DeTri: Debiasing General-Purpose LLMs for Zero-Shot Relation Triplet Extraction via Structural Expert

Zehan Li, **Fu Zhang**, **Jiawei Li**, **Wenqing Zhang** and **Jingwei Cheng**
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
lizehan1999@163.com, zhangfu@neu.edu.cn,

Abstract

Zero-Shot Relation Triplet Extraction (ZSRTE) aims to extract relation triplets for unseen relation types without any annotated training data. Recent advancements in Large Language Models (LLMs) have significantly enhanced ZSRTE performance, enabling the direct generation of relational triplets from unstructured text. However, LLMs often introduce biases such as entity shift, relation confusion, and over-prediction, which limit the reliability of the extracted triplets.

In this paper, we introduce DeTri, a novel debiasing framework for ZSRTE that addresses these biases by leveraging a discriminative structural expert model. DeTri treats LLMs as inspiration generators and refines their outputs via a three-stage debiasing pipeline, consisting of inspiration-based re-prediction, confidence-based filtering, and entity shift correction through span perturbation. The framework improves LLM-generated triplets without requiring fine-tuning the LLM, thus offering a transferable and scalable solution. Extensive experiments on two datasets show that DeTri outperforms state-of-the-art methods by reducing bias and improving extraction accuracy, achieving 2.89% F1 improvement over fine-tuned LLM methods while keeping the LLM parameters frozen. Our approach shows strong generalization across different LLMs, offering a solution to mitigate biases.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved impressive progress in Relation Triplet Extraction (RTE) tasks [Xu *et al.*, 2024b]. Using instruction-tuning [Sainz *et al.*, 2023] or prompt-learning [Ma *et al.*, 2023], LLMs can directly generate `subject-relation-object` triplets from raw text without the need for annotated data on unseen relations [Chia *et al.*, 2022], showcasing strong Zero-Shot RTE (ZSRTE) capabilities. Although LLMs benefit from the vast background knowledge encoded during pre-training [Dagdelen *et al.*, 2024], they also introduce potential biases [Zhang *et al.*, 2024; Zhou *et al.*, 2024], which limit their practical applications. As illustrated in Figure 1, we analyze three common types of bias in the triplets generated by LLMs compared to the ground truth.

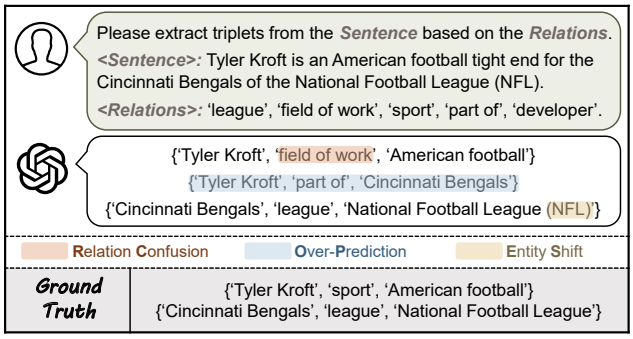

Figure 1: Three common types of bias introduced by general-purpose LLMs in Zero-Shot RTE.

Relation Confusion. LLMs may confuse similar relations, leading to classification errors. This bias directly impacts the reliability of the results.

Over-Prediction. LLMs tend to generate seemingly plausible but weakly connected triplets. Hallucinatory content can negatively impact the performance of downstream tasks.

Entity Shift. The entities generated by general-purpose LLMs are often still semantically relevant to the ground truth but exhibit slight deviations.

These biases stem from the inherent nature of LLMs [Li *et al.*, 2025a], which generate triplets based on linguistic patterns learned from large-scale data without explicit logical validation or semantic constraints. Strategies such as confidence estimation [Zhang *et al.*, 2025a] or self-consistency [Li *et al.*, 2023] can partially mitigate these issues, but their characteristics limit their application on closed-weight models, such as those in the GPT series [Brown *et al.*, 2020]. Recent advances in ZSRTE mainly leverage transfer learning, which includes two primary methods: in-context learning [Shi *et al.*, 2024] and instruction-tuning [Li *et al.*, 2024; Zhang *et al.*, 2025b]. While the former is constrained by lim-

* Corresponding author.

ited context length and prompt sensitivity [Guo *et al.*, 2025], the latter suffers from high retraining costs and poor adaptability to evolving LLMs [Wu *et al.*, 2024].

Given the limitations of current methods, a natural question arises: **How can we fully harness the relational semantic generalization capabilities of *general-purpose LLMs* without retraining?** This question has led to the exploration of novel strategies that combine the generative power of LLMs with the structural precision of other models. In this paper, we propose **DeTri**, a novel framework that aims to **De**bias LLM-generated **Tri**plets through a structural expert model. Unlike traditional methods that directly fine-tune the LLMs, DeTri treats the frozen LLM as an inspiration generator, serving as a semantically rich but potentially noisy generator for candidate triplets that are then refined by a discriminative structural expert model.

