Towards Annotation Bias in Information Extraction Across Datasets

Anonymous ACL submission

Abstract

Annotation bias is a negative phenomenon that can mislead models. However, annotation bias in information extraction appears not only across datasets from different domains but also within datasets sharing the same domain. We identify two types of annotation bias in IE: bias among information extraction datasets and bias between information extraction datasets and instruction tuning datasets. To systematically investigate annotation bias, we conduct three probing experiments to quantitatively analyze it and discover the limitations of unified information extraction and large language models in solving annotation bias. To mitigate annotation bias in information extraction, we propose a multi-stage framework consisting of annotation bias measurement, bias-aware fine-tuning, and task-specific bias mitigation. Experimental results demonstrate the effectiveness of our framework in addressing annotation bias.

1 Introduction

Annotation bias in machine learning refers to the skewed or inconsistent labeling of data in the process of training. This bias can occur when the datasets used for training are not representative of the real-world scenarios or contain inherent biases. Generally speaking, annotation bias is viewed as a negative phenomenon since it may lead to inaccurate and non-generalizable models. Annotation bias is initially identified in computer vision studies, particularly in the analysis of stereotypical biases within facial expression datasets(Chen and Joo, 2021). It is also studied in Natural Language Processing (NLP) such as the instruction bias caused by crowd-sourcing problem in multiple Natural Language Understanding (NLU) benchmarks (Parmar et al., 2022).

Annotation bias is also a prevalent issue in information extraction (IE). As the fast development of Unified Information Extraction (UIE) and Large Language Models (LLMs) in recent years, two novel annotation bias emerge, which are: *Bias among IE datasets* and *Bias between IE and instruction tuning (IFT) datasets*. Regarding *Bias among IE datasets*, it refers to the annotation differences between different data sets under the same annotation schema. As illustrated in Fig 1, different datasets have different annotation results to the same input for both Named Entity Recognition (NER) and Relation Extraction (RE) tasks. Regarding *Bias between IE and instruction tuning datasets*, it highlights the mismatch between information extraction task and general task. As depicted in Fig 1, although GPT-4(OpenAI, 2023) is capable of extracting entities or relational triples in accordance with the specified task description without providing extra examples, its annotations differ from those in the existing datasets.

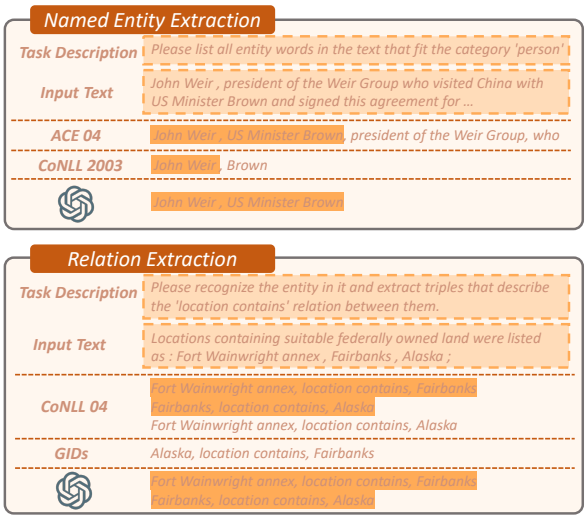

Figure 1: Annotation bias among different datasets and LLMs even when they share the same entity type (for NER) or the same relation type (for RE)

To systematically investigate annotation bias in IE, we devise a series of probing experiments. First, we analyze *whether annotation bias exists and how it varies among datasets sharing the same tasks*.

