A Self-Reflective LLM-based Architecture for Semi-Open Event Extraction

Hristo Tanev , Michel De Bollivier , Bertrand De Longueville
Joint Research Centre, European Commission
{hristo.tanev, michel.de-bollivier, bertrand.de-longueville}@ec.europa.eu

Abstract

We present a multi-agent *reflective* architecture for event extraction based on generative large language models (LLMs). Our architecture is the first of its kind to perform **Semi-Open Event Extraction (SOEE)**, a hybrid framework that combines a fixed set of event-template fields with dynamically generated attributes produced using self-reflection. A further contribution of the system is its operationalization of reflection as an internal question-generation and answering process. It is defined as the generation of questions about missing or implicit event information and finding their answers within the system itself. We model event extraction as an iterative dialogue between a **reflective** LLM-based agent, which generates questions to uncover missing event information and a set of **expert** agents, which provide domain-aware answers to these questions. The expert agents also generate the initial event template using a generative LLM. Across all evaluation experiments on articles from the health domain, **MAREA** demonstrates strong core-field extraction and effective reflective template expansion, with its three question-generation strategies producing useful additional event attributes. The observed errors were mainly due to imprecise prompt interpretation or inaccurate interpretation of the source article by the LLM, rather than to hallucination or an intrinsic inability to retrieve the requested information.

1 Introduction

LLMs are increasingly adopted within the event extraction community (Li et al., 2025), where recent work demonstrates that prompting and generative approaches can produce structured outputs in zero- and few-shot settings, creating new prompt-driven extraction paradigms. Moreover, the successful application of LLM in NLP has motivated approaches that move beyond single-pass prompting toward more iterative and collaborative forms of processing.

The integration of LLMs into multi-agent systems (MAS) represents a further evolution in addressing complex NLP tasks, including event extraction (Zhang et al., 2026), (Guo et al., 2026), (Wang and Huang, 2024), question answering (Zong et al., 2024), summarization (Kim and Kim, 2025), (Celikyilmaz et al., 2018), fact checking (Lin et al., 2025), and scientific text generation. Within these architectures, multiple agents—each associated with specialized reasoning roles—collaborate through the exchange of intermediate representations, hypothesis generation and evaluation, and other structured steps in the analytical process. Recent studies, such as (Wang and Huang, 2024), suggest that such distributed reasoning mechanisms enable multi-agent systems to achieve deeper and more robust analyses compared to single-agent LLM models.

In our work, we also address **self-reflection** as a specific type of collaborative reasoning inside a MAS. During such reasoning, a system generates questions about missing or implicit information (in our case event-related) and answers them within the MAS itself. This system aspect is related to a growing body of work on formulating event extraction as question answering (Lu et al., 2025, 2023; Hong and Liu, 2024).

In this paper, we propose a multi-agent architecture specifically designed for **Semi-Open Event Extraction (SOEE)**.

In SOEE, a sub-set of event fields — such as event type, date, and location, specified by the system user — remains fixed in order to preserve cross-domain comparability. The remaining fields are defined dynamically at runtime, depending on the content of the article, the nature and type of the event, and other event-specific attributes that may be important but are not covered by the predefined schema. In our architecture, these flexible

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 72–82

attributes are inferred through a reflective reasoning process in which the LLM generates questions and then answers them. This process is implemented as an iterative dialogue: an initial event template is first constructed and then progressively refined and expanded through question generation and answering.

Our method is organized as a three-layer architecture: (1) an expert layer, which is responsible for constructing the initial event representation and answering the questions generated during reflection; (2) a reflective layer, which is responsible for generating natural-language questions aimed at identifying missing event information and expanding or refining the current event template; (3) a coordination layer, usually containing one manager agent who coordinates the activities of the other agents.

To demonstrate the feasibility of our proposed method, we developed a prototype called **MAREA** , **M**ulti-**A**gent **R**eflective architecture for **E**vent **A**nalysis.

