

DrDEE: A Deconstruction-Reconstruction Framework for Document-Level Event Extraction

Kuofei Fang, Yuting Wei, Zexuan Wan, Bo Wang, Siyan Wu, Bin Wu*
Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications
{coffeyf, yuting_wei, wanzexuan, wangbo, aiyana, wubin}@bupt.edu.cn

Abstract

Document-level Event Extraction (DEE) aims to extract event records from given documents. Previous works typically process documents by constructing complex feature fusion matrices or graph structures, aiming to predict complete event records via entity selection or role filling. However, this holistic structure extraction paradigm often suffers from cascaded error propagation due to rigid entity dependency, multi-role structural omissions limited by selection mechanisms, and consequent scalability bottlenecks. In this paper, we propose a **Deconstruction-Reconstruction Framework for Document-Level Event Extraction (DrDEE)**. This framework deconstructs the holistic event record extraction task into atomic Token-Event-Role (TER) triplet verification, and subsequently reconstructs event records from valid tokens. It realizes a paradigm shift in DEE from Many-to-Many Selection to One-to-One Verification, significantly simplifying the task formulation. Experimental results show that the proposed method outperforms state-of-the-art methods, confirming its effectiveness. Our code is available at <https://github.com/CoffeyF/DrDEE>.

1 Introduction

Document-Level Event Extraction (DEE) is the task of determining event types present in a document, identifying all event records, and extracting arguments corresponding to each role (Liu et al., 2023; Ebner et al., 2020; Du et al., 2021). Existing DEE methods first employ interaction matrices or graphs to aggregate context-aware features across documents, sentences, and entities. Based on their decoding strategies, these methods can be categorized into two streams: one selects extracted named entities to fill each role, while the other identifies the corresponding role for each entity. Essentially,

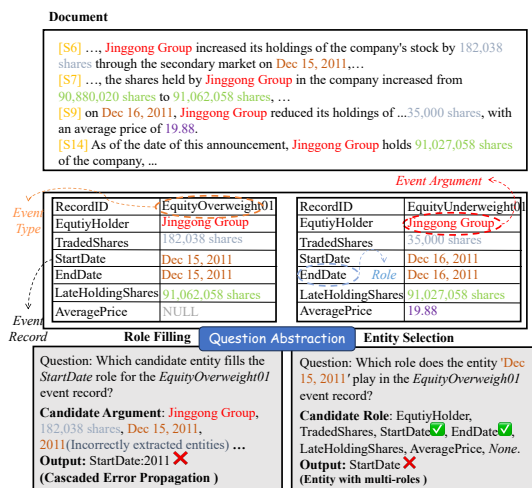


Figure 1: An example of document-level event extraction and the working pattern of previous works.

both approaches treat DEE as a many-to-many selection problem, establishing role-entity connections to directly formulate the event records in a holistic manner.

However, the structure of document-level events is highly complex, characterized by varying numbers of records and roles with different event types. Consequently, the previous extraction paradigm inevitably suffers from three intrinsic limitations. First, due to the rigid dependency on upstream named entity recognition (NER) (Nadeau and Sekine, 2007), errors in NER directly lead to cascaded error propagation. Second, constrained by the exclusivity of the selection mechanism, several approaches struggle to resolve complex scenarios where a single entity serves multiple roles within the same record (Figure 1). Finally, as the number of roles grows, the many-to-many selection mechanism leads to search space explosion and class imbalance, resulting in scalability bottlenecks that hinder performance in complex scenarios.

To tackle the numerous issues caused by holistic extraction methods, this paper proposes a

* Corresponding author.

Deconstruction-Reconstruction paradigm. This paradigm reformulates the extraction of complex event records into fine-grained Token-Event-Role (TER) triplet verification. Specifically, it determines whether a token serves as a specific role in an event record and then assembles valid tokens into complete event records.

Concretely, we first introduce a Heuristic-Driven In-Context Learning (HD-ICL) pruning strategy to avoid introducing meaningless negative samples. It utilizes Large Language Models (LLMs) to identify event types present in the document, enabling us to construct TER triplets exclusively on positive event types. Subsequently, we project records and roles of different event types into a unified semantic space to integrate their semantic information. This not only avoids the search space explosion caused by discrete combinations, but also provides clear semantic references for subsequent triplet verification. In parallel, unlike previous methods that are rigidly dependent on NER results, we adopt fine-grained tokens as basic verification units. By integrating entity information as entity constraints into token embeddings, we effectively utilize entity semantics to enhance reasoning robustness. Finally, we concatenate context-enriched token representations and record-specific role embeddings to form TER triplets. These triplets are evaluated via independent validity verification and subsequently pass through a record reconstruction mechanism to eliminate conflicting tokens and merge extracted arguments, finally reconstructing complete event structures within the document. Combining these innovations, we propose the Deconstruction-Reconstruction Framework for Document-Level Event Extraction. We term it DrDEE, homophonous with *Doctor DEE*, to symbolize its role as a domain expert in this field. This framework offers a comprehensive solution capable of effectively tackling the full spectrum of complex scenarios in document-level event extraction.

Our contributions are summarized as follows:

- We propose an innovative DEE framework, DrDEE, which effectively overcomes the limitations such as error propagation, complex role omission, and scalability bottlenecks.
- We design a novel data structure, the TER triplet, which effectively integrates the semantic information of tokens, event, and roles.
- Experiments confirm that DrDEE achieves

state-of-the-art performance on the DEE task, demonstrating its effectiveness.

2 Related Work

Early research on Event Extraction (EE) primarily focused on sentence-level tasks (Du and Cardie, 2020; Zhou et al., 2021; Xiangyu et al., 2021), limiting applicability to documents where arguments often span multiple sentences. Consequently, Document-level Event Extraction (DEE) emerged. Existing methods can be categorized into role filling and entity-driven role selection approaches.

Role filling approaches fill each role with arguments in records sequentially or in parallel. Doc2EDAG (Zheng et al., 2019) serializes records via entity-based path expansion, while GIT (Xu et al., 2021) augments this with a heterogeneous graph for global tracking. For parallel extraction, DE-PPN (Yang et al., 2021) employs a multi-decoder mechanism, which IPGPF (Huang et al., 2023) further optimizes using a role pre-filling strategy to enhance decoding efficiency.

Entity-driven role selection approaches aim to determine the specific roles of candidate entities by modeling structural dependencies using graphs or clustering networks. SCDEE (Huang and Jia, 2021) utilizes sentence communities to model distant connections between entities and events. PTPCG (Zhu et al., 2022) treats event detection as identifying maximal cliques formed by pseudo-triggers. ProCNET (Wang et al., 2023) represents events via proxy nodes and used a clustering network. LAAP (Xu et al., 2025) employs a location-aware mechanism to enhance argument role distinction. SIAT (Tao et al., 2024) incorporates spatial distribution features and adaptive thresholding to optimize interaction decoding. SEELE (Xu et al., 2024) leverages schema-aware descriptions to guide the construction of event complete graphs. DSH (Ren et al., 2025) designs a dynamic structure hypergraph to model the evolving interactions among document elements. CAINet (Pan et al., 2024) facilitated information interaction by explicitly exploiting event relations and argument correlations.