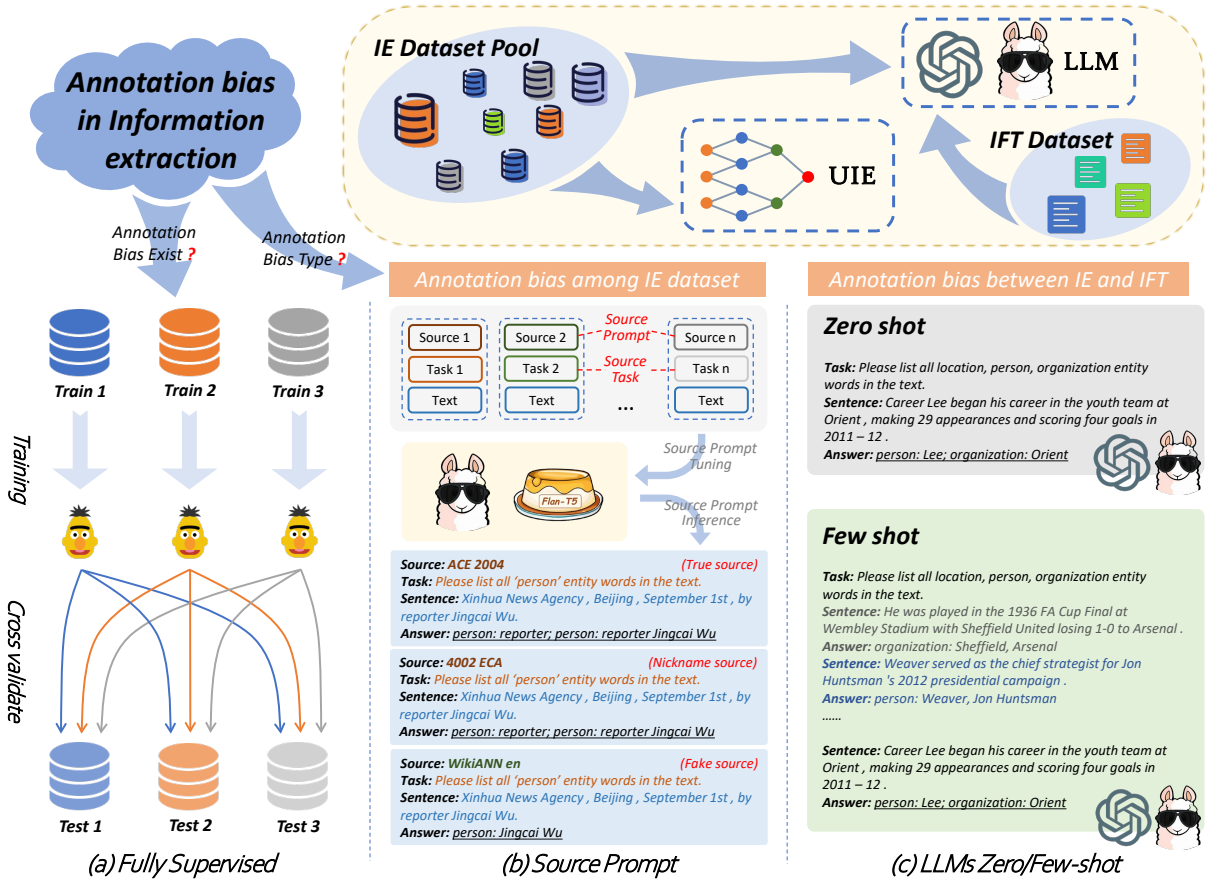

Figure 2: Three settings for the probing tasks on annotation bias across datasets, including (a) fully supervised, (b) source prompt and (c) LLMs zero/few-shot.

By conducting cross-validation experiment among various dataset in the NER and RE, we observe a significant decrease in performance, indicating that annotation bias negatively impacts the transferability of a fully-supervised model. An intuitive way to alleviate annotation bias is unified information extraction, which is trained across multiple IE dataset. Hence, we analyze *in the unified information extraction mode, does annotation bias still exist?* By introducing source prompt that apply true or fake source name for the UIE models, we discover the inconsistency of the UIE for extraction, which indicates that UIE suffer from annotation bias among IE dataset. The other way to mitigate annotation bias is LLM, which is able to understand a wide range of human instructions. Thereupon, we analyze *Can LLMs address the challenge of annotation bias?* By conducting experiments on few-shot settings on NER and RE task with in-context learning, we find that it's difficult for LLMs without parameter updates to attain satisfactory performance, which indicates that LLM still suffers from annotation bias between IE and instruction tuning dataset.

According to our probing experiments, it is imperative to address annotation bias when proposing a universal solution for IE tasks. However, mitigating annotation bias is non-trivial, primarily owing to the following three challenges. 1. Enhancing the capacity of LLMs in general information extraction tasks is vital to reduce the annotation bias between information extraction datasets and instruction tuning datasets; 2. It is essential to mitigate the annotation bias during the tuning of LLMs with diverse datasets; 3. Learning from new data over time, adapting to new tasks while ensuring the model remains relevant and less biased, is a significant challenge.

To address these challenges, we propose a framework to alleviate annotation bias, which consist of annotation bias measurement, bias-aware fine-tuning and task-specific bias mitigation. With using Fleiss' Kappa(Fleiss, 1971), we measure the two type of annotation bias above. Then we conduct bias-aware fine-tuning with multiple information extraction instructions to enhance the extraction

capabilities with less annotation bias. Ultimately, we conduct the task-specific bias mitigation, with low-rank adaptation technique (LoRA)(Hu et al., 2021) for specific information extraction tasks to further align the LLMs with annotation.

Our main contributions are as follows:

- We propose several probing experiments to demonstrate the annotation bias among various information extraction task, which highly affect the performance of large language models and universal information extraction framework.
- We introduce a framework that consists of annotation bias measurement, bias-aware fine-tuning and task-specific bias mitigation, which can effectively mitigate the annotation bias in IE.
- We conduct comprehensive experimental study on several IE datasets, which demonstrates that our framework outperforms SoTA IE baselines.

2 The Annotation Bias Probing Task

We initially propose an experiment employing cross-validation to investigate whether the issue of annotation bias exists in information extraction task. Subsequently, we design two specific detection tasks: source prompt detection and few-shot prompting in LLMs, to inspect two categories of annotation bias: bias among IE dataset and bias between IE and instruction fine-tuning dataset. These tasks aim to explore the effectiveness of UIE and LLM frameworks in resolving the annotation bias problem.

2.1 Whether annotation bias exists?

To better illustrate the annotation bias among different information extraction task, we design cross extraction task. As Figure 2(a) shows, we train multiple fully-supervised model with different datasets on the same task respectively (NER and RE), and test them on the other dataset to evaluate whether the annotation bias exists.

We first introduce two BERT-based extraction framework to handle the NER and RE task respectively.