The remainder of this paper is organized as follows. Section 2 introduces the concept of SOEE. Section 3 reviews related research on multi-agent reasoning, LLM-based, and open event extraction. Section 4 presents the design of the **MAREA** system, detailing the interactions between the coordination, reflective, and expert layers. Section 5 describes the evaluation of **MAREA** on a dataset of health-related news reports and discusses the observed improvements in template completeness and attribute discovery. Section 6 discusses the limitations of our approach. Section 7 presents an overview of the achieved results and future directions.

2 Semi-Open Event Extraction (SOEE)

A central challenge in event extraction is to balance the structural consistency of event templates with the richness of the information they encode.Traditional closed-domain extraction systems rely on ontologies and taxonomies, such as ACE or ERE (Song et al., 2015) and event templates with fixed structures such as the ones presented in (Grishman et al., 2002). Such an approach ensures consistency in the event template fields, but limits generalization to unseen event types. In contrast, open event extraction (OEE) approaches (Deng et al., 2022) offer domain flexibility but may potentially generate heterogeneous or inconsistent representations. To address this trade-off, we introduce the concept of SOEE, a hybrid form of event extraction that combines a fixed core schema with dynamically extensible attributes.

In SOEE, a sub-set of event fields — such as `event_type`, `date`, `location`, and several others — remains fixed to preserve cross-domain comparability, while other fields are contextually defined by the system at runtime. These flexible attributes are inferred through a reflective reasoning process, introduced in the previous section.

3 Related work

The research most closely related to our work falls into two main research lines: (1) the application of LLMs, agents, and multi-agent systems to event extraction; and (2) approaches to Open Event Extraction.

Recently, LLMs have proven to be effective tools for event detection and extraction (Meng et al., 2024; Tanev et al., 2025b; Wang et al., 2025). Beyond single-LLM approaches, recent work has explored multi-agent LLM architectures for event extraction to improve extraction quality. These include debate-based event template refinement (Wang and Huang, 2024), multi-agent generation-and-extraction for document-level event argument extraction (Zhang et al., 2026), and programming-style agent decomposition for zero-shot event extraction (Guo et al., 2026). All the works mentioned so far, however, assume a fixed event template. Our approach, on the other hand, is the first to experiment with a semi-open event schema, using a multi-agent LLM architecture.

Another line of research relevant to our work is Open Event Extraction (OEE), which does not assume a fixed event template or pre-defined event types. Instead, it typically extracts less structured event information from text without targeting specific event classes. The extracted information is often represented as tuples, such as (time, location, keyword set), or pairs, such as an event name and a date. OEE systems may also extract event triggers aimed at broad classes of events (Tong et al., 2020). Works such as (Deng et al., 2022) and (Wang et al., 2019) show that OEE can provide effective solutions for event annotation and extraction. In our implementation of SOEE, similarly to OEE, we enrich the event template with new, context-dependent event arguments, however our approach relies on a predefined core schema, which

increases its usefulness. Many event extraction related applications, like automatic event database filling, assume the presence of standard arguments and a fixed core event structure.

4 System Architecture

As already explained in the introduction, the **MAREA** event extraction architecture is made up of three layers: Expert, Reflection, and Coordination. The input of the event extraction process is a news article, while the output is a set of event templates (Aone and Ramos-Santacruz, 2000), structurally describing the events from the article.

