European Commission

Joint Research Centre

Recent Advancements on Artificial Intelligence for Public Health Threats

The Digital Health Perspectives

Consoli, S., Bertolini, L., Biazzo, I., Ceresa, M., Comte, V., Markov, P., Orfei, L., Ronco, M., Schuh, L., Stefanovitch, N., Stilianakis, N.

2026

Contact information
Name: Sergio Consoli
Address: JRC.F7 Digital Health, Via E. Fermi 2749, I - 21027 Ispra
Email: sergio.consoli@ec.europa.eu
Tel.: +39 0332-786584

Joint Research Centre
https://joint-research-centre.ec.europa.eu

JRC142475
EUR 40768

PDF ISBN 978-92-68-41024-0 ISSN 1831-9424 doi:10.2760/6953152 KJ-01-26-276-EN-N

Luxembourg: Publications Office of the European Union, 2026

How to cite this report: Consoli, S., Bertolini, L., Biazzo, I., Ceresa, M., Comte, V. et al., *Recent Advancements on Artificial Intelligence for Public Health Threats - The Digital Health Perspectives,* Publications Office of the European Union, Luxembourg, 2026, https://data.europa.eu/doi/10.2760/6953152, JRC142475.

Contents

Abstract . . . 3

Acknowledgements . . . 4

Executive summary . . . 6

1. Introduction . . . 9

    1.1. The Shift from Traditional to Generative Approaches . . . 10

    1.2. Related Work . . . 10

2. Epidemic Information Extraction for Event-Based Surveillance using Large Language Models . . . 13

    2.1. Data . . . 13
        2.1.1. ProMED . . . 14
        2.1.2. WHO Disease Outbreaks News . . . 14

    2.2. Methods . . . 14
        2.2.1. EpiTator . . . 15
        2.2.2. Pythia-12b . . . 15
        2.2.3. Mpt-30b-chat . . . 15
        2.2.4. Llama-2-70b-chat . . . 16
        2.2.5. Mistral-7b-openorca . . . 16
        2.2.6. Zephyr-7b-alpha . . . 16
        2.2.7. Gpt-3.5-turbo-16k . . . 16
        2.2.8. Gpt-4-32k . . . 17
        2.2.9. Gpt-4-FewShots . . . 17
        2.2.10. Open-Ensemble . . . 17

    2.3. Computational Experiments and Discussion . . . 18

3. Development of the Epidemiology Knowledge Graph . . . 23

    3.1. Methods . . . 24
        3.1.1. LLMs for Epidemic IE . . . 25
        3.1.2. eKG FAIR Publishing . . . 27
        3.1.3. eKG Services & Interfaces . . . 30

    3.2. Data Records . . . 30
        3.2.1. Data Statistics . . . 34

    3.3. Validation & Usage . . . 35

    3.4. Discussion . . . 41

4. Comparing In-Context Learning and Fine-Tuning Approaches . . . 43

    4.1. Methods . . . 44
        4.1.1. In-Context Learning . . . 44
        4.1.2. Fine-Tuning . . . 46

4.2. Results . . . 47
4.2.1. In-Context Learning Analysis . . . 47
4.2.2. Fine-Tuning Analysis . . . 51
4.3. Discussion . . . 52
5. AI-Generated Health Threat and Disaster Storylines from Global News with Large Language Models and Retrieval-Augmented Generation . . . 54
5.1. Data Sources for Retrieval . . . 55
5.2. Generation of Storylines and Graphs . . . 57
5.3. Data Records . . . 58
5.4. Validation & Usage . . . 60
6. Open issues and Critical Analysis on AI for Public Health: Challenges, Limitations, and Ethical Imperatives . . . 64
6.1. Technical and Data-Related Limitations . . . 64
6.2. Ethical and Societal Considerations . . . 65
6.3. Privacy, Security, and Regulatory Compliance . . . 66
7. Conclusions . . . 67
References . . . 68
List of Figures . . . 80
List of Tables . . . 82

Abstract

In recent years, the confluence of artificial intelligence and public health has paved the way for unprecedented advancements in epidemic intelligence and health threat surveillance. This technical report, titled "Recent Advancements on Artificial Intelligence for Public Health Threats - The Digital Health Perspectives", encapsulates the pioneering work conducted under the auspices of two projects of the JRC: DigLife (Digital Innovation in Life and Health Sciences) and ETOHA (Emerging Health Threats and the One Health Approach) in the JRC Digital Health Unit. These initiatives, in collaboration with other JRC units, DG SANTE, ECDC, WHO and DG HERA through recurrent touchpoints, embark on a journey that began in early 2024 to explore the transformative applications of AI in epidemic intelligence.

In particular, the report provides detailed insights of how generative AI, particularly Large Language Models (LLMs), is being leveraged to extract and analyse epidemiological information. At the forefront of this exploration is the innovative use of AI to harness unstructured epidemiological data from the Epidemic Intelligence from Open Sources (EIOS) system, which includes platforms such as ProMED and WHO Disease Outbreak News. This report delves into the methodologies and technologies developed to extract and structure critical epidemiological information, thereby enhancing the surveillance and response capabilities for health threats. The analysis confirms that generative AI holds immense potential to fundamentally support public health surveillance, enabling more timely, accurate, and proactive responses to disease outbreaks.

The report highlights core applications, including the use of LLMs for Epidemic Information Extraction, which has culminated in the creation of an epidemiology knowledge graph. A key methodological advancement is the use of an ensemble approach with multiple LLMs, which enhances reliability and mitigates the limitations of individual models. Through comparative analyses of Prompt Engineering approaches, including In-context Learning and Fine-tuning, we introduce some fine-tuned LLMs optimised for granular and rapid epidemic intelligence extraction from EIOS data. Furthermore, this report elucidates the integration of Retrieval Augmented Generation (RAG) techniques over epidemiological reports, exemplified by WHO Disease Outbreak News, to generate AI-driven health threat and disaster storylines. By leveraging AI to analyse, surveil, and respond to disease outbreaks, this report underscores the potential for digital health perspectives to innovate public health surveillance and response strategies. In addition, it provides several functional AI-based prototypes that underline the above-mentioned use cases, and can be further enhanced into more operational use.

Despite this high potential, the report also provides a critical analysis of significant challenges. These include the technical limitations of LLMs, such as the "black box" problem, a propensity for "hallucinations," and slow knowledge adaptation. Moreover, substantial ethical and societal challenges are identified, including data equity concerns that could reinforce health disparities, acute data privacy and surveillance risks, and the amplification of medical misinformation.

The report concludes that the future of generative AI in public health is a collaborative one, where technology serves as a powerful accelerator but does not displace ultimate human responsibility. A trustworthy future will require a "problem-first, technology-second" approach, prioritising ethical design, human oversight, and global collaboration to ensure the benefits of this technology are accessible and equitable for all.

Acknowledgements

We would like to thank the colleagues of the Digital Health Unit (JRC.F7) at the Joint Research Centre of the European Commission for helpful guidance and support. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission. This research did not receive any specific grant from funding agencies in the public, commercial, or nonprofit sectors. We would like to acknowledge the GPT@JRC initiative for providing access to the LLMs used in this study. We extend our gratitude also to the JRC Big Data Analytics Platform for providing secure access to a variety of open-source AI models. We would also like to extend our special gratitude to Gianfranco Spiteri, Laura Espinosa and Alexakis Leonidas from the Global Epidemic Intelligence and Health Security section at the European Centre for Disease Prevention and Control (ECDC) for their valuable support, fruitful collaboration, and helpful discussions throughout the development of this work. Finally, we would like to thank all the colleagues contributing to the Epidemic Intelligence from Open Sources (EIOS) initiative for the helpful suggestions and support during the development of this work.

Authors

Sergio Consoli
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Sergio.CONSOLI@ec.europa.eu

Lorenzo Bertolini
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Lorenzo.BERTOLINI@ec.europa.eu

Indaco Biazzo
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Indaco.BIAZZO@ec.europa.eu

Mario Ceresa
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Mario.CERESA@ec.europa.eu

Valentin Comte
European Commission, Joint Research Centre (JRC), 2440 Geel, Belgium
Valentin.COMTE@ec.europa.eu

Peter V. Markov
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Peter.MARKOV@ec.europa.eu

Lia Orfei
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Lia.ORFEI@ec.europa.eu

Michele Ronco
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Michele.RONCO@ec.europa.eu

Lea Schuh
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Lea.SCHUH@ec.europa.eu

Nicolas Stefanovitch
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Nicolas.STEFANOVITCH@ec.europa.eu

Nikolaos I. Stilianakis
European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy
Nikolaos.STILIANAKIS@ec.europa.eu

Executive summary

In an era where information is as abundant as it is critical, the confluence of artificial intelligence and public health marks a transformative journey toward smarter epidemic intelligence and health threat management. This report, "Recent Advancements on Artificial Intelligence for Public Health Threats - The Digital Health Perspectives", presents the exploratory and prototype-stage work conducted within the JRC projects DigLife (Digital Innovation in Life and Health Sciences) and ETOHA (Emerging Health Threats and the One Health Approach)[1], in collaboration with public health stakeholders and institutions including the ECDC, WHO, DG SANTE, and DG HERA, among others. The AI methods and prototypes described throughout the report are intended as research and innovation activities and have not yet been validated in operational public-health settings.

