Unexpected Phenomenon: LLMs' Spurious Associations in Information Extraction

Weiyan Zhang[1], **Wanpeng Lu**[1], **Jiacheng Wang**[1], **Yating Wang**[1], **Lihan Chen**[2], **Haiyun Jiang**[3], **Jingping Liu**[1]*, **and Tong Ruan**[1]*
[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
[2]Beijing Institute of Control Engineering, Beijing, China [3]Tencent AI Lab, Shenzhen, China
y10190118@mail.ecust.edu.cn, lhc825@gmail.com, haiyunjiang@tencent.com, {jingpingliu,ruantong}@ecust.edu.cn

Abstract

Information extraction plays a critical role in natural language processing. When applying large language models (LLMs) to this domain, we discover an unexpected phenomenon: LLMs' spurious associations. In tasks such as relation extraction, LLMs can accurately identify entity pairs, even if the given relation (label) is semantically unrelated to the pre-defined original one. To find these labels, we design two strategies in this study, including forward label extension and backward label validation. We also leverage the extended labels to improve model performance. Our comprehensive experiments show that spurious associations occur consistently in both Chinese and English datasets across various LLM sizes. Moreover, the use of extended labels significantly enhances LLM performance in information extraction tasks. Remarkably, there is a performance increase of 9.55%, 11.42%, and 21.27% in F1 scores on the SciERC, ACE05, and DuEE datasets, respectively.[1]

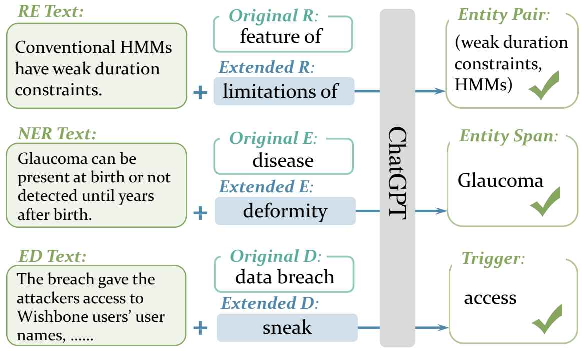

Figure 1: The phenomenon of LLMs' spurious associations in RE, NER, and ED tasks. Taking the RE task as an example, even if we provide the text with a relation that has no semantic relevance to the original relation, the output remains unchanged from when the original relation is used as input.

1 Introduction

Information Extraction (IE) plays a vital role in natural language processing (NLP), aiming to extract pre-defined types of information from unstructured text sources. Typical tasks in IE include Relation Extraction (RE) (Shang et al., 2022), Named Entity Recognition (NER) (Li et al., 2022), and Event Detection (ED) (Xie and Tu, 2022). Despite its importance, IE often faces obstacles in limited-data scenarios, such as zero-shot or few-shot settings, where traditional models struggle to achieve effective performance (Agrawal et al., 2022).

Phenomenon Definition. Recently, Large Language Models (LLMs) like ChatGPT[2] have emerged as a fundamental backbone in the field of NLP. Their remarkable capability lies in achieving impressive performance without parameter tuning, relying instead on a limited number of example instructions. Hence, we also explore their potential in IE tasks. In this study, we uniformly define the IE task as the prediction of *A-B pairs* for the given textual data. "A" represents a pre-defined type label, while "B" refers to a single or multiple spans extracted directly from the text. Specifically, these pairs in RE, NER, and ED take the forms of *relation-(head entity, tail entity)*, *entity type-entity span*, and *event type-event trigger*, respectively. During our exploration, we discover an intriguing phenomenon, i.e., LLMs' spurious associations. This phenomenon reveals that the models are capable of correctly predict the answer "B" even when confronted with a different "A′" that is semantically unrelated to the original type label "A". As illustrated in Figure 1, the model successfully identifies the entity pair (`weak duration constraints`, `HMMs`) ("B"), even when provided with another relation like `limitations of` ("A′"), which is semantically unrelated to `feature of`

*Corresponding authors.
[1]The codes are publicly available at `https://github.com/TreMila/SaIE`
[2]`https://chat.openai.com/`

Findings of the Association for Computational Linguistics: ACL 2024, pages 9176–9190
August 11-16, 2024

("A"). This phenomenon is not limited to RE but is also observable in NER and ED tasks, as shown in Figure 1.

Phenomenon Origins. We take RE as an example to describe the process of discovering spurious association phenomena. We first feed a sentence and a pre-defined set of relations to ChatGPT, and ask the model to generate triplets in the form of (head entity, relation, tail entity). Both entities are derived from the provided sentence, and the relation originates from the relation set. When examining the error results, we observe that a significant portion of inaccuracies stems from generated relations that do not align with the pre-defined set. This is due to ChatGPT's generative nature, which sometimes generates relations that differ significantly in semantics from the intended original relations. Furthermore, in our attempts to employ these generated semantically unrelated relations for identifying the head and tail entities in other sentences with the original relation, we find that the large model can utilize these relations to effectively extract the correct head-tail entity pairs.

Phenomenon Application. We utilize the spurious association phenomenon to enhance the model performance in IE tasks. We still consider RE as an illustrative case. First, we select the Top-K (K=1 in the experiments) extended relations based on the highest F1 scores on the verification dataset from those semantically unrelated to the original relation, yet capable of accurately identifying the correct head and tail entity pairs. Then, we integrate them with all pre-defined original relations to create a new set of relations. This augmented set, along with each test sample, is then fed into the model. To facilitate the extraction process, we design Chain-of-Thought (CoT) prompts that guide the model to extract triplets from the text. An improvement in the quality of the extracted triplets, compared to those obtained without incorporating the extended relations, confirms the positive impact of spurious association on model performance.

To investigate the aforementioned intriguing phenomenon, we conduct a comprehensive set of experiments using LLMs of varying parameter sizes: ChatGLM (6B) (Du et al., 2022), BaiChuan (13B) (Yang et al., 2023), Alpaca (33B) (Taori et al., 2023), LLaMA-2 (70B) (Touvron et al., 2023), ChatGPT, and GPT-4 (OpenAI, 2023). These experiments contain diverse IE tasks: RE, NER, and ED, and are conducted on datasets in both Chinese and English languages. After experimental analysis, several significant conclusions have been drawn:

- **Finding 1:** Regardless of the size of LLMs, spurious associations occur in both Chinese and English datasets across the RE, NER, and ED tasks.
- **Finding 2:** The phenomenon of LLMs' spurious associations is more pronounced in IE tasks. Despite over 60% of extended labels differing from original labels, the model accurately predicts entity pairs, spans, or triggers associated with original labels using these extended labels.
- **Finding 3:** The semantic representations of labels in spurious associations are closer to those of the original labels compared to other extended labels.
- **Finding 4:** Extended labels prove to be valuable for enhancing the LLMs' performance on IE tasks. Notably, the model performance has improved by 9.55%, 11.42%, and 21.27% in terms of F1 scores on the SciERC (RE task), ACE05 (NER task), and DuEE (ED task) datasets, respectively.

2 Related Work

Related work of applying LLMs to IE tasks (Yu et al., 2023; Zhao et al., 2023; Wang et al., 2022; Xu et al., 2023) can be roughly divided into four categories: 1) directly employing LLMs for inference, 2) incorporating LLMs and small language models (SLMs), 3) leveraging SLMs with knowledge distilled from LLMs, and 4) utilizing LLMs with instruction tuning.

