Exploring the Performance of Large Language Models for Event Detection and Extraction in the Health Domain

Hristo Tanev
Joint Research Centre
European Commission
hristo.tanev
@ec.europa.eu

Nicolas Stefanovitch
Joint Research Centre
European Commission
nicolas.stefanovitch
@ec.europa.eu

Tomáš Harmatha
UniSystems
Ispra, Italy

Diana F. Sousa
Joint Research Centre
European Commission
diana.francisco-de-sousa
@ec.europa.eu

Abstract

In this paper, we present an experiment evaluating several state-of-the-art open Large Language Models (LLMs) for the task of zero-shot event detection and event metadata extraction in the health domain against GLiNER, a lightweight zero-shot classifier, and state-of-the-art rule-based systems. For evaluation, we used a set of 854 health-related news articles, containing the title and lead sentences. We manually annotated them for the type of event they contained and its arguments (e.g., number of cases, victims, and animal cases) for a coarse typology of events. Additionally, we used as silver dataset the fine-grained annotations produced by the pandemics event classifier of Piskorski et al. (2023). Using this dataset, we conducted additional experiments on the capacity of models to suggest new labels and the position of event-label carrying sentences in the abstract of an article, comparing the results obtained when processing the article directly and per-sentence.

1 Introduction

Large Language Models (LLMs) have entered the Natural Language Processing (NLP) field at remarkable speed (Qin et al., 2024), demonstrating high efficiency across a variety of NLP tasks, including machine translation (Kocmi and Federmann, 2023), summarization (Wang et al., 2023), named entity recognition (Tan et al., 2023; Ye et al., 2024), and sentiment analysis (Kabaev et al., 2023). In addition to these core NLP tasks, LLMs have gained attention for their ability to handle complex intellectual challenges, such as passing standardized exams (OpenAI, 2023), generating human-like text (Science, 2025), solving mathematical problems (Trinh et al., 2024), assisting in programming (Guardian, 2025), and even producing creative content, such as, co-authoring essays and stories (Vara, 2025).

Despite these advances, the application of LLMs to socio-political event extraction remains relatively underexplored. Only recently a few studies begun to investigate the potential of LLMs at various stages of event extraction pipelines. For instance, Raiyaan et al. (2024) introduced *Political-RAG*, a framework combining Retrieval-Augmented Generation (RAG) with LLMs to improve political information extraction from media content.

Furthermore, the open LLMs Llama (Grattafiori et al., 2024) and Mistral (Jiang et al., 2023) play an increasingly important role in research due to their accessibility and performances. Nevertheless, studies assessing the performance of open LLMs in specialized NLP tasks like event extraction remain scarce.

To address this gap, we conducted an experiment focused on open LLMs-based event detection and extraction from news articles in the health domain. This includes disease outbreak reports, vaccination campaign news, and other articles discussing events affecting public health. Specifically, we used four of the open LLMs provided by the GPT@JRC project (De Longueville et al., 2025), which is an in-house API that provides privacy-friendly and effective access to both open and commercial models. The models evaluated were: LLama-3.1-70B-Instruct (Grattafiori et al., 2024), Mistral-7B Open Orca (Jiang et al., 2023), Zephyr-7B-Beta (Tunstall et al., 2023), Nous-Hermes-2-Mixtral-8x7B-DPO (Nous Research, 2024). We also evaluated Gliner-Multitask-Large-v0.5 (Stepanov and Shtopko, 2024), which is not an LLM, but a lightweigth zero-shot classifier that claims to have performance on par with LLMs.

In our experiment, we prompted each LLM to identify event mentions in sentences extracted from the first lines (snippet) of health-related news articles. Specifically, the models were instructed to

Proceedings of Recent Advances in Natural Language Processing,pages 1237–1247
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-143

detect events that pose a threat to human health. To ensure input relevance, we first applied an XLM-RoBERTa-based text classifier trained on an infectious disease news corpus and taxonomy (Piskorski et al., 2023) to filter health-related articles and then to perform a more detailed analysis of the behavior of the classifier.

Our contributions are the following:

- The primary objective of the study was to evaluate the accuracy of open LLMs against state-of-the-art rule-based systems in three key areas: event detection, argument extraction, and event location identification;
- Additionally, we also evaluated the capacity of open LLMs to propose new categories when the pre-existing one did not fit the data well;
- We evaluated the potential and limitations of GLiNER, a lightweight zero-shot model, for both coarse-grained and fine-grained classification;
- Furthermore, we compared the classification of articles snippets with the classification of each sentence in the snippet to better assess the location of event information.