Since embarking on this venture in February 2024, our collective focus has been on leveraging AI to harness the untapped potential of unstructured epidemiological data from sources like the Epidemic Intelligence from Open Sources (EIOS) system. This initiative aims to elevate public health responses by extracting, structuring, and analysing critical information from platforms such as ProMED and WHO Disease Outbreak News.

Our exploration unfolds through several tracks. First, the deployment of Large Language Models (LLMs) for precise epidemic information extraction, leading to the creation of a dynamic epidemiology knowledge graph. The second is a detailed comparison of Prompt Engineering approaches, resulting in a customised set of fine-tuned LLMs for swift and detailed epidemic intelligence. Then, the application of Retrieval Augmented Generation (RAG) techniques over epidemiological reports creates AI-generated health threat narratives, enhancing situational awareness.

The culmination of these efforts are several functional AI-based prototypes, poised to redefine public health research and policy-making. By transforming how we analyse, surveil, and respond to health threats, we set the stage for a future where digital innovation drives resilience and preparedness in global health. This report not only documents these advancements but also inspires a vision for harnessing AI's full potential to safeguard public health.

Policy context

This technical report directly supports the EU's commitment to enhancing public health through digital innovation, focusing on the European Health Union's objectives to improve preparedness and response to cross-border health crises. The report underscores the role of AI in epidemic intelligence, enhancing the EU's capacity to manage health threats and inform policy frameworks. By providing several AI-based prototypes for public health surveillance, the report adds significant value to previous policy efforts, particularly in the context of emerging health threats and the need for rapid, data-driven responses. Its relevance extends to broader digital and health policies, promoting resilience and proactive measures across the EU.

Key conclusions

The integration of AI in health threats surveillance, preparedness and response, as explored in this report, has substantial implications for enhancing the European Union's health security and response strategies. The findings demonstrate AI's transformative potential in extracting valuable insights from unstructured

[1] While the ETOHA initiative adopts a broader One Health perspective encompassing human, animal, and environmental health, the AI applications and use cases described in this report primarily focus on human public health surveillance and epidemic intelligence.

epidemiological data, suggesting that its incorporation into the EU's health threat surveillance systems could significantly enhance early detection and response capabilities.

A key recommendation is the implementation of AI-driven tools for more timely and scalable epidemic intelligence workflows. Additionally, the report highlights the importance of promoting standardised data formats and interoperability among EU Member States, together with robust governance mechanisms for validation and bias mitigation, to ensure the safe integration of AI technologies in public health. These findings align with the EU's broader policy direction on leveraging digital innovation for health security, including initiatives such as the European Health Data Space and the AI Act. However, they also highlight the need for more robust, domain-specific AI governance frameworks in public health, particularly regarding validation, transparency, and safe deployment of AI systems in epidemiological surveillance.

The report identifies a new challenge: the underutilisation of unstructured data sources — such as news feeds and outbreak reports from ProMED and WHO Disease Outbreak News — in current surveillance systems. Addressing this issue is crucial for maximising the benefits that AI can offer. To this end, there is a call for increased investment in AI research focusing on public health surveillance applications, which would drive innovation and improve the EU's capacity to respond to emerging health threats.

The problem statement should be revised to emphasise AI's role in processing unconventional data sources. While the potential innovations in developing AI models tailored to specific epidemiological challenges are promising, they also present risks related to data privacy and algorithmic bias, which require careful management.

Although this report reduces uncertainties regarding AI's applicability in public health threats surveillance, preparedness and response, significant knowledge gaps persist, especially concerning the long-term impacts of AI integration in health systems. Continued research is necessary to address these uncertainties and optimise AI's contribution to public health policy.

In conclusion, the report advocates for a strategic inclusion of AI-driven public health surveillance measures, emphasising the need for policy adaptations to effectively harness technological advancements. By doing so, the EU can better prepare for, and respond to, health threats, ultimately enhancing public health outcomes.

Main findings

The report identifies several key findings in the application of AI to public health, particularly in epidemic intelligence. AI technologies, especially Large Language Models (LLMs), have shown remarkable effectiveness in extracting and structuring epidemiological information from unstructured sources. By harnessing data from platforms like the Epidemic Intelligence from Open Sources (EIOS), which includes ProMED and WHO Disease Outbreak News, among many other information sources, AI offers timely and actionable insights crucial for public health responses.

One of the main achievements is the development of an epidemiology knowledge graph. This tool integrates data from various sources, providing a comprehensive overview of disease outbreaks and supporting more informed decision-making processes. Additionally, the report highlights the successful creation of fine-tuned LLMs for optimised, granular and rapid epidemic information extraction from EIOS data, emerged from a comparative analysis of different prompt engineering approaches, demonstrating superior performance. The application of Retrieval Augmented Generation (RAG) over epidemiological reports represents another significant advancement. This technique improves the generation of detailed health threat narratives, thereby enhancing situational awareness and preparedness.

The findings underscore the transformative potential of integrating AI into health threat surveillance. Such integration can revolutionise the analysis and utilisation of public health surveillance data, leading to improved timely responses and more informed policy-making. By providing a robust evidence base, these enhanced surveillance capabilities may contribute to shaping public health threats policies and practices. In conclusion, the report advocates for the broader adoption of AI in public health surveillance to enhance preparedness and response to health threats, highlighting its pivotal role in transforming surveillance strategies.

Related and future JRC work

The Joint Research Centre will continue to explore AI advancements in health threat surveillance, focusing on integrating AI with more timely and scalable data systems. From an implementation perspective, a phased deployment approach may be the most feasible pathway. AI-based tools could first be piloted at regional or national public health agency level, where surveillance workflows and data access arrangements can be evaluated under operational conditions. Successful pilots could then inform broader adoption across Member States through common interoperability standards, governance mechanisms, and capacity-building initiatives coordinated at EU level. In this model, AI systems would complement rather than replace existing surveillance infrastructures, providing enhanced analysis of unstructured information sources while preserving established reporting and validation processes. Such pilot deployments would also provide valuable evidence on the organisational, technical, and resource requirements necessary for the sustainable integration of AI-based analysis of unstructured data into national and European public health surveillance workflows.

Quick guide

This report explores the use of artificial intelligence (AI) in enhancing health threats surveillance, preparedness and response through improved epidemic intelligence. Key terms include Large Language Models (LLMs), used for processing unstructured data, and the Epidemic Intelligence from Open Sources (EIOS), a vital data source. Assumptions include the effectiveness of AI in data processing and its potential for more timely and scalable applications. Uncertainties remain regarding long-term impacts and integration challenges in existing health systems.

1. Introduction

Generative AI represents a subfield of artificial intelligence focused on creating new, original content rather than merely analysing or classifying existing data [2]. It distinguishes itself from traditional AI by its ability to synthesise data and generate content such as text, images, video, and code [151, 47]. A key component of this paradigm is given by Large Language Models (LLMs) [36, 19]. LLMs are deep learning architectures, often built on the Transformer architecture, that are trained on vast text corpora to predict word sequences and, in doing so, develop a nuanced understanding of human language at scale [136, 19].

The proliferation of unstructured data—from clinical notes and medical records to social media posts and news reports—has created a significant challenge for traditional epidemiological surveillance. Generative AI offers a scalable solution to this problem, enabling public health professionals to process and derive meaningful insights from this overwhelming quantity of information [48, 15]. This capability marks a pivotal moment in epidemiology, promising to support how infectious disease outbreaks are tracked and managed. This report delves into the innovative application of AI to enhance our ability to monitor, understand, and respond to emerging health threats. Engagements with stakeholders, including the Joint Research Centre (JRC), the European Centre for Disease Prevention and Control (ECDC), and the World Health Organization (WHO), among others, have contributed to exploring the potential of AI to process unstructured epidemiological data from the Epidemic Intelligence from Open Sources (EIOS) system, encompassing platforms such as ProMED and WHO Disease Outbreak News. By transforming such data into structured outputs, AI has the potential to support public health response and surveillance activities. Beyond infectious disease surveillance, these capabilities may also support the detection and monitoring of emerging Chemical, Biological, Radiological and Nuclear (CBRN) threats, where traditional surveillance infrastructures are often limited and situational awareness relies heavily on the analysis of unstructured reports and open-source intelligence.

A key area of exploration is the use of Large Language Models (LLMs) for Epidemic Information Extraction. We present a novel approach to epidemic surveillance, leveraging the power of Artificial Intelligence and Large Language Models (LLMs) for effective interpretation of unstructured big data sources, like the popular ProMED and WHO Disease Outbreak News. We explore several LLMs, evaluating their capabilities in extracting valuable epidemic information. We further enhance the capabilities of the LLMs using in-context learning, and test the performance of an ensemble model incorporating multiple open-source LLMs. The findings indicate that LLMs can significantly enhance the accuracy and timeliness of epidemic modelling and forecasting, offering a promising tool for managing future pandemic events. These models, designed to sift through and interpret vast amounts of text-based data, are able to identify critical signals and trends that might otherwise go unnoticed. This capability is exemplified in the creation of an epidemiology knowledge graph, which synthesises diverse data inputs into a coherent framework, providing a holistic view of ongoing and potential health threats.