- **Expert Layer**: This layer consists of two types of agents responsible for constructing the initial event representation and answering the questions generated during reflection:
    1. **A Template-Proposing Agent** generates the initial event template. It formulates the LLM prompts that specify the extraction task, define the initial template structure, and, when needed, provide few-shot examples consisting of source texts and their corresponding event annotations.
    2. **Answering Agent** responds to the questions generated by the Reflective Layer. It formulates prompts that guide the LLM toward locating and extracting the requested information from the source text. This agent also manages the interaction with the LLM and, in some cases, it may resolve specific sub-tasks without relying on the LLM.
- **Reflective Layer**: This layer is responsible for expanding the event template with contextually relevant out-of-schema fields and fill in their values. It is the core component that enables the semi-open character of the architecture, allowing the system to move beyond a fixed schema and introduce new, context-dependent event attributes. In MAREA, this layer contains two agents:
    1. A generic question formulating agenh uses thr,e different strategies, presented in 4.1, to generate natural-language questions aimed at identifying missing event information and expanding or refining the event template, created by the template-proposing agent.
    2. A spatial reasoning agent, which identifies the location names inside the input article using spaCy (Vasiliev, 2020) and then asks the LLM to identify the semantic role of each location.
- **Coordinating layer with a manager agent**: The manager agent in this layer coordinates the activities of the other agents, controls the execution flow of the extraction process, and ensures effective interaction with the user or with the software system in which the SOEE module is embedded.

The basic sequence of text processing stages is outlined below.

1. A news article is passed to the manager agent which forwards it to the template-proposing and to the reflective agents.
2. The template-proposing agent creates an initial set of event templates, one for each event within the article, by prompting the LLM using few-shot learning settings:
   The templates are created by prompting the LLM to produce event templates with a structure defined in the prompt. Several input and output examples are also provided in it. This prompt has the following structure:
   "You are an information extraction assistant, specialized in health-related events, such as [event_type_list]. Given the following news article, extract all distinct health-related events mentioned: [news_article]. Consider as an example this news article as sample input [sample_news_article]! Consider the following extracted event templates as a sample output: [sample_event_templates_in_JSON]".
3. After the initial templates are generated, they are passed to the reflective agent layer:
    (a) The question formulating agent generates a set of questions, e.g. "What measures are being taken to prevent the spread of COVID-19?", aimed at finding new event attributes and their values.
    (b) The spatial reasoning agent discovers the places mentioned inside the news article and their semantic functions for each extracted event ("place of infection", "event place", "hospital location", "responding

authority location", etc.) Each semantic role for a location constitutes a new candidate field for the event template.

4. The answering agent forwards the questions generated by the reflective agents to the LLM, using prompts that instruct the model to provide concise and precise answers. Then, it parses the LLM answer into a JSON structure.

In summary, **MAREA** first generates an initial event template through few-shot learning, followed by a reflection phase in which clarifying questions are formulated to identify potential event attributes beyond the predefined schema, thereby improving template completeness.

4.1 Question formulation strategies

Our system uses three question formulation or *reflective* strategies, that are implemented in the question formulation agent from the reflective layer:

1. Mapping event text to a predefined set of questions: We have trained a BERT model, given a sentence to map it to a sub-set of 29 frequent questions and corresponding fields from the domain of health events. Example of such a questions-field pair is ("Where did the infection occur?", "infection-place"). Due to lack of space, we cannot give the details of the BERT model training here. We used this BERT model to map all the sentences from the LLM-generated event summary to question-field sets and then we group them into one question set for each event.

2. Generic prompt-based question generation: We created a few-shot prompt for asking the LLM itself to suggest questions for expanding the event template, see Table 4. In this prompt, we give as parameters the article text and the event. The prompt asks the LLM to suggest questions which can discover new information and new event fields. The few-shot prompt itself was optimized using the MIPROv2 prompt optimization algorithm (Opsahl-Ong et al., 2024), implemented in the DSPy prompt programming and optimization framework. We used a small training set containing 5 tuples having the structure (article; event; relevant additional question set) to run the MIPROv2 otimizer with the **LLaMA-3.1-70b-instruct** model.

3. Keyword-based question generation: In this strategy, we first prompt the LLM to identify a set of keywords and key phrases for the event. Then, for each keyword or phrase, we prompt the LLM using the question: "How does keyword/phrase relate to the event?" The prompt also asks the LLM to propose a new template field name to accommodate the answer.

