PHE^2:A Chinese Document-level Dataset for Public Health Event Extraction

Qian Chen[1,2,*], Jiaju Ren[1], Xin Guo[1], Suge Wang[1,2]
[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
Email: chenqian@sxu.edu.cn, renjiaju@sxu.edu.cn, guoxinjsj@sxu.edu.cn, wsg@sxu.edu.cn

***Abstract*—Given the critical role of public health in national security and social stability, effective surveillance of emerging threats is essential, which necessitates Document-level Event Extraction (DEE) to derive structured event knowledge from public health news for timely risk assessment and decision-making. However, research on DEE in public health remains limited, primarily due to the prohibitive cost of manual annotation that hampers the construction of large-scale, high-quality datasets. To bridge this gap, we propose a modular and extensible LLM-assisted annotation framework for Event Extraction, with which we construct a Chinese document-level dataset for Public Health Event Extraction(PHE^2), a large-scale Chinese dataset comprising 10,028 documents, over 21,000 events, and 90,000 arguments. Our proposed framework achieves F1 scores of 80.68% for argument extraction on a human-annotated test set, verifying its effectiveness. We benchmark several DEE models on PHE^2, and results show that existing models achieve substantially lower performance on PHE^2, highlighting its unique challenges and strong potential to advancein research DEE. PHE^2 is publicly available at https://gitee.com/mike29/PHE2.**

***Index Terms*—Public Health, Document-level Event Extraction, Dataset Construction, LLM**

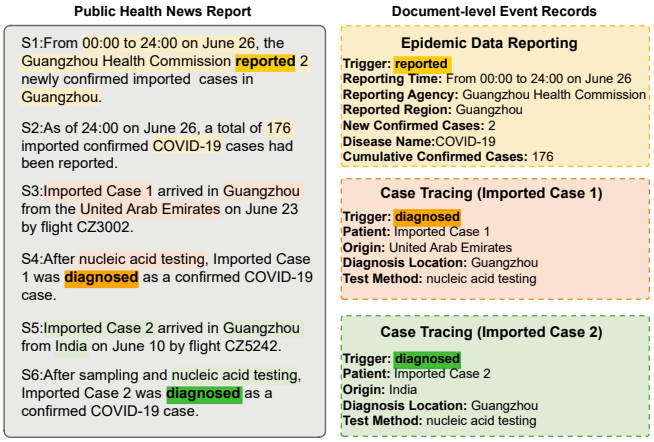

Fig. 1. An example from PHE^2.

I. Introduction

Public health is not only related to individual health but also directly affects national security and social stability. Public health news reports serve as a primary source for monitoring emerging public health threats, providing abundant yet largely unstructured information on epidemics, environmental hazards, and health incidents. Consequently, converting such reports into structured event knowledge is essential for timely risk evaluation and policy decision-making [1] [2]. DEE identifies event triggers and extracts their corresponding arguments across entire documents [3], making it particularly suitable for public health news reports that are lengthy and exhibit information scattered across sentences.

However, the lack of dedicated datasets and the high cost of manual annotation constitute the primary factors limiting research on DEE in public health domain. Existing studies predominantly focus on general news [4] [5], finance [6], [7], and military domains [8], with comparatively limited attention paid to public health domain. An illustrative sample from PHE^2 is shown in Figure 1 in which the left panel presents a simplified document containing three events, and the right panel shows the corresponding document-level event records, distinguished by different colors. Specifically, for the Epidemic Data Reporting event, the trigger *reported* appears in sentence S1, with the arguments *new confirmed cases* and *cumulative confirmed cases* distributed across sentences S1 and S2, respectively. This example highlights that document-level annotation in public health domain must handle multiple interrelated events within a single document and identify event arguments scattered across multiple sentences, requiring long-context reasoning and domain expertise and thus making large-scale manual annotation costly. Although recent studies have explored distant supervision [9] and LLM-based collaborative annotation [10], there remains a need for a low-cost and effective methodology to construct document-level datasets for public health domain.

In this paper, we propose a modular and extensible LLM-assisted annotation framework. To balance annotation quality and scalability in complex document-level scenarios, the framework adopts a decomposition strategy that separates the annotation process into four modules: (1) event schema

Qian Chen is the corresponding author. This work is supported by the National Natural Science Foundation of China (NO.U24A20335, NO.6237073346)

definition, (2) data collection and prompt formulation, (3) multi-stage LLM annotation, and (4) human-in-the-loop quality control. In module 1, the event schema is defined based on Winslow's public health definition. In module 2, public health news reports are crawled and cleaned. Representative reports are selected from the cleaned corpus and manually annotated to construct event type-specific few-shot examples for prompt. In module 3, the annotation process is carried out in two stages: first, event triggers are identified and assigned to their corresponding event types at the document level; second, schema-constrained argument extraction is performed based on these event types. In module 4, human evaluation and refinement are conducted to ensure data quality through comparison with expert annotations and systematic corrections. Using this framework, we construct PHE^2, a large-scale public health event extraction dataset.