The first branch is to directly employ LLMs for inference (Min et al., 2022). Typical works along this line include ChatIE (Wei et al., 2023) and ChatEE (Gao et al., 2023). For example, ChatIE transforms the zero-shot IE task into a multi-turn question-answering problem with a two-stage framework. In this framework, the method is designed to first determine relations, entity types, or event types, and then to extract the corresponding entity pairs, entity spans, or triggers from the given text. The second branch is to incorporate LLMs and SLMs for the IE tasks. For instance, the filter-then-rerank (Ma et al., 2023) method is proposed, employing SLMs as filters and LLMs as

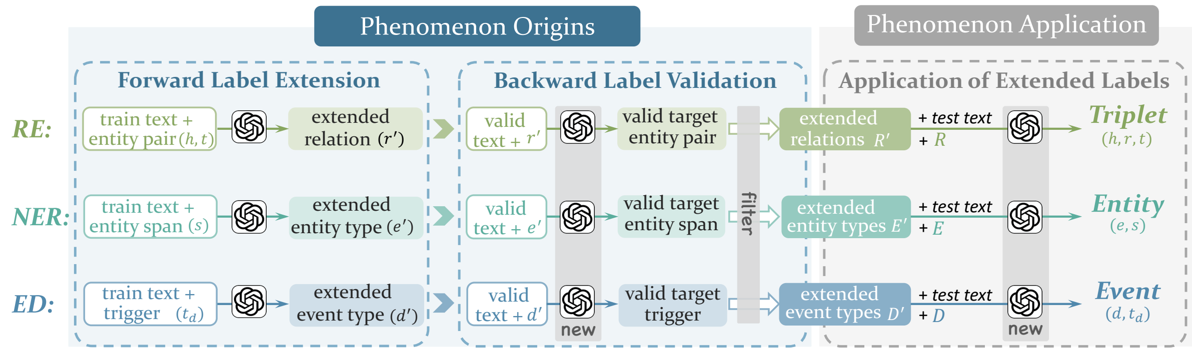

Figure 2: Our study framework is designed for the RE, NER, and ED tasks, covering both spurious association phenomenon origins and application. Taking the RE task as an example, the process begins with phenomenon origins. Given the text and entity pair in the training set, we first perform the forward label extension, generating an extended relation set. Then, we move to the backward label validation step, which involves the selection of accurate extended relations capable of extracting the entity pair aligned with the target relation within the validation set. Finally, in terms of phenomenon application, we use the refined extended relation set to enhance the LLMs' performance on the test set of the RE task.

rerankers. This is achieved by prompting LLMs to rerank a small subset of challenging samples identified by SLMs. The third branch is to use SLMs with knowledge distilled from LLMs for the tasks. This type of method regards LLMs as annotators and generates abundant samples with (pseudo) labels. Then, SLMs are trained using augmented data to achieve superior performance (Josifoski et al., 2023). The fourth branch is to use LLMs based on supervised instruction tuning. For example, InstructUIE (Wang et al., 2023) is a multi-task learning framework for universal IE which enables the use of human-readable instructions to guide LLMs for IE tasks.

In summary, the prevailing tendency is to employ large models for IE tasks, yet there remains room for performance enhancement. In this paper, we unveil an intriguing phenomenon, i.e., the spurious associations of LLMs, and leverage this discovery to improve the model's performance on IE tasks.

3 Study Design

In this section, we elaborate on the spurious association phenomenon in LLMs, including its definition, origins, and application.

3.1 Phenomenon Definition

Definition 1: RE task. Given a text $C_r = [c_1, c_2, \ldots, c_{n_r}]$ and the pre-defined relation types $\mathcal{R} = \{r_1, r_2, \ldots, r_{m_r}\}$, where n_r denotes the number of tokens in C_r and m_r is the number of relations in $\mathcal{R}$, RE task aims to obtain a triplet set $\mathcal{T} = \{(h, r, t)\}^m$ from C_r, where m is the number of extracted triplets and r represents the relation between the head entity h and tail entity t.

Definition 2: NER task. Given a text $C_e = [c_1, c_2, \ldots, c_{n_e}]$ and the pre-defined entity types $\mathcal{E} = \{e_1, e_2, \ldots, e_{m_e}\}$, NER task aims to detect the mention spans $\mathcal{S} = \{s_1, s_2, \ldots, s_{w_e}\}$ from C_e and the entity type $e \in \mathcal{E}$ (e.g., PERSON, LOCATION, etc) for each extracted span.

Definition 3: ED task. Given a text $C_d = [c_1, c_2, \ldots, c_{n_d}]$ and the pre-defined event types $\mathcal{D} = \{d_1, d_2, \ldots, d_{m_d}\}$, ED task aims to identify the event trigger t_d for C_d and the event type $d \in \mathcal{D}$ of t_d.

Definition 4: Spurious associations. In RE, for a training sample with C_r and (h, r, t), LLMs would predict (h, t) based on C_r and r', even if r' is semantically unrelated to r. Similarly, in NER, with a sample containing C_e, s, and e, LLMs would predict s using C_e and e', even when there is no semantic connection between e' and e. In ED, when considering a sample comprising C_d, t_d, and d, LLMs would predict t_d given C_d and d', even if d' lacks semantic relevance to d.

3.2 Phenomenon Origins

As illustrated in Figure 2, we describe the origins of the spurious association phenomenon for three tasks: RE, NER, and ED. Since the phenomenon exhibits a uniform pattern across these tasks, we take RE as an example to detail the process, which is structured into two steps: 1) Forward label extension (on the training set), utilizing an LLM to extend the pre-defined original relations, and 2) Backward label validation (on the validation set), The prompt of forward relation extension

Role definition

You are currently a senior relation generation expert.

Instruction

Your objective is to generate relation types for head-tail entity pairs within a given text, adhering to the following rules:
1. Utilize head-tail entity pairs, e.g., (noisy-channel architecture, transformation model).
2. Incorporate the context of the given text.
3. Generate relation types.

Demonstrations

The following is an example of instruction to help you understand the task:
Input: "We present a text mining method for finding synonymous expressions based on the distributional hypothesis in a set of coherent corpora. "*Head-tail entity pair*: (text mining method, synonymous expressions) "
Output: [USED-FOR]

Input: "Conventional HMMs have weak duration constraints." *Head-tail entity pair*: (weak duration constraints, HMMs)"

Figure 3: A prompt example in forward relation extension.

selecting the extended relations that can assist the model in precisely identifying head-tail entity pairs corresponding to the original relations.

3.2.1 Forward Label Extension

This step is designed to extend a new relation set R'_r for each $r \in \mathcal{R}$. Specifically, we select all samples with the original relation r from the training set. For each sample, we concatenate the sentence with its corresponding head-tail entity pair. This combination then serves as the input for the LLM, which is tasked with outputting a semantic relation (or "Na" if no relation is found) between the head-tail entity pair. To further enhance the model's potential, we incorporate role definitions, instructions, and demonstrations. An illustrative example of this process is provided in Figure 3. Through the above process, we obtain the extended set R'_r for the specific relation r. Notably, the relations extended by different $r \in R$ may be the same, such as $r_1 \rightarrow r'$ and $r_2 \rightarrow r'$. If these identical relations emerge within the set R' from which all relations are extended, it becomes challenging to ascertain their original corresponding relations. Hence, we eliminate these duplicate relations that appear across various extension sets for every $r \in R$, ensuring distinctiveness among the sets.

3.2.2 Backward Label Validation

In this step, we aim to evaluate the validity of each extended relation $r' \in R'_r$ derived from the original relation r. Specifically, we select all samples associated with r from the validation set. For each r', we concatenate the sentence from each selected

The prompt of backward relation validation

Role definition

You are currently a senior expert in relation extraction.

Instruction

Your objective is to extract head-tail entity pairs given a text and the specified relation, adhering to the following rules:
1. Utilize the context of the given text with the specified relation to extract head-tail entity pairs that represent the relation.
2. Verify the yes-no question sentence formed by concatenating the specified relation and the extracted head-tail entity pairs, and output only "yes" or "no".
3. Generate a list of triplet (head entity, relation, tail entity) based on the answer of yes-no question, where the relation must be the specified relation.