Specifically, we introduce a lightweight expert model as a medium for knowledge transfer. During training, we construct bias-enhanced data using representative biased negative samples from LLMs. By incorporating prefix sequences enriched with candidate relation semantics, the expert model is guided to reassemble potential entity pairs inspired by LLMs and predict their compatibility with candidate relations. This approach mitigates issues such as relation confusion and prediction redundancy. In addition to improving the model's ability to detect and correct entity shifts, we introduce a span perturbation mechanism that generates combinations of neighboring entities by perturbing the entity's boundaries, providing additional contrastive signals that strengthen the model's debiasing capability during training. Our framework effectively reduces LLM-induced biases while avoiding resource-intensive and repetitive instruction fine-tuning. Our main contributions are as follows:

- We propose a debiasing framework named DeTri for common bias types in ZSRTE that takes the extraction results of LLMs as inspiration. This framework is model-agnostic and continuously benefits from the evolving capabilities of general-purpose LLMs.
- We introduce a structural expert model for re-prediction and filtering, along with a span perturbation augmentation mechanism that enhances the model's ability to detect and correct entity shifts.
- Extensive experiments demonstrate that DeTri outperforms state-of-the-art approaches on two widely used datasets, reduces the bias introduced by general-purpose LLMs, and achieves a 2.89 absolute F1 improvement over fine-tuned LLM methods while keeping the LLM parameters frozen.

2 Related Work

2.1 Transfer Learning for Zero-shot RTE

Relation Triplet Extraction (RTE) aims to identify triplets from unstructured text [Li *et al.*, 2025b]. In the zero-shot setting, target relation types are unseen during training [Li *et al.*, 2025c], making zero-shot RTE a typical knowledge transfer problem [Zhuang *et al.*, 2020]. Transfer learning [Zhang *et al.*, 2019] plays a key role in improving generalization to unseen relations [Huang *et al.*, 2018] by leveraging knowledge from seen relations.

Generative LLMs have demonstrated strong performance in extraction tasks due to their deep semantic understanding [Dagdelen *et al.*, 2024]. Some methods fine-tune LLMs using instruction tuning [Li *et al.*, 2024; Sun *et al.*, 2024] or multitask learning [Zhang *et al.*, 2025b] to transfer knowledge from seen data, though these are resource-intensive. Alternatively, prompt engineering techniques guide LLMs in extracting triplets without requiring model updates [Wei *et al.*, 2024], but these methods are highly sensitive to prompt design. Discriminative methods [Gong and Eldardiry, 2024; Lan *et al.*, 2024] have shown great potential in ZSRTE, largely due to the structured nature of the task, which benefits from discriminative modeling. However, these methods exhibit limited generalization to unseen relations, especially when there is a lack of sufficient annotated data. Motivated by this, our approach for the first time combines the strengths of both paradigms, optimizing the extraction results of LLMs through a discriminative expert model for debiasing.

2.2 Debiasing Methods for LLMs

Due to potential biases and inaccuracies in pretraining data, LLMs are prone to hallucinations in tasks such as information extraction [McKenna *et al.*, 2023]. To address this issue, some studies integrate external knowledge during fine-tuning [Ma *et al.*, 2024]. Zhu et al. [Zhu *et al.*, 2025] adopt a two-stage instruction tuning strategy to enhance LLMs' debiasing capabilities. Xu et al. [Xu *et al.*, 2024a] constrain LLM outputs by injecting structured examples from supervised models as in-context demonstrations.

To avoid costly parameter updates, other works propose structured prompting frameworks that mitigate bias without fine-tuning [Furniturewala *et al.*, 2024]. In the context of information extraction, over-generation and positional bias pose further challenges. Zhang et al. [Zhang *et al.*, 2024] train a Flan-T5 model to judge hallucinated outputs, reducing entity errors in event argument extraction. Meanwhile, to address uncertainty-induced bias in LLMs, Li et al. [Li *et al.*, 2023] apply self-consistency decoding. Despite these efforts, generative models remain inherently limited in reliability.

Some research focuses on evaluating ambiguous extraction results [Jiang *et al.*, 2024], but such approaches may introduce risks for downstream applications. Therefore, we propose a complementary paradigm that treats LLMs as inspiration generators rather than final predictors. By coupling their generative flexibility with the structural precision of lightweight discriminative models, we achieve more robust and controllable extraction in ZSRTE.

3 Methodology

3.1 Task Formulation

Problem Statement 1 (Relation Triplet Extraction, RTE). Given a dataset $D = ((X_i, \Gamma_i)_{i=1}^{|D|}, R)$, where X_i denotes the i-th input sentence and Γ_i represents the corresponding set of relation triplets, formulated as $\Gamma_i = \{s_j, r_j, o_j\}_{j=1}^{|\Gamma_i|}$. Here, $s_j, o_j \in E$, where E is the set of all entities in X_i, and

