Multi-knowledge Enhanced Graph Neural Network for Multi-trait Essay Scoring

Shiman Zhao, Siyuan Liu*, Zhiqi Shen

College of Computing and Data Science, Nanyang Technological University, Singapore
{shiman.zhao, syliu, zqshen}@ntu.edu.sg

Abstract

Multi-trait Essay Scoring (MES) aims to evaluate the quality of essays across multiple traits (e.g., Language, Content, and Organization). The task can be summarized into three crucial steps: essay content encoding, trait feature learning, and multi-trait scoring. However, previous methods fall short in these steps due to neglecting essential scoring-oriented knowledge, leading to suboptimal performance. To solve these issues, we propose a novel multi-trait scoring framework with multi-knowledge enhancement. Specifically, linguistic knowledge is used to model syntactic structural relations between words, highlighting structurally-informed essay encoding. We learn trait knowledge by capturing the knowledge dependencies between traits to enhance trait-specific features. Further, score-aware ordinal knowledge is integrated to promote ordinal alignment in trait-specific features associated with score rankings, improving scoring performance. Extensive experiments show that our proposed method achieves significant performance.

Introduction

In education systems, traditional essay evaluation (Fiacco, Adamson, and Rose 2023) struggles to provide instant feedback on a large number of essays due to the teacher-student ratios in classrooms. Recently, Automated Essay Scoring (AES) (Jiang et al. 2023; Boquio and Naval 2024) has been developed to enhance the efficiency and effectiveness of essay evaluation. AES not only provides timely and personalised feedback (Wang and Jin 2025) on writing exercises but also reduces the inconsistency and subjective bias of human evaluators. However, AES methods mostly focus on holistic essay scoring (Yang et al. 2020; Xie et al. 2022), providing an overall score for an essay. The single prediction has limited application in real-world scenarios.

Multi-trait Essay Scoring (MES) (Li and Ng 2024a) is non-trivial for education systems, which aims to cultivate students' writing skills across different traits, such as Language, Content, and Organization. Recently, MES has attracted widespread attention in academia and industry. The MES task consists of three main components: essay content encoding, trait feature learning, and multi-trait scoring. However, there remain some challenges.

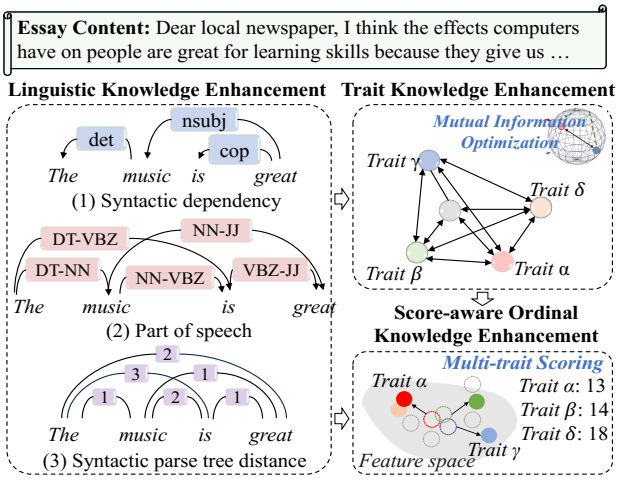


Figure 1: A simple illustration of our proposed method. It contains three core components.

(1) **Insufficient essay content encoding**. Existing methods mostly rely on hierarchical CNN-LSTM encoders (Kumar et al. 2022; He et al. 2022) or finetuned BERT encoders (Cho, Huang, and Kwon 2024; Li et al. 2025) to encode essay content. However, they focus on implicit essay semantics and ignore explicit structural relations between words, leading to limited performance. Xu et al. (2025) attempt to learn syntax-aware word features by directly concatenating part-of-speech tags (i.e., NN, VB, JJ) and dependency tags (i.e., det, cop, nsubj). However, they focus on single words and struggle to encode contextual essay content.

(2) **Poor trait feature representation**. Previous methods (Li and Pan 2025; Faseeh et al. 2024) mostly use separated dense layers (Chen and Li 2023; Do, Kim, and Lee 2023) or attention networks (Wang and Liu 2025) to learn trait features. However, they struggle to share knowledge between traits and fail to capture the relation dependencies between them. For the MES task, each trait is closely related to the other traits. For example, a well-organized essay often contributes positively to content clarity, indicating that the trait Organization can have a significant impact on the trait Content. To capture knowledge dependency, Do, Kim, and Lee (2024) formulate the MES task as an autoregressive genera-

*Corresponding author

tion by considering all trait scores as a sequence. However, they heavily rely on sequential interactions between traits, with each trait conditioned on the outputs of previous traits. Therefore, they share knowledge only in a fixed, unidirectional order, failing to obtain expressive trait features.

(3) **Low scoring reliability**. Existing methods (Li and Ng 2024b) focus on mean squared error or cross-entropy loss, which struggle to distinguish subtle feature differences between adjacent scores for ordinal scoring tasks. Therefore, they fall short in several scenarios with continuous and wide-ranging scores. Do, Ryu, and Lee (2024) attempt to handle the MES task using reinforcement learning with score-aware rewards. However, the reward functions are not sufficiently sensitive to subtle differences between adjacent scores, leading to unreliable scoring performance.

Therefore, the main challenges are summarized as follows: (1) How to effectively model the structural relations between words to enhance essay content? (2) How to learn the knowledge interactions between traits to capture expressive trait-specific features? (3) How to enhance scoring efficiency and effectiveness in complex scenarios with wide-ranging scores?