CoT Demonstrations

The following is an example of a chain of thought that guides you step by step to generate the results.
Input: "We propose and analyze a block minimization framework for data larger than the memory size." *Relation*: COMPARE

Head-tail entity pair: [(data, memory size)]
Answer: Is the relation between "data" and "memory size" "COMPARE"? Yes
Generate a list of triplet:
``
(data, COMPARE, memory size)
``

Input: "Synchronous dependency insertion grammars are a version of synchronous grammars defined on dependency trees." *Relation*: LIMITATIONS-OF

Figure 4: An example of the CoT prompt in backward relation validation.

sample with r' to form the input of the LLM and ask the model to generate the head-tail entity pair[3]. The model's response is considered correct if its outputs exactly match the ground truth pairs. Accuracy is defined as the proportion of consistent responses to the total count of ground truth pairs. With this approach, we compute the F1 score for the extended relation r' over all samples related to r. If the F1 score is zero, the extended relation r' is removed from R'_r. Otherwise, we retain it. Notably, in the LLM's input, in addition to the sentence and extended relation, we introduce the CoT process in the demonstration, as illustrated in Figure 4. That is, we first ask the model to produce the head-tail entity pair. Then, we integrate this output with the extended relation and request the model to assess whether these two entities exhibit this extended relation. The model is expected to generate results that align with real-world facts.

3.3 Phenomenon Application

To verify the impact of the extended relations, we incorporate them to enhance model performance on the RE task. Specifically, for each $r \in R$, we first select the Top-k extended relations from R'_r according to the F1 scores obtained in the validation

[3]Despite our requirement for the model to output entity pairs, the inherent generative nature of LLMs may lead to unexpected results, such as "Null". See Appendix E for the detailed analysis.

Table 1: The statistics of six datasets used for RE, NER, and ED. "#" denotes the number of samples in the specific dataset. Note that "*" indicates that the dataset is preprocessed. That is, we select 10 pre-defined relations from the original 44 provided by CMeIE and then divide the samples in the training and validation set into three subsets based on the selected relations in a ratio of 8:1:1.

Task	Dataset	Lang.	Type	# Train	# Valid	# Test
RE	SciERC	en	7	1366	187	397
	CMeIE*	zh	10*	8680*	1053*	1053*
NER	ACE05	en	7	7299	971	1060
	CMeEE	zh	9	15000	5000	3000
ED	CASIE	en	5	3751	778	1500
	DuEE	zh	9	11958	1498	3500

set. Next, we merge R with the selected extended relations of all relations in R. After this, we feed each sentence from the test set and the merged relation set into the LLM and design CoT to guide the LLM through the following steps to produce triplets. The model first identifies a set of head and tail entity pairs from the sentence. It then selects a relation for each pair from the provided relation set to form triplets. Finally, the model evaluates the reasonableness of each triplet. Only the triplets that are judged as reasonable are kept. The details of the prompt design are described in Appendix C.

4 Experiments

We conduct extensive experiments to demonstrate the phenomenon of LLMs' spurious associations in IE tasks. Then, we leverage this phenomenon to improve model performance in these tasks.

4.1 Experimental Setup

Datasets. To illustrate the universality of the phenomenon in IE tasks, we conduct experiments on six public datasets: SciERC (Luan et al., 2018) and CMeIE[4] (Guan et al., 2020) for RE, ACE05[5] (Walker et al., 2006) and CMeEE[4] (Zhang et al., 2022) for NER, CASIE (Satyapanich et al., 2020) and DuEE1.0 (Li et al., 2020) for ED. The statistics of these datasets are detailed in Table 1. Notably, for every relation, entity type, or event type, we select 100 training samples in forward label extension to ensure efficiency. The entity pairs/entity spans/triggers[6] are restricted to a single label in each sample. In addition, we utilize 10 and 20 validation samples for the English and Chinese datasets separately in backward label validation.

Models. We experiment LLMs with various parameter sizes, including ChatGLM (6B) (Du et al., 2022), BaiChuan (13B)[7] (Yang et al., 2023), Alpaca (33B) (Taori et al., 2023), LLaMA-2 (70B)[8] (Touvron et al., 2023), ChatGPT[9] and GPT-4[10] (OpenAI, 2023). Notably, for GPT-4, due to its higher access costs, we randomly selected half of the total samples from every dataset for our experiments. Our experiments are conducted on a workstation running Ubuntu 20.04.6 LTS, with two Intel(R) Xeon(R) Platinum 8336C CPUs, four NVIDIA A800 GPUs, and 1.0TiB of memory.

Evaluation metrics. Following the previous works (Wei et al., 2023; Zhang et al., 2023), we employ three standard evaluation metrics, i.e., micro Precision (P), Recall (R), and strict Micro-F1 score (F1). Notably, in RE, a triplet is considered correct only if the relation type, along with the types and the boundaries of the head-tail entities are precisely determined. In NER, only when both the span and the type of the predicted entity are accurately predicted, we consider it correct. In ED, an event is considered correct only if both the event trigger and event type are accurately identified.

4.2 Study Results

S1: Does the phenomenon of spurious associations manifest across various scales of LLMs? After performing forward label extension and backward label validation, we present the results in Table 2. We observe that: 1) *Spurious associations exist across various scales of LLMs in both Chinese and English datasets for the RE, NER, and ED tasks.* This is evident from the consistently higher DIS-T results. 2) The results from SIM-T illustrate that even when the extended relation/entity/event types closely resemble the original, the performance based on these labels is inferior to that of the dissimilar label (DIS-T). This phenomenon appears counterintuitive and the underlying reasons will be explored in future work. 3) In scenarios where the extended labels diverge from the original, it is notable that the count of extended labels

[4]https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414

[5]catalog.ldc.upenn.edu/LDC2006T06

[6]Notably, every entity pair/entity spans/triggers cannot have other relations/entity types/event types in the training sample. However, we observe that such instances are relatively rare in the IE datasets. Refer to Appendix A for more details.

[7]https://github.com/baichuan-inc/Baichuan2

[8]https://ai.meta.com/llama/

[9]gpt-3.5-turbo

[10]gpt-4-0314

Table 2: The results of spurious associations in LLMs for all original labels. "# S1O" means the count of relation, entity, or event type labels Output from Step 1. "SIM (DIS)" denotes extended labels that are similar (dissimilar) to the ground truth label through human annotators. "T" indicates that the predictions for entity pairs in RE, entities in NER, or triggers in ED are true. "F" denotes that the predictions for these same elements are false. "Count" represents the number of extended labels, and $Ratio = \frac{Count}{\#\ S1O}$. *The column shaded in light grey indicates the prevalence of the phenomenon of LLMs' spurious associations, and a higher value signifies a greater occurrence.* The reason LLaMA-2 is not applied to the Chinese datasets is due to the absence of a Chinese version for the model at present.