We compared the output of the open LLMs with a baseline provided by two knowledge-based event extraction systems, NEXUS (Hristo Tanev et al., 2008) and Medical NEXUS (Linge et al., 2011), which have been running as an integral part of the Europe Media Monitoring (EMM) platform (Steinberger et al., 2013). These systems were compared at sentence level with a gold standard obtained by manually annotating 854 articles. A silver standard for fine-grained event classification was also created using the pandemics event classifier.

The performance of the LLMs surpassed all the other systems for event metadata extraction. While for event detection they all performed similarly on coarse-grained event detection, for fine-grained zero-shot event classification only some of the LLM achieved acceptable performances. The most significant improvement was observed in extracting the number of human fatalities: LLama achieved an F1 score of 0.84, compared to 0.64 F1 for NEXUS. This model also showed the highest performance for most of the tasks.

2 Related work

Event extraction is an emerging domain of LLMs applications. Currently, event extraction has been approached by machine learning techniques for text classification (Nguyen et al., 2016), sequence labeling (Chen et al., 2015), cascaded grammars for semantic argument extraction and event detection (Hristo Tanev et al., 2008), as well as transformers for full structure generation (Chen et al., 2015).

In this landscape, enhancing the event extraction process via LLMs is a new trend. Several recent research works have addressed this topic:

In Gao et al. (2024) and Zhu et al. (2024) LLMs are used to assess and correct the output of an event extraction algorithm. In the first case, they use reinforcement learning, and in the second, it is an automatic correction of the event extractor output. Researchers have also explored how event schemes can be harvested from LLMs without manual annotations (Tang et al., 2023). Schema-aware approaches using LLMs have also been explored in Shiri et al. (2024).

Another rapidly growing area of research shows that LLMs based data augmentation can boost event extraction performance by synthesizing additional training examples, especially for low resource cases or long-tailed distributions. Several recent works represent this trend. In some experiments, such as by Cartier and Tanev (2024), the LLMs have been instructed to generate training examples from sample sentences containing specific event types. In contrast, Meng et al. (2024) asked LLMs to create paraphrases of the training texts, based on a set of paraphrase patterns.

LLMs have been used not only for event detection and extraction, but also for tracking relations between events. For example, Hu et al. (2025) uses rationales generated by an LLM to extract event relations. LLMs can also be use to build knowledge graph and reason on them which has recently been studied in the health domain by Consoli et al. (2025).

Our work examines zero shot learning for event extraction in the healthcare domain. Previous studies have explored aspects of zero shot LLM-based event annotation. For instance, Chen et al. (2024) investigates zero-shot event argument extraction and the generation of novel event-containing sentences, although these experiments are constrained to the ACE event types (Doddington et al., 2004). The work presented in Consoli et al. (2024) ap-

plies zero-shot LLM-based extraction of epidemic events. Our work focuses specifically on health-related events and targets a broader and more comprehensive set of event arguments and event types than these previous studies. Additionally, our study emphasizes the use of open LLMs.

Zero-shot event extraction approaches outside the LLM context are also related to our work. These approaches are mainly based on transfer learning. Lyu et al. (2021) present the event extraction as a chain of other tasks, such as, question answering and entailment. In another line of work, Huang et al. (2018) and Zhang et al. (2021) train a semantic representation model on mentions of "seen" event types and encode "unseen" event type mentions with the same model.

Finally, our work is linked to research in event extraction for detection of disease outbreaks. One of the early examples of a system for disease outbreak detection has been described in Grishman et al. (2002). Other event extraction systems for the same domain have been presented in Lejeune et al. (2015), Abbood et al. (2020), Fisichella et al. (2010), and Linge et al. (2011).

3 Approach

Our event extraction approach has four main steps:

1. **Filtering**: We use a pre-trained XLM-RoBERTa-based classifier (Piskorski et al., 2023) to select only those news abstracts that are relevant to infectious disease or other health-related events. This step helps reduce noise and focus the event extraction process on meaningful content.

2. **Splitting**: Firstly, a news snippet is extracted from the full text, in order to obtain on average about 1024 characters while respecting sentences boundaries. Secondly, snippets are cleaned of noise artifacts and split into sentences.

