-base	bl	1.3468	1.4108	1.3905	0.943
	ord	1.2518	1.2759	1.2767	1.000
	Δ	−0.0950	−0.1349	−0.1138	0.057
DistilBERT	bl	1.3663	1.4591	1.4022	0.943
	ord	1.3133	1.3570	1.3398	1.000
	Δ	−0.0530	−0.1021	−0.0624	0.057
TextCNN+GloVe	bl	1.4100	1.5160	1.5160	1.000
	ord	1.4006	1.4840	1.4731	1.000
	Δ	−0.0094	−0.0320	−0.0429	0.000
DeBERTa-large	bl	1.2790	1.4404	1.4154	0.943
	ord	1.1777	1.2354	1.1832	1.000
	Δ	−0.1013	−0.2050	−0.2322	0.057

Table 15: Probe results and ordinal axis projection of misinformation detection methods

Method	Rep.	Model MAE	Flat Probe MAE	Ordinal Probe MAE	Ordinal Axis ρ
DeBERTa+LSTM	bl	0.5075	0.5295	0.5263	1.000
	ord	0.5013	0.4932	0.4894	1.000
	Δ	−0.0062	−0.0363	−0.0369	0.000
GEAR	bl	1.1497	1.1573	1.1429	1.000
	ord	1.1230	1.1549	1.1369	1.000
	Δ	−0.0267	−0.0024	−0.0060	0.000
dEFEND	bl	0.3539	0.3352	0.3398	1.000
	ord	0.3275	0.3407	0.3474	1.000
	Δ	−0.0264	0.0055	0.0076	0.000
CompareNet	bl	0.3542	0.3595	0.3483	1.000
	ord	0.3341	0.3968	0.4031	1.000
	Δ	−0.0201	0.0373	0.0548	0.000
FDHN	bl	0.9712	1.1091	1.1052	0.943
	ord	0.9283	1.0779	1.0694	0.943
	Δ	−0.0429	−0.0312	−0.0358	0.000
Speaker context	bl	0.9322	1.0522	1.0483	1.000
	ord	0.9150	1.0140	1.0125	1.000
	Δ	−0.0172	−0.0382	−0.0358	0.000

Table 16: Probe results and ordinal axis projection of ablation methods

Encoder	Config.	Model MAE	Flat Probe MAE	Ordinal Probe MAE	Ordinal Axis ρ
RoBERTa -base	Flat baseline	0.5925	0.5978	0.5964	1.000
	CORN	0.5667	0.5827	0.5784	1.000
	CORN + HEGN	0.5547	0.5634	0.5588	1.000
	CORN + COMDA	0.5605	0.5684	0.5651	1.000
	BSOR	0.5590	0.5714	0.5632	1.000
TextCNN +GloVe	Flat baseline	1.0525	1.1073	1.1002	1.000
	CORN	1.0149	1.0926	1.0249	1.000
	CORN + HEGN	1.0121	1.1277	1.0399	1.000
	CORN + COMDA	0.9974	1.0498	1.0451	1.000
	BSOR	0.9923	1.0329	1.0234	1.000
BERT -base	Flat baseline	0.7395	0.7497	0.7477	1.000
	CORN	0.6991	0.7176	0.7176	1.000
	CORN + HEGN	0.6967	0.6991	0.6935	1.000
	CORN + COMDA	0.6959	0.7067	0.7061	1.000
	BSOR	0.7015	0.7011	0.7050	1.000

Table 17: Ordinal-axis class-center projections on the RDE and LIAR datasets (the LIAR-side entries of BiLSTM+Att and WordAvgMLP are excluded: their flat baselines collapse on LIAR, see Appendix F; their RDE-side entries are valid and retained)

