Adaptive Schema-aware Event Extraction with Retrieval-Augmented Generation

Sheng Liang[1], Hang Lv[2], Zhihao Wen[1], Yaxiong Wu[1], Yongyue Zhang[1], Hao Wang[2], Yong Liu[1]*

[1]Huawei Technologies Co., Ltd., [2]University of Science and Technology of China

liangsheng16@huawei.com, lvhang1001@mail.ustc.edu.cn

wanghao3@ustc.edu.cn, liu.yong6@huawei.com

Abstract

Event extraction (EE) is a fundamental task in natural language processing (NLP) that involves identifying and extracting event information from unstructured text. Effective EE in real-world scenarios requires two key steps: selecting appropriate schemas from hundreds of candidates and executing the extraction process. Existing research exhibits two critical gaps: (1) the rigid schema fixation in existing pipeline systems, and (2) the absence of benchmarks for evaluating joint schema matching and extraction. While large language models (LLMs) show promise, challenges such as schema hallucination and the inefficient use of long contexts for schema selection hinder their practical deployment. In response, we propose **A**daptive **S**chema-aware **E**vent **E**xtraction (**ASEE**), a retrieval-augmented generation paradigm. ASEE first enriches schemas via **schema paraphrasing** to improve retrieval accuracy and then performs targeted, schema-guided extraction. To facilitate rigorous evaluation, we construct the **M**ulti-**D**imensional **S**chema-aware **E**vent **E**xtraction (**MD-SEE**) benchmark, which systematically consolidates 12 datasets across diverse domains, complexity levels, and language settings. Extensive evaluations on MD-SEE show that our proposed ASEE demonstrates strong adaptability across various scenarios, significantly improving the performance of joint schema retrieval and event extraction. Our codes and datasets are available at https://github.com/USTC-StarTeam/ASEE.git

1 Introduction

Event extraction (EE) is an essential task in information extraction (IE) (Xu et al., 2023; Lu et al., 2022) that aims to identify event triggers (i.e., Event Detection (ED)) and their associated arguments (i.e., Event Arguments Extraction (EAE)) from unstructured text, thereby transforming raw text into structured event representations (Xu et al., 2023). Event extraction plays an important role in various natural language processing applications, such as knowledge graph construction, question answering, information retrieval, and event prediction, by providing structured representations of real-world occurrences (Lai, 2022).

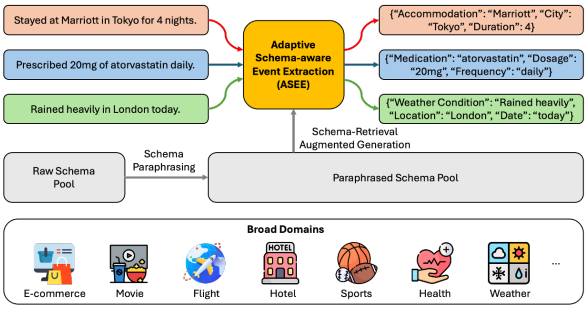

Figure 1: An example of Adaptive Schema-aware Event Extraction (ASEE) in broad domains.

While existing EE studies are effective, they typically operate under the assumption of a small, predefined set of event schemas. However, real-world EE deployments, especially in broad-domain applications, present a more complex challenge: practitioners must first select the appropriate schema from hundreds or even thousands of domain-specific candidates before executing the extraction. This dynamic, two-stage paradigm is critical in interactive scenarios. For instance, in building an on-device personalized assistant (Wu et al., 2025a; Lv et al., 2025), the system needs to construct a user's profile by extracting diverse information (e.g., interests, appointments, contacts) from a long-term conversational history. In such cases, a conventional approach with a rigid, fixed schema is incompatible with the dynamic nature of user interactions. This operational gap between academic research and industrial needs exposes two fundamental limitations: **one in system design and the other in evaluation benchmarks**.

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7927–7946
November 4-9, 2025

First, in terms of system design, existing approaches lack adaptability. Traditional pipeline systems are inherently rigid: they are either hard-coded for a fixed schema, which fails in cross-domain scenarios, or handle multiple schemas via naive concatenation, leading to conflicts and error propagation. While Large Language Models (LLMs) seem to resolve this with their generalization abilities (Huang et al., 2025a), their practical application reveals a new dilemma. A common LLM-based strategy involves loading all candidate schemas into the context window. Even as context windows expand, this "brute-force" method is suboptimal, incurring significant computational costs and suffering from performance decay due to effects like "lost in the middle" (Liu et al., 2023), where models struggle to utilize information within long inputs(Wu et al., 2025b; Nie et al., 2023). Furthermore, LLMs can exhibit schema hallucination, inventing event types not specified in the provided schemas (Huang et al., 2025a). Retrieval-Augmented Generation (RAG) (Gao et al., 2023; Gu et al., 2025) presents a more targeted alternative. However, its effectiveness is often hampered because current implementations rely on oversimplified schema definitions that lack the semantic depth required for precise retrieval and reliable schema-guided extraction.

Second, the absence of joint evaluation benchmarks creates an academic-industrial disconnect. While recent studies explore LLM-based extraction (Shiri et al., 2024), existing datasets (Lai, 2022) presuppose perfect schema selection—an assumption invalid in real-world scenarios where schema retrieval constitutes a mission-critical prerequisite. This discrepancy leaves the combined capability of schema retrieval and event extraction unevaluated.

To address the first limitation, we propose **A**daptive **S**chema-aware **E**vent **E**xtraction (**ASEE**), a novel paradigm integrating schema paraphrasing with schema retrieval-augmented generation, by decomposing the event extraction task into schema retrieval and schema-aware extraction. Figure 1 shows an example of ASEE in broad domains. In particular, ASEE extensively builds event extraction schemas, adaptively retrieves specific schemas, automatically assembles event extraction prompts, and accurately generates targeted structures.

To resolve the second limitation, we construct the **M**ulti-**D**imensional **S**chema-aware **E**vent **E**xtraction (**MD-SEE**) benchmark by systematically consolidating 12 datasets, enabling rigorous joint evaluation on schema retrieval and extraction accuracy across various domains, complexities, and language settings, addressing the current evaluation vacuum.

Our principal contributions are trifold:

- **Adaptive Schema-aware EE Framework**: We propose ASEE, the first framework that jointly enhances the event extraction capability through LLM-based schema paraphrasing and schema-retrieval augmented generation.
- **Multi-Dimensional EE Benchmark**: We construct MD-SEE, the first benchmark evaluating both schema matching and extraction accuracy across domains, complexity, and language settings.
- **Empirical Insights**: Through extensive experiments and analysis, we provide insightful results for event extraction with LLMs across broad domains, providing actionable guidelines for industrial deployment.

2 Related Work

Information Extraction Task Information extraction (IE) aims to automatically extract structured knowledge from unstructured texts, typically involving three core tasks: (1) Named Entity Recognition (NER) for identifying entity boundaries and types (Ye et al., 2024), (2) Relation Extraction (RE) for detecting semantic relationships between entities (Wang et al., 2023b), and (3) Event Extraction (EE) for recognizing event triggers and their associated arguments (Wang et al., 2022a). These tasks are conventionally categorized as closed IE (with predefined schemas) or open IE (schema-agnostic). Closed IE relies on predefined schemas specifying target entities, relations, or event structures, enabling precise extraction in constrained domains (Lu et al., 2022). Open IE systems, conversely, extract open-domain triples without schema constraints, sacrificing precision for broader coverage (Wang et al., 2023b). Recent approaches in OIE have explored learning from syntactic structures or chunking to improve extraction quality without predefined schemas (Dong et al., 2022, 2023).

Existing event extraction methods predominantly follow the closed IE paradigm, requiring predefined schemas that specify event types, roles, and constraints. Though effective in controlled settings,

such rigid schemas face two key challenges: (1) *Oversimplification* - Most schemas abstract away domain-specific nuances (e.g., using generic role labels like "participant"), making them insufficient for guiding LLMs in real-world extraction; (2) *Semantic Gap* - Predefined schemas often mismatch actual context semantics, especially when processing cross-domain or emerging event types.