Our contributions can be summarized as follows:

- We introduce a modular and extensible LLM-assisted annotation framework for event extraction that integrates event schema definition, data collection and prompt formulation, multi-stage LLM annotation, and human-in-the-loop quality control for efficient construction of high-quality event datasets.
- We construct PHE^2, a large-scale Chinese DEE dataset tailored for public health domain. It comprises 10,028 documents, 21,005 event instances, and 90,693 arguments, covering 11 event types and 57 argument roles.
- We conduct systematic investigation on PHE^2 and identify four salient characteristics, including multi-event co-occurrence, long-span arguments, argument scattering, and generative arguments. Benchmark experiments indicate that existing DEE models exhibit limited performance on this dataset. These results position PHE^2 as a challenging benchmark for future research.

II. Related Work

Event Extraction Datasets. Early datasets like ACE 2005 [11] focused on sentence-level extraction, which fails to capture the cross-sentence dependencies inherent in complex documents. The shift to DEE led to datasets such as RAMS [4] and WikiEvents [5]. However, RAMS imposes a 5-sentence window constraint around triggers, and WikiEvents, while densely annotated, is limited in scale (246 documents). In specialized domains, ChFinAnn [6] introduces a large-scale financial corpus that adopts a trigger-free design, effectively reformulating the extraction task as direct table-filling. Similarly, in the military domain, CMNEE [8] addresses data scarcity by offering a large-scale, manually annotated dataset of 17,000 documents that captures both event triggers and argument roles. PHEE [12] is a dataset for pharmacovigilance event extraction from clinical texts. However, despite these advances, a large-scale DEE dataset specifically designed for public health remains unavailable.

III. Construction of PHE^2

To construct a large-scale, high-quality DEE dataset for public health domain, we design a modular and extensible LLM-assisted annotation framework that integrates event schema definition, data collection and prompt formulation, multi-stage LLM annotation, and human-in-the-loop quality control. The overall framework is illustrated in Figure 2.

A. Event Schema Definition

To support systematic annotation in public health, we define an event schema tailored for DEE. The schema is grounded in Winslow's classical definition of public health, which characterizes public health as organized societal efforts to prevent disease, prolong life, and promote health [13].

Layer 1: **Public Health Domain.** This layer defines the overall scope of the dataset, ensuring that all annotated events fall within public health.

Layer 2: **Key Topic.** Guided by Winslow's definition and common reporting practices in public health news, we organize the domain into five topics: Infectious Diseases, Environmental Health, Food Safety, Drug Safety, and Health Promotion.

Layer 3: **Event Type.** Within each topic, we define fine-grained event types that serve as the classification targets for event extraction. In total, the schema includes 11 event types corresponding to canonical public health events.

Layer 4: **Event Instance.** Event types may be instantiated by multiple real-world event instances, each representing a specific public health event. An event instance may be mentioned across multiple documents and evolve over time as new information is reported.

Layer 5: **Event Mention.** Event mentions denote the textual realizations of event instances in documents. Multiple mentions may refer to the same event instance and can be distributed across sentences. This layer bridges real-world event instances and their linguistic expressions, which is essential for DEE with cross-sentence argument aggregation.

For each event type, we define a set of argument roles capturing its essential participants and attributes. The event schema is iteratively refined through team discussion during annotation to resolve inconsistencies and ensure coverage. The final schema consists of 11 event types and 57 argument roles, summarized in Table I.

B. Data Collection and Prompt Formulation

Our dataset consists of news articles on public health, crawled from major Chinese news portals (*People's Daily Online*, *Xinhua News Agency*, *The Paper News*, and *Sohu*) with publication dates spanning 2020 to 2024. The collected documents were processed through a standard preprocessing pipeline, including deduplication, removal of irrelevant content, and filtering of incomplete or corrupted documents. After cleaning, we obtained a curated corpus of 10,028 public health news documents.

To facilitate downstream annotation, the cleaned corpus are pre-grouped into 11 coarse event-type subsets using keyword matching. This step is solely used to reduce annotation cost