LLM	Task	Dataset	# S1O	SIM-T		SIM-F		DIS-T		DIS-F	
				Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)
ChatGLM (6B)	RE	SciERC	352	6	1.70	12	3.41	170	48.30	164	46.59
		CMeIE	255	13	5.10	3	1.18	181	70.98	58	22.74
	NER	ACE05	124	18	14.52	5	4.03	83	66.94	18	14.52
		CMeEE	431	35	8.12	5	1.16	335	77.73	56	12.99
	ED	CASIE	237	13	5.49	2	0.84	177	74.68	45	18.99
		DuEE	502	83	16.53	9	1.79	340	67.73	70	13.95
BaiChuan (13B)	RE	SciERC	733	32	4.37	33	4.50	392	53.48	276	37.65
		CMeIE	210	13	6.19	15	7.14	119	56.67	63	30.00
	NER	ACE05	188	33	17.55	17	9.04	88	46.81	50	26.60
		CMeEE	2358	105	4.45	9	0.38	1585	67.22	659	27.95
	ED	CASIE	335	61	18.21	2	0.60	247	73.73	25	7.46
		DuEE	653	163	24.96	13	1.99	460	70.45	17	2.60
Alpaca (33B)	RE	SciERC	1093	78	7.14	35	3.20	546	49.95	434	39.71
		CMeIE	698	48	6.87	2	0.00	623	89.25	25	3.58
	NER	ACE05	240	29	12.08	40	16.67	34	14.17	137	58.08
		CMeEE	1055	176	16.68	0	0.00	831	78.77	48	4.55
	ED	CASIE	374	53	14.17	0	0.00	296	79.15	25	6.68
		DuEE	1343	304	22.64	30	2.23	838	62.40	171	12.73
LLaMA-2 (70B)	RE	SciERC	380	48	12.63	0	0.00	312	82.11	20	5.26
		CMeIE	–	–	–	–	–	–	–	–	–
	NER	ACE05	171	51	29.83	5	2.92	83	48.54	32	18.71
		CMeEE	–	–	–	–	–	–	–	–	–
	ED	CASIE	154	9	5.84	0	0.00	143	92.86	2	1.30
		DuEE	–	–	–	–	–	–	–	–	–
ChatGPT	RE	SciERC	862	116	13.46	25	2.90	512	59.40	209	24.24
		CMeIE	510	121	23.72	0	0.00	312	61.18	77	15.10
	NER	ACE05	281	78	27.76	8	2.85	149	53.02	46	16.37
		CMeEE	597	158	26.47	10	1.67	354	59.30	75	12.56
	ED	CASIE	271	23	8.49	0	0.00	236	87.08	12	4.43
		DuEE	870	218	25.06	0	0.00	650	74.71	2	0.23
GPT-4	RE	SciERC	122	15	12.30	6	4.92	54	44.26	47	38.52
		CMeIE	80	20	25.00	0	0.00	47	58.75	13	16.25
	NER	ACE05	65	8	12.31	2	3.08	30	46.15	25	38.46
		CMeEE	210	42	20.00	5	2.38	154	73.33	9	4.29
	ED	CASIE	66	9	13.64	0	0.00	52	78.79	5	7.57
		DuEE	99	20	20.20	2	2.02	71	71.72	6	6.06

associated with accurate predictions significantly surpasses the count linked to inaccurate predictions. This shows that most of the labels extended and validated from the LLM are effective for identifying head-tail entities/entity spans/triggers.

S2: What is the extent of LLMs' spurious associations? To analyze the extent of LLMs' spurious associations, we conduct the following experiments, continuing to employ the RE task as an example. First, we extract triplets from the samples in the test set using every relation extended by the validation set. Then, we retain the extended relation if at least one accurately extracted triplet is found in all results generated by the samples associated with this original relation. Finally, we ask the previous human annotators to determine

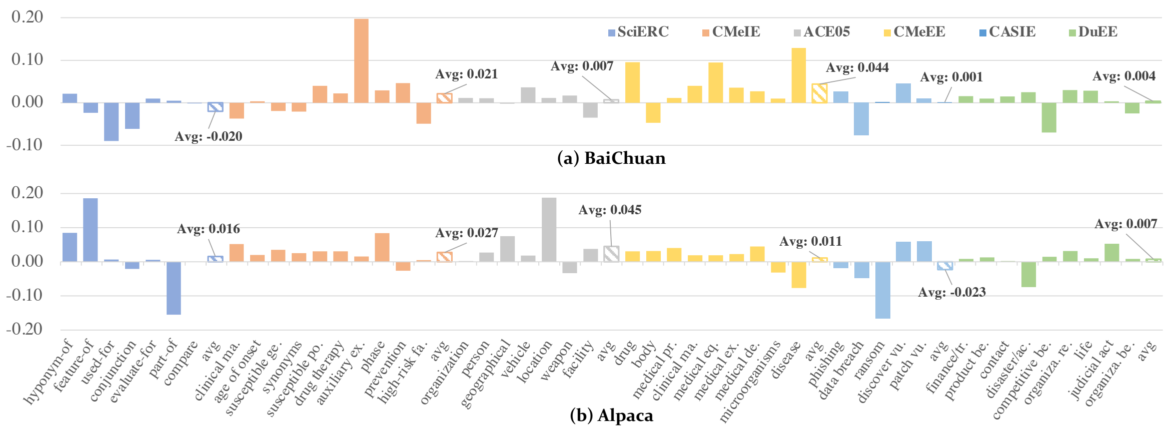

Figure 5: Similarities between the original label and different types of extended labels under the same context. Avg denotes the mean value of $Sim_{D_C} - Sim_{V_T}$ across all pre-defined type labels in the task.

Table 3: Analysis of the extent of LLMs' spurious association. "V" refers to the set of extended relations produced from the Validation set. "T" indicates the set of extended relations in V that yield at least one correct output in the Test set. D_C denotes the number of relations in T that diverge from the ground truths judged by human annotators. $D_R = \frac{\# D_C}{\# T}$. *A higher D_R indicates a more prominent occurrence of LLMs' spurious association phenomenon.* Due to space limitations, the detailed results of ACE05, CMeEE, CASIE, and DuEE are reported in the Appendix B.

Dataset	**Original** Relation	ChatGPT			
		# V	# T	# D_C	D_R (%)
SciERC	feature-of	17	10	9	90.00
	hyponym-of	63	52	40	76.92
	conjunction	83	30	21	70.00
	part-of	62	12	10	83.33
	used-for	319	318	273	85.85
	compare	55	55	48	87.27
	evaluate-for	29	12	8	66.67
	All	628	489	409	83.64
CMeIE	synonyms	44	29	20	68.97
	clinical manifestations	66	62	35	56.45
	age of onset	21	18	8	44.44
	high-risk factor	41	41	27	65.85
	susceptible population	95	95	73	76.84
	prevention	33	33	26	78.79
	auxiliary examination	27	27	23	85.19
	drug therapy	14	14	14	100.00
	susceptible gender	44	44	33	75.00
	phase	48	47	33	70.21
	All	433	410	292	71.22
ACE05	All	241	224	143	63.84
CMeEE	All	512	512	354	69.14
CASIE	All	271	257	234	91.05
DuEE	All	868	852	657	77.11

whether the pre-defined relation aligns with the remaining corresponding relations in semantics. The presence of a significant number of extended relations that semantically diverge from the original relations indicates the substantial extent of the phenomenon. The experimental results are shown in Table 3. By analyzing the results, we notice that a significant portion of the valid extended labels chosen in the test set are considered dissimilar to the pre-defined labels by human annotators. This observation highlights the noticeable presence of spurious associations in LLMs. In particular, the overall ratio of spurious associations for ChatGPT consistently exceeds 60% across the six datasets.

***S3: How relevant are extension labels and original* labels in specific contexts?** We design the experiments as follows: First, the similarity between the original label and the three extended labels randomly selected from D_C (referenced in Table 3) respectively, which can accurately predict the results on the validation set but are regarded as dissimilar to the original label, is calculated using the same text. The mean of these three similarity scores is denoted as Sim_{D_C}. Second, the similarity assessment is repeated for the original label against three extended labels randomly selected from V-T (Table 3) respectively, which incorrectly predict the results and are considered dissimilar, using the same textual content. The average of these scores is recorded as Sim_{V-T}. In cases where there are fewer than three labels, additional labels are randomly selected from those extended by the validation set to complete the set of three. The difference, $Sim_{D_C} - Sim_{V_T}$, is then calculated for each predefined label and the results are illustrated in Figure 5. We observe that for the extended labels considered dissimilar to pre-defined labels by humans, the Sim_{D_C} for most labels correctly

Table 4: Application of extended labels on the test sets. △ represents the results of our method minus the results of the baseline with the highest F1 score. Due to space limitations, detailed experimental results for ACE05, CMeEE, CASIE, and DuEE are provided in Appendix D. The Top-1 extended label for each original label used in our method is provided in Appendix F.