To overcome these challenges, we introduce a novel multi-trait scoring framework with multi-knowledge enhancement. As illustrated in Figure 1, the framework consists of three core components: linguistic knowledge enhancement, trait knowledge enhancement, and score-aware ordinal knowledge enhancement. **For the first challenge**, our proposed method utilizes linguistic knowledge to construct diverse syntactic structural relations between words (see the left side of Figure 1), enhancing contextual essay content and promoting fine-grained essay comprehension. **For the second challenge**, our proposed method explicitly models the relation interactions between multiple traits. With a graph structure, the proposed method leverages mutual information to capture knowledge dependencies and propagate relevant trait knowledge to learn discriminative trait-specific features. **For the final challenge**, the score-aware ordinal knowledge is utilized to explicitly align trait-specific features with their corresponding score rankings, improving scoring performance. The proposed method promotes the model to distinguish subtle differences between adjacent scores effectively, enhancing scoring performance in complex scenarios with wide-ranging scores. The main contributions are summarized as follows:

- We propose a novel multi-knowledge enhanced framework for the MES task, which utilizes linguistic knowledge enhancement, trait knowledge enhancement, and score-aware ordinal knowledge enhancement to solve the challenges derived from essay content encoding, trait feature learning, and multi-trait scoring.
- We construct syntactic structural relations between words to enhance essay content and model the knowledge interactions between traits to learn expressive trait-specific features. Ordinal knowledge is utilized to align trait-specific features with score rankings, improving performance in complex scenarios with continuous and wide-ranging scores.
- Extensive experiments show that our proposed method outperforms strong baselines and obtains significant performance in the MES task. The results demonstrate the effectiveness of our proposed method across different prompts and traits.

Related Work

Multi-trait Essay Scoring Existing methods (Hussein, Hassan, and Nassef 2020; Xue, Tang, and Zheng 2021; Kumar et al. 2022) could be divided into two mainstream lines: encoder-only models and encoder-decoder models. Encoder-only models mostly design a dedicated dense layer for each trait, but such shallow network structures limit their ability to model complex semantic context. Therefore, various attention networks (Ridley et al. 2021; Li and Pan 2025) are adopted to emphasize the context relevant to each trait. However, these methods ignore the syntactic structures between words. Xu et al. (2025) attempt to leverage syntactic knowledge to learn word features. However, they focus on single words and ignore the relation interactions between words, leading to limited performance. Encoder-decoder models (i.e., T5) (Do, Kim, and Lee 2024) use a sequence generation to model the knowledge flow between traits, capturing dependency relations across traits. However, they are limited to a fixed generation order and fail to model the flexible interactions between traits. Therefore, these methods fall short in the MES task.

Graph Neural Network Graph Neural Networks (GNNs) (Kong, Guo, and Liu 2024; Chen et al. 2022; Sun et al. 2023; Zhang et al. 2023a; Yu et al. 2025) have shown significant progress in modeling the relationship between instances. Using graph structures, GNNs could not only explicitly exploit the syntactic structure information to enhance essay semantics but also propagate knowledge between traits. Recently, GNNs have shown promising performance in sentiment analysis (Yin and Zhong 2024; Yu et al. 2024) and intent detection (Cui et al. 2024). Zhang et al. (2023a) construct an instance-level and a class-level GNN to propagate label information and feature structure. Zhao, Chen, and Wang (2025) introduce an instance relation learning network to learn the interaction relations between instances with class information, propagating label knowledge across instances. Zhao et al. (2024) propose a dual relations propagation network, which models similarity and diversity among instances to enhance instance features. However, these methods are rarely applied to the MES task, and the potential of GNNs has not been fully explored.

To fill this gap, we propose a multi-knowledge enhanced graph neural network, which utilizes linguistic knowledge to model the fine-grained relations between words and learns flexible knowledge interactions between traits to enhance trait-specific features. We further introduce score-aware ordinal knowledge to align trait-specific features with score rankings, improving scoring performance. Our proposed method solves the aforementioned challenges and achieves significant improvements.

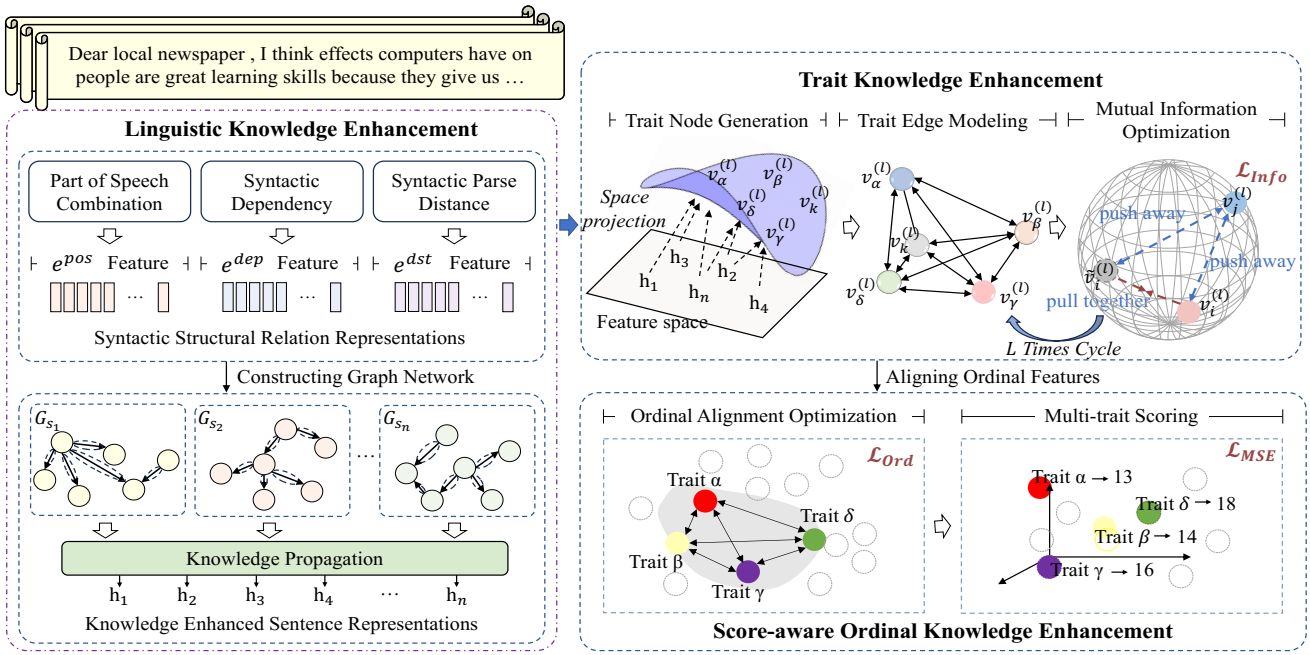


Figure 2: The overall architecture of the proposed method.