		RDE		LIAR	
Encoder	**Class**	**Baseline**	**BSOR**	**Baseline**	**BSOR**
RoBERTa-large	0	−2.153	−2.148	−0.881	−0.880
	1	−1.320	−1.317	−0.258	−0.183
	2	−0.570	−0.577	−0.167	−0.159
	3	0.120	0.051	0.124	0.117
	4	1.084	1.061	0.434	0.397
	5	1.828	1.810	0.424	0.492
RoBERTa-large-mnli	0	−2.127	−2.111	−0.401	−0.897
	1	−1.284	−1.286	−0.210	−0.298
	2	−0.545	−0.566	−0.092	−0.046
	3	0.093	0.077	0.052	0.118
	4	1.071	1.034	0.366	0.525
	5	1.791	1.767	0.235	0.536
RoBERTa-base	0	−1.994	−2.029	−0.871	−0.751
	1	−1.192	−1.208	−0.239	−0.178
	2	−0.587	−0.592	−0.241	−0.138
	3	0.066	0.026	0.161	0.139
	4	0.999	0.992	0.492	0.613
	5	1.743	1.743	0.367	0.545
BERT-base	0	−1.769	−1.806	−0.703	−0.827
	1	−0.975	−1.034	−0.123	−0.266
	2	−0.477	−0.519	−0.150	−0.222
	3	0.044	0.059	0.060	0.074
	4	0.834	0.843	0.400	0.423
	5	1.649	1.657	0.503	0.468
DistilBERT	0	−1.692	−1.764	−0.649	−0.658
	1	−0.937	−1.051	−0.148	−0.208
	2	−0.480	−0.561	−0.154	−0.167
	3	0.036	−0.025	0.072	0.127
	4	0.819	0.792	0.387	0.419
	5	1.633	1.619	0.471	0.474
BiLSTM+Att	0	−1.206	−1.281	–	–
	1	−0.776	−0.768	–	–
	2	−0.565	−0.527	–	–
	3	−0.161	−0.094	–	–
	4	0.686	0.729	–	–
	5	1.400	1.492	–	–
WordAvgMLP	0	−1.217	−1.234	–	–
	1	−0.813	−0.810	–	–
	2	−0.487	−0.497	–	–
	3	−0.072	−0.090	–	–
	4	0.714	0.734	–	–
	5	1.438	1.424	–	–
TextCNN+GloVe	0	−0.997	−1.036	−0.322	−0.326
	1	−0.661	−0.707	−0.176	−0.281
	2	−0.481	−0.525	−0.115	−0.136
	3	−0.168	−0.185	−0.032	0.097
	4	0.559	0.583	0.184	0.299
	5	1.128	1.259	0.187	0.321
DeBERTa-large	0	−2.178	−2.179	−0.691	−0.885
	1	−1.323	−1.293	−0.219	−0.336
	2	−0.563	−0.608	−0.195	−0.207
	3	0.113	0.073	0.110	0.064
	4	1.087	1.040	0.424	0.506
	5	1.823	1.780	0.423	0.523

Table 18: Ordinal-axis class-center projections of representative methods

Dataset	Method	Class	Baseline	BSOR
RDE	DeBERTa+LSTM	0	−2.056	−2.153
		1	−1.224	−1.331
		2	−0.567	−0.623
		3	0.095	0.083
		4	1.042	1.061
		5	1.685	1.785
	GEAR	0	−0.758	−0.766
		1	−0.699	−0.698
		2	−0.213	−0.212
		3	0.018	0.022
		4	0.588	0.611
		5	1.171	1.193
	dEFEND	0	−1.961	−1.873
		1	−1.552	−1.555
		2	−0.624	−0.654
		3	0.232	0.229
		4	0.830	0.810
		5	2.089	2.126
	CompareNet	0	−1.683	−1.716
		1	−1.562	−1.138
		2	−0.755	−0.905
		3	0.198	0.280
		4	0.872	1.030
		5	2.054	2.158
LIAR	FDHN	0	−1.203	−1.398
		1	−0.630	−0.603
		2	−0.186	−0.241
		3	0.178	0.224
		4	0.617	0.697
		5	0.535	0.590
	Speaker context	0	−1.567	−1.623
		1	−0.624	−0.637
		2	−0.256	−0.263
		3	0.233	0.240
		4	0.692	0.740
		5	0.751	0.805

Table 19: Ordinal-axis class-center projections of the ablation encoders (RDE)