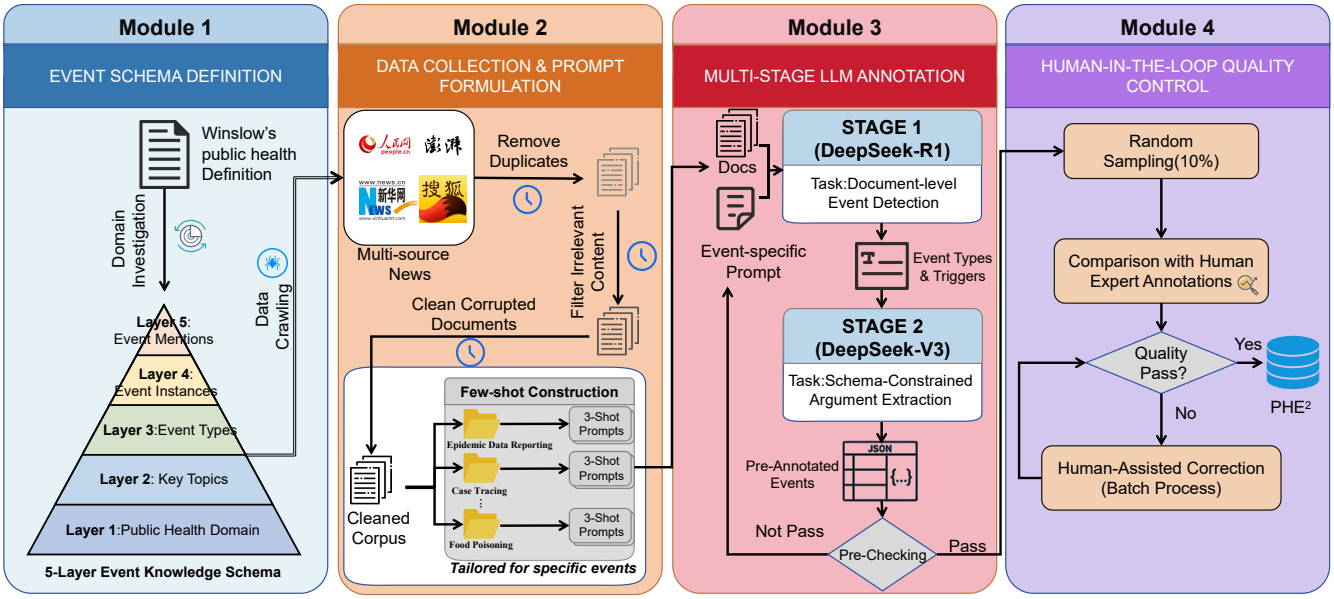

Fig. 2. A Modular and Extensible LLM-assisted Annotation Framework.

and does not preclude documents from containing multiple event types, which are still determined by the LLM during annotation. Then, we manually annotated a small set of representative documents from corpus for both Event Detection(ED) and Event Argument Extraction(EAE). These annotated examples were used to construct event type-specific few-shot prompts, which serve as demonstrations for the annotation module described in the following section.

C. Multi-Stage LLM annotation

We design a multi-stage LLM annotation module. This module decomposes the DEE task into two stages, namely document-level ED and schema-constrained argument extraction, enabling structured annotation in multi-event documents.

1) Stage 1: Document-level Event Detection: The goal is to identify all events in a document and locate their corresponding triggers. The corpus are organized into 11 event type-specific subsets. During annotation, these subsets are processed sequentially, with each subset annotated using a prompt specifically designed for its corresponding event type.

We employ *DeepSeek-R1-671B* [14] to perform ED. An event type-specific prompt includes all event definition, workflow, and manually annotated few-shot examples. Given a document, the model identifies all event types and extracts their corresponding triggers, which are used as the output of this stage. These outputs are normalized through format standardization and are used as input to Stage 2.

2) Stage 2: Schema-Constrained Argument Extraction: In Stage 2, we perform schema-constrained argument extraction for the detected events from Stage 1. We adopt *DeepSeek-V3-671B* [15] for this stage, as it provides a favorable balance between efficiency and fine-grained extraction capability required for large-scale annotation.

For each document, *DeepSeek-V3-671B* takes as input the document text, event types and triggers identified in Stage 1, as well as event schemas and few-shot examples. It then extracts all arguments associated with the event instance from the document. This process enables accurate aggregation of event arguments at the document level.

The extracted results are organized as event records, each consisting of an event type, its trigger, and the corresponding arguments. These raw event records are then post-processed and normalized into a structured JSON format for downstream use.

3) Pre-Check procedure: After the completion of annotation stages, we apply a lightweight pre-check procedure to ensure the validity of the annotations. The pre-check covers empty outputs, violations of the predefined event schema, such as undefined event types or undefined argument roles, failures in JSON parsing, and cases where the predicted event trigger cannot be found in the original document text. Annotations exhibiting any of these issues are sent back for re-annotation. If the output remains empty after re-annotation, manual inspection is conducted. Samples are excluded from the final dataset only when human annotators confirm that no event types defined in the schema are present in the document.