Methodology

Task Formulation

Given an essay set $E = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the i^{th} essay, N is the number of essays. We define $y_i = \{y_{i1}, y_{i2}, ..., y_{ik}\}$ as the set of trait scores assigned to x_i, where k is the number of traits, and $y_{ik} \in \mathbb{R}^{\mathcal{Y}}$ takes a continuous score value in the label space $\mathcal{Y}$. The goal of the MES task is to predict each trait score for a given essay.

Overall Framework

Figure 2 provides an overview of our proposed method, which consists of linguistic knowledge enhancement, trait knowledge enhancement, and score-aware ordinal knowledge enhancement. These important components are introduced in detail.

Linguistic Knowledge Enhancement

Given an essay x, it is denoted as $x = \{s_1, s_2, ..., s_n\}$, where n is the number of sentences. To enhance essay content, we construct a graph network for each sentence by analyzing its syntactic structures and semantic correlations. Each sentence-level graph is denoted as $\mathcal{G}_s = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. The nodes in $\mathcal{V}$ correspond to word features with contextual semantics, and the edge $e_{ij} \in \mathcal{E}$ represents the syntactic structural relations between the i^{th} word and the j^{th} word. Briefly, we represent words as nodes to model diverse structural relations between them. With linguistic knowledge propagation, our proposed method achieves structurally-informed essay encoding.

Node Initialization For a sentence with m words, we define $s = \{w_1, w_2, ..., w_m\}$ and use the encoder (e.g., BERT) to derive contextualized semantic features for each word. The nodes in the graph $\mathcal{G}_s$ are defined as $\mathcal{V} = \{v_i\}_{i=1}^{m}$, where v_i is the feature vector of the i^{th} word in $\mathcal{G}_s$. The node is represented as $v_i = f_{emb}(w_i) \in \mathbb{R}^d$, where $f_{emb}(\cdot)$ is the encoder function, and d is the feature dimension.

Edge Initialization Traditional graph networks (Zhang et al. 2023a; Zhao et al. 2024) use a numerical value (e.g., a weight) to represent the edge between two nodes, reflecting the strength or similarity between them. Unlike these methods, we utilize linguistic knowledge to construct multiple syntactic structural relations between words, including part-of-speech, syntactic dependency, and syntactic parse tree distance. To effectively represent edges, we define these structural relation features as: $e^{pos} \in \mathbb{R}^{d'}$, $e^{dep} \in \mathbb{R}^{d'}$, and $e^{dst} \in \mathbb{R}^{d'}$, where d' is the feature dimension. Each edge is represented as $e_{ij} = \{e_{ij}^{pos}, e_{ij}^{dep}, e_{ij}^{dst}\} \in \mathbb{R}^{3\times d'}$.

Linguistic Knowledge Propagation With the graph network, we propagate linguistic knowledge to enhance word features.

$$\tilde{v}_i^{\tau,(l)} = \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{\tau} W^{\tau} v_j^{(l-1)} + b^{\tau}, \tag{1}$$

$$\alpha_{ij}^{\tau} = \frac{\exp(e_{ij}^{\tau})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik}^{\tau})}, \tag{2}$$

where $e_{ij}^{\tau} \in \mathbb{R}^{d'}$ represents the edge feature vectors for structural relation $\tau \in \{pos, dep, dst\}$. The attention weights α_{ij}^{τ} are obtained by normalizing the edge features. $\mathcal{N}(i)$ represents the set of neighbors of the i^{th} node. W^{τ} and b^{τ} are learnable parameters. Then, we use element-wise

max pooling to obtain the final word feature vectors from different structural relations.

$$v_i^{(l)} = \max(\tilde{v}_i^{pos,(l)}, \tilde{v}_i^{dep,(l)}, \tilde{v}_i^{dst,(l)}) \tag{3}$$

Then, we learn sentence features from these word features to capture higher-level linguistic knowledge. Specifically, a mean pooling layer (Chen and Sun 2023) is utilized to calculate the average of feature vectors across all words, serving as the feature vector of the sentence.

$$h = \text{MeanPooling}(\mathcal{V}), \tag{4}$$

where $\mathcal{V}$ is the set of word feature vectors. h is the corresponding sentence feature vector. For an essay with n sentences, we could obtain enhanced sentence features: $H = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n \times d}$.

Trait Knowledge Enhancement

As shown in Figure 2, our proposed method captures relevant knowledge from different sentences and explicitly models the relation interactions between traits. Then, mutual information is leveraged to promote knowledge propagation across traits, learning expressive trait-specific features.