Building on this, the report introduces fine-tuned LLMs specifically developed for rapid and granular information extraction from EIOS data. These models emerged from a comparative analysis of prompt engineering approaches, demonstrating superior performance in handling complex epidemiological data. Their developments underscore the potential of AI to enhance the speed and accuracy of epidemic intelligence, facilitating faster and more informed decision-making.

Another innovative application explored in the report is Retrieval Augmented Generation (RAG). This approach leverages AI to generate detailed health threat narratives from epidemiological reports, such as those provided by WHO Disease Outbreak News. By augmenting data retrieval with narrative generation, RAG enhances situational awareness and preparedness, enabling public health officials to better anticipate and respond to emerging threats.

As the European Union continues to prioritise digital innovation in its health security strategies, this report provides a comprehensive overview of the opportunities and challenges associated with integrating AI into existing public health surveillance frameworks. By examining these recent advancements, the report underscores the necessity of strategic policy adaptations to fully leverage AI's potential, setting the stage for a future where digital health perspectives are integral to enhancing resilience and preparedness in public health systems.

1.1. The Shift from Traditional to Generative Approaches

Historically, public health and clinical data extraction have relied on either manual processes, which are slow and costly, or on traditional, rule-based computational methods [71]. For example, the Centers for Disease Control and Prevention (CDC) uses a dictionary-based approach for cancer surveillance that compares pathology reports against a predefined dictionary of terms [26]. This method provides good results but requires significant manual effort to maintain and update the dictionary to account for new terms and variations [26]. Similarly, traditional statistical NLP approaches, which use supervised machine learning, work best when a large volume of structured data is available for training [26].

In contrast, generative AI's core capability is its adaptability to dynamic inputs [129, 151]. Unlike traditional AI models that operate on explicit rules, generative models learn patterns and relationships from massive, unstructured datasets, allowing them to handle complex, nuanced tasks with a greater degree of flexibility [36, 15]. For instance, domain-specific LLMs can detect not only medical entities but also linguistic nuances like negations ("no sign of pneumonia") and uncertainty ("possibly viral origin"), which would be challenging for a rule-based system to capture without extensive manual intervention [20, 15].

This shift represents a fundamental methodological change in how public health surveillance is conducted. Traditional methods are often reactive and dependent on pre-structured data, which can lead to delays in analysis and response [116, 151]. Generative AI, as demonstrated by its application in processing dynamic sources like the World Health Organization's (WHO) Disease Outbreak News (DONs), can automate the continuous processing of new information to provide more timely and adaptable epidemic intelligence [20, 102]. This capability moves the function of public health surveillance from being a reactive, post-event reporting mechanism to a proactive, continuous intelligence system that can inform rapid response strategies and policy-making [152, 102]. In future surveillance architectures, information extracted from unstructured sources could complement traditional structured surveillance data in a convergent manner, providing early signals, contextual insights, and cross-validation mechanisms that enhance the timeliness and completeness of public health intelligence. Table 1 further delineates the key differences between these two approaches.

1.2. Related Work

The application of AI, particularly LLMs, in health threats surveillance and epidemiological research has been explored in various contexts, offering insights into the potential and challenges of leveraging these technologies for effective outbreak response and data extraction [15, 151]. The earlier research in [55] aims at improving epidemic response through innovative data sources [13], by introducing in the specific an internet-based computational model for timely surveillance of influenza using search engine data. Keller et al. [73] discussed automated surveillance methods for disease outbreaks using online data, focusing on geo-parsing and real-time monitoring. Espinosa et al. [49] evaluated an automated early warning tool using Twitter data, demonstrating its effective performance in detecting public health threats

compared to manual tweets monitoring. Building on these efforts, the work in [82] presents a multilingual event extraction system for epidemic detection, highlighting the importance of language diversity in surveillance efforts. Torii et al. [133] investigated instead the application of various machine learning classifiers for the automated detection of disease outbreak-related articles. Shifting from detection to prediction, Chowell et al. [28] introduced an ensemble model for forecasting epidemic trajectories, focusing on addressing data and model uncertainty. The exploration of cost-effective training methods and advanced modelling techniques adds valuable insights into the challenges of more timely and scalable epidemic monitoring. For more details on the various methods and data sources in the literature for disease outbreak detection and forecasting the reader is referred to the comprehensive review in [13].

Table 1. Traditional AI vs. Generative AI

Feature	Traditional AI/NLP	Generative AI
Core Capability	Analyzes data, makes predictions, automates rule-based tasks	Generates original content, synthesizes data, and serves as an assistant
Data Type	Relies on structured data and predefined algorithms	Cope with large, unstructured datasets; learns patterns and relationships
Adaptability	Rigid and rule-based; requires manual intervention for new scenarios	Highly adaptable to dynamic inputs; learns to learn
Implementation	Simpler to implement in specific business processes	More complex and resource-intensive due to large datasets and computational power requirements
Typical Use Cases	Automated registry reporting, predictive analytics, rule-based tasks	Information extraction from unstructured text, content synthesis, conversational assistance

Source: Elastic, 2024 [47].

Jingxian et al. [148] emphasised the importance of extracting information from unstructured data sources. In particular, they used text mining techniques to analyse reports from ProMED-mail [91], focusing on outbreak trends. More recently, the study by Chang et al. [27] introduced a method for summarising epidemic alerts from ProMED-mail using natural language processing techniques [102]. This work demonstrated the potential for automated methods to enhance the effectiveness of epidemic monitoring by improving the summarisation of epidemic alerts. The findings highlight the broader applicability of AI in optimising public health surveillance systems.

In recent years, there has been a significant shift towards the application of generative AI technologies in epidemic surveillance and response, demonstrating their potential to improve the timeliness and accuracy of information processing during public health emergencies [102]. For example, Kwok et al. [79] investigated the role of ChatGPT in developing disease transmission models, showcasing how LLMs can assist public health practitioners in shaping effective strategies. In [128], the authors performed a com-

parative analysis to evaluate the performance of various LLMs against traditional information extraction methods on symptom detection from clinical records in the context of COVID-19, providing insights into the strengths and limitations of LLMs in this area. Similarly, Bizel-Bizellot et al. [18] explored the use of generative AI models to extract records of COVID-19 transmission from free-text survey responses, highlighting a practical application of LLMs in epidemiological research. While these studies provide valuable information on the application of LLMs for COVID-19 symptom extraction, our work extends these explorations to a broader range of functionalities and diseases relevant to public health.

In [48], the authors investigated the effectiveness of LLMs in accurately extracting public opinions on vaccination from social media, highlighting their potential as scalable tools for public health surveillance. Du et al. [45] presented a framework that utilises multimodal LLMs for infectious disease forecasting, aligning closely with the focus of our work. The innovative approach proposed to reformulate forecasting as a text-reasoning problem complements the findings of the present chapter, offering additional insights into the application of LLMs in public health.

For an in-depth review on the exploration of diverse applications of natural language processing and large language models in infectious disease management, the reader is referred to [102]. This systematic review highlights areas where LLMs can further contribute to diagnosis and prediction, providing a broader context for their utility in epidemic surveillance.

2. Epidemic Information Extraction for Event-Based Surveillance using Large Language Models

The use of novel unstructured data sets in epidemic modelling has been increasingly encouraged, as they can significantly contribute to improving early warning systems [116]. The recent COVID-19 pandemic has underscored the importance of these novel approaches [93]. The pandemic revealed the long delays in the release of official statistics, emphasising the need to track pandemic information in alternative ways and at a higher frequency.

These opportunities and challenges in infectious disease epidemiology are inspiring some of the research activities at the Digital Health unit of the Joint Research Centre (JRC)[1] of the European Commission (EC). Ongoing work, described in this technical report, aims at tracking pandemic outbreaks in the European Union (EU) using data sets which are considered unconventional in classic epidemiological modelling. The project aims at exploring novel (big) data sources to provide a better response to epidemics. It links to other ongoing relevant initiatives on the subject[2]. An important source of such information is represented by moderated news and reports, since they discuss important events and experts opinions, and can substantively inform policy-making decisions [83]. However, translating this new source of data into valuable information is challenging, given that the data derived from these sources are often unstructured and large. This data also exhibits non-linear relationships across variables, which adds to the complexity of its interpretation. Despite these challenges, the potential benefits of utilising these unconventional data sources are considerable. By effectively translating this data into actionable insights, we can significantly improve our understanding of epidemics and, consequently, our response to them.