D. Human-In-The-Loop Quality Control

To assess the quality of the dataset, we follow the evaluation protocol proposed in [10], adopting a comparison-based approach against a gold-standard annotation constructed from a small manually labeled subset. Specifically, we select 1,003 documents from the test set, accounting for 10% of the full

TABLE I
EVENT SCHEMA OF PHE^2

Event Type	Event Definition	Argument Roles
Project Bid Winning	Enterprises obtain the qualification to undertake environmental protection engineering projects.	Winning Time, Winning Company, Winning Project, Bid Amount
Water Pollution	Pollution sources lead to the deterioration of water environmental quality.	Time, Pollution Source, Involved Party, Polluted Area
Epidemic Data Reporting	Official agencies release newly reported epidemic-related statistical information.	Reporting Time, Disease Name, Publishing Agency, Reported Region, New Confirmed Cases, New Suspected Cases, New Deaths, Cumulative Confirmed Cases
Case Tracing	Investigation of the origin and infection pathways of confirmed cases.	Patient, Origin, Confirmation Time, Confirmation Location, Symptoms, Testing Method
Drug Launch	Newly developed drugs pass regulatory approval and officially enter the market for sale.	Research Institution, Drug Name, Indication, Trade Name, Approval Agency, Approval Time
Food Sampling	Regulatory authorities conduct food quality sampling inspections and publish the results.	Sampling Time, Sampling Agency, Total Batches, Qualified Batches, Unqualified Batches, Sampling Object
Food Poisoning	Group health symptoms caused by the ingestion of contaminated food.	Time, Location, Poisoned Count, Poison Source, Symptoms, Death Count
Health Campaign	Organized activities aimed at promoting public health awareness.	Event Time, Organizer, Theme, Related Health Day
Hero Commemoration	Honoring individuals who made significant contributions in important events.	Name, Occupation, Affiliation, Deeds
Donation	Social organizations or individuals provide public welfare assistance to those in need.	Donation Time, Donor, Recipient, Donated Content, Donation Purpose
Vaccine Development	Institutions conduct scientific research and experimental work on vaccines.	Time, Institution Name, Research Phase, Vaccine Project Name

dataset, and ask 6 graduate students in the field of NLP to annotate this subset. The aggregated annotations are used to construct a human-annotated gold standard, which serves as a reference for evaluating the accuracy and reliability of the proposed LLM-assisted annotation framework.

In the assessment procedure, We evaluate the annotation results by comparing them with the human-annotated gold standard using the F1-score as the metric. The proposed framework achieves F1-scores of 81.52% for ED and 80.68% for EAE, indicating that the LLM-generated annotations are of high quality and can effectively assist human annotation in large-scale dataset construction.

Consequently, In the correction procedure, We further extracted and audited 30% of the entire dataset using non-replacement sampling. This large-scale auditing process allowed us to identify and rectify systematic errors, significantly elevating the overall fidelity of the PHE^2 dataset for downstream model training.

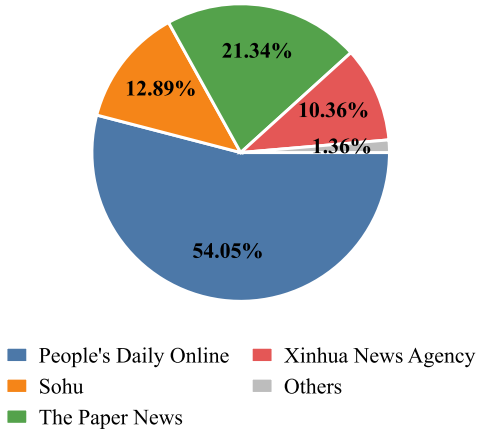

Fig. 3. Data Sources.

IV. DATA INVESTIGATION OF PHE^2

In this section, we investigate the statistical properties and task challenges of PHE^2. The raw data was crawled from multiple mainstream news portals, and the distribution of data sources is shown in Figure 3. Overall, PHE^2 includes 11 event types and 57 argument roles. It contains 10,028 public health news documents, totaling 104,651 sentences. PHE^2, with an average length of 512.72 Chinese characters and 10.44 sentences for each document, is annotated with 21,005 events and 90,693 event arguments. On average, each document contains 2.09 events and 9.04 event arguments, reflecting a high information density.

We compare PHE^2 with representative event extraction datasets in Table II, including ACE2005 and several widely used document-l evel benchmarks such as WikiEvents, RAMS, ChFinAnn, DuEE-Fin, DocEE, and CMNEE. PHE^2 exhibits higher event and argument density than most existing datasets, leading to more complex document-level event structures.

A. Event Types Distribution

The distribution of 11 event types is shown in Figure 4. Epidemic Data Reporting, Donation, and Case Tracing collectively account for 82% of all event instances, whereas the remaining 8 event types constitute only 18%, reflecting the long-tailed distribution of the PHE^2. Such a long-tailed distribution increases the difficulty of learning reliable representations for infrequent event types.