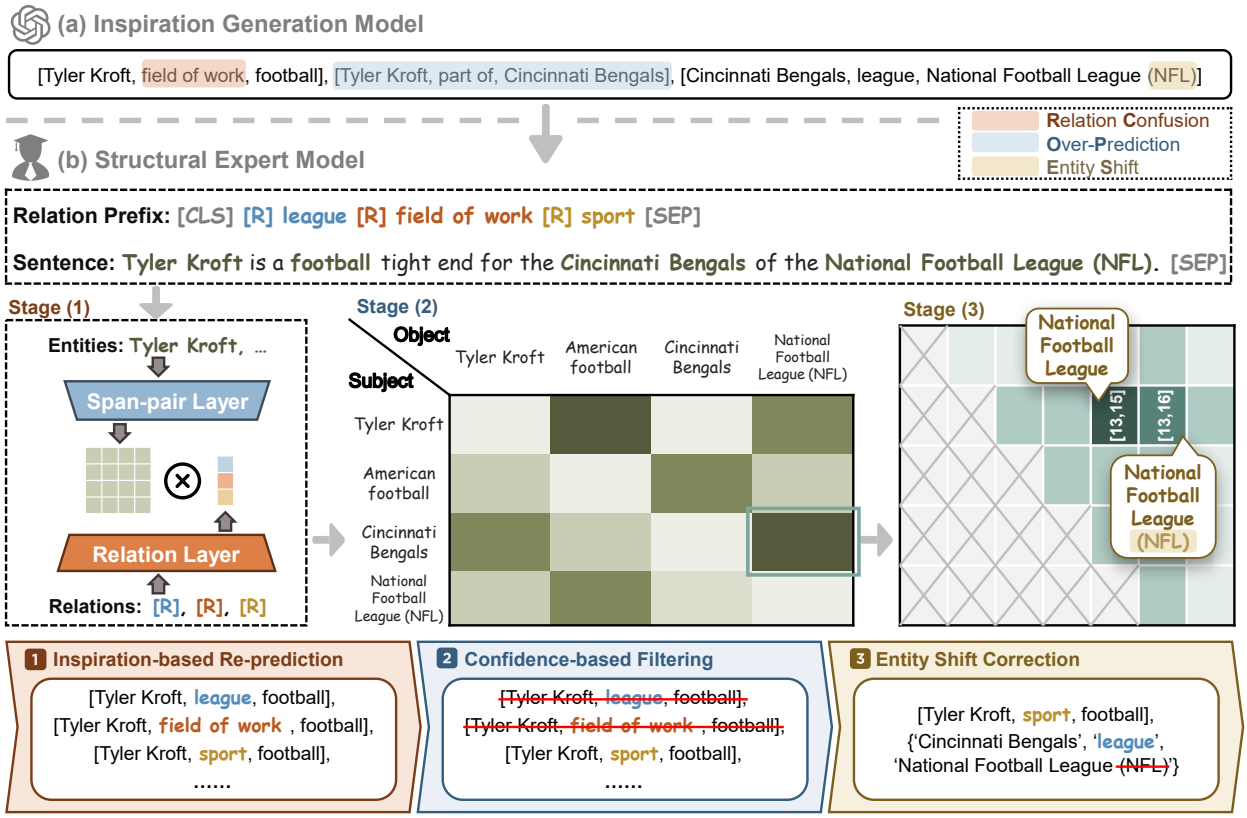

Figure 2: An illustration of the inference phase in the proposed DETRI framework. **(a)** The LLM first acts as an inspiration generator, producing candidate triplets with potential biases. **(b)** These outputs are then refined by the structural expert model through a three-stage debiasing process: *(1) Inspiration-based Re-prediction*, which reconstructs entity pairs under candidate relations; *(2) Confidence-based Filtering*, which retains semantically reliable results; *(3) Entity Shift Correction*, which addresses entity shifts by expanding the span boundary neighborhood.

$r_j \in R$, where R denotes the predefined set of relation labels. The goal of RTE is to extract all valid triplets.

Problem Statement 2 (Zero-Shot RTE, ZSRTE). Under the zero-shot setting, the relation set R is split into disjoint seen (R_{seen}) and unseen (R_{unseen}) subsets. The model learns from triplets in the seen subset D_{seen} and transfers knowledge to handle unseen relations in D_{unseen}. Only D_{seen} is accessible during training, emphasizing transfer learning over exhaustive relation coverage.

3.2 Framework Overview

DETRI leverages general-purpose LLMs as an **Inspiration Generator** to produce candidate triplets potentially affected by structural or semantic biases and introduces a **Structural Expert Model** to refine them. First, in the sample construction stage (§3.3), biased negative samples are generated using LLMs to build an augmented training set. Then, a discriminative structural expert model (§3.4) is trained to capture fine-grained relational semantics through interactions between entity spans and relation prefixes. Finally, at the inference phase (§3.5), the model corrects LLM-generated triplets via a three-stage pipeline as shown in Figure 2.

3.3 Biased Negative Samples Construction

To enhance the generalization and debiasing of the expert model, we simulate typical LLM-induced biases by constructing structurally biased hard negative samples. These samples imitate the coarse semantics and fine-grained confusion commonly exhibited by LLMs, thus providing a challenging signal to encourage discriminative training.

Concretely, we employ `GPT-4o-mini` as the generation engine due to its low inference cost and stable generation quality. Given D_{seen}, the LLM is prompted to generate triplets $\widehat{\Gamma}_i$ without being restricted to the predefined relations. This allows the LLM to output triplets freely, simulating real-world ambiguity. The entities involved in the generated triplets are referred to as *potential entities*. We then combine Γ_i with the biased samples to build the training set:

$$\widetilde{D}_{\text{seen}} = \left\{ \left(X_i, \Gamma_i \cup \widehat{\Gamma}_i \right) \right\}_{i=1}^{|D_{\text{seen}}|}. \tag{1}$$

We also discuss scenarios that do not rely on LLM-generated negative samples in Section 5.5, demonstrating the flexibility in constrained environments.

3.4 Structural Expert Model

The expert model aims to learn debiasing capabilities from the structure-bias-enhanced $\widetilde{D}_{\text{seen}}$.

Input Encoding. To enhance relational awareness, we prepend a relation-aware prefix sequence P to each input. This sequence consists of learnable markers $[\mathcal{R}]$ interleaved with candidate relations:

$$P = \{[\mathcal{R}], r_1, [\mathcal{R}], r_2, \ldots, [\mathcal{R}], r_n\}. \tag{2}$$

The final input sequence $X' = P \oplus X$, where $X = w_1, w_2, \ldots, w_l$ is the original sentence and $\oplus$ denotes concatenation. We then encode X' using a bidirectional Transformer encoder to obtain contextual representations $\mathbf{H}$.