3. **Prompting the LLM**: Each Open LLM, used in our experiments (Meta LLama-3.1-70B-Instruct, Mistral-7B Open Orca, Zephyr-7B-Beta, Nous-Hermes-2-Mixtral-8x7B-DPO) was prompted with a structured, carefully designed prompt as shown in Figure 1. The prompt asks the LLMs to extract the event name, country name and geographical coordinates, as well as the number of infected people and animals, and human fatalities, for each sentence.

4. **Parsing**: The output of each LLMs is then parsed and an event template is filled with the above mentioned information.

4 Experiments and Evaluation

4.1 Data

Our experiments were carried out on 854 news items, for which we considered separately the snippet, containing the title and the lead sentences part of it. We manually annotated 2160 snippets elements. The news articles were filtered automatically and only health related news have been selected. For this purpose we have used the pandemic event classifier of Piskorski et al. (2023), in order to select only articles that have a label, including the labels "miscellaneous" and "other". These articles covered the year 2024 and were downloaded from the Europe Media Monitoring platform (Steinberger et al., 2013). Then, 854 articles were randomly selected, after which were split in 2160 sentences. These sentences were then manually annotated with event information by two annotators and one curator. It is important to note that even if the classifier detects a health related event in an article, the individual sentences inside it do not necessarily contain such an event.

We focused on the lead sentences of each article, based on the well-established journalistic principle of the inverted pyramid reporting style (Pöttker, 2003), that the most important information is typically presented at the beginning of a news article. Considering the snippets and the sentences separately allowed us to do a more precise focus-based evaluation of the performance, and study how the event information was spread across sentences.

We have annotated each sentence for the presence of a health-related or other type of event which endangers the life of people, such as accidents and disasters. In each sentence we also annotated the place, the country, the number of human and animal cases, as well as the number of human fatalities, as they are mentioned in the same sentence. In more detail, we annotated each sentence with the following information:

- Event flag: 0 if there are no events, 1 if it is an disease-related event, 2 if it is an health related, but not a disease-related event and 3

You are an experienced data analyst specialising in extracting relevant information about events from health-related news articles. From text below (which is UTF-8 encoded and can be in any language), delimited by triple backticks, kindly extract all identifiable events and for each one, please extract following items:
1 - Name of the event that happened. Summarise to max. 7 words. Translate to English.
2 - Name of the country where this event happened, if present. Do not invent. Translate to English. In front of the country name, prepend ISO 3166-1 alpha-2 code of the country enclosed in square brackets and delimited with a space.
3 - Name of the place where this event happened, if present. Do not invent. Translate to English. Try to identify as precise place name as possible, down to official settlement name.
4 - Geographical location. For the information in previous point (3), try and geo-locate, i.e. identify longitude and latitude in WGS84 (EPSG: 4326) coordinate reference system. Present this element as a point in WKT (Well Known Text) format.
5 - Date when this event happened, if present. Do not invent. Show the date in YYYY-mm-dd format.
6 - Number of cases in humans (i.e. afflicted persons) derived exclusively from the particular event mentioned in the text, if present. Do not invent. Absolutely always summarise to one integer number only, no text.
7 - Number of fatalities among humans (i.e. persons dead) caused exclusively by the event mentioned in the text, if present. Do not invent. Absolutely always summarise to one integer number only, no text.
8 - Number of cases among animals caused exclusively by the event mentioned in the text, if present. Do not invent. Absolutely always summarise to one integer number only, no text.
9 - Category. One or more of the capitalised labels from the following taxonomy (inside ¡taxonomy¿¡/taxonomy¿ tag), best describing the event. You cannot introduce new labels, always choose one of these provided. If unsure, choose MISCELLANEOUS-OTHER.
10 - Category suggestion. Try and suggest a new category label which would best represent the content of the text you are analysing. Use a form consistent with the other labels, i.e. this format: "COARSE-FINE" where "COARSE" is the main-level category and "FINE" is a sub-category of the main-level one.
The taxonomy is a list of categories where a ⟨category ⟩always contains a pair of ⟨label⟩and ⟨explanation⟩elements, as follows: {Path('taxonomy-pandemic.txt').read_text()}
Format your response as an array of JSON objects with the following keys and only those keys (under no circumstance can you introduce other keys):
event_name, country, place_name, geo_point, date, cases_human, fatalities_human, cases_animal, category, category_suggestion.
If the information is not present, do not invent and instead use "None" as the value (unquoted).
Express cases and fatalities as integer numbers. Write no explanations nor notes and do not repeat the wider context back to me.
If you cannot extract any of the requested data, still include the JSON structure with all fields' values set to None.
Please order the array of results in the decreasing order of importance where importance is determined by the overall impact of the event on the health of the subjects (humans, animals) and area of impact (global > nation-wide > regional > local).
Please double-check your response so that you output only valid JSON array of JSON objects and nothing else.