TABLE II
COMPARISON OF PHE^2 WITH EXISTING EVENT EXTRACTION DATASETS

Dataset	#isDocEvent	#EvTyp	#ArgTyp	#Doc	#Event	#Arg	#Ev/Doc	#Arg/Doc
ACE2005	×	33	35	599	-	9,590	-	16
WikiEvents	✓	50	59	246	3,951	5,536	16	22.5
RAMS	✓	139	65	9,124	9,124	21,237	1	2.3
ChFinAnn	✓	5	35	32,040	47,824	289,871	1.5	9.1
DuEE-Fin	✓	13	92	11,699	15,850	81,632	1.4	6.9
DocEE	✓	59	356	27,485	27,485	180,528	1	6.6
CMNEE	✓	8	11	17,000	29,223	93,348	1.7	5.5
PHE^2(ours)	✓	11	57	10,028	21,005	90,693	2.09	9.04

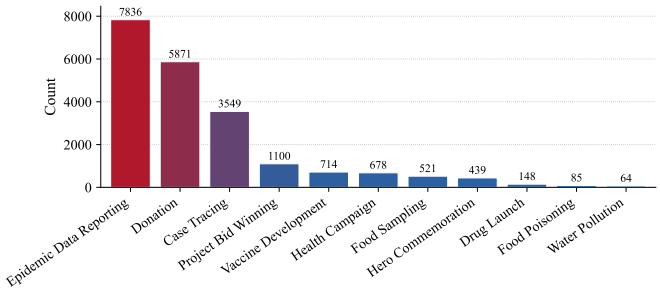

Fig. 4. Event Type Distribution.

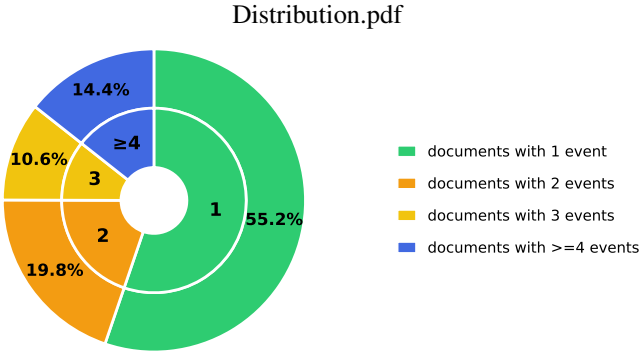

Fig. 5. Multi-events Distribution.

Multi-event co-occurrence. Figure 5 illustrates the proportion of documents containing different numbers of events. Nearly 45% of the documents in PHE^2 contain two or more events. This proportion is significantly higher than in datasets like ChFinAnn at 29% and DuEE-Fin at 29.2%. Among multi-event documents, 41.29% contain events of different types. Handling multiple events in a single document requires models to effectively identify and differentiate between events, even when they are contextually linked.

Overlapping Events. PHE^2 also contains a considerable number of overlapping events. Overlapping events refer to cases where the same word serve as triggers for different events, the same argument plays different roles in multiple events, or the same argument takes on multiple roles within a single event [16]. In PHE^2, 13.8% of documents exhibit trigger overlap, accounting for 30.8% among multi-event documents. This requires models to not only identify multiple events but also resolve the relationships between them when they share triggers or arguments.

B. Event Arguments Analysis

PHE^2 presents several challenges for event argument extraction, including long-span arguments, generative arguments, and argument scattering.

Long-span Arguments. PHE^2 features a significant number of long-span arguments, with over 20% of arguments spanning 10 characters or more and the longest reaching 243 characters. These long-span arguments often consist of entire clauses or descriptive phrases, which increases the difficulty of accurately identifying their syntactic boundaries.

Generative Arguments. A unique feature of our LLM-assisted annotation framework is its ability to capture arguments with generative properties. As a result, 1.33% of the arguments in PHE^2 are generative, where the argument content is constructed by aggregating and summarizing information from multiple locations in the document instead of being extracted from a single text span.

For example, the document contains sentences such as "During multiple epidemic response periods, the individual remained at the frontline of prevention and control work," and "The individual's contributions to public health were recognized by relevant authorities. For a Hero Commemoration event, arguments such as *Occupation* and *name* can be extracted from individual sentences, whereas the *Deeds* require generative modeling. By integrating information across sentences, Deeds can be summarized as "sustained frontline epidemic response efforts and significant contributions to public health." This requirement pushes the task beyond traditional extraction and into the realm of abstractive reasoning, a capability that most current event extraction systems lack.

Argument Scattering. The problem of argument scattering is prominent. We measure this with ArgScat, defined as the number of sentences in which event arguments of the same event are scattered [17]. PHE^2 has an average ArgScat value of 4.79. This means that to construct a complete event record, a model must identify and connect pieces of information scattered across nearly five different sentences on average. This directly challenges a model's ability to handle long-range dependencies and aggregate information from across the entire document.

TABLE III
STATISTICS OF SPLITTING PHE^2

Subset	Docs	Events	Args
Train	8,023	16,834	71,960
Dev	1,002	2,221	10,008
Test	1,003	1,950	8,725
Total	**10,028**	**21,005**	**90,693**

V. EXPERIMENTS ON PHE^2

A. Experimental setting

Benchmark setting Similar to ChFinAnn, we split PHE^2 into training, validation, and test sets following an 8:1:1 ratio. The distribution of event types is largely consistent across the three subsets. Table III presents the statistics of the data split.

Models We select eight representative DEE models that cover a wide range of technical paradigms.