Trait Node Generation The proposed method generates trait features from linguistic-enhanced sentences by space projection.

$$P = \text{softmax}(W_2 \tanh(W_1 H + b_1) + b_2), \tag{5}$$

$$\mathcal{V}_t^{(l)} = P^T H, \tag{6}$$

where $\mathcal{V}_t^{(l)} \in R^{k \times d}$ indicates the set of trait nodes. k is the number of traits. In Figure 2, we use $v_\tau^{(l)}$ ($\tau \in \{\alpha, \beta, \gamma, \delta\}$) to represent each trait feature vector. Besides, P is a probability matrix with size $n \times k$ from sentence nodes to trait nodes. W_1, W_2, b_1 and b_2 are trainable parameters. The softmax($\cdot$) function ensures that the elements in each row of P are in the range [0, 1], and the sum of elements in each row is 1.

Trait Edge Modeling The edges represent the relevance between two traits. Following VS, Oza, and Patel (2023), we use key-query to model the relationships between traits, where the trait features are projected into key and query spaces. For the i^{th} trait node, its key (k_i), query (q_i), and node pairwise logits (e_{ij}) are as follows: $k_i^{(l)} = W_k^{(l)} \cdot v_i^{(l)}$, $q_i^{(l)} = W_q^{(l)} \cdot v_i^{(l)}$, and $e_{ij}^{(l)} = q_i^{(l)}(k_j^{(l)})^T$. $W_k^{(l)}$ and $W_q^{(l)}$ are trainable parameters.

Trait Knowledge Propagation The proposed method propagates knowledge from related traits.

$$\tilde{v}_i^{(l)} = \sum_j \frac{e_{ij}^{(l-1)}}{\sum_k e_{ik}^{(l-1)}} v_j^{(l-1)}, \tag{7}$$

$$v_i^{(l)} = W_3 \text{ReLU}(\text{MLP}(\tilde{v}_i^{(l)}) + v_i^{(l-1)}) + b_3, \tag{8}$$

where $v_i^{(l)}$ is the i^{th} node feature in the l^{th} layer. MLP($\cdot$) is a multi-layer perceptron. W_3 and b_3 are trainable parameters. Moreover, the proposed method could adjust edge features from the latest node features.

Mutual Information Optimization Different traits often exhibit inherent correlations (Do, Ryu, and Lee 2024), such as the trait Content is positively associated with the trait Organization. Therefore, we maximize mutual information between traits to encourage each trait node to learn the contextual knowledge from its neighborhood traits, promoting relevant knowledge propagation across traits. We denote the Probability Density Function (PDF) of the trait node v over the feature space as $p(v)$, and the PDF of $\tilde{v}$ (see Equation 7) as $p(\tilde{v})$. The mutual information is defined as:

$$\mathcal{I}(\tilde{v}^{(l)}; v^{(l)}) = \mathbb{E}_{p(\tilde{v}^{(l)}, v^{(l)})}(\log \frac{p(\tilde{v}^{(l)}, v^{(l)})}{p(\tilde{v}^{(l)}) \cdot p(v^{(l)})}) \tag{9}$$

However, computing mutual information is notoriously difficult in continuous and high-dimensional feature spaces. Following Dong et al. (2022), we convert mutual information maximization into minimizing the contrastive loss:

$$\mathcal{L}_{\text{Info}} = -\frac{1}{k} \sum_{i=1}^{k} \log \frac{\exp\left(\text{sim}(\tilde{v}_i^{(l)}, v_i^{(l)})\right)}{\sum_{j \in \mathcal{N}(i)} \exp\left(\text{sim}(v_i^{(l)}, v_j^{(l)})\right)}, \tag{10}$$

where, sim($\cdot, \cdot$) denotes the cosine similarity. $\mathcal{L}_{\text{Info}}$ contributes to graph smoothing and promotes knowledge propagation between traits, which has demonstrated strong performance in node or graph prediction tasks (Zhang et al. 2021).

Inspired by Yifan et al. (2020), we revisit the feature smoothness metric to demonstrate its effect.

$$\delta_f^{(l)} = \frac{\| \sum_{i \in \mathcal{V}_t} (\sum_{j \in \mathcal{N}(i)} (\tilde{v}_i^{(l)} - v_j^{(l)}))^2 \|}{|\mathcal{E}| \cdot D^{(l)}}, \tag{11}$$

where $|\mathcal{E}|$ is the number of edges, $D^{(l)}$ is the feature dimension at the l^{th} layer, and $||\cdot||$ denotes the L1 norm. Obviously, the feature smoothness metric $\delta_f^{(l)}$ is positively correlated to Kullback-Leibler (KL) divergence, i.e., $\mathcal{D}_{\text{KL}}(\tilde{v}^{(l)} \parallel v^{(l)}) \sim \delta_f^{(l)}$. Therefore, a large $\delta_f^{(l)}$ represents significant disagreement between $\tilde{v}^{(l)}$ and $v^{(l)}$. The KL divergence $\mathcal{D}_{\text{KL}}(\tilde{v}^{(l)} \parallel v^{(l)})$ reflects the difference between the neighborhood and trait nodes distributions. The KL divergence is negatively correlated with the mutual information, i.e., $\mathcal{I}(\tilde{v}^{(l)}; v^{(l)}) \sim 1/\mathcal{D}_{\text{KL}}(\tilde{v}^{(l)} \parallel v^{(l)})$. Therefore, maximizing $\mathcal{I}(\tilde{v}^{(l)}; v^{(l)})$ is equivalent to minimizing the KL divergence $\mathcal{D}_{\text{KL}}(\tilde{v}^{(l)} \parallel v^{(l)})$ and feature smoothness value $\delta_f^{(l)}$, which encourages knowledge propagation and facilitates smoothing over the graph structure.

Score-aware Ordinal Knowledge Enhancement

We introduce score-aware ordinal knowledge to encourage ordinal alignment within trait-specific features, enhancing further scoring effectiveness.