Figure 1: Prompt for Extracting Events and Geolocation Data from News Articles

for any crisis event, not health-related such as natural disaster, transport accident, etc.

- Place name, the annotator used the same guidelines for the place name as the guidelines given to the LLM (Figure 1, point 3).
- Country name, where the event happened, using the two-letter country codes according to the ISO 3166-1 alpha-2 standard. This is the same format that the LLMs were instructed to use for country names (Figure 1, point 2).
- Number of human fatalities.
- Number of humans cases.
- Number of animals cases.

4.2 Knowledge-based Baseline

As a baseline we have considered a joint run of two knowledge-based systems, namely NEXUS (Hristo Tanev et al., 2008) and Medical NEXUS (Linge et al., 2011). Both systems have been pivotal in structured event extraction, particularly in crisis monitoring scenarios. Their rule-based architectures, leveraging cascaded grammars and domain specific semantic dictionaries, ensure high precision in identifying predefined event types. For instance, NEXUS is specialized in detecting sociopolitical events such as armed conflicts and protests, while Medical NEXUS is specialized in identifying disease outbreaks and reporting related statistics. Their deterministic nature ensures consistent outputs, which is crucial in the real-world environments in which they were tested, such as the situation rooms of some international organizations.

However, the rigidity of rule-based systems like NEXUS and Medical NEXUS can be a double-edged sword. While they offer consistency, they often lack the flexibility to adapt to novel or evolving event types without manual rule updates. In contrast, LLM-based approaches bring a level of adaptability and contextual understanding that rule-based systems struggle to achieve.

We selected NEXUS and Medical NEXUS as baseline systems for comparison with Large Language Models (LLM) due to their established performance in rule-based event extraction within crisis monitoring contexts.

We have used NEXUS as the baseline, and combined it with Medical NEXUS which is the only version of NEXUS able to extract human cases.

4.3 GLiNER Baseline

We use GLiNER (Stepanov and Shtopko, 2024), a zero-shot lightweight model that can be used for different information extraction tasks. The model uses an open label set, and it is able to extract the relevant corresponding span in the text. GLiNER lacks in generalization but is fast, and outputs performances comparable to LLM. In practice, the specific naming of labels must have a concise wording for better performance. GliNER, sometimes requires to use several multi-word expression to capture an idea which must then mapped to the one intended generic label. This is what did it the event codes, using this mapping: 'epidemic outbreak': 1, 'health disaster': 2, 'natural disaster': 3, 'transport disaster': 3, 'industrial disaster': 3, 'man made disaster': 3.

We specifically used a larger more general purpose variant of GliNER, the model: `gliner-multitask-large-v0.5`.

We used three more sets of labels: *label_slots* (country, location name, human fatalities, human cases, animal cases, and date), *labels_geo* (country, region, city, and location name) and pandemic events type which has been adapted by discarding some labels (others and miscellaneous) that can not be captured by GLiNER, and by rewriting some of the labels to make them better suited (e.g. "communication-meeting", becomes "meeting"): communication instrument, meeting, event cancellation, people displacement, impact on economy, impact on health system, authority regulation recommendation, facility closure, travel restriction, vaccine or medicine rollout, reporting cases, reporting situation, research funding, research progress, research phenomena, financial support, supply chain or provision, fake product or fraud, misinformation, and restriction violation or unrest.

For geo-location, we used only the labels "country" and use any of the labels selecting the most specific one. GLiNER has issues correctly detecting the word endings, as such it was necessary to find a workaround to have a fair evaluation.

4.4 Evaluation

We have evaluated the performance of all the open LLMs discussed so far and the joint NEXUS run, as well as the GLiNER model. Evaluation had two stages: (1) Evaluation of the accuracy of each model to extract event arguments and spatial parameters. (2) Accuracy of detecting health related events.

GLiNER is unable to produce ISO codes for countries. As such, it was necessary to create a mapping the raw GLiNER output to the ISO code, which was done using the Llama model. Given that other models were able to detect different countries, and reported in different formats, these were also harmonized into the same format.