Named Entity Recognition We adopt GlobalPointer(Su et al., 2022), an efficient span-based approach that models the beginning and the end positions to predict the entity with a 2-dimension scoring matrix. With the extended softmax and cross-entropy, GlobalPointer can better learn from class imbalance scenarios.

Relation Extraction We adopt RERE(Xie et al., 2021) as the basic model for relation extraction, which is a pipeline approach that performs sentence-level relation detection then subject/object extraction. Specifically, RERE model the former sub-task as multi-class classification task and the latter as span detection task.

In the process of cross-validation, due to the presence of label type biases in different datasets, (e.g. in ACE 2004 dataset, extraction of the `weapon` entity is required, while CoNLL 2003 not), we focus solely on the types of labels (entity types in NER and relationship types in RE) that are annotated in both the training and testing datasets (e.g. `person` for ACE 2004 and CoNLL 2003).

To avoid the impact of text distribution shift on the experimental results, we sample a subset of sentences with similar semantics as cross-validation set. Specifically, we measure the semantic similarity between two sentences through calculating the cosine similarity of their sentence embedding, and we define the semantic similarity of the sentence $sent_i$ to the dataset $\mathcal{D}$. Finally, we filter out all sentences below the $threshold(\mathcal{D})$.

$$sim(sent_i, \mathcal{D}) = \max_{ref_j \in \mathcal{D}} cosine(V_{sent_i}, V_{ref_j}) \quad (1)$$

$$threshold(\mathcal{D}) = \sigma \cdot \frac{1}{|\mathcal{D}|} \sum_{s_i \in \mathcal{D}} sim(s_i, \mathcal{D} \setminus \{s_i\}) \quad (2)$$

where V_S denotes the embedding vector of a sentence S encoded by a sentence model[1], and σ denotes the hyper-parameters that adjust the threshold, which is set 0.7 empirically.

2.2 Can UIE address annotation bias?

Unified information extraction, which encodes different extraction structures with a pre-defined structured extraction language, can precisely recognize the extraction instruction. Inspired by (Li et al., 2022), which introduces a novel prompt-based method in a transferable setting on text generation task, we adopt a source prompt settings for probing. Briefly, in our experiment setting, source can be denoted as the name of the dataset (i.e ACE 2004).

[1]We adopt MPNet(Song et al., 2020) as our sentence embedding encoder, which is commonly used for retrieval

Presenting UIE with various sources by indicating which dataset the instance is from, we can guide it to yield different extraction results, thereby assessing whether it can keep the same result with different source prompt.

As Figure 2(b) shows, the probing experiment consists of two part: *source prompt tuning* and *source prompt inference*. Initially, we undertake a source prompt tuning process to improve the UIE model's ability to recognize different sources. Subsequently, we examine the annotation bias within the UIE model by introducing various sources.

Source Prompt Tuning The source prompt process can be regarded as a general multi-task learning framework. First, we define a set of source information extraction task $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_n\}$, where the k-th task $\mathcal{S}_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ contains N_k tuples of the input text $x_i^k \in \mathcal{X}_k$ and its corresponding output text $y_i^k \in \mathcal{Y}_k$. For a target information extraction task $\mathcal{T}$, the goal of multi-task learning is to use previously learned task-specific knowledge of the source tasks $\mathcal{S}$ to better predict the extraction result. Compared to the traditional multi-task fine-tuning scenario, we learn an independent *source prompt* p_k for each source information extraction task $\mathcal{S}_k$ in source prompt tuning, where x_i^k consist of extraction task source name s_k, information extraction task description t_k, and the sentence $sent_i^k$.

To clearly clarify that UIE with instruction tuning can implicitly learn the annotation principle through source prompt, we assign a nickname p'_k for every dataset and randomly replace p_k with p'_k. For simplify, we merely reverse the order of the original dataset names, thereby generating a non-natural language nickname. This procedure is designed to eliminate the influence caused by the differences in learning various source names in the UIE and ensure that the discrepancies in results between true and fake settings are solely due to dataset annotation bias.

Specifically, we adopt Llama-v2-13B (Touvron et al., 2023) and FlanT5-11B (Chung et al., 2022) as our backbone models in source prompt tuning settings because of their powerful instruction understanding and instruction following capability. Based on multiple datasets in NER and RE, we add an additional *source prompt* to every extraction instance to indicate which it belongs to. Further details on source prompt tuning are described in the Appendix A.

Source Prompt Inference In reference, we respectively offer different source prompt with the same extraction instance in testing set to our UIE that are fine-tuned on the dataset with source prompt. To probing the annotation bias in universal generative information extraction bias, UIE predicts the extraction result with ***True source*** (extraction case with the origin source name), ***Nickname source*** (nickname of the original source name) and ***Fake source*** (extraction case with a fake source name). With different source name, UIE generates different extraction result following different annotation principle learning from source prompt tuning.

2.3 Can LLMs address annotation bias?

Large language models show remarkable instruction understanding capability, which help achieve extraordinary performance on various tasks. Because of the annotation bias between IE dataset and instruction tuning dataset, there is a significant performance gap in LLMs when it comes to information extraction (IE) task (Wadhwa et al., 2023). In context learning, where LLMs make predictions only based on contexts augmented with a few examples, is a training-free learning framework for the model adapting to new task(Dong et al., 2023). It is considered as a solution to address the annotation bias between IE dataset and instruction tuning dataset.