Encoder	Class	Baseline	BSOR
RoBERTa-base	0	−1.994	−2.029
	1	−1.192	−1.208
	2	−0.587	−0.592
	3	0.066	0.026
	4	0.999	0.992
	5	1.743	1.743
TextCNN+GloVe	0	−0.997	−1.036
	1	−0.661	−0.707
	2	−0.481	−0.525
	3	−0.168	−0.185
	4	0.559	0.583
	5	1.128	1.259
BERT-base	0	−1.769	−1.806
	1	−0.975	−1.034
	2	−0.477	−0.519
	3	0.044	0.059
	4	0.834	0.843
	5	1.649	1.657

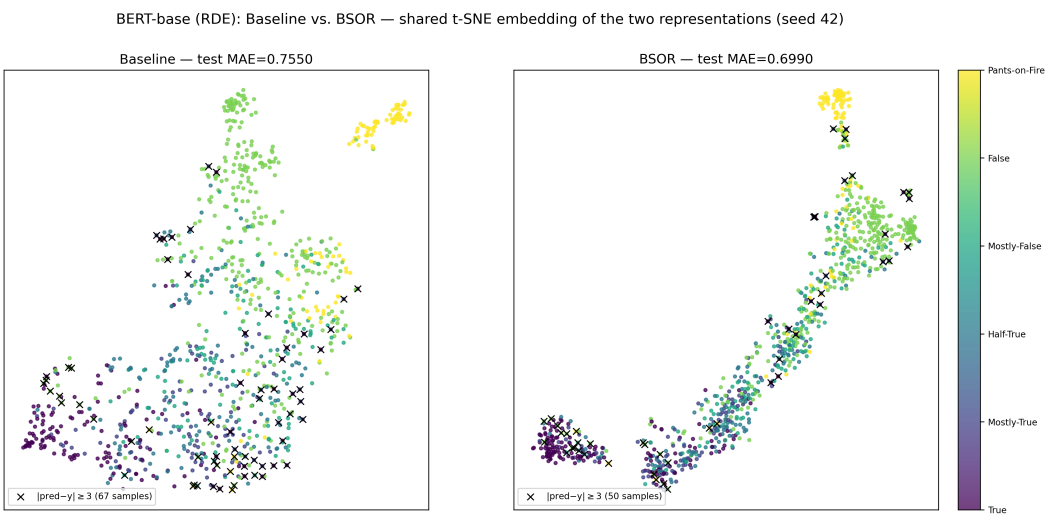


Figure 8: t-SNE visualization of the flat-baseline (left) and BSOR (right) representations of BERT-base on RDE (1,000 stratified test samples, seed 42), embedded in one shared space; point colors denote the true classes, and black $\times$ marks severe errors with $|\mathrm{pred} - y| \geq 3$. The baseline spreads into a Y-shaped topology with the two extremes close to each other, whereas the BSOR representation arranges the classes along an ordinal band with the severe errors concentrated near its two ends.