Multi-trait Scoring Loss Given the i^{th} essay, we utilize a regression function $f_\theta(\cdot)$ to map the j^{th} trait-specific feature v_{ij} to a predicted score $\hat{y}_{ij} = f_\theta(v_{ij})$. Mean Squared Error (MSE) is used to optimize the following training objective:

$$\mathcal{L}_{\text{MSE}}(y, \hat{y}) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \hat{y}_{ij})^2, \tag{12}$$

where M represents the number of traits for N essays, y is the ground truth, and $\hat{y}$ is the prediction. Although MSE is widely used in the essay evaluation, it struggles to distinguish the feature differences between adjacent scores (Zhang et al. 2023b). Therefore, we introduce score-aware ordinal knowledge to enhance scoring performance across continuous and wide-ranging score scales.

Score-aware Ordinal Knowledge To improve feature discriminability across continuous scores, we encourage latent feature vectors v to carry key information about their associated scores c. The goal can be formulated as maximizing the mutual information between the latent features and the scores: $\mathcal{I}(\mathbf{v}; c) = \mathcal{H}(c) - \mathcal{H}(c \mid \mathbf{v})$. We define $\mathcal{H}(c)$ as $\mathcal{H}(c) = -\int p(c) \log p(c)\, dc$, and it denotes the entropy of the score distribution. $\mathcal{H}(c \mid \mathbf{v}) = -\int p(\mathbf{v}, c) \log p(c \mid \mathbf{v})\, dv\, dc$ denotes the conditional entropy. It reflects the uncertainty of the score c given the latent feature vector v, During training, it is commonly assumed that the prior distribution of scores $p(c)$ is fixed, i.e., $\mathcal{H}(c) = \text{const}$. Therefore, maximizing $\mathcal{I}(\mathbf{v}; c)$ is equivalent to minimizing the conditional entropy $\mathcal{H}(c \mid \mathbf{v})$:

$$\max_{\phi} \mathcal{I}(\mathbf{v}; c) = \max_{\phi} [\mathcal{H}(c) - \mathcal{H}(c \mid \mathbf{v}_\phi)] \Rightarrow \min_{\phi} \mathcal{H}(c \mid \mathbf{v}_\phi), \tag{13}$$

where $\mathbf{v}_\phi$ denotes the feature vectors by the model parameterized by ϕ. The process encourages the conditional distribution $p(c \mid \mathbf{v})$ to become more deterministic, reducing uncertainty in the prediction of continuous scores. However, directly minimizing $\mathcal{H}(c \mid \mathbf{v})$ is typically intractable, since the posterior distribution $p(c \mid \mathbf{v})$ is not directly accessible. Therefore, we introduce a distance-based surrogate objective following Zhang et al. (2023b). For the i^{th} essay, we define the feature v_{i,c_j} corresponding to the j^{th} trait at score value c_j.

$$\mathcal{L}_{\text{Ord}} = -\frac{1}{NM(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k \in \mathcal{N}(j)} w_{ij} \left\| v_{i,c_j} - v_{i,c_k} \right\|_2 \tag{14}$$

where $w_{ij} = |y_{i,c_j} - y_{i,c_k}|$ is a weighting term that reflects the ordinal distance between score values. The loss penalizes large distances between the j^{th} trait and the k^{th} trait.

Training Objective

The overall training objective is written as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Info}} + \beta \mathcal{L}_{\text{Ord}} + \mathcal{L}_{\text{MSE}}, \tag{15}$$

where α and β are hyper-parameters.

Experiments

Experimental Setup

Implementation Details The proposed method is implemented with PyTorch (version 1.10.0) on a single GPU (RTX 3090 Ti) with CUDA version 11.3. We use the uncased English version of BERT-base (Devlin et al. 2019) as our backbone. We utilize NLP tools such as NLTK to extract part-of-speech tags and Stanza to obtain syntactic dependencies and syntactic parse tree distances. For syntactic structural relations, the feature dimension d' is 50. The AdamW optimizer trains the model with a learning rate of 5e-6. Meanwhile, we use the GradualWarmupScheduler to optimize the learning rate and set the warmup proportion to 0.05. We set L to 2 as the number of cycles for the updates in the trait knowledge propagation. Following Wang and Liu (2025), the experimental results are obtained by computing the average performance over five-fold cross-validation. We randomly select a subset (e.g., 100 samples) as the validation set and perform grid search to determine hyperparameters (i.e., α = 0.1, β = 0.1).

Dataset Extensive experiments are conducted on ASAP and ASAP++ (Mathias and Bhattacharyya 2018), which have been widely used for the MES task. The dataset contains eight sets of essays, each of which belongs to an essay prompt with multiple traits.

Baselines We divide strong baselines into encoder-only models and encoder-decoder models. For encoder-only models, **MTL** (Kumar et al. 2022) and **DualTrans** (Cho, Huang, and Kwon 2024) adopt separated dense layers to predict each trait score. **T-MES** (Wang and Liu 2025) introduces a trait-expert network to learn distinct expert weights for each trait. Encoder-decoder models formulate the MES task as a sequence generation. **T5** (i.e., **ArTS**), **BART**, **Pegasus**, and **LED** (Beltagy, Peters, and Cohan 2020; Do, Kim, and Lee 2024; Chu et al. 2025), as strong baselines, are used to compare scoring performance with our proposed method. Further, **SaMRL** (Do, Ryu, and Lee 2024) employs reinforcement learning with a scoring-aware reward based on the encoder-decoder architecture.

Evaluation Metric Following previous works (Wang and Liu 2025), we use Quadratic Weighted Kappa (QWK) as the main metric to evaluate the MES task.

Main Experimental Results

We conduct extensive experiments in Table 1 and Table 2, with the following observations.