4.2 Application of MAREA to Health-Related Event Extraction

Within the **MAREA** architecture, only the template-proposing agent in the Expert layer is specialized for the health domain. Its domain knowledge comprises definitions of the core event fields and a sample news article annotated with two example events, used to support few-shot prompting. All remaining agents—including the question formulation and the spatial-reasoning components—are intentionally designed to remain domain-independent, enabling reuse across domains without architectural modification.

4.2.1 Event Template

As we mentioned, we adopt a SOEE strategy, in which a sub-set of the event template fields is fixed, following established approaches in health-related event extraction (Piskorski et al., 2023; Linge et al., 2012). This design preserves structural consistency while allowing the system to dynamically extend the template in the reflective reasoning phase.

For our experiments, we define a set of health-related event types grounded in prior research (Tanev et al., 2025b; Piskorski et al., 2023):

Outbreak: Sudden rise in disease cases; *Product_recall*: Withdrawal of unsafe medical or food products; *Study*: Research activities related to diseases, treatments, or vaccines; *Drug_approval*: Regulatory approval of drugs or vaccines; *Health_policy_change*: Changes to health-related laws or guidelines; *Disease_statistics_update*: Updated incidence or mortality data; *Biological_threat*: Bioterrorism incidents or emerging high-risk pathogens; *Pandemic_response*: Measures addressing large-scale epidemics; *Other_health*: Miscellaneous health-related events.

The fixed portion of the event template consists of the core fields shown in Table 1.

These event types and fixed template fields are explicitly provided to the LLM during prompting and serve as a structural basis for the initial event

Field name	Description
event_type	Category of the health-related event (e.g., outbreak, policy change, biological threat).
actors	People, organizations, or groups directly involved in or affected by the event.
description	Concise natural-language summary of the event.
event_text	Longer textual description preserving the main details and context of the event.
disease	Name of the disease or health condition associated with the event, if applicable.
biological_agent	Pathogen or biological agent responsible for the disease (e.g., virus, bacterium).
symptoms	Reported symptoms associated with the disease or health event.
where_place	Specific place or locality where the event occurred or was reported.
where_country	Country in which the event took place.
when_start	Date or time period marking the beginning of the event.
when_end	Date or time period marking the end of the event, if specified.
number_of_cases	Reported number of confirmed or suspected cases associated with the event.
number_of_deaths	Reported number of fatalities related to the event.
main_reason	Primary cause, trigger, or motivating factor underlying the event.
measures_taken	Actions, interventions, or policies implemented in response to the event.

Table 1: Core fields of the semi-open health event template and their semantics.

template generation.

5 Experiments and Evaluation

To evaluate the proposed approach, we compiled a corpus of health-related news articles from three sources: The first sub-setconsists of 300 articles collected from the *Europe Media Monitor* (Steinberger et al., 2013) over several months in 2020. The second sub-setincludes approximately 120 articles gathered from the Fox News Health RSS feed [1] over several weeks in September 2025. The third source is a random sample of 60 articles from March 2026, retrieved from the Medical Express RSS (section Infection Diseases) [2]. All articles were processed using **MAREA** , backed by the **LLaMA 3.1–70B–Instruct** large language model.

For the final evaluation, we selected 65 articles in random from the 451 articles corpus. **MAREA** extracted 70 events from these 65 articles.

[1] https://moxie.foxnews.com/google-publisher/health.xml

[2] https://medicalxpress.com/rss-feed/breaking/infectious-diseases-news

One expert evaluator has manually inspected the output of the **MAREA** system. The annotator has labeled all extracted field values as correct or not and additionally has indicated the missing values from each extracted event template. Additionally, the evaluator has searched for events which were mentioned but not extracted by the system.

5.1 Event detection

The evaluation showed that event detection achieved 100% precision and recall on the 65-article test set. All 70 events extracted by**MAREA**were judged to be relevant, and the expert annotator did not identify any additional events that were missed by the system. Event identity was determined by considering the event summary generated by the LLM, encoded in the ‘description‘ field, together with the other event attributes. Under this matching criterion, all system-generated events corresponded to manually identified events, yielding perfect event-level detection performance on this dataset.

