MMUIE: Massive Multi-Domain Universal Information Extraction for Long Documents

Shuyi Zhang[1,2], Zhenbin Chen[3], Shuting Li[4], Kewei Tu[3],
Li Jing[1,†], Zixia Jia[2,†], Zilong Zheng[2,†],
[1]School of Computer Science, Wuhan University,
[2]State Key Laboratory of General Artificial Intelligence, BIGAI,
[3]ShanghaiTech University,
[4]University of Science and Technology of China,
[†]Correspondence

Abstract

We present **MMUIE**, a large-scale universal dataset for multi-domain, document-level information extraction (IE) from long texts. Existing IE systems predominantly operate at the sentence level or within narrow domains due to annotation constraints. MMUIE addresses this gap by introducing an automated annotation pipeline that integrates traditional knowledge bases with large language models to extract fine-grained entities, aliases, and relation triples across 34 domains. The dataset comprises a weakly-supervised training set and a manually verified test set, featuring 723 entity types and 456 relation types. Empirical evaluations reveal that existing sentence-level IE models and even advanced LLMs underperform on this task, highlighting the need for better domain-aware document-level models. To this end, we develop DocUIE, a universal IE model fine-tuned on MMUIE, which achieves strong generalization and transferability across domains. MMUIE lays the foundation for robust, scalable, and universal information extraction from long-form text in diverse real-world scenarios. All code, data, and models are available in https://github.com/Shuyi-zsy/Massive-Multi-Domain-UIE.

1 Introduction

Information Extraction (IE) (Xu et al., 2024; Dagdelen et al., 2024; Bouziani et al., 2024; Huguet Cabot and Navigli, 2021; Giorgi et al., 2022; Li et al., 2025a), which aims to identify and extract structured information from unstructured text, has long played a vital role in knowledge graph study (Zhong et al., 2023; Yu et al., 2020; Fu et al., 2019). Moreover, numerous domains increasingly demand the structured representation of key information extracted from lengthy texts to effectively address domain-specific challenges. For instance, in the *legal* domain, the legal case retrieval task (Feng et al., 2024) focuses on identifying and matching legal essentials, often using legal graphs instead of entire texts (Tang et al., 2024b,a; Donabauer and Kruschwitz, 2025). Similarly, in the *financial* domain, stock trend forecasting task often relies on relationship graphs that capture financial concepts and inter-stock relationships (Niu et al., 2024; Wang et al., 2021).

Domains: Health
Document: The world Health Organization assisted ....since existing methods for heavy metal poisoning were not particularly effective. Dimercaprol was administered to several patients...for this sort of poisoning following the outbreak. Polythiol resins, penicillamine and dimercaprol sulfonate all helped..... Different treatements for mercury poisoning have been developed...a joint FAO and WHO meeting made several recommendations...
① **relation types:** [*drug or therapy used for treatment*, has part, country, *medical condition treated*, official language...]
② **entity types:** [location, *designated intractable/rare disease*, organization, *class of disease*, *type of chemical entity*...]
Relations:
<Baghdad, country, Iraq>
<United Kingdom, official language, English>
<dimercaprol sulfonate, *drug or therapy used for treatment*, mercury poisoning>
<dimercaprol, *drug or therapy used for treatment*, mercury poisoning>
Entities:
class of disease: (mercury poisoning)
organization: (WHO, World Health Organization)
type of chemical entity: (dimercaprol sulfonate, dimercaprol)...

Figure 1: Practical application cases in *Health* domain. Domain-specific entity and relation types are highlighted in *italics and underlined*.

Addressing such downstream tasks typically requires models capable of comprehending lengthy texts and identifying domain-specific relationships beyond general ones. This necessitates the development of document-level, fine-grained information extractors that can automatically extract entities, entity aliases, and relations of interest across diverse domains. Figure 1 illustrates practical application cases within a specific domain. In the *Health* domain, extracting relations like *drug or therapy used for treatment* is more worthy of attention than generic ones such as *part of*. In addition, considering the characteristics of lengthy texts, accurately identifying co-referential aliases will affect the extraction of both entities and relations. For instance, recognizing that *dimercaprol sulfonate* and *dimercaprol* refer to the same entity in this document

Findings of the Association for Computational Linguistics: EACL 2026, pages 6338–6370
March 24-29, 2026

Dataset	# Train	# Test	Avg.Entity	Avg.Triple	Inter-sent Tri.	*Avg.Word*	*# Rel-type*	*# Ent-type*	*Domain*
DocRED	4053	1000	26	12.5	40.7%	198	96	5	4 *domains*
DocRED†	101873	-	25	14.8	-	210	96	5	-
Re-DocRED	3553	500	19	29.6	53.47%	198	96	5	4 *domains*
DocRED-FE	1596	1000	19.47	12.5	-	199	96	119	4 *domains*
MMUIE	-	343	47.5	30.1	26.54%	875	245	461	34 *domains*
MMUIE †	8357	-	80	59.7	51.4%	1081	456	723	34 *domains*
DWIE	700	99	27	24.4	41.61%	625	63	10	*News*
SciERC	400	100	16	9.4	-	130	7	6	*Science*
SciREX	372	66	368	21	-	5735	9	4	*Science*
BioRED	600	-	34	10.8	-	303	4	6	*Biology*

Table 1: Existing document-level information extraction datasets and their characteristics. **Train / Test**: The number of documents used for training/testing. **Inter-sent Tri.**: the proportion of inter-sentence relation triples. **Rel-type / Ent-type**: the amount of relation/entity types. **Notation** †: weakly supervised training data.

enables the extraction of two relation triples of type *drug or therapy used for treatment*.

However, recent advances have predominantly focused on developing *universal sentence-level* generative information extractors (Li et al., 2025b; Gui et al., 2024; Li et al., 2024c; Sainz et al., 2023; Li et al., 2023a; Wang et al., 2023b; Lou et al., 2023), largely due to the prevalence of sentence-level annotations in existing datasets. Other studies have explored available document-level datasets, which remain scarce and are typically confined to *a single domain* (Luo et al., 2022; Jia et al., 2019; Luan et al., 2018; Jain et al., 2020) or encompass only *general entity and relation types across a limited number of domains* (e.g., only 4 domains included in DocRED (Yao et al., 2019), Re-DocRED (Tan et al., 2022), and DocRED-FE (Wang et al., 2023a)).

To alleviate the scarcity of multi-domain document-level IE datasets, we first develop an automated annotation pipeline that combines traditional Knowledge Bases (KBs) with advanced LLMs. Our approach focuses on analyzing lengthy texts to uncover domain-specific entity and relation types across diverse domains, thereby facilitating the construction of fine-grained, domain-aware label sets, leading to a **m**assive **m**ulti-domain **u**niversal **i**nformation **e**xtraction dataset (MMUIE). By incorporating complementary annotated techniques, we integrate previously separate functions into a unified structure information annotation framework, where components interact with one another, thereby improving the accuracy of entity and relation annotations.

We present a rigorously validated benchmark for long-context, cross-domain information extraction, covering 34 domains with fine-grained entity types and relation types (see Table 1). This benchmark enables fine-grained evaluation of entity and relation extraction at the document level. Initial results in Table 2 reveal that both large-scale sentence-level extractors and even advanced pre-trained LLMs with reasoning capabilities perform poorly, highlighting the substantial gap between current methods and real-world requirements.

Furthermore, we introduce DocUIE, a universal information extractor fine-tuned on domain-aware data from MMUIE, developed to handle long texts and capture domain-specific types. Experimental results demonstrate that jointly training from increasing domains consistently enhances our extractor's long text comprehension and domain transferability, underscoring the value of MMUIE in advancing document-level IE research.

In summary, our main contributions include:

- We develop an automated annotation pipeline that combines traditional KBs with LLMs to extract fine-grained structured information from raw web content, especially for lengthy texts.
- We introduce MMUIE, a high-quality *multi-domain* dataset for *fine-grained* document-level information extraction, addressing the scarcity of such datasets and enabling research in this area. The dataset encompasses 34 domains, 723 entity types, and 456 relation types.
- We propose a domain-agnostic DocUIE model which, guided by domain relevance, simultaneously demonstrates generalization across diverse domains and transferability between them.

2 MMUIE Construction from Raw Texts

Although lengthy texts are abundant, document-level information extraction (DocIE) is still hindered by the scarcity of annotated data and limited domain coverage, owing to the high cost and inherent challenges of manual annotation. Additionally,

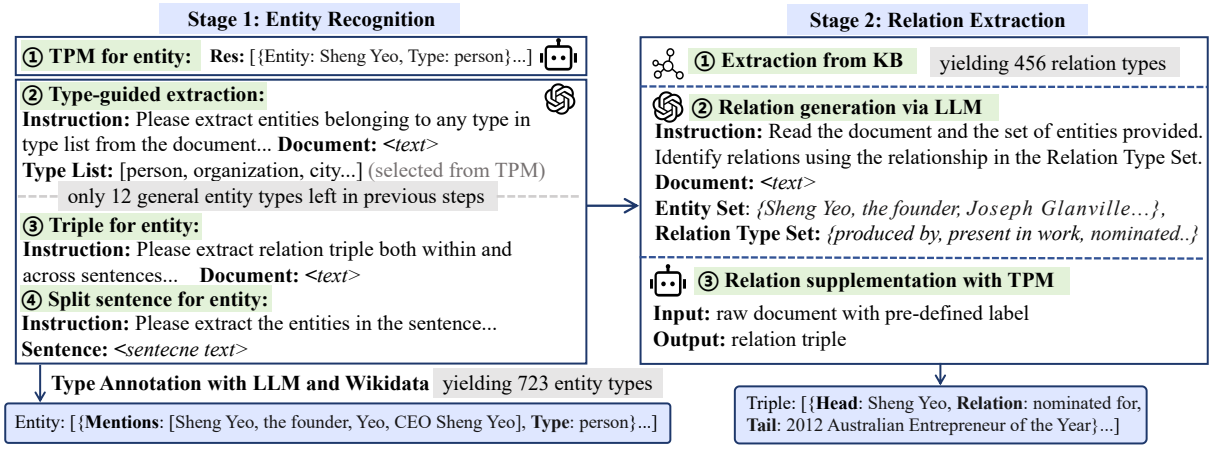


Figure 2: Overview of our automatic annotation pipeline to construct our multi-domain fine-grained information extraction dataset MMUIE. TPM means Task-specific Pretrained Models. See appendix E for all prompt details.

the uncontrollability of LLMs in content generation makes it difficult to convert information-rich long texts into fine-grained structured information through a single instruction, as shown by the results in Table 2. To address this gap, we propose an automated annotation pipeline capable of constructing a multi-domain, fine-grained DocIE dataset that captures long-context dependencies. Spanning 34 diverse domains, MMUIE comprehensively annotates core structure information elements, including entities, co-referential aliases, and relation triples.

2.1 Data Collection

Inspired by (Yao et al., 2019), we collect articles from Wikipedia [1] as raw texts. To construct a multi-domain dataset, we select articles by category and ensure that their length is in the range of 500-3,000 words, a carefully curated scope that meets the long text demands while mitigating manual annotation difficulties for the test set (no length limitation in automated annotation itself). All texts contain only the main content, excluding irrelevant information such as references and external links. As for the domain selection, we choose 34 domains that are both practically relevant and cover the 13 main Wikipedia categories (Appendix D.1).

2.2 Pipeline Architecture

Our annotation pipeline is primarily derived from three sources: Traditional Knowledge Base, Wikidata[2]; An advanced closed-source LLM, GPT4o-mini[3] (Considering the trade-off between performance and cost, the comparative experiments with different models are shown in Table 9); Task-specific Pretrained Models (TPMs). An overview of the pipeline is shown in Figure 2. We describe the entity (containing co-referential aliases) and relation annotation processes in the following subsections.

2.2.1 Stage 1: Entity Recognition (ER)

In DocIE, an entity typically comprises a series of mentions that represent various aliases referring to the same underlying entity. Consequently, entity recognition involves identifying all associated aliases of an entity and its corresponding type. Our objective in ER is to identify both general and domain-specific entities that extend beyond the popular entity types typically covered in existing datasets (Yao et al., 2019; Cabot et al., 2023).

Annotating General Entities We follow the entity annotation process of DocRED (Yao et al., 2019), leveraging a TPM spaCy[4] to annotate general types of entities (**TPM for entity**). However, DocRED's annotation has two limitations: *1) While SpaCy provides 18 entity types, DocRED selects only 5 general types, such as City, Person, and Organization, limiting expressiveness and coverage; 2) For lengthy documents, using SpaCy, which is primarily trained on splitting sentences, still misses many entities, even for these general types, due to insufficient context.* To facilitate these limitations: (1) We retain 12 SpaCy types for their universality and alignment with Wikidata (excluding quantita-

[1] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[4] https://spacy.io/models/en#en_core_web_lg

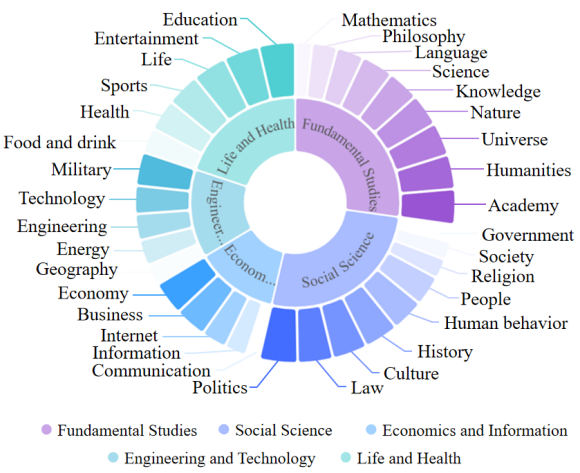


Figure 3: The proportion of documents distributed among 34 domains. For correlation analysis across domains, see Appendix D.5.

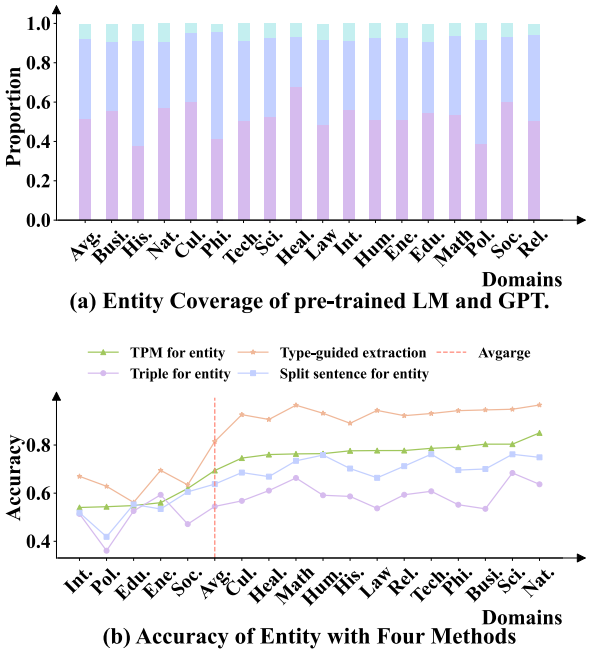


Figure 4: Entity Recognition Analysis considering both diversity and accuracy of entities extracted by different methods.(a) shows the coverage of entity extraction using TPM and LLM: entities extracted only by LLM (bottom); entities extracted only by spaCy (top); entities captured by both (middle).(b) presents the accuracy of entity identification using four different methods. The analysis of other domains is included in Appendix C.4.

tive types such as Percent, Quantity, and Ordinal, which are error-prone and less informative), and apply SpaCy to infer entity mentions on our raw texts; (2) We leverage the strong generalization capability of LLMs to discover additional entities of these general types, supplementing the initial extraction, termed as **Type-guided extraction**. Specifically, we provide the full document and the 12 predefined entity types, and instruct the LLM to extract all corresponding entity mentions.