3 Method

As shown in Figure 2, there are three steps in DrDEE. First, we deconstruct the task by utilizing HD-ICL to detect event types and constructing Token-Event-Role triplets based on the pruning results. Finally, we compute the truth value score of

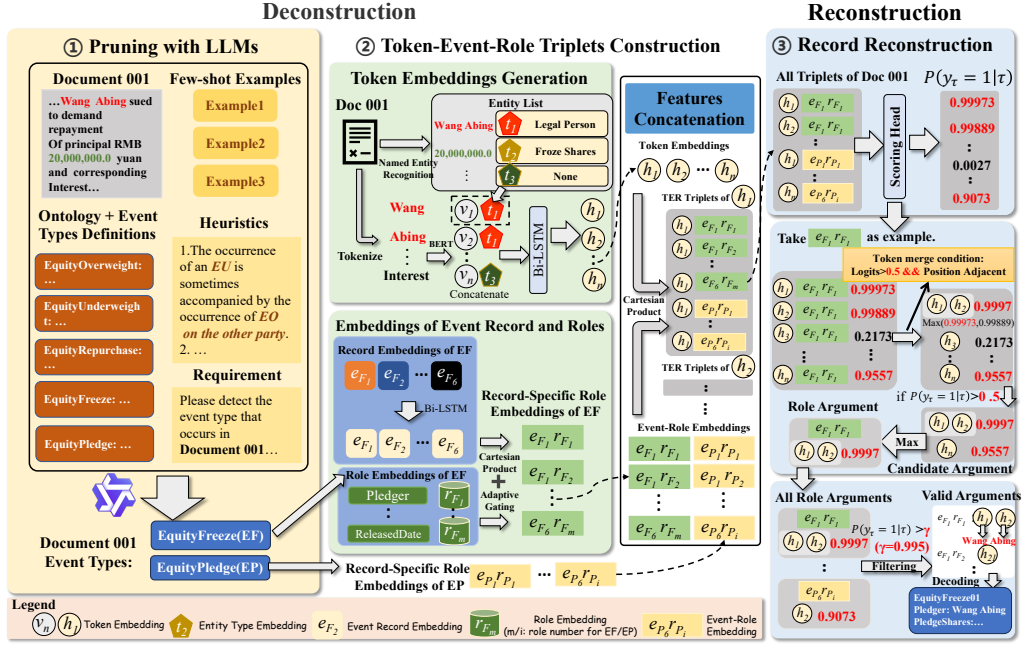


Figure 2: The overall architecture of DrDEE. Given a document, the framework first utilizes HD-ICL to identify specific event types. Triggered by these detected types, the model initiates the construction of TER triplets. Subsequently, the framework computes the truth value score of each triplet. Finally, the valid triplets are reconstructed into complete event records using a conflict resolution decoding strategy.

each triplet and employ a conflict resolution mechanism to merge tokens and filter invalid ones to reconstruct complete event records.

3.1 Pruning with LLMs

To minimize invalid negative samples and avoid decoding ambiguity arising from unnecessary false positives, we introduce a Heuristic-Driven In-Context Learning (HD-ICL) prompting scheme to detect the event types occurring in the document and prune irrelevant event types. The design rationale of HD-ICL is discussed in Appendix A.4.

3.1.1 In-context Learning Setup

In-context learning (ICL) (Liu et al., 2022; Zhou et al.) leverages the contextual understanding capabilities of LLMs to enhance inference performance without parameter updates. In HD-ICL, we construct the context by incorporating event type definitions and task examples.

Event Type Definitions. To enhance semantic comprehension of specific event types, we incorporate a set of LLM-generated event definitions, which provide domain-specific knowledge. The content of definitions is provided in Appendix A.1.

Task Example Selection. To standardize the output format and facilitate the LLM’s comprehension

of specific task patterns, we construct a task example set $\mathcal{D}_e$ from the training set. The example selection strategy is defined as follows:

$$\begin{aligned} \mathcal{D}_e = \operatorname{argmin}_{\mathcal{D}' \subseteq \mathcal{D}_{train}} |\mathcal{D}'|, \\ \text{s.t. } \forall d \in \mathcal{D}', |\mathcal{E}_d| \geq 2, \bigcup_{d' \in \mathcal{D}'} \mathcal{E}_{d'} = \mathcal{E}_{\text{ont}}. \end{aligned} \quad (1)$$

Here, $\mathcal{E}_d$ and $\mathcal{E}_{\text{ont}}$ represent the set of event types in document d and the complete event ontology, respectively. The first constraint ensures that the model captures the co-occurrence patterns of multiple event types within a single document, while the second constraint guarantees that the example set covers all event types defined in the ontology. Finally, minimizing the size of the example set ensures the conciseness of the prompt.

3.1.2 Heuristics Generation via LLMs

Recent studies (Zhou et al., 2024) have demonstrated that adding explicit heuristics to prompts can effectively address complex or ambiguous patterns. To generate heuristics for large-scale datasets, we propose an automated method utilizing the LLM as an introspective analyst, instead of time-consuming manual summarization. Specifically, we first apply a prompt containing only the

ICL context to detect event types of documents on the training dataset, and collect error cases. Then, we prompt the LLM to act as an introspective analyst to compare the erroneous predictions with the ground truth. The model analyzes the underlying reasons for these failures and derives explicit heuristics to prevent recurrence. We then manually review the generated heuristics to filter out noise and hallucinations. Details of this process are available in Appendix A.2.

3.1.3 Prompt Construction

We construct the final prompt by concatenating the ICL context, validated heuristics, and the target document as the input for the LLM. The detailed prompt template is provided in Appendix A.3.

3.2 Token-Event-Role Triplets Construction

Building upon the pruning of irrelevant event types in HD-ICL, we decompose the extraction task into the verification of atomic Token-Event-Role triplets. The construction of these triplets involves three key steps. First, we generate token embeddings. Next, we construct embeddings for event records and roles via semantic feature fusion. Finally, we concatenate each part to form the triplet.

3.2.1 Token Embeddings Generation.

We first segment the document d using NLP tools to obtain the token set $\mathbf{w} = \{w_i\}_{i=1}^n$, where n denotes the number of tokens. Next, we use BERT (Devlin et al., 2019) to encode the token sequence into a sequence of vectors $\mathbf{v} = \{v_i\}_{i=1}^n$, where $v_i \in \mathbb{R}^{d_v}$. BERT’s contextual encoding effectively captures both absolute positional information and relative semantic features.

Since named entities serve as candidate arguments for event roles, we utilize the information of named entity types as a soft constraint, providing a buffer for potential NER errors. We employ an independent NER module to identify entities and their categories within the document (Appendix A.5). Specifically, we concatenate each token’s semantic feature with the entity type of the entity it belongs to. This concatenated representation is then encoded by a Bi-LSTM network to capture bidirectional contextual dependencies. The formulation is as follows:

$$\ell_i = \begin{cases} \text{Type}(e) & \text{if } \exists e \in \mathcal{E}, w_i \in \text{Span}(e) \\ \text{None} & \text{otherwise} \end{cases}, \quad (2)$$

$$x_i = [v_i; \mathbf{W}_e(\ell_i)],$$

$$[h_1, \dots, h_n] = \text{BiLSTM}([x_1, \dots, x_n]),$$

where $\mathcal{E}$ represents the set of extracted named entities, $\mathbf{W}_e$ denotes the learnable entity type embedding matrix, and $[\cdot; \cdot]$ denotes concatenation.

3.2.2 Embeddings of Event Record and Roles

We initialize a set of learnable event record embeddings $\mathbf{e} = \{e_i\}_{i=1}^{H_E}$, where H_E denotes the number of event types, and $e_i \in \mathbb{R}^{k \times d_e}$ contains k candidate record embeddings for event type i . To capture latent dependencies between records of the same event type, we process the sequence of record embeddings for each event type through a Bi-LSTM:

$$e'_i = \text{BiLSTM}([e_{i1}, e_{i2}, \dots, e_{ik}]), \quad (3)$$

where e_{ij} represents the embedding of the j -th event record for event type i .

Simultaneously, we generate learnable role embeddings $\mathbf{r}_i \in \mathbb{R}^{H_r^i \times d_r}$ for each event type i , where H_r^i denotes the number of roles associated with the i -th event type. Distinct embeddings are assigned to roles in different event types even if they share the same name, ensuring that role representations inherently contain event-type specific semantics.

Finally, we connect each event record to its corresponding role embeddings. To construct record-specific role representations with enhanced semantic discriminability for triplet verification, we employ a gated fusion mechanism, which is formulated as:

$$\begin{aligned} \mathbf{z}_{ijm} &= [e'_{ij}; r_{im}], \\ \mathbf{g}_{ijm} &= \sigma(\mathbf{W}_g \mathbf{z}_{ijm} + \mathbf{b}_g), \\ \tilde{\mathbf{u}}_{ijm} &= \mathbf{g}_{ijm} \odot r_{im} + (1 - \mathbf{g}_{ijm}) \odot e'_{ij}, \\ \mathbf{u}_{ijm} &= \tilde{\mathbf{u}}_{ijm} + \text{MLP}(\mathbf{z}_{ijm}), \end{aligned} \quad (4)$$

where $e'_{ij} \in \mathbf{e}'_i$ and $r_{im} \in \mathbf{r}_i$, σ denotes the sigmoid activation function, and $\mathbf{u}_{ijm}$ is the final representation of the m -th role of the j -th record for the i -th event type.

3.2.3 Features Concatenation

We perform a full cartesian product connection between the sequence of refined token embeddings $\mathbf{h} = \{h_l\}_{l=1}^n$ and the set of record-specific role embeddings generated in the previous step to generate TER triplets. The final set of candidate TER triplets $\mathcal{T}$ for the document is formulated as:

$$\mathcal{T} = \left\{ [h_l; \mathbf{u}_{ijm}] \mid i \in \mathcal{E}_d, \right. \quad (5)$$

$$\left. 1 \leq j \leq k, 1 \leq m \leq H_r^i, 1 \leq l \leq n \right\},$$

where $[h_l; \mathbf{u}_{ijm}]$ represents the unified feature vector for a candidate triplet corresponding to the l -th token, j -th record, and m -th role of event type i . A concrete walkthrough example is provided in Appendix A.6.

3.3 Record Reconstruction

3.3.1 Triplet Verification

For each candidate triplet $\tau \in \mathcal{T}$, we compute the validity probability distribution $P(y_\tau|\tau)$ via an MLP layer as follows:

$$P(y_\tau|\tau) = \text{softmax}(\mathbf{W}_c \text{MLP}(\tau) + \mathbf{b}_c), \quad (6)$$

where $\mathbf{W}_c, \mathbf{b}_c$ are learnable parameters. $y_\tau = 1$ indicates that triplet is valid, and $y_\tau = 0$ otherwise.

3.3.2 Model Training

During the training phase, we do not perform event record reconstruction but focus exclusively on optimizing the triplet verification task.

Following previous studies (Wan et al., 2023a; Chen et al., 2018), given the significant class imbalance where negative samples ($y = 0$) far outnumber positive ones ($y = 1$), we employ the weighted cross-entropy loss to mitigate the class imbalance problem. The objective function is formulated as:

$$\mathcal{J} = - \sum_{\tau \in \mathcal{T}} \omega_{y_\tau} \log P(y_\tau | \tau), \quad (7)$$

where ω_{y_τ} denotes the class weight, computed as:

$$\omega_y = \begin{cases} \frac{N}{2N_1} & \text{if } y = 1 \\ \frac{N}{2N_0} & \text{if } y = 0 \end{cases}, \quad (8)$$

where N is the total number of triplets in the training set, while N_1 and N_0 represent the counts of positive and negative triplets, respectively.

3.3.3 Event Record Reconstruction

Since the extraction task was decomposed into the verification of atomic TER triplets, we need to re-assemble candidate triplets to reconstruct the complete event records. The reconstruction process consists of three stages:

Argument Merging. For each role of each event record, we first filter its corresponding triplets with validity probabilities ($P(y_\tau = 1|\tau)$) exceeding 0.5 (standard binary classification threshold). Subsequently, the filtered tokens that are adjacent in position are merged to form continuous candidate arguments, with their validity probability defined as the maximum value among its constituent tokens.

Conflict Resolution. In the DEE task, each role of a record corresponds to a unique argument. Since multiple candidate arguments may be simultaneously predicted for the same role, we resolve this conflict by selecting the candidate with the highest validity probability as the final argument for the corresponding role.

Record Filtering. Since candidate triplets are constructed for k pre-allocated event records per event type, this mechanism inevitably introduces numerous false-positive predictions. To address this, we enforce a validity constraint where only arguments with probabilities exceeding the threshold γ are considered valid, and their corresponding triplets are retained. Finally, the event records are reconstructed based on the information from these valid TER triplets. The detailed algorithm of this process will be shown in Appendix A.7.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on two widely used document-level event extraction benchmarks: ChFinAnn (Zheng et al., 2019) and DuEE-fin (Han et al., 2022). For DuEE-fin, we follow the data processing method described in Wang et al. (2023). Details are provided in Appendix A.8. Following (Zheng et al., 2019), we use micro precision, recall, and F1-score over all arguments.

4.2 Baselines

To comprehensively evaluate the performance of DrDEE, we compare it with a series of state-of-the-art methods. Detailed descriptions of these baselines are provided in Section 2. For EDEE (Wan et al., 2023b), the original implementation assumes that all entities are known and injects ground-truth entity information into the tokenizer. We removed the entity information injection and corrected its erroneous evaluation for a fair comparison.

4.3 Experiment Settings

We chose the Adam optimizer in experiments, and set batch size = 1, learning rate = $1e-3$, dropout = 0.2, and epochs = 50. The embedding dimensions of tokens/entity types/event records/roles are 768/50/128/128. Considering the superior performance of the Qwen (Bai et al., 2023) series in Chinese, we used Qwen-plus API for HD-ICL. The hidden layer size and layers of Bi-LSTM are 200 and 4. We set the maximum number of event

Models	ChFinAnn										DuEE-fin				
	P	R	EF	ER	EU	EO	EP	F1	S.	M.	P	R	F1	S.	M.
Doc2EDAG (Zheng et al., 2019)	82.7	75.2	70.2	87.3	71.8	75.0	77.3	78.8	83.9	67.3	67.1	60.1	63.4	69.1	58.7
DE-PPN (Yang et al., 2021)	83.7	76.4	73.5	87.4	74.4	75.8	78.4	79.9	85.9	68.4	69.9	33.5	45.1	54.2	21.8
GIT (Xu et al., 2021)	83.6	76.9	73.4	90.8	74.3	76.3	77.7	80.1	87.5	72.1	69.8	65.9	67.8	73.7	63.8
PTPCG (Zhu et al., 2022)	83.4	75.4	71.4	91.6	71.5	72.2	76.4	79.2	87.4	71.3	71.0	61.7	66.0	72.6	64.1
EDEE [†] (Wan et al., 2023b)	75.1	84.5	80.3	81.0	79.2	76.4	75.5	79.5	86.0	71.1	-	-	-	-	-
IPGPF (Huang et al., 2023)	85.7	77.3	76.4	93.0	76.1	74.8	80.9	81.3	91.0	70.0	68.2	61.8	64.8	-	-
ProCNET (Wang et al., 2023)	84.4	80.9	75.7	93.7	76.0	72.0	81.3	82.7	89.5	75.3	78.6	72.6	75.5	79.9	72.0
SIAT (Tao et al., 2024)	85.9	78.8	73.3	92.6	78.9	78.5	79.4	82.2	90.0	73.2	68.1	61.5	64.6	-	-
DSH (Ren et al., 2025)	84.8	79.9	-	-	-	-	-	82.3	89.2	74.6	78.2	70.6	74.2	78.4	70.5
CAINet (Pan et al., 2024)	84.3	82.9	76.0	93.6	77.2	72.9	81.7	83.6	89.8	76.5	79.1	72.9	75.8	80.3	72.2
SEELE (Xu et al., 2024)	-	-	78.8	92.0	77.7	82.4	83.1	85.1	-	-	-	-	80.8	-	-
DrDEE (Ours)	86.2	88.7	88.4	92.7	87.6	85.1	85.0	87.4	93.5	79.7	79.3	83.5	81.3	86.5	77.9

Table 1: Overall performance on ChFinAnn and DuEE-fin test datasets. **P**, **R**, and **F1** denote Micro Precision, Recall, and F1-score. **EF**, **ER**, **EU**, **EO**, and **EP** denote the Micro F1 scores for different event types in ChFinAnn, respectively. **S.** and **M.** denote F1-scores on single-record and multi-record documents. EDEE[†] indicates the result of our re-implementation under the fair setting.

records k to 6 and 2, and the validity filtering threshold γ to 0.998 and 0.995 for ChFinAnn and DuEE-fin respectively. Other settings are Python 3.12.9, PyTorch 2.6, and NVIDIA L20. We use the COARSE_ELECTRA_SMALL_ZH model of HanLP (He and Choi, 2021) as the NLP tool. Details of k and γ can be referred to Appendix A.9.

4.4 Main Result and Analysis

Overall performance. As shown in Table 1, DrDEE surpasses the previous best baseline (SEELE) by **2.3** and **0.5** F1 points on ChFinAnn and DuEE-fin, respectively. Distinct from prior methods, DrDEE consistently exhibits higher recall than precision. This trend stems from our "Decomposition-Reconstruction" paradigm: by explicitly enumerating all potential cases via TER triplets, our model effectively retrieves subtle arguments often missed by holistic graph generation, ensuring comprehensive extraction.

Generally, role filling approaches perform poorly compared to entity-driven role selection methods. While recent role selection methods have achieved progress by devising mechanisms to enhance semantic fusion, they remain constrained by the inherent limitations of the holistic extraction paradigm.

Furthermore, DuEE-fin poses a scalability challenge with significantly more event types (13 vs. 5) and roles (92 vs. 35) than ChFinAnn. While baselines suffer severe degradation on this complex dataset, DrDEE maintains high efficacy. This robustness is achieved by transforming the complex prediction space into a binary (0/1) verification task, effectively circumventing scalability bottlenecks as the complexity of the schema increases.

Single vs. Multi Records. To analyze perfor-

mance across varying complexities, we split the test sets into single-record and multi-record subsets. Critically, DrDEE significantly outperforms baselines on multi-record documents. This is primarily attributed to the exhaustive enumeration of TER triplets, which ensures comprehensive coverage of concurrent events. This mechanism allows for the independent verification of each possibility, disentangling inter-record dependencies and resolving conflicts in complex multi-event scenarios.

Performance Across Event Types. Existing methods tend to overfit simple patterns such as Equity Repurchase (ER), which is characterized by a simple role composition and is easily recognizable. As shown in Table 2, ER dominates the test set, whereas complex types like Equity Overweight and Underweight are rare. Consequently, baseline methods struggle to effectively handle complex event extraction, resulting in performance imbalance. In contrast, DrDEE maintains balanced performance across all categories. This stability stems from our Deconstruction-Reconstruction paradigm, which projects all potential combinations into a unified representation space. By constructing highly distinctive TER triplets for verification, DrDEE narrows the prediction space, ensuring robust generalization regardless of pattern complexity.

Event Type	EP	ER	EO	EU	EF	Total
Count	2,299	1,234	504	382	330	4,749

Table 2: Statistics of event records in the test set.

Model	EF	ER	EU	EO	EP	Avg
Deep Learning	98.7	99.9	96.5	93.2	99.6	97.4
LLM Solely	90.5	93.2	87.1	85.4	94.8	90.2
Heuristic	96.5	98.2	94.5	92.1	99.0	96.1
ICL	99.0	99.5	97.5	97.0	99.6	98.8
HD-ICL (Ours)	99.5	99.8	98.0	97.9	99.6	99.4

Table 3: Performance comparison of different methods on event type detection.

4.5 Additional Analysis and Discussion

4.5.1 Performance of HD-ICL

In Section 3.1, we introduce HD-ICL as a pivotal pruning mechanism for TER triplet construction. To validate its efficacy, we evaluate Precision on the ChFinAnn test set against two primary paradigms: 1) **Deep Learning methods** (e.g., Doc2EDAG), utilizing neural encoders with complex joint training; and 2) **Prompting Strategies**, including *LLM Solely* (zero-shot), *Heuristic* (heuristics only), and *Standard ICL* (HD-ICL without heuristics). As shown in Table 3, HD-ICL achieves superior performance (99.4 Avg). Compared to deep learning methods, HD-ICL delivers higher accuracy by bypassing complex joint training dependencies. Moreover, it consistently outperforms other prompting baselines, demonstrating that integrating heuristic knowledge with ICL offers a robust paradigm for complex semantic reasoning tasks. Crucially, this remarkable precision ensures that the downstream verification module receives high-fidelity candidates, significantly reducing the search space while minimizing the introduction of noise.

4.5.2 Robustness Analysis

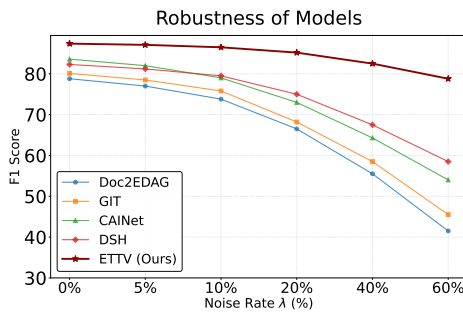


Figure 3: Performance comparison under varying entity noise rates λ on ChFinAnn.

Mitigating Error Propagation. To investigate the robustness of the proposed paradigm against error propagation, we simulate upstream NER errors by constructing noisy inputs where a ratio λ of

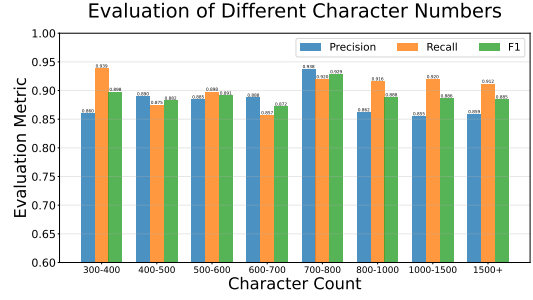


Figure 4: Performance stability of DrDEE across increasing document lengths.

extracted named entities is randomly masked or replaced. Evaluated across $\lambda \in [0\%, 60\%]$, we compare DrDEE with representative entity-dependent models. As shown in Figure 3, while these baselines suffer precipitous performance drops due to their partial or direct reliance on NER results, DrDEE exhibits resilience with minimal decay. This validates that integrating entity type constraints with token representations effectively mitigates the cascade of errors. By establishing tokens as the fundamental predictive units, our mechanism provides a buffer against NER errors, ensuring full content coverage while simultaneously leveraging entity constraints for semantic enhancement.

Stability across Document Lengths. Generally, as the number of characters in a document increases, the volume of TER triples inherently rises. To assess the stability of DrDEE under such scaling conditions, we conducted a stratified evaluation by randomly sampling ten documents for each varying length interval (ranging from 300 to 1500+ characters) from the test set. As depicted in Figure 4, despite the substantial increase in document length, DrDEE maintains consistent performance. This demonstrates the stability of our Decomposition-Aggregation paradigm in effectively handling complex, long-context DEE scenarios.

4.5.3 Resource Consumption

We evaluate the overall resource consumption of DrDEE by analyzing two key aspects: training cost and inference efficiency.

Training Cost. As detailed in Table 4, DrDEE is the lightest among the evaluated models, with only 21.36M total parameters and a mere 5.13M trainable parameters (excluding embeddings). Evaluated under the same configuration as PTPCG (Zhu et al., 2022), DrDEE completes the training in just 28.3 GPU hours. This is highly competitive with the fastest baseline. Importantly, operating with a

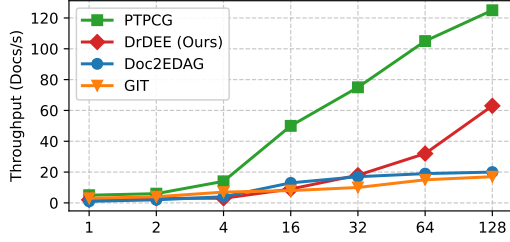


Figure 5: **Inference throughput comparison with varying batch sizes.** Experiment are recorded the same as PTPCG (Zhu et al., 2022).

batch size of 64 and $k = 6$ caps the GPU memory footprint at roughly 6GB, effectively removing the barrier of large-scale computing requirements.

Model	Params (M)	Trainable (M)	GPU Hours
GreedyDec	64	48	604.8
Doc2EDAG	64	48	604.8
GIT	97	81	633.6
PTPCG	32	16	24.0
DrDEE (Ours)	21.36	5.13	28.3

Table 4: **Training resource comparison.** Params denotes total parameters, Trainable denotes parameters excluding embeddings. GPU Hours are measured on a single NVIDIA V100 GPU.

Inference Efficiency. While the time complexity of triplet verification scales with the number of TER triplets, the verification model itself remains exceptionally lightweight. Therefore, the overall inference latency is primarily dominated by the LLM-based event type detection. As illustrated in Figure 5, with a sufficiently large batch size, DrDEE achieves a competitive throughput of 65.05 docs/s, substantially outperforming Doc2EDAG (Zheng et al., 2019) and GIT (Xu et al., 2021).

4.6 Ablation Study

Model	ChFinAnn	DuEE-fin
w/o Entity Type Constraints	79.8	71.0
w/o Pruning Mechanism	82.0	74.4
w/o Feature Fusion	84.7	78.5
DrDEE (Oracle Pruning)	90.2	85.7
DrDEE	87.4	81.3

Table 5: Ablation study of DrDEE variants.

To validate the effectiveness of each component in DrDEE, we conducted an ablation study on both datasets. The results are summarized in Table 5.

First, **w/o Entity Type Constraints** generates token representations without concatenating entity type embeddings. The performance decline

confirms that under our problem decomposition paradigm, entity type constraints serve as a vital information supplement for processing the massive volume of candidate triplets. Second, **w/o Pruning Mechanism** omits the HD-ICL based filtering, forcing the model to process all potential event types. The decline in F1 score confirms that our pruning strategy effectively filters out substantial training noise, allowing the model to focus on valid event structures. Third, **w/o Feature Fusion** substitutes the gating fusion with a simple concatenation of record and role embeddings. As shown in Table 5, this leads to a notable performance drop. This confirms that the fusion mechanism effectively aligns the semantic space, enhancing the semantic coherence of triplets. Finally, we introduce **DrDEE (Oracle Pruning)**, which assumes perfect event type detection to isolate the contributions of our Verification and Reconstruction modules. Yielding a great performance on both dataset, this result confirms that the TER triplet verification mechanism alone adeptly extracts complex arguments and resolves inter-record conflicts, underscoring its role as the core performance driver of our framework.

4.7 Case Study

A persistent bottleneck in DEE is the **Role Overlap** phenomenon, where a single entity functions as multiple roles within a unified record. As illustrated in Figure 1, traditional methods often fail when an entity like *Dec 16, 2011* acts as both *StartDate* and *EndDate* in *EquityUnderweight01*. Bound by rigid mapping constraints or exclusive decoding paths, prior baselines typically cannot assign multiple roles to a single entity simultaneously. In contrast, DrDEE naturally resolves this ambiguity. As detailed in Section 4.3, we employ the COARSE_ELECTRA_SMALL_ZH model of HanLP, which effectively segments complex financial entities (e.g., dates and companies) into individual tokens. Consequently, regarding *Dec 16, 2011* as a single token, DrDEE constructs parallel candidates: (*Dec 16, 2011*, *EquityUnderweight01*, *StartDate*) and (*Dec 16, 2011*, *EquityUnderweight01*, *EndDate*). Since each triplet is evaluated independently within the unified representation space, the model successfully identifies both as valid.

5 Conclusion

In this paper, we proposed DrDEE, a framework established on the Deconstruction-Reconstruction

paradigm. By deconstructing the DEE task into Token-Event-Role triplet verification and reconstructing records from valid tokens, our approach overcomes the fundamental bottlenecks of the holistic structure extraction paradigm, offering a robust solution for complex information extraction.

Limitations

The widely used benchmarks in this field (e.g., ChFinAnn) consist primarily of financial announcements, which typically follow a formalized and structured reporting style. While these documents present challenges in logical reasoning, they may lack the linguistic diversity (e.g., heavy metaphors, informal pragmatics, or highly irregular syntactic nesting) found in open-domain texts. Consequently, the robustness of DrDEE in substantially more linguistically complex or unstructured contexts remains to be fully verified. Future work will be dedicated to constructing a comprehensive, multi-domain DEE benchmark to rigorously stress-test our framework and facilitate broader community evaluation.

Ethics Statement

This research utilizes the ChFinAnn and DuEE-fin datasets, which are established, publicly available benchmarks derived from official financial announcements. The data contains no personally identifiable information or private content.

Regarding social impact, our work aims to enhance the efficiency of financial information processing by automating complex extraction tasks. While the proposed DrDEE framework demonstrates superior performance and stability, we advocate for a human-in-the-loop approach in real-world applications. Given the high-stakes nature of the financial domain, automated outputs should serve as an auxiliary reference for professionals to ensure maximum decision-making safety.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (62372060).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin,

Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. [Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. [Event extraction by answering \(almost\) natural questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. [Template filling with generative transformers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. [Multi-sentence argument linking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. [Ducee-fin: A large-scale dataset for document-level event extraction](#). In *Natural Language Processing and Chinese Computing*, pages 172–183, Cham. Springer International Publishing.

Han He and Jinho D. Choi. 2021. [The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang, and Weinan E. 2023. [An iteratively parallel generation method with the pre-filling strategy for document-level event extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10834–10852, Singapore. Association for Computational Linguistics.

- Yusheng Huang and Weijia Jia. 2021. [Exploring sentence community for document-level event extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 340–351, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Qi Liu, Zhen Luan, Kunlong Wang, Ye Zou, Bing Liu, and Yang Zhang. 2023. [Document-level event extraction - a survey of methods and applications](#). *Journal of Physics: Conference Series*, 2504(1):012008.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Bangze Pan, Yang Li, Suge Wang, Xiaoli Li, Deyu Li, Jian Liao, and Jianxing Zheng. 2024. [Document-level event extraction via information interaction based on event relation and argument correlation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5156–5166, Torino, Italia. ELRA and ICCL.
- Qi Ren, Weihua Wang, Jie Yu, and Guanglai Gao. 2025. [Dynamic structure hypergraph for document-level event extraction](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zekun Tao, Changjian Wang, Zhiliang Tian, Kele Xu, Yong Guo, Shanshan Li, Yanru Bai, and Da Xie. 2024. [Siat: Document-level event extraction via spatiality-augmented interaction model with adaptive thresholding](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(10).
- Qizhi Wan, Changxuan Wan, Keli Xiao, Rong Hu, and Dexi Liu. 2023a. [A multi-channel hierarchical graph attention network for open event extraction](#). *ACM Trans. Inf. Syst.*, 41(1).
- Qizhi Wan, Changxuan Wan, Keli Xiao, Dexi Liu, Chenliang Li, Bolong Zheng, Xiping Liu, and Rong Hu. 2023b. [Joint document-level event extraction via token-token bidirectional event completed graph](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10481–10492, Toronto, Canada. Association for Computational Linguistics.
- Xinyu Wang, Lin Gui, and Yulan He. 2023. [Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10118–10133, Toronto, Canada. Association for Computational Linguistics.
- Xi Xiangyu, Wei Ye, Shikun Zhang, Quanxiu Wang, Huixing Jiang, and Wei Wu. 2021. [Capturing event argument interaction via a bi-directional entity-level recurrent decoder](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 210–219, Online. Association for Computational Linguistics.
- Jinghan Xu, Cheng Yang, and Xiaojun Kang. 2025. [Laap: Learning the argument of an entity with event prompts for document-level event extraction](#). *Neurocomputing*, 613:128584.
- Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. [Document-level event extraction via heterogeneous graph-based interaction model with a tracker](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3533–3546, Online. Association for Computational Linguistics.
- Zijie Xu, Peng Wang, Wenjun Ke, Guozheng Li, Jiajun Liu, Ke Ji, Xiye Chen, and Chenxiao Wu. 2024. [Incorporating schema-aware description into document-level event extraction](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*.
- Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. [Document-level event extraction via parallel prediction networks](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6298–6308, Online. Association for Computational Linguistics.
- Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. [Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.
- Hanzhang Zhou, Junlang Qian, Zijian Feng, Lu Hui, Zixiao Zhu, and Kezhi Mao. 2024. [LLMs learn task](#)

heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11972–11990, Bangkok, Thailand. Association for Computational Linguistics.

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. [What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14638–14646.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Min Zhang. 2022. [Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph](#). *Preprint*, arXiv:2112.06013.

A Appendix

A.1 Event Type Definitions

Figure 6 presents the event type definitions for ChFinAnn. Notably, these definitions are generated by the same LLM employed in our HD-ICL inference. This strategy ensures strict semantic consistency, guaranteeing that the model’s understanding of event schema aligns perfectly with its reasoning logic during the pruning phase. The prompt details for the DuEE-fin dataset will be released on our github repository upon acceptance.

A.2 Heuristics

Figure 7 presents the heuristic rules generated for the ChFinAnn dataset.

A.3 Prompt Construction

We construct the example set $\mathcal{D}_e$ by randomly selecting candidate documents that satisfy the structural constraints defined in Eq. 1. The resulting example set used for context is shown in Figure 13.

Subsequently, we assemble the Event Type Definitions, the constructed Example Set, the generated Heuristics, and the target document into a unified input sequence. The complete structure of the HD-ICL prompt is illustrated in Figure 8.

A.4 Design Rationale of HD-ICL Pruning

Prior DEE methods typically embed event detection within jointly trained pipelines, where detection accuracy is tightly coupled with upstream document encoding through shared loss optimization. This coupling introduces two limitations: (1) errors in downstream argument extraction can back-propagate and degrade detection quality, and (2)

Dataset	Original	After Pruning	Reduction
ChFinAnn	728.5M (0.73B)	152.3M	79.09%

Table 6: Effect of HD-ICL pruning on the TER triplet space.

the detection component behaves as a black box whose reasoning process is neither interpretable nor independently controllable.

DrDEE instead decouples event detection from the extraction pipeline. By leveraging LLMs as an independent semantic reasoner equipped with event definitions, task examples, and automatically generated heuristic rules, HD-ICL achieves 99.4% average precision (Table 3) without relying on upstream feature engineering or downstream loss signals.

We characterize this step as **Pruning** based on its functional role within our architecture. In DrDEE, the full Cartesian product of tokens, event records, and roles produces a combinatorially large candidate space. HD-ICL operates as an explicit hard constraint on this space: by filtering out irrelevant event types *before* triplet construction, it physically eliminates the vast majority of invalid candidates. As shown in Table 6, on ChFinAnn the total triplet space is reduced from 728.5M to 152.3M, a 79.09% reduction. This is functionally analogous to branch pruning in search trees—it does not provide a soft signal for downstream modules, but structurally removes entire invalid branches from the verification space. This hard-filtering property positions HD-ICL as a structural pruning mechanism within the Deconstruction phase, enabling efficient and focused triplet verification.

A.5 Named Entity Recognition

Following the standard sequence labeling paradigm, we treat the input document d as a collection of sentences $D = \{s_j\}_{j=1}^{N_s}$. For each sentence $s_j = (w_1, w_2, \dots, w_m)$, we utilize BERT as the backbone encoder to capture contextual semantic features. The hidden representation is obtained as follows:

$$\mathbf{H} = \text{BERT}([w_1, w_2, \dots, w_m]), \quad (9)$$

where $\mathbf{H} \in \mathbb{R}^{m \times d_h}$ denotes the sequence of hidden states. To explicitly model the dependencies between adjacent labels, we employ a Conditional Random Field (CRF) layer on top of the encoder. We adopt the BIO (Begin, Inside, Other) tagging

Event Type	Definitions
<i>EquityFreeze</i>	<i>Equity Freeze event refers to the judicial or authorized agencies legally freezing shares held by company shareholders, restricting their transfer or disposal. This typically occurs when shareholders are involved in legal disputes or debt defaults, serving as a legal measure to protect creditors' interests.</i>
<i>EquityUnderweight</i>	<i>Equity Underweight event describes shareholders decreasing their holdings in a company. This typically indicates cautious outlook on the company's future development or reflects capital needs prompting the divestment.</i>
<i>EquityOverweight</i>	<i>Equity Overweight event involves shareholders increasing their ownership stake. This generally signals confidence in corporate prospects or strategic investment objectives driving the accumulation.</i>
<i>EquityPledge</i>	<i>Equity Pledge event occurs when shareholders mortgage their holdings as collateral to financial institutions or creditors for financing. This reflects capital-raising needs while potentially impacting corporate ownership structure stability.</i>
<i>EquityRepurchase</i>	<i>Equity Repurchase event entails companies using retained earnings to buy back issued shares from the market. This typically signals perceived market undervaluation or serves purposes like equity incentives and market capitalization management.</i>

Figure 6: The event type definitions of ChfinAnn.

Pay attention to **title information**, which usually specifies the main event types contained in financial announcements; in addition to title information, focus on **uncovering hidden financial event information in the document content**, such as equity changes, counterparties, and other key elements.

Note the implicit **co-occurring event** information in documents. For example, equity transfers often involve both overweight and underweight events. It must be emphasized: When a shareholder reduces holdings, the counterparty receiving these shares inevitably triggers an overweight event. For instance:

- If Shareholder A reduces holdings to B and C, both B and C trigger overweight events;
- Even if the document title only mentions reduction, analyze whether recipients trigger overweight events;
- Each counterparty receiving shares should be labeled as an overweight event.

Attention must be paid to relationships between subject entities and counterparties mentioned in announcements, as these may imply multiple event types.

Extracted financial event types should be explicitly recorded historical events in announcements, excluding **unoccurred or planned events**. For example, in a repurchase report, if only plans are mentioned without actual repurchase, it is not considered a repurchase event.

It must be clarified: There is no "**pledge release**" event type. If an announcement title mentions pledge release, verify whether historical **equity freeze** events are described in the content. Strictly distinguish:

- Equity pledge: Subject entity is financial institutions (pledge registration/pledgee)
- Equity freeze: Subject entity is judicial authorities (rolling freeze/executing court)

Figure 7: Automatically generated heuristic rules for the ChFinAnn dataset.

schema to assign labels to tokens. The model is trained by minimizing the negative log-likelihood of the ground truth tag sequence y^* :

$$\mathcal{L}_{NER} = - \sum_{s \in D} \log P(y_s^* | s), \quad (10)$$

where $P(y_s^* | s)$ represents the probability of the true label sequence given the input sentence, calculated via the CRF layer.

A.6 Walkthrough Example

We illustrate the complete extraction process using the example document in Figure 9, which contains an *EquityPledge* event.

TER Triplet Construction. Given the document, HD-ICL first identifies *EquityPledge* as the active event type. The framework then initializes $k = 6$ candidate record embeddings $(e_1, \dots, e_6)$ for this

You are a professional financial event analyst. Please carefully analyze the following text and identify the types of equity-related events it contains.

All types of financial events include: **Ontology of Event Types**

The event types detection should be strictly limited to the event types included in the above ontology. The following are the professional definitions of each financial event. Please refer to the definitions to detect event types: **Event Types Definitions**

Please output the results strictly in the following format:
["Event Type 1", "Event Type 2", ...]

Task Examples are below: **Example Set**

Examples analysis: (Evidence for the above examples' event type detection)

Event type detection should follow the following heuristics: **Heuristics Set**

Note:

1. Only output the event type list; do not include any other content.
2. If no events are detected, output an empty list [].
3. Event types must exactly match the event type names defined above
4. Do not add any explanations or additional notes.
5. If the same event type appears multiple times, only output it once.
6. Carefully analyze the document content and do not omit any implied event information.

Document content:

Figure 8: The structure of HD-ICL prompt.

type and retrieves its associated role set (e.g., *Pledger*, *TradedShares*, *StartDate*, etc.). Meanwhile, the document is tokenized and encoded via BERT with entity type constraints to produce token embeddings $\{h_l\}_{l=1}^n$.

The triplet set is formed by connecting each token embedding with every record-specific role embedding. For instance, the token "182,038" yields triplets such as $(h_{182,038}, e_1, r_{Pledger})$, $(h_{182,038}, e_1, r_{TradedShares})$, through to $(h_{182,038}, e_6, r_{Pledger})$, and similarly for all other tokens. Each triplet is then scored by the verification module to obtain $P(y_\tau = 1 | \tau)$.

Record Reconstruction. Consider the role *TradedShares* in record slot e_1 . The verification module assigns high validity scores to adjacent tokens "182,038" (0.999) and "shares" (0.998), both exceeding the binary threshold of 0.5. Since these tokens are positionally adjacent, the MERGEADJACENT operation consolidates them into a single candidate argument span "182,038 shares", whose score is defined as the maximum among its constituents (0.999).

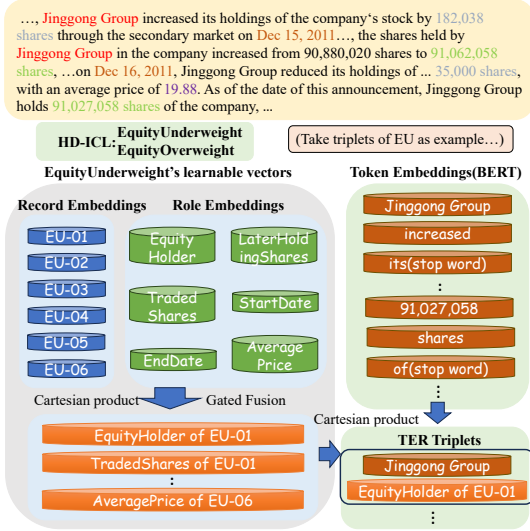


Figure 9: Illustration of the walkthrough example

If multiple candidate spans compete for the same role—for example, “182,038 shares” (0.999) and a spurious token “2011” (0.6)—the conflict resolution step selects the candidate with the highest score as the final argument.

This merging and resolution procedure is applied iteratively across all roles and all k record slots. Since many of the $k = 6$ pre-allocated slots correspond to non-existent records, a validity threshold γ (e.g., 0.995) is enforced to suppress false positives. Only arguments whose scores exceed γ are retained: “182,038 shares” with score 0.999 passes the filter and is assigned to *TradedShares* of e_1 , while a spurious prediction “Jinggong Group” in slot e_6 with score 0.45 is discarded. The surviving arguments across all roles and slots are assembled into the final structured event records.

A.7 Record Reconstruction

To provide a rigorous formulation of the inference logic, Algorithm 1 details the step-by-step procedure for reassembling atomic TER triplets into complete event records. The process operates on the probability matrix P (specifically focusing on the validity probability where $y = 1$) and iterates through each pre-allocated record slot j (up to k) for every event type i . The reconstruction unfolds in three critical phases. First, during **Argument Merging** (Lines 6–8), the algorithm isolates valid tokens where the predicted probability exceeds the standard binary threshold (0.5). Adjacent valid tokens are then consolidated via the `MergeAdjacent( $\cdot$ )` function to form a set of continuous candidate argument spans $\mathcal{S}$. Subsequently,

Algorithm 1 Event Record Reconstruction

Require: Probabilities P (for class $y = 1$), Threshold γ , Max Records k

Ensure: Final Event Records $\mathcal{R}_{final}$

```

1:  $\mathcal{R}_{final} \leftarrow \emptyset$ 
2: for each event type  $i \in \mathcal{E}$  do
3:   for  $j \leftarrow 1$  to  $k$  do
4:      $R_{ij} \leftarrow \emptyset$ 
5:     for each role  $m \in \mathcal{A}_i$  do
6:       // 1. Argument Merging
7:        $L_{valid} \leftarrow \{l \mid P(y_{ijml}) > 0.5\}$ 
8:       //  $l$  denotes the token in the TER triplet
9:        $\mathcal{S} \leftarrow \text{MergeAdjacent}(L_{valid})$ 
10:      // 2. Conflict Resolution
11:      if  $\mathcal{S} \neq \emptyset$  then
12:         $Score(s) \leftarrow \max_{l \in \mathcal{S}} P(y_{ijml}), \forall s \in \mathcal{S}$ 
13:         $s^* \leftarrow \arg \max_{s \in \mathcal{S}} Score(s)$ 
14:        // 3. Validity Filtering
15:        if  $Score(s^*) > \gamma$  then
16:           $R_{ij}[m] \leftarrow s^*$ 
17:        end if
18:      end if
19:    end for
20:    if  $R_{ij} \neq \emptyset$  then
21:       $\mathcal{R}_{final} \leftarrow \mathcal{R}_{final} \cup \{R_{ij}\}$ 
22:    end if
23:  end for
24: end for
25: return  $\mathcal{R}_{final}$ 

```

the **Conflict Resolution** phase (Lines 11–13) addresses the constraint that a specific role must map to a unique argument. Here, the algorithm evaluates all candidate spans in $\mathcal{S}$ and identifies the optimal candidate s^* by maximizing the validity score, defined as the peak probability among its constituent tokens. Finally, via **Validity Filtering** (Lines 15–17), the algorithm mitigates the redundancy introduced by pre-allocating k records. A filtering threshold γ is employed to prune false positives, ensuring that only the candidate s^* whose score exceeds γ is retained in the final record R_{ij} .

A.8 Dataset Descriptions

We evaluate our framework on two widely used benchmarks in the field of Document-level Event Extraction (DEE).

ChFinAnn. Constructed via remote supervision, ChFinAnn (Zheng et al., 2019) is a large-scale dataset focusing on financial equity announcements. It defines a schema containing 5 event types and 35 argument roles. The dataset comprises 32,040 documents, divided into 25,632 for training, 3,204 for development, and 3,204 for testing. A significant characteristic of this corpus is that 29% of the documents contain multiple event records. To provide a comprehensive view of the data characteristics, we present the *overall* distri-

bution of event types across the entire dataset in Table 7.

Event Type	Distribution (Overall)
Equity Freeze	4.2%
Equity Repurchase	9.5%
Equity Underweight	16.0%
Equity Overweight	18.3%
Equity Pledge	52.0%

Table 7: The overall distribution of event types in the ChFinAnn dataset. Note that this distribution reflects the entire corpus statistics, which differs slightly from the test-set-specific distribution discussed in Table 2.

DuEE-fin. Released by Baidu, DuEE-fin (Han et al., 2022) is a manually annotated dataset sourced from Chinese financial news and announcements. It features a more complex schema with 13 event types and 92 argument roles. Since the ground truth for the official test set is not publicly available, we follow the preprocessing setting established by Wang et al. (2023). Specifically, we utilize a split of 6,515, 500, and 1,171 documents for the training, development, and test sets, respectively. Approximately 33% of the documents involve multiple events.

Critically, the DuEE-fin dataset lacks explicit entity type annotations. Consequently, previous studies typically utilize argument roles directly as pseudo-entity labels to train the NER module. Although this substitution introduces a potential semantic granularity mismatch since specific roles are not strictly equivalent to ontological entity categories, we adopt this setting to ensure a fair and consistent comparison with baseline methods. However, we suggest that future research should incorporate precise entity annotations to further enhance extraction accuracy.

A.9 Hyperparameter Analysis

We investigate the impact of two critical hyperparameters on the ChFinAnn dataset: the maximum event records per document (k) and the validity filtering threshold (γ). Both hyperparameters are selected exclusively on the development set prior to test evaluation. The following analysis serves as a post-hoc sensitivity study to illustrate their impact on model behavior.

Impact of Maximum Records k . The selection of k represents a trade-off between theoretical coverage and inference precision. We determine the optimal k through a two-step process on the train-

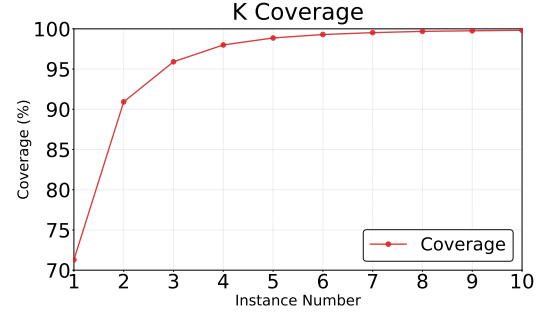


Figure 10: The cumulative percentage of documents covered by setting different maximum record counts k . Setting $k = 6$ covers 99.3% of instances.

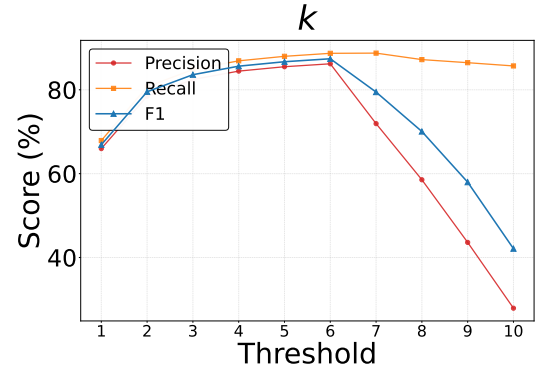


Figure 11: The impact of k on extraction performance. Precision drops significantly when $k > 6$ due to introduced noise.

ing and development sets. First, we analyze the intrinsic data distribution. As shown in Figure 10, the cumulative coverage curve rises sharply and plateaus around $k = 6$, which effectively covers 99.3% of the document instances in the training set. This provides strong statistical support for setting $k = 6$. Second, we validate this choice on the dev set, and select $k = 6$. Figure 11 further illustrates the Precision, Recall, and F1 trends on the test set as k varies. While Recall improves with higher k , Precision suffers a steep decline beyond $k = 6$, as forcing the model to predict surplus records in documents that typically contain fewer events inevitably introduces sparse, low-confidence predictions. This consistent pattern confirms the robustness of our selection.

Impact of Validity Threshold γ . The threshold γ acts as a gatekeeper to suppress false-positive noise generated by the pre-allocated k slots. Since we intentionally set $k = 6$ to maximize recall, many slots in single-event documents are redundant and must be filtered. Figure 12 demonstrates the model’s sensitivity to γ . Lower thresholds

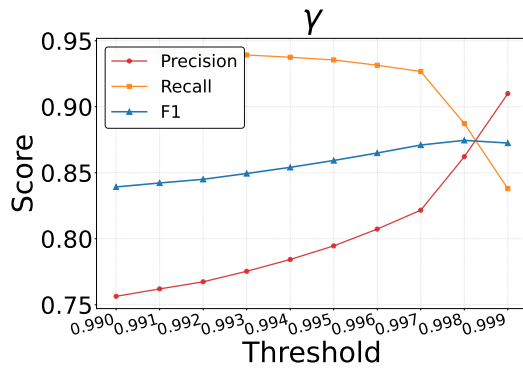


Figure 12: The impact of the filtering threshold γ on performance metrics. A high threshold is required to filter false positives.

($\gamma < 0.99$) allow too much noise, maintaining Recall but severely damaging Precision. By progressively increasing γ , we effectively prune these ambiguous predictions. The results show that a strict threshold (around $\gamma = 0.998$) achieves the best balance, filtering invalid records while retaining high-confidence extractions.

Generalizability of Hyperparameters. Notably, we adopt the identical setting of $k = 6$ across both ChFinAnn and DuEE-fin without dataset-specific tuning. Since k is determined by data distribution rather than model capacity, this setting generalizes naturally to documents with multiple event types (e.g., 5 types with $k = 6$ yields up to 30 candidate records). For γ , it serves as a standard decision boundary applied exclusively during inference and can be easily calibrated on the development set. As shown in Figure 12, performance remains stable across a wide range ($\gamma > 0.99$), indicating low sensitivity and minimal tuning effort.

Example 1 Securities Code: 002278 Securities Abbreviation: Shenkai Shares Announcement No.: 2016-116 Announcement Regarding Shareholder Equity Rolling Freeze The Company and all members of the Board warrant the authenticity, accuracy, and completeness of information disclosure, with no false records, misleading statements, or material omissions. Shanghai Shenkai Petroleum & Chemical Equipment Co., Ltd. (hereinafter "the Company") received notice on the evening of November 15, 2016, from shareholder Shanghai Yexiang Investment Management Co., Ltd. (hereinafter "Yexiang Investment") regarding judicial rolling freeze of its shares. Details are as follows: Yexiang Investment recently received a summons from Binjiang District People's Court of Hangzhou, Zhejiang Province, indicating that Mr. Wang Abing has filed a lawsuit against Yexiang Investment. As Yexiang Investment failed to repay loans upon maturity, Wang Abing sued to demand repayment of principal RMB 20,000,000.0 yuan and corresponding interest. On November 14, 2016, Yexiang Investment received a Litigation Preservation Notice from Binjiang District People's Court, imposing rolling freeze on 47,577,481 Shenkai shares held by Yexiang Investment and their accrued interest during the freezing period, effective from October 31, 2016, to October 30, 2019. Notice No.: (2016) Zhe 0108 Zhi Bao 575. Yexiang Investment is currently the largest shareholder, holding 47,577,481 shares (13.07% of total equity). As of this announcement, Yexiang Investment's share status is as follows: (1) On February 2, 2016, pledged 42,000,000 shares (11.54% of total equity) to Haitong Securities Co., Ltd., with repurchase date February 1, 2017 (88.27% of its holdings; Announcement No. 2016-004). (2) On September 14, 2016, 47,577,481 shares were judicially frozen by Changning Branch of Shanghai Public Security Bureau for 2 years (Announcement No. 2016-083). (3) On October 31, 2016, 47,577,481 shares were judicially rolling frozen by Binjiang District People's Court for 3 years. The Company will closely monitor subsequent developments and urge shareholders to fulfill information disclosure obligations promptly. Investors are advised to note investment risks. Hereby Announced. Shanghai Shenkai Petroleum & Chemical Equipment Co., Ltd. Board of Directors November 16, 2016 EventTypes: "EquityFreeze", "EquityPledge"	
Example 2 Securities Code: 601933 Securities Abbreviation: Yonghui Superstores Announcement No.: Lin 2018-6 Announcement on Commencement and Results of Share Repurchase under 2017 Restricted Stock Incentive Plan The Company and all members of the Board warrant that the announcement contains no false records, misleading statements, or material omissions, and assume individual and joint liability for its authenticity, accuracy, and completeness. Important Notice: • Repurchased shares: 166,343,400 • Repurchase price: ¥9.77/share The Company's Third 27th Board Meeting (November 30, 2017) and 2017 Second Interim Shareholders' Meeting (December 18, 2017) approved the "Proposal on Share Repurchase". I. Share Repurchase Quantity and Amount On January 22, 2018, the Company's dedicated securities account acquired 166,343,400 shares from Mr. Zhang Xuansong, representing 1.74% of total share capital, with actual expenditure of ¥1,625,695,074.01. According to the Proposal, the planned repurchase quantity was no more than 166,671,300 shares. The actual repurchase of 166,343,400 shares fell short by 327,900 shares due to the departure of one incentive recipient (core business backbone). II. Share Repurchase Period The repurchase commenced and concluded on January 22, 2018. III. Transaction Price The repurchase price was ¥9.77 per share. IV. Status of Repurchased Shares The repurchased shares are held in the Company's dedicated securities account. These shares will be used for the 2017 Restricted Stock Incentive Plan. During holding in the dedicated account, the shares carry no voting rights and do not participate in profit distribution. As of January 22, 2018, the balance of the dedicated account was ¥33,304,925.99. The Company will transfer it back to the bank account, with remaining funds continuing as bank deposits. V. Compliance Statement The share repurchase strictly complies with the Company Law and relevant regulations of the incentive plan. Hereby Announced. Board of Directors of Yonghui Superstores Co., Ltd. January 24, 2018 EventTypes: "EquityRepurchase", "EquityUnderweight"	
Example 3 Securities Code: 002178 Securities Abbreviation: Yanhua Intelligent Announcement No.: 2009-001 The Company, its directors, supervisors, and senior management warrant the authenticity, accuracy, and completeness of this announcement and assume liability for any false records, misleading statements, or material omissions. Ms. Hu Meizhen, a shareholder of Shanghai Yanhua Intelligent Technology Co., Ltd. (hereinafter "the Company"), originally held 10,740,000 unrestricted tradable shares, representing 13.425% of the Company's total shares. On January 8, 2009, the Company received notice from Hu Meizhen that as of market close on January 7, 2009, she reduced 4,000,000 unrestricted tradable shares (5% of total shares) through the block trading system of Shenzhen Stock Exchange. Among these, 1,000,000 shares were transferred to Shanghai Jiuge Investment Management Co., Ltd. and 3,000,000 shares to individual Zhang Chunji via the block trading system. Shanghai Jiuge Investment Management Co., Ltd. and Zhang Chunji confirmed no affiliation or concerted action. After reduction, Hu Meizhen holds 6,740,000 unrestricted tradable shares (8.425% of total shares). Having resigned as Director and Vice Chairman on June 18, 2008, she committed to transferring no more than 50% of her holdings (5,370,000 shares) within twelve months after six months from resignation. For details on equity changes, please refer to the "Simplified Equity Change Report" disclosed on cninfo.com.cn (http://www.cninfo.com.cn). The Company announces this pursuant to relevant regulations. Hereby Announced. Shanghai Yanhua Intelligent Technology Co., Ltd. Board of Directors January 9, 2009 EventTypes: "EquityOverweight", "EquityUnderweight"	

Figure 13: The example set of ICL in ChFinAnn.