5.2 Core fields extraction

Table 5 shows the precision, recall and F1 score for the accuracy of the core fields extraction, that is the fields predefined by the health event schema, defined in Table 1.

Performance (F1) is very high (over 0.85) for important event fields such as *event type*, *actors*, *disease*, *where-country*, and *measures taken*. The dataset did not contain enough data for evaluating the *number of deaths* field, since most of the articles were about clinical studies and disease statistics update, reporting only number of new cases. Fatalities were reported in only one of the 65 articles from the test set. Another significant field, *number of cases* was evaluated excluding the events of type STUDY for which the number of cases was not relevant. Instead, articles about these events typically report annual disease rates for the studied diseases, which in this case is ambiguous. Therefore, in order to ensure clarity in the evaluation process and to avoid ambiguity, we excluded the events classified as STUDY from our evaluation.

Notably, *event type* obtained a very high accuracy of 90%, demonstrating that our approach our approach achieves very strong performance in event classification.

Performance is lower, but remains satisfactory, for key spatial and temporal attributes: *where-country* and the event starting date in the *where-start* field. The *where-place* field for which the**MAREA**system has a relatively low performance, was intended to store the populated place name. Although the LLaMA model was instructed to do so, it has extracted instead locations such as names of universities and in some cases even countries. Even if these values can formally be considered to be partially correct, their level of granularity was different from the required one. Consequently, we measured low levels of precision and recall for this field. Another problematic field was found to be *when end* field, containing the end date of the event. The recall of extracting this attribute was notably low, as well as the precision. We have inspected the errors and found that this information is not always obviously stated and may even require temporal inference.

Altogether, the errors in our core fields extraction stem from incorrect interpretation of the prompt by the LLaMA model. For example, the **where place** mistakes are places which are correct as event locations, but do not correspond to the required geographic granularity. Similarly, the **biological agent** field often captured organisms, which were not disease vectors, as required by the prompt.

Taken together, these results indicate that MAREA, when powered by **LLaMA-3.1-70B-Instruct**, holds considerable promise for event extraction from real-world news data. Although the system shows lower performance on some core event attributes, largely due to incorrect prompt interpretation, our analysis suggests that improvements in post-processing and prompt design could substantially enhance overall extraction performance. It is also noteworthy that the event attribute extraction accuracy is comparable to that reported in previous work on a dataset from the same genre, namely medical news (Tanev et al., 2025a). Although the test corpora differ, our achieved accuracy is considerably higher than the baseline performance reported in that study. Thus, our work further supports the conclusion of the aforementioned study that LLMs can reliably perform event extraction and classification.

5.3 Additional fields added to the template

In order to assess the reflective capacity of our architecture and the question- and field-formulating strategies, described in section 4.1, we have randomly selected **31** event templates from 31 different articles from the test set. For these templates the reflective agent layer (both the question formulating and the spatial reasoning agents) has generated **147** additional event fields in total.

We have classified these new fields and their values into 4 relevance categories: (a) **Irrelevant**: Irrelevant information or incorrect field names or values; (b) **Low relevance**: formally correct field name and value; information is new, but it is not strongly connected to the event; (c) **Medium relevance**: relevant field name and a correct value; the field-value pair brings new information, but field name could be improved; (d) **High relevance**: relevant field name and a correct value, bringing new information, relevant to the event; (e) **Duplicate**: the field and value are correct, but they duplicate information already present in the template. Table 3 represents the distribution of the 147 event attribute-value pairs across these relevance categories. It is important also to note that the values of the fields are correct except few, classified as **Irrelevant**. Considering this, the evaluation of the attribute (field name) - value pairs is driven by the field name relevance.

It is noteworthy that 56% of the new generated fields have high or medium relevance (the last two rows in the Table 3). Highly relevant fields are completely correct as name and value, highly relevant to the event, and contain new information w.r.t. the other part of the template. These fields are 36% of all additionally-generated event fields. They are complemented by the medium-relevance ones, which are still correct and introduce relevant and innovative information, although their names could be improved.