- **Doc2EDAG** [6]: An end-to-end model that generates an entity-based directed acyclic graph to explicitly handle argument scattering and multi-event challenges.
- **GreedyDec** [6]: A variant of Doc2EDAG that employs a simplified greedy decoding strategy. Instead of generating a full graph, it fills the event table by greedily finding the first matching entity for each role.
- **GIT** [18]: A state-of-the-art model and modification of Doc2EDAG that uses a heterogeneous graph interaction network and maintains a global tracker during decoding.
- **DE-PPN** [19]: A parallel prediction network that uses a multi-granularity non-autoregressive decoder to generate multiple event records in parallel, trained with a matching loss function.
- **PTPCG** [20]: An efficient non-autoregressive DEE model that uses pseudo-trigger-aware pruned complete graphs to rapidly combine arguments.
- **ProCNet** [21]: A model utilizing event proxy nodes to capture global dependencies and Hausdorff distance minimization to directly optimize the alignment between predicted and gold event sets.
- **EPAL** [22]: A method using event-specific probes and argument libraries to extract events, employing probe-label alignment to distinguish fine-grained differences among homogeneous events.
- **Text2Event** [23]: An end-to-end sequence-to-structure model based on T5. It linearizes the structured event record into a target text sequence and frames DEE as a generation task.

Evaluation We adopted the same evaluation settings as previous studies [6] [18]. we report the micro Precision, Recall, and F1-score for ED and EAE for Text2Event, the remaining models are designed to perform EAE only and do not explicitly perform ED; therefore, their scores are not reported and marked as "–" in Table IV.

B. Overall Experimental Results

Overall experimental results can be seen in Table IV. From these results, we find that: 1) Almost all models exhibit relatively low performance on PHE^2, where the best F1 score achieved by EPAL is only 66.21%. In contrast, EPAL attains F1 scores of 83.4% on ChFinAnn and 76.4% on DuEE-Fin, indicating that PHE^2 is substantially more challenging than existing DEE benchmarks. 2) Most models exhibit a pronounced precision–recall imbalance, with substantially higher precision than recall. This consistently low recall can be largely attributed to the presence of generative arguments and argument scattering, which make exhaustive argument extraction particularly challenging. 3) It is worth noting that EPAL performs the best on PHE^2. Its Recall of 63.45% significantly surpasses graph-based baselines. This indicates that EPAL's event-specific probe mechanism is conducive to recovering scattered arguments and distinguishing complex events.

TABLE IV
OVERALL EXPERIMENTAL RESULTS ON THE PHE^2 TEST SET

Model	Argument Extraction			Event Detection		
	P	R	F1	P	R	F1
Doc2EDAG	69.33	44.47	54.19	-	-	-
GreedyDec	65.64	38.08	48.20	-	-	-
DE-PPN	56.86	47.20	51.58	-	-	-
GIT	67.45	47.08	55.45	-	-	-
PTPCG	**72.81**	40.03	51.66	-	-	-
ProCNet	70.74	60.21	65.05	-	-	-
EPAL	69.23	**63.45**	**66.21**	-	-	-
Text2Event	68.08	55.09	60.90	58.45	44.16	50.31

C. Single and Multi-Event Results

To investigate the model in complex scenarios, we compare performance on single-event versus multi-event documents in Table V. The results reveal a universal performance degradation when processing documents with multiple co-occurring events. However, the magnitude of this decline varies significantly across model paradigms.

EPAL demonstrates the strongest resilience, achieving the highest multi-event F1 score of 60.8%, followed by ProCNet. This significantly outperforms traditional graph-based architectures like GIT and Doc2EDAG. This suggests that its mechanism of jointly learning event-specific probes and argument libraries in EPAL effectively disentangles overlapping arguments and distinguishes between multiple events better than global graph interactions alone. In contrast, PTPCG and GreedyDec suffer the most severe performance drops. Specifically, PTPCG's performance decreases from 69.2% (Single) to 38.9% (Multi). This indicates that while greedy decoding or aggressive graph pruning strategies are efficient for simple texts, they fail to retain sufficient context for resolving complex, multi-event structures.

In multi-event settings, as shown in Table V, models achieve reasonable performance on high-frequency event types such as Epidemic Data Reporting, whereas zero F1-scores are observed for rare event types such as Water Pollution and Drug Launch. These near-zero scores are primarily attributable to extreme data sparsity and the limited presence of these