In this chapter, we aim to leverage the vast capabilities of Artificial Intelligence (AI) [20], specifically the power of Generative Pre-trained Large Language Models (LLMs) [136, 19], for the the exploration and effective interpretation of such innovative big data sources, with the ultimate goal of improving epidemic response. LLMs are a type of generative AI models that utilises deep-learning (DL) algorithms (involving billions of parameters) to calculate the likelihoods of word sequences. These probabilities are determined based on substantial text corpora that the model has previously learned from. Significant advancements to LLMs have been made with the advent of the Transformers architecture [136]. Transformers are designed to handle sequential data by using a mechanism called attention, which allows the model to weigh and prioritise different parts of the input data [126]. Unlike traditional sequential models such as Recurrent Neural Networks (RNNs), Transformers process all input data concurrently, which allows for more efficient computation and the ability to handle longer sequences.

Considering the rapid advancements in LLMs, this study aims to evaluate the most significant and recent open-source and commercial models for the specific task of epidemic information extraction. The use of the information extracted from large-scale, unstructured data sets in epidemic infectious disease, coupled with the integration to surveillance systems, can significantly enhance the accuracy and timeliness of epidemic outbreaks modelling and forecasting.

2.1. Data

In this section, we describe the unstructured data sets used in the study. Although we have focused on these two data sets, the approach is completely generalisable to any textual source concerning infectious disease epidemics.

[1] https://ec.europa.eu/info/departments/joint-research-centre_en

[2] WHO EIOS: Epidemic Intelligence from Open Sources, https://www.who.int/initiatives/eios

2.1.1. ProMED

ProMED[3], the Program for Monitoring Emerging Diseases, is an initiative by the International Society for Infectious Diseases (ISID). As the largest publicly accessible system for infectious disease outbreak reporting, it serves as a critical resource for healthcare professionals and the public alike. Its platform offers a wealth of daily posts detailing the latest developments in the field of infectious diseases. Launched in 1994, ProMED has been at the forefront of outbreak reporting, being the first to report on numerous disease outbreaks, including SARS, Chikungunya, Ebola, Zika, MERS, and most recently, COVID-19 [91]. Its users comprise international public health leaders, government officials, physicians, veterinarians, researchers, companies, journalists, and the general public worldwide. Reports are produced and commentary provided by a multidisciplinary global team of subject matter experts in a variety of fields, including virology, parasitology, epidemiology, and entomology. Since its inception, ProMED has generated so far more than 66,000 mail posts, from 1994-08-19 to date.

2.1.2. WHO Disease Outbreaks News

The World Health Organization (WHO)'s Disease Outbreak News (DONs)[4] serve as a vital source of information on confirmed acute public health threats or potential events of concern. The platform has been providing crucial updates since January 1996. The essence of DONs lies in its commitment to sharing news about confirmed or potential public health events. This includes incidents of unknown cause that carry significant or potential international health concerns and could affect international travel or trade. It also encompasses diseases of known cause which have shown the capacity to cause a serious public health impact and spread internationally. Since its inception, DONs has published over 3,000 pieces of curated epidemic news, from 1996-01-22 to date. It is worth noting, however, that DONs does not provide exhaustive coverage of all epidemic events globally; reporting is contingent on national authorities formally notifying the WHO, which may result in underreporting of outbreaks in regions with limited surveillance infrastructure or weaker institutional reporting mechanisms. This inherent incompleteness should be taken into account when interpreting findings derived from this data source.

2.2. Methods

In this section, we describe the epidemic information extraction models considered in our study. These consist in the most popular and recent LLMs, spanning from open-source to commercial ones, along with a significant semantic method from the literature for epidemic information extraction, namely EpiTator. Please note that the LLMs employed in this study have been used through the GPT@JRC initiative of the Joint Research Centre (JRC) of the European Commission, which enables JRC staff to explore the potential uses of AI pre-trained LLMs. The initiative, which is part of a broader study on the new technology's applications within the European Commission, is a central hub offering secure access to various AI models. GPT@JRC is hosted at the JRC datacentre and supports both open-source AI models, deployed on-premises at the JRC Big Data Analytics Platform[5] and commercial OpenAI's GPT models running in the European Cloud under a Commission contract with an opt-out clause on prompt analysis by third parties.

[3] https://promedmail.org/
[4] https://www.who.int/emergencies/disease-outbreak-news/
[5] https://jeodpp.jrc.ec.europa.eu/bdap/

2.2.1. EpiTator

EpiTator[6] is a comprehensive open-source epidemiological annotation tool, specifically designed for the extraction of epidemiological information from texts[1]. This tool is essentially a Python script that leverages the powerful Natural Language Processing (NLP) spaCy library[7] to extract critical named-entities that are of particular interest in epidemiology applications. These entities include diseases, locations, dates, and counts. One of the remarkable features of EpiTator is the *Resolved Keyword Annotator*. It utilises an SQLite database of entities, which ensures that multiple synonyms for an entity are resolved to a single id, thereby enhancing the accuracy and consistency of the extracted data. EpiTator also imports information about infectious diseases and animal species from several reliable sources, including Disease Ontology, Wikidata, and the Integrated Taxonomic Information System (ITIS). The *Count Annotator* feature of EpiTator extracts and parses count values, identifying attributes such as whether the count refers to cases or deaths, or if the value is approximate. This provides a more nuanced understanding of the data. Instead, the *Date Annotator* feature identifies and parses dates and date ranges. EpiTator employs a key entity filtering process that condenses the output to a single entity per class that best describes the corresponding event. Despite EpiTator's capability to return all entities of an entity class (e.g., disease) found in a text, this filtering process is necessary to distill the most pertinent information. It achieves this using the most-frequent approach filtering, which identifies the key entity from all the entities by selecting the most frequently mentioned one per class.

2.2.2. Pythia-12b

The Pythia-12b model is a state-of-the-art transformer-based LLM that follows the architecture of the highly popular GPT3 model. This open-source model is owned by EleutherAI[8], a collective of AI specialists aiming at advancing AI in an open and collaborative manner. Pythia-12b has been primarily designed to facilitate research on the behaviour, functionality, and limitations of LLMs. The model operates within a context length of 4,096 tokens and consists of a whopping 12 billion parameters.

2.2.3. Mpt-30b-chat

The Mpt-30b-chat model is a state-of-the-art, open-source LLM owned by MosaicML[9]. One of the key features is its specialisation in chat and dialogue generation. It has been fine-tuned to handle conversational dynamics, making it an excellent choice for creating interactive AI applications. It was fine-tuned on several datasets including ShareGPT-Vicuna, Camel-AI, GPTeacher, Guanaco, Baize, and some more generated datasets. In terms of technical specifics, the Mpt-30b-chat model boasts a context length of 8,192 tokens, and is built with an impressive 30 billion parameters. This vast number of parameters allows the model to capture and learn from a wide variety of linguistic nuances, thereby greatly enhancing its conversational capabilities and predictive performance.

[6] https://github.com/ecohealthalliance/EpiTator
[7] https://spacy.io/
[8] https://huggingface.co/EleutherAI/pythia-12b
[9] https://huggingface.co/mosaicml/mpt-30b-chat

2.2.4. Llama-2-70b-chat

The Llama-2-70b-chat model [135] is an advanced open-source LLM owned by Meta[10]. It is one of the most powerful open LLMs currently available. The base model, Llama 2, was pretrained on a diverse range of publicly available online data sources. The fine-tuned version of this model, known as Llama-2-70b-chat, further enhances this understanding by leveraging publicly available instruction datasets and over 1 million human annotations. This extensive fine-tuning process ensures a high level of precision and adaptability in its responses. In terms of its technical specifications, the Llama-2-70b-chat model operates with a context length of 4,096 tokens. Moreover, the model is built with an impressive 70 billion parameters.

2.2.5. Mistral-7b-openorca

The Mistral-7b-openorca model[11] is an open-source model, owned and fine-tuned by OpenOrca using its datasets[12] on top of the Mistral LLM [99]. Despite its smaller size with 7 billion parameters, the model is fast at inference and delivers robust performance across a wide range of tasks. With a context length of 4,096 tokens, the Mistral-7b-openorca is one of the best performing open-source models available for models under 30 billion parameters. This makes it a strong candidate for many API use-cases due to its combination of size, speed, and performance.

2.2.6. Zephyr-7b-alpha

The Zephyr-7b-alpha model is an open-source model owned by HuggingFace[13]. This model has been fine-tuned by HuggingFace using a combination of publicly available and synthetic datasets on top of the Mistral LLM. Despite its small size with only 7 billion parameters, it often outperforms the larger Llama-2-13B model, demonstrating performance comparable to several models in the 20-30B range.

2.2.7. Gpt-3.5-turbo-16k

The Gpt-3.5-turbo-16k model is a robust commercial model owned by OpenAI. This model is an advanced version of OpenAI's Gpt-3.5-turbo, colloquially known as ChatGPT, but with four times the context. The versatility of this model is apparent in its ability to perform a wide range of tasks. With its context length of 16,384 tokens, the Gpt-3.5-turbo-16k model offers an expansive reach for various applications. However, being a commercial model, its extensive usage comes with a considerable cost and some usage constraints.