Span-Pair Representation. We represent all possible entity spans as basic semantic units. For every span from token i to j with a maximum span length K, we compute:

$$E = \left\{ e^{(i,j)} = \text{FFN}_e([\mathbf{h}_i; \mathbf{h}_j]) \;\middle|\; j - i < K \right\}, \tag{3}$$

where $\mathbf{h}$ is its token representation in $\mathbf{H}$, $[;]$ denotes the embedding concatenation operation, and FFN is a feed-forward network. We then construct the span-pair matrix $\boldsymbol{\mathcal{T}}$ by combining all entity spans:

$$\boldsymbol{\mathcal{T}}(E) = \left\{ t^{(p,q)} = \text{FFN}_t\left([e_s^p; e_o^q]\right) \mid e_s, e_o \in E \right\}. \tag{4}$$

For all candidate relations r_k, we match their representations $h_{[\mathcal{R}],c}$ with all span-pairs t to evaluate whether the combination aligns with the relational semantics:

$$\mathcal{P}_{\boldsymbol{\mathcal{T}}(E)} = \Big\|_{c=1}^{n} \left(\sum_{d} t^{(i,j)} \times \text{FFN}_{\text{rel}}(h_{[\mathcal{R}],c}) \right), t^{(i,j)} \in \boldsymbol{\mathcal{T}}(E). \tag{5}$$

Span perturbation Augmentation. To improve entity shift correction, we define a span neighborhood $\mathcal{N}(\cdot)$ through local perturbations. Specifically, for any span $e_{i,j}$:

$$\begin{aligned} \mathcal{N}(e_{i,j}) = &\{e_{i-\delta,j},\ e_{i,j+\delta} \mid 1 \le \delta \le k\} \\ &\cup \{e_{i',j'} \mid i \le i' \le j' \le j\} \cup \{e_{i,j}\}, \end{aligned} \tag{6}$$

where k is the maximum extension step. Based on the above perturbation set, we construct the span-pair matrix $\boldsymbol{\mathcal{T}}(\mathcal{B})$, where $\mathcal{B}$:

$$\mathcal{B} = \bigcup_{(s,o)\in\boldsymbol{\mathcal{T}}} \left(\left\{ (s', o) \mid s' \in \mathcal{N}(s) \right\} \cup \left\{ (s, o') \mid o' \in \mathcal{N}(o) \right\} \right). \tag{7}$$

By training on these span-perturbed structures, the expert model learns to recover correct span boundaries and mitigate structural biases.

3.5 Training and Inference

Training. We reformulate the original extraction task into a discrimination-and-correction task for the expert model trained on $\widetilde{D}_{\text{seen}}$, which consists of structurally biased samples. The training objective is to enable the model to accurately distinguish relation types while adjusting misaligned entity boundaries. The loss function is defined as:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}\Big(\mathcal{P}_{\boldsymbol{\mathcal{T}}(E)}\Big) + (1-\alpha) \cdot \mathcal{L}\Big(\mathcal{P}_{\boldsymbol{\mathcal{T}}(\mathcal{B})}\Big), \tag{8}$$

where $\mathcal{L}(\cdot)$ is the binary cross-entropy loss:

$$\mathcal{L}(\mathcal{P}) = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log(p_i) + (1-y_i) \log(1-p_i) \Big], \tag{9}$$

where $p_i \in \mathcal{P}$, $y_i \in \{0, 1\}$ indicates whether the i-th span-pair matches the given relation, and $\alpha \in [0, 1]$ controls the weight of the losses. Within each batch, we randomly select negative relations for potential entity pairs to construct contrastive samples, enhancing the model's ability to discriminate between semantically similar relations.

Inference. During inference, DETRI improves RTE performance by integrating an LLM with a structural expert model. As shown in Figure 2, the LLM first serves as an inspiration generator, producing candidate triplets from the input sentence. To examine the debiasing capability of DETRI for general-purpose LLMs, we adopt a simple RTE prompt enriched with candidate relation types to generate relation-related initial inspirations:

Relation Triplet Extraction Prompt

```
Given a sentence and candidate
relations, extract all triplets and
output them in the following JSON
format: {Triplet: [{Subject:'',
Relation:'', Object:''}]}
<Sentence>: ...
<Candidate Relations>: ...
```

While the generated triplets are often semantically reasonable, they typically suffer from three structural biases. To mitigate these, the expert model performs multi-stage refinement: (1) **Inspiration-based Re-prediction**: Rebuild span pairs based on LLM outputs and re-predict relation types. (2) **Confidence-based Filtering**: Eliminate low-confidence or redundant triplets based on a threshold τ. (3) **Entity Shift Correction**: Search boundary neighborhoods $\mathcal{N}$ to adjust entity spans and recover locally optimal boundaries. This collaborative pipeline combines the LLM's strong semantic reasoning with the structural expert's precision, thereby significantly improving ZSRTE performance.

4 Experiments Setup

4.1 Datasets

We evaluate DETRI on two widely used ZSRTE benchmarks: FewRel [Han *et al.*, 2018] and Wiki-ZSL [Chen and Li, 2021]. FewRel is a human-annotated dataset with 80 relations and 56,000 instances. Wiki-ZSL is a larger, distantly supervised dataset built from Wikipedia, containing 113 relations and 94,383 instances. Although more realistic, it introduces label noise that may affect model robustness. We adopt the split strategy of Chia et al. [Chia *et al.*, 2022], ensuring disjoint relation labels across training, validation, and test sets. Unseen validation relations are used for early stopping and tuning, while the test set contains $m \in 5, 10, 15$ unseen relations to evaluate generalization.

Methods	FewRel			Wiki-ZSL			
	m=5	m=10	m=15	m=5	m=10	m=15	*Avg.*
TableSequence [Wang and Lu, 2020]	11.82	12.54	11.65	14.47	9.61	9.20	11.55
RelationPrompt [Chia *et al.*, 2022]	22.27	23.18	18.97	16.64	16.48	16.16	18.95
ZETT [Kim *et al.*, 2023]	30.71	27.90	26.17	21.49	17.27	12.78	22.72
TAG [Xu *et al.*, 2024c]	28.94	28.16	22.53	23.12	17.24	16.41	22.73
MICRE [Li *et al.*, 2024]	<u>37.53</u>	<u>34.77</u>	<u>32.42</u>	27.74	24.64	22.23	29.89
Vanilla Prompt -GPT-4o-mini	12.37	7.88	8.68	5.11	5.05	4.77	7.31
In-Context Learning -GPT-4o-mini	21.33	18.67	15.33	16.22	12.12	13.10	16.13
ChatIE -GPT-4o-mini [Wei *et al.*, 2024]	22.73	23.23	15.66	8.59	8.08	11.62	14.99
AgentRE -GPT-4o-mini [Shi *et al.*, 2024]	18.18	18.69	16.16	14.14	14.65	18.37	16.70
DeTri -GPT-4o-mini	36.68	34.53	31.76	<u>32.40</u>	<u>27.80</u>	<u>25.03</u>	<u>31.37</u>
DeTri -Llama-4-Scout	**37.96**	**35.56**	**32.52**	**34.12**	**29.77**	**26.77**	**32.78**

Table 1: Main results on FewRel and Wiki-ZSL under the single-triplet setting. m represents the number of unseen relations in the candidate set. **Bold** marks the highest score. <u>Underline</u> marks the second-best score.

4.2 Implementation Details

We use `GPT-4o-mini` to generate inspiration triplets and adopt `DeBERTa-v3-small` [He *et al.*, 2023] as the backbone of our structural expert model. The model is trained using AdamW with learning rates of $1e{-}5$ for the encoder and $5e{-}4$ for other components. Training is conducted with a batch size of 16 for up to 10 epochs. The maximum extension step k for triplet debiasing is set to 2 to effectively correct entity boundary shifts while limiting the number of candidate spans, thereby reducing noise and computational overhead. Hyperparameters are tuned via grid search, with optimal values of $\alpha = 0.3$ and $\tau = 0.4$.

For evaluation, we follow prior ZSRTE protocols. In the **Single Triplet** setting (one triplet per sentence), we report accuracy (*Acc.*), while in the **Multi Triplet** setting (two or more triplets per sentence), we report micro precision ($P.$), recall ($R.$), and F_1 score. All metrics are averaged over five random splits to ensure robustness. To control experimental costs, we randomly sample 1,000 test instances under each setting for evaluation.

4.3 Compared Baselines

We compare DeTri with two categories of ZSRTE methods.

Fine-tuning-based Methods. **TableSequence** [Wang and Lu, 2020] jointly extracts triplets using a table-sequence encoder. **RelationPrompt** [Chia *et al.*, 2022] trains GPT-2 on synthetic data generated for unseen relations via prompt learning. **TAG** [Xu *et al.*, 2024c] introduces a two-agent reinforcement learning framework to improve synthetic data quality. **ZETT** [Kim *et al.*, 2023] employs T5 with relation templates to directly generate entities. **MICRE** [Li *et al.*, 2024] fine-tunes LLMs like Llama [Touvron *et al.*, 2023] on 12 additional datasets using tabular prompt.

Frozen LLM-based Methods. **Vanilla Prompt** uses a single-turn instruction with candidate relations. **In-Context Learning** extends this by incorporating three demonstration examples. **ChatIE** [Wei *et al.*, 2024] builds a multi-turn, two-stage chain-of-thought prompting framework. **AgentRE** [Shi *et al.*, 2024] is an agent-based method that retrieves relevant knowledge, and we adopt its zero-shot setting.

Method	m=5	m=10	m=15
DeTri	**36.68**	**34.53**	**31.76**
- w/o Re-prediction	28.59	26.49	24.31
- w/o Span Perturbation	30.81	28.21	27.16
- w/o Expert Model	12.37	7.88	8.68

Table 2: Ablation study of DeTri on FewRel. We report single-triplet accuracy under m = 5, 10, 15.

DeTri does not require fine-tuning the LLM. Instead, it augments the frozen LLM with an expert model to filter and revise generated triplets. For fair comparison, we use the same backbone model across all methods and either cite official results or reproduce them using publicly available code.

5 Analysis and Discussion

5.1 Main Results

We evaluate DeTri under the single-triplet setting on two benchmarks: FewRel and Wiki-ZSL. As shown in Table 1, the main findings are as follows:

First, with the same frozen LLM backbone (`GPT-4o-mini`), DeTri significantly outperforms existing baselines. Compared to CoT reasoning (ChatIE) and agent prompting (AgentRE), it achieves **16.38%** and **14.67%** higher average accuracy, respectively. This demonstrates the persistent reasoning bias in LLMs, which DeTri successfully mitigates using a structural expert model, without updating any LLM parameters, leading to notable improvements in ZSRTE performance.