As Figure 2(c) shows, we conduct the probing experiment with multiple LLMs in both zero and few shot settings.

We use open-source LLMs Llama-v2-chat-70B(Touvron et al., 2023), and close-source LLMs ChatGPT, GPT4(OpenAI, 2023) as backbone. In zero-shot settings, we prompt LLMs with task description, which can probe the annotation bias between IE and IFT dataset. While in few-shot settings, we prompt LLMs with task description and another 4 cases randomly sampled from the corresponding training set, to probe whether in-context learning address annotation bias. For fair comparison, we sample 200 cases in each dataset, and test them in zero-shot and few-shot respectively.

3 Empirical Study of Annotation Bias

3.1 Whether annotation bias exists?

Following the cross validation setting described in Section 2.1, we conduct experiment separately on NER and RE tasks in general domain.

	A04[1]	A05[2]	C03[3]	Ont[4]	Wie[5]	TN7[6]	WiN[7]	PoN[8]
A04[1]	85.10	82.19	35.77	28.89	49.89	28.06	30.54	17.64
A05[2]	83.44	84.45	37.80	26.43	46.53	26.94	29.09	18.23
C03[3]	24.10	16.57	92.19	55.82	55.10	78.26	92.08	53.67
Ont[4]	32.53	21.20	60.60	89.69	49.76	34.75	61.23	37.58
Wie[5]	23.09	8.42	67.10	41.14	86.60	61.99	70.96	44.13
TN7[6]	25.60	21.07	76.16	56.15	73.95	63.39	82.70	54.45
WiN[7]	25.48	20.61	80.10	58.69	57.33	63.44	95.21	51.96
PoN[8]	14.58	10.84	44.36	35.28	40.26	32.65	69.66	77.77

[1] ACE 2004 [3] CoNLL 2003 [5] WikiANN en [7] WikiNeural
[2] ACE 2005 [4] Ontonotes [6] TweetNER 7 [8] PolyglotNER

Table 1: Annotation bias among different NER tasks. For each dataset we train a model fully-supervised on training set and evaluate them on other testing set.

	CoNLL 04	NYT10	NYT11	GIDs	WikiKBP
CoNLL 04	61.12	10.20	12.07	-	26.98
NYT10	14.36	89.68	52.29	14.33	30.32
NYT11	8.78	83.32	56.82	10.70	32.64
GIDs	-	7.77	6.45	65.12	55.65
WikiKBP	0.00	15.05	2.53	26.49	36.57

Table 2: Annotation bias among different RE tasks. For each dataset we train a model fully-supervised on training set and evaluate them on other testing set.

Table 1,2 show the validation result in fully-supervised settings.

Briefly, we denote the model that train and test on the same dataset *reference model*. The numbers in the cells of the table represent the F1 values of compared with the golden label, while the depth of color in each cell indicates the relative quality of extraction compared to the reference model. In other words, the darker the color, the more consistent the extraction results are with the reference model.

Intuitively, the deepest red cells are distributed along the diagonal of the entire table, which illustrate the annotation bias exists among different datasets even they share the some same types. Especially for the NER tasks, even there are several datasets that focus the common entity type such as `person`, `location` and `date`, the annotation bias can lead to significant variations in the model's extraction capabilities.

3.2 Can UIE address annotation bias?

Following the source prompt setting described in Section 2.2, we tuning Llama-13b and Flan-T5

Model	Llama-13b			Flan-T5		
Source	*True*	*Nickname*	*Fake*	*True*	*Nickname*	*Fake*
Named Entity Recognition						
ACE 04	84.93	84.89	60.85	77.82	78.41	45.79
ACE 05	84.85	85.16	61.56	79.20	79.59	44.10
CoNLL 03	81.02	80.87	73.34	78.94	78.84	69.23
Ontonotes	91.85	91.81	81.79	91.03	91.04	78.71
WikiANN en	89.54	89.65	81.43	76.26	76.07	66.08
TweetNER 7	68.92	69.11	66.19	68.35	68.45	60.44
WikiNeural	96.03	95.93	83.51	94.03	94.03	74.30
PolyglotNER	80.21	80.41	68.34	74.00	74.03	54.24
avg	-	-	-12.6	-	-	-18.4
Relation Extraction						
CoNLL 04	69.88	69.51	61.73	67.09	67.00	57.34
NYT10	97.80	97.78	94.82	96.20	96.20	90.54
NYT11	76.14	76.24	72.82	76.14	76.41	71.94
GIDs	80.49	80.15	78.69	76.41	76.34	74.26
WikiKBP	64.68	65.67	63.50	63.78	63.94	59.64
avg	-	-	-3.5	-	-	-5.2

Table 3: Different extraction result by prompting source prompt tuning UIE with *true*, *nickname* and *fake* source name. Nickname source refers to an alternative representation of the original dataset's name, fake source refers to a randomly sampled source name.