[10] https://ai.meta.com/llama/
[11] https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca
[12] https://huggingface.co/datasets/Open-Orca/OpenOrca
[13] https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha

2.2.8. Gpt-4-32k

The Gpt-4-32k model, another commercial offering from OpenAI, is a powerful tool which sets itself apart in terms of its problem-solving capabilities. It can solve difficult problems with a greater level of accuracy than any of OpenAI's previous models, leading to its reputation as the best LLM available for any task. While it may come at a higher cost, the Gpt-4-32k model's superior performance and its ability to manage approximately 80 pages of text make it one of the best LLM options available. Its context length of 32,768 tokens provides an even greater scope for processing and managing large amounts of data, making it a perfect option to handle large texts.

2.2.9. Gpt-4-FewShots

The Gpt-4-FewShots is a variant to the Gpt-4-32k model obtained by passing three examples (Three-Shots) of epidemic information extraction in the context to the Gpt-4-32k prompt, with the goal of specialising the model to handle the specific epidemic information extraction task. In-context learning (ICL) [19] refers to a learning approach where an AI model learns from its context. LLMs have already shown an ability for ICL as they scale in terms of model and corpus sizes [44]. The fundamental principle of in-context learning is learning through analogy. Initially, ICL needs a handful of examples to create a demonstration context, usually framed in natural language. Following this, ICL combines a query question with a demonstration context to create a prompt. This prompt is then input into the language model for prediction. We have leveraged the ICL approach on the Gpt-4-32k model[14], leading to Gpt-4-FewShots, to allow for a more efficient and accurate way to extract and process information on epidemics.

2.2.10. Open-Ensemble

The Open-Ensemble model is an AI approach that leverages the power of multiple high-performing open-source models to generate outputs. It uses a majority voting system to determine the best result from the contributed models. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use [115].

The models included in this ensemble are Llama-2-70b-chat, Mistral-7b-openorca, and Zephyr-7b-alpha, which, as shown in the following sections, resulted in the best-performing open-source LLMs for epidemic IE. Each of these models brings complementary strengths, enhancing the overall robustness of the ensemble. This Open-Ensemble model, with its majority voting mechanism, ensures that the most agreed-upon output from these three models is chosen, thereby increasing the likelihood of accuracy and reliability in its predictions.

However, this approach introduces certain trade-offs. In particular, inference is inherently slower, as the extraction process must be executed independently across all three models before aggregation. In addition, while the models are open-source and do not incur direct monetary costs, the approach requires increased computational resources. Despite these limitations, the ensemble strategy provides improved robustness and consistency compared to individual models, and allows us to assess whether it can compete with the performance of more advanced commercial GPT models.

[14] Please note that ICL was only applied to Gpt-4-32k due its large context capability, which was not amenable in the other studied LLMs which are characterised by a way more lower context.

2.3. Computational Experiments and Discussion

In this section, we delve into the application of LLMs for the purpose of epidemic information extraction. The extracted entities include the name of the disease, the country where the cases were reported, the confirmed case count, and the date of the case count. We compare the performance of the different LLMs for infectious disease epidemic information extraction, along with EpiTator. The comparison has been made over a subset of the Incident Database (IDB), a repository developed for the purpose of event-based surveillance[15]. The IDB is structured to house a comprehensive collection of data related to epidemic events, making it an important resource for researchers and healthcare professionals. The database includes a sample of ProMED and WHO DONs that have been meticulously annotated by domain experts. For the purpose of comparing the performance of the tested epidemic information extraction models, a gold standard subset of the IDB comprising 171 carefully selected samples has been selected. This subset provides a benchmark against which the performance of the models have been assessed. It should be noted that, due to the limited transparency of LLM pre-training datasets, it is not possible to fully exclude potential overlap between the evaluation data and the models' training data, which represents a general limitation in such assessments.

We have evaluated the information extraction task as a binary classification problem where the negative class means that no information ("None") is associated to the IDB data sample, while the the positive class indicates that the IDB is labelled with a specific epidemic information. The models have been tested on their ability to recognise such positive and negative classes by considering the widely adopted Precision, Recall, and F1 score metrics [33].In words, the Precision is the accuracy for the positive class predictions. The Recall (Sensitivity) is the ratio of the positive class instances that are correctly detected as such by the classifier. The F1 score is the harmonic mean of precision and recall: whereas the regular mean treats all values equally, the harmonic one gives much weight to low values, and for this reason is a metric preferred to the classical accuracy in an imbalanced problem setting.

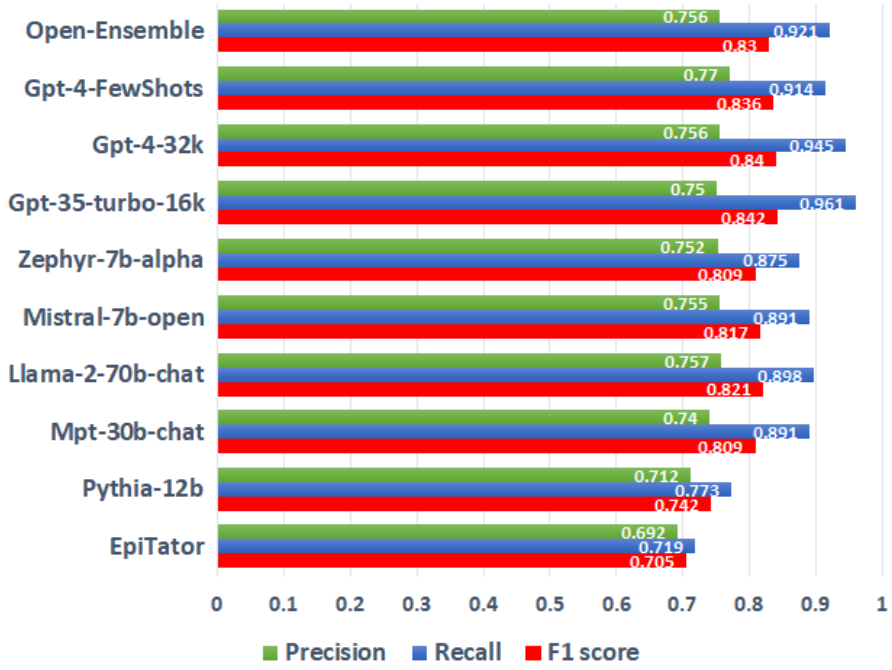
Our computational results are reported in Figure 1, Figure 2, Figure 3 and Figure 4, which report the comparison of the different algorithms for the extraction, respectively, of the pandemic name, the country where this outbreak occurred, the related date, and the number of cases associated to the specific outbreak. In particular, each illustration reports the considered metric scores obtained by each method (on the left), also with the corresponding graphical performance illustration, for a better understanding (on the right). Focusing on the performance of the LLMs, two models, namely Pythia-12b and Mpt-30b-chat, have shown underwhelming performance, falling below the standard NLP-based method, EpiTator. However, this was not the case for all LLMs. The majority of them have demonstrated a remarkable ability to outperform EpiTator by a significant margin. Among the LLMs in the evaluation, Gpt-4-32k stood out by achieving the best overall performance, as expected. Another very well performer was Gpt-3.5-turbo-16k, which also managed to achieve a significant performance level, as shown in the tables of the results. Furthermore, experimenting with in-context learning via the 3-shots method (Gpt-4-32k-FewShots), we have managed to improve the performance of Gpt-4, although this was not always the case.

Although the superior performance of OpenAI GPT models relative to the other tested methods, their use comes with its own set of challenges. These models are costly to implement, and some of their usage restrictions make them unsuitable in practice for a extensive deployment on a large dataset, like the full ProMED and DONs. On the brighter side, some open-source models performed very well, albeit slightly below the level of the OpenAI GPTs. These include Llama-2-70b-chat, Mistral-7b-openorca, and Zephyr-7b-alpha. It's important to note that Llama-2-70b-chat was found to be slower in terms of computational times compared to the rest, given with large number of parameters (70 billions) of its underlying

[15] https://github.com/aauss/EventEpi

Figure 1. Comparison of the models for the extraction of the pandemic name.