This paper proposes Adaptive Schema-aware Event Extraction (ASEE) to address real-world EE scenarios requiring schema retrieval before extraction. Unlike traditional closed EE with rigid predefined schemas, our framework dynamically adapts schemas through : (1) Schema Paraphrasing that rewrites schema elements using contextual knowledge, and (2) Schema-Retrieval Augmented Generation guided by retrieving the relevant schemas and extracting events with the matched schemas.

Event Extraction with LLMs Recently, a significant amount of work has focused on utilizing large language models (LLMs) to implement and enhance the performance of information extraction tasks. These works can be primarily divided into four aspects: (1) data augmentation: leveraging LLMs to generate synthetic data or augment existing datasets for information extraction tasks, avoiding the introduction of unrealistic, misleading, and offset patterns (Wang et al., 2023a). (2) prompt engineering: crafting effective prompts for LLMs to guide them in extracting relevant information from text (Ashok and Lipton, 2023). (3) code LLMs: utilizing code-based LLMs to automate and enhance the information extraction process (Guo et al., 2023). (4) LLM fine-tuning: fine-tuning LLMs on domain-specific data to optimize their performance on targeted information extraction tasks (Zhou et al., 2023).

However, when using LLMs for event extraction (EE), they tend to experience significant hallucinations due to insufficient task-specific optimization, and their limited context window makes it impractical to rely solely on LLMs to handle event extraction tasks across diverse scenarios. Retrieval-augmented generation (RAG) methods have emerged as a promising solution, enhancing extraction accuracy by incorporating external knowledge (Gao et al., 2023; Huang et al., 2025b). For instance, Shiri et al. (2024) decompose event extraction into two subtasks: Event Detection (ED), which retrieves relevant event examples, and Event Argument Extraction (EAE), which extracts events based on the retrieved examples.

In this paper, we aim to mitigate the hallucination issue of LLMs in EE by adopting schema-retrieval augmented generation, while also addressing the issue of LLM context length limitations.

3 Methodology

3.1 Preliminaries

Schema-aware Event Extraction (SEE) (Shiri et al., 2024; Gui et al., 2024) leverages predefined schemas to guide the extraction of structured information from unstructured text. A schema serves as a formal representation of the types of information to be extracted, encompassing entities, relationships, events, and their attributes. By defining the structure and constraints of the desired data, schemas enable the extraction system to identify and organize relevant information systematically.

Given a query text q, ..., and a schema s [1] that specifies a particular type of information to be extracted along with its associated arguments, the goal of SEE is to extract relevant information from q according to s. Each schema s consists of a set of arguments $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$, where K is the number of arguments in schema s.

Formally, the Schema-aware Event Extraction (SEE) task can be represented as $\mathcal{V} = \theta(s, q)$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ denotes the set of extracted values corresponding to the arguments in schema s, and θ represents the extraction model that takes schema s and query q as inputs to produce the extracted information.

SEE assumes the availability of predefined schemas, which may not always be feasible in real-world scenarios. Acquiring accurate and comprehensive schemas from a wide range of domains and languages can be resource-intensive and challenging. Our proposed Adaptive Schema-aware Event Extraction (ASEE) framework addresses this issue by incorporating schema paraphrasing and retrieval, enabling robust and versatile event extraction across diverse scenarios.

3.2 Adaptive Schema-aware EE (ASEE)

Building upon the foundational concepts of SEE (Shiri et al., 2024; Gui et al., 2024), we introduce our proposed framework, Adaptive Schema-

[1] In this work, a schema refers to the definition of a single event type, including its name, a description, and a set of arguments (or roles) to be filled. For example, a "Research" schema might include arguments like Topic, Institution, and Location

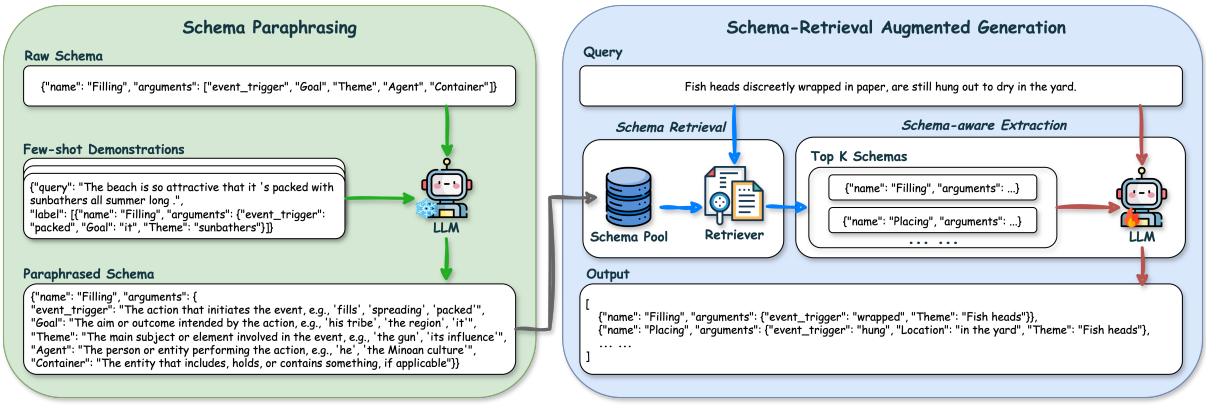

Figure 2: The architecture of ASEE comprises two primary components: Schema Paraphrasing (SP) and Schema-Retrieval Augmented Generation (SRAG, including Schema Retrieval (SR) and Schema-aware Extraction (SE)).

aware Event Extraction (ASEE). ASEE addresses the inherent challenges of SEE by incorporating adaptive mechanisms for schema retrieval and extraction. The framework is visually represented in Figure 2 and comprises two primary components: Schema Paraphrasing (SP) and Schema-Retrieval Augmented Generation (SRAG, including Schema Retrieval (SR) and Schema-aware Extraction (SE)).

The ASEE framework can be formally represented by the following sequence of operations, with each step explained alongside its corresponding formula:

• **Step 1 - Schema Paraphrasing (SP):**

$$\mathcal{S} = \bigcup_{s \in \mathcal{S}_0} \{\phi_{\text{LLM}}(s, D_s) \,|\, D_s \subseteq \mathcal{D}_{\text{train}}\} \quad (1)$$

Equation 1 represents the schema paraphrasing process. Here, ϕ_{LLM} denotes the schema paraphrasing function, which takes each initial schema s from the set $\mathcal{S}_0$ and generates a paraphrased schema pool $\mathcal{S}$ using a subset of the training data D_s as few-shot examples.

• **Step 2 - Schema Retrieval (SR):**

$$\mathcal{R}_q = \psi_{\text{retriever}}(q, \mathcal{S}) \quad (2)$$

Equation 2 describes the schema retrieval step. The function $\psi_{\text{retriever}}$ is the retrieval model that takes the query text q and the paraphrased schema pool $\mathcal{S}$ to retrieve the top-k relevant schemas, resulting in the set $\mathcal{R}_q$.

• **Step 3 - Schema-aware Extraction (SE):**

$$\mathcal{V} = \theta_{\text{LLM}}(q, \mathcal{R}_q) \quad (3)$$

Equation 3 outlines the information extraction process. The function θ_{LLM} represents the LLM fine-tuned for extraction, which takes the query q and the retrieved schemas $\mathcal{R}_q$ to produce the final set of extracted values $\mathcal{V}$.