Annotating Domain-Specific Entities Recall that the primary objective of our task is to extract domain-specific entities with fine-grained types and to subsequently annotate the various domain-aware relations between them. To this end, we undertake the following measures to optimize the identification of domain-specific entities. Given the absence of datasets or ontologies that define domain-specific entity type sets across various domains, we leverage LLMs by prompting them with task-relevant instructions (refer to Section E for details). Specifically, we adopt a two-phase approach: a comprehensive identification of all entity mentions, followed by a fine-grained classification of domain-specific entity types. [1]

1. *Domain-Specific Entity Identification with LLMs* We propose a **Triple for entity** strategy for identifying domain-specific mentions. Specifically, we instruct LLMs to extract open-vocabulary relation triples from the entire document, using the resulting subjects and objects as additional entity mentions. These entities are generally more relevant to the specific domain and context. The motivation for this strategy is threefold: *1) LLMs show high precision in open-vocabulary relation extraction, where types are directly represented by predicates (Li et al., 2023b); 2) Domain-specific texts typically contain predicates reflecting domain-relevant concepts; and 3) Entities involved in these relations are likely to be domain-specific.* However, we observe two main issues: *1) Without explicit type constraints, LLMs may produce boundary errors in long contexts (see Section C.2); and 2) certain domain-specific entities may not participate in explicit relations.* To address these challenges, we introduce a **Split Sentence for Entity** step, decomposing the document into sentences (a finer textual granularity), and then prompting the LLM to directly extract mentions independently of type or relation constraints. Finally, to improve the accuracy of entities, we utilize LLMs to perform **Secondary verification** on the extracted entities, ensuring that they are present in the text.

2. *Type Annotation with LLMs and Wikidata* As previously mentioned, there are no existing reference sets for domain-specific entity types in

[1] Note that in this step, LLMs may also identify general entities, some of which are new general types, and others that overlap with entities predicted by SpaCy. We retain all newly identified general entities.

the literature. Therefore, our first objective is to construct domain-specific type sets. We begin by leveraging "instance of" property in KBs to build a fine-grained entity type set, and then use LLMs to classify mentions accordingly. Specifically, entity mentions are first linked to Wikidata to retrieve their item IDs and corresponding types, as detailed in Appendix C.6. Mentions with the same ID and type are regarded as aliases of the same entity. For entities not present in the KB, the LLM is instructed to select the most appropriate type based on the entity's contextual snippets and the pre-constructed type set. Type annotation, combining both Wikidata and LLMs rather than relying solely on LLM-generated mention types, offers two advantages: *1) it mitigates the risk of producing synonymous types in various forms; 2) it minimizes the occurrence of untyped entities (though some miscellaneous ones may remain), thereby enhancing the consistency and quality of the entity type classification.*

2.2.2 Stage 2: Relation Extraction (RE)

As the entity recognition process identifies domain-specific entities (with quantitative analysis provided in Section 2.3), the subsequent relation extraction naturally yields a broad set of domain-aware relation types. Following a similar rationale to the annotation process of domain-specific entities, we first construct a fine-grained set of relation types, encompassing both general and domain-specific categories, by utilizing the relations recorded in KBs. We subsequently guide LLMs with these pre-constructed type sets to extract relation triples. However, as relation extraction requires a more precise understanding of semantics, LLM performance remains limited (see supporting experiments in Appendix F.2). As a result, we fine-tune a Task-specific Pretrained Model (TPM) using existing, well-annotated document-level datasets to mitigate the incomplete labeling issues associated with LLMs.

Extraction from KB: For entities linked to Wikidata, we query their recorded relations (details in Appendix C.6).

Relation generation via LLM: We provide the entity set, our constructed relation types set, and the original document, prompting LLM to generate potential relation triples in a question-answering format. Upon further inspection, we find that the model occasionally confuses the order of subject and object, likely due to an imprecise understanding of relational semantics. To address this, we adopt the correction method proposed in DocGNRE (Li et al., 2023b), employing a Natural Language Inference (NLI) model to rectify such errors.

Relation supplementation with TPM: Taking full advantage of existing manually-refined document-level training set (*i.e.,* Re-DocRED (Tan et al., 2022), DWIE (Zaporojets et al., 2021), SciERC (Luan et al., 2018)), we fine-tune an open-source LLM-based information extractor. This extractor demonstrates enhanced capabilities comparable to those of closed-source LLMs (experiments for verification are shown in Appendix F.2) and is used to supplement triples based on our constructed label set. After these three steps, all extracted triples undergo **secondary verification**, similar to entity recognition.

Following this pipeline, we construct the first multi-domain dataset **MMUIE** for long-document information extraction. ***Notably, the two characteristics of our dataset, domain-specific and fine-grained, are complementary***: the former emphasizes that domain-relevant information deserves greater attention, while the latter focuses on precise semantic expression. Together, they are indispensable in information extraction from specialized documents. Figure 3 depicts the domain distribution, and Table 1 presents the main statistics. From the 8700 annotated documents (additional data can be automatically generated with our framework if needed), 343 were carefully selected as the test set, ensuring at least 5 documents per domain. All entities, types, and relation triples in the test set were manually verified for accuracy (with a detailed description of the process and quality verification of manual annotation provided in Appendix D.3). The remaining 8,357 documents serve as a weakly supervised training set for demonstration and model training purposes. The version of the dataset we used has filtered out extremely low-frequency (appearing less than 20 times) relation and entity types. Detailed data analysis is provided in Appendix D.2.

2.3 Rationality of Annotation Framework

Although previous information extraction methods can be used for data annotation (as discussed in Appendix B), our approach synthesizes the advantages of knowledge bases (KBs) and large language models (LLMs) within a cohesive pipeline, where the two are mutually complementary: the KB provides diverse and precise type sets, enabling the LLM to generate concise, domain-aware entities and triples.

To evaluate the effectiveness of each component in our annotation framework, we use the manually refined test set for a rationality analysis, illustrated with examples and varying numbers of entities or triples in Table 10. Three key findings emerge from this quantitative assessment.

LLM-generated annotations encompass a diverse range of entity types, effectively supplementing domain-specific entities. Figure 4 (a) compares the coverage of entity annotation between TPM and LLM, showing that LLM achieves an average coverage of 92.26%, far exceeding the 48.2% of TPM. We further analyzed the accuracy of entities obtained at different steps and showed in Figure 4 (b). Through entity types annotation by LLM, we manually check[1] that 35.13% of the entity types are domain-specific.

LLM extracts more accurate relation triples than KB. Although relations recorded in KB are objectively correct, most cannot necessarily be inferred from the context. For example, the triple <*Bahir Dar*, *located in the administrative territorial entity*, *Amhara region*>is factually accurate, but the document does not mention the administrative region of the school. This explains why only 38.6% of the relations extracted from the KB are manually verified to be correct, while those generated by LLM are more accurate at 48.11%. Besides, 19.96% of the relation types are domain-specific. Detailed analyses are provided in Appendices C.5 and D.4.

Secondary verification is helpful. After further refinement by LLM, entity accuracy increased to 84.35%, a 39.78% improvement over the original 60.35%.[2] Among the correct entities, those with correct types accounted for 74.69% of all extracted entities. For relation triples, the average accuracy rate across all domains has also increased from 43.63% to 56.75%.

In our automated annotation framework, we make every effort to incorporate multiple effective steps to annotate fine-grained, domain-aware entities and relation triples. However, as observed in the analysis of DocRED from Re-DocRED, the omission of certain entities and relations remains unavoidable. Potential approaches to mitigate this issue are discussed in Appendix C.3 and are left for future work. Nevertheless, even with the current version of our automatically annotated weakly supervised training set, consistency improvements are demonstrated in the following experiments.

[1]Types that appear exclusively in a single domain are considered domain-specific.

[2]Note that high accuracy in entity identification is crucial for reliable downstream relation extraction.

3 Experiments on MMUIE

Leveraging MMUIE, we conduct two category experiments: training-free evaluation on the test set of a series of models; fine-tuning a universal document-level information extractor. To better compare and analyze models' performance, we additionally leverage two existing representative datasets: Re-DocRED (Tan et al., 2022) is the most commonly used dataset in the field of document-level relation extraction, which contains four domains, as well as entity types from five fields (see Appendix D.4). DWIE (Zaporojets et al., 2021) is a single-domain dataset that contains 63 relation types and 10 entities in the specific *news* domain.

Evaluation We used three evaluation metrics: Precision, Recall, and F1 score. In the document-level setting, an entity may have different aliases within lengthy texts; we adopted both ordinary and strict evaluation criteria. Specifically, **NER** refers to the evaluation of whether a specific entity *mention* and its corresponding type are correct, while **NER.coref** requires that all mentions must be found for an entity recognition to be considered correct. **RE** requires that the entities and relation types in the triples must be the same as the ground truth, whereas **RE.coref/RE.coref** ^{GT} allows for the substitution of subject or object entities with their predicted/ground-truth aliases to be considered correct. Our **RE** score serves as a combined metric for evaluating entity recognition and relation extraction.

3.1 Training-free Evaluation

Models We select three different types of models to evaluate their training-free performance on our MMUIE test set.

- **API-based LLMs**: We select GPT4o, GPT4.1[3], and Deepseek R1[4]. It should be noted that these models all support structured output, avoiding the impact of incorrect structured generation on the final results.
- **Open-source LLMs**: We select Llama3-8B-Instruct[5] and Qwen3-8B (Yang et al., 2025). We

[3]https://openai.com/

[4]https://www.deepseek.com/

[5]https://ai.meta.com/blog/meta-llama-3/

Datasets	MMUIE			Re-DocRED			DWIE		
Eval. Setting	**RE**	**RE.coref** ^{GT}	**NER**	**RE**	**RE.coref** ^{GT}	**NER**	**RE**	**RE.coref** ^{GT}	**NER**
GPT4o	5.18	7.12	22.44	14.22	17.29	52.85	3.88	5.02	43.67
GPT4.1	5.01	6.99	21.32	11.97	13.50	64.50	14.03	14.51	43.35
Deepseek R1	6.87	7.23	21.70	14.09	15.88	67.64	12.20	15.45	47.73
Llama3-8B-Instruct	0.54	0.54	7.52	1.02	1.02	14.83	3.12	3.12	48.13
Qwen3-8B	2.21	2.49	16.76	0.33	0.37	25.27	6.97	9.48	13.73
LangExtract (Qwen3-8B)	2.22	2.55	21.90	7.67	8.75	20.53	3.57	5.14	35.52
IEPile	0.92	3.34	19.60	7.70	11.03	61.17	0.46	4.28	12.22

Table 2: Performance of API-based LLM, open-source LLM, and sentence-level UIE model without fine-tuning. All results shown are F1 scores. Considering that existing work (Xue et al., 2024) uses ground-truth co-reference information when evaluating the ability of relation extraction, we show it here as **RE.coref** ^{GT}.

Task (Eval.Setting)	**Name Entity Recognition** (NER / NER.coref)			**Relation Extraction** (RE / RE.coref)		
Datasets	MMUIE	Re-DocRED	DWIE	MMUIE	Re-DocRED	DWIE
Single (Llama)	37.43 / 32.08	**86.36 / 79.99**	86.79 / **83.26**	15.10 / 16.49	51.44 / 54.16	74.37 / 75.18
DocUIE‡ (Llama)	- / -	85.85 / 79.25	75.74 / 70.69	- / -	50.96 / 54.50	67.27 / 69.94
DocUIE (Llama)	**43.22 / 37.36**	85.59 / 76.66	**88.04** / 72.55	**23.35 / 24.83**	**54.59 / 60.50**	**77.65 / 79.76**
Single (Qwen)	38.46 / 33.76	**88.69** / 81.08	86.37 / 80.93	17.92 / 19.91	54.43 / 58.31	75.56 / 76.89
DocUIE‡ (Qwen)	- / -	88.22 / **81.12**	88.96 / 85.48	- / -	54.70 / 57.90	76.13 / 78.47
DocUIE (Qwen)	**40.64 / 34.79**	88.52 / 81.03	**89.97 / 86.47**	**19.50 / 21.52**	**55.54 / 60.01**	**78.10 / 80.12**

Table 3: F1 scores of models fine-tuning on MMUIE. **Single** indicates training and testing solely on the corresponding dataset, while notation ‡ refers to training on a combination of existing datasets without our MMUIE.

Input **Tokens** Number	Precision	Recall	F1 score
less than 1000	27.22	29.81	28.43
[1000, 1500)	22.34	23.88	23.08
[1500, 2000)	15.7	15.62	15.66
[2000, 2500)	17.01	16.67	16.84
[2500, 3000)	14.01	15.59	14.76
no less than 3000	3.65	3.44	3.54

Table 4: Relation extraction performance of DocUIE on different document lengths within MMUIE dataset.

only instructed them to output the results in JSON format without any post-processing.

- **UIE Models**: We select IEpile(Gui et al., 2024) as a representative sentence-level model, which is fine-tuned on 33 sentence-level IE datasets. We believe it has multi-domain generalization; thus, we explore whether it can still maintain good capabilities for long documents. LangExtract[1] is an advanced and widely used tool for document-level IE, encompassing both NER and RE tasks.

Results Table 2 presents the models' performance on both MMUIE and previous datasets. Due to their limited ability to identify aliases and the lack of a significant difference between the RE and RE.coref settings, we only report results for RE and NER. Consistent with the findings of previous studies (Xue et al., 2024; Ozyurt et al., 2023), LLMs have not achieved good results in structured information extraction tasks. Even for the most advanced tool, LangExtract, the performance remains limited, despite being provided with the ground-truth type in the user-defined prompt (More usage is presented in Appendix B). Compared to the other two datasets, the performance on MMUIE has significantly deteriorated, a result that is consistent with the characteristics of our data. *The longer text length and the diverse fine-grained entity and relation types have both increased the difficulty of information extraction*. For instance, distinguishing between fine-grained subtypes requires nuanced semantic understanding (Ling and Weld, 2012). Additionally, *the models trained on a large number of sentence-level datasets exhibit limited effectiveness in document-level information extraction*, highlighting the need for a universal multi-domain document-level dataset. Besides the zero-shot methods, we supplement a few-shot (another training-free setting) experiment in Appendix F.2.

3.2 Universal Information Extractor

In this section, we aim to develop a universal document-level information extractor by enhancing the information extraction capabilities of open-source LLMs and enabling them to learn a unified output format across tasks through fine-tuning.