Test Sets	Original Label			Definition			Paraphrase			Our Method			△		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P (%)	R (%)	F1(%)
# SciERC															
feature-of	0.00	0.00	0.00	5.26	33.33	9.09	6.67	33.33	11.11	8.70	25.00	12.77	+2.03	-8.33	+1.66
hyponym-of	2.13	3.85	2.74	1.10	3.85	1.71	6.02	19.23	9.17	12.12	15.38	13.56	+6.10	-3.85	+4.39
conjunction	8.89	12.50	10.39	5.49	15.63	8.13	6.74	18.75	9.92	13.33	18.75	14.81	+4.44	+6.25	+4.43
part-of	0.00	0.00	0.00	1.23	6.25	2.06	1.28	6.25	2.13	11.54	31.25	14.29	+10.26	+25.00	+12.16
used-for	18.97	18.64	18.80	13.75	18.64	15.83	14.49	16.95	15.63	39.29	28.81	30.48	+20.32	+10.17	+11.67
compare	32.00	72.73	44.44	18.60	72.73	29.63	14.71	45.45	22.22	45.45	72.73	51.85	+13.45	+0.00	+7.41
evaluate-for	10.00	8.70	9.30	1.14	4.35	1.80	5.26	8.70	6.56	20.00	17.39	18.60	+10.00	+8.70	+9.30
overall evaluation	14.53	9.85	11.74	17.32	5.64	8.50	18.44	7.32	10.48	24.02	19.11	21.29	+9.50	+9.26	+9.55
# CMeIE															
synonyms	29.41	27.78	28.57	28.57	25.00	26.67	13.64	12.50	13.04	16.67	16.67	16.67	-12.75	-11.11	-11.90
clinical manifestations	46.81	57.89	51.76	19.67	31.58	24.24	52.00	68.42	59.09	55.00	68.42	56.82	+3.00	+0.00	-2.27
age of onset	31.25	45.45	37.04	33.33	63.64	43.75	40.00	54.55	46.15	50.00	63.64	56.00	+10.00	+9.09	+9.85
high-risk factor	29.03	60.00	39.13	23.53	53.33	32.65	28.13	60.00	38.30	61.90	86.67	72.22	+32.87	+26.67	+33.09
susceptible population	40.00	54.55	46.15	42.86	54.55	48.00	46.15	54.55	50.00	53.85	72.73	61.54	+7.70	+18.18	+11.54
prevention	19.05	40.00	25.81	23.81	50.00	32.26	23.81	50.00	32.26	31.58	60.00	41.38	+7.77	+10.00	+9.12
auxiliary examination	38.10	80.00	51.61	37.50	60.00	46.15	30.00	60.00	40.00	46.15	80.00	52.17	+8.06	0.00	+0.56
drug therapy	35.00	41.18	37.84	17.65	17.65	17.65	42.86	52.94	47.37	47.62	58.82	52.63	+4.76	+5.88	+5.26
susceptible gender	41.18	63.64	50.00	53.85	63.64	58.33	42.11	72.73	53.33	50.00	72.73	57.14	-3.85	+9.09	-1.19
phase	52.94	45.00	48.65	33.33	40.00	36.36	24.24	40.00	30.19	46.67	35.00	40.00	-6.27	-10.00	-8.65
overall evaluation	50.93	36.94	42.82	40.72	28.10	33.25	51.50	34.96	41.65	57.76	45.81	51.10	+6.83	+8.88	+8.28
# ACE05	49.54	60.00	54.27	53.08	51.89	49.81	43.12	51.09	46.77	55.05	68.97	61.22	+1.97	+17.08	+11.42
# CMeEE	63.98	35.10	45.33	63.98	53.60	58.33	72.04	57.26	63.81	81.18	65.09	72.25	+9.14	+7.82	+8.44
# CASIE	76.00	43.68	55.47	64.00	41.03	50.00	68.00	49.28	57.14	82.00	62.12	70.69	+14.00	+12.85	+13.55
# DuEE	77.45	39.50	52.32	72.55	42.29	53.43	81.37	42.13	55.52	84.31	70.49	76.79	+2.94	+28.36	+21.27

predicted by the model is higher than the Sim_{V-T} for those incorrectly predicted. This suggests that the labels in D_C are closer in vector space to the original labels than the labels in V-T.

S4: Do the extended labels improve the model performance on the test set? To further evaluate the usefulness of the extended labels, we incorporate the Top-1 extended label of each type into the pre-defined set of all types to enhance the model performance on the test set (refer to Section 3.3). In this experiment, we design three baselines. The first one considers the text and all pre-defined types as the input, and the model predicts the results (triplets in IE, entity and its type in NER, and trigger and its type in ED). The second baseline adds the type definition derived from GPT-4 based on the first baseline. The third baseline is to use GPT-4 to paraphrase the pre-defined type also based on the first baseline. The experimental results are listed in Table 4. We observe that our method consistently outperforms all baselines across all datasets and metrics in the overall evaluation, which illustrates the effectiveness of our extended labels. In particular, compared with the baselines, our method achieves a substantial improvement in F1 score by 9.55%, 11.42%, and 21.27% on the SciERC (RE task), ACE05 (NER task), and DuEE (ED task) datasets, respectively. In addition, our method outperforms the baseline in terms of P, R, and F1 on most relationship/entity/event types. However, there are also cases where the extraction results based on extended labels are inferior to those produced by baselines, such as `synonyms` and `phase` in CMeIE.

5 Conclusion

In this paper, we observe an intriguing phenomenon: LLMs' spurious associations, when utilizing the LLM-based method for accomplishing IE tasks. To explore this phenomenon, we design two strategies in this study, including forward label extension and backward label validation. Moreover, we leverage these extended labels to enhance model performance. Following the procedures described, we conduct extensive experiments to validate this intriguing phenomenon of LLMs with varying parameter sizes. Furthermore, we perform experiments on downstream tasks, confirming that the extended labels have a positive impact on all IE sub-tasks.

Limitations

This study focuses on discovering the phenomenon of spurious associations in LLMs and utilizing this insight to enhance the model's performance in IE tasks. However, it's crucial to acknowledge a limitation: we do not provide an in-depth analysis of the underlying causes of this phenomenon. This limitation stems from the inherent black-box nature of LLMs. Therefore, we identify the exploration of the causes as a topic for future research. In addition, the phenomenon we discovered is also limited to the tasks that can be characterized as the A-B pair prediction problem, as described in the Introduction.

Acknowledgments

We would like to thank the anonymous reviewers for their excellent feedback. This work was supported by the National Natural Science Foundation of China (No. 62306112), the Shanghai Sailing Program (No. 23YF1409400), and the Shanghai Pilot Program for Basic Research (No. 22TQ1400100-20).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Tongfeng Guan, Hongying Zan, Xiabing Zhou, Hongfei Xu, and Kunli Zhang. 2020. Cmeie: Construction and evaluation of chinese medical information extraction dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 270–282. Springer.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

OpenAI. 2023. Gpt-4 technical report.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8749–8757.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11285–11293.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

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*.

Zhipeng Xie and Yumin Tu. 2022. A graph convolutional network with adaptive graph generation and channel selection for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11522–11529.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? *arXiv preprint arXiv:2305.01555*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915.

Weiyan Zhang, Chuang Chen, Jiacheng Wang, Jingping Liu, and Tong Ruan. 2023. A co-adaptive duality-aware framework for biomedical relation extraction. *Bioinformatics*, 39(5):btad301.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Exploring Multi-Label Situations

In Footnote 5, we ignore the situation where an entity pair corresponds to multiple relations in a sample due to its infrequent occurrence. To prove this statement, we quantify the instances of this situation in the RE datasets. Similarly, we perform similar operations on the NER and ED datasets. The results are listed in Table 5. From the table, we observe that the number of samples that meet the above situation is very small, less than 1%. This shows that the strategy we adopted is reasonable.