	Prec	*Rec*	*F1*
EpiTator	0.692	0.719	0.705
Pythia-12b	0.712	0.773	0.742
Mpt-30b-chat	0.74	0.891	0.809
Llama-2-70b-chat	0.757	0.898	0.821
Mistral-7b-open	0.755	0.891	0.817
Zephyr-7b-alpha	0.752	0.875	0.809
Gpt-3.5-turbo-16k	0.75	0.961	0.842
Gpt-4-32k	0.756	0.945	0.84
Gpt-4-FewShots	0.77	0.914	0.836
Open-Ensemble	0.756	0.921	0.83



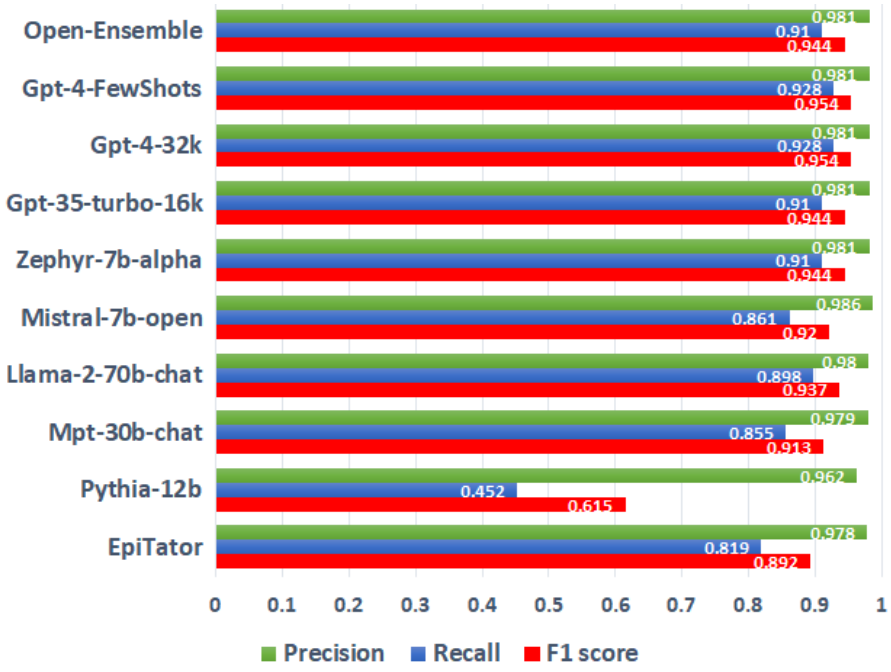
Source: Consoli et al., 2024 [32].

model. Finally, the adopted ensemble approach (OpenLLMs-Ensemble) on these three open-source LLMs produces an overall robust and satisfactory performance on all the four epidemic information extraction tasks, resulting to be comparable to the GPT models. Therefore, OpenLLMs-Ensemble allows for a full deployment on the entire ProMED and DONs data, spanning approximately 70,000 documents, thereby providing a comprehensive and efficient solution for infectious disease epidemic information extraction.

This use case has demonstrated the significant potential of LLMs in epidemic surveillance, particularly in the extraction of valuable epidemic information from unstructured big data sources. Several LLMs were evaluated and their capabilities in extracting epidemic information were assessed. The findings of this study suggest that LLMs, particularly when used in an ensemble, can significantly enhance the accuracy and timeliness of epidemic modelling and forecasting. This application of LLMs not only streamlines the data gathering process but also has the potential to improve early warning systems and epidemic re-

Figure 2. Comparison of the models for the extraction of the country name.

	Prec	*Rec*	*F1*
EpiTator	0.978	0.819	0.892
Pythia-12b	0.962	0.452	0.615
Mpt-30b-chat	0.979	0.855	0.913
Llama-2-70b-chat	0.98	0.898	0.937
Mistral-7b-open	0.986	0.861	0.92
Zephyr-7b-alpha	0.981	0.91	0.944
Gpt-3.5-turbo-16k	0.981	0.91	0.944
Gpt-4-32k	0.981	0.928	0.954
Gpt-4-FewShots	0.981	0.928	0.954
Open-Ensemble	0.981	0.91	0.944

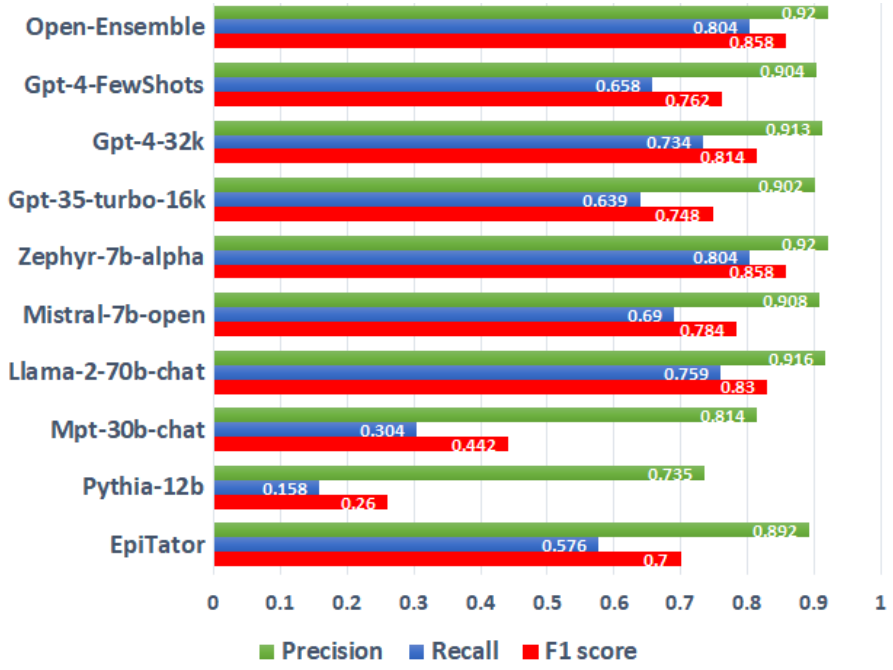


Source: Consoli et al., 2024 [32].

sponse strategies. These models provide a promising tool for managing future pandemic events, reducing their impact on society and economies, and, finally, saving lives.

Figure 3. Comparison of the models for the extraction of the pandemic date.

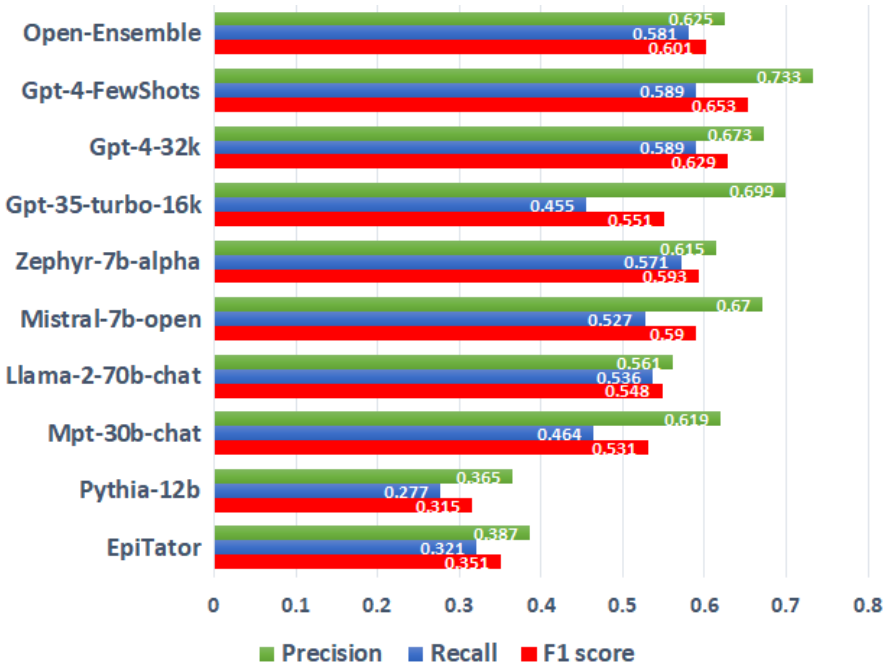
	Prec	*Rec*	*F1*
EpiTator	0.892	0.576	0.7
Pythia-12b	0.735	0.158	0.26
Mpt-30b-chat	0.814	0.304	0.442
Llama-2-70b-chat	0.916	0.759	0.83
Mistral-7b-open	0.908	0.69	0.784
Zephyr-7b-alpha	0.92	0.804	0.858
Gpt-3.5-turbo-16k	0.902	0.639	0.748
Gpt-4-32k	0.913	0.734	0.814
Gpt-4-FewShots	0.904	0.658	0.762
Open-Ensemble	0.92	0.804	0.858



Source: Consoli et al., 2024 [32].

Figure 4. Comparison of the models for the extraction of the number of cases engender by the pandemic.

	Prec	*Rec*	*F1*
EpiTator	0.387	0.321	0.351
Pythia-12b	0.365	0.277	0.315
Mpt-30b-chat	0.619	0.464	0.531
Llama-2-70b-chat	0.561	0.536	0.548
Mistral-7b-open	0.67	0.527	0.59
Zephyr-7b-alpha	0.615	0.571	0.593
Gpt-3.5-turbo-16k	0.699	0.455	0.551
Gpt-4-32k	0.673	0.589	0.629
Gpt-4-FewShots	0.733	0.589	0.653
Open-Ensemble	0.625	0.581	0.601



Source: Consoli et al., 2024 [32].

3. Development of the Epidemiology Knowledge Graph

Processing the vast quantity of unstructured epidemiological datasets poses significant challenges [108, 138]. In this chapter we focus on the World Health Organization (WHO) Disease Outbreak News (DONs, see Section 2.1.2.) This publicly accessible system disseminates information on events related to health emergencies, reported by national authorities and various collaborators [103, 96]. The WHO compiles these data into narrative descriptions that detail the specifics of each outbreak. Typically, these narratives pinpoint the affected geographical area, the disease in question or a syndromic occurrence with an undetermined cause, and may include details of the response efforts, such as the status of laboratory testing for confirmation. Despite their considerable information value, DONs have been underutilised in academic research, primarily due to their unstructured, prose-based format, which poses significant barriers to systematic analysis and integration into automated health informatics systems [90].