If we exclude the 24% duplicates from the evaluation, the percentage of the medium and high relevance template fields becomes **77%**. Some interesting template-expanding questions and suggested new fields, generated by the question formulation agent are presented in Table 4.

5.4 Evaluation of the BERT question-generation module

The BERT-based question-proposing module suggests template-expansion questions in parallel with the prompt and keyword -based question-generation strategies. The module relies on a BERT model that maps each sentence to a set of 29 predefined questions, as described in Section 4.1.

We evaluated partially the accuracy of the BERT question-proposing module as follows: A proposed question and field was considered correct, if its answer (field value) could be found in the corresponding article; otherwise, it was considered incorrect. For this evaluation, we selected 68 articles from our corpus of 451 news articles. These articles were not used in the other evaluation experiments and all concerned infectious-disease outbreaks. This topic was chosen because outbreak-related articles typically contain information relevant to many of the 29 predefined questions used by the BERT module.

We ran the**MAREA**system on this dataset, extracting 68 events, one from each article, and recorded the questions suggested by the BERT module. Then, for each event–article pair, we manually identified which of the 29 predefined questions are answerable from the article text. Finally, we measured the overlap between the correctly suggested questions and the manually identified answerable questions. This allowed us to compute precision, recall, and F1-score for each event, as well as the overall micro-averaged precision, recall, and F1-score.

Table 2 summarizes the results. The precision, namely the proportion of answerable questions among all questions generated by BERT, is above 0.70. However, recall, which measures how well the module covers all relevant answerable questions, is considerably lower, at approximately 0.43.

The fact that more than half of the relevant questions are missed is compensated for by the other question-generation strategies. At the same time, the macro precision of 0.75 suggests that, when the module does propose a question, it is often relevant and answerable, providing evidence for the usefulness of this question-proposing strategy as a complementary component.

5.5 Missing information

Since the recall is not addressed in our evaluation, due to lack of annotated data, we have asked one evaluator to estimate how many facts are missing from each template from a sub-set of randomly selected **20 templates**. The evaluator considered the information inside each template and the corresponding article. He counted the number of missing facts from each template, where each fact could be positioned in one event field. Out of 20 event templates, the evaluator has found **8 missing facts**, which shows that on average our current implementation of **MAREA** in the health domain has a probability of missing a fact of **0.4**. In our test settings, this number can be plausibly interpreted as "in 40% of the cases our system may miss one important fact". This means, however, that most of the information is successfully captured in the output template, since the generated event templates in this test set have approximately 13 fields on average.

5.6 Evaluation overview

Overall,**MAREA**demonstrates strong core-field extraction and useful semi-open template expansion through reflective question generation, while the main remaining challenges concern prompt interpretation, field-name normalization, and incomplete question coverage.

6 Limitations

This study has several limitations. First, although**MAREA**performs well on many core event fields, some errors result from prompt misinterpretation by the underlying LLM, especially for fields requiring fine-grained spatial and temporal interpretation. Second, the dynamically generated fields are not always equally informative: some

Averaging	TP	FP	FN	Precision	Recall	F1
Macro	–	–	–	0.7468	0.4364	0.4713
Micro	283	110	363	0.7201	0.4381	0.5448

Table 2: Evaluation results for BERT suggested questions

Macro scores are computed as averages over records. Undefined per-record F1 scores were treated as 0. Micro scores are computed from pooled TP, FP, and FN across all records.

are duplicate, low-relevance, or require better field name. Finally, some evaluations were conducted on small manually inspected samples, and all evaluation experiments were carried out by a single annotator, which introduces a subjective bias and prevents us from estimating inter-annotator agreement. Broader testing across domains, event types, LLM models, and multiple annotators is therefore needed.

7 Conclusions

In this paper, we introduced **MAREA** , a reflective multi-agent architecture designed to perform SOEE. The proposed system models event extraction as a combination of an LLM-based few-shot learning template generation and an internal question–answer process in which specialized agents collaborate to enrich event representations. By combining these two approaches, **MAREA** is an attempt to address the trade-off between structural consistency and informational completeness in event extraction.

The experimental results in the health domain show that **MAREA** achieves consistently good performance across a wide range of core event fields, including event type, actors, disease information, countries, and public health measures. At the same time, the reflective agent successfully identifies and populates additional fields that provide complementary semantic context. The majority of the newly introduced fields are relevant and significant for the event template completeness. The probability to miss a fact for this approach was found to be 0.4, still this experiment was done on a very small scale. This number, however means that this approach tends to capture the larger part of the information in the generated event templates.

Future work will focus on three main directions. First, we plan to improve both the prompting strategy and the post-processing of generated templates. We will further optimize the prompts used by the template-proposing and answering agents, possibly using automatic prompt optimization methods such as MIPROv2. This may reduce errors caused by prompt misinterpretation. We also plan to introduce post-processing procedures for normalizing and filtering generated templates. These may include merging information from semantically similar fields, removing duplicate or low-informative attributes, and applying NLP-based filters to enforce field-specific constraints. For example, for the ‘where-place‘ field, named-entity recognition could be used to retain only entities identified as locations.

Second, we plan to conduct more extensive evaluations of the SOEE approach. This includes testing on benchmark data, when available for health-related news, providing second annotator evaluations, and experimenting with different LLM backends, such as various GPT models, open-weight models, and smaller locally installable LLMs. We also plan to carry out qualitative analyses of the system's behavior, including checking whether similar events across different articles, or different runs on the same article, lead to consistent outputs. In addition, we will investigate the causes of errors, such as prompt misinterpretation, inaccurate text interpretation, or possible LLM hallucinations, with particular attention to the reflective layer.

Third, we plan to exploit the most accurately extracted fields for the creation of synthetic or silver-standard event annotations. Fields such as ‘event type‘, ‘description‘, ‘actors‘, ‘where-country‘, ‘disease‘, ‘measures taken‘, and others which achieved high extraction accuracy in our evaluation, could be used as reliable anchors for automatically constructing larger weakly supervised corpora. Such silver-standard datasets could support further training, prompt optimization, and evaluation of event extraction systems in domains where manually annotated data are scarce.

Finally, producing new versions of **MAREA** for domains like security or disaster relief is an exciting research direction.

Field correctness and informativeness	Relative quantity
Irrelevant	0.05
Low	0.12
Medium	0.20
High	0.36
Duplicate	0.24

Table 3: Relevance of the fields added by reflection layer

Question	Suggested Field
What are the factors that could contribute to outbreak resurgence?	resurgence_factors
What is the expected timeline for implementing the new economic stimulus efforts?	implementation_timeline
What is the trend of new cases?	case_trend
What is the current severity of the outbreak?	outbreak_severity
Where did Robert O'Brien contract the viral infection?	infection_location
Who is criticizing the government's handling of the pandemic?	criticism_source
What journal published the findings of the study?	publication_venue
How does "flu Vaccine" relates to the event?	vaccine_type
What is the purpose of the early testing of the new vaccine candidate?	vaccine_purpose

Table 4: Questions and suggested template fields by reflection agent.

Field	Precision	Recall	F1-score
event_type	0.901	0.901	0.901
description	0.933	0.933	0.933
event_text	0.796	0.796	0.796
disease	0.922	0.870	0.895
biological-agent	0.600	0.600	0.600
symptoms	0.789	0.714	0.750
main-reason	0.864	0.704	0.776
actors	0.931	0.931	0.931
where-place	0.538	0.583	0.560
where-country	1.000	0.972	0.986
when-start	0.833	0.714	0.769
when-end	0.500	0.429	0.462
number_of_cases*	0.714	0.714	0.714
measures taken	0.960	0.923	0.941

Table 5: Core fields extraction accuracy.
*Computed after excluding STUDY events, for which number_of_cases is not applicable.