Training Formats Given the limitations of LLMs in handling long-context inputs, we adopt a strategy inspired by Gui et al. (2024). Specifically, we constrain the input to a fixed number of target types per instance and instruct the model to extract

[1] https://langextract.com/

corresponding entities or relation triples. Consequently, each training instance includes a hyperparameter k (set to 8 according to the ablation study) number of types, and each document is split into multiple such instances. We present the prompt format used for entity recognition and relation extraction during training in Appendix E. In the output, entities with aliases are extracted and presented as grouped lists. For RE, we adopt a nearly identical prompting strategy to ensure consistency across tasks. For complete experimental details (including configurations), refer to Appendix F.1.

Results Table 3 shows the results under different training settings. On the RE task, DocUIE has an F1 score of 60.5 on the Re-DocRED dataset, surpassing the best result of the current State-of-the-art (SOTA) generative model (AutoRE (Xue et al., 2024): 53.84). On the DWIE dataset, DocUIE also exceeds the SOTA model REXEL (65.8) (Bouziani et al., 2024). These results demonstrate: *1) Our dataset is of high quality, as mixed training enhances model performance on both our newly constructed test set and existing benchmarks for the challenging joint entity and relation extraction task; and 2) mixed-domain training significantly boosts model performance compared to training on a single domain (as evidenced by comparisons between DocUIE and the Single settings).* Figure 5 shows the performance improvements brought by our DocUIE in different domain categories. To further examine the remaining challenges in MMUIE and DocUIE, we conduct a qualitative error analysis, with detailed examples and quantitative results provided in Appendix F.3.

Further evaluation on relation extraction Focusing on the more challenging RE task, we analyze MMUIE and DocUIE from two aspects. (1) ***Domain-specific Relations***: MMUIE contains fine-grained, domain-specific relations that are infrequent and harder to learn. We created an easy version by removing these relations (see details in Appendix F.4). Figure 6 shows improved Recall and F1 score (especially Recall), though DocUIE's Precision slightly declined. (2) ***Document Length Impact***: Longer documents have more diverse types and complex relations (e.g., inter-sentence). As shown in Table 4, performance declines significantly with document length, further illustrating why sentence-level UIE models underperform on document-level tasks.

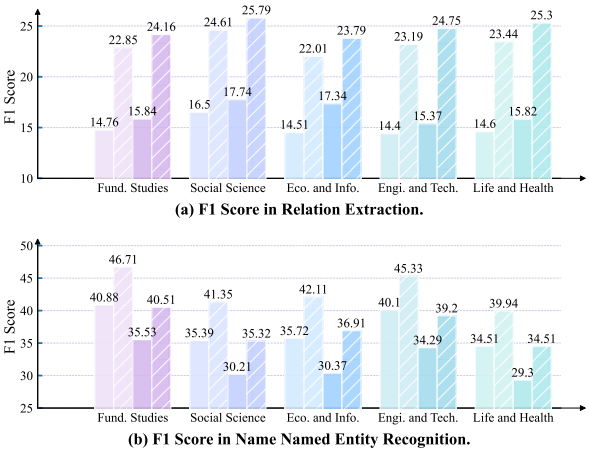
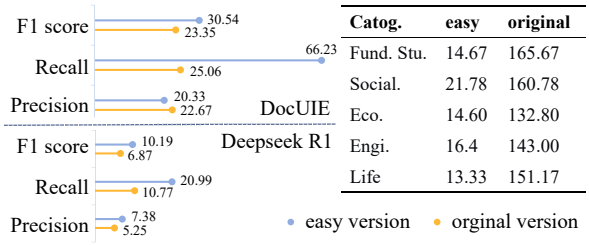


Figure 5: Performance improvements brought by DocUIE in five domain categories. Bars with slashes are DocUIE results, while the others are single train results. The four bars in each category show the results of RE (Single), RE (DocUIE), RE.coref (Single), and RE.coref (DocUIE) from left to right, respectively.



Catog.	easy	original
Fund. Stu.	14.67	165.67
Social.	21.78	160.78
Eco.	14.60	132.80
Engi.	16.4	143.00
Life	13.33	151.17

Figure 6: Results of the easy version of MMUIE and comparison with the original version in three metrics. The table shows the average number of relation types for the easy and original versions.

4 Related Work

Recent efforts have highlighted the limitations of the original document-level relation extraction dataset (Zhang et al., 2023b). To address this issue, several studies (Huang et al., 2022; Tan et al., 2022; Li et al., 2023b) have proposed re-annotated versions of the most popular dataset DocRED (Yao et al., 2019). However, there still remain significant challenges in building document IE models due to the lack of data and adapted models. With the advancement of LLMs, the emergent capabilities of LLMs have shown surprising performance in universal IE (Gui et al., 2024; Sainz et al., 2023; Li et al., 2024c). Although they have achieved promising performance on NER and RE tasks, most of them are trained on sentence-level datasets, leading to suboptimal extraction results on long texts. See more in Appendix A.

5 Conclusion

We present MMUIE, the first fine-grained, multi-domain, document-level IE benchmark, and DocUIE, which demonstrates strong cross-domain generalization and effective long text IE. Despite some noise in the distantly supervised data, DocUIE achieves consistent improvements across datasets. Our findings highlight the challenge of extracting fine-grained information from long documents and the promise of LLM-based annotation for enhancing the MMUIE 's completeness.

Limitations

Constrained by time and cost, we developed a weak supervision training set with limited documents for both demonstration and model training purposes. Should the need arise, this pipeline can be readily employed to swiftly generate a substantial volume of training data. Meanwhile, the domains in MMUIE can not cover all possible nested sub-categories in Wikipedia. Though our DocUIE achieves consistent performance gains with noisy weakly supervised data as training corpora, the overall improvement remains constrained by the inherent noise and incompleteness of the annotations. We plan to explore denoising techniques as part of future work to further enhance data quality and model robustness.

Ethics Statement

The development of MMUIE and DocUIE adheres to the principles of responsible AI research. All datasets used or constructed in this work are derived from publicly available sources, and no personally identifiable or sensitive information is included.

We acknowledge that information extraction models trained on weak supervision and large-scale LLMs may inadvertently reinforce domain-specific or systemic biases present in the underlying sources or training data. To mitigate this, we encourage future researchers and practitioners to perform bias audits and domain-specific fairness evaluations before deployment, especially in sensitive domains like healthcare, legal systems, or financial decision-making.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.62372335, No.62376031.

References

Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. Rexel: An end-to-end model for document-level relation extraction and entity linking. *arXiv preprint arXiv:2404.12788*.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. Red^{FM}: a filtered and multilingual relation extraction dataset. *arXiv preprint arXiv:2306.09802*.

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

Gregor Donabauer and Udo Kruschwitz. 2025. A reproducibility study of graph-based legal case retrieval. *arXiv preprint arXiv:2504.08400*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485.

Cong Fu, Tong Chen, Meng Qu, Woojeong Jin, and Xiang Ren. 2019. Collaborative policy learning for open knowledge graph reasoning. *arXiv preprint arXiv:1909.00230*.

John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, Dublin, Ireland. Association for Computational Linguistics.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. Iepile: Unearthing large-scale schema-based information extraction corpus. *arXiv preprint arXiv:2402.14710*.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, J. Guo, and Xueqi Cheng. 2023. Retrieval-augmented code generation for universal information extraction. In *Natural Language Processing and Chinese Computing*.

Wenlong Hou, Weidong Zhao, Xianhui Liu, and Wenyan Guo. 2024. Knowledge-enriched prompt for low-resource named entity recognition. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(5).

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. *arXiv preprint arXiv:2204.07980*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.

Rebii Jamal, Mounir Ourekouch, and Mohammed Erradi. 2025. UOREX: Towards uncertainty-aware open relation extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6027–6040, Albuquerque, New Mexico. Association for Computational Linguistics.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. *ArXiv*, abs/1904.02347.

Guochao Jiang, Ziqin Luo, Yuchen Shi, Dixuan Wang, Jiaqing Liang, and Deqing Yang. 2024. ToNER: Type-oriented named entity recognition with generative language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16251–16262, Torino, Italia. ELRA and ICCL.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138.

Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024a. Meta in-context learning makes large language models better zero and few-shot relation extractors. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

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

Huahang Li, Shuangyin Li, Fei Hao, Chen Jason Zhang, Yuanfeng Song, and Lei Chen. 2024b. Booster: Leveraging large language models for enhancing entity resolution. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1043–1046.

Jiangnan Li, Yice Zhang, Bin Liang, Kam-Fai Wong, and Ruifeng Xu. 2023a. Set learning for generative information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13043–13052.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023b. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023c. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Zhongqiu Li, Shiquan Wang, Ruiyu Fang, Mengjiao Bao, Zhenhe Wu, Shuangyong Song, Yongxiang Li, and Zhongjiang He. 2025b. Mr-uie: Multi-perspective reasoning with reinforcement learning for universal information extraction. *arXiv preprint arXiv:2509.09082*.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, et al. 2024c. Knowcoder: Coding structured knowledge into llms for universal information extraction. *arXiv preprint arXiv:2403.07969*.

Xiao Ling and Daniel Weld. 2012. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 94–100.

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

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. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 13318–13326.

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. *arXiv preprint arXiv:2203.12277*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Xuanfan Ni and Piji Li. 2023. Unified text structuralization with instruction-tuned language models. *ArXiv*, abs/2303.14956.

Yingjie Niu, Lanxin Lu, Rian Dolphin, Valerio Poti, and Ruihai Dong. 2024. Evaluating financial relational graphs: Interpretation before prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 564–572.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.

Yilmazcan Ozyurt, Stefan Feuerriegel, and Ce Zhang. 2023. Document-level in-context few-shot relation extraction via pre-trained language models.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, JIE MA, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages.

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. *arXiv preprint arXiv:2310.03668*.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting docred–addressing the false negative problem in relation extraction. *arXiv preprint arXiv:2205.12696*.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024a. Casegnn++: Graph contrastive learning for legal case retrieval with graph augmentation. *arXiv preprint arXiv:2405.11791*.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024b. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. In *European Conference on Information Retrieval*, pages 80–95. Springer.

Hongbo Wang, Weimin Xiong, Yifan Song, Dawei Zhu, Yu Xia, and Sujian Li. 2023a. Docred-fe: a document-level fine-grained entity and relation extraction dataset. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. 2021. A review on graph neural network methods in financial applications. *arXiv preprint arXiv:2111.15367*.

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

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *ArXiv*, abs/2312.15548.

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

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220.

Faren Yan, Peng Yu, and Xin Chen. 2024. Ltner: Large language model tagging for named entity recognition with contextualized entity marking. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XIX*, page 399–411, Berlin, Heidelberg. Springer-Verlag.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *ArXiv*, abs/2505.09388.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2):735–753.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. LLMaAA: Making large language models as active annotators. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pages 13088–13103, Singapore. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023b. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10853–10865.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.

A Related Work

A.1 Universal Information Extraction Datasets

High-quality annotated data is essential for training robust information extraction (IE) models. IEPile (Gui et al., 2024) proposed an entity-centric UIE dataset, which is obtained by collecting and cleaning existing IE datasets and reconstructing them. GoLLIE (Sainz et al., 2023) and Knowcoder (Li et al., 2024c) reorganized the schema into Python-style code format to construct hierarchical relationships between labels. However, all existing universal datasets are limited to the sentence level.

A.2 Annotation Methods

Recent efforts have highlighted the limitations of the original document-level relation extraction dataset DocRED (Yao et al., 2019), particularly its high false-negative rate. To address this issue, several studies have proposed re-annotated versions of the dataset by adding substantial numbers of missing relation triples. Notably, Huang et al. (2022) performed manual re-annotation from scratch using two expert annotators to label 96 documents. In contrast, Tan et al. (2022) introduced Re-DocRED by leveraging pre-trained relation extraction models followed by manual revision. Building on these re-annotation efforts, DocNGRE (Li et al., 2023b) further enhanced Re-DocRED by incorporating LLMs and natural language inference techniques to augment the dataset. Despite these efforts, there still remain significant challenges in building universal document IE models due to the lack of comprehensive training data and adapted models.

A.2.1 Previous IE Methods for Annotation

To tackle the challenges of information extraction, numerous studies have explored diverse approaches for name entity recognition (Jiang et al., 2024; Hou et al., 2024; Yan et al., 2024), relation extraction (Orlando et al., 2024; Zhang et al., 2023a; Cabot et al., 2023; Ni and Li, 2023; Li et al., 2024a), and co-reference resolution (Li et al., 2024b). While prior models rely either on traditional knowledge bases (Elsahar et al., 2018) or solely on LLMs, our automated annotation pipeline integrates both: using LLMs to identify context-dependent entities and relations, and knowledge graphs to provide predefined schemas that reduce issues of uncontrolled generation (e.g., semantically equivalent but inconsistent relation types). In line with our annotation requirements, we selected some readily reproducible methods (Orlando et al., 2024; Huguet Cabot and Navigli, 2021) (such as those that do not require additional training) and the tool LangExtract[1] for systematic experimentation in constructing a ***multi-domain***, ***fine-grained***, ***document-level*** information extraction dataset. The results are presented in Section B.

A.3 Universal Information Extraction Methods

The goal of Universal Information Extraction (UIE) is to integrate traditionally fragmented subtasks in IE into a unified generative framework, which eliminates boundaries between tasks, enhances model generalization, and improves cross-task transfer efficiency. UIE methods based on large language models can be categorized into two types: those that use natural language text (Gui et al., 2024; Paolini et al., 2021; Li et al., 2025b; Wang et al., 2023b; Lu et al., 2022; Xiao et al., 2023) and those that use code data (Sainz et al., 2023; Li et al., 2024c; Guo et al., 2023; Li et al., 2023c) for training and inference. IEPile (Gui et al., 2024) conducted schema-based instruction generation to fine-tune the LLMs and improve the performance in various subtasks. GoLLINE (Sainz et al., 2023) and KnowCoder (Li et al., 2024c) converted the schema into Python classes and used the capabilities of LLM to obtain correctly structured information extraction results. Although the aforementioned methods have achieved promising performance on NER and RE tasks, most of them are trained on sentence-level datasets, leading to suboptimal extraction results on long texts.

[1] https://langextract.com/

B Experiments in Annotation Methods

LangExtract[1] is an advanced tool that uses LLMs to extract structured information from unstructured documents based on user-defined prompts and can be adapted to any domain without model fine-tuning. Table 5 compares the usage of LangExtract and our pipeline for annotation. It is worth noting that LangExtract represents an open-vocabulary information extraction approach (Jamal et al., 2025), where both the entity and relation types to be extracted, as well as the output structure, are entirely determined by the prompt (see usage details in Table 6). In contrast, our target dataset requires a fixed set of fine-grained types to support the training of DocUIE. In addition, we argue that LangExtract is better suited for scenarios where the general content of a document is already known and the goal is to facilitate deeper analysis, such as extracting specific information (depending on both the document content and user-defined prompts) and highlighting it with visualization tools, rather than for use in data annotation.