Dataset	Llama-chat-70B	ChatGPT	GPT4
ACE04	8.56 | 30.42	19.68 | 32.81	13.70 | 35.16
ACE05	17.64 | 33.48	20.83 | 34.32	16.13 | 45.30
CoNLL 03	33.89 | 49.36	39.70 | 55.90	46.66 | 64.99
Ontonotes	11.86 | 27.56	22.14 | 28.83	31.70 | 40.57
WikiANN en	32.87 | 50.00	50.83 | 57.90	51.57 | 59.03
TweetNER 7	31.77 | 35.68	32.98 | 38.13	36.62 | 47.88
WikiNeural	42.98 | 57.03	50.00 | 59.83	65.23 | 70.66
PolyglotNER	21.44 | 30.91	42.20 | 44.88	45.14 | 43.23
CoNLL 04	3.36 | 18.77	9.22 | 23.86	24.62 | 29.86
NYT10	2.97 | 13.17	2.13 | 13.64	16.67 | 20.13
NYT11	2.03 | 5.33	1.93 | 6.50	8.00 | 12.00
GIDs	11.36 | 7.92	7.89 | 19.45	6.82 | 24.54
WikiKBP	18.55 | 29.56	17.25 | 32.41	25.00 | 45.85

Table 4: Performance of Open-source LLM and close-source LLM on various information extraction task in zero-shot and few-shot settings.

with source prompt instruction and prompting them with three source settings.

The table 3 shows the extraction result evaluated by F1 scores. By replacing true source name with fake source name, the F1 score in all NER and RE task drop on average 12.6/3.5 and 18.4/5.2, while replacing true source names with nicknames, there is virtually no difference in the results. The distinct performance gap demonstrates that UIE is unable to mitigate annotation bias while in multi-task learning process. The implicitly annotation bias would diffuse the model, which leads to inconsistent extraction result with the same extraction

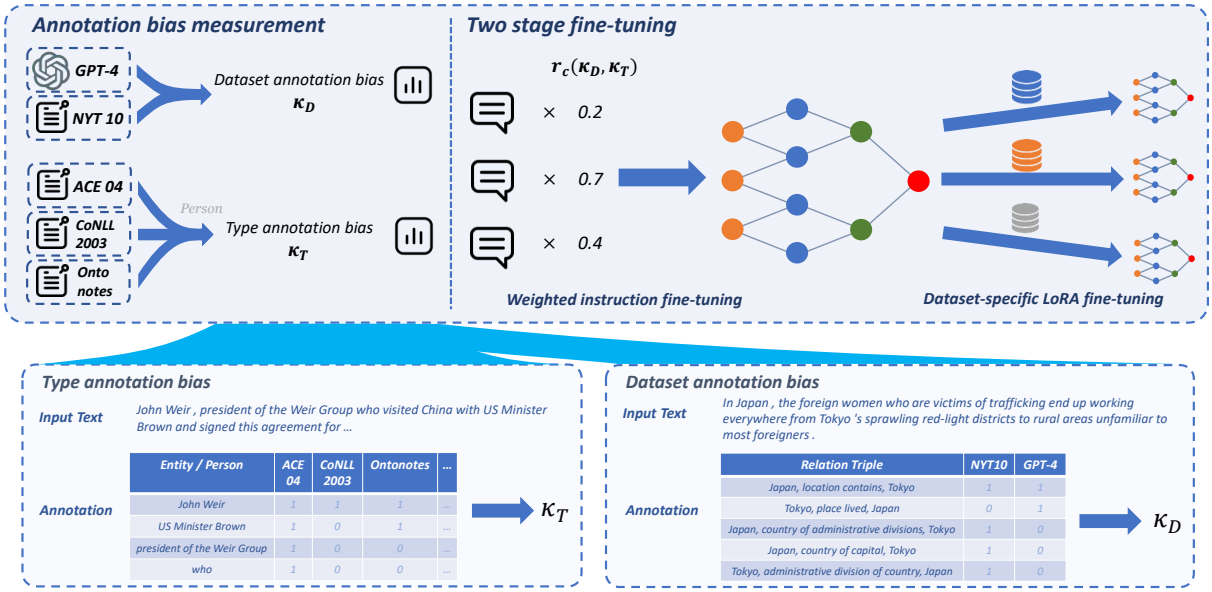

Figure 3: Existing information extraction datasets focus on various task with different schema. However, there is annotation bias between different datasets even when they share the same entity type (for NER) or the same relation type (for RE). Besides, the annotation of the dataset varies from LLMs.

task instruction.

3.3 Can LLM address annotation bias?

The performance of the models on different tasks is shown in Tab. 4

Among the models assessed, GPT-4 almost achieve all best performance in every dataset in both zero-shot and few-shot settings. Besides, few-shot settings providing similar case of same dataset in the context can improve about 9.82 on average compared to zero-shot settings. This suggests that in-context learning can partly mitigate the annotation bias.

Nevertheless, it remains challenging for a standard, off-the-shelf method to achieve the same level of performance as that of a fully supervised approach, which indicates that there are a huge annotation bias between IE and instruction fine-tuning dataset. And there are two more restrictions for applying LLMs to IE. First, limited by the context length, it is impossible to provide all the cases with annotation in the context. Second, the annotation bias between information extraction dataset close the door for designing a comprehensive prompting to perfectly describe the extraction task.

4 Alleviate Annotation Bias

In this section, we illustrate how to enhance information extraction capability of large language models (LLMs).

Based on the probing task and conclusion in section 2,3, annotation bias among different dataset highly affect the performance of UIE and LLMs, which indicates that framework with one-stage and parameter-free update can not address the annotation bias. Consequently, we introduce a two-stage fine-tuning framework as a solution. Moreover, building upon the two types of annotation bias we have identified, we explicitly measure these biases and integrate them into the fine-tuning framework to effectively mitigate the impact of annotation bias.

4.1 Annotation Bias Measurement

First, we introduce the **Fleiss' Kappa**, a statistical measure used for assessing the reliability of agreement between multiple raters when assigning categorical ratings to a number of items, which can help in identifying and mitigating annotation bias.

$$\kappa = 1 - \frac{1 - p_0}{1 - p_e} = \frac{p_0 - p_e}{1 - p_e} \tag{3}$$

where p_o denotes the *Observed Agreement*, the proportion of times that the raters actually agree, and p_e denotes the *Expected Agreement*, which represents the agreement that could be expected purely by chance. Suppose there are N cases for a task, and each data is labeled n times, and k is the number of categories. They can be calculated

according to the following formula.

$$p_e = \sum_{j=1}^{k} p_j^2, \;\; p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{4}$$

$$p_o = \frac{1}{N} \sum_{i=1}^{N} p_i, \;\; p_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{5}$$

where n_{ij} denotes the number of annotator that label case i as catogory j.

Specifically, we focus on the annotation bias in information extraction and split the annotation bias into two type: ***dataset annotation bias*** and ***type annotation bias***.

Dataset Annotation bias κ_D Recognized as the agreement between GPT4 and the annotation of dataset, serving as a measure of reliability for transforming information extraction into instruction tuning dataset. It is conducted by calculating the Fleiss' Kappa between the GPT4 extraction results and the golden annotation of the dataset.

Type Annotation bias κ_T Considered as the agreement among information extraction datasets with the same type either entity or relationship, and serves as a metric to evaluate the reliability of these types in terms of consistent annotation.

4.2 Bias-Aware Fine-Tuning

We further fine-tuning the LLM with information extraction dataset through C-RLFT(), which enables leveraging mixed-quality training data. We define the quality of the training samples as metrics based on κ_D and κ_T. Suppose there are N entity or relation triples in a case, we calculate the coarse-grained rewards of each case $r_c(x_i, y_i)$ by the formula below.

$$r_c(x_i, y_i) = (1 + \kappa_D)\frac{1}{N} \sum_{i=1}^{N} \kappa_{T_i} \tag{6}$$

4.3 Task-Specific Bias Mitigation

To better enhance the performance of LLM on a specific information extraction dataset, we adopt low rank adaptation(LoRA) for further instruction tuning. We hypothesize the updates to the weights for a dataset have a low intrinsic rank. These low intrinsic dimension adaptation can mitigate the annotation bias between a multi-task learning model and the dataset. Specifically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we constrain its update with a low rank decomposition.

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{7}$$

The model updates its parameter knowledge through further fine-tuning on specific dataset to align the annotation principle.

5 Experiments

Dataset	UIE	USM	InstructUIE	Ours
ACE 04	86.89	**87.62**	-	79.07
ACE 05	85.78	**87.14**	86.66	78.88
CoNLL 03	92.99	**93.16**	92.94	91.26
Ontonotes	-	-	**90.19**	86.08
WikiANN en	-	-	85.13	**86.36**
TweetNER 7	-	-	**64.97**	58.68
WikiNeural	-	-	91.36	**92.73**
PolyglotNER	-	-	70.15	**71.41**
CoNLL 04	75.00	**78.84**	78.48	65.42
NYT10	-	-	**90.47**	82.37
NYT11	-	-	**56.06**	54.81
GIDs	-	-	**81.98**	77.32

Table 5: Main result for comparing with other models on NER and RE tasks.

5.1 Experimental Setup

Our baseline contains: UIE(Lu et al., 2022), USM(Lou et al., 2023), and InstructUIE(Wang et al., 2023). All of them are trained with full-parameter updating on specific dataset.

5.2 Main Results

Table 5 presents the result on different dataset with SoTA model. Although trained on several information extraction datasets in gerneral domain, which is unfair for comparing the baseline that are trained with other dataset, our framework achieves competitive performance with the baseline on many dataset. It is worth noting that in the dataset that only focus on `person`, `location`, `organization` (type list in Table 9), our framework achieve the best performance on WikiANN en, WikiNeural and PolyglotNER. It proves the effectiveness of our framework on mitigating the annotation bias among different dataset. Our framework for measuring and integrating them remains

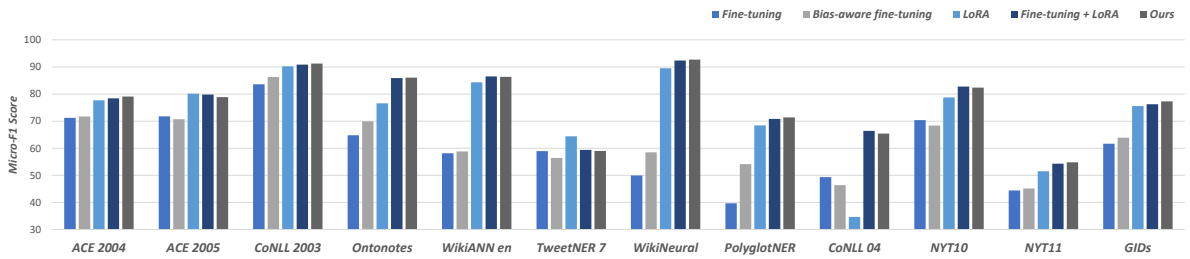

Figure 4: Ablation Study on 12 information extraction dataset (NER and IE)

5.3 Experiment with two-stage fine-tuning

To better improve the effectiveness of our two-stage fine-tuning framework, we conduct ablation study comparing with the following baseline: 1.***Fine-tuning***: fine-tuning the model with information extraction; 2.***Bias-aware fine-tuning***: first stage fine-tuning in section 4.2; 3.***LoRA***: instruction-tuning with LoRA on specific dataset; 4.***Fine-tuning+LoRA***: data-specific instruction-tuning with LoRA on the weight of baseline 1; 5.***Ours***: our two-stage fine-tuning framework.

The result is shown in Figure 4. Overall, our framework almost achieve the best performance compared to the baseline above, which demonstrate its effectiveness. By comparing baseline 1 and 2, it prove that our bias-aware fine-tuning can alleviate annotation bias among IE datasets and help models better align with GPT4. It is also noticeable that two-stage fine-tuning can consistently improve the performance on the specific dataset, which is attributed to the task-specific bias mitigation.

6 Related Work

6.1 LLMs for information extraction

Large language models has shown remarkable performance in instruction following (OpenAI, 2023). To better align the natural instruction task from pre-trained and instruction tuning task, (Wei et al., 2023; Wadhwa et al., 2023; Zhang et al., 2023) convert structural information extraction task into natural instruction task such as question answering, multi-choice and etc. While (Li et al., 2023; Guo et al., 2023) recast the structured output in the form of code to better leverage the LLMs of code to address the complex structure. Although LLMs show impressive performance in various information extraction task by designing fine-grained instruction, they still fail to address annotation bias without further tuning.

6.2 Annotation bias

Annotation bias is widely study in other field. (Chen and Joo, 2021) demonstrate that many expression datasets contains significant annotation biases between genders, while (Parmar et al., 2022) study the bias in annotation instruction. They mainly focus on the annotation bias in a single dataset, but fail to transfer the framework to multi-task learning settings, which is vital important in enhancing large language model with further training.

6.3 Universal Information extraction

Unified Information extraction, proposed by (Lu, 2022), uniformly encodes various information extraction task with a pre-defined structured extraction language(SEL), and enhance the common IE abilities via a large-scale pre-trained generation model. (Lou et al., 2023) further introduce USM to model different IE tasks, while (**?**) unified tasks into natural language instruction. GoLLIE convert IE schema into code-style structural description and add guidelines to improve zero-shot results (Sainz et al., 2023). However, they mainly focus on how to encode different extraction task into a uniform structure but fail to notice and detect the annotation bias among various datasets.

7 Conclusion

In the paper, we propose the annotation bias problem in information extraction task. We conduct several probing task to comprehensively demonstrate the existences of annotation bias. Mean-times, we find that UIE and LLMs with zero/few-shot still hard to address annotation bias problem. We propose a two-stage tuning framework, which consist of multi-task learning and task-specific tuning, to alleviate the annotation bias in specific task. Experimental results shows that our method is efficient for mitigating annotation bias.

Limitation

We systematically investigate annotation bias in IE with devising a series of probing experiments. And we propose a multi-stage framework to mitigate annotation bias in IE. However, there are still some limits of our probing experiment and the solution framework.

First, our probing task only focus on the the annotation bias among NER and RE tasks, which doesn't cover all the task in information extraction, which remains improvement for the future work.

Second, the performance of our solution framework is restricted by two main reason: 1.more diverse dataset can be used for the bias-aware finetuning dataset; 2.the choice on backbone model also plays an important role in model performance. More experiments can more effectively validate the effectiveness of the proposed framework.

Ethic statement

We hereby declare that all authors of this article are aware of and adhere to the provided ACL Code of Ethics and honor the code of conduct.

Use of Human Annotations Human annotations are only utilized in the early stages of methodological research to assess the feasibility of the proposed solution. All annotators have provided consent for the use of their data for research purposes. We guarantee the security of all annotators throughout the annotation process, and they are justly remunerated according to local standards. Human annotations are not employed during the evaluation of our method.

Risks The datasets used in the paper have been obtained from public sources and anonymized to protect against any offensive information. Though we have taken measures to do so, we cannot guarantee that the datasets do not contain any socially harmful or toxic language.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Yunliang Chen and Jungseock Joo. 2021. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M Strassel, and Jonathan Wright. 2012. Linguistic resources for 2013 knowledge base population evaluations. In *TAC*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, et al. 2023. Retrieval-augmented code generation for universal information extraction. *arXiv preprint arXiv:2311.02962*.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Learning to transfer prompts for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3506–3518.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. Codeie: Large code generation models are better few-shot information extractors. *arXiv preprint arXiv:2305.05711*.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching.

Daming Lu. 2022. daminglu123 at SemEval-2022 task 2: Using BERT and LSTM to do text classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 186–189, Seattle, United States. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7072–7079.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco Barbieri, and Jose Camacho-Collados. 2022. Named entity recognition in twitter: A dataset and analysis on short-term temporal shifts. *arXiv preprint arXiv:2210.03797*.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

C. Walker and Linguistic Data Consortium. 2005. *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised 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 1: Long Papers)*, pages 3572–3581, Online. Association for Computational Linguistics.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *arXiv preprint arXiv:2305.11159*.