Table 5: Statistics of the six datasets in a multi-label scenario. “Count” refers to the number of samples in the dataset. “M-Count” represents the count of samples where an entity pair (entity span or trigger) in the sample corresponds to multiple relations (entity types or event types). Ratio $= \frac{\text{M-Count}}{\text{Count}}$.

Taks	Datasets	Types	Count	in single-sample	
				M-Count	Ratio (%)
RE	SciREC	train	1366	0	0.00
		valid	187	0	0.00
	CMeIE	train	8680	18	0.21
		valid	1053	2	0.19
NER	ACE05	train	7299	12	0.16
		valid	971	4	0.41
	CMeEE	train	15000	67	0.45
		valid	5000	17	0.34
ED	CASIE	train	3571	0	0.00
		valid	788	0	0.00
	DuEE	train	11958	1	0.01
		valid	1498	0	0.00

B Detailed Results of Table 3

This section reports the detailed results of Table 3 for ACE05, CMeEE, CASIE, and DuEE. The results are listed in Tables 6 and 7. We notice that the model utilizing extended labels effectively extracts the same results as the model employing actual labels in both NER and ED tasks. However, the annotation experts consider these two labels distinct, as indicated by a relatively high dissimilarity ratio. In particular, the overall ratios of unexpected associations for ChatGPT across the ACE05, CMeEE, CASIE, and DuEE datasets stand at approximately 64%, 69%, 91%, and 77%, respectively.

C Prompt Details in Application

In this section, we provide an overview of the prompts used in Section 3.3. Taking RE as an example, the prompt is outlined in Figure 6. To enhance model performance, we augment the inputs

Table 6: Analysis of the extent of LLMs’ spurious association on NER task. “V” refers to the set of extended entity types produced from the Validation set. “T” indicates the set of extended entity types in V that yield at least one correct output in the Test set. D_C denotes the number of extended entity types in T that diverge from the ground truths judged by human annotators. $\mathrm{D}_R = \frac{\#\,\mathrm{D}_C}{\#\,\mathrm{T}}$. *A higher D_R indicates a more prominent occurrence of LLMs’ spurious association phenomenon.*

Dataset	Original Entity Type	ChatGPT			
		# T	# V	D_C	D_R (%)
ACE05	facility	32	30	18	60.00
	geographical soci.	27	27	13	48.15
	vehicle	25	25	12	48.00
	weapon	23	23	16	69.57
	organization	33	33	14	42.42
	person	79	68	58	85.30
	location	22	18	12	66.67
	All	241	224	143	63.84
CMeEE	medical department	31	31	27	87.10
	medical procedure	80	80	52	65.00
	body	127	127	87	68.50
	medical examinations	49	49	28	57.14
	medical equipment	36	36	25	69.44
	disease	37	37	24	64.86
	microorganisms	49	49	41	83.67
	clinical manifestations	71	71	53	74.65
	drug	32	32	17	53.13
	All	512	512	354	69.14

Table 7: Analysis of the extent of LLMs’ spurious association on ED task. “V” refers to the set of extended event types produced from the Validation set. “T” indicates the set of extended event types in V that yield at least one correct output in the Test set. D_C denotes the number of extended event types in T that diverge from the ground truths judged by human annotators. $\mathrm{D}_R = \frac{\#\,\mathrm{D}_C}{\#\,\mathrm{T}}$. *A higher D_R indicates a more prominent occurrence of LLMs’ spurious association phenomenon.*

Dataset	Original Event Type	ChatGPT			
		# V	# T	D_C	D_R (%)
CASIE	phishing	50	47	42	89.36
	data breach	32	28	26	92.86
	ransom	41	40	36	90.00
	patch vulnerability	55	51	45	88.24
	discover vulnerability	93	91	85	93.41
	All	271	257	234	91.05
DuEE	product behavior	102	102	65	63.73
	judicial act	126	126	72	57.14
	life	129	128	121	94.53
	organizational behavior	70	70	53	75.71
	organizational relation	67	67	54	80.60
	competitive behavior	127	112	102	91.07
	contact	81	81	61	75.31
	finance/trading	85	85	71	83.53
	disaster/accident	81	81	58	71.60
	All	868	852	657	77.11

Table 8: Results of the application of extended labels on the NER task. △ represents the results of our method minus the results of the baseline with the highest F1 score.

Test Sets	Original Label			Definition			Paraphrase			Our Method			△		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P (%)	R (%)	F1(%)
# ACE05															
facility	38.46	31.25	34.48	40.00	25.00	30.77	12.50	6.25	8.33	63.64	50.00	51.85	+25.17	+18.75	+17.37
geographical soci.	63.64	73.68	68.29	83.33	52.63	64.52	60.87	73.68	66.67	90.00	63.16	64.52	+26.36	-10.53	-3.78
vehicle	87.50	50.00	63.64	70.00	50.00	58.33	35.29	42.86	38.71	90.00	64.29	75.00	+2.50	+14.29	+11.36
weapon	63.64	58.33	60.87	52.00	68.42	59.09	46.67	58.33	51.85	90.00	91.67	81.82	+26.36	+33.33	+20.95
organization	75.00	69.23	72.00	38.10	80.00	51.61	77.78	53.85	63.64	66.67	69.23	62.07	-8.33	+0.00	-9.93
person	60.00	24.00	34.29	50.00	24.00	32.43	70.00	28.00	40.00	100.00	44.00	51.28	+30.00	+16.00	+11.28
location	42.86	60.00	50.00	41.67	50.00	45.45	50.00	50.00	50.00	45.45	60.00	50.00	+2.60	+0.00	+0.00
overall evaluation	49.54	60.00	54.27	53.08	51.89	49.81	43.12	51.09	46.77	55.05	68.97	61.22	+1.97	+17.08	+11.42
# CMeEE															
medical department	7.55	34.78	12.40	41.03	69.57	51.61	48.39	65.22	55.56	50.00	73.91	58.62	+1.61	+8.70	+3.07
medical procedure	21.43	42.86	28.57	40.74	78.57	53.66	34.48	71.43	46.51	55.00	78.57	64.71	+14.26	+0.00	+11.05
body	41.46	65.38	50.75	42.86	57.69	49.18	42.55	76.92	54.79	80.00	88.46	76.00	+37.45	+11.54	+21.21
medical examinations	39.02	57.14	46.38	43.48	35.71	39.22	60.71	60.71	60.71	76.92	89.29	74.07	+16.21	+28.57	+13.36
medical equipment	43.48	66.67	52.63	66.67	66.67	66.67	76.92	66.67	71.43	76.92	93.33	74.29	+0.00	+26.67	+2.86
disease	58.82	80.00	67.80	59.26	64.00	61.54	78.26	72.00	75.00	75.00	96.00	75.00	-3.26	+24.00	+0.00
microorganisms	62.50	75.00	68.18	61.90	86.67	72.22	81.25	65.00	72.22	91.67	90.00	75.00	+10.42	+25.00	+2.78
clinical manifestations	53.13	70.83	60.71	53.57	62.50	57.69	55.56	83.33	66.67	60.87	100.00	68.85	+5.31	+16.67	+2.19
drug	100.0(	90.91	95.24	100.00	90.91	95.24	100.00	90.91	95.24	100.00	100.00	100.00	+0.00	+9.09	+4.76
overall evaluation	63.98	35.10	45.33	63.98	53.60	58.33	72.04	57.26	63.81	81.18	65.09	72.25	+9.14	+7.82	+8.44

The prompt of relation extraction

Role definition

You are currently a senior expert in relation extraction.