(1) Overall, our proposed method achieves the best results compared with strong baselines. In terms of encoder-only models, our proposed method achieves an average improvement of 1% across prompts and 1.2% across traits. Most encoder-only models use separated dense layers to predict each trait score, but they ignore the knowledge interactions between traits, leading to suboptimal performance. We utilize the graph structure to model the interactions between traits and leverage mutual information to optimize knowledge propagation. In terms of encoder-decoder models, our method achieves up to an average of 0.9% improvement

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑
Pegasus	0.639	0.520	0.518	0.562	0.636	0.597	0.539	0.478	0.561
BART	0.647	0.602	0.658	0.727	0.713	0.713	0.624	0.534	0.652
LED	0.704	0.650	0.679	0.705	0.701	0.707	0.638	0.520	0.663
ArTS	0.712	0.680	0.713	0.771	0.730	0.775	0.747	0.595	0.715
SaMRL	0.717	0.703	0.715	0.773	0.729	**0.778**	0.745	0.604	0.720
MTL	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665
DualTrans	0.712	0.671	0.690	0.760	0.714	0.740	0.748	0.620	0.707
T-MES	0.728	0.684	0.702	0.771	0.726	0.754	**0.755**	0.629	0.719
Ours	**0.732**	**0.704**	**0.719**	**0.782**	**0.734**	0.771	**0.755**	**0.636**	**0.729**

Table 1: Average QWK scores across all traits for each prompt. The results are divided into two parts: encoder-decoder models (upper) and encoder-only models (lower).

Model	Over	Cont	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑
Pegasus	0.536	0.584	0.608	0.586	0.629	0.578	0.515	0.559	0.519	0.578	0.388	0.553
BART	0.701	0.672	0.711	0.664	0.705	0.600	0.588	0.624	0.601	0.646	0.547	0.642
LED	0.709	0.677	0.706	0.666	0.707	0.627	0.633	0.643	0.640	0.655	0.522	0.653
ArTS	0.754	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	0.721	0.570	0.695
SaMRL	0.750	0.732	0.754	0.704	0.740	0.670	0.684	0.681	0.685	**0.726**	0.558	0.699
MTL	0.764	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.639
DualTrans	**0.778**	0.726	0.732	0.660	0.704	0.682	0.668	0.674	0.663	0.689	0.619	0.690
T-MES	0.774	0.730	0.750	0.702	0.730	0.685	**0.686**	0.679	0.675	0.693	0.590	0.699
Ours	0.770	**0.736**	**0.759**	**0.708**	**0.744**	**0.691**	**0.686**	**0.691**	**0.699**	0.714	**0.623**	**0.711**

Table 2: Average QWK scores across all prompts for each trait. Over: Overall, Cont: Content, Org: Organization, WC: Word Choice, SF: Sentence Fluency, Conv: Conventions, PA: Prompt Adherence, Lang: Language, Nar: Narrativity.

across prompts and 1.2% across traits. The strong baseline, SaMRL, achieves competitive results on Prompt P6 and the trait Style, but its overall performance remains significantly lower than our proposed method. SaMRL is built upon T5 with a large number of parameters, whereas our method is based on BERT, a lighter encoder-only model. Compared to SaMRL, our proposed method still achieves maximum improvements of 3.2% and 6.5% in Table 1 and Table 2, respectively. The results demonstrate the effectiveness of our proposed method on the MES task.

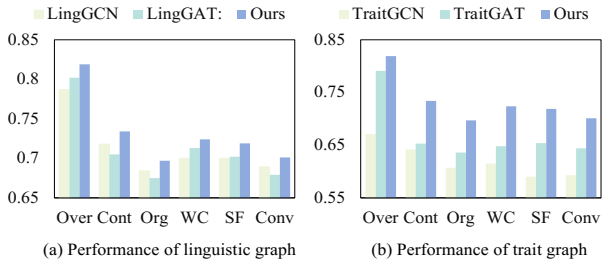
(2) For all the mentioned methods, Prompt P8 consistently shows lower performance than other prompts. The reason lies in Prompt P8 providing significantly less data, while the other prompts have more than twice the amount of data. The limited amount of data makes it challenging for baselines, whereas our proposed method leverages linguistic knowledge, trait knowledge, and ordinal knowledge to enhance multi-trait scoring performance. The results demonstrate our robustness and superior effectiveness compared to baselines. Compared to other traits, the trait Voice is exclusive to Prompt P8, where our method achieves superior performance over encoder-only models by 3.3%-4.1% and encoder-decoder models by 6.5%-23.5%. The results also demonstrate the robustness of our proposed method in low-resource settings, where it consistently outperforms baselines. In conclusion, our proposed method utilizes multi-knowledge enhancement to improve scoring performance for the MES task.

Traits	Xlnet		Bert		Roberta	
	T-MES	Ours	T-MES	Ours	T-MES	Ours
Over	0.672	**0.674**	0.792	**0.819**	0.817	**0.821**
Cont	0.615	**0.628**	0.726	**0.734**	0.714	**0.716**
Org	0.595	**0.596**	0.687	**0.690**	0.666	**0.699**
WC	0.611	**0.622**	0.721	**0.724**	0.702	**0.710**
SF	0.630	**0.630**	0.714	**0.719**	0.691	**0.715**
Conv	0.620	**0.622**	0.700	**0.701**	0.674	**0.692**

Table 3: Comparison of QWK scores for Prompt P1 across different encoders.

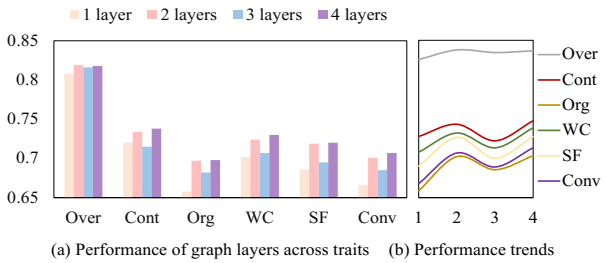
Performance across Encoders

We further conduct experiments to investigate performance across different encoders, as shown in Table 3. Compared to the best encoder-only model, T-MES, we consistently outperform it across various encoder architectures. Our proposed method not only exhibits strong generalizability across different encoders but also robustly maintains its effectiveness, demonstrating clear superiority in the MES task. Notably, the average 1.48% improvement under Roberta indicates that our proposed method adapts well to differ-



(a) Performance of linguistic graph (b) Performance of trait graph

Figure 3: Impact of graph structure on Prompt P1.