In response to these challenges, our study employs an ensemble of advanced generative AI techniques, leveraging the strengths of multiple Large Language Models (LLMs) [136, 19], to extract and structure relevant epidemiological information from the WHO DONs [32], as already described in the previous Section 2. The epidemiological information (i.e., disease and pathogen name, involved country, date of event, cases total, and mortality) extracted via this novel approach is made publicly available adopting *FAIR* (*F*indable, *A*ccessible, *I*nteroperable, and *R*eusable) principles and following the paradigms of Linked Open Data (LOD) [60].

Extracting and integrating heterogeneous information for knowledge discovery can be a complex task for researchers and public health surveillance experts. To address these issues, we adopt a knowledge-centric paradigm involving the creation of a structured, interconnected, and formal representation of the WHO DONs into a knowledge graph (KG) [9, 106], in order to improve the capabilities of conducting complex and extensive analysis of disease outbreaks. In recent years, KGs have become a powerful tool for organising domain knowledge of various nature and integrating the contained information to reveal hidden associations [68]. They have been successfully used by the research community to create semantically enriched representations of data across various domains [9] as well as to offer significant support to AI systems across multiple domains [132]. KGs are data structures describing the key entities in a domain and their relationships, presenting information in a format accessible and interpretable by both machines and humans [61]. The relationship between two entities is typically formalised as a triple in the format *<subject, predicate, object>* following the Resource Description Framework (RDF[1] as a standard model [68] (e.g., *<Chikungunya virus, cause, 109 deaths>* or *<Yersinia Pestis, occur, China>*), on top of which different kinds of reasoning activities can be conducted using expressive languages, such as Description Logic (DL) [12], Web Ontology Language (OWL) [6] or SPARQL [110].

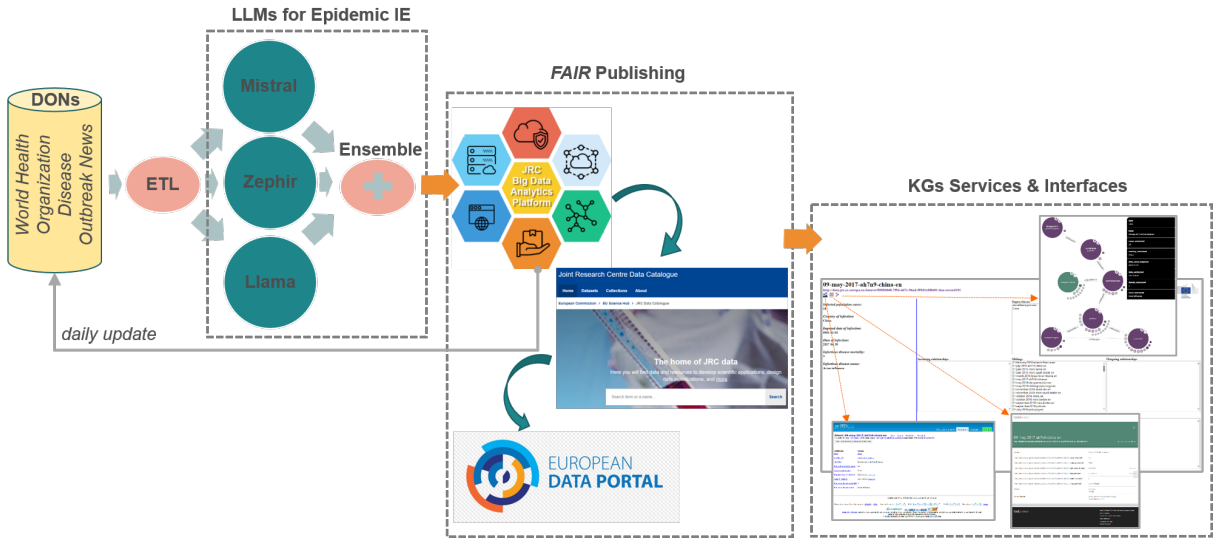
While other KG architectures are possible, such as through graph DBMS (such as Neo4j[2]), each with its own advantages and disadvantages [4], we have chosen to adopt RDF as our data model due to its ability to provide a robust framework for the semantic representation and integration of complex datasets [104]. RDF is ideal for ensuring interoperability and extensibility, as it follows widely accepted standards that enable seamless data integration from diverse sources [92]. This is particularly beneficial in the context of epidemiological data, where the integration of various ontologies and datasets can provide deeper insights and more accurate analysis. Moreover, RDF's compatibility with OWL ontologies allows us to leverage semantic reasoning capabilities, enabling more sophisticated data interpretations and inferencing [6]. These features are essential for enhancing the utility and usability of the produced epidemiological knowledge graph (*eKG*), as they support advanced queries and analysis that go beyond simple data retrieval. By

[1] Resource Description Framework (RDF): https://www.w3.org/RDF/
[2] http://neo4j.org/

choosing RDF, we ensure that our dataset is not only semantically rich and coherent but also well aligned with the principles of FAIR data, thus promoting broader data sharing and reuse within the scientific and public health communities [60]. By following this approach, we construct the eKG from DONs, offering a nuanced representation of the public health domain of knowledge. This formalised framework effectively integrates extracted epidemiological information from DONs, enhancing the analysis and understanding of public health surveillance data.

Figure 5 provides a schematic overview of the pipeline, which will be described in more detail in the "Methods" section. A process is triggered to extract, transform and load (ETL) new reports from WHO DONs, which are then processed by the ensemble of LLMs for the epidemiological information extraction (IE) task. The extracted information is then published using FAIR principles along with the derived eKG. This enables a number of semantic services and interfaces to be deployed above, allowing access and exploration of the derived knowledge graph. These resources have significant potentials to improve the utility and usability of WHO DONs for both researchers and public health professionals, and promote new research opportunities in public health surveillance and epidemiology for the analysis of disease outbreaks, and response to public health threats.

Figure 5. Schematic overview of the pipeline to extract and publish relevant epidemiological information from the WHO's Disease Outbreak News.



Source: Consoli et al., 2025 [31].

In summary, the work described in this chapter provides a comprehensive and structured dataset derived from the WHO DONs. Our work showcases the value of extracting epidemiological data from narrative reports. The establishment of a knowledge graph of epidemiological information and the application of LLMs in an ensemble fashion [32, 154] to process and analyse unstructured data mark a transformative step towards enhancing our global understanding and management of disease outbreaks.

3.1. Methods

This section describes how we constructed the knowledge graph of epidemiological information from DONs. The released resource was built using the automatic pipeline depicted in Figure 5, composed of several modules that can be grouped into three main categories: *i) LLMs for Epidemiological Information*

Extraction (IE), where the important epidemiological information is extracted from DONs using an ensemble of open-source LLMs, *ii) eKG FAIR Publishing*, which converts the extracted information into the eKG knowledge graph following the paradigm of Linked Open Data, adds additional contextual information to the extractions, and makes the resource available to local authorities, research organisations, national and European governmental institutions and international organisations through a public repository, and *iii) eKG Services & Interfaces*, consisting of several additional services and tools that we are building on top to expand the final user experience.

3.1.1. LLMs for Epidemic IE

As already described in the previous Section 2, this study utilises LLMs to extract useful epidemiological information for analysis. These predictions are informed by the model's learning from extensive text datasets.

Building on the assessment of recent LLMs for epidemiological information extraction [32], our work employs an ensemble approach, which was shown to be comparable or even superior to commercial techniques for extracting structured epidemiological information from text, including outbreak disease names, involved countries, event dates, total cases, and deaths (Section 2). A solution based on open-source LLMs is ideal in this context because it overcomes specific usage constraints of commercial models, e.g. maximum number of tokens that can be processed per day, maximum number of API requests per units of time etc., allowing for a full deployment on the entire DONs data.

By leveraging the extracted information in the context of infectious disease outbreaks from DONs, and integrating it into a surveillance system, we aim to improve the precision and speed of epidemiological analysis. After extracting, transforming and loading (ETL) the new reports from WHO DONs, cleaned and normalised, they are processed by the ensemble of those LLMs. The adopted LLMs for the ensemble are: *Mistral-7B-OpenOrca*, *Zephyr-7B-Beta*, and *Meta-Llama-3-70B-Instruct*[3].