TABLE V
MODEL PERFORMANCE (ROLE-F1) ON SINGLE-EVENT (S) VS. MULTI-EVENT (M) DOCUMENTS

Model	Split	Event Type (F1 %)											Micro F1
		Bid	**Camp.**	**Pollu.**	**Samp.**	**Pois.**	**Donat.**	**Hero**	**Launch**	**Rep.**	**Trace**	**Vacc.**	
Doc2EDAG	Single	67.5	36.5	25.0	67.0	46.7	37.6	39.0	81.0	76.3	40.5	32.3	65.6
	Multi	43.2	13.9	0.0	39.3	9.5	31.7	6.0	0.0	57.2	49.4	31.6	47.0
GreedyDec	Single	67.1	41.0	10.5	66.2	57.1	39.9	30.8	73.9	74.2	37.8	39.2	64.5
	Multi	37.5	15.0	0.0	50.0	10.5	21.4	9.1	0.0	52.0	27.3	16.8	36.7
DE-PPN	Single	60.3	32.7	43.5	65.8	42.1	40.0	22.9	66.7	72.6	38.3	41.2	62.5
	Multi	47.7	18.0	0.0	53.6	27.3	27.4	9.9	0.0	57.0	44.1	21.3	44.9
GIT	Single	67.3	38.5	47.1	68.6	51.9	43.4	16.2	65.3	74.6	32.8	37.0	65.6
	Multi	48.9	20.0	0.0	57.7	10.5	35.8	9.5	0.0	58.9	50.3	30.2	49.1
PTPCG	Single	71.3	56.6	47.6	77.5	43.8	48.6	**55.0**	86.3	76.5	**43.8**	42.9	69.2
	Multi	45.9	26.0	0.0	54.5	45.5	25.7	11.6	0.0	52.3	30.2	18.1	38.9
ProCNet	Single	**82.8**	**67.0**	**69.2**	**81.7**	**75.7**	**59.7**	36.4	88.5	83.6	35.0	**44.0**	**76.4**
	Multi	55.2	**40.4**	0.0	**59.6**	**50.0**	45.5	28.0	0.0	68.7	56.9	**39.6**	57.9
EPAL	Single	80.3	59.4	63.6	76.0	58.1	58.1	29.5	**90.2**	**83.8**	27.3	41.2	75.2
	Multi	**65.3**	31.1	0.0	56.6	28.6	**48.7**	**29.3**	**12.5**	**71.4**	**61.2**	33.1	**60.8**

Note: Event abbreviations: **Bid**: Project Bid Winning; **Camp.**: Health Campaign; **Pollu.**: Water Pollution; **Samp.**: Food Sampling; **Pois.**: Food Poisoning; **Donat.**: Donation; **Hero**: Hero Commemoration; **Launch**: Drug Launch; **Rep.**: Epidemic Data Reporting; **Trace**: Case Tracing; **Vacc.**: Vaccine Development.

event types in multi-event documents, rather than to systematic deficiencies in the models' reasoning capabilities.

D. Error Analysis

To gain a deeper understanding of the challenges posed by the PHE^2 dataset, we conducted a detailed error analysis. This analysis was based on 50 randomly sampled instances, for which we examined the prediction results of the Text2Event model to identify and categorize errors. After comprehensive analysis, we conclude six types of errors: 1) **Event Type Error**: The model incorrectly predicts the event type. 2) **Non-Existent Role Error**: The document does not contain an argument corresponding to a particular event role, but the model incorrectly predicts such an argument. 3) **Missing Role Error**: The document contains an argument for a specific event role, but the model fails to predict it. 4) **Argument Boundary Error**: Both the predicted event type and role are correct, while the predicted argument only partially matches the gold standard. 5) **Event Argument Error**: Unlike Argument Boundary Error, the predicted event argument is completely incorrect. 6) **Event Number Error**: The predicted number of events is either greater or fewer than the actual number of events in one document.

Event Type Error and Event Argument Error are the two most frequent error types, accounting for 30.0% and 28.7% of all errors. These errors are categorized as semantic errors, as they stem from incorrect interpretation of event semantics. Such errors commonly occur in documents containing multiple events, where contextual overlap makes it difficult for models to accurately distinguish event boundaries and assign arguments to the correct events.

Non-Existent Role Error and Missing Role Error account for 14.7% and 6.7% of the total errors. These errors are characterized by a mismatch between predicted arguments and the explicit textual cues present in the document.

Event Number Error, which constitutes 11.3% of all errors, occurs when the model predicts too many or too few event instances in a document, indicating difficulties in consistently identifying and separating multiple event instances in long documents.

Argument Boundary Error accounts for the remaining 8.7% of cases. It is characterized by imprecise identification of argument spans. Although less frequent, it highlights the difficulty of accurately determining the boundaries of long-span or structurally complex arguments.

VI. CONCLUSION

In this paper, we introduced PHE^2, a novel large-scale dataset for DEE in the public health domain, characterized by multi-event co-occurrence, long-span arguments, argument scattering, and generative arguments. We introduce a modular and extensible LLM-assisted annotation framework that significantly reduces the cost and effort of manual annotation while maintaining high reliability. Our investigation and experimental results indicate that PHE^2 serves as a challenging benchmark. We expect PHE^2 to advance the development of DEE methodologies and provide support for practical public health information extraction applications.

Despite its scale, PHE^2 exhibits a pronounced long-tailed distribution over event types and covers a relatively limited scope of predefined event types, which may not fully capture the diversity and evolving nature of public health events in real-world scenarios. In addition, while the proposed LLM-assisted annotation framework substantially reduces manual annotation cost, ensuring annotation quality still requires human-in-the-loop sampling-based verification and correction, which inevitably incurs additional manual effort.

REFERENCES

[1] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection—harnessing the web for public health surveillance," *The New England journal of medicine*, vol. 360, no. 21, p. 2153, 2009.

[2] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi, "Biocaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 10 2008. [Online]. Available: https://doi.org/10.1093/bioinformatics/btn534

[3] H. Yang, Y. Chen, K. Liu, Y. Xiao, and J. Zhao, "DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data," in *Proceedings of ACL 2018, System Demonstrations*, F. Liu and T. Solorio, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 50–55. [Online]. Available: https://aclanthology.org/P18-4009/

[4] S. Ebner, P. Xia, R. Culkin, K. Rawlins, and B. Van Durme, "Multi-sentence argument linking," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8057–8077. [Online]. Available: https://aclanthology.org/2020.acl-main.718/

[5] S. Li, H. Ji, and J. Han, "Document-level event argument extraction by conditional generation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 894–908. [Online]. Available: https://aclanthology.org/2021.naacl-main.69/

[6] S. Zheng, W. Cao, W. Xu, and J. Bian, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 337–346. [Online]. Available: https://aclanthology.org/D19-1032/

[7] C. Han, J. Zhang, X. Li, G. Xu, W. Peng, and Z. Zeng, "Duee-fin: A large-scale dataset for document-level event extraction," in *Natural Language Processing and Chinese Computing*, W. Lu, S. Huang, Y. Hong, and X. Zhou, Eds. Cham: Springer International Publishing, 2022, pp. 172–183.

[8] M. Zhu, Z. Xu, K. Zeng, K. Xiao, M. Wang, W. Ke, and H. Huang, "CMNEE:a large-scale document-level event extraction dataset based on open-source Chinese military news," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 3367–3379. [Online]. Available: https://aclanthology.org/2024.lrec-main.299/

[9] S. Li, Q. Zhan, K. Conger, M. Palmer, H. Ji, and J. Han, "GLEN: General-purpose event detection for thousands of types," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2823–2838. [Online]. Available: https://aclanthology.org/2023.emnlp-main.170/

[10] W. Liu, Z. Li, L. Bai, Y. Zuo, D. Xu, X. Jin, J. Guo, and X. Cheng, "Towards event extraction with massive types: LLM-based collaborative annotation and partitioning extraction," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 34 377–34 399. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1743/

[11] C. Walker, S. Strassel, J. Medero, and K. Maeda, "Ace 2005 multilingual training corpus," *(No Title)*, 2006.

[12] Z. Sun, J. Li, G. Pergola, B. Wallace, B. John, N. Greene, J. Kim, and Y. He, "PHEE: A dataset for pharmacovigilance event extraction from text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5571–5587. [Online]. Available: https://aclanthology.org/2022.emnlp-main.376/

[13] C.-E. A. Winslow, "The untilled fields of public health," *Science*, vol. 51, no. 1306, pp. 23–33, 1920. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.51.1306.23

[14] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[15] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[16] J. Sheng, S. Guo, B. Yu, Q. Li, Y. Hei, L. Wang, T. Liu, and H. Xu, "CasEE: A joint learning framework with cascade decoding for overlapping event extraction," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 164–174. [Online]. Available: https://aclanthology.org/2021.findings-acl.14/

[17] M. Tong, B. Xu, S. Wang, M. Han, Y. Cao, J. Zhu, S. Chen, L. Hou, and J. Li, "DocEE: A large-scale and fine-grained benchmark for document-level event extraction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3970–3982. [Online]. Available: https://aclanthology.org/2022.naacl-main.291/

[18] R. Xu, T. Liu, L. Li, and B. Chang, "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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 3533–3546. [Online]. Available: https://aclanthology.org/2021.acl-long.274/

[19] H. Yang, D. Sui, Y. Chen, K. Liu, J. Zhao, and T. Wang, "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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 6298–6308. [Online]. Available: https://aclanthology.org/2021.acl-long.492/

[20] T. Zhu, X. Qu, W. Chen, Z. Wang, B. Huai, N. Yuan, and M. Zhang, "Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 4552–4558, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/632

[21] X. Wang, L. Gui, and Y. He, "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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 118–10 133. [Online]. Available: https://aclanthology.org/2023.acl-long.563/

[22] J. Hu, C. Xue, C. Yu, J. Xu, and C. Tan, "Joint learning event-specific probe and argument library with differential optimization for document-level multi-event extraction," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 714–726. [Online]. Available: https://aclanthology.org/2025.findings-naacl.42/

[23] Y. Lu, H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen, "Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2795–2806. [Online]. Available: https://aclanthology.org/2021.acl-long.217/