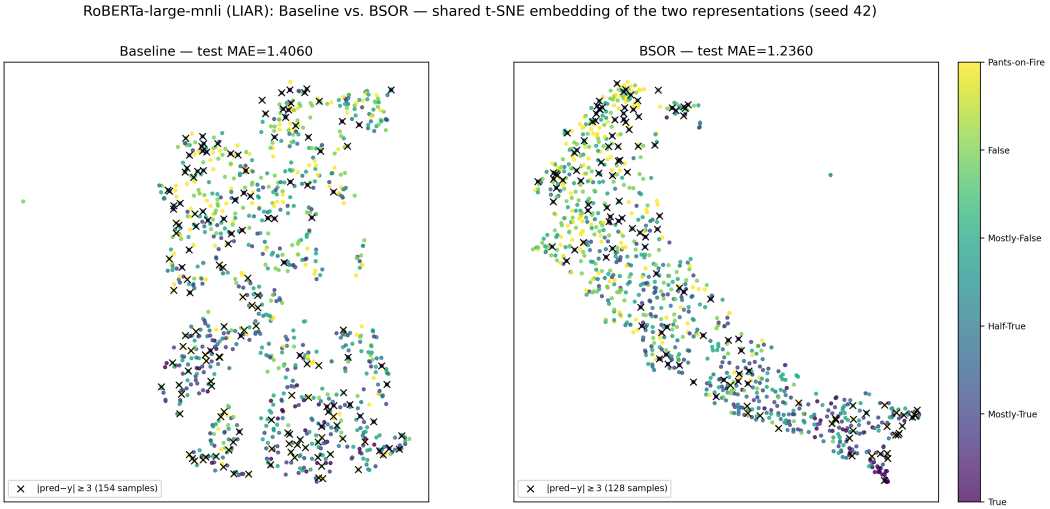


Figure 9: t-SNE visualization of the flat-baseline and BSOR representations of RoBERTa-large-mnli on LIAR, under the same protocol as Figure 8; the BSOR panel arranges the classes into a graded band from True to Pants-on-Fire and removes part of the severe errors scattered inside the dense middle region.

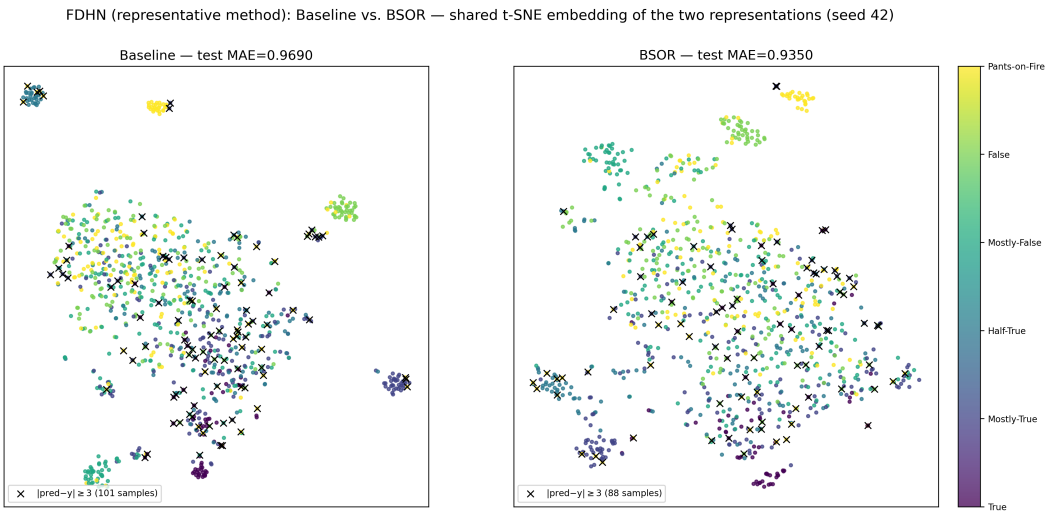


Figure 10: t-SNE visualization of the flat-baseline and BSOR representations of the representative detection method FDHN, under the same protocol as Figure 8; severe errors inside the extreme-class clusters are removed in the BSOR panel.

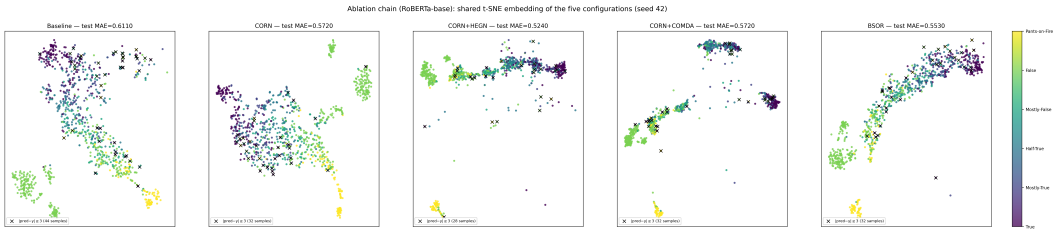


Figure 11: t-SNE visualization along the RoBERTa-base ablation chain (1,000 stratified test samples, seed 42). The five configurations share one embedding, so the panels are directly comparable; point colors denote the true classes, and black $\times$ marks severe errors with $|\mathrm{pred} - y| \geq 3$. The severe-error count drops from the flat baseline (44) to every ordinal configuration (32, 28, 32, and 32), consistent with the reduced long-range mispredictions of the ordinal configurations.