(a) Performance of graph layers across traits (b) Performance trends

Figure 4: The impact of graph layers on Prompt P1.

ent encoder backbones, especially those with stronger pretrained knowledge. In conclusion, experimental results further demonstrate that our proposed method achieves convincing performance on the MES task.

Impact of Graph Structure

Our proposed method includes two distinct types of graphs: one to enhance linguistic knowledge and another to enhance trait knowledge. In Figure 3, we conduct experiments to evaluate the effectiveness of these two graph structures. Specifically, we compare these two graphs with representative Graph Attention Network (GAT) (Vrahatis, Lazaros, and Kotsiantis 2024) and Graph Convolutional Network (GCN) (Li et al. 2021).

Linguistic Graph Structure Figure 3(a) presents the performance across diverse graph structures for linguistic knowledge enhancement. Our graph structure achieves the best performance. The improvement stems from explicitly modeling syntactic structural relations between words. While GAT adaptively assigns weights to neighboring nodes via attention mechanisms, GCN employs fixed weights through graph convolutions, our proposed method offers more effective and structurally-informed essay encoding.

Trait Graph Structure Figure 3(b) presents the performance across diverse graph structures for trait knowledge enhancement. The results demonstrate consistent improvements over GAT and GCN, verifying the effectiveness of the proposed method. Our proposed method utilizes the key-query to model the relationships between traits, contributing to significant performance on the MES task.

Impact of Knowledge Propagation Layer

To investigate the impact of the knowledge propagation layer, we evaluate the proposed method with one to four layers in trait knowledge propagation. As illustrated in Figure 4, the performance exhibits a generally increasing trend with fluctuations as the number of layers increases. Compared to other traits, the overall scoring remains relatively stable and consistently achieves high performance. The results indicate that trait knowledge propagation has a positive effect, though excessive layers may introduce local saturation and feature redundancy. Briefly, our proposed method could capture intricate trait interactions and enhance scoring performance for the MES task.

Model	Over	Cont	Org	WC	SF	Conv
Full model	**0.819**	**0.734**	**0.697**	**0.724**	**0.719**	**0.701**
w/o LKE	0.799	0.724	0.683	0.722	0.710	0.694
w/o TKE	0.679	0.666	0.640	0.651	0.619	0.596
w/o OKE	0.816	0.727	0.681	0.711	0.699	0.683
w/o MIO	0.813	0.731	0.687	0.718	0.707	0.689

Table 4: An ablation study on Prompt P1. LKE: linguistic knowledge enhancement, TKE: trait knowledge enhancement, OKE: ordinal knowledge enhancement, and MIO: mutual information optimization.

Ablation Study

In Table 4, we conduct an ablation study. (1) "w/o LKE". We remove linguistic knowledge from the essay content encoding. The negative results highlight its crucial role in multi-trait evaluation. (2) "w/o TKE". We remove trait graph construction and discard inter-trait knowledge propagation. The observed performance degradation suggests that trait knowledge enhancement has a positive effect by sharing knowledge between traits. (3) "w/o OKE". After we remove ordinal knowledge, the degraded performance highlights the importance of aligning features with score rankings in multi-trait essay scoring. (4) "w/o MIO". When mutual information optimization is removed from trait knowledge enhancement, the performance drops a lot. The fact suggests that the loss term $\mathcal{L}_{\text{Info}}$ is beneficial for promoting effective knowledge propagation. In conclusion, the complete model significantly outperforms all ablation studies and achieves the best performance.

Conclusion

We propose a novel multi-knowledge enhanced graph neural network for the MES task. Specifically, we construct diverse syntactic structural relations between words to learn structurally-informed essay encoding and enhance fine-grained essay comprehension. We model the knowledge interactions between traits and share knowledge across them, learning expressive and discriminative trait-specific features. To improve scoring performance, ordinal knowledge is utilized to align trait-specific features with score rankings, enhancing its effectiveness in complex scenarios with continuous and wide-ranging scores. Extensive experiments show that we significantly outperform strong baselines.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG22/23).

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Boquio, E. N. V.; and Naval, P. C., Jr. 2024. Beyond Canonical Fine-tuning: Leveraging Hybrid Multi-Layer Pooled Representations of BERT for Automated Essay Scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2285–2295.

Chen, J.; Yang, Y.; Yu, T.; Fan, Y.; Mo, X.; and Yang, C. 2022. Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning. In *Proceedings of ACM SIGKDD*, 2741–2751.

Chen, Y.; and Li, X. 2023. PMAES: Prompt-mapping Contrastive Learning for Cross-prompt Automated Essay Scoring. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1489–1503.

Chen, Z.; and Sun, Q. 2023. Extracting class activation maps from non-discriminative features as well. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3135–3144.

Cho, M.; Huang, J.-X.; and Kwon, O.-W. 2024. Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal*, 46(1): 82–95.

Chu, S.; Kim, J. W.; Wong, B.; and Yi, M. Y. 2025. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 5796–5814.

Cui, M.; Zhang, R.; Xiang, H.; and Xue, S. 2024. Enhanced Multi-Intent Recognition with BERT Embeddings and Graph-based Decoding. In *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 165–172. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 57st Annual Meeting of the Association for Computational Linguistics*.

Do, H.; Kim, Y.; and Lee, G. 2024. Autoregressive Score Generation for Multi-trait Essay Scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, 1659–1666.