ReLiK (Orlando et al., 2024) adopts a Retriever–Reader architecture, where the Retriever identifies candidate or relations likely to appear in the text, and the Reader verifies these candidates by aligning them with the corresponding textual spans. For relation annotation, ReLiK can also be used as an open-vocabulary IE method, formulated as Open-Domain Question Answering, where the input text itself is treated as the question from which relations are extracted. Besides, since the training data for ReLiK consists of 32-word chunks, we first split each document into sentences before annotation. Table 7 presents an example of usage and result on a document from our MMUIE. Although ReLiK is a fast relation extractor, its limited training context length prevents it from directly capturing cross-sentence relations without retraining or additional strategies for document-level extraction. Furthermore, when contextual information is insufficient, sentence-level inputs may lead to misinterpretation of proper nouns. For instance, with sent.1 as input, the model incorrectly identified *Anthropocene*, which is a conceptual term, as an organization, resulting in the erroneous triple <*Paul Crutzen, company, Anthropocene*>. The actual organization, *AWG*, only appears in subsequent text.

REBEL (Huguet Cabot and Navigli, 2021) is a seq2seq model that performs end-to-end relation extraction for more than 200 different relation types, which aligns more closely with our need for fine-grained labels. We use model *rebel-large* to annotate our document for the relation extraction task. From the results presented in the Table 8, we draw the following three observations. ***First, constructing a fine-grained set of relation types proves crucial for extracting semantically precise relation triples.*** For instance, both <*NASA, subsidiary, Kennedy Space Center*> and <*Kennedy Space Center, operator, NASA*> implicitly convey a hierarchical relationship. However, *subsidiary* applies only to corporate ownership and misrepresents the Kennedy Space Center as an independent entity. In contrast, *operator* precisely reflects NASA's role in managing the facility, aligning with real-world structures and knowledge graph practices. Thus, using fine-grained, domain-adapted relation types (e.g., *operator* instead of the general one *subsidiary*) greatly enhances the accuracy and semantic validity of knowledge extraction.

Second, large language models (LLMs) exhibit a greater ability than traditional pretrained models(TPMs) to uncover domain-relevant relations. Both <*Kennedy Space Center, operator, NASA*> and <*Kennedy Space Center, parent organization, NASA*> are valid relations. While *parent organization* represents a general organizational link, *operator* captures the actual functional relationship between subject and object, which is commonly observed in *politics* and *business* domains. However, because training corpora may not include such relation types, TPMs often fail to capture these domain-specific distinctions.

Finally, increasing the richness of entity coverage enhances the diversity of extracted relations. Although the full text was provided as input, REBEL produced only a single set of synonymous triples involving two entities. This limitation highlights the necessity of designing multiple entity recognition strategies to maximize the coverage of all potentially relevant entities.

C More Analysis of Annotation Pipeline

C.1 Base model selection

When designing our annotation pipeline, we conducted comparative experiments using both GPT-4o-mini and GPT-4o, and found that GPT-4o did not yield a higher number of valid relation triples than GPT-4o-mini. Considering the quality and

[1] https://langextract.com/

		LangExtract	Our Pipeline
Sources	*Knowledge graph*	no	yes
	Large language model	yes	yes
Tasks	*Entity recognition*	yes	yes
	Relation extraction	yes	yes
	Co-reference resolution	yes	yes
Characteristics	*Document-level*	chunked text, parallel processing	whole text, split sentence
	fine-grained Schema	user-defined in prompt	pre-defined (extract from Wikidata KB)
	Domain	adaptable to any domain	adaptable to any domain

Table 5: Comparison of LangExtract and Our Pipeline. We compare these two annotation methods from three aspects: **Sources**, which show the different knowledge source used for extracting information; **Tasks**, which represent the sub-tasks each method can accomplish; and **Characteristics**, which highlight key features of our target dataset and illustrate how each method addresses them.

Task	Key	Content
Entity recognition	***prompt***	Extract all *class of disease*, *person*, *scholarly article*... from the following text.
	examples	***text***="Patient was given 250 mg IV Cefazolin TID for one week.", ***extractions***=[lx.data.Extraction(*extraction_class="dosage", extraction_text="250 mg"*)]
	max char buffer	500
Relation extraction	***prompt***	Extract all relationships with types: *influenced by*, *risk factor*, *has part(s)* from the following text. Provide meaningful attributes for each relationship.
	examples	***text***=("ROMEO. But soft! What light through yonder window breaks? It is" " the east, and Juliet is the sun."), ***extractions***=[lx.data.Extraction(extraction_class="relationship", extraction_text="Juliet is the sun", attributes=*{"type": "metaphor", "subject": "Juliet", "object": "sun"}*,)]
	max char buffer	1000
Co-reference resolution	***prompt***	Extract all mentions refer to the same entity from the following text. Use exact text for extractions. Do not paraphrase or overlap entities. Provide meaningful attributes for each entity.
	examples	***text***=("Michael Jordan, known as the 'Air Jordan', was a legendary basketball player."), ***extractions***=[lx.data.Extraction(extraction_class="person", extraction_text="Michael Jordan", attributes=*{"class": "person", "text": "['Michael Jordan', 'Air Jordan']"}*)]
	max char buffer	1000

Table 6: **Usage Examples of LangExtract for annotation.** ***prompt***, ***examples***, and ***max char buffer*** are the key elements in LangExtract for document-level IE.

cost trade-off, we selected GPT-4o-mini as the base model for our pipeline. When used for annotation, GPT-4o and GPT-4o-mini show comparable performance (Figure 9), and it is reasonable to expect that stronger models can yield even better annotation quality. Overall, our automated annotation pipeline can employ different LLMs as the base model, ensuring reproducibility.

C.2 Analysis of Boundary Error

The boundary error occurs in the entity mentions with the wrong boundary. For example, "*Thinking in Time*" (mention a) and "*Thinking in Time: The Uses of History for Decision-Makers*" (mention b) refer to the same entity (a book or the title of a book). Since mention b is the full name of the book and mention a is a part of it (mention a doesn't appear alone elsewhere in the document), we believe that mention b is the result of a more reasonable entity boundary segmentation. Another example is "*Dyson*" and "*Dr.Dyson*" where the former has the obvious boundary error. We evaluate that our strategy ***split sentence for entity*** can find entity mentions with more reasonable boundaries, thereby increasing the proportion of correct entities. So we adopt this strategy (sentence-by-sentence prediction rather than the whole document's prediction, which is reasonable because there are no

Content	Method	chunk	Triple (Candidate Relation)
(1) The late Nobel Prize-winning Paul Crutzen, who popularized the word 'Anthropocene' in 2000, had also been a member of the group until he died on January 28, 2021. **(2)** The main goal of the AWG is providing scientific evidence robust enough for the Anthropocene to be formally ratified by the International Union of Geological Sciences (IUGS) as an epoch within the Geologic time scale. **(3)** Prior to the establishment of the Anthropocene Working Group in 2009, no research program dedicated to the formalization of the Anthropocene in the geologic time scale existed.	**ReLik**	sent.1	<Paul Crutzen, ***company***, Anthropocene> (company, ***founded by***, residence...)
		sent.2	- (contains, ***company***, shareholder of...)
		sent.3	- (contains, ***company***, founded by...)
		all sent.	<Paul Crutzen, ***company***, International Union of Geological Sciences > (company, residence, founded by...)
	Ours	all sent.	<Paul Crutzen, ***member of***, Anthropocene Working Group >; <Paul Crutzen, ***member of***, AWG >; <Paul Crutzen, ***award received***, Nobel Prize >...

Table 7: Comparison of ReLik and our pipeline for annotation. **-** means there is no relation triple extracted, though various candidates are identified by the Retriever. The term **chunk** refers to whether the input corresponds to a single sentence or a relatively complete text, irrespective of ReLiK's context length limitation.

Content	Method	Triple
.....Being largely successful, this netted Dell a $20 million yearly contract to keep on managing the School Districts systems.NASA's Kennedy Space Center had its IT services, approximately 22,000 devices, outsourced in 1998 for $30 million a year on a 3-year contract. The US Treasury outsourced their 1643 desktop and 700 portable seats in 1999 for around $27 million yearly.....	**REBEL**	<NASA, ***subsidiary***, Kennedy Space Center >; <Kennedy Space Center, ***parent organization***, NASA >
	Ours	<Dell, ***country***, U.S. >; <Kennedy Space Center, ***operator***, NASA >; <US Treasury, ***applies to jurisdiction***, U.S. >; <NASA, ***country***, U.S. >; .....

Table 8: Comparison of REBEL and our pipeline for annotation. We use a 502-token document as an example to ensure a complete input rather than segmented chunks. This table illustrates that (1) fine-grained relation types enhance semantic precision, (2) large language models outperform traditional pretrained models in capturing domain-specific relations, and (3) broader entity coverage increases relation diversity.

	Entity	Relation	Avg.
GPT-4o-mini	0.944	0.619	0.782
GPT-4o	0.870	0.762	0.816

Table 9: **Comparison of entities and relations** annotated by the GPT-4o-2024-11-20 and the GPT-4o-mini in three documents. The values in the table are the Recall of annotation results with different base models.

mention boundary cross-sentences) to directly find more mentions.

C.3 Analysis on Annotation Result

Incomplete Annotations Although LLMs supplement a substantial number of entities and relations, omissions remain inevitable. Besides, Since manual annotation only removed incorrect entity and relation labels without adding any new ones (due to the difficulty of annotating long texts), the test set of MMUIE may still miss certain correct entity and relation triples. Given sufficient time and resources, the test set could be progressively refined following the strategies used to improve the raw DocRED (Yao et al., 2019) dataset, as adopted by Re-DocRED (Tan et al., 2022), where models trained on MMUIE are employed to generate predictions that are subsequently verified and corrected through manual annotation.

Error Analysis and Mitigation By examining the portion of relation triples identified as incorrect during manual annotation, we found two types of errors that are difficult to eliminate through automated annotation, even with the inclusion of NLI model and secondary verification components. (1) ***Relation triples unrelated to the textual content (as discussed in Section 2.3).*** For this type of error, we believe that the LLM used for secondary verification have learned a large amount of factual knowledge during pretraining, while its instruction-following capability remains limited. As a result, it struggles to determine whether a given triple is both factually correct and explicitly supported by the text. However, such errors in the training set do not substantially impair the model's performance, particularly when the base model has already been exposed to extensive factual knowledge. (2) ***Errors in the subject/object form or order.*** Although the NLI component helps identify some inconsisten-

Methods	Specific	Quantity	Examples	Design Motivation
TPM for entity	No	34 (34)	<Digital Public Library of America, organization>	Follow the previous work (Yao et al., 2019) and utilize the capability of TPM model spaCy to generate general-type entities.
Type-guided extraction	No	13 (37)	<16 May 2017, date>	Increase the number of ***general-type entities*** and supplement TPM annotations.
Triple for entity†	Yes	25 (57)	<social sciences, academic discipline>derived from relation<sociology, subclass of, social sciences>	***Intuition***: (1) LLMs achieve high precision in extracting open-vocabulary relations where types are directly represented by predicates (Li et al., 2023b); (2) domain-specific documents generally contain predicates indicative of domain concepts; (3) entities involved in such relations are likely ***domain-relevant***.
Split sentence for entity†	Yes	54 (81)	<Dyson, person> to <Dr.Dyson, person>	Identify entities don't participate in any explicit relations, and ***increase the probability of entities with correct boundary***.
Extraction from KB	Yes	7 (7)	<twentieth century, has part(s), 1990s>	Utilize the relations recorded in KB, and construct a ***fine-grained set of relation types*** containing both general and domain-specific types.
Rel. generation via LLM	Yes	18 (25)	<New Economy, influenced by, Internet>	Harness LLMs' superior contextual comprehension to identify ***content-aware relations*** within lengthy documents.
Rel. supplement with TPM	Yes	47 (64)	<National Health Service, located in, United Kingdom >	Take full advantage of the existing document-level datasets to fine-tune an information extractor for ***mitigating the incomplete labeling*** of LLMs (Li et al., 2023b).

Table 10: **The design motivation of each step in the automated annotation pipeline.** We conduct a detailed analysis of each component's functionality and advantages through illustrative examples and statistical data. **Specific** refers to whether the method accounts for domain-relevant entity and relation types. Since relation extraction relies on entity recognition results (which include both general and domain-specific entities), each method has the potential to extract domain-specific relations. **Quantity** denotes the number of entity mentions or relations obtained in applying the corresponding strategy, and the number in parentheses is the quantity of new ones accumulated across the sequential steps. Appendix D.4 provides a detailed presentation of domain-specific entity and relation types across 34 domains. † means the type of entity extracted in this step is supplemented by the following **Type Annotation**.

cies, it is still difficult to detect triples in which the subject and object are mistakenly swapped or partially misidentified, especially when both entities legitimately co-occur in the text and are semantically related. Though these errors exist in our weakly supervised training set, our DocUIE still achieves consistent performance improvement in three different test sets.

C.4 Entity Recognition

Figure 7 is the **automated entity annotation analysis** in the remaining domains. (a) shows the coverage of entity extraction using spaCy and GPT4o-mini: The bottom purple indicates entities extracted only by LLM; The top section denotes entities extracted only by spaCy; The middle blue represents entities captured by both. while (b) shows the accuracy of entity identification using four different methods.

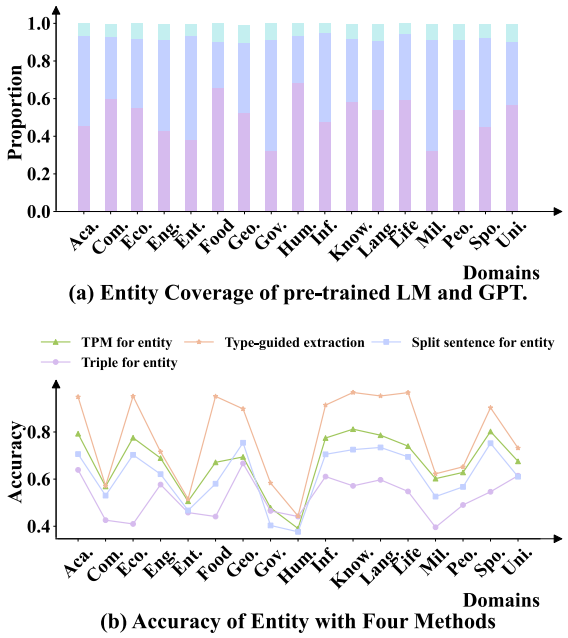


Figure 7: Automated Entity Annotation Analysis in remaining domains.

C.5 Relation Extraction

Table 18 shows the quantitative analysis of relation triples across different domains.

C.6 The Usage of Wikipedia

In the annotation of entity types, we take a textual mention (every entity mention extracted from previous steps), then query the Wikidata[1] API to find the corresponding Wikidata entity ID. Specifically, we first construct a GET request to the Wikidata API endpoint[2] with these parameters: (1) *action=wbsearchentities* tells the API to search for entities. (2) *format=json* asks for results in JSON. (3) *search=mention* is the text we are looking for. (4) *type=item* means the return of the function is Wikidata *"items"*, not *"properties"* or *"lexemes"*. Then, we send the GET request and hand back the corresponding Wikidata ID or *None* if nothing is found. It is worth noting that the type of these entities with a *None* response will be provided by the LLM in the following step. Besides, since the mentions refer to the same entity have the same IDs and types in Wikidata, they are combined as aliases of a canonical entity.