Second, DeTri outperforms even the fine-tuned state-of-the-art method MICRE by an average of **2.89%**, without requiring any fine-tuning on LLMs. This underscores the strong generalization ability and efficiency of DeTri achieved through purely external debiasing.

Third, DeTri performs robustly on Wiki-ZSL, confirming its effectiveness on noisy samples. Moreover, preliminary results suggest continued performance gains as the base LLM improves (e.g., switching from `GPT-4o-mini`

Methods	m=5			m=10			m=15		
	$P.$	$R.$	F_1	$P.$	$R.$	F_1	$P.$	$R.$	F_1
RelationPrompt [Chia *et al.*, 2022]	20.80	24.32	22.34	21.59	28.68	24.61	17.73	23.20	20.08
ZETT [Kim *et al.*, 2023]	38.14	30.58	33.71	30.65	32.44	31.28	22.50	27.09	24.39
RSED [Lan *et al.*, 2024]	43.91	34.97	38.93	30.89	29.90	30.39	27.00	23.55	25.16
ZS-SKA [Gong and Eldardiry, 2024]	57.50	26.24	36.04	60.48	23.22	33.28	37.29	19.13	25.29
TAG [Xu *et al.*, 2024c]	37.56	40.24	38.81	31.04	33.49	32.18	25.35	25.88	25.59
$\text{FT}_{\text{Llama2-13B-Chat}}$ [Zhang *et al.*, 2025b]	50.77	52.86	51.79	34.98	32.45	33.67	31.28	24.87	27.71
$\text{FT}_{\text{Qwen1.5-14B-Chat}}$ [Zhang *et al.*, 2025b]	53.10	52.66	**52.88**	35.46	32.17	33.74	30.28	25.03	27.41
$\textbf{DETRI}_{\text{GPT-4o-mini}}$	43.40	46.16	44.58	39.35	36.65	**37.62**	32.20	31.88	**31.81**

Table 3: Performance on the multi-triplet setting of FewRel. We compare seven competitive approaches. Four are generative methods, including **FT**, which represents LLMs fine-tuned on seen data using Low-Rank Adaptation (LoRA). The remaining two are discriminative methods, namely **RSED** and **ZS-SKA**. For fair comparison, we report results either directly from the original papers. **Bold** indicates the best score under each setting.

to `Llama-4-Scout`), demonstrating strong scalability and transferability.

5.2 Ablation Experiment

To evaluate the contribution of each component in DETRI, we conduct ablation studies, with results shown in Table 2.

Removing **Re-prediction**, which skips entity recombination and filters initial triplets based solely on confidence, leads to a clear performance drop. This demonstrates that restructuring entities to enrich the triplet space plays a critical role in correcting errors generated by LLMs.

Excluding **Span Perturbation**, i.e., removing the perturbation loss $\mathcal{L}(\mathcal{P}_{\mathcal{T}(\mathcal{B})})$ during training and omitting the entity shift correction, results in noticeable accuracy degradation, especially for larger m. This confirms the mechanism's role in enhancing robustness to entity boundary shifts and correcting plausible but inaccurate predictions.

Eliminating the **Expert Model** and relying solely on LLM outputs causes a sharp drop to 10% accuracy, indicating LLMs struggle with generalization in ZSRTE without additional structural and semantic guidance.

5.3 Performance on Multi Triplets

We evaluate DETRI on multi-triplet setting on FewRel under $m = 5, 10, 15$, as shown in Table 3. When $m = 5$, Fine-Tuned (FT) LLMs perform best, demonstrating their effectiveness in scenarios with limited candidate relations. However, as m increases to 10 and 15, all methods show performance degradation. Notably, DETRI consistently achieves the highest F_1 scores, outperforming fine-tuned LLMs by **3.88%** and **4.40%**, respectively. These results indicate that generative methods are more prone to bias in complex settings, whereas DETRI's discriminative structure offers better robustness by effectively mitigating such biases.

5.4 Performance on Different LLMs

To evaluate the generalizability of DETRI, we assess its performance across five widely used general-purpose large language models: `GPT-4o-mini`[1], `GPT-4.1-mini`[1], `DeepSeek-Chat`[2], `Llama-3.1-40B`[3], and `Llama-4-Scout`[3]. As shown in Figure 3, we compare the accuracy of each LLM under two settings: the basic vanilla prompt and our proposed DETRI framework.

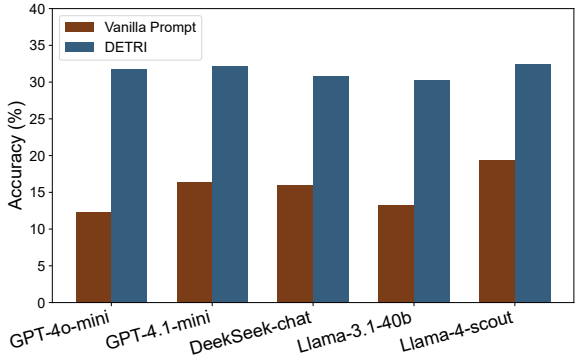

Figure 3: Accuracy (%) comparison between Vanilla Prompting and DETRI across five LLMs, evaluated on FewRel under the single-triplet setting with $m = 15$.

Under the vanilla prompting baseline, all models perform poorly, with none exceeding 20% accuracy. `Llama-4-Scout` achieves the highest at 19.43%, revealing the limitations of direct LLM generation in ZSRTE due to biases such as entity shift and relation confusion. In contrast, when combined with DETRI, all LLMs show significant improvements.

Importantly, as LLMs continue to improve, for example with `Llama-4-Scout` outperforming `Llama-3.1`, our framework consistently yields greater performance gains. DETRI increases their accuracies to 32.52% and 30.24%, respectively, demonstrating strong compatibility with more advanced models.

These results validate that DETRI can effectively reduce structural and semantic biases across various LLMs without additional fine-tuning. Its plug-and-play model-agnostic design ensures continued effectiveness as LLMs evolve, making it a scalable and efficient solution for ZSRTE.

[1]https://openai.com/api/
[2]https://api-docs.deepseek.com/
[3]https://www.llama.com/products/llama-api/

`Case 1:` **Sentence**	At the 2011 World Amateur Boxing Championships he beat Artur Beterbiev and Teymur Mammadov to win the Heavyweight title.
Ground Truth	{Artur Beterbiev, *competition class*, Heavyweight}, {Teymur Mammadov, *competition class*, Heavyweight}
LLM Prediction	{he, *competition class*, Heavyweight title}, {Artur Beterbiev, *competition class*, Heavyweight title}
Relation Debiasing	{Artur Beterbiev, *competition class*, Heavyweight title}
Entity Debiasing	{Artur Beterbiev, *competition class*, Heavyweight}
`Case 2:` **Sentence**	A building that also can be considered an example of a blob is Peter Cook and Colin Fournier's Kunsthaus (2003) in Graz, Austria.
Ground Truth	{Kunsthaus, *architect*, Peter Cook}, {Kunsthaus, *architect*, Colin Fournier}
LLM Prediction	{Kunsthaus, *located in or next to body of water*, Graz}, {Kunsthaus, *architect*, Peter Cook and Colin Fournier}
Relation Debiasing	{Kunsthaus, *architect*, Peter Cook and Colin Fournier}
Entity Debiasing	{Kunsthaus, *architect*, Peter Cook}, {Kunsthaus, *architect*, Colin Fournier}

Table 4: Two cases comparing the ground truth, LLM predictions, and debiasing results. The **Relation Debiasing** refers to the results after the structural expert model filters the re-predicted triplets, while the **Entity Debiasing** refers to the results obtained by applying local perturbations to the previous output and making new predictions.

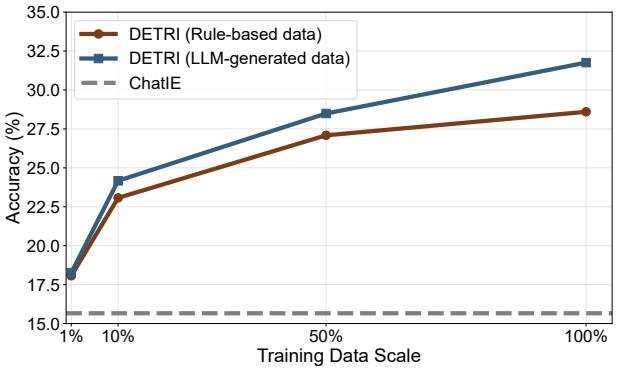

Figure 4: Accuracy (%) comparison under different scales and qualities of seen data on FewRel in the single-triplet setting with $m = 15$.

5.5 Data Scale Experiment

To evaluate how the quantity and quality of seen data affect the performance of DETRI, we randomly sample 1%, 10%, and 50% of the seen data to train the structural expert model. To further reduce reliance on LLMs during data preparation, we also manually construct high-quality negative samples based on common bias types identified in LLM outputs.

As shown in Figure 4, the structural expert model's accuracy improves significantly with more training data, indicating effective use of biased supervision. Under the same data scale, negative samples derived from actual LLM outputs consistently outperform rule-based ones. This suggests that real extraction errors better reflect practical scenarios and help the model learn more robust discriminative patterns. Notably, even with just 1% of training data, DETRI significantly outperforms ChatIE [Wei *et al.*, 2024], demonstrating strong robustness to noise and high data efficiency, highlighting the value of debiasing in ZSRTE.

5.6 Case Study

We analyze two representative cases in Table 4.

In `Case 1`, the LLM incorrectly identifies "he" as the subject and treats "Heavyweight title" as a redundant object, reflecting typical issues of over-prediction and entity shift. Although such redundant extractions are not entirely incorrect, they can negatively impact downstream tasks such as knowledge graph construction. In the first stage of DeTri, the model accurately retains the correct triplet and filters out the clearly incorrect ones. The second stage further refines the object by simplifying it to "Heavyweight," yielding a more precise result and demonstrating the structural expert's ability to correct fine-grained entity shift. However, due to the absence of information about "Teymur Mammadov" in the original LLM output, this entity is missing in the final prediction, indicating that error propagation may still lead to extraction omissions.

In `Case 2`, the LLM mistakenly aligns "Graz" with the relation "located in or next to body of water" and fails to correctly decompose "Peter Cook and Colin Fournier" into two separate triplets, illustrating relation confusion and entity shift problems. In the first stage, DETRI identifies the correct triplets and removes incorrect relations through re-prediction and confidence-based filtering. In the second stage, the span perturbation mechanism successfully separates conjunctive entities, demonstrating strong generalization capabilities.

6 Conclusion

In this work, we introduce DETRI, a novel framework designed to address the bias inherent in general-purpose LLMs when performing zero-shot relation triplet extraction. LLMs are leveraged as inspiration generators rather than definitive predictors. By constructing a synthetic biased dataset and designing a structural expert model, DETRI effectively captures relational semantics independent of LLM-induced biases. We further enhance the model with a span perturbation mechanism and a three-stage reasoning strategy to iteratively refine triplet predictions.

Extensive experiments on multiple benchmarks demonstrate that DETRI significantly outperforms state-of-the-art baselines, validating its effectiveness in extracting more accurate and unbiased relation triplets. In future work, we plan to further verify the effectiveness and generalization of our debiasing framework on other zero-shot information extraction tasks.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper. This work is supported by the National Natural Science Foundation of China (62276057).

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen and Li, 2021] Chih-Yao Chen and Cheng-Te Li. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[Chia *et al.*, 2022] Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

[Dagdelen *et al.*, 2024] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

[Furniturewala *et al.*, 2024] Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, 2024.

[Gong and Eldardiry, 2024] Jiaying Gong and Hoda Eldardiry. Prompt-based zero-shot relation extraction with semantic knowledge augmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13143–13156, 2024.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Han *et al.*, 2018] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, 2018.

[He *et al.*, 2023] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

[Huang *et al.*, 2018] Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2160–2170, 2018.

[Jiang *et al.*, 2024] Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. In *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, pages 2820–2837, 2024.

[Kim *et al.*, 2023] Bosung Kim, Hayate Iso, Nikita Bhutani, Estevam Hruschka, Ndapandula Nakashole, and Tom Mitchell. Zero-shot triplet extraction by template infilling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 272–284, 2023.

[Lan *et al.*, 2024] Yuquan Lan, Dongxu Li, Yunqi Zhang, Hui Zhao, and Gang Zhao. Rsed: Zero-shot relation triplet extraction via relation selection and entity boundary detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11256–11260. IEEE, 2024.

[Li *et al.*, 2023] Guozheng Li, Peng Wang, and Wenjun Ke. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, 2023.

[Li *et al.*, 2024] Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. Meta in-context learning makes large language models better zero and few-shot relation extractors. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6350–6358, 2024.

[Li *et al.*, 2025a] Hao Li, Yubing Ren, Yanan Cao, Yingjie Li, Fang Fang, Zheng Lin, and Shi Wang. Bridging the gap: Aligning language model generation with structured information extraction via controllable state transition. In *Proceedings of the ACM on Web Conference 2025*, pages 1811–1821, 2025.

[Li *et al.*, 2025b] Zehan Li, Fu Zhang, Kailun Lyu, Jingwei Cheng, and Tianyue Peng. Re-cent: A relation-centric framework for joint zero-shot relation triplet extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7344–7354, 2025.

[Li *et al.*, 2025c] Zehan Li, Fu Zhang, Wenqing Zhang, Jiawei Li, Zhou Li, Jingwei Cheng, and Tianyue Peng. Frame first, then extract: A frame-semantic reasoning pipeline for zero-shot relation triplet extraction. In *Pro-*

ceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27351–27364, 2025.

[Ma *et al.*, 2023] Xilai Ma, Jing Li, and Min Zhang. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, 2023.

[Ma *et al.*, 2024] Congda Ma, Tianyu Zhao, and Manabu Okumura. Debiasing large language models with structured knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10274–10287, 2024.

[McKenna *et al.*, 2023] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[Sainz *et al.*, 2023] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*, 2023.

[Shi *et al.*, 2024] Yuchen Shi, Guochao Jiang, Tian Qiu, and Deqing Yang. Agentre: An agent-based framework for navigating complex information landscapes in relation extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2045–2055, 2024.

[Sun *et al.*, 2024] Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM Web Conference 2024*, pages 4407–4416, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang and Lu, 2020] Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, 2020.

[Wei *et al.*, 2024] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. Chatie: Zero-shot information extraction via chatting with chatgpt, 2024.

[Wu *et al.*, 2024] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

[Xu *et al.*, 2024a] Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. Small models are valuable plug-ins for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 283–294, 2024.

[Xu *et al.*, 2024b] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357, 2024.

[Xu *et al.*, 2024c] Ting Xu, Haiqin Yang, Fei Zhao, Zhen Wu, and Xinyu Dai. A two-agent game for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7510–7527, 2024.

[Zhang *et al.*, 2019] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Jiaoyan Chen, Wei Zhang, and Huajun Chen. Transfer learning for relation extraction via relation-gated adversarial learning. *arXiv preprint arXiv:1908.08507*, 2019.

[Zhang *et al.*, 2024] Xinliang Frederick Zhang, Carter Blum, Temma Choji, Shalin Shah, and Alakananda Vempala. Ultra: Unleash llms' potential for event argument extraction through hierarchical modeling and pair-wise self-refinement. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8172–8185, 2024.

[Zhang *et al.*, 2025a] Fu Zhang, Xinlong Jin, Jingwei Cheng, Hongsen Yu, and Huangming Xu. Rethinking the role of llms for document-level relation extraction: a refiner with task distribution and probability fusion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 6293–6312, 2025.

[Zhang *et al.*, 2025b] Yifang Zhang, Pengfei Duan, Yiwen Yang, and Shengwu Xiong. Enhancing zero-shot relation extraction through staged interaction with large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Zhou *et al.*, 2024] Hanzhang Zhou, Junlang Qian, Zijian Feng, Lu Hui, Zixiao Zhu, and Kezhi Mao. Llms learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11972–11990, 2024.

[Zhu *et al.*, 2025] Senbin Zhu, ChenYuan He, Hongde Liu, Pengcheng Dong, Hanjie Zhao, Yuchen Yan, Yuxiang Jia, Hongying Zan, and Min Peng. Silc-efsa: Self-aware in-context learning correction for entity-level financial sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4980–4992, 2025.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.