Do, H.; Kim, Y.; and Lee, G. G. 2023. Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1538–1551.

Do, H.; Ryu, S.; and Lee, G. 2024. Autoregressive Multi-trait Essay Scoring via Reinforcement Learning with Scoring-aware Multiple Rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16427–16438.

Dong, W.; Wu, J.; Luo, Y.; Ge, Z.; and Wang, P. 2022. Node representation learning in graph via node-to-neighbourhood mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16620–16629.

Faseeh, M.; Jaleel, A.; Iqbal, N.; Ghani, A.; Abdusalomov, A.; Mehmood, A.; and Cho, Y.-I. 2024. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics*, 12(21): 3416.

Fiacco, J.; Adamson, D.; and Rose, C. 2023. Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 232–241.

He, Y.; Jiang, F.; Chu, X.; and Li, P. 2022. Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, 3007–3016.

Hussein, M. A.; Hassan, H. A.; and Nassef, M. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).

Jiang, Z.; Gao, T.; Yin, Y.; Liu, M.; Yu, H.; Cheng, Z.; and Gu, Q. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12456–12470.

Kong, W.; Guo, Z.; and Liu, Y. 2024. Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8627–8635.

Kumar, R.; Mathias, S.; Saha, S.; and Bhattacharyya, P. 2022. Many Hands Make Light Work: Using Essay Traits to Automatically Score Essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1485–1495.

Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; and Hovy, E. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. 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)*, 6319–6329.

Li, S.; and Ng, V. 2024a. Automated Essay Scoring: A Reflection on the State of the Art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17876–17888.

Li, S.; and Ng, V. 2024b. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8458–8478.

Li, S.; Zhao, S.; Zhang, Z.; Fang, Z.; Chen, W.; and Wang, T. 2025. Basis is also explanation: Interpretable Legal Judgment Reasoning prompted by multi-source knowledge. *Information Processing Management*, 62(3): 103996.

Li, X.; and Pan, W. 2025. KAES: Multi-aspect Shared Knowledge Finding and Aligning for Cross-prompt Automated Scoring of Essay Traits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24476–24484.

Mathias, S.; and Bhattacharyya, P. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Ridley, R.; He, L.; Dai, X.-y.; Huang, S.; and Chen, J. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 13745–13753.

Sun, X.; Cheng, H.; Liu, B.; Li, J.; Chen, H.; Xu, G.; and Yin, H. 2023. Self-supervised hypergraph representation learning for sociological analysis. *IEEE Transactions on Knowledge and Data Engineering*.

Vrahatis, A. G.; Lazaros, K.; and Kotsiantis, S. 2024. Graph attention networks: a comprehensive review of methods and applications. *Future Internet*, 16(9): 318.

VS, V.; Oza, P.; and Patel, V. M. 2023. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3520–3530.

Wang, D.; and Jin, W. 2025. The impact mechanism of AES on improving English writing achievement. *Scientific Reports*, 15(1): 3928.

Wang, J.; and Liu, J. 2025. T-MES: Trait-Aware Mix-of-Experts Representation Learning for Multi-trait Essay Scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, 1224–1236.

Xie, J.; Cai, K.; Kong, L.; Zhou, J.; and Qu, W. 2022. Automated Essay Scoring via Pairwise Contrastive Regression. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 2724–2733.

Xu, J.; Liu, J.; Lin, M.; Lin, J.; Yu, S.; Zhao, L.; and Shen, J. 2025. EPCTS: Enhanced Prompt-Aware Cross-Prompt Essay Trait Scoring. *Neurocomputing*, 621: 129283.

Xue, J.; Tang, X.; and Zheng, L. 2021. A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring. *Ieee Access*, 9: 125403–125415.

Yang, R.; Cao, J.; Wen, Z.; Wu, Y.; and He, X. 2020. Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1560–1569. Online.

Yifan, H.; Jian, Z.; James, C.; Kaili, M.; TB, M. R.; Hongzhi, C.; and Ming-Chang, Y. 2020. Measuring and improving the use of graph information in graph neural network. In *The Eighth international conference on learning representations (ICLR 2020), Addis Ababa*.

Yin, S.; and Zhong, G. 2024. Textgt: A double-view graph transformer on text for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 19404–19412.

Yu, M.; Peng, F.; Zhao, Y.; Zhang, W.; Yu, J.; and Zhao, M. 2024. IDSV-GCN: integrating dual syntactic views graph convolutional network for aspect-based sentiment analysis. *Knowledge-Based Systems*, 305: 112656.

Yu, P.; Gu, J.; Pi, D.; Zhou, Q.; and Wang, Q. 2025. Aspect-aware graph interaction attention network for aspect category sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Zhang, F.; Chen, W.; Ding, F.; and Wang, T. 2023a. Dual Class Knowledge Propagation Network for Multi-label Few-shot Intent Detection. In *Proc. of ACL*, 8605–8618.

Zhang, S.; Yang, L.; Mi, M. B.; Zheng, X.; and Yao, A. 2023b. Improving deep regression with ordinal entropy. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Zhang, W.; Yang, M.; Sheng, Z.; Li, Y.; Ouyang, W.; Tao, Y.; Yang, Z.; and Cui, B. 2021. Node dependent local smoothing for scalable graph learning. *Advances in Neural Information Processing Systems*, 34: 20321–20332.

Zhao, S.; Chen, W.; and Wang, T. 2025. Instance Relation Learning Network with Label Knowledge Propagation for Few-shot Multi-label Intent Detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Zhao, S.; Xie, Y.; Chen, W.; Wang, T.; Yao, J.; and Zheng, J. 2024. Metric-Free Learning Network with Dual Relations Propagation for Few-Shot Aspect Category Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, 12: 100–119.