Since the Wikidata knowledge base contains a substantial volume of factual data, it serves as a valuable source for both relation triples and fine-grained relation types. Using the previously acquired entity IDs, we construct SPARQL queries for all possible entity pairs and submit these queries to the Wikidata SPARQL endpoint[3]. When a predefined relationship exists between two entities, the corresponding relation ID is returned. Following the same methodology employed in entity annotation, we then retrieve the textual representation of each relation by issuing a GET request with its relation ID.

Leveraging the Wikidata knowledge base, we enrich entity type information, extract recorded relation triples, and construct a fine-grained set of entity and relation types.

D Detailed Information of the Datasets

D.1 Wikipedia Categories

Wikipedia[4] is organized into 13 top-level categories:: *General reference*, *Culture and the arts*, *Geography and places*, *Health and fitness*, *History and events*, *Human activities*, *Mathematics and logic*, *Natural and physical sciences*, *People and self*, *Philosophy and thinking*, *Religion and belief systems*, *Society and social sciences*, and *Technology and applied science*. Each of these top-level categories encompasses multiple second-level categories, which contain a large number of nested subcategories, forming a hierarchical taxonomy. To construct our multi-domain dataset, we primarily sample from the second-level categories, systematically selecting two to three representative subcategories from each top-level category to ensure both domain diversity and balanced coverage. We focus on second-level categories because they are more specific than broad top-level categories, yet broad enough to include diverse and representative documents within each domain.

D.2 Basic Statistics of MMUIE

Table 23 shows detailed data for all domains, including the number of articles in the training set, development set, and test set, the number of relation types and entity types, and the proportion of domain-specific types. Table 24 shows the data after filtering out the low-frequency relation and entity types. And Table 11 compares the average data analysis across all domains before and after filtering out types that appeared fewer than 20 times. *Unique Rel.* refers to the count of relation types that are exclusive to a single domain, as well as the *Unique Ent.*

Data Information		Before	After	Δ
Document	Train set	221.53	220.76	0.35%↓
	Dev set	25.12	25.03	0.36%↓
	Test set	10.09	10.09	-
Type	Relation	233.91	153.68	34.3%↓
	Unique Rel.	4.74	2.68	43.5%↓
	Entity	702.91	154.32	78.1%↓
	Unique Ent.	87.97	7.47	91.5%↓
Entity Number		21784	18643	14.4%↓
Triple Number		15815	14045	11.2%↓

Table 11: Average data analysis across all domains before and after filtering out types that appeared fewer than 20 times.

D.3 Manual Verification

From the 8700 documents collected, 343 samples were carefully selected as the test set, ensuring at least 5 documents per domain. All entities, types, and relation triples in the test set were manually verified for accuracy. The remaining 8,357 documents serve as a weakly supervised training set. We

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

[2] https://www.wikidata.org/w/api.php

[3] https://query.wikidata.org/sparql

[4] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories#General_reference

uploaded all the test data to the Mechanical Turk[1] platform for verification, where each data point was evaluated by three annotators. The annotators were required to determine 1) whether entities could be extracted from the text; 2) whether the entity types were correct; 3) whether the relation triples were accurate and could be inferred from the text. To ensure the effectiveness of human annotation, we set the following thresholds for annotators: (i)The number of accepted HITs should be >2000; (ii) The historical acceptance rate of HITs should >98; (iii) The annotators must be native English speakers. The Kappa coefficient for all manually annotated data points reached 0.99, demonstrating a high level of consistency among annotators. Additionally, an annotation can only be retained if at least 2/3 of the annotators agree on it.

D.4 Domain Unique Types Distribution

Table 26 and 21 respectively show the domain-specific types of MMUIE across 34 domains and DocRED (Yao et al., 2019)/Re-DocRED (Tan et al., 2022) across four domains.

The types marked in red are those that exist both in the train set and test set. Table 21 presents the distribution of relation labels in the DocRED dataset across four specific domains, Art, Personal Life, Science, and Time, while the remaining labels are grouped under a general domain category. Our statistical analysis reveals that all named entity recognition (NER) labels in DocRED are associated exclusively with the general domain, indicating the absence of domain-specific entity type annotations. In the case of relation extraction (RE), approximately 50% of the relation labels are also categorized under the general domain. This skewed annotation distribution highlights a pronounced domain imbalance in the dataset: domain-specific relation types constitute a minority, while general-domain relations are predominant. Such a long-tail distribution can hinder the effectiveness of universal information extraction (UIE) models trained on DocRED, particularly when these models are applied to domain-specific corpora. The insufficient representation of domain-specific patterns in training data limits the model's ability to learn and generalize semantic relationships unique to specialized domains.

Moreover, the domain classification of relation labels in DocRED is often ambiguous, lacking well-defined boundaries. As illustrated in Table 20, certain relation types, such as "publication date", can be interpreted in multiple domain contexts. For example, this label may be seen as a temporal attribute, aligning with the Time domain, or as metadata associated with creative works, thus fitting the Art domain. This ambiguity in annotation further complicates domain adaptation, as UIE models may struggle to disambiguate relation types based on context, leading to degraded performance in domain-specific applications.

To improve the utility of DocRED for domain-adaptive tasks, a more balanced and delineated annotation scheme may be necessary. Enhancing domain coverage and refining label definitions could facilitate the development of models capable of capturing nuanced semantic patterns across both general and specialized domains.

D.5 Correlation between different domains of MMUIE

To better utilize our dataset and improve the model's cross-domain information extraction capabilities, we deeply analyzed the correlation between different domains from the perspective of mutual information (MI) (Kraskov et al., 2004). MI is a statistic used to quantify the amount of shared information between two variables. We employed mutual information to assess the extent to which data from one domain reduces the uncertainty of data from another domain and used it to measure the relevance between different domains. Specifically, we utilize the pre-trained model RoBERTa (Liu et al., 2019) to encode the semantics of all relation types and entity types, obtaining the vector representation of each of them. We then weighted and aggregated all type vectors within a domain based on their frequency of occurrence to derive the vector representation for each domain. By treating each domain as a random variable, we calculated the joint and marginal probability distributions between pairs of domains to obtain the mutual information matrix. On this basis, the Figure 8 shows the relevance between different domains.

To analyze the cross-domain transferability, we choose two different goal domains *Academy* and *Politics*. Specifically, the model was fine-tuned on the test sets of the five domains with the highest and lowest relevance to the goal domain. The results in Table 12 indicate that, despite the limited amount of training data, following domain

[1] https://www.mturk.com/

Model	RE		NER	
	Academy	Politics	Academy	Politics
Deepseek R1	7.02	6.25	21.42	19.32
Single FT	19.74	19.04	48.83	36.95
DocUIE	37.06	25.32	54.24	40.87
Relevance^{top}	14.85	**39.45**	42.51	35.57
Relevance^{bottom}	12.11	10.21	36.81	29.95

Table 12: Domain Relevance Analysis. Relevance^{top} refers to training using the test sets from the five most relevant domains, while Relevance^{bottom} represents the five least relevant domains.

relevance can significantly enhance the model's performance in the target domain. Additionally, in *Politics*, Relevance^{top} trained with a small but clean dataset outperformed DocUIE trained with a large amount of noisy data. Future research can further leverage our MMUIE dataset to explore the cross-domain capabilities of IE models.

E Different Prompts

Table 17 and 16 show the training example of the DocUIE. And Table 22 shows the different Prompts used in the automated Annotation Pipeline.

F Experiments

F.1 Configurations

Due to the relatively long input of document-level information extraction, which includes instructions, text, and target types, we employed the model's maximum context length of 8k (the maximum in-context limit in Llama3-8B-Instruct) for LoRA fine-tuning. We divide the weak supervision data into a training set and a development set in a ratio of 9:1. Hyperparameters are selected based on performance on the development set. During training, we utilized the AdamW optimizer with a learning rate of 1×10^{-4}, and a batch size of 2 per device. The number of training epochs ranged from 3 to 10, varying in different training settings. We used one A100 GPU for inference and six A100 GPUs for fine-tuning. When training on datasets including MMUIE, Re-DocRED (Tan et al., 2022), DWIE (Zaporojets et al., 2021), and SciERC (Luan et al., 2018), the process took nearly five days to complete.

F.2 Supporting Experiments

(1) To explain the reason for designing the methods *Relation supplementation with TPM* and present the limitation of closed-source LLMs in relation extraction, a more complex task than entity recognition , we train a model using the weak supervision data from MMUIE and test it on the Re-DocRED and DWIE datasets. Table 13 shows that the model fine-tuned on our MMUIE dataset(DocUIE) performs better than the closed-source model GPT-4o in the RE task.

Dataset	Model	F1	Recall	Precision
Re-DocRED	DocUIE	**52.1**	45.33	61.26
	GPT-4o	14.22	13.72	14.76
DWIE	DocUIE	**16.27**	36.25	16.27
	GPT-4o	3.88	4.02	3.75

Table 13: The comparison between GPT-4o and DocUIE, which is fine-tuned with the distant supervision data from MMUIE.

(2) Training-free evaluation of a series of models Table 25 shows the zero-shot results of the filtered test set in each domain. The few-shot relation extraction frameworks proposed by existing studies not only significantly outperform zero-shot methods but also achieve better performance than fine-tuned models. REPLM (Ozyurt et al., 2023) presents a novel framework for document-level in-context few-shot relation extraction and achieves outstanding performance in DocRED (67.47 in F1) and Re-DocRED (41.48[1] in F1) with GPT-4o as backbones. The table 15 shows the results of this few-shot framework in the MMUIE dataset.

(3) Analyze the performance of different architecture models on the MMUIE dataset For fair comparison, we subconsciously chose Llama3-8B-Instruct following previous work. But to find the best base model for the MMUIE, we supplemented zero-shot experiments on high-performing architectures model, including Mistral-7B-Instruct[2], Qwen2.5-7B[3], and Qwen3-8B[4]. Table 14 shows the average F1 scores across all domains. We observe that both models perform similarly, except for Qwen3-8B, which is a reasoning model. Qwen3-8B is concurrent with our work. We plan to train DocUIE based on the recently published reasoning model in future work.

[1] the result reported in *Cognitive Mirroring for DocRE*

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[3] https://huggingface.co/Qwen/Qwen2.5-7B

[4] https://huggingface.co/Qwen/Qwen3-8B

Task	Llama3	Qwen2.5	Qwen3	Mistral
RE	0.54	0.88	2.21	0.49
NER	7.52	8.98	16.76	0.41

Table 14: The average F1 scores across all domains testing in different architectures models.

F.3 Qualitative Analysis

To further analyze the remaining challenges in MMUIE and DocUIE, we provide the following qualitative analysis, including error analysis, as follows:

- **There is an imbalance in the capabilities of the DocUIE model across different domains.** As shown in Table 19, for example, in the RE task, the F1 score of the DocUIE in the *Academy* domain is 37.06, while it is only 5.59 in the *Universe*. In the NER task, the F1 score in the *Academy* is 54.24, while it is only 30.12 in the *Food and drink*. We suspect that the possible reason may lie in the imbalanced domain knowledge of the pretraining backbone. The richer the domain-specific pretraining corpus, the stronger the backbone's capabilities in that domain.
- **The extracted relations are not relevant to the content of the text.** For example, (*Burma*, *shares border with*, *Laos*): although the entities *Burma* and *Laos* appear in the document, the relation *shares border with* cannot be directly inferred from the context. This kind of error originates from the model generating relations based more on its own knowledge rather than strictly following the context.
- **Long-tail relation types are hard to learn and predict.** The model tends to generate relations and entities of general types, which is the general observation of all long-tail distribution problems. In the RE task, according to statistical analysis, the DocUIE model tends to predict relations like *country*, *located in the administrative territorial entity*, and *country of citizenship* (top 3 among all relations). However, some domain-specific relations are not predicted at all, such as *filming location*, *director*, and *military branch*. This phenomenon is due to the fact that general types occur more frequently than domain-specific ones.
- **Error occurs in the format of subject or object.** For example, (*People's Bank of China*, *country*, *Chinese*) is correct if the object is *China*, not *Chinese*.

F.4 Construction of easy version

A major characteristic of our MMUIE is the abundance of fine-grained entity and relation types, some of which have distinct domain-specific features. These domain-specific types occur less frequently and are inherently more challenging for models to learn. To analyze this effect, we separated the relation triples that appear more frequently across all domains to construct an *easy* version of MMUIE. Specifically, triples containing domain-specific types (as listed in Table 26), or types appearing fewer than 5 times within their respective domains were excluded from this version. Figure 6 compares the results in these two versions. The Recall and F1 score are improved in both models, especially the recall, while the Precision is slightly reduced in DocUIE. The notable gain in Recall indicates that the DocUIE tends to predict general types that appear more frequently in the training data.

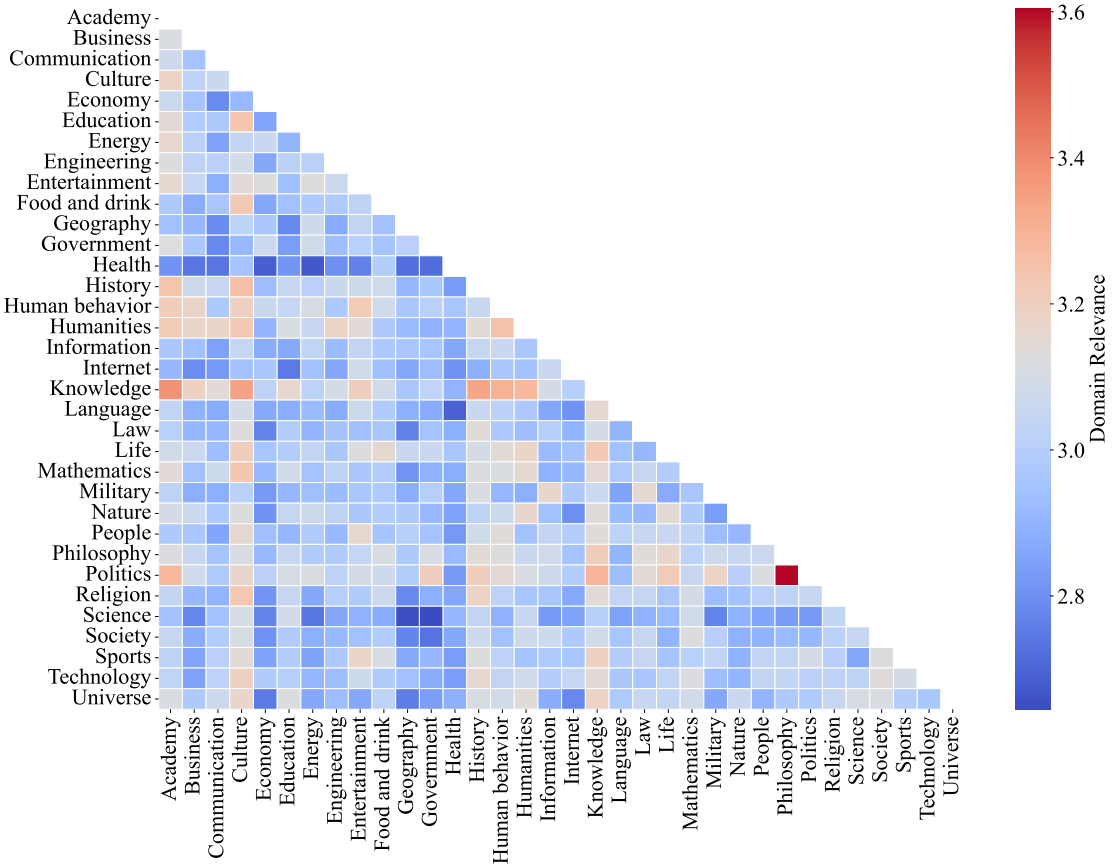


Figure 8: Domain Relevance Heatmap

Category	Domains	$RE.coref^{GT}$ GPT-4o	Llama3-8B
Fundemental Stadies	Mathematics	5.14	2.81
	Philosophy	7.10	3.29
	Language	5.01	6.09
	Science	0.66	2.34
	Knowledge	1.04	4.19
	Nature	1.34	4.07
	Universe	0.89	3.30
	Humanities	4.98	4.21
	Academy	8.50	6.72
Social Science	Government	6.70	4.07
	Society	4.12	9.60
	Religion	0.39	7.72
	People	4.37	7.81
	Human behavior	5.36	3.19
	History	6.80	5.90
	Culture	1.19	0.00
	Law	3.20	1.88
	Politics	3.23	4.12
Economics and Information	Communication	1.78	3.39
	Information	4.70	4.38
	Internet	4.31	7.14
	Business	2.33	1.87
	Economy	2.81	5.41
Engineering and Technology	Geography	5.48	4.41
	Energy	1.85	5.57
	Engineering	5.15	3.58
	Technology	3.36	3.14
	Military	4.81	3.83
Life and Health	Food and drink	6.18	9.73
	Health	4.44	6.09
	Sports	1.70	7.06
	Life	1.03	1.88
	Entertainment	5.60	7.24
	Education	5.52	7.92
Avg.		3.67	4.65