Afterwards, for the extraction of the key epidemiological entities (name of disease outbreaks, involved country, date of the outbreak, number of confirmed cases and deaths) for each DON report, the LLMs are supplied with the prompt depicted below [32]:
"*From the text below extract the following items:*
1 - The name of the disease that caused the outbreak.
2 - The name of the country where this disease outbreak occurred, if present.
3 - The date when this disease outbreak occurred, if present. Show the date in the format YYYY-mm-dd.
4 - The number of deaths caused exclusively by the disease outbreak mentioned in the text, if present.
Format your response as a JSON object with the following keys: disease name, country, date, cases. If the information is not present, do not invent and use "None" as the value.
Text: … "
The used prompts were selected based on their ability to effectively extract key epidemiological entities with high accuracy, as demonstrated through iterative testing and validation against a benchmark portion of the data. This ensures that the extracted information is both precise and reliable, aligning with the objectives of our study. As an example, consider the WHO DON "Nipah virus - India" of the 31st May 2018 (https://www.who.int/emergencies/disease-outbreak-news/item/31-may-2018-nipah-virus-india-en), reporting a Nipah virus outbreak in Kerala, India, resulting in 15 laboratory-confirmed cases, 13 deaths and

[3] Please note that the original LLMs adopted for the ensemble in the previous section were: *Mistral-7B-OpenOrca*, *Zephyr-7B-Alpha*, and *Meta-Llama-2-70B-Chat*. Here we adopt the upgraded, larger versions of these models.

2 hospitalisations, with bats linked to the initial transmission, and prompting a robust local and WHO-supported public health response, including surveillance and infection control measures. The extraction answer provided in JSON by the *Mistral-7B-OpenOrca* model using the detailed prompt was:

{ *"disease": "Nipah virus", "country": "India", "date": "2018-05-19", "cases": "15", "deaths": "13"* },

which we can note correctly structures the main epidemiological information from this disease outbreak.

In addressing the challenge of ensuring valid JSON outputs from LLMs, we opted for an algorithmic validation approach. This involves programmatically verifying JSON validity and employing synthetic heuristics to attempt reconstruction in case of JSON formatting issues. These heuristics involve checking for common JSON structural issues such as typos, missing or extra characters, and unexpected attributes. The process involves:
1. Checking if the output is a valid JSON using a validation function.
2. If invalid, attempting to clean and correct common errors such as misplaced or absent quotation marks, misplaced brackets, and removing irrelevant text.
3. Ensuring all necessary attributes are present and properly formatted.
4. Removing any unexpected attributes not required in the JSON structure.
The outputs we were unable to clean, which accounted for less than 1% of the overall data from our calculations, were discarded. This method allows us to effectively manage and correct JSON outputs, ensuring they meet the necessary structural requirements for further elaboration.

After a report is processed by all the contributed LLMs, majority voting is used among their outputs for each of the extracted epidemiological information, forming in this way the output of the *Ensemble*. This majority voting technique is particularly effective in mitigating the hallucination issue commonly associated with LLMs. Hallucination refers to the generation of information that is not present in the input data [11]. In machine learning, ensemble methods aggregating the predictions from multiple approaches are able to improve the overall performance. This is indeed a well-documented phenomenon in machine learning [115, 154]. The underlying principle is that by pooling together the strengths and mitigating the weaknesses of various models, the ensemble can achieve better accuracy and robustness than any single model alone. By employing majority voting on the extracted epidemiological information from the LLMs, the ensemble approach ensures that only the most consistent and agreed-upon information is retained, thereby reducing the potential for hallucinated outputs and improving the reliability of the extracted epidemiological data. For example, if the extracted number of cases from the three LLMs for a report is: 15, 17, and 15, the *Ensemble* would give 15 as the output for the number of cases for that report. Please note that if each model had given a different answer, the ensemble randomly selects one. When textual epidemiological information, that is disease names and countries, need to be evaluated by the majority voting system, the process is slightly more complicated, as two terms representing semantically the same concept should be grouped together. This was achieved by compiling a dictionary of synonyms of disease and pathogen names, and another of country names synonyms. Synonyms are created by (i) first checking for syntactic similarity among the terms to be compared (e.g. "Trinidad and Tobago" and "Trinidad & Tobago", or "Florida, USA" and "USA (Florida)"; (ii) checking for synonymy by looking for overlapping synsets in WordNet [95]; (iii) checking for semantic similarity among the terms to be compared (e.g. "Middle East respiratory syndrome" and "MERS-CoV", or "Dengue" and "Breakbone fever". To calculate the semantic similarity we used Sentence Transformers, also known as SBERT (Sentence-BERT) [112], which is a modification of the pre-trained BERT (Bidirectional Encoder Representations from Transformers) [42] network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings. These embeddings can then be compared using a metric like the cosine similarity [76] to determine the semantic similarity between sentences. SBERT fine-tunes the BERT network on

a dataset of sentence pairs, training it to produce embeddings that bring semantically similar sentences closer together in the embedding space. This allows for much faster performance in downstream tasks, as the sentence embeddings can be precomputed and easily compared without the need for further BERT processing. In particular, to evaluate the semantic similarity among country names, we experimentally selected the *all-mpnet-base-v2*[4]. SBERT model, which is an all-round model trained on a large and diverse dataset of over 1 billion training pairs and tuned for many diverse use-cases. For the disease outbreaks we used instead BioBERT [80], a domain-specific version of the BERT model pre-trained on biomedical literature to enhance performance on biomedical natural language processing tasks [38]. While alternatives such as SciBERT could also be considered, BioBERT was selected due to its specific focus on biomedical corpora, which are more closely aligned with the characteristics of disease outbreak data. We set an experimental threshold of 0.8 for the semantic similarity $\in$ [0,1], meaning that two terms having a semantic similarity larger than this threshold were considered to represent the same concept. Choosing the most appropriate semantic similarity score can be challenging and often requires empirical experimentation to determine the best fit for a specific scenario. A typical range for semantic similarity values is often between 0.7 and 0.9, which has been supported by numerous studies (see e.g. [57, 153, 72]). For our use case, we experimentally selected a threshold of 0.8, as it demonstrated an effective balance in differentiating terms while maintaining homogeneous groups of synonyms for disease and pathogen names, as well as country names. In this way, clusters of disease names and countries synonyms were obtained, forming dictionaries of similar concepts.

Using the majority voting approach, the *Ensemble* model determines the most precise output from the three participating models, thus boosting the chances of accurate and reliable predictions. This approach enables the model to compete with the performance of more advanced commercial language models for epidemiological IE [32]. By combining the advantages of multiple models, it emerges as an effective strategy for the task of epidemiological information extraction.

3.1.2. eKG FAIR Publishing

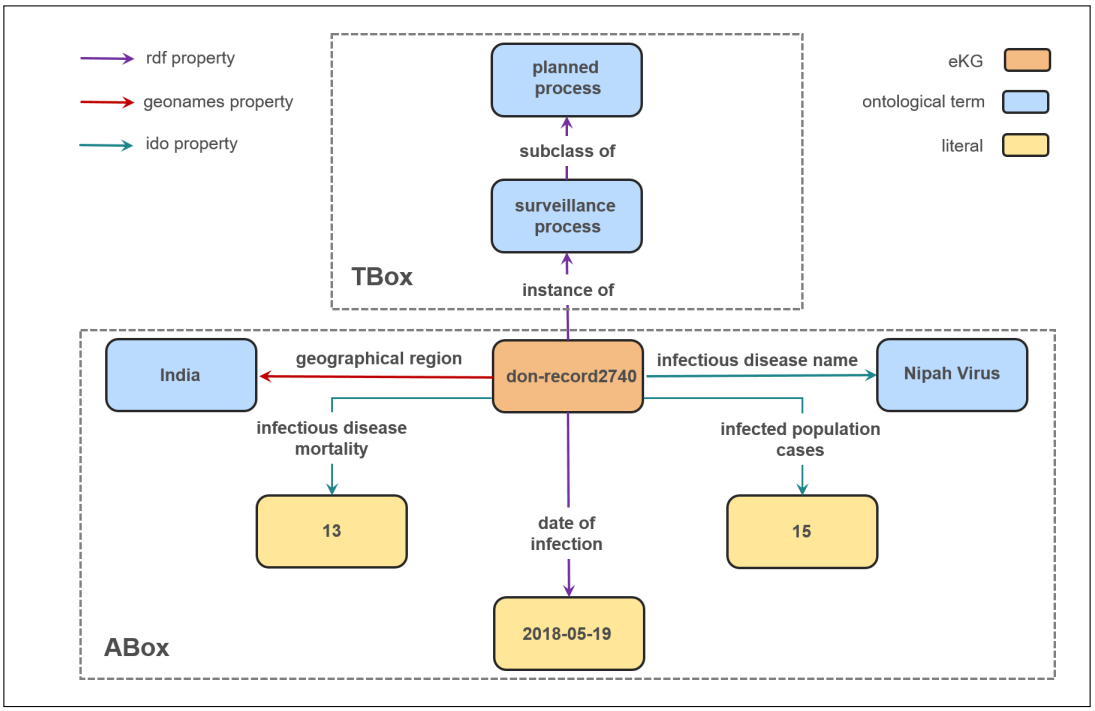
The epidemiological knowledge graph, eKG, has been produced from the information extracted by the WHO DONs, providing a rich data model for the public health surveillance domain. It has been implemented by using Semantic Web technologies, such as RDF[5] and the Web Ontