GDLLM: A Global Distance-aware Modeling Approach Based on Large Language Models for Event Temporal Relation Extraction

Anonymous ACL submission

Abstract

In Natural Language Processing(NLP), Event Temporal Relation Extraction (ETRE) is to recognize the temporal relations of two events. Prior studies have noted the importance of language models for ETRE. However, the restricted pre-trained knowledge of Small Language Models(SLMs) limits their capability to handle minority class relations in imbalanced classification datasets. For Large Language Models(LLMs), researchers adopt manually designed prompts or instructions, which may introduce extra noise, leading to interference with the model's judgment of the long-distance dependencies between events. To address these issues, we propose **GDLLM**, a **G**lobal **D**istance-aware modeling approach based on **LLM**s. We first present a distance-aware graph structure utilizing Graph Attention Network(GAT) to assist the LLMs in capturing long-distance dependency features. Additionally, we design a temporal feature learning paradigm based on soft inference to augment the identification of relations with short-distance proximity band, which supplements the probabilistic information generated by LLMs into the multi-head attention mechanism. Since the global feature can be captured effectively, our framework substantially enhances the performance of minority relation classes and improves the overall learning ability. Experiments on two publicly available datasets, TB-Dense and MATRES, demonstrate that our approach achieves state-of-the-art (SOTA) performance. Our code will be available after the paper is accepted.

1 Introduction

In Natural Language Processing (NLP), Event Temporal Relation Extraction (ETRE) aims to identify temporal connections between event pairs. As illustrated in Figure 1(a), in the given sentence, the relation between the target Event1 **continues** and Event2 **grip** is *IS_INCLUDED*.

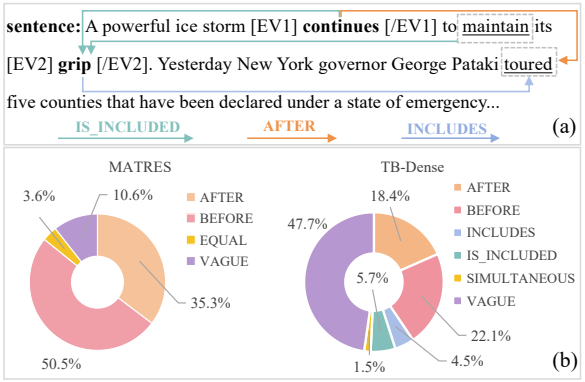

Figure 1: (a) is an example of the ETRE task. Above the arrows in the legend are the corresponding relation categories. "$[EV_i]$" is the hand-crafted symbol that can explicitly mark event boundaries in such examples. (b) is the relation distribution on two datasets.

Much of the existing studies pay attention to the crucial role of language models for ETRE, especially Small Language Models(SLMs). Some research utilizes SLMs to form certain rules for temporal realtion(Zhang et al., 2022; Man et al., 2022; Zhuang et al., 2023). Prior SOTA model MulCo(Yao et al., 2024) combines GNNs and the model of BERT variants via multi-scale knowledge distillation to enhance the performance of ETRE. However, the restricted pre-trained knowledge of SLMs limits their capability to handle minority class relations in imbalanced classification datasets(UzZaman et al., 2013; Guan et al., 2021). Although some researchers have invested substantial effort in it (Han et al., 2019; Ning et al., 2024; Yuan et al., 2024), the performance of their models is still suboptimal on two popular datasets, MATRES(Ning et al., 2019) and TB-Dense(Cassidy et al., 2014). As depicted in Figure 1(b), the relation "*SIMULTANEOUS*" that refers to two events happening simultaneously only takes 1.5% in the TB-Dense dataset while "*VAGUE*" has 47.7%(Yuan et al., 2024).

Recent advancements have noted the impressive capabilities of Large Language Models(LLMs) for ETRE. However, based on the powerful learning ability for contextual knowledge, prior studies rely on manually designed prompts and instructions to fine-tune LLMs(Hu et al., 2025; Xu et al., 2025), leading to noise accumulation(Chen et al., 2024) that interferes with the model's judgment of the global event relation feature. As shown in Figure 1(a), unlike most event pairs among Events "**continues**", "**grip**" and "**toured**", two another events occur between Events "**continues**" and "**toured**" in the text and make their distance of the occurrence order is longer. This indicates that there are two different event relation features that constitute the global feature: long-distance dependency and short-distance proximity band. Since modeling global event relation feature poses a challenge for researchers, they often neglect the recognition of long-distance dependency features when adopting manually designed prompts and instructions, which is also not conducive to handling minority class relations in imbalanced classification datasets.

To resolve the aforementioned problems, we propose **GDLLM**, a **G**lobal **D**istance-aware modeling approach based on **LLM**s, enabling the effective identification of event relations with the global feature to alleviate the impact of data imbalance on classification results. To be specific, we select the Graph Attention Network(GAT) to assist the fine-tuned LLMs in capturing event relations with long-distance dependency features, which circumvents the limitations of manually designed prompts or instruction templates. Compared to the "hard classification" (0/1 decision labels) as graph edge features, we integrate the probability distribution generated by LLMs into GAT to learn more comprehensive relation information. Both the probabilistic information and the multi-head attention mechanism augment the identification of relations with a short-distance proximity band. Since the global feature can be captured effectively, our framework substantially enhances the performance of minority relation classes and the overall learning ability.

Our contributions can be summarized as follows:

- We propose **GDLLM**, a **G**lobal **D**istance-aware modeling approach based on **LLM**s. Specifically, we introduce a global modeling method integrating LLMs and GAT, which is to identify the minority categories more effectively in imbalanced classification datasets.
- We present a distance-aware graph structure utilizing Graph Attention Network to assist the fine-tuned LLMs in capturing event relations with long-distance dependency features, which circumvents the limitations imposed by manually designed prompts or instructions.
- We design a temporal feature learning paradigm based on soft inference to augment the relation extraction with a short-distance proximity band. Rather than 0/1 decision labels, the probability distribution we selected as edge features enables the GAT to learn more comprehensive relation information.
- We conduct extensive experiments on two public datasets, TB-Dense and MATRES, which demonstrate that our approach outperforms all existing LLM-based and GNN-based benchmarks, achieving state-of-the-art (SOTA) performance without manually designed prompts or instructions for LLMs.

2 Method

In this section, we introduce the overall architecture of our proposed **GDLLM** method, which is depicted in Figure 2. Firstly, we formulate the ETRE task. Secondly, we introduce the LLM-based probability distribution prediction module. Finally, we present the distance-aware graph attention module.

2.1 Problem Formulation

Following Previous work, we define ETRE as a text classification task. Given a sentence T that contains two events E_1 and E_2, our aim is to identify the temporal relation between these two events. The output of our model is the particular temporal relation label prediction.

2.2 Probability Distribution Prediction

Input and Fine-tuning for the LLM. In our work, the unified format defined that input to LLMs from datasets contains manually designed symbols of the form $[EV_i]$ in the given sentence T, where i denotes the ordinal number of an event pair. It serves as a marker to annotate the boundaries of the event. The Appendix A takes Llama and Qwen as examples to show the specific input details. Before generating probabilistic information, we fine-tune the LLM based on LoRA(Hu et al., 2022). As depicted in Figure 2(a), the LoRA fine-tuned technique is

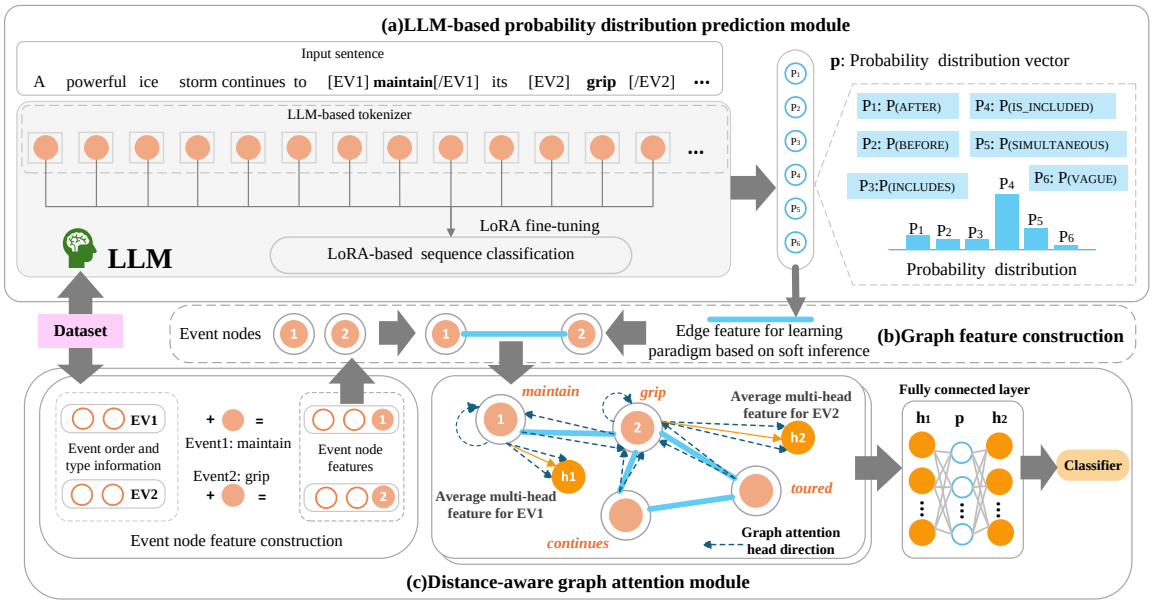

Figure 2: Overall architecture of our proposed method.

used for sequence classification, which is adopted for parameter-efficient fine-tuning.

Probability generation. As shown in Figure 2(a), while applying the LoRA fine-tuning, the model is ready to make predictions of probabilistic information generated by the LLM to construct edge features of graph structure, forming the soft inference-based temporal feature learning paradigm we designed. For each pair of events (E_i, E_j) in the document, the LLM outputs a probability distribution over a set of predefined and annotated event relation classes.

Specifically, we define c as the number of event relation classes, and the output of the LLM for the event pair (E_i, E_j) is a vector $\mathbf{p}_{ij} \in \mathbb{R}^c$. In the inference process of LoRA tuning, the model first generates a set of logits for each event pair. These logits are then passed through the softmax function. This operation converts the logits into probabilities, which represent the likelihood of each event pair belonging to different relation labels. For c kinds of relation types, and a specific event pair (E_i, E_j), the probability of it belonging to relation r is denoted as $P(p = r|E_i, E_j)$. Mathematically, the logits for an event pair are $z_1, z_2, \cdots, z_n$, then the probability $P(p = r|E_i, E_j)$ is calculated as:

$$P(p = r|E_i, E_j) = \frac{e^{z_r}}{\sum_{n=1}^{c} e^{z_n}}, \qquad (1)$$

which normalizes the logits so that the sum of probabilities for all relation classes is equal to 1, and they are all stored in the probability distribution vectors $\mathbf{p}_{ij} \in \mathbb{R}^C$. As depicted in Figure 2(a), rather than determining the most likely temporal relation between events, these probabilities are made to be a vector sequence distribution to provide more comprehensive pre-trained information for the subsequent module. For the TB-Dense dataset, which is shown in Figure 2(a) as an instance, the LLM provides the prediction distribution of the six labels it has, while the MATRES dataset does so for the four labels it possesses.

The training objective is the cross-entropy loss for multi-class classification based on LLM, which does not participate in the final loss calculation, and the calculation details of the loss function are similar to the final classification.

Notably, a topic that deserves discussion is why we choose LLMs to be the main language models. We generate probability distributions through different language models and visualize these distributions in scatter plots for comparison of the accuracy of the probability value. To be specific, we randomly select a sample sequence from the TB-Dense dataset and compare the distribution of probabilities for positive samples. It can be seen from Figure 3 that LLMs can present probability values with higher accuracy, which always assign a value closer to 1 for the relation category with the highest probability, while for low probability prediction values, their distributions tend to be closer to zero. In Appendix C, we compare the distribution of probabilities on the MATRES dataset, from which we can draw the same conclusions as above.

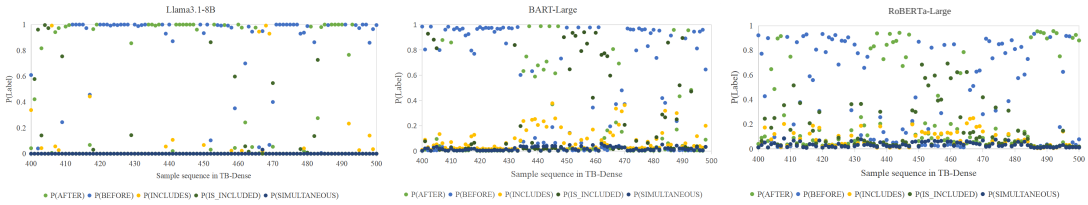

Figure 3: The distribution of probabilities generated by different language models on the TB-Dense dataset.

2.3 Distance-aware Graph Attention Module

From the previous section, we have obtained the probability distribution vectors $\mathbf{p}_{ij} \in \mathbb{R}^C$ for event pair predictions generated by the LLM. Next, we will introduce the construction of graph features first, followed by the temporal feature learning paradigm based on soft inference. Since our graph structure is a solution for capturing relation features at different distances, we define the proposed GAT-based architecture as a distance-aware approach.

Graph feature construction. As depicted in Figure 2(c), we construct a graph to model the relations between events based on every complete document. Compared with traditional graph construction methods, our approach aims to be more conducive to enabling the graph structure to learn accurate global relational features at an earlier stage. Each event E_i in the document and its order and type information are both represented as a node $v_i \in V$. And the node features $h_i^{(0)} \in \mathbb{R}^{d_h}$ are obtained from the dataset corresponding to the event.

For edge feature, which is shown as Figure 2(b), it exists between every pair of nodes, and the edge features are initialized as the probability distribution vectors $\mathbf{p}_{ij} \in \mathbb{R}^C$ for the event pair (E_i, E_j), which is to form our temporal feature learning paradigm based on soft inference.

Temporal feature learning paradigm. We design a temporal feature learning paradigm based on soft inference as depicted in Figure 2(b), which is to supplement the probabilistic information generated by LLMs into the multi-head attention mechanism. This paradigm shifts the edge feature representation from the previous 0/1 decision label to a probability distribution for "soft inference", which augments the identification of relations with short-distance proximity band. To achieve this, we apply this paradigm to the edge feature learning of GAT, which constructs a graph structure to model event relations with a multi-head attention mechanism.

Specifically, our Graph Attention Network architecture consists of multiple layers of multi-head attention mechanisms. In our implementation, in order to enable the model to learn more diverse feature combinations and interaction information, we adopt two layers with $K = 8$ attention heads. In the first GAT Layer, for each node v_i, the output of the k-th attention head is computed as:

$$\hat{\mathbf{h}}_{i,k}^{(1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij,k} \mathbf{W}_k^{(1)} \mathbf{h}_j^{(0)}\right), \quad (2)$$

where $\mathcal{N}(i)$ is the set of neighboring nodes of v_i, $\mathbf{W}_k^{(1)} \in \mathbb{R}^{d_h \times d_{h1}}$ is the weight matrix for the k - th head, σ is the activation function LeakyReLU, and the attention coefficients $\alpha_{ij,k}$ are calculated as:

$$\mathbf{z}_{ij,k} = \mathbf{a}_k^\top [\mathbf{W}_k^{(1)} \mathbf{h}_i^{(0)} \parallel \mathbf{W}_k^{(1)} \mathbf{h}_j^{(0)} \parallel \mathbf{p}_{i,j}], \quad (3)$$

$$\mathbf{z}_{im,k} = \mathbf{a}_k^\top [\mathbf{W}_k^{(1)} \mathbf{h}_i^{(0)} \parallel \mathbf{W}_k^{(1)} \mathbf{h}_m^{(0)} \parallel \mathbf{p}_{i,m}], \quad (4)$$

$$\alpha_{ij,k} = \frac{\exp\left(\text{LeakyReLU}(\mathbf{z}_{ij,k})\right)}{\sum_m \exp\left(\text{LeakyReLU}(\mathbf{z}_{im,k})\right)}, \quad (5)$$

where $\mathbf{a}_k \in \mathbb{R}^{3d_{h1}}$ is a learnable attention vector, "$\parallel$" denotes concatenation, and $m \in \mathcal{N}(i)$. Afterwards, the output of the first layer for node v_i is then concatenated with the outputs of all heads: $\mathbf{h}_i^{(1)} = \text{Concat}(\hat{\mathbf{h}}_{i,1}^{(1)}, \cdots, \hat{\mathbf{h}}_{i,K}^{(1)})$. For the second GAT layer, following a similar process of the first layer, we get the output of the k-th attention head $\hat{\mathbf{h}}_{i,k}^{(2)}$, and the final average multi-head feature $\mathbf{h}_i^{(2)}$.

Final Classification. In the final classification stage, as depicted in Figure 2(c), we integrate the output of the second GAT layer and the processed edge features $\mathbf{p}_{i,j}$. We concatenate these two types of features as: $\mathbf{h}_o = [\mathbf{h}_i^{(2)} \parallel \mathbf{p}_j \parallel \mathbf{h}_j^{(2)}]$. The concatenated feature vector $\mathbf{h}_o$ is then fed into a fully-connected layer. The output of the fully-connected layer is calculated as follows:

$$\mathbf{s} = \mathbf{W}_{\text{cls}} \mathbf{h}_o + \mathbf{b}_{\text{cls}}, \quad (6)$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{d_{h2} \times C}$ is the weight matrix of the classification layer, and $\mathbf{b}_{\text{cls}} \in \mathbb{R}^C$ is the bias vector of the classification layer.

Subsequently, we apply the softmax function for the output of the fully-connected layer to obtain the predicted probability distribution over the classes. The softmax function is defined as:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{s}), \tag{7}$$

where $\hat{\mathbf{y}}$ is the predicted probability for an event pair (E_i, E_j).

We employ the cross-entropy loss function to measure the difference between the final predicted probability and the true label. Given the true label $\mathbf{y} = (y_1, y_2, \cdots, y_C)$, the cross-entropy loss for this event pair is calculated as:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{C} y_k \log(\hat{y}_k). \tag{8}$$

3 Experiments and Results

3.1 Datasets and Metrics

We validate our approach on two widely adopted datasets: MATRES (Ning et al., 2019), and TB-Dense(Cassidy et al., 2014). The data splits, along with relation pairs statistics, are reported in Appendix B. The input format in zero-shot learning scenarios is illustrated in Appendix A. In accordance with prior study(Han et al., 2019), we adopt the micro-F1 score, with the *VAGUE* label excluded, as the evaluative metric for the datasets.

3.2 Experimental Setup

We compare our method with the following baselines: **1) LLM-based approaches:** these methods leverage LLMs to encode contextual information and perform temporal reasoning through prompt or instruction tuning(Xu et al., 2025; Hu et al., 2025). Prior studies also explore zero-shot temporal relation extraction using different prompt strategies(Yuan et al., 2023; Xu et al., 2025). Following this work, we conduct zero-shot experiments on two LLMs, the closed-source GPT4o(Hurst et al., 2024) and the open-source Llama3.1. **2) Graph-based approaches:** These models construct event graphs to capture temporal information, often using Graph Neural Networks (GNNs) to propagate information(Mathur et al., 2021; Zhang et al., 2022; Zhou et al., 2022; Yao et al., 2024). **3) Other benchmarks:** Methods that do not fit into the above categories but have shown strong performance, often combining neural networks or heuristic features(Huang et al., 2023; Tan et al., 2023; Ning et al., 2024; Yuan et al., 2024). In addition, we employ RoBERTa-Large (Liu et al., 2019) and BART-Large (Lewis et al., 2020), as two baseline models for comparison of SLMs.

As for fine-tuning LLMs, the LoRA rank is set to 16. All experiments are trained on NVIDIA A800 GPUs with 80GB of memory. In this paper, following previous work for hyperparameter optimization(Yao et al., 2024), we employ the HEBO (Heuristic-Efficient Bayesian Optimization) algorithm. We show the experiment implementation details aforementioned in Appendix B.

3.3 Main Results

As shown in Table 1, our method achieves SOTA performance in all existing methods and baseline methods. It is apparent from this table that very few models utilize LLMs as their language model to chase superior performance, but our methods adopt LLMs and outperform all previous models without manually designed prompts or instructions tuning. Meanwhile, unlike previous approaches, we arrange LLMs not as a standalone reasoning model, which also shows the effectiveness of utilizing distance-aware graph structure to form our approach, and is validated to capture the global feature of temporal event relations.

Additionally, our method *GDLLM(Ours)* outperforms the previous SOTA model(Yao et al., 2024) that only adopts SLM as its language model, achieving an increase of 1.9% on the micro-F1 comparison of the TB-Dense dataset. Although the relatively smaller data scale of the MATRES dataset and its characteristic of extremely imbalanced class distribution may limit the model's ability to fully learn the event categories, our method still effectively outperforms the existing best result by 0.5%. This not only validates that our temporal feature learning paradigm based on distance-aware modeling enables the model to learn global features with different proximity more effectively, but also indicates the impressive capabilities of the LLM we employed. We also observe significant advantages of our method compared with the two baseline models we developed, the RoBERTa-Large and BART-Large, which also confirms the superiority of our method.

3.4 Performance on Minority Categories

To confirm that our method is valid to identify the minority categories more effectively in the situation of imbalanced data, we also compare the micro-F1

Model	Language Model	TB-Dense			MATRES		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
TIMERS*(Mathur et al., 2021)	BERT-Base	48.1	65.2	67.8	81.1	84.6	82.3
SGT*(Zhang et al., 2022)	BERT-Large	-	-	67.1	-	-	80.3
RSGT*(Zhou et al., 2022)	RoBERTa-Base	68.7	68.7	68.7	82.2	85.8	84.0
Bayesian (Tan et al., 2023)	BART-Large	-	-	65.0	79.6	86.0	82.7
Unified (Huang et al., 2023)	RoBERTa-Large	-	-	68.1	-	-	82.6
TCT (Ning et al., 2024)	BART-Large	70.3	71.6	70.9	79.0	87.2	82.9
CPTRE (Yuan et al., 2024)	BERT-Base	73.4	69.5	71.4	81.3	86.3	84.2
MulCo* (Yao et al., 2024)	RoBERTa-Large	-	-	85.6	-	-	90.4
MAQInstruct (Xu et al., 2025)	Llama2-7B	-	-	-	85.5	83.9	84.7
LLMERE (Hu et al., 2025)	Llama3.1-8B	-	-	-	82.6	88.7	85.5
GDLLM_BART(Ours)	BART-Large	75.8	68.9	71.3	80.6	84.7	81.2
GDLLM_RoBERTa(Ours)	RoBERTa-Large	70.8	68.8	69.2	82.4	91.7	86.4
GDLLM_Qwen(Ours)	Qwen2.5-7B	85.3	86.5	86.1	**86.8**	94.8	90.6
GDLLM(Ours)	**Llama3.1-8B**	**88.3**	**86.6**	**87.5**	86.5	**95.9**	**90.9**

Table 1: The overall experiment results on the two datasets. Models marked with a * use the GNN-based approach. The F1 score means micro-F1.

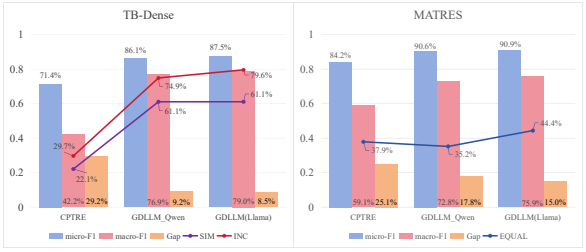

Figure 4: The performance of micro-F1, macro-F1, and the F1 score of some minority categories between our methods and the selected study. SIM: *SIMULTANEOUS*. INC: *INCLUDES*. Gap: the difference between micro-F1 and macro-F1. A lower Gap value indicates better performance of the model on minority categories.

Method	LLMs	P(%)	R(%)	F1(%)
GDLLM	Llama3.1	**86.5**	**95.9**	**90.9**
w/o LP	-	64.6	73.4	68.7
w/o GD	Llama3.1	77.2	79.0	78.1
w/o PI	Llama3.1	78.9	86.7	82.6
GDLLM	Qwen2.5	**86.8**	**94.8**	**90.6**
w/o LP	-	64.6	73.4	68.7
w/o GD	Qwen2.5	74.0	82.7	77.1
w/o PI	Qwen2.5	75.3	82.1	79.5

Table 2: Ablation study results on the MATRES. *w/o LP* only uses GAT-based multi-head attention mechanism.

score and macro-F1 score between our methods and the most recent study on a similar issue(Yuan et al., 2024), which reports that their results outperform earlier studies on the macro-F1 score. According to respective definitions, macro-F1 gives equal weight to each category, while micro-F1 gives equal weight to each sample. This ensures that if a model achieves a severe gap between micro and macro, the model cannot perform well on minority categories.

It can be seen from the data in Figure 4 that the "Gap" scores on our methods are obviously lower than those in the model CPTRE. In general, our GDLLM (Llama3.1) model outperforms CPTRE on all minority categories. Although our method GDLLM_Qwen performs suboptimally on the *EQUAL* class when using Qwen as the language model, we think that is because the *EQUAL* class has an exceptionally low count, causing the model's severely biased prediction on the categories with extremely high proportions during training. On the basis of the analysis above, our proposed model achieves significantly better performance on all datasets regarding macro-F1 scores, and it indeed improves the model's performance on minority temporal relation classes.

3.5 Ablation Study

Table 2 illustrates the ablation experimental results on the MATRES dataset(Appendix D shows

ablation results on the TB-Dense dataset). Our experiments are based on two LLMs (Llama3.1 and Qwen2.5). When analyzing the impact of removing components from the **GDLLM** method, we observe that "*w/o* ***LP***" (without **L**LM-based **P**robability Generation), "*w/o* ***GD***" (without **G**AT-based **D**istance-aware Structure), and "*w/o* ***PI***" (without **P**robabilistic Soft **I**nference Learning Paradigm) lead to a decrease in performance.

Analysis of *LP*. Through the comparison of the *w/o LP* module, the micro-F1 scores decrease by 22.2% and 21.9%, respectively. This illustrates it is challenging for GAT to identify event relations without the probabilistic information generated by LLMs, because the model has been deprived of the powerful capability to capture relation features with a short-distance proximity band.

Analysis of *GD*. Comparing the *w/o GD* module, the micro-F1 scores drop by 12.8% and 13.5% based on Llama and Qwen, respectively. This indicates the limitation of utilizing LLMs standalone for ETRE, and further demonstrates that our GAT-based distance-aware structure indeed aids the LLMs to better learn the relation features with long-distance dependency.

Analysis of *PI*. We also remove the probabilistic soft inference paradigm for temporal feature learning. That is, we make the LLMs only generate corresponding "0/1" label prediction values for edge features, transforming the entire process into a dual-stage hard classification. Comparing the *w/o PI* module, the micro-F1 scores decline by 8.3% and 11.1% on the two models. This suggests that enabling the model to learn probabilistic distribution information improves the identification of the event relation of the short-distance proximity band.

3.6 Performance on Distance Features

We also test the performance with modules *w/o GD* and *w/o PI* under different distance conditions utilizing Llama3.1. Specifically, we define the distance feature as follows: If there are n other events between the target event pair (E_i, E_j), the distance between them is set to n. As illustrated in Table 3, when the distance is progressively increased, the performance of the *w/o GD* models becomes lower than the *w/o PI* models. This indicates that our distance-aware graph structure can more effectively identify temporal relations with longer event distances. Meanwhile, when we remove the *PI* approach, the decline of micro-F1 scores becomes less pronounced as the event distance increases. Notably, when the distance increases to 5, our method only outperforms the model *w/o PI* by 0.7%. This suggests that the proposed feature learning paradigm based on soft inference can more effectively enhance the performance for events with shorter distances.

Method	Distance			
	2	3	4	5
w/o GD	79.3	80.8	75.7	81.8
w/o PI	78.1	86.3	87.8	90.2
Ours	**87.3**	**93.1**	**95.7**	**90.9**

Table 3: The comparison of micro-F1 scores(%) of subsets divided based on different distance conditions on the MATRES dataset. The data in bold and with underlines represent the optimal and suboptimal results under each distance condition, respectively.

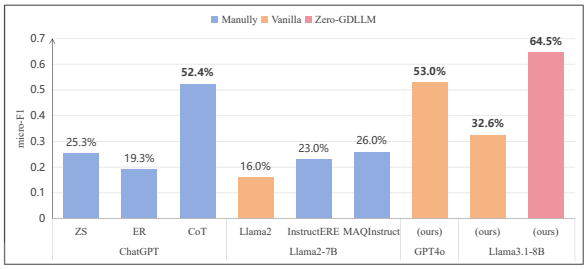

Figure 5: The micro-F1 score of the performance comparison on the MATRES dataset between our methods and other benchmarks based on zero-shot.

3.7 Analysis of Zero-Shot Experiment

As depicted in Figure 5, we conduct various experiments to compare the zero-shot performance on the MATRES dataset with different benchmarks(Yuan et al., 2023; Xu et al., 2025). **1)For *Manually* and *Vanilla***, *Manually* means giving manually designed prompts or instructions to the "*Vanilla*" LLMs which are not fine-tuned. Early work(Yuan et al., 2023) designs three kinds of prompt techniques (ZS, ER, and CoT) to evaluate ChatGPT, which gives their best performance on the CoT prompts at 52.4%. We report the result on vanilla GPT4o, which is higher than the CoT method. It suggests the importance of the scale of different LLMs and the limitation of manually designed prompts. **2)For *Zero-GDLLM***, it is to directly generate probability distribution from Llama3.1 to GAT without LoRA tuning and the GAT operates with fixed parameters. We can see our *Zero-GDLLM* method in Figure 5 outperforms all previous results above. That indi-

cates the superior capacity of our distance-aware modeling approach in zero-shot learning scenarios.

3.8 Case Study for Minority Categories

To evaluate the effectiveness of clustering minority categories, we visualize the final prediction result representations of positive samples in high-dimensional space. Specifically, we first obtain all representations on the testing set of the TB-dense dataset, which features highly imbalanced classes. Given the complex and non-linear nature of the data, we choose t-Distributed Stochastic Neighbor Embedding (t-SNE) as the dimension reduction technique to project the high-dimensional representations onto a two-dimensional space for visualization. We employ three baseline models following the ablation study.

As depicted in Figure 6(b) and Figure 6(c), the representation distribution of all positive examples has almost no obvious boundaries, which indicates the model performs poorly in clustering. Compared with Figure 6(a), it can be seen from Figure 6(d) that the t-SNE visualization of the proposed approach clearly separates and clusters minority relation classes, such as *INCLUDES* and *IS_INCLUDED*, although there is still some minor overlap between classes, the distinct clustering patterns indicate that the model effectively captures the unique characteristics of these minority categories. This demonstrates that our approach effectively augments the capacity of capturing the global relation feature. Overall, the comparison results from the t-SNE visualization strongly demonstrate the superiority of the proposed model in handling minority temporal relation classes.

4 Related Work

Earlier studies for ETRE predominantly rely on machine learning(Mani et al., 2006; Yoshikawa et al., 2009). Afterwards, some research integrates Pre-trained Language Models to capture temporal semantics in the context (Cheng et al., 2020; Wen and Ji, 2021; Mathur et al., 2021; Man et al., 2022). It is also worth noting that more and more studies focus on the special structure of event temporal relations. One of the widely employed graph-based methods is GNNs. Different GNN-based methods have been proposed to better learn the relation cues (Mathur et al., 2021; Man et al., 2022). Differently, other researchers embed events in hyperbolic spaces for better hierarchical structure

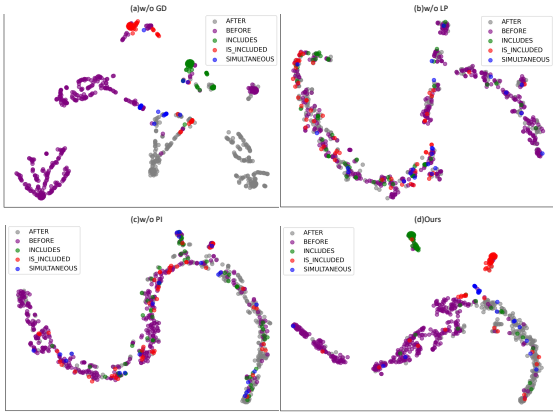

Figure 6: The visualized clustering comparison results of the ablation study based on Llama3.1-8B.

modeling(Tan et al., 2021). Prior SOTA model MulCo(Yao et al., 2024) combines GNNs and the model of BERT variants via multi-scale knowledge distillation. There are also studies that tackle data scarcity or imbalance(UzZaman et al., 2013; Wang et al., 2020; Han et al., 2020; Guan et al., 2021; Tan et al., 2023; Yuan et al., 2024), while some work designs certain temporal rules (Ballesteros et al., 2020; Zhuang et al., 2023; Ning et al., 2024).

With the rapid development of LLMs, researchers pay great attention to the Question-Answer (QA) mechanism(Xu et al., 2025; Hu et al., 2025). Similar to the zero-shot studies, another work proposes a variety of valuable prompt explanations(Yuan et al., 2023) or utilizes a unified framework(Huang et al., 2023). Appendix E reports the results comparison on GNN-based and LLM-based benchmarks.

5 Conclusion

In this paper, we propose **GDLLM**, a **G**lobal **D**istance-aware modeling approach based on **LLM**s. Specifically, we present a distance-aware graph structure utilizing GAT to assist LLMs in capturing long-distance dependency features. Additionally, we design a temporal feature learning paradigm based on soft inference to augment the event relation extraction with a short-distance proximity band. Our framework also substantially enhances the performance of minority relation classes and improves the overall learning ability. Extensive experiments on two public datasets, TB-Dense and MATRES, demonstrate that our approach outperforms all LLM-based and GNN-based benchmarks, achieving SOTA performance without manually designed prompts or instructions for LLMs.

Limitations

Although our method has already achieved the current state-of-the-art performance, the limitations may still exist. Due to the different category choices of LLMs, their inherent adaptability to task diversity or bias may pose challenges to our model training or performance. For example, on the minority class *EQUAL*, the baseline utilizing the Qwen model exhibits suboptimal performance compared to the model CPTRE. Meanwhile, future work is needed to explore more effective and diverse modeling or training methods for Large Language Models.

References

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen Mckeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Fei Cheng, Masayuki Asahara, Ichiro Kobayashi, and Sadao Kurohashi. 2020. Dynamically updating event representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357.

Hong Guan, Jianfu Li, Hua Xu, and Murthy Devarakonda. 2021. Robustly pre-trained neural model for direct temporal relation extraction. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 501–502. IEEE.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11058–11066.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chungmin Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. Timers: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. 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)*, pages 6203–6209.

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, and Jingyao Tang. 2024. Temporal cognitive tree: A hierarchical modeling approach for event temporal relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 855–864.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second joint conference on lexical and computational semantics (* SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 1–9.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437.

Jun Xu, Mengshu Sun, Zhiqiang Zhang, and Jun Zhou. 2025. Maqinstruct: Instruction-based unified event relation extraction. *arXiv preprint arXiv:2502.03954*.

Hao-Ren Yao, Luke Breitfeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rose. 2024. Distilling multi-scale knowledge for event temporal relation extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2971–2980.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 405–413.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2024. Temporal relation extraction with contrastive prototypical sampling. *Knowledge-Based Systems*, 286:111410.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. Rsgt: relational structure guided temporal relation extraction. In *Proceedings of the 29th international conference on computational linguistics*, pages 2001–2010.

Ling Zhuang, Hao Fei, and Po Hu. 2023. Knowledge-enhanced event relation extraction via event ontology prompt. *Information Fusion*, 100:101919.

A Input Formats Details

In the experiment, we make the input format of our LLMs as following arrangements :

A)Llama: In the given sentence T, manually designed symbols of the form $[EV_i]$, where i denotes the ordinal number of the event in a pair, serve as explicit in-text markers to annotate the boundaries of the event to facilitate model focus.

B)Qwen: The internal structure of Qwen determines that the model has a greater inclination towards a dialogue-based model. Our inputs to the Qwen model are carefully structured as "Input = T", and the processing of symbolic marking for the events in sentence T is the same as that of the Llama model.

C)Zero-Shot: For zero-shot scenarios, we utilize hand-crafted prompts (e.g., "I will give you a paragraph that uses $[EV1]$, $[/EV1]$, $[EV2]$ and $[/EV2]$ to, respectively mark two events, with the event relations divided into '*BEFORE*', '*AFTER*', '*VAGUE*' and '*EQUAL*'. You only need to provide the final judgment result of the event relation") without task-specific training.

B Experiment Details

Our experiment details are reported as follows:

A)Datasets: Data splits statistics are reported in Table 4.

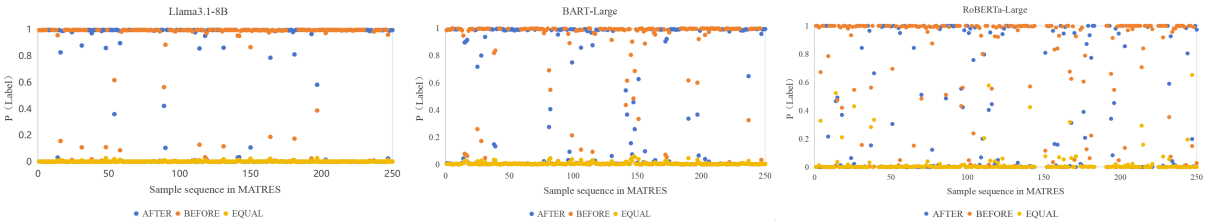

Figure 7: The distribution of probabilities generated by different language models on the MATRES dataset.

Dataset	Train:Validation:Test
TB-Dense	4,032:629:1,427
MATRES	182:73:20

Table 4: Data splits and relation statistics.

B)HEBO Algorithm for Hyperparameter Optimization: HEBO is a Bayesian optimization-based algorithm designed to efficiently search for optimal hyperparameter combinations in a high-dimensional space. The algorithm details are as follows:

Suppose x be a vector of hyperparameters, and $y = f(x)$ be the objective function, which is the evaluation metric of the model on the validation set. We utilize the Gaussian Process surrogate model of HEBO, which is $\hat{f}(x)$ that has a mean function $\mu(x)$ and a variance function $\sigma^2(x)$, such that $\hat{f}(x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$. The acquisition function, such as expected value $\mathbb{E}$, is defined as:

$$a(x) = \mathbb{E}[\max(0, f(x^*) - \hat{f}(x))], \quad (9)$$

where x^* is the optimal hyperparameter point currently. The next hyperparameter point x_{next} to evaluate is selected by maximizing the acquisition function:

$$x_{next} = \arg\max_x a(x). \quad (10)$$

Specifically, we use the HEBOSearch implementation. The hyperparameter search space includes parameters such as the dropout rates, class weights, and the learning rate. We initialize the search process with a set of randomly sampled hyperparameter points. For each iteration, the HEBO algorithm calculates the acquisition function values for all points in the search space based on the current surrogate model. The hyperparameter point with the maximum acquisition function value is then selected and evaluated on the model. After obtaining the evaluation result, the surrogate model is updated to incorporate this new information. Compared to traditional hyperparameter optimization methods such as random search and grid search, HEBO can more efficiently explore the hyperparameter space by leveraging the information from previously evaluated points.

Method	LLMs	P(%)	R(%)	F1(%)
GDLLM	Llama3.1	**88.3**	**86.6**	**87.5**
w/o LP	-	47.3	69.1	53.2
w/o GD	Llama3.1	67.8	58.1	62.5
w/o PI	Llama3.1	62.4	72.6	66.0
GDLLM	Qwen2.5	**85.3**	**86.5**	**86.1**
w/o LP	-	47.3	69.1	53.2
w/o GD	Qwen2.5	68.0	72.7	70.8
w/o PI	Qwen2.5	63.6	71.5	66.0

Table 5: The ablation experimental results on the TB-Dense dataset. "*w/o LP*" only adopts multi-head attention for ETRE.

C Distribution of Generated Probabilities on the MATRES Dataset.

Figure 7 depicts the distribution of probabilities generated by different language models on the MATRES dataset.

D The Ablation Experimental Results on the TB-Dense Dataset.

As shown in Table 5, the ablation experimental results on the TB-Dense dataset also reveal the importance of different components.

E The Performance Comparison on GNN-based and LLM-based Benchmarks

We analyze the performance of GNN-based methods with different benchmarks, which is depicted in Figure 8. The existing SOTA model MulCo(Yao et al., 2024) contributes various GNN-based results.

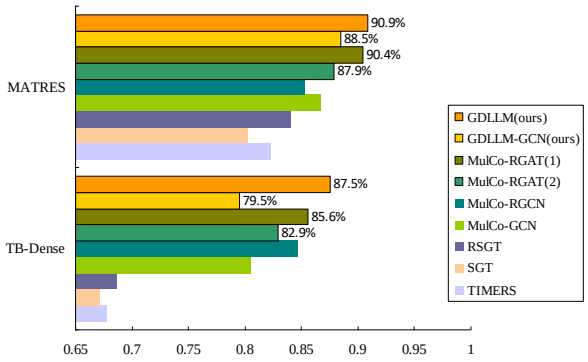

Figure 8: The micro-F1 score of the previous GNN-based method versus our approach. "MulCo-RGAT(n)" represents the model adopts n GNN layers.

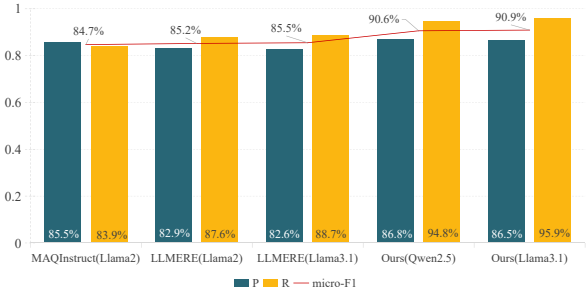

Figure 9: The performance comparison on the MATRES dataset between our method and other benchmarks based on LLMs.

Our method, based on two layers of GAT, outperforms MulCo-RGAT(2), highlighting the effectiveness of our **GDLLM** proposed in the GNN-based approaches. We also test the performance of the GCN-based method, the results suggest that GCN lacks the capacity of multi-head attention, which fails to effectively learn the probabilistic relation features for the short-distance proximity band.

As shown in Table 9, our method reports the experimental results based on Llama3.1 and Qwen2.5. Our results outperform all other LLM-based benchmarks(Xu et al., 2025; Hu et al., 2025) on the MATRES dataset.