Table 15: F1 score of MMUIE dataset with few-shot relation extraction framework *REPLM (Ozyurt et al., 2023)*. We use distantly supervised training data as candidate context documents. Considering that the documents in MMUIE are longer, we set the number of context examples to 2.

I/O	Content
Input	You are an expert in information extraction within the **Health** domain. **Instruction:** Please extract relation triples that match the goal types from the document. Return an empty list if the relation type does not exist. Please respond in the format of a JSON string. **Goal types:** [*drug or therapy used for treatment*, has part, country, *medical condition treated*, official language, diplomatic relation, named after, *symptoms and signs*] **Document**: <Text>
Output	drug or therapy used for treatment: [{head: dimercaprol sulfonate, tail: mercury poisoning}, {head: dimercaprol, tail: mercury poisoning}...] has part: [] country: [{head: Kirkuk, tail: Iraq}, {head: Basra, tail: Iraq}...] *medical condition treated*: [] official language: [{head: World Health Organization, tail: English}] diplomatic relation: [{head: Mexico, tail: Sweden}...] named after: [] *symptoms and signs*: [{head: ataxia, tail: blindness}...]

Table 16: A training example for Relation Extraction.

I/O	Content
Input	You are an expert in information extraction within the **Law** domain. **Instruction:** Please extract entities that match the goal types from the document. Return an empty list if the entity type does not exist. Please respond in the format of a JSON string. **Goal types:** [event, occupation, organization, *type of crime*, location, *social theory*, group of humans, *legal concept*] **Document**: Carceral feminism is a critical term for types of feminism that advocate for enhancing and increasing prison sentences that deal with feminist and gender issues...
Output	event: [] occupation: [[feminist sociologist]...] organization: [[ACLU, American Civil Liberties Union], [Survived and Punished Organization]...] *type of crime*: [[domestic violence]...] location: [] *social theory*: [[Carceral feminism]...] group of humans: [] *legal concept*: []

Table 17: A training example for Entity Recognition.

Domain	KB		LLM		BOTH		NLI	Auto	Human	Kappa
	before	after	before	after	before	after				
Academy	0.5682	0.75	0.4539	0.5	0.5019	0.6045	209/306	34.93	17.53	1
Business	0.2352	0.3125	0.3296	0.3642	0.3005	0.3518	208/269	38.6	11.6	1
Communication	0.4605	0.8462	0.6766	0.8614	0.6173	0.8583	165/201	27.7	17.1	1
Culture	0.2328	0.3833	0.4244	0.509	0.3457	0.4758	197/274	46	15.9	0.9985
Economy	0.2222	0.4894	0.4474	0.5063	0.3683	0.5024	239/269	29.29	10.79	1
Education	0.2085	0.3836	0.5579	0.728	0.3919	0.601	174/237	44.4	17.4	0.9991
Energy	0.2159	0.4857	0.5267	0.78	0.3792	0.6588	196/253	48	18.2	1
Engineering	0.1338	0.3256	0.3716	0.3925	0.2358	0.3651	326/338	59.38	14	1
Entertainment	0.3913	0.5217	0.6551	0.8229	0.5568	0.697	185/234	30.83	17.17	1
Food and drink	0.2444	0.4225	0.4817	0.6167	0.3666	0.541	153/194	37.1	13.6	1
Geography	0.3333	0.5	0.538	0.6774	0.4708	0.6341	129/185	34.25	16.13	1
Government	0.2953	0.4127	0.5613	0.7699	0.3939	0.5233	200/259	75.89	29.89	1
Health	0.5	0.6667	0.4809	0.5495	0.485	0.5714	140/184	29.13	14.13	1
History	0.4061	0.5725	0.4023	0.5728	0.4042	0.5726	116/269	52.7	21.3	1
Human behavior	0.4737	0.537	0.6706	0.7815	0.6171	0.7052	168/256	26.92	16.62	1
Humanities	0.4639	0.5957	0.502	0.4805	0.4913	0.5075	205/252	31.45	15.45	1
Information	0.1684	0.875	0.3725	0.4286	0.2471	0.5278	216/306	59.46	14.69	1
Internet	0.4286	0.619	0.6402	0.8333	0.5873	0.7928	144/190	25.2	14.8	1
Knowledge	0.5263	0.5735	0.4701	0.5577	0.4841	0.5616	297/405	33.44	16.19	1
Language	0.5349	0.6667	0.3885	0.562	0.4249	0.5917	193/263	34.6	14.7	1
Law	0.5483	0.623	0.4932	0.6014	0.5205	0.6133	297/336	58.6	30.5	0.9926
Life	0.3644	0.5417	0.5188	0.609	0.4679	0.5912	201/242	32.45	15.18	1
Mathematics	0.4561	0.5526	0.4474	0.5614	0.4503	0.5579	94/117	34.2	15.4	1
Military	0.348	0.5263	0.3889	0.5368	0.3661	0.5307	166/206	49.78	18.22	1
Nature	0.3762	0.3529	0.4041	0.4718	0.396	0.4333	170/248	38.44	15.22	1
People	0.4054	0.5139	0.4667	0.5495	0.4362	0.5337	118/151	29.8	13	1
Philosophy	0.7347	0.8824	0.3734	0.4103	0.4589	0.5166	139/170	29.57	13.57	1
Politics	0.3234	0.4439	0.5526	0.7791	0.4061	0.5461	156/192	52.7	21.4	1
Religion	0.4286	0.5313	0.3589	0.4545	0.3854	0.4868	146/196	31.4	12.1	1
Science	0.5034	0.6322	0.4799	0.6268	0.488	0.6288	213/276	46.44	22.67	1
Society	0.4214	0.5588	0.6301	0.7218	0.5423	0.6667	171/220	34.36	18.64	1
Sports	0.3795	0.6406	0.3598	0.4865	0.3697	0.558	115/165	36.67	13.56	1
Technology	0.3545	0.4528	0.4406	0.4953	0.4103	0.4813	167/204	52	21.33	1
Universe	0.4211	0.4762	0.4925	0.525	0.4667	0.5082	119/137	42	19.6	1
All	0.3855	0.5491	0.4811	0.5919	0.4363	0.5675	6132/8004	40.39	16.99	0.9997

Table 18: Quantitative Analysis of Automated Annotation of Relation Triples from KB and LLM. **KB** refers to the triples extracted from the Wikidata knowledge base. **LLM** refers to the triples generated by LLM, especially GPT-4o-mini. *before* and *after* mean the accuracy before and after ***secondary verification*** by GPT-4o. **BOTH** reflects the overall accuracy of all the triples. **Auto** counts the number of triples extracted by the automated annotation pipeline, and **Human** counts the number of remaining triples after manual verification. **Kappa** refers to the Kappa coefficient, which is a statistical measure of inter-rater agreement for qualitative (categorical) items.

Domains	RE			RE.coref			NER			NER.coref		
	Single	Mix	Δ(%)	Single	Mix	Δ(%)	Single	Mix	Δ(%)	Single	Mix	Δ(%)
Mathematics	23.28	30.45	30.8	25.40	33.86	33.3	44.52	43.73	-1.8	38.13	37.01	-2.9
Philosophy	14.28	22.7	59.0	14.85	23.31	57.0	43.77	52.96	21.0	39.53	47.52	20.2
Language	10.79	17.24	59.8	12.45	18.30	47.0	38.90	48.26	24.1	34.30	40.13	17.0
Science	14.22	17.83	25.4	17.31	22.18	28.1	38.51	48.42	25.7	30.39	40.59	33.6
Knowledge	14.53	19.40	33.5	16.06	20.15	25.5	38.09	41.86	9.9	32.79	34.67	5.7
Nature	17.94	23.99	33.7	18.27	24.72	35.3	38.13	44.15	15.8	32.83	39.20	19.4
Universe	5.02	5.59	11.4	5.02	5.59	11.4	36.87	40.69	10.4	32.55	34.79	6.9
Humanities	13.02	28.80	121.2	13.28	28.80	116.9	40.26	46.12	14.6	37.12	41.79	12.6
Academy	19.74	37.06	87.7	19.94	38.53	93.2	48.83	54.24	11.1	42.11	48.92	16.2
Government	4.45	17.64	296.4	4.97	18.64	275.1	30.85	36.47	18.2	23.00	28.72	24.9
Society	22.94	29.70	29.5	23.20	30.52	31.6	23.40	31.91	36.4	20.67	26.99	30.6
Religion	12.77	17.07	33.7	12.77	18.87	47.8	41.61	49.69	19.4	37.85	45.08	19.1
People	18.24	30.12	65.1	22.00	30.50	38.6	35.74	38.61	8.0	31.05	33.64	8.3
Human behavior	12.77	24.55	92.2	13.59	24.91	83.3	17.74	24.36	37.3	14.50	21.82	50.5
History	20.92	32.15	53.7	21.11	32.65	54.7	46.35	54.22	17.0	39.20	44.66	13.9
Culture	15.55	17.27	11.1	16.85	18.34	8.8	39.20	44.16	12.7	34.37	38.50	12.0
Law	21.83	27.64	26.6	24.06	28.48	18.4	46.69	51.90	11.2	38.33	44.07	15.0
Politics	19.04	25.32	33.0	21.11	29.17	38.2	36.95	40.87	10.6	32.90	34.38	4.5
Communication	5.73	12.50	118.2	7.16	14.29	99.6	25.82	32.81	27.1	22.52	28.85	28.1
Information	21.80	29.21	34.0	24.56	31.03	26.3	43.79	48.71	11.2	37.45	42.34	13.1
Internet	24.18	33.04	36.6	29.80	36.21	21.5	32.34	35.36	9.3	26.12	31.01	18.7
Business	9.19	12.92	40.6	9.19	13.37	45.5	35.40	40.86	15.4	32.28	35.06	8.6
Economy	11.64	22.40	92.4	16.01	24.07	50.3	41.25	52.80	28.0	33.46	47.31	41.4
Geography	17.59	36.26	106.1	17.93	37.01	106.4	36.05	41.82	16.0	31.01	36.12	16.5
Energy	12.79	15.86	24.0	13.07	16.88	29.2	30.07	35.12	16.8	23.99	29.55	23.2
Engineering	13.91	16.17	16.2	15.83	20.15	27.3	39.89	45.28	13.5	34.44	38.14	10.7
Technology	13.36	21.31	59.5	13.36	21.86	63.6	43.65	49.21	12.7	38.79	43.70	12.7
Military	14.36	26.35	83.5	16.65	27.83	67.1	50.83	55.20	8.6	43.22	48.48	12.2
Food and drink	22.38	31.24	39.6	22.38	31.24	39.6	24.64	30.12	22.2	20.60	27.94	35.6
Health	15.66	28.25	80.4	18.79	31.09	65.5	33.76	37.65	11.5	25.16	30.11	19.7
Sports	14.64	22.22	51.8	16.82	23.93	42.3	48.24	51.66	7.1	42.23	44.74	5.9
Life	10.45	14.19	35.8	10.45	16.02	53.3	32.57	40.83	25.4	29.86	36.67	22.8
Entertainment	12.09	21.93	81.4	12.46	24.02	92.8	37.66	41.74	10.8	29.94	34.88	16.5
Education	12.40	22.78	83.7	13.99	25.51	82.3	30.18	37.63	24.7	28.01	32.72	16.8
Avg.	15.10	23.35	61.9	16.49	24.83	58.0	37.43	43.22	16.5	32.08	37.36	17.9

Table 19: F1 score of each domain under single-domain train setting and mix-domain training with Llama3-8B as base model.

Labels	Domain 1	Domain 2	Domain 3
original language of work	Art	Science	-
languages spoken, written or signed	Art	politics	-
located in the administrative territorial entity	General	politics	-
contains administrative territorial entity	General	politics	-
applies to jurisdiction	General	Law	-
legislative body	General	Law	-
developer	General	Business	-
subsidiary	General	Business	-
parent organization	General	Business	-
operator	General	Business	-
member of sports team	General	Sport	-
head of government	General	politics	-
head of state	General	politics	-
ethnic group	General	politics	-
headquarters location	General	politics	Business
country of origin	General	Business	-
chairperson	General	politics	Business
employer	Personal Life	Business	-
founded by	Personal Life	Business	-
member of	Personal Life	politics	Business
publication date	Time	Art	-

Table 20: Labels with ambiguous domains in DocRED/Re-DocRED

Domain	Num. of Label	Unique Relation Type
Art	20	performer, composer, record label, lyrics by, director, screenwriter, cast member, producer, production company, creator
Personal Life	42	father, mother, spouse, child, sibling, educated at, residence
Science	17	parent taxon, located on terrain feature, instance of, subclass of
Time	10	start time, end time, point in time, publication date, date of death, date of birth

Table 21: Unique Relation Distribution in DocRED/Re-DocRED.