4.4.1 Event Argument Extraction

We evaluated the systems and the LLM only on sentences which have been annotated as containing disease reports and health related events (event flag 1 or 2). For each sentence we have compared the annotated values with the output of each model and the baseline system for the fields place name, country code, number of human and animal cases, and human fatalities. Table 1 shows the F1 score for detecting the value for each field considered by each model and the NEXUS baseline.

Before checking if a value output from the model is the same as a manually annotated value, we performed several processing steps on the output of the LLMs. Our observations were based on experiments with a small data set, different from the test set.

- When a value of a numeric field (*number cases*, *number animal cases*, *number fatalities*) is an empty string, it is considered to be equivalent to the number "0".
- All LLM models consistently mismatched the ISO 3166-1 alpha-2 codes of the following countries: USA, Congo Brazzaville, Democratic Republic of Congo, and United Arab Emirates. For example, the Democratic Republic of Congo's code was often given the code "DCG", while the correct code was "CD", and the United Arab Emirates' correct 2 letter code "AE" was often substitute with "UAE". We have written procedures which corrected the output of the LLM for the "Country" field before matching it with the annotated value.
- Matching place name values turned out to be a challenging part of the evaluation, since place names can be written in different languages. Although we have explicitly instructed the LLM NOT to translate the names into English, in some cases the results were not according to

Table 1: Model performance on metadata extraction measured by F1 score

Model	Fatalities	Cases	Animal Cases	Country	Place	Cat.
llama-3.1-70b-instruct	0.9744	0.8365	0.9881	0.9447	0.7048	0.6655
mistral-7b-openorca	0.9608	0.7700	0.9319	0.8300	0.7169	0.2586
zephyr-7b-beta	0.9216	0.6780	0.9421	0.5700	0.5060	0.0814
nous-hermes-2-mixtral-8x7b-dpo	0.9625	0.7581	0.9693	0.9234	0.8373	N/A
gliner	0.8842	0.5724	0.9489	0.3320	0.3614	0.0579
nexus+mednexus	0.9199	0.6065	N/A	N/A	N/A	N/A

the instructions. Moreover, sometimes more than one place was mentioned in the event sentence and consequently these place names were put together in a list by the LLM. Annotators also annotated more than one place names on several occasions.

To improve matching in the presence of name variants, an automatic search was made for each place name in the OpenStreetMap database (Mooney et al., 2017). From there we have extracted for each place its name variants. We then identified intersections between the name variants of the annotated place and the variants of the place name proposed by the LLM.

4.4.2 Health Event Detection Evaluation

In the second stage of our evaluation, we assessed each model's accuracy in identifying articles that reported health-related events.

In this experiment we have aggregated the sentences which belong to one news article abstract and in case any of its sentences were annotated as containing a health-related event, we annotated the whole abstract as containing a health event.

Regarding the output of LLMs and the NEXUS, if an LLM had come up with an event name, event argument, or a location for any of the sentences in the abstract, the output of the LLM was considered to be positive for this abstract. The NEXUS baseline was considered to detect an event only when the Medical NEXUS detected number of cases or number of fatalities bigger than zero.

Considering this, we have calculated health event detection precision, recall and F1 score for each LLM model and the baseline. Results are shown in Table 2.

Both evaluation tables clearly show that all LLMs outperform the NEXUS baseline. The largest improvement was in the number of human cases detection 0.8365 vs. 0.6065 F1 score (0.23) by the LLama model. On the other hand, the difference in event detection between the baseline and all the models is not significant. All models and the NEXUS baseline demonstrate high recall (100% except Zephyr) and low precision, around 0.4, which shows that the models and the baseline both successfully detect health related events, but also erroneously detect other non-health related ones. The last fact may be caused by low performance of the pre-filtering module and the fact the prompt did not ask explicitly for health related, but "significant" events. The best performing model for event argument extraction and location detection was found to be LLama, while Zephyr showed the lowest performance for all fields, apart from "animal cases", still above the baseline.

Table 2: Precision (P), Recall (R), and F1 Score (F1) of the event detection task for different models

Model	P	R	F1
llama	0.4019	1.0000	0.5734
mistral	0.3971	1.0000	0.5685
zephyr	0.4168	0.9333	0.5762
nous-hermes	0.4115	1.0000	0.5830
gliner	0.4188	0.9606	**0.5833**
NEXUS	0.3966	1.0000	0.5680

4.4.3 Event Classification Evaluation

In Table 3 we report the classification performance of the different models measure as the weighted F1, comparing against the output of the pandemics event classifier. We compare the models for different focus: snippet and sentence level, and for different level of coarseness: fine- and coarse-grained. For GLiNER, the evaluation was performed only on the subset of labels retained, and these were mapped to the original taxonomy. The best model both for snippet and sentence level

model	focus	grain	P	R	F1
llama	SNI	fine	0.56	0.33	0.31
llama	SNI	coarse	0.71	0.59	**0.61**
mistral	SNI	fine	0.44	0.26	0.28
mistral	SNI	coarse	0.53	0.37	**0.39**
zephyr	SNI	fine	0.47	0.09	0.12
zephyr	SNI	coarse	0.53	0.19	0.25
gliner	SNI	fine	0.09	0.04	0.04
gliner	SNI	coarse	0.68	0.07	0.10
llama	SEN	fine	0.56	0.24	0.23
llama	SEN	coarse	0.65	0.64	**0.63**
mistral	SEN	fine	0.52	0.18	0.17
mistral	SEN	coarse	0.53	0.37	**0.39**
zephyr	SEN	fine	0.35	0.07	0.09
zephyr	SEN	coarse	0.55	0.17	0.22
gliner	SEN	fine	0.08	0.03	0.03
gliner	SEN	coarse	0.36	0.05	0.07

Table 3: Classification performance: precision, recall and weighted F1, measured against the pandemics event classifier output used as the ground truth

size	1	2	3	4	5	6	7
count	135	227	336	113	16	4	1
prop	.16	.27	.39	.13	.02	0	0

Table 4: Statistics on the number of sentence per snippet

were Llama for fine-grained and Mistral for coarse-grained. The performance of GLiNER was extremely poor, which is in contrast to the results for event-detection were it was the best model. This indicates that GLiNER is not a viable solution for domain-specific ad-hoc fine-grained categorization, but it is good for general purpose coarse-grained topic classification. We can also observe that GLiNER has at snippet level one of the best precisions, indicating that it lacks in generalization power, performance might therefore be improved by including more ad-hoc labels, which we will explore in future work.

index	1	2	3	4	5	6	7
count	741	634	436	123	20	4	1
prop	0.87	0.74	0.51	0.14	02	0	0

Table 5: Statistics on the sentence carrying an event category at a specific index

4.5 Event Category Suggestion

The LLMs were prompted to suggest category labels in case the existing taxonomy did not provide a good fit. This was done for two reasons: reduce the hallucinations when predicting the categories and to get an insight in which health-related topic are not covered well by the taxonomy. Mistral provided only three suggestions about cyberattacks and meteorological phenomenon. Zephyr provided a total of 624 labels, the 20 most common being about cybersecurity of water facilities, waterborne diseases, infrastructure failure, mass causalities, food security, and food safety. Llama provided 2825 suggestions, the 20 most common of which were about: animal to human transmission, food contamination, water contamination, natural disaster, clinical trials, animal health, and waterborne diseases. While these very specific labels reflect the small sample of article studied, it nevertheless shows different directions to expand the taxonomy for health-related events that are not directly linked to disease outbreaks. While Llama suggested the most suitable labels for this task, it also provided a long tail of suggestions that are extremely precise and not necessarily about event or health-related. This shows that LLM-based zero-shot classification is better used to propose new taxonomy rather than being applied directly on data.

4.6 Position of Event Information

We used the annotated sentence data to explore where event information is found in news article snippets.

In Table 4, we report statistics on the number of sentences in snippets and in Table 5 we report statistics on the index of sentence carrying event label information, in this last case we considered the classification made by the pandemics event classifier. We can see that 85% of the snippets are up to 3 sentences, and that 87% of the first sentences carry label information against 51% of the third sentences. This means that it is necessary to process the whole snippet and that the attention can not be restricted to the first sentence. In Table 6 we report the histogram of location statistics for coarse label of the pandemics event taxonomy. We can observe that the "miscellaneous" label is the one that has the most sentences; that for almost all the labels the first sentence is the most frequent one to carry the information, and that all the other labels have snippets under five sentences, with the

label	count	prop	list of (index,count)
COMMUNICATION	18	0.02	[(0, 5), (1, 7), (2, 6)]
RESEARCH	333	0.39	[(0, 156), (1, 105), (2, 51), (3, 20), (4, 1)]
MISCELLANEOUS	829	0.97	[(0, 257), (1, 249), (2, 223), (3, 80), (4, 15), (5, 4), (6, 1)]
REPORTING	479	0.56	[(0, 185), (1, 169), (2, 103), (3, 18), (4, 4)]
SUPPORT	30	0.04	[(0, 18), (1, 11), (2, 1)]
MEASURE	212	0.25	[(0, 95), (1, 75), (2, 38), (3, 4)]
VIOLATION	12	0.01	[(0, 5), (1, 5), (2, 2)]
IMPACT	46	0.05	[(0, 20), (1, 13), (2, 12), (3, 1)]

Table 6: Statistics of coarse grained labels of the pandemics event classifier index of the sentence in the snippet

last one very unlikely to convey event label information. For the label "reporting", which is the most critical and second most frequent label, the first three sentences convey 95% of the label information, as such this seems the optimal number of sentences to processes in article snippets for event extraction in the health domain.

Finally, we want to reflect on the pandemics event classifier: it did not generate a label for 7% of the sentences. Moreover, in only 20% of the cases the main label generated for snippet is strictly equal to the ones generated for the set of sentences, meaning that there is more information to obtain by analyzing at the level of sentence. However and more significantly, in 30% of the cases the main label of a snippet is not found in the labels of any of its constituent sentences. The most common of this labels are: "reporting case", "reporting situation" and "research phenomena finding", meaning that in order not to miss reporting of case, which is the most critical when monitoring news in the health domain, it is preferable to process articles headers as a whole. This effect could be due to event information being scattered across sentences, we however leave this investigation for future work.

5 Conclusions and Future Work

Our experimental analysis highlights the performance and limitations of open LLMs for zero-shot classification and traditional knowledge-based systems in the task of event extraction from health-related news articles.

The results indicate that LLMs, particularly the LLama-3.1-70b-instruct model, outperform NEXUS and Medical NEXUS state-of-the-art knowledge-based baselines in terms of F1 scores for extracting event arguments, showcasing their capability in handling complex and nuanced text extraction tasks. However, for event identification all the models, including keyword-based and GLiNER, have similar performances. The comparison of fine-grained and coarse-grained classification shows that models vary significantly in their ability to categorize events accurately given the precision of the classification task. The GLiNER model demonstrated the best performance in coarse-grained event detection and the worst performance in fine-grained event detection and event argument extraction. This study validates the performance of LLMs for event detection and extraction, however building an efficient event extraction pipeline requires to make use of the strength of each available solutions: LLMs are the slowest and most costly, followed by GLiNER and NEXUS, the fastest of all systems considered.

Our investigation into the placement of event information within news article snippets reveals that while the initial sentences are critical, significant information can be distributed throughout the snippet. This finding underscores the importance of processing entire snippets for effective event extraction, however it requires future works in order to assess how much exactly, notably taking into account that these could refer to different events. Moreover, the analysis of event category suggestions by the LLMs presents a promising avenue for expanding existing taxonomies to include more specific and relevant health-related categories, but shows that each model behave in very specific way, and that the best usage after identifying a model with good suggestions is to use it to refine a taxonomy.

Overall, the study demonstrates the potential of LLMs in improving event extraction accuracy in health-related news monitoring, while also identifying areas for improvement, particularly in refining prompts and enhancing precision in event detection. Future work could focus on refining LLM prompts to increase precision, exploring the integration of LLMs with traditional systems for optimized performance, and further expanding taxonomies to encompass a wider range of health-related events.

References

Auss Abbood, Alexander Ullrich, Rüdiger Busche, and Stéphane Ghozzi. 2020. Eventepi—a natural language processing framework for event-based surveillance. *PLoS computational biology*, 16(11):e1008277.

Emmanuel Cartier and Hristo Tanev. 2024. Event detection in the socio political domain. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 12–21, Torino, Italia. ELRA and ICCL.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Yubo Chen, Shulin Liu, Xiang Zhou, and Man Lan. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL-IJCNLP 2015*, page 167–176. Uses BIO sequence labelling for trigger and argument spans.

S. Consoli, P. Coletti, P. Markov, L. Orfei, I. Biazzo, L. Schuh, N. Stefanovitch, L. Bertolini, M. Ceresa, and N. I. Stilianakis. 2025. An epidemiological knowledge graph extracted from the world health organization's disease outbreak news. *Scientific Data*, 12:970.

Sergio Consoli, Peter Markov, Nikolaos I Stilianakis, Lorenzo Bertolini, Antonio Puertas Gallardo, and Mario Ceresa. 2024. Epidemic information extraction for event-based surveillance using large language models. In *International Congress on Information and Communication Technology*, pages 241–252. Springer Nature Singapore Singapore.

Bertrand De Longueville, Ignacio Sanchez, Snezha Kazakova, Stefano Luoni, Fabrizio Zaro, Kalliopi Daskalaki, and Marco Inchingolo. 2025. The proof is in the eating: Lessons learnt from one year of generative ai adoption in a science-for-policy organisation.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Marco Fisichella, Avaré Stewart, Kerstin Denecke, and Wolfgang Nejdl. 2010. Unsupervised public health event detection for epidemic intelligence. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1881–1884.

Jun Gao, Huan Zhao, Wei Wang, Changlong Yu, and Ruifeng Xu. 2024. Eventrl: Enhancing event extraction with outcome supervision for large language models. *arXiv preprint arXiv:2402.11430*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002. Real-time event extraction for infectious disease outbreaks. In *Proceedings of Human Language Technology Conference (HLT)*, pages 366–369.

The Guardian. 2025. Now you don't even need code to be a programmer. but you do still need expertise. *The Guardian*.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Proceedings of COLING 2008*, page 1129–1136. Describes NEXUS, a cascaded-grammar system combining pattern-based NER with event templates.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 7484–7496. Association for Computational Linguistics.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170.

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.

Anton Kabaev, Pavel Podberezko, Andrey Kaznacheev, and Sabina Abdullayeva. 2023. Half-masked model for named entity sentiment analysis.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143.

Jens P. Linge, Marco Verile, Hristo Tanev, Vanni Zavarella, Flavio Fuart, and Erik van der Goot. 2011. Media monitoring of public health threats with medisys. In *Living in Surveillance Societies*, Iaşi, Romania. Editura Universităţii Alexandru Ioan Cuza. Presented at the Living in Surveillance Societies conference, 2012.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. 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.

Peter Mooney, Marco Minghini, et al. 2017. A review of openstreetmap data. *Mapping and the citizen sensor*, pages 37–59.

Ngo Khai Cong Nguyen, Richárd Zsély, Grégoire Dupuy, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT 2016*, page 300–309. Treats each sentence as a classification instance to predict ACE event types.

Nous Research. 2024. Nous-hermes-2-mixtral-8x7b-dpo. `https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO`. Accessed: 2025-04-16.

OpenAI. 2023. Gpt-4 technical report.

Jakub Piskorski, Nicolas Stefanovitch, Jens P Linge, Sopho Kharazi, Jas Mantero, Guillaume Jacquet, Alessio Spadaro, and Giulia Teodori. 2023. Multilabel infectious disease news event corpus. In *Proceedings of the Text2Story'23 Workshop*, pages 171–183, Dublin, Republic of Ireland. Elsevier.

Horst Pöttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Muhammad Arslan Raiyaan, Saba Munawar, and Christophe Cruz. 2024. Political-rag: Using generative ai to extract political information from media content. *International Journal of Public Administration in the Digital Age*, 11(1):1–15.

Live Science. 2025. Gpt-4.5 is the first ai model to pass an authentic turing test, scientists say. *Live Science*.

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. *arXiv preprint arXiv:2406.01045*.

Ralf Steinberger, Bruno Pouliquen, and Erik Van der Goot. 2013. An introduction to the europe media monitor family of applications. *arXiv preprint arXiv:1309.5290*.

Ihor Stepanov and Mykhailo Shtopko. 2024. Gliner multi-task: Generalist lightweight model for various information extraction tasks. *arXiv preprint arXiv:2406.12925*.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, et al. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. *arXiv preprint arXiv:2305.03688*.

Jialong Tang, Hongyu Lin, Zhuoqun Li, Yaojie Lu, Xianpei Han, and Le Sun. 2023. Harvesting event schemas from large language models. In *China Conference on Knowledge Graph and Semantic Computing*, pages 57–69. Springer.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Vauhini Vara. 2025. Can a.i. writing be more than a gimmick? *The New Yorker*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

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*.

Hongming Zhang, Haoyu Wang, and Dan Roth. 2021. Zero-shot label-aware event trigger and argument classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1331–1340.

Mengna Zhu, Kaisheng Zeng, Jibing Wu, Lihua Liu, Hongbin Huang, Lei Hou, and Juanzi Li. 2024. Lc4ee: Llms as good corrector for event extraction.

In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12028–12038. Association for Computational Linguistics.