References

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Sixth applied natural language processing conference*, pages 76–83.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, and 1 others. 2022. 2event: Benchmarking open event extraction with a large-scale chinese title dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6511–6524.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002. Information extraction for enhanced access

to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246.

Quanjiang Guo, Sijie Wang, Jinchuan Zhang, Ben Zhang, Zhao Kang, Ling Tian, and Ke Yan. 2026. Extracting events like code: A multi-agent programming framework for zero-shot event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30880–30887.

Zijin Hong and Jian Liu. 2024. Towards better question generation in QA-based event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9025–9038, Bangkok, Thailand. Association for Computational Linguistics.

Hyuntak Kim and Byung-Hak Kim. 2025. NexusSum: Hierarchical LLM agents for long-form narrative summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10120–10157, Vienna, Austria. Association for Computational Linguistics.

Bobo Li, Xudong Han, Jiang Liu, Yuzhe Ding, Liqiang Jing, Zhaoqi Zhang, Jinheng Li, Xinya Du, Fei Li, Meishan Zhang, and 1 others. 2025. Event extraction in large language model:a holistic survey of method, modality, and future. *arXiv preprint arXiv:2512.19537*.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. 2025. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 360–381.

Jens P Linge, Marco Verile, Hristo Tanev, Vanni Zavarella, Flavio Fuart, and Erik van der Goot. 2012. Media monitoring of public health threats with medisys. *C. WILLIAM, CWR. WEB-STER, D. BALAHUR, et al*, pages 17–31.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688, Toronto, Canada. Association for Computational Linguistics.

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, and Xuan Wang. 2025. Crossagentie: Cross-type and cross-task multi-agent llm collaboration for zero-shot information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Zihao Meng, Tao Liu, Heng Zhang, Kai Feng, and Peng Zhao. 2024. Cean: Contrastive event aggregation network with llm-based augmentation for event extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 321–333.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366.

Jakub Piskorski, Nicolas Stefanovitch, Brian Doherty, Jens P Linge, Sopho Kharazi, Jas Mantero, Guillaume Jacquet, Alessio Spadaro, Giulia Teodori, and 1 others. 2023. Multi-label infectious disease news event corpus. In *Text2Story@ ECIR*, pages 171–183.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation*, pages 89–98.

Ralf Steinberger, Bruno Pouliquen, and Erik Van der Goot. 2013. An introduction to the europe media monitor family of applications. *Information Access in a Multilingual World*.

Hristo Tanev, Nicolas Stefanovitch, Tomáš Harmatha, and Diana F. Sousa. 2025a. Exploring the performance of large language models for event detection and extraction in the health domain. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1237–1247, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Hristo Tanev, Nicolas Stefanovitch, Tomáš Harmatha, and Diana F Sousa. 2025b. Exploring the performance of large language models for event detection and extraction in the health domain. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain event trigger knowledge. Association for Computational Linguistics.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press.

Rui Wang, Jiaoli Liu, Yu Yan, Liwei Zang, Huimin Wang, and Jianyi Liu. 2025. Document-level event extraction framework based on prompt learning. In *International Conference on Computer Application and Information Security (ICCAIS 2024)*, volume 13562, pages 996–1001. SPIE.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Open event extraction from online text using a generative adversarial network. *arXiv preprint arXiv:1908.09246*.

Sijia Wang and Lifu Huang. 2024. Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16422–16435, Miami, Florida, USA. Association for Computational Linguistics.

Guangjun Zhang, Hu Zhang, Yazhou Han, Yue Fan, Yuhang Shao, Hongye Tan, and Ru Li. 2026. Learning to generate and extract: A multi-agent collaboration framework for zero-shot document-level event arguments extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34665–34673.

Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Yongfeng Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1710, Miami, Florida, USA. Association for Computational Linguistics.