Iuput	Content
Type-guided extraction	**Instruction:** You are an expert in document-level named entity recognition. Please extract entities belonging to any type in the type list from the document. Please respond in the format of a JSON string as follows: [{entity: entity1, "type": type1},...]. If you cannot find any entity with the type, please respond with an empty list. **type list:** [person, facilities, location, products, work of art, data...] **Document:** <Text>
Triple for entity	**Instruction:** You are an expert in document-level relationship extraction, both within and across sentence relationships. Please extract the underlying relationship from the document. Please respond in the format of a JSON string as follows: ["head": "head1", "relation": "relation1", "tail": "tail1",...]. If you cannot find any relationship, please respond with an empty list. **Document:** <Text>
Split sentence for entity	**Instruction:** Please extract the entities in the sentence. Please respond in the format of a JSON string as follows: ['entity': 'entity1', 'type': 'type1',...]. If you cannot find any entity, please respond with an empty list. **Sentence:** <Sentence Text>
Secondary verification	**Instruction:** I have a piece of text and a list of entities extracted from it. I need you to help me determine whether each entity is a correct entity extracted from the text. If it is correct, please mark it as "Correct" and identify the type of entity. If it is not correct, please mark it as "Incorrect" and explain the issue (e.g., spelling error, incomplete, does not match the text, etc.). **Format:** The format of the response should in JSON format. For each entity, it should be "entity": "entity1", "status": Correct, "type": "type1" or "entity": "entity1", "status": Incorrect, "reason":"reason". **Definitions :** An entity is a specific item, object, or concept that has a distinct and significant meaning within the context of the text. Examples include names of people, organizations, locations, dates, events, and other specific items. For the purpose of Named Entity Recognition (NER), a correct entity must: 1. Accurately represent the entity: The content should clearly and precisely refer to the entity it represents. 2. Match the text exactly: The content must appear in the text exactly as it is written, including case sensitivity, punctuation, and spacing. 3. Be contextually relevant: The content should be relevant and meaningful within the context of the text. **Document:** <Text> **Entities:** [All Entities extracted through the above steps]
Type annotation with LLM	**Instruction:** I have a list of entities extracted from a text, and I need to assign a type in types to each entity. Please help me identify the type of each entity based on the context and types provided. Read the list of entities provided. 1. Identify the type of each entity based on the context provided. 2. Assign a type to each entity based on the context. 3. Provide the types in the following format: Entity 1: Type 1, Entity 2: Type 2, ... **Entities with context:** <Entity, Sentence> **Types:** [type1, type2...]
Relation generation via LLM	**Instruction:** I have a document and a list of entities extracted from it. I need you to help me identify and describe the relationships between these entities based on the context in the document and the relation label set provided. Read the document and the list of entities provided. 1. Identify relationships between the entities based on the context provided in the document. 2. Please use the relationship in the labelSet. 3. You must respond with a complete list of triples. Provide the types in the following JSON format: ["head": "entity1", "relation": "relationship", "tail": "entity2",...]. **Document:** <Text> **Label Set:** [relation1, relation2...] **Entity List:** [entity1, entity2, entity3...]
Secondary verification	**Instruction:** I have a document and a list of relation triples extracted from it. I need you to help me determine whether the triple is correct and can be directly extracted or inferred from the text content. Generate the response in the following JSON format: [{"triple": {"subject": "entity1", "relation": "relation1", "object": "entity2"}, "status": "Correct/Incorrect", "reason": "reason"}] **Document:** <Text> **Relation triple List:** [{"subject": "entity1", "relation": "relation1", "object": "entity2"},...]

Table 22: **Prompts in Annotation with LLM.**

Domain	train	dev	test	rel.	uniq.(%)	ent.	uniq.(%)	entities	alias	triples	inter(%)	sent.
Academy	270	30	15	253	2.77	955	10.37	27180	567	19592	51.59	13427
Business	243	27	10	224	0.45	554	12.45	21884	438	16418	47.42	10647
Communication	100	12	10	259	1.16	549	9.65	9584	211	7183	48.34	5925
Culture	268	30	10	252	1.98	753	16.75	28737	616	21252	50.47	14077
Economy	255	29	14	256	3.52	107	4.67	24801	488	19513	44.69	13230
Education	279	31	10	227	2.20	335	8.36	28157	476	20324	48.69	13310
Energy	195	22	10	153	1.96	538	16.91	21110	339	13287	41.09	9793
Engineering	202	23	13	190	4.21	947	13.83	18338	286	10422	40.71	11313
Entertainment	273	31	12	244	2.46	533	12.95	27775	415	20737	46.35	13485
Food and drink	192	22	10	150	1.33	355	10.14	21713	160	10406	42.66	11035
Geography	160	18	8	211	1.90	592	11.82	16034	372	15202	44.28	7678
Government	129	15	9	139	0.72	402	9.45	12491	341	6723	38.27	5856
Health	241	27	8	242	1.65	884	11.88	23870	526	14641	45.02	13539
History	262	30	10	262	1.91	842	15.20	31241	734	26026	48.25	15885
Human behavior	248	28	13	274	1.09	1001	12.78	22486	449	16966	52.09	13597
Humanities	263	30	11	265	0.75	994	12.78	26721	536	20538	47.42	13891
Information	162	18	13	221	0.45	669	11.51	15294	375	9616	44.86	8232
Internet	192	22	10	217	0.46	572	11.01	18623	321	13489	44.07	9389
Knowledge	239	27	16	229	0.87	834	11.15	21789	469	16304	47.80	12051
Language	207	23	10	193	0.00	781	18.31	19075	604	14356	38.96	11716
Law	270	31	10	285	0.83	456	10.09	23837	603	20280	39.93	13539
Life	263	30	11	285	4.91	493	7.91	26308	767	17341	45.76	14678
Mathematics	129	15	5	191	3.14	462	9.52	11784	203	8616	49.45	6168
Military	256	29	9	268	2.61	776	16.49	26645	605	20060	48.65	13362
Nature	252	29	9	274	1.82	1024	15.14	23385	623	17071	44.62	13631
People	229	26	10	284	2.46	782	10.36	23717	550	20510	52.52	11992
Philosophy	192	22	7	215	0.93	689	7.84	16994	264	11986	48.69	9904
Politics	293	33	10	289	1.38	954	11.95	28598	730	23678	51.06	15146
Religion	148	17	10	266	3.01	671	9.24	16982	429	12186	47.97	8636
Science	244	28	9	240	1.25	924	11.80	26230	526	17270	54.49	12839
Society	134	15	11	224	0.45	641	8.74	12240	236	8991	45.93	7032
Sports	254	29	9	224	3.13	737	13.57	23212	480	18326	46.90	13874
Technology	221	25	6	229	3.93	889	14.06	20773	427	13603	46.28	11364
Universe	267	30	5	263	5.32	1204	19.02	23050	466	14802	41.75	15033
All	7532	854	343	703	22.9	6018	49.7	740658	15632	537715	46.86	395274

Table 23: Raw Data Detail before filtering low-frequency entity and relation types. **entities** and **triples** refer to the number of entities and triples in the corresponding domain. And **inter(%)** refers to the proportion of cross-sentence ones in the total number of triples. **alias** means the number of entities with an alias. **sent.** reflects the length of the document.

Domain	train	dev	test	rel.	uniq.(%)	ent.	uniq.(%)	entities	alias	triples	inter(%)	sent.
Academy	266	30	15	142	2.82	106	13.21	22360	2504	16555	49.60	13162
Business	243	27	10	202	0.99	200	4.00	19155	2581	15714	42.48	10647
Communication	100	12	10	75	1.33	75	1.33	7748	820	4916	50.53	5925
Culture	268	30	10	151	1.32	111	1.80	24757	2962	18783	48.35	14077
Economy	255	29	14	133	3.01	27	0.00	22132	3089	18377	42.12	13230
Education	279	31	10	139	3.60	98	0.00	24910	3209	19215	46.28	13310
Energy	195	22	10	102	2.94	116	2.59	17939	2434	12202	38.40	9793
Engineering	195	22	13	105	4.76	199	16.08	14733	1562	7733	38.38	10817
Entertainment	273	31	12	155	2.58	134	2.24	24892	2613	19502	43.73	13485
Food and drink	192	22	10	88	1.14	113	0.88	19309	2200	9590	38.87	11035
Geography	160	18	8	121	1.65	169	2.37	13965	1515	13484	42.24	7678
Government	129	15	9	78	0.00	147	2.72	10721	1653	5945	35.34	5856
Health	241	27	8	130	4.62	253	5.93	20465	2274	12238	42.89	13539
History	262	30	10	157	0.64	234	4.27	27414	3370	24019	46.65	15885
Human behavior	248	28	13	142	0.00	74	1.35	18259	1977	13414	51.22	15397
Humanities	263	30	11	146	0.00	128	1.56	22638	2367	17306	44.67	13891
Information	162	18	13	112	0.89	120	0.83	12919	1660	7846	42.68	8232
Internet	192	22	10	142	1.41	123	0.00	15941	2390	12387	39.74	9389
Knowledge	239	27	16	171	1.75	179	2.23	18537	2236	14646	44.97	12051
Language	207	23	10	103	0.97	71	23.94	16589	1865	12134	36.39	11716
Law	270	31	10	168	2.38	141	1.42	20753	2985	19258	37.42	13539
Life	263	30	11	193	3.13	158	0.63	23496	2969	16253	43.38	14678
Mathematics	129	15	5	153	1.96	167	3.59	10116	1166	7834	46.73	6168
Military	256	29	9	203	1.97	203	3.94	22847	3133	18562	46.05	13362
Nature	250	28	9	222	1.80	282	11.35	19979	2127	14986	43.53	13443
People	229	26	10	160	0.00	70	1.43	19786	2409	17694	48.29	11992
Philosophy	192	22	7	149	0.00	100	0.00	14184	1483	10162	43.66	9904
Politics	293	33	10	208	0.96	141	0.71	24075	3265	21219	48.96	15146
Religion	148	17	10	194	2.58	160	2.50	14609	1761	11023	45.01	8636
Science	244	28	9	188	0.00	200	1.50	22399	2677	15887	49.53	12839
Society	133	15	11	189	0.53	201	2.49	10363	1076	7851	45.51	6968
Sports	254	29	9	202	2.58	203	3.94	20173	2381	16983	45.12	13874
Technology	218	25	6	184	1.09	221	4.52	17342	2223	11730	43.86	11199
Universe	258	29	5	217	3.69	323	15.79	19105	1750	12068	40.20	14560
All	7506	851	343	456	19.96	723	35.13	633880	76686	477516	44.30	939623

Table 24: Data Detail after filtering the low frequent relation and entity types($<$ 20 times). **entities** and **triples** refer to the number of entities and triples in the corresponding domain. And **inter(%)** refers to the proportion of cross-sentence ones in the total number of triples. **alias** means the number of entities with an alias. **sent.** reflects the length of the document.