Instruction

Your objective is to extract triplets given a text and a list of relation, adhering to the following rules:
1. Generate a keyword pair list from the given text
2.Extract potential relations for keyword pairs from the given relation list
3. Verify the yes-no question sentence formed by concatenating the potential relations and the extracted keyword pairs, and output only "yes" or "no".
4. Generate a list of triplet (head entity, relation, tail entity) based on the answer of yes-no question, where the relation must be in the given relation list {extend_label}.

CoT Demonstrations

The following is an example of a chain of thought that guides you step by step to generate the results.
Input: We present a text mining method for finding synonymous expressions based on the distributional hypothesis in a set of corpora.

Intermediate_keyword pair: [(text mining method, synonymous expressions), (distributional hypothesis, text mining method)]
Intermediate_relation: [USED-FOR, EVALUATE-FOR]
Answer:
Is the relation between "text mining method" and "synonymous expressions" the "USED-FOR"? yes
Is the relation between "distributional hypothesis" and "text mining method" the "USED-FOR"? Yes
Is the relation between "text mining method" and "synonymous expressions" the "EVALUATE-FOR R"? no
Is the relation between "distributional hypothesis" and "text mining method" the "EVALUATE-FOR "? no

Generate a list of triplet:
``
(text mining method, USED-FOR, synonymous expressions)
(distributional hypothesis, USED-FOR, text mining method)
``

Input: "An entity-oriented approach to restricted-domain parsing is proposed."

Figure 6: The prompt for the application of extended relations on the RE task.

with role definition, instruction, and demonstration. It should be noted that we introduce CoT in the demonstration. That is, we first ask the model to produce the keyword pairs. Then, based on the keyword pairs, we instruct the model to identify potential relations from the given relation list. Next, we integrate the keyword pairs with the identified relations and ask the LLM to assess the factual accuracy of these three elements. Finally, the model retains the correct factual triplets as outputs. Note that in our prompt, we employ the term "keyword pair" instead of directly utilizing "entity pairs". This strategy aims to stimulate ChatGPT to generate more candidate subject-object pairs in the initial step, effectively enhancing recall.

D Detailed Results of Table 4

This section presents the detailed results of Table 4 for ACE05, CMeEE, CASIE, and DuEE. These results are shown in Tables 8 and 9. We notice that the model with our extended labels consistently outperforms the competitors, indicating the effectiveness of these labels. In particular, the model performance has improved by 11.42%, 8.44%, 13.55%, and 21.27% on ACE05 (NER task), CMeEE (NER task), CASIE (ED task), and DuEE (ED task) datasets, respectively. In addition, even when the F1 based on the original labels (such as `location`) is 0, the model optimized by extended labels also can extract the correct results.

Table 9: Results of the application of extended labels on the ED task. △ represents the results of our method minus the results of the baseline with the highest F1 score.

Test Sets	Original Label			Definition			Paraphrase			Our Method			△		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P (%)	R (%)	F1(%)
# CASIE															
phishing	45.00	90.00	60.00	42.86	60.00	50.00	35.29	60.00	44.44	47.37	90.00	62.07	+2.37	+0.00	+2.07
data breach	47.06	80.00	59.26	46.15	60.00	52.17	75.00	90.00	81.82	66.67	90.00	72.73	-8.33	0.00	-9.09
ransom	81.82	90.00	85.71	88.89	80.00	84.21	66.67	80.00	72.73	80.00	80.00	80.00	-1.82	-10.00	-5.71
patch vulnerability	29.41	50.00	37.04	29.41	50.00	37.04	35.71	50.00	41.67	58.33	80.00	63.64	+22.62	+30.00	+21.97
discover vulnerability	31.82	70.00	43.75	28.00	70.00	40.00	42.86	60.00	50.00	70.00	90.00	78.26	+27.14	+30.00	+28.26
overall evaluation	76.00	43.68	55.47	64.00	41.03	50.00	68.00	49.28	57.14	82.00	62.12	70.69	+14.00	+12.85	+13.55
# DuEE															
product behavior	66.67	100.00	80.00	81.82	90.00	85.71	71.43	100.00	83.33	90.00	100.00	90.00	+8.18	+10.00	+4.29
judicial act	40.91	75.00	52.94	34.62	75.00	47.37	32.14	75.00	45.00	71.43	100.00	82.76	+36.81	+25.00	+35.39
life	30.00	69.23	41.86	28.57	46.15	35.29	42.86	69.23	52.94	53.85	69.23	53.85	+10.99	+0.00	+0.90
organizational behavior	52.94	81.82	64.29	44.44	72.73	55.17	47.62	90.91	62.50	90.00	90.91	86.96	+37.06	+9.09	+22.67
organizational relation	33.33	66.67	44.44	44.44	100.00	61.54	46.15	100.00	63.16	100.00	100.00	100.00	+53.85	+0.00	+36.84
competitive behavior	21.21	58.33	31.11	20.83	41.67	27.78	14.29	33.33	20.00	61.54	75.00	66.67	+40.33	+16.67	+35.56
contact	55.56	100.00	71.43	64.29	90.00	75.00	52.63	100.00	68.97	75.00	100.00	81.82	+10.71	+10.00	+6.82
finance/trading	50.00	90.00	64.29	50.00	90.00	64.29	50.00	100.00	66.67	66.67	100.00	80.00	+16.67	+0.00	+13.33
disaster/accident	34.78	66.67	45.71	43.75	58.33	50.00	45.00	75.00	56.25	50.00	83.33	57.14	+5.00	+8.33	+0.89
overall evaluation	77.45	39.50	52.32	72.55	42.29	53.43	81.37	42.13	55.52	84.31	70.49	76.79	+2.94	+28.36	+21.27

Table 10: Unexpected outputs in backward label validation on RE task. "Tot." represents the total number of samples in the output. "Un_1" and "Un_2" denote the number of samples where the model output is empty and inconsistent with the expected format, respectively. $\text{R}_1 = \frac{\text{Un}_1}{\text{Tot.}}$ and $\text{R}_2 = \frac{\text{Un}_2}{\text{Tot.}}$.

Original Relations	Tot.	Un_1	R_1 (%)	Un_2	R_2 (%)
SciERC					
feature-of	1130	458	40.53	23	2.04
hyponym-of	890	330	37.08	8	0.90
conjunction	1040	487	46.83	7	0.67
part-of	790	277	35.06	9	1.14
used-for	3460	1795	51.88	28	0.81
compare	580	207	35.69	3	0.52
evaluate-for	730	189	25.89	5	0.68
All	8620	3743	43.42	83	0.96
CMeIE					
synonyms	500	14	2.80	48	9.60
clinical manifestations	920	21	2.28	343	37.28
age of onset	250	5	2.00	26	10.40
high-risk factor	500	10	2.00	81	16.20
susceptible population	950	9	0.95	30	3.16
prevention	410	12	2.93	92	22.44
auxiliary examination	290	7	2.41	56	19.31
drug therapy	150	7	4.67	48	32.00
susceptible gender	490	29	5.92	81	16.53
phase	640	29	4.53	122	19.06
All	5100	143	2.80	927	18.18

Table 11: Unexpected outputs in backward label validation on NER task. "Tot." represents the total number of samples in the output. "Un_1" and "Un_2" denote the number of samples where the model output is empty and inconsistent with the expected format, respectively. $\text{R}_1 = \frac{\text{Un}_1}{\text{Tot.}}$ and $\text{R}_2 = \frac{\text{Un}_2}{\text{Tot.}}$.

Original Entity Types	Tot.	Un_1	R_1 (%)	Un_2	R_2 (%)
ACE05					
facility	340	36	10.59	57	16.76
geographical soci.	250	21	8.40	36	14.40
vehicle	340	67	19.71	40	11.76
weapon	240	12	5.00	18	7.50
organization	370	68	18.38	59	15.95
person	1010	304	30.10	167	16.53
location	260	32	12.31	29	11.15
All	2810	540	19.22	406	14.45
CMeEE					
medical department	360	53	14.72	68	18.89
medical procedure	870	162	18.62	252	28.97
body	1610	368	22.86	393	24.41
medical examinations	520	71	13.65	79	15.19
medical equipment	360	81	22.50	63	17.50
disease	480	111	23.13	118	24.58
microorganisms	550	90	16.36	104	18.91
clinical manifestations	890	206	23.15	199	22.36
drug	330	43	13.03	49	14.85
All	5970	1185	19.85	1325	22.19

However, there are also cases where the extraction results based on extended labels are inferior to those baselines, such as `geographical social political` and `organization` in ACE05, `data breach` and `ransom` in CASIE.

E Unexpected Outputs in Backward Label Validation

In footnote 3, we mention that, despite our requirement for the model to output entity pairs, the inherent generative nature of LLMs may lead to unexpected results. These unexpected results include two aspects: 1) LLMs would generate an empty

Table 12: Unexpected outputs in backward label validation on ED task. "Tot." represents the total number of samples in the output. "Un_1" and "Un_2" denote the number of samples where the model output is empty and inconsistent with the expected format, respectively. $R_1 = \frac{\text{Un}_1}{\text{Tot.}}$ and $R_2 = \frac{\text{Un}_2}{\text{Tot.}}$.

Original Event Types	Tot.	Un_1	R_1 (%)	Un_2	R_2 (%)
		CASIE			
phishing	500	11	2.20	3	0.60
data breach	320	8	2.50	38	11.88
ransom	410	12	2.93	6	1.46
patch vulnerability	610	17	2.79	6	0.98
discover vulnerability	870	69	7.93	5	0.57
All	2710	117	4.32	58	2.14
		DuEE			
product behavior	1020	1	0.10	0	0.00
judicial act	1260	0	0.00	0	0.00
life	1300	5	0.38	1	0.08
organizational behavior	700	3	0.43	0	0.00
organizational relation	670	0	0.00	0	0.00
competitive behavior	1270	6	0.47	0	0.00
contact	820	0	0.00	1	0.12
finance/trading	850	5	0.59	1	0.12
disaster/accident	810	13	1.60	0	0.00
All	8700	33	0.38	3	0.03

output, and 2) LLMs would fail to produce output in the expected format. We take ChatGPT as an example to provide statistics on these two situations in Tables 10, 11, and 12. The results reveal that regardless of the RE, NER, or ED datasets, both situations are present, with irregular proportions. Moreover, combining Table 2 and Table 4, we conclude that the spurious association phenomenon of LLMs and the positive effect of extended labels on downstream tasks remain unaffected by these situations.

F Detailed Results of Extended Labels

In this section, we provide the three extended labels with the highest F1 scores for each pre-defined type during the application of the extended labels. The results for the RE, NER, and ED tasks are presented in Tables 13, 14, and 15, respectively. To facilitate better understanding, we also provide the original Chinese words for the extended labels on the three Chinese datasets CMeIE, CMeEE, and DuEE at the URL https://github.com/TreMila/SaIE.

Table 13: Three extended labels with the highest F1 in the application of RE task.

Relation Types	Extended Relation Labels	F1 (%)
	SciERC	
hyponym-of	version-of	13.56
	instance-of	10.96
	exemplify	10.39
feature-of	modifier-of	12.77
	embedded-in	11.43
	attribute	11.11
used-for	applied-to	30.48
	applies-to	29.47
	use_as	28.57
conjunction	coordination	14.81
	sequence	12.90
	coordinate	9.52
evaluate-for	measure-of	18.60
	result-from	12.24
	indicator-of	9.09
part-of	additional constraint	14.29
	expand/extend	12.05
	combine-and	11.90
compare	outperform	51.85
	outperforms	45.45
	comparison/contrast	43.24
	CMeIE	
clinical manifestations	possible symptoms	56.82
	common symptoms	56.41
	accompanying symptoms	55.81
age of onset	predisposing age	56.00
	onset time	48.28
	time of occurrence	43.75
susceptible gender	sex differences in onset	57.14
	incidence sex ratio	56.00
	disease gender bias	56.00
synonyms	analogy	16.67
	subclass relationship	16.22
	disease_alias	15.79
susceptible population	disease onset age	61.54
	risk of disease	58.33
	incidence group	50.00
drug therapy	treatment measures	52.63
	treatment programs	50.00
	treatment equipment	43.90
auxiliary examination	diagnosis methods	52.17
	confirmation methods	51.85
	check for complications	51.61
phase	disease level	40.00
	duration	30.77
	symptoms/manifestations	30.00
prevention	substitute	41.38
	prevention/treatment	40.00
	slow down progress	36.36
high-risk factor	susceptible groups	72.22
	uncertain relevance	56.41
	comorbidities	55.56

Table 14: Three extended labels with the highest F1 in the application of NER task.

Entity Types	Extended Entity Labels	F1 (%)
	ACE05	
organization	group	62.07
	governmental organization	60.00
	sports team	59.26
person	living_being	51.28
	person/organization	47.37
	entity	44.44
geographical soci.	geopolitical location	64.52
	geopolitical entity	64.52
	other	63.16
vehicle	type or vehicle	75.00
	transportation	75.00
	machine or equipment	69.57
location	geographic_area	50.00
	geographic location	47.62
	geographical entity	45.45
weapon	weapon category	81.82
	weapon_type	81.82
	weapon/tool	72.00
facility	sentence	51.85
	physical object	50.00
	infrastructure	46.67
	CMeEE	
drug	substance	100.00
	brand	100.00
	medicinal	100.00
body	body parts	76.00
	parts	75.47
	human organs	73.33
medical procedure	medical behavior	64.71
	route of administration	60.61
	dosing method	58.82
clinical manifestations	abnormal behavior	68.85
	indicator results	66.67
	phenomenon	66.67
medical equipment	instrument	74.29
	laboratory equipment	71.43
	infrastructure	70.97
medical examinations	laboratory test results	74.07
	biological indicators	71.64
	biochemical indicators	69.84
medical department	field expertise	58.62
	department/agency	58.18
	academic area	55.17
micro-organisms	microbial subtype	75.00
	microbial drugs	75.00
	source of infection	73.91
disease	disease characteristics	75.00
	vaccination history	74.07
	disease cause	73.47

Table 15: Three extended labels with the highest F1 in the application of ED task.

Event Types	Extended Event Labels	F1 (%)
	CASIE	
phishing	trick	62.07
	deceive	52.94
	trap	52.17
data breach	steal	72.73
	data theft	69.57
	theft	66.67
ransom	extortion	80.00
	financial crime	80.00
	event type: ransom	69.57
discover vulnerability	detection	78.26
	discover	72.73
	vulnerability discovery	70.00
patch vulnerability	update	63.64
	solution	61.54
	software_patch	60.87
	DuEE	
finance/ trading	transaction-pick	80.00
	economic-transfer	66.67
	capital markets-listing	64.29
product behavior	business activities-release	90.00
	product-launch	90.00
	financial business-launch	86.96
contact	relationships-apology	81.82
	emotion-visiting class	80.00
	personal connection-thanks	80.00
disaster/ accident	traffic accident-distress	57.14
	unexpected event-distress	55.56
	natural disaster-accident	52.63
competitive behavior	match result-beat	66.67
	match-beat	66.67
	contest result-defeated	64.00
organizational behavior	personal relationships-exit	100.00
	movement-leave	100.00
	sports competition-exit	96.00
life	personnel status-deceased	53.85
	death-remains	53.33
	health-death	51.61
judicial act	legal action-detention	82.76
	crime-arrested	76.92
	behavior-arrested	76.92
organizational relation	meeting-opening	86.96
	sports competition-unveiling	86.96
	events-unveiling	86.96