A Further detail about source prompt settings

To eliminate the instruction bias that different dataset focus on different types of entity or relation, we employ a task decomposition approach, which involves constructing separate instructions for every entity type or relationship type. It helps decompose a task instruction with many types into atomic task instruction, which can share across different datasets. Such setting compels the UIE model to focus solely on the source name and apply distinct extraction principle for different source prompts. Table 6 show the case of task decomposition.

Original Extraction Instruction
Instruction: Please list all entity words in the text that fit the category. Here's the category list: *[person,organization,location]* And then output the result in the format of "'type1: entity1; type2: entity2; ...'" Input: *[Input text for NER]* Output:
Decomposed Extraction Instruction
Instruction: Please list all entity words in the text that fit the category. Here's the category list: *[person]* And then output the result in the format of "'type1: entity1; type2: entity2; ...'" Input: *[Input text for NER]* Output:
Instruction: Please list all entity words in the text that fit the category. Here's the category list: *[organization]* And then output the result in the format of "'type1: entity1; type2: entity2; ...'" */*Input text*/* Input: *[Input text for NER]* Output:
Instruction: Please list all entity words in the text that fit the category. Here's the category list: *[location]* And then output the result in the format of "'type1: entity1; type2: entity2; ...'" Input: *[Input text for NER]* Output:

Table 6: A case for decomposing NER tasks instruction which focus on the entity type: `person`, `organization` and `location`

B Prompt template

We use the prompt in Table7 for probing experiments and multi-stage fine-tuning.

Prompt for Named Entity Recognition
*/*Task prompt*/* Instruction: Please list all entity words in the text that fit the category. Here's the category list: */*Entity type List*/* *[List of the entity type]* */*Output Format*/* And then output the result in the format of "'type1: entity1; type2: entity2; ...'" */*In-context learning cases*/* */*Input text*/* Input: *[Input text for NER]* Output:
Prompt for Relation Extraction
*/*Task prompt*/* Instruction: Given a sentence or paragraph, and a given relationship set that describe the relation between entities. Here's the relation set: */*Relation type List*/* *[List of the relationship type]* */*Output Format*/* Output the result in the format of "'(subject1, relation1, object1), (subject2, relation2, object2), ...'" */*In-context learning cases*/* */*Input text*/* Input: *[Input text for RE]* Output:

Table 7: The prompts for two type of information extraction task: NER and RE. The prompts

C Fleiss' Kappa

Table 8 show the κ with measure between dataset annotation and GPT-4 output.

D Detail on the dataset

We use 13 dataset in named entity recognition and relation extraction. For NER task, the used dataset include ACE04, ACE05(Walker and Consortium, 2005), CoNLL2003(Sang and De Meulder, 2003), Ontonotes(Hovy et al., 2006),PolyglotNER(Al-Rfou et al., 2015), TweetNER(Ushio et al., 2022), WikiNeural(Tedeschi et al., 2021), WikiANN(Pan et al., 2017). For RE task, the used dataset include CoNLL 2004(Roth and Yih, 2004), GIDS(Jat et al., 2018), NYT10(Riedel et al., 2010), NYT11-HRL(Takanobu et al., 2019), WikiKBP(Ellis et al., 2012).

The pre-defined entity or relation types of every dataset is shown in Table 9

Dataset	Fleiss' Kappa
ACE 2004	-0.648
ACE 2005	-0.546
CoNLL 2003	-0.350
Ontonotes	-0.594
PolyglotNER	-0.567
TweetNER7	-0.521
WikiANN en	-0.409
WikiNeural	-0.293
conll04	-0.701
GIDS	-0.748
NYT10	-0.799
NYT11	-0.879
WikiKBP	-0.541

Table 8: Caption

Dataset	Annotation type
	Named Entity Recognition
ACE 2004	geographical social political, organization, person, location, facility, vehicle, weapon
ACE 2005	organization, person, geographical social political, vehicle, location, weapon, facility
CoNLL 03	location, else, organization, person
Ontonotes	date, organization, person, geographical social political, national religious political, facility, cardinal, location, work of art, law, event, product, ordinal, percent, time, quantity, money, language
PolyglotNER	location, person, organization
TweetNER 7	group, creative work, person, event, product, location, corporation
WikiANN en	location, person, organization
WikiNeural	location, person, organization
	Relation Extraction
CoNLL 04	company founded place, location contains, place lived, person of company, kill
GIDs	place of death, place of birth, education degree, education institution
NYT10	ethnicity, place lived, geographic distribution, company industry, country of administrative divisions, administrative division of country, location contains, person of company, profession, ethnicity of people, company shareholder among major shareholders, sports team of location, religion, neighborhood of, company major shareholders, place of death, nationality, children, company founders, company founded place, country of capital, company advisors, sports team location of teams, place of birth
NYT11	nationality, country capital, place of death, children, location contains, place of birt, place lived, administrative division of country, country of administrative divisions, company, neighborhood of, company founders
WikiKBP	parent, children, person of company, place of birth, place of death, place lived, religion

Table 9: The type of entity or relationship in each dataset.