Task	Domain	gpt4o			gpt4.1			deepseek-r1			iepile			llama3-8b			qwen2.5-7b			qwen3-8b		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RE	Academic	3.5	4.56	3.96	2.45	9.13	3.86	3.13	11.79	4.95	12	1.14	2.08	0.32	1.35	0.52	0.38	1.14	0.57	0.61	2.58	0.99
	Business	2.42	2.59	2.5	0.78	3.45	1.27	1.63	5.17	2.48	25	0.86	1.66	0	0	0	0	0	0	0.80	3.30	1.29
	Communication	6.31	4.09	4.96	2.61	6.26	3.49	1.91	5.85	2.88	0	0	0	0.71	2.8	1.13	1.56	2.34	1.87	1.84	2.84	2.23
	Culture	5.26	3.77	4.39	2.17	6.29	3.23	3.89	11.95	5.87	28.57	1.26	2.41	0.18	1.00	0.31	0.77	1.26	0.96	0.25	0.69	0.37
	Economy	1.36	1.32	1.34	3.29	11.26	5.09	1.97	6.62	3.04	0	0	0	0.29	1.82	0.50	0	0	0	2.20	5.41	3.13
	Education	3.74	2.3	2.85	0.87	1.72	1.16	5.07	10.92	6.92	0	0	0	2.17	3.06	2.54	1.96	0.57	0.88	0	0	0
	Energy	5.17	4.95	5.06	3.9	7.14	5.04	2.85	7.14	4.07	12.5	0.55	1.05	0.30	1.22	0.48	1.43	0.55	0.79	4.30	7.48	5.46
	Engineering	1.76	3.3	2.3	1.56	6.59	2.52	1.25	6.04	2.07	0	0	0	0.21	1.01	0.35	0	0	0	0.25	0.47	0.33
	Entertainment	3.23	3.4	3.31	1.59	3.88	2.26	3.19	10.19	4.86	12.5	0.49	0.94	0.36	1.25	0.56	0.73	0.97	0.83	0.71	2.76	1.13
	Food_and_drink	2.14	5.56	3.09	0.85	5.15	1.47	0.4	2.21	0.68	0	0	0	0.70	3.64	1.17	0.75	0.74	0.74	1.06	2.70	1.52
	Geography	5.56	1.55	2.42	8.7	9.3	8.99	10	11.63	10.75	12.5	0.78	1.47	1.34	4.65	2.08	3.79	3.88	3.83	8.14	14.29	10.37
	Government	5	5.95	5.43	2.45	9.29	3.88	2.26	9.67	3.66	11.11	0.37	0.72	0	0	0	0	0	0	1.02	2.23	1.40
	Health	1.74	1.77	1.75	1.49	4.42	2.23	1.69	4.42	2.45	0	0	0	0.20	1.61	0.36	0.40	0.88	0.55	0	0	0
	History	6.67	5.16	5.82	3.2	7.98	4.57	3.92	11.27	5.82	12.5	0.47	0.91	0.64	1.29	0.86	0.45	0.47	0.46	0.72	2.27	1.09
	Human_behavior	1.95	1.39	1.62	2.59	6.48	3.7	2.62	6.94	3.8	33.33	0.93	1.81	1.18	1.71	1.4	1.08	1.39	1.22	0.70	2.03	1.04
	Humanities	5.62	5.29	5.45	1.81	4.71	2.62	1.96	6.47	3.01	0	0	0	0	0	0	0	0	0	0.42	1.18	0.62
	Information	25.98	17.28	20.76	9.34	14.14	11.25	10.6	16.75	12.98	18.18	1.05	1.99	0	0	0	0	0	0	2.36	4.49	3.09
	Internet	18.46	8.11	11.27	9.15	9.46	9.3	7.54	10.14	8.65	0	0	0	0.70	3.38	1.16	2.56	0.68	1.07	1.43	1.49	1.46
	Knowledge	2.94	2.7	2.81	1.42	3.09	1.95	3.14	8.88	4.64	0	0	0	0.13	0.44	0.20	0	0	0	0	0	0
	Language	8.24	9.52	8.83	4.61	13.61	6.89	4.11	10.2	5.86	16.67	0.68	1.31	1.07	2.04	1.40	0.77	1.36	0.98	3.68	5.22	4.32
	Law	2.76	1.31	1.78	4.86	7.21	5.81	7.38	14.43	9.77	12.5	0.33	0.64	0	0	0	0.45	0.33	0.38	1.43	1.97	1.66
	Life	5.56	2.99	3.89	3.45	8.38	4.89	4.08	13.17	6.23	0	0	0	0.26	1.10	0.42	0	0	0	0.77	1.20	0.94
	Mathematics	8.75	9.09	8.92	11.72	22.08	15.31	9.13	24.68	13.33	20	1.3	2.44	1.21	6.06	0.36	33.33	1.3	2.5	5.04	10.61	6.83
	Military	5.07	6.71	5.78	3.84	10.37	5.6	4.66	13.41	6.92	5.88	0.61	1.11	0.23	0.79	2.02	0.38	0.61	0.47	0.93	4.10	1.52
	Nature	6.98	4.38	5.38	4.27	8.76	5.74	4.53	9.49	6.13	12.5	0.73	1.38	0.74	1.30	0.94	0	0	0	0.89	2.74	1.34
	People	4.76	4.62	4.69	2.88	6.92	4.07	4.51	14.62	6.89	0	0	0	0.16	1.96	0.3	0.98	0.77	0.86	0.80	2.08	1.16
	Philosophy	15.19	12.63	13.79	10.85	24.21	14.98	9.28	32.63	14.45	12.5	1.05	1.94	8.11	4.23	5.56	0.98	1.05	1.01	9.23	12.00	10.43
	Politics	2.63	2.8	2.71	2.27	5.61	3.23	2.64	8.41	4.02	0	0	0	0	0	0	0	0	0	2.56	4.00	3.12
	Religion	7.3	8.26	7.75	5.54	15.7	8.19	4.76	17.36	7.47	10	0.83	1.53	0	0	0	2.88	5.79	3.85	0.68	2.02	1.02
	Science	17.65	8.82	11.76	6.45	11.76	8.33	5.71	11.27	7.58	25	0.49	0.96	0.42	2.46	0.72	3.94	2.45	3.02	0.37	0.78	0.50
	Society	0.9	1.59	1.15	1.44	3.9	2.1	2.49	6.83	3.65	0	0	0	0	0	0	0.52	0.49	0.50	0.66	1.69	0.95
	Sports	4.37	6.65	5.25	1.16	3.28	1.71	3.31	14.75	5.41	11.11	0.82	1.53	0.37	2.88	0.66	0	0	0	3.73	8.20	5.13
	Technology	2 2.34	2.16	2.18	7.81	3.41	3.9	14.06	6.11	50	0.78	1.54		0	0	0	2.68	2.34	2.50	0.56	1.56	0.82
	Universe	1.02	1.02	1.02	1.52	5.1	2.34	1.42	6.12	2.31	25	1.02	1.96	0	0	0	0	0	0	0	0	0
NER	Academic	24.6	38.96	30.16	15.61	52.39	24.05	13.69	49.2	21.42	26.72	30.99	28.7	17.15	17.38	17.26	7.64	21.12	11.22	18.81	17.09	17.91
	Business	10.03	30.43	15.09	1.74	37.46	3.33	10.91	40.64	17.2	19.08	23.22	20.95	4.38	6.34	5.18	3.34	7.00	4.52	12.50	15.89	13.99
	Communication	17.59	24.7	20.55	13.09	35.35	19.11	13.58	34.87	19.55	18.59	21.07	19.75	3.30	3.51	3.40	5.98	12.25	8.04	17.25	13.08	14.88
	Culture	17.61	25.32	20.77	14.37	41.77	21.38	13.54	41.05	20.36	22.27	20.45	21.37	6.04	5.66	5.84	5.47	14.79	7.99	12.32	17.36	14.41
	Economy	6.72	33.33	11.18	9.89	49.11	16.46	9.47	48.72	15.86	28.33	32.13	30.11	16.23	16.45	15.29	15.76	18.71	17.11	7.72	11.64	9.28
	Education	13.38	20.9	16.32	9.2	33.92	14.47	7.8	30.68	12.44	24.35	31.13	27.33	12.45	13.45	12.93	9.55	18.58	12.62	11.63	23.60	15.58
	Energy	10.15	19.32	13.31	11.16	41.87	17.62	9.64	38.29	15.4	15.33	19.6	17.2	3.03	4.65	3.67	2.85	8.81	4.31	7.81	13.29	9.84
	Engineering	22.66	31.96	26.52	9.94	48.55	16.5	17.81	42.01	25.01	17.24	26.49	20.89	5.05	8.49	6.33	4.80	17.29	7.51	20.21	18.28	19.20
	Entertainment	21.92	30.93	25.66	14.38	44.23	21.7	15.4	43.19	22.7	17.59	26.96	21.29	4.78	7.74	5.91	5.20	27.21	8.73	14.99	15.30	15.14
	Food_and_drink	7.37	17.94	10.45	5.58	30.84	9.45	6.62	31.71	10.95	12.87	11.31	12.04	4.99	4.61	4.79	4.55	9.96	6.25	6.59	11.42	8.36
	Geography	19.95	19.95	19.95	18.47	44.17	26.05	18.01	47.32	26.07	17.79	21.14	19.32	6.06	13.91	8.44	5.83	12.17	7.88	17.61	14.72	16.04
	Government	17.91	27.66	21.74	17.33	44.47	24.96	16.08	41.91	23.24	16.06	30.45	21.03	4.99	5.68	5.31	4.23	16.01	6.69	16.31	19.57	17.79
	Health	18.88	20.13	19.48	13.38	29.87	18.48	12.73	27.88	17.48	10.28	13.52	11.68	3.12	4.95	3.83	2.82	13.68	4.68	12.48	15.71	13.91
	History	18.02	32.27	23.13	14.81	46.69	22.46	13.93	45.53	21.33	16.93	32.32	22.22	11.95	12.56	12.25	12.20	25.40	16.48	12.83	22.36	16.30
	Human_behavior	8.88	24.82	13.08	6.25	34.06	10.56	5.98	33.33	10.14	9.13	16.45	11.74	6.41	4.55	5.32	4.62	13.63	6.90	6.99	18.25	10.11
	Humanities	24.01	33.2	27.87	20.79	43.93	28.22	17.09	44.53	24.7	18.09	25.26	21.08	8.54	15.11	10.91	8.64	25.67	12.93	19.59	21.46	20.48
	Information	28.17	34.9	31.2	16.23	47.75	24.23	17.45	43.94	24.98	21.42	27.31	24.01	8.57	17.53	11.51	6.03	16.42	8.82	17.43	30.80	22.26
	Internet	25.14	31.75	28.06	19.66	42.34	26.85	16.19	37.59	22.63	13.86	22.97	17.29	4.73	9.33	6.28	3.38	11.19	5.18	19.13	25.55	21.88
	Knowledge	13.23	27.65	17.9	10.45	42.98	16.81	10.83	43.7	17.36	12.42	16.89	14.31	7.22	11.14	8.76	3.84	13.20	5.95	13.16	18.62	15.42
	Language	15.51	36.57	21.78	20.66	51.49	29.49	20.54	43.78	27.96	13.54	18.11	15.5	11.34	12.56	11.92	10.01	19.49	13.23	15.30	17.66	16.40
	Law	19.3	31.95	24.06	14.1	39.82	20.83	12.46	43.24	19.35	18.96	22.12	20.42	8.22	22.29	12.01	8.23	16.98	11.09	10.82	16.34	13.02
	Life	18.13	27.86	21.97	12.19	43.82	19.07	12.68	43.35	19.62	19.06	23.2	20.93	10.54	9.12	9.78	5.04	11.86	7.07	10.80	21.91	14.47
	Mathematics	31.72	31.38	31.55	23.84	40.96	30.14	25	50.53	33.45	14.15	33.52	19.9	4.89	10.06	6.58	4.27	16.76	6.81	24.12	25.53	24.80
	Military	23.02	45.82	30.64	23.38	47.26	31.28	23.95	52.98	32.99	17.56	37.62	23.94	4.57	10.14	6.30	5.08	21.98	8.25	20.86	30.07	24.63
	Nature	20.36	24.35	22.18	18.07	42.46	25.35	17.67	37.28	23.98	12.17	16.85	14.13	7.6	5.6	6.45	13.67	21.12	16.60	15.59	21.34	18.02
	People	17.79	23.63	20.3	14.97	42.36	22.12	13.72	40.06	20.44	20.52	25	22.54	3.95	4.59	4.25	4.92	10.79	6.76	11.39	6.67	8.41
	Philosophy	37.03	47.88	41.76	29.85	46.58	36.38	30.48	55.7	39.4	18.94	34.44	24.44	21.05	5.76	9.04	22.51	34.53	27.25	37.65	40.72	39.12
	Politics	14.62	29.62	19.58	4.25	50.78	7.84	12.16	46.99	19.32	11	23.28	14.94	5.57	11.95	7.60	3.03	14.73	5.03	14.69	28.73	19.44
	Religion	12.85	40.71	19.53	18.89	46.1	26.8	15.91	46.84	23.75	13.49	25.57	17.66	6.33	21.88	9.82	5.66	22.61	9.05	15.93	24.72	19.37
	Science	18.02	30.32	22.61	12.97	38.74	19.43	13.62	43.37	20.73	15.87	22.65	18.66	4.85	6.99	5.73	5.33	13.20	7.59	14.29	22.53	17.49
	Society	16.15	31.54	21.36	11	39.72	17.23	10.02	37.85	15.85	11.18	21.03	14.6	3.91	6.62	4.92	2.25	10.64	3.71	10.84	24.30	14.99
	Sports	24.33	34.52	28.54	21.12	44.68	28.68	19.09	47.74	27.28	18.24	28.85	22.35	4.78	16.35	7.40	5.04	16.63	7.74	14.97	19.86	17.07
	Technology	24.3	35.19	28.75	19.57	45.57	27.38	18.37	41.77	25.52	17.61	25	20.66	3.36	6.60	4.45	5.26	14.21	7.68	14.87	22.03	17.76
	Universe	9.26	29.22	14.06	13.09	44.42	20.22	12.58	42.99	19.46	9.93	20.83	13.45	3.92	5.58	4.60	2.59	6.88	3.76	10.11	15.20	12.14

Table 25: Zero-shot evaluation Results of various models in Each Domain.

Domain	Unique Entity Type	Unique Relation Type
Academy	art term, mathematical constant, geographic coordinate system, mathematical object, periodization, accounting term, unix, essentially contested concept, trigonometric function, aircraft undercarriage class, wikidata property change frequency, affinity, type of regulation and control, wikimedia list of lists	academic appointment, archives at, investigated by, studied in
Business	accounting standard, convention, educational organization, street, fellowship, research program, software company, book series	transport network, legal form
Communication	form of communication	mascot
Culture	speculative fiction genre, variation	subdivision of this unit, objects of occurrence have role
Economy	-	destroyed, associated electoral district, has certification, maintains linking to
Education	-	choreographer, first performance by, category's main topic, academic calendar type, exhibition history
Energy	electricity generation, knowledge base, infrastructure	next crossing downstream, designed to carry, dam
Engineering	bulk carrier, unit of pressure, unit of speed, function, federal law enforcement agency of the united states, numeral, optimization algorithm, hyperbolic function, unary operation, real-valued function of a real variable, type of statistical model, medical device type, software design pattern, abstraction layer, computer network protocol, technical process, automobile model, material property, iec standard, failure cause, orientation, worldcon, chemical data page, inflection class, stochastic partial differential equation, weather warning, relative quality, stochastic process, research center, unit of area, standards organization, communication science term	fabrication method, carries, track gauge, flag, is metaclass for
Entertainment	theatrical genre, class of award, performing arts award	narrator, trained by, production company, location of first performance
Food and drink	advertises	national cuisine
Geography	branch of geography, year, study type, district	individual of taxon, drainage basin
Government	bicameral legislature, aviation regiment, financial measure, profession and socioprofessional category in france	-

Domain	Unique Entity Type	Unique Relation Type
Health	type of medical treatment, surgical procedure, chebi ontology term, vital statistics, goal, health profession, group or class of strains, state, medical test type, group of chemical entities, musical, act of the parliament of england, performance indicator, anatomical structure, food group	possible treatment, prohibits, clinical trial phase, vaccine for, risk factor, health specialty
History	city, tribe, human settlement, village, dynasty, isolated human group, calendar, application programming interface, persian empire, road	'has seal, badge, or sigil'
Human behavior	demon	-
Humanities	archaeological sub-discipline, stylistic device	-
Information	technical specification	vessel class
Internet	-	copyright license, funding scheme
Knowledge	school district in the united states, conceptual model, division, primary school	software engine, comorbidity, host
Language	dialect, natural writing system, latin-script alphabet, writing system, alphabet, unicode block, language group, abugida, type of language, christian prayer, question-and-answer site, phonemic transcription, insult, specification edition, unicode plane, character encoding, protein family	has grammatical mood
Law	act of the oireachtas, reformism	produced sound, mandates, voted on by, nominated by
Life	execution method	type locality (biology), habitat, foods traditionally associated, co-driver, contributing factor of, approved by
Mathematics	theorem, part of a work, mathematical problem, type of work of art, distinctive feature, type of number	proved by, solved by, statement describes
Military	aircraft model, military unit type-size class, military unit size class, ship class, armed organization, computer form factor, military division, watercraft class	position holder, film crew member, character designer, damaged
Nature	conservation status, plant life-form, meteorological phenomenon, enterprise, ecological concept, chapter, type of security, natural environment, period, gift, landscape type, geosphere, statistic, province of canada, homogamy, file format, yield, cardinal direction, greek letter, ship type, sector of rwanda, group or class of proteins, physics term, mathematical model, notion, exact solutions in general relativity, basketball position, metric function, motions of the earth, unit of acceleration, generator, copyright determination method	facilitates flow of, research intervention, greater than, located on astronomical body

Domain	Unique Entity Type	Unique Relation Type
People	human y-chromosome dna haplogroup	-
Philosophy	-	-
Politics	territory	candidate, crew member
Religion	legal doctrine, christian movement, civil liberties, religious community	does not have characteristic, observed in, cathedral, political coalition, exhibited creator
Science	school of thought, knowledge type, scientific concept	-
Society	political identity, demographic indicator, unit of analysis, communication technology, social relation	student organization of
Sports	competition class, olympic sport, ball game, symptom type, sport with racquet/stick/club, national association football supercup, chess term, acrobatic element	league level below, sports discipline competed in, competition class, tournament format, drafted by
Technology	fixed expression, artistic duo, muscle organ zone type, weapon family, plate, computer security technique, computer software term, type of infrastructure, spacecraft family, european standard	therapeutic area, investor
Universe	kind of quantity, continuum, universe, manga series, fundamental interaction, physical theory, type of quantum particle, god, medical test, relativistic wave equation, physical constant, electromagnetic wave, type of object, theonym, cosmological model, type of relation, conservation law, stellar evolution, record label, state of matter, year bc, isotope of hydrogen, hypothetical entity, natural phenomenon, cognitive process, legislation, mechanical property, approximation, physical law, exotic atom, idea, physical quantity, ucum constant, knowledge graph, deity, object genre, philosophical work, class of chemical substances by use, ordinal scale, scientific model, graph property, astronomical survey, visa, group or class of enzymes, character, nomenclatural term, hypersurface, toy model, mathematical analysis, hypothetical scientific object, philosophical terminology	measured physical quantity, shape, interaction, powered by, religious order, calculated from, is invariant under, decays to

Table 26: Unique Relation and Entity Distribution in Each Domain of MMUIE