3.2.1 Schema Paraphrasing (SP)

Schema Paraphrasing (SP) serves as the preparation stage and backbone for the extraction process, establishing a robust and comprehensive schema pool. For a given event extraction task, we first collect all relevant schemas $\mathcal{S}_0$. For each schema $s \in \mathcal{S}_0$, we utilize data samples from the training set that adhere to schema s as few-shot demonstrations to guide a frozen LLM in generating paraphrased versions of the original schema. These paraphrased schemas introduce detailed argument descriptions, enhancing the semantic clarity of the schemas and facilitating more effective retrieval. The resulting paraphrased schemas are aggregated to form the schema pool $\mathcal{S}$, which serves as a repository of diverse schemas tailored for various extraction tasks, as defined in Equation 1.

3.2.2 Schema-Retrieval Augmented Generation (SRAG)

The extraction component encompasses the operational aspects of the ASEE framework, divided into Schema Retrieval (SR) and Schema-aware Extraction (SE) processes.

Schema Retrieval (SR) Upon receiving a new query text q, the first task is to identify the most relevant schemas from the schema pool $\mathcal{S}$. This is achieved through a schema retrieval mechanism that employs a specialized retrieval model ψ. The retriever processes the query q to search the schema pool and retrieves the top-k schemas $\mathcal{R}_q = \{s_1, s_2, \ldots, s_k\}$ that are most pertinent to the information contained within q, as defined in Equation 2. This retrieval step ensures that the subsequent extraction is guided by schemas that are contextually aligned with the query, enhancing extraction accuracy and relevance.

Schema-aware Extraction (SE) With the relevant schemas $\mathcal{R}_q$ identified, the next step involves the information extraction process. While large language models (LLMs) can directly perform extraction based on schemas, their zero-shot performance may be suboptimal due to challenges in strictly adhering to schema constraints or interpreting complex argument definitions. To address this, we employ **Supervised Fine-Tuning (SFT)** to adapt the base LLM for schema-guided extraction.

Given a dataset $\mathcal{D}_{\text{train}} = \{(q_i, s_i, \mathcal{V}_i)\}_{i=1}^{N}$ containing query texts, schemas, and ground-truth structured outputs, we fine-tune the LLM to minimize the discrepancy between its predictions and the target values. Formally, for each sample $(q, s, \mathcal{V})$, the model θ is optimized to maximize the likelihood of generating the correct argument values conditioned on the input query q and schema s. The SFT objective is defined as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\sim\mathcal{D}_{\text{train}}} \sum_{k=1}^{K} \log P_\theta \left(v_k \mid q, s, v_{<k}\right) \quad (4)$$

where v_k denotes the k-th argument value in $\mathcal{V}$, and $v_{<k}$ represents previously generated values. This autoregressive training enables the model to learn schema-specific dependencies and formatting constraints. After fine-tuning, the LLM θ_{LLM} in Equation 3 becomes specialized in generating structured outputs that strictly comply with the retrieved schemas $\mathcal{R}_q$. The final output $\mathcal{V}$ comprises structured information extracted from q, organized according to the specifications of $\mathcal{R}_q$.

4 Benchmark Construction

We construct the **M**ulti-**D**imensional **S**chema-aware **E**vent **E**xtraction (**MD-SEE**) benchmark to facilitate rigorous evaluation, which systematically consolidates 12 datasets across diverse domains, complexity levels, and language settings. Due to space constraints, additional details are provided in the Appendix A.

4.1 Data Collection

Datasets were collected from various sources, focusing on those that are fully open-source to avoid licensing restrictions. The collected datasets include DocEE (Tong et al., 2022), MAVEN-Arg (Wang et al., 2024), GENEVA (Parekh et al., 2023), CrudeOilNews (Lee et al., 2022), and the event extraction datasets from IEPILE (Gui et al., 2024), which contains out-of-domain testing tests to evaluate the generalize capability. Notably, IEPILE is a collection that itself consolidates multiple individual datasets (e.g., CASIE, PHEE, RAMS, DuEE-fin). For clarity in our main experiments, we treat the English and Chinese subsets of IEPILE as two unified macro-datasets. Table 5 in Appendix A provides a summary of the collected datasets.

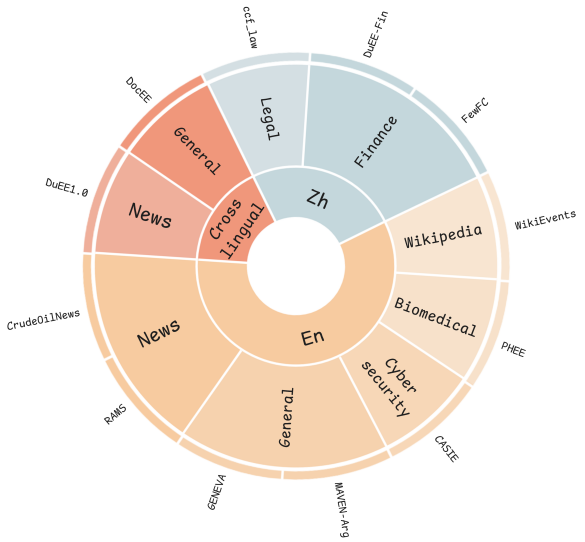

Figure 3: Dataset used in MD-SEE.

We perform initial cleaning to ensure data quality, detailed procedures are provided in Appendix A.1 for schema processing and Appendix A.2 for dataset processing.

4.2 MD-SEE Dataset

To demonstrate the capabilities of ASEE, we developed the **M**ulti-**D**imensional **S**chema-aware **E**vent **E**xtraction (**MD-SEE**) dataset. Constructed by aggregating the collected datasets, MD-SEE ensures comprehensive coverage across a wide range of non-overlapping schemas. The multidimensional nature of MD-SEE is characterized by:

- **Various Query Lengths**: Supports sentence-level to document-level queries, enabling evaluation across varying textual granularities.
- **Multiple Domains**: Incorporates datasets from diverse domains such as news, cybersecurity, biomedical, finance, and legal sectors, ensuring generalization across varied contexts.
- **Diverse Event Complexity**: Accommodates both single-event and multi-event extraction scenarios, testing the system's adaptability in handling

complex event structures.

• **Multiple Language Settings**: Encompasses English-only, Chinese-only, and cross-lingual extraction subsets, highlighting proficiency in multilingual and cross-lingual information extraction.

4.2.1 Schema Consolidation

To enhance schema diversity and reduce redundancy, we performed schema consolidation. We merged all schemas from individual datasets and conducted a manual merging step to unify nearly identical schemas and cross-lingual duplicates (English and Chinese). Detailed procedures are the same as in Appendix A.1.

Subsequently, we used the multilingual sentence embedding model BGE-M3 (Chen et al., 2024) to encode schemas. We constructed a graph where each node represents a schema, and edges connect schemas with cosine similarity above 0.85. Applying a Greedy Maximum Independent Set algorithm (Algorithm 1 in Appendix A.3), we identified the largest possible subset of diverse schemas by removing highly similar ones. We then filtered the dataset to include only the samples associated with these schemas, ensuring relevance and alignment within MD-SEE.

4.2.2 Cross-Lingual Subset

To incorporate cross-lingual extraction capabilities, we processed subsets from DocEE and DuEE1.0:

• **DocEE**: After consolidation, only English queries remained. We translated all schemas from English to Chinese and adjusted the labels, resulting in English queries paired with Chinese schemas.

• **DuEE1.0**: Originally in Chinese, we translated schemas from Chinese to English and modified the labels, producing Chinese queries associated with English schemas.

These cross-lingual subsets require the model to interpret schemas in one language while extracting information from queries in another, preserving the extracted values in the query's original language. This setup evaluates the model's ability to generalize across languages and handle multilingual scenarios common in real-world applications.

This setup evaluates the model's ability to generalize across languages and handle multilingual scenarios common in real-world applications, a challenge actively explored in recent research (Devlin et al., 2019; Liang et al., 2020; Conneau et al., 2020; Nie et al., 2023; Li et al., 2023a).

4.2.3 Dataset Statistics

Shown as Figure 3, the consolidated MD-SEE dataset comprises 300 schemas covering multiple dimensions, 12,817 training samples, 1,775 development samples, and 7,686 test samples. Detailed statistics are provided in Table 6 (Appendix A.4). This aggregation ensures that MD-SEE is robust and versatile, catering to various event extraction tasks across different contexts.

5 Experiments

5.1 Evaluation Settings

Our evaluation considers three main aspects: 1) schema retrieval evaluation, 2) schema-aware extraction evaluation, and 3) end-to-end evaluation, providing a comprehensive assessment across various dimensions.

5.1.1 Schema Retrieval Evaluation

Given a query ranging from a sentence to a full document, the schema retrieval task expects to retrieve the most relevant schemas from a schema pool.

Metric We use **Recall@K** to measure the proportion of ground-truth schemas included in the top-K retrieved schemas for each query. A high Recall@K indicates effective identification of potential schemas.

5.1.2 Schema-aware Extraction Evaluation

Provided with both the query q and the corresponding ground-truth schemas S_q, the objective of the extraction evaluation is to extract information accurately according to the paraphrased schemas.

Metric We evaluate extraction performance by calculating the **F1** score for each argument within each schema and then averaging these scores over all arguments and schemas.

5.1.3 End-to-End Evaluation

The end-to-end evaluation assesses the system's performance in autonomously retrieving relevant schemas and extracting information based on the matched schemas without prior knowledge of the ground-truth schemas. For each query, the system:

- **Schema Retrieval**: Retrieve a set of schemas R_q based on the query q.
- **Schema-aware Extraction**: Extract information from the query q according to retrieved schemas R_q.

Metric We use a modified End-to-End F1 Score (**E2E-F1**), defined as:

$$F1 = \frac{1}{N|S_q|} \sum_{q=1}^{N} \sum_{s \in S_q} F1(s, q) \cdot \mathbb{I}(s \in R_q) \quad (5)$$

Here, $\mathbb{I}(s \in R_q)$ is an indicator function that equals 1 if schema s is retrieved for query q, and 0 otherwise. The E2E-F1 considers the following cases:
• **Schema Retrieved and Relevant**: If a ground-truth schema $s \in S_q$ is retrieved ($s \in R_q$), we evaluate the extraction for that schema using $F1(s, q)$.
• **Schema Not Retrieved**: If a ground-truth schema $s \in S_q$ is not retrieved ($s \notin R_q$), we assign an F1 score of zero for that schema.
• **Schema Retrieved but Not Relevant**: Retrieved schemas not in the ground truth ($s \in R_q$, $s \notin S_q$) are ignored in the evaluation.

This evaluation framework mirrors real-world scenarios where both retrieval and extraction accuracy are critical, emphasizing the importance of effectively identifying and extracting relevant information.

5.2 Experimental Settings

Retrieval Models. We employed the following seven retrieval models, which are commonly used in RAG scenarios and support multiple languages, including: BM25 (Robertson and Zaragoza, 2009), BGE-M3, BGE-Reranker-Base (BGE-RB), and BGE-Reranker-Large (BGE-RL) (Chen et al., 2024), E5-large-v2 (E5-LV2) (Wang et al., 2022b), GTE-Large (GTE-L) (Li et al., 2023b) LLM-Embedder (LLM-E) (Zhang et al., 2024).

LLMs. Within the limits of our computational resources, we employed the following state-of-the-art large language models and information extraction models to carry out our information extraction tasks, including: Phi-3.5-mini (Abdin et al., 2024), Llama-3.2-3B and Llama-3.1-8B (Dubey et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2.5-7B and Qwen2.5-14B (Yang et al., 2024), YAYI-UIE (Xiao et al., 2023), GPT-4-turbo (Wada et al., 2024)

Datasets We conduct experiments on several collected datasets, including MAVEN-Arg, GENEVA, CrudeOilNews, DocEE (-zh & -en), IEPILE (-zh & -en), and our newly developed MD-SEE dataset. These datasets span multiple scenarios, providing comprehensive insights into our ASEE framework.

5.3 Schema Retrieval Evaluation

To verify the effectiveness of our proposed ASEE method in improving retrieval performance by paraphrasing the raw schema, we conducted the following experiments. First, we performed the retrieval experiments to the collected individual datasets, computing Recall@10 for both raw schema (denoted as "Raw") and schema paraphrasing (denoted as "Paraph."). In most cases presented in Table 1, the "Paraph." consistently showed marked improvements over "Raw" across different retrieval models.

As an addition, we calculated the Recall@10, Recall@20, and Recall@50 for both the raw schema and the schema paraphrasing on the MD-SEE dataset. The results, as shown in Table 2, indicate that "Paraph." outperforms "Raw" across all metrics—Recall@10, Recall@20, and Recall@50—for all retrieval models. These experiments collectively demonstrate that paraphrasing the raw schema using our proposed few-shot demonstration paraphrasing can effectively enhance schema retrieval performance.

5.4 Schema-aware Extraction Evaluation

To benchmark the core extraction capabilities of various LLMs on our consolidated datasets, we designed a schema-aware extraction evaluation. This evaluation is intentionally isolated from the retrieval task; its purpose is to assess how well different models can perform extraction when provided with the ground-truth paraphrased schemas. By bypassing the schema retrieval step, we can measure the upper-bound performance of the extraction component alone. As shown in Table 3, we used six state-of-the-art open-source LLMs (i.e., Phi-3.5-mini, Llama-3.2-3B, Llama-3.1-8B, Mistral-7B-v0.3, Qwen2.5-7B, Qwen2.5-14B), an information extraction model based on an LLM (i.e., YAYI-UIE), and the popular closed-source model GPT-4 (i.e., GPT-4-turbo). Notably, we *bypassed the schema matching step and directly provided each extraction task with an optimal event extraction schema*.

The results clearly indicate that GPT-4 has the strongest zero-shot event extraction capabilities, outperforming all other models on most datasets. However, in the Chinese data sets DocEE-zh and IEPILE-zh, it was closely matched by Qwen2.5-14B, which specializes in Chinese. We also observed that YAYI-UIE, a universal information extraction method, did not achieve satisfactory event

	CrudeOilNews		DocEE-en		DocEE-zh		GENEVA		IEPILE-en		IEPILE-zh		MAVEN-Arg	
	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.
BM25	**0.35**	0.31	0.21	**0.67**	0.55	**0.77**	0.11	**0.41**	0.61	**0.78**	0.75	**0.81**	0.09	**0.22**
BGE-M3	**0.35**	0.32	0.85	**0.92**	0.73	**0.94**	0.43	**0.50**	0.77	**0.88**	0.84	**0.91**	0.13	**0.33**
E5-LV2	0.35	**0.36**	0.84	**0.87**	0.41	**0.51**	0.43	**0.51**	0.83	**0.90**	**0.46**	0.30	0.20	**0.31**
GTE-L	0.33	**0.37**	0.95	0.95	0.47	**0.50**	0.40	**0.48**	0.84	**0.89**	0.34	**0.50**	0.19	**0.32**
LLM-E	0.30	**0.31**	0.88	**0.90**	**0.34**	0.33	0.40	**0.51**	0.82	**0.88**	0.27	**0.53**	0.17	**0.30**
BGE-RB	0.31	**0.33**	0.43	**0.49**	**0.48**	0.44	0.33	**0.47**	0.66	**0.83**	**0.89**	0.87	0.10	**0.16**
BGE-RL	0.28	**0.31**	0.57	**0.77**	0.66	**0.69**	0.41	0.41	0.78	**0.81**	0.89	**0.90**	0.14	**0.20**

Table 1: Schema retrieval evaluation results on individual datasets, using Recall@10 as the metric. Better results are highlighted in **bold**.

	Recall@10		Recall@20		Recall@50	
	Raw	Paraph.	Raw	Paraph.	Raw	Paraph.
BM25	0.33	**0.58**	0.39	**0.67**	0.49	**0.76**
BGE-M3	0.61	**0.78**	0.69	**0.86**	0.78	**0.94**
E5-LV2	0.35	**0.61**	0.43	**0.72**	0.62	**0.86**
GTE-L	0.28	**0.57**	0.36	**0.68**	0.51	**0.82**
LLM-E	0.34	**0.71**	0.49	**0.82**	0.68	**0.91**
BGE-RB	0.51	**0.59**	0.60	**0.66**	0.70	**0.76**
BGE-RL	0.56	**0.61**	0.64	**0.69**	0.75	**0.79**

Table 2: Schema retrieval evaluation results on MD-SEE. Better results are highlighted in **bold**.

extraction results compared to the open-source LLM-based methods. Despite having 14 billion parameters, its performance lagged behind due to the limited capabilities of its backbone LLM, performing even worse than Llama-3.2-3B.

5.5 End-to-End Evaluation

To assess the practical performance of our proposed **ASEE framework**, we conducted an end-to-end evaluation on the MD-SEE benchmark. Unlike the previous evaluation which focused solely on retrieval or extraction, this setup mirrors a real-world scenario by integrating both schema retrieval and schema-aware extraction. The system is tasked with autonomously retrieving relevant schemas and then extracting information based on them. The results from this evaluation, presented in Table 4, serve as a strong baseline for our MD-SEE benchmark and demonstrate the effectiveness of the ASEE pipeline as a whole.

Table 4 shows that ASEE, using BGE-M3 as the schema retriever, achieves the best overall performance in E2E-F1 compared to other retrieval models (e.g., BM25) across both Llama-3.2-3B and Llama-3.1-8B, with and without SFT, for schema-aware event extraction. The improved retrieval enabled by schema paraphrasing allows ASEE to identify more relevant schemas, thereby providing a stronger foundation for accurate event extraction.

In addition, ASEE with Llama-3.1-8B achieves better E2E-F1 performance than ASEE with Llama-3.2-3B, which is a smaller and less powerful LLM. The end-to-end event extraction performance of ASEE can be further enhanced when the extraction LLM (e.g., Llama-3.2-3B or Llama-3.1-8B) is trained with supervised fine-tuning (SFT). These results demonstrate that ASEE benefits from using a larger and/or fine-tuned LLM for schema-aware extraction.

Overall, our proposed ASEE method can be enhanced either by a **strong retrieval model** for paraphrased schema matching or by a **larger** and/or **fine-tuned** LLM for schema-aware extraction.

6 Conclusion

In this paper, we introduced Adaptive Schema-aware Event Extraction (ASEE), a novel framework designed to enhance event extraction across diverse domains using large language models. By decomposing the extraction process into schema paraphrasing and schema retrieval-augmented extraction, ASEE effectively mitigates challenges such as hallucinations and context length limitations inherent in LLM-based approaches. We develop the MD-SEE dataset, which consists of high-quality schemas across various dimensions, providing a comprehensive resource for evaluating event extraction systems. We conducted extensive experiments on multiple datasets including our newly developed MD-SEE dataset, demonstrating ASEE's adaptability and superior extraction performance. The framework's ability to build and leverage a comprehensive schema pool enables more precise and scalable event extraction, making it a robust solution for a wide range of real-world applications. Future work will explore integrating relation extraction and named entity recognition, fine-tuning larger LLMs, and extending ASEE to more complex multilingual and cross-lingual scenarios.

	CrudeOilNews	DocEE-en	DocEE-zh	GENEVA	IEPILE-en	IEPILE-zh	MAVEN-Arg
Phi-3.5-mini	0.22	0.43	0.46	0.31	0.33	0.60	0.29
Llama-3.2-3B	0.35	0.47	0.60	0.47	0.44	0.71	0.37
Llama-3.1-8B	0.43	0.48	0.61	0.58	0.55	0.71	0.51
Mistral-7B-v0.3	0.37	0.45	0.57	0.57	0.49	0.71	0.47
Qwen2.5-7B	0.42	0.45	0.41	0.60	0.55	0.73	0.48
Qwen2.5-14B	0.30	0.49	**0.62**	0.49	0.53	**0.78**	0.40
YAYI-UIE	0.35	0.37	0.30	0.43	0.41	0.34	0.33
GPT-4-turbo	**0.50**	**0.56**	**0.62**	**0.67**	**0.62**	0.77	**0.56**

Table 3: Zero-shot schema-aware extraction results, using F1 score as the metric.The best results are marked in **bold**, while the second underlined.

	BM25	BGE-M3	E5-LV2	GTE-LG	LLM-E	BGE-RB	BGE-RL
Llama-3.2-3B	0.46	0.62	0.48	0.46	0.57	0.47	0.48
w/ SFT	**0.47**	**0.63**	**0.49**	**0.47**	**0.58**	**0.48**	**0.49**
Llama-3.1-8B	0.48	0.65	0.51	0.47	0.59	0.49	0.51
w/ SFT	**0.50**	**0.68**	**0.53**	**0.50**	**0.62**	**0.51**	**0.53**

Table 4: End-to-end evaluation of our proposed **ASEE** framework on the MD-SEE benchmark. Results are shown across different retrieval and extraction model configurations using E2E-F1 score as the metric. Better results are highlighted in **bold**.

7 Limitations

Our Adaptive Schema-aware Event Extraction (ASEE) framework has several limitations. First, we did not incorporate relation extraction (RE) and named entity recognition (NER) due to our focus on schema-based extraction tasks, for which event extraction is more suited. Second, we were unable to fine-tune larger language models, such as those with 32B or 70B parameters, owing to computational resource constraints. Additionally, our current implementation does not address more complex multilingual and cross-lingual scenarios, which presents further challenges for scalable and versatile information extraction. We hope to address the above limitations in the follow-up work.

References

Marah Abdin, Sam Ade Jacobs, and Ammar Ahmad Awan. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

D Ashok and ZC Lipton. 2023. Promptner: Prompting for named entity recognition. arxiv 2023. *arXiv preprint arXiv:2305.15444*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, 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.

Kuicai Dong, Aixin Sun, Jung-Jae Kim, and Xiaoli Li. 2022. Syntactic multi-view learning for open information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4072–4083, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kuicai Dong, Aixin Sun, Jung-jae Kim, and Xiaoli Li. 2023. Open information extraction via chunks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15390–15404, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. 2025. RAPID: Efficient retrieval-augmented long text generation with writing planning and information discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16742–16763, Vienna, Austria. Association for Computational Linguistics.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. IEPile: Unearthing large scale schema-conditioned information extraction corpus. pages 127–146, Bangkok, Thailand.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yuqing Huang, Rongyang Zhang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, Yao Hu, Xuyang Zhi, Guiquan Liu, Xin Li, et al. 2025b. Selfaug: Mitigating catastrophic forgetting in retrieval-augmented generation via distribution self-alignment. *arXiv preprint arXiv:2509.03934*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Viet Dac Lai. 2022. Event extraction: A survey. *arXiv preprint arXiv:2210.03419*.

Meisin Lee, Lay-Ki Soon, Eu Gene Siew, and Ly Fie Sugianto. 2022. CrudeOilNews: An annotated crude oil news corpus for event extraction. pages 465–479, Marseille, France. European Language Resources Association.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023a. From classification to generation: Insights into crosslingual retrieval augmented icl. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *ArXiv*, abs/2308.03281.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

Hang Lv, Sheng Liang, Hao Wang, Hongchao Gu, Yaxiong Wu, Wei Guo, Defu Lian, Yong Liu, and Enhong Chen. 2025. Costeer: Collaborative decoding-time personalization via local delta steering. *arXiv preprint arXiv:2507.04756*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. pages 3664–3686, Toronto, Canada.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Fatemeh Shiri, Van Nguyen, Farhad Moghimifar, John Yoo, Gholamreza Haffari, and Yuan-Fang Li. 2024. Decompose, enrich, and extract! schema-aware event extraction using llms. *Preprint*, arXiv:2406.01045.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. pages 3970–3982, Seattle, United States.

Akihiko Wada, Toshiaki Akashi, George Shih, Akifumi Hagiwara, Mitsuo Nishizawa, Yayoi K. Hayakawa, Junko Kikuta, Keigo Shimoji, Katsuhiro Sano, Koji Kamagata, Atsushi Nakanishi, and Shigeki Aoki. 2024. Optimizing gpt-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. *Diagnostics*, 14.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. Deepstruct: Pretraining of language models for structure prediction. *arXiv preprint arXiv:2205.10475*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. Improving unsupervised relation extraction by augmenting diverse sentence pairs.

X Wang, W Zhou, C Zu, H Xia, T Chen, Y Zhang, R Zheng, J Ye, Q Zhang, T Gui, et al. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. arxiv 2023. *arXiv preprint arXiv:2304.08085*.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. pages 4072–4091, Bangkok, Thailand.

Bowen Wu, Wenqing Wang, Haoran Li, Ying Li, Jingsong Yu, and Baoxun Wang. 2025a. Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history. *Preprint*, arXiv:2503.05150.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. 2025b. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *Preprint*, arXiv:2504.15965.

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

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

An Yang, Baosong Yang, and Binyuan Hui. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, and Jian-Yun Nie. 2024. A multi-task embedder for retrieval augmented LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3537–3553, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Dataset Collection

We present the details of MD-SEE dataset collection, including schema processing (Appendix A.1), dataset processing (Appendix A.2), schema consolidation algorithm (Appendix A.3), and dataset statistics (Appendix A.4).

A.1 Schema Processing

Each dataset was examined to extract unique schemas, which include the schema name, description, arguments, and relevant metadata. We employed a heuristic merging process to address potential duplications. This included:

1. **Character Similarity:** Schemas were merged if their names and arguments shared over 80% character similarity, thus reducing redundancy due to variations in tense or plural forms.

2. **Numerical and Variant Arguments:** For arguments with numerical variants (e.g., place, place1, place2), we consolidated them into a single argument with values stored in a list, to enhances the schema's clarity and usability.

These efforts ensure that our schema pool reflects a rich and diverse set of scenarios, ultimately enhancing the robustness of the data.

A.2 Dataset Processing

We follow these principles:

1. **Split Handling:** If a dataset contains original train/dev/test splits, we adhered to them even if some splits were missing. For datasets with only a train split, we implemented an 80/10/10 split for consistency.

2. **Filtering Criteria:** To maintain quality, we filtered out data instances with more than 15 extracted event labels, especially in MAVEN-Arg, where instances with excessive labels could skew results.

3. **Query Length Diversity:** We ensured the datasets included varied lengths of queries, from single sentences to longer documents, enriching the task complexity and addressing different real-world scenarios.

A.3 Greedy Maximum Independent Set

Algorithm 1 shows the algorithm for identifying the largest possible subset of diverse schemas for schema consolidation.

Algorithm 1 Greedy Maximum Independent Set

Require: Adjacency list adj_list representing schema similarities
Ensure: Maximum independent set of schemas

remaining $\leftarrow$ {all schema indices in adj_list}
independent_set $\leftarrow$ {}
degrees $\leftarrow$ $\{i : \text{len}(\text{adj\_list}[i]) \mid i \in \text{adj\_list}\}$
while remaining is not empty **do**
  node $\leftarrow$ $\text{argmin}_{i \in \text{remaining}}(\text{degrees}[i])$
  independent_set $\leftarrow$ independent_set $\cup$ {node}
  remaining $\leftarrow$ remaining $\setminus$ {node}
  remaining $\leftarrow$ remaining $\setminus$ adj_list[node]
  for each neighbor $n \in$ adj_list[node] **do**
    degrees $\leftarrow$ degrees $\setminus$ $\{n\}$
    for each neighbor $m \in$ adj_list$[n]$ **do**
      if m in remaining **then**
        degrees$[m] \leftarrow$ degrees$[m] - 1$
      end if
    end for
  end for
  degrees $\leftarrow$ degrees $\setminus$ {node}
end while
return independent_set

A.4 Dataset Statistics

Table 5 shows the statistics of the collected datasets before processing, such as CrudeOilNews, GENEVA, MAVEN-Arg, DocEE-en, DocEE-zh, IEPILE-en, and IEPILE-zh. Table 6 shows the statistics of the processed MD-SEE dataset, including source, language, domain, split distributions, and maximum number of labels per sample.

B End-to-End Experimental Results

The following tables show the end-to-end results of CrudeOilNews (Table 7), DocEE-en (Table 8), DocEE-zh (Table 9), GENEVA (Table 10), IEPILE-en (Table 11), IEPILE-zh (Table 12), and MAVEN-Arg (Table 13).

C Examples

Dataset	Source	Language	Domain	#Schemas	#Train	#Dev	#Test	#Max_Labels
CrudeOilNews	-	en	Oil News	18	1489	-	265	10
GENEVA	-	en	General	115	1922	778	931	10
MAVEN-Arg	-	en	General	162	2913	-	710	15
DocEE-en	DocEE	en	General	59	21966	2748	2771	1
DocEE-zh	DocEE	zh	General	58	29383	3672	3674	1
IEPILE-en	CASIE†	en	Cybersecurity	5	3732	777	1492	1
	PHEE†	en	Biomedical	2	2897	960	968	1
	RAMS†	en	News	106	-	-	887	1
	WikiEvents†	en	Wikipedia	31	-	-	249	5
IEPILE-zh	DuEE-fin†	zh	Finance	13	7015	-	1171	14
	DuEE1.0†	zh	News	65	11908	-	1492	15
	FewFC†	zh	Finance	5	-	-	2879	5
	ccf_law†	zh	Legal	9	-	-	971	10

Table 5: Statistics of the collected datasets including language, domain, split distributions, and maximum number of labels per sample. Datasets marked with † were collected from IEPILE and further processed in this work.

Source	Language	Domain	#Schemas	#Train	#Dev	#Test	#Max_Labels
DocEE	en/zh	General	52	2000	200	1000	1
DuEE1.0	zh/en	News	53	2000	-	1000	8
CrudeOilNews	en	Oil News	11	949	-	198	6
GENEVA	en	General	61	1278	200	600	7
MAVEN-Arg	en	General	30	590	-	150	8
CASIE	en	Cybersecurity	4	2000	200	1000	1
PHEE	en	Biomedical	1	2000	200	867	1
RAMS	en	News	52	-	-	286	1
WikiEvents	en	Wikipedia	18	-	-	66	2
DuEE-fin	zh	Finance	7	2000	-	558	11
FewFC	zh	Finance	5	-	-	1000	5
ccf_law	zh	Legal	6	-	-	961	9
Total	-	-	300	12817	800	7686	11

Table 6: Statistics of the MD-SEE dataset, including source, language, domain, split distributions, and maximum number of labels per sample. Mixed languages (en/zh or zh/en) indicate the language of queries and schemas respectively, representing subsets for cross-lingual extraction.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.08	0.12	0.15	0.13	0.15	0.10	0.12	0.17
BGE-M3	0.08	0.12	0.15	0.13	0.15	0.10	0.12	0.17
E5-LV2	0.08	0.12	0.15	0.13	0.15	0.10	0.12	0.17
GTE-L	0.07	0.12	0.14	0.12	0.14	0.10	0.12	0.17
LLM-E	0.07	0.11	0.13	0.11	0.13	0.09	0.11	0.15
BGE-RB	0.07	0.11	0.13	0.11	0.13	0.09	0.11	0.15
BGE-RL	0.06	0.10	0.12	0.10	0.12	0.08	0.10	0.14
				Paraph.				
BM25	0.07	0.11	0.13	0.11	0.13	0.09	0.11	0.15
BGE-M3	0.07	0.11	0.14	0.12	0.14	0.10	0.11	0.16
E5-LV2	0.08	0.13	0.16	0.13	0.15	0.11	0.13	0.18
GTE-L	0.08	0.13	0.16	0.14	0.15	0.11	0.13	0.18
LLM-E	0.07	0.11	0.14	0.12	0.13	0.09	0.11	0.16
BGE-RB	0.07	0.11	0.14	0.12	0.14	0.10	0.11	0.16
BGE-RL	0.07	0.11	0.13	0.11	0.13	0.09	0.11	0.15

Table 7: End to end results of CrudeOilNews.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.09	0.10	0.10	0.09	0.09	0.10	0.08	0.12
BGE-M3	0.36	0.40	0.41	0.38	0.38	0.41	0.31	0.47
E5-LV2	0.36	0.40	0.40	0.38	0.38	0.41	0.31	0.47
GTE-L	0.41	0.45	0.46	0.43	0.43	0.47	0.35	0.53
LLM-E	0.38	0.41	0.42	0.39	0.39	0.43	0.32	0.49
BGE-RB	0.19	0.20	0.21	0.19	0.19	0.21	0.16	0.24
BGE-RL	0.25	0.27	0.27	0.26	0.26	0.28	0.21	0.32
				Paraph.				
BM25	0.29	0.31	0.32	0.30	0.30	0.33	0.25	0.37
BGE-M3	0.39	0.43	0.44	0.41	0.41	0.45	0.34	0.51
E5-LV2	0.37	0.41	0.42	0.39	0.39	0.42	0.32	0.49
GTE-L	0.41	0.45	0.45	0.43	0.43	0.46	0.35	0.53
LLM-E	0.38	0.42	0.43	0.40	0.40	0.44	0.33	0.50
BGE-RB	0.21	0.23	0.23	0.22	0.22	0.24	0.18	0.27
BGE-RL	0.33	0.36	0.37	0.34	0.34	0.37	0.28	0.43

Table 8: End to end results of DocEE-en.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.25	0.33	0.33	0.31	0.22	0.34	0.16	0.31
BGE-M3	0.34	0.44	0.45	0.42	0.30	0.45	0.22	0.44
E5-LV2	0.19	0.24	0.25	0.23	0.17	0.25	0.12	0.23
GTE-L	0.22	0.28	0.29	0.27	0.19	0.29	0.14	0.26
LLM-E	0.16	0.20	0.21	0.19	0.14	0.21	0.10	0.19
BGE-RB	0.22	0.29	0.29	0.27	0.20	0.30	0.14	0.28
BGE-RL	0.31	0.40	0.40	0.38	0.27	0.41	0.21	0.37
				Paraph.				
BM25	0.36	0.46	0.47	0.44	0.32	0.48	0.23	0.43
BGE-M3	0.43	0.56	0.57	0.53	0.38	0.58	0.28	0.53
E5-LV2	0.24	0.30	0.31	0.29	0.21	0.31	0.15	0.28
GTE-L	0.23	0.28	0.29	0.27	0.21	0.29	0.14	0.27
LLM-E	0.15	0.19	0.20	0.18	0.14	0.20	0.10	0.18
BGE-RB	0.20	0.26	0.27	0.25	0.18	0.27	0.13	0.25
BGE-RL	0.26	0.35	0.36	0.33	0.28	0.38	0.21	0.43

Table 9: End to end results of DocEE-zh.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.03	0.05	0.06	0.06	0.07	0.05	0.05	0.07
BGE-M3	0.13	0.20	0.25	0.24	0.26	0.21	0.18	0.29
E5-LV2	0.13	0.20	0.25	0.24	0.26	0.21	0.18	0.29
GTE-L	0.13	0.19	0.23	0.22	0.24	0.20	0.15	0.27
LLM-E	0.12	0.19	0.23	0.23	0.24	0.20	0.15	0.27
BGE-RB	0.10	0.16	0.19	0.18	0.20	0.15	0.13	0.22
BGE-RL	0.13	0.19	0.24	0.23	0.25	0.20	0.15	0.27
				Paraph.				
BM25	0.13	0.19	0.24	0.23	0.25	0.20	0.18	0.27
BGE-M3	0.16	0.24	0.29	0.29	0.30	0.25	0.22	0.34
E5-LV2	0.16	0.24	0.29	0.29	0.31	0.25	0.22	0.34
GTE-L	0.15	0.22	0.28	0.27	0.29	0.24	0.20	0.32
LLM-E	0.16	0.24	0.29	0.29	0.31	0.25	0.22	0.34
BGE-RB	0.14	0.22	0.27	0.26	0.28	0.23	0.20	0.32
BGE-RL	0.13	0.19	0.24	0.24	0.25	0.20	0.18	0.27

Table 10: End to end results of GENEVA.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.20	0.27	0.33	0.30	0.33	0.32	0.25	0.38
BGE-M3	0.25	0.34	0.42	0.38	0.42	0.41	0.31	0.47
E5-LV2	0.27	0.37	0.46	0.41	0.46	0.44	0.34	0.51
GTE-L	0.28	0.37	0.46	0.41	0.46	0.44	0.34	0.52
LLM-E	0.27	0.36	0.45	0.40	0.45	0.43	0.33	0.50
BGE-RB	0.22	0.29	0.36	0.32	0.36	0.35	0.27	0.41
BGE-RL	0.26	0.34	0.43	0.38	0.43	0.41	0.32	0.48
				Paraph.				
BM25	0.26	0.34	0.43	0.38	0.43	0.41	0.32	0.48
BGE-M3	0.29	0.39	0.49	0.43	0.49	0.47	0.36	0.55
E5-LV2	0.30	0.40	0.50	0.44	0.50	0.48	0.37	0.56
GTE-L	0.29	0.39	0.49	0.43	0.49	0.47	0.36	0.55
LLM-E	0.29	0.39	0.49	0.43	0.49	0.47	0.36	0.55
BGE-RB	0.27	0.37	0.46	0.41	0.46	0.43	0.33	0.52
BGE-RL	0.27	0.36	0.44	0.40	0.45	0.43	0.33	0.50

Table 11: End to end results of IEPILE-en.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.45	0.53	0.53	0.53	0.55	0.58	0.25	0.57
BGE-M3	0.50	0.59	0.59	0.59	0.61	0.65	0.28	0.64
E5-LV2	0.28	0.33	0.33	0.33	0.34	0.36	0.16	0.36
GTE-L	0.20	0.24	0.24	0.24	0.25	0.27	0.12	0.26
LLM-E	0.16	0.19	0.19	0.19	0.20	0.21	0.09	0.19
BGE-RB	0.54	0.63	0.63	0.63	0.65	0.70	0.30	0.69
BGE-RL	0.54	0.63	0.63	0.63	0.65	0.70	0.30	0.69
				Paraph.				
BM25	0.49	0.58	0.58	0.58	0.59	0.63	0.28	0.62
BGE-M3	0.55	0.65	0.65	0.65	0.67	0.71	0.31	0.70
E5-LV2	0.18	0.21	0.21	0.21	0.22	0.23	0.10	0.23
GTE-L	0.26	0.31	0.31	0.31	0.32	0.34	0.14	0.34
LLM-E	0.32	0.37	0.37	0.37	0.39	0.41	0.18	0.40
BGE-RB	0.53	0.62	0.62	0.62	0.64	0.68	0.29	0.67
BGE-RL	0.54	0.64	0.64	0.64	0.67	0.70	0.30	0.69

Table 12: End to end results of IEPILE-zh.

	Phi-3.5-mini	Llama-3.2-3B	Llama-3.1-8B	Mistral-7B-v0.3	Qwen2.5-7B	Qwen2.5-14B	YAYI-UIE	GPT-4-turbo
				Raw				
BM25	0.03	0.04	0.05	0.04	0.04	0.04	0.03	0.05
BGE-M3	0.04	0.05	0.07	0.06	0.06	0.05	0.04	0.08
E5-LV2	0.06	0.07	0.10	0.09	0.09	0.08	0.06	0.11
GTE-L	0.06	0.07	0.10	0.09	0.09	0.08	0.06	0.11
LLM-E	0.05	0.06	0.09	0.08	0.08	0.07	0.06	0.09
BGE-RB	0.03	0.04	0.05	0.05	0.05	0.04	0.03	0.06
BGE-RL	0.04	0.05	0.07	0.06	0.07	0.06	0.05	0.08
				Paraph.				
BM25	0.06	0.08	0.11	0.10	0.10	0.09	0.07	0.12
BGE-M3	0.10	0.12	0.17	0.16	0.16	0.13	0.11	0.18
E5-LV2	0.09	0.12	0.16	0.15	0.15	0.12	0.10	0.17
GTE-L	0.09	0.12	0.16	0.15	0.15	0.13	0.10	0.18
LLM-E	0.09	0.12	0.16	0.15	0.15	0.12	0.10	0.17
BGE-RB	0.06	0.08	0.11	0.10	0.10	0.09	0.07	0.13
BGE-RL	0.06	0.08	0.10	0.10	0.10	0.08	0.07	0.11

Table 13: End to end results of MAVEN-Arg.

Schema Paraphrasing

[Instruction]
Generate natural language descriptions for the given event extraction schema, which consists of an event name and its associated argument roles, each with an empty description to be completed. Using the provided reference queries as contextual guidance, your description must:
Clearly delineate the event's semantic scope and triggers,
Specify each argument's functional role with domain-relevant examples,
Maintain schema-task alignment through query-aware contextualization.

[Schema]
{"name": "Preventing exit of a person", "arguments": "{'event_trigger': '', 'destination': '', 'origin': '', 'preventer': '', 'passenger': '', 'transporter': ''}"}

[Relevant Queries]
["Customs officers intercepted the migrant group attempting to cross the Montenegro border using forged passports last night.",
"A mother blocked her teenage daughter from boarding the Chicago-bound Greyhound bus after discovering suspicious text messages.",
"The automated facial recognition system at Dubai Airport Terminal 3 flagged and detained a passenger on the Emirates flight to London Heathrow."]

[The Start of Assistant's Response]
{"name": "Preventing exit of a person", "description": "This schema captures incidents where an individual/entity actively obstructs a person's movement from an origin point to an intended destination. The event inherently implies physical or systemic intervention to stop transit.", "arguments": "{'event_trigger': 'Verbs/Nouns indicating prevention: intercepted, blocked, detained, barred, lockdown', "destination": "The intended end location for the subject or object, e.g., 'US'", "origin": "The initial location from which the subject or object originates, e.g., 'Iran', 'jail'", "preventer": "The entity that acts to prevent the event from occurring, e.g., 'Iran', 'jail'", "passenger": "The person or object being transported or affected by the event, e.g., 'Americans', 'Harold Martin'", "transporter": "The entity responsible for moving or transporting the subject, e.g., 'Greyhound bus'}}
[The End of Assistant's Response]

Figure 4: Prompt and Example of Schema Paraphrasing

Schema Retrieval

[Query q]
The chief facility for researching and producing CW agents was a military front company called Delta G Scientific, located between Johannesburg and Pretoria, and several other facilities were set up to develop protective clothing, manufacture exotic assassination devices, and " weaponize " irritants (Riot Control Agents such as CS and CR) and incapacitants by placing them in artillery shells, mortar bombs, and grenades.

[Retrieve from the Paraphrased Schema Pool $\mathcal{S}$]
[{"name": "Construction of a structure or artifact", "description": "This event extraction task focuses on physical building processes involving permanent structures (e.g., monuments, buildings) or complex artifacts. Focuses on geographic specificity (exact construction sites) and physical tools/methods.", "arguments": "{"event_trigger": "The action that initiates the event, e.g., 'constructed', 'assembled', 'construct'", "instrument": "The tool or means used to carry out the action, if applicable", "manufacturer": "The maker or creator of the artifact or structure, e.g., 'European manufacturer'", "artifact": "The object or structure that is created, e.g., 'pathetic monuments'", "place": "The location where the event occurs, e.g., 'Ur'"}"},

{"name": "Creation or manufacturing of goods", "description": "This schema identifies production events of movable goods/technological systems (e.g., spacecraft, electronics). Emphasizes technical instrumentation and industrial actors, including virtual/digital creations.", "arguments": "{ "event_trigger": "The action that initiates the event, e.g., 'assembled'", "instrument": "An apparatus used during the event, e.g., 'scientific instruments', 'radar', 'computers'", "manufacturer": "The maker or constructor of the artifact, e.g., 'UK', 'European manufacturer Thales Alenia Space'", "artifact": "The object or item that is created, e.g., 'rover', 'Schiaparelli'", "place": "The location targeted or concerned, e.g., 'Meridiani Planum'"}"},
.....]

Figure 5: Example Input of Schema Retrieval

Schema-aware Extraction

[Instruction]
Extract from the query any fields defined in the given schema and output exactly one dictionary. Only include schema arguments that appear in the query, using their exact field names and verbatim values.

[Schema s]
{ "name": "Research", "description": "The schema focuses on identifying events related to scientific research activities.", "arguments": { "event_trigger": "Action verbs/gerunds explicitly indicating scientific research activities (e.g., researching, developing, studying)", "Topic": "Specific subject or technical domain being researched", "Institution": "Name of organization or entity that conducting the research", "Location": "Geographic location of research facilities (city, country, or geocoordinates)", "Application": "Specific applications of research outcomes" } }

[Query q]
The chief facility for researching and producing CW agents was a military front company called Delta G Scientific, located between Johannesburg and Pretoria, and several other facilities were set up to develop protective clothing, manufacture exotic assassination devices, and “ weaponize ” irritants (Riot Control Agents such as CS and CR) and incapacitants by placing them in artillery shells, mortar bombs, and grenades.

[The Start of Assistant’s Response]
{"name": "Research", "arguments": {"event_trigger": "researching", "Topic": "CW agents"}}
[The End of Assistant’s Response]

Figure 6: Prompt and Example of Schema-aware Extraction

MD-SEE Examples

[schema example]
{"name": "phishing", "description": "The information extraction task focuses on identifying key elements related to phishing attacks and incidents.", "arguments": {"damage amount": " Details on any financial or material loss resulting from the phishing attack.", "attack pattern": "Information on the method or technique used in the phishing attempt", "tool": "Identification of any software or technology used to facilitate the phishing attack.", "victim": "The individual or group targeted by the phishing scheme.", "place": "The geographical or virtual location where the phishing attack occurred.", "attacker": "Information on the person or group responsible for conducting the phishing attack.", "purpose": "The objective or goal behind the phishing attempt.", "trusted entity": "The legitimate entity being impersonated by the attacker to deceive the victim.", "time": "The timeframe during which the phishing attack took place."}, "dataset": "CASIE", "task": "EE", "lang": "en"}

[test example_1]
{"query": "There ’s likely not a person reading this online who has n’t received a phishing attack , in which someone pretending to be a bank sends an email or text message , hoping to trick you into enter or re-enter account information or a credit card number .", "label": ["name": "phishing", "arguments": "Event_trigger": "pretending to be", "attacker": "someone", "trusted entity": "a bank"], "dataset": "CASIE", "task": "EE", "lang": "en"}

[test example_2]
{"query": "From there , the attacker could spoof the website , using it to lure in victims and potentially gather credentials or spread malware .", "label": ["name": "phishing", "arguments": "Event_trigger": "spoof", "trusted entity": "the website", "attacker": "the attacker"], "dataset": "CASIE", "task": "EE", "lang": "en"}

Figure 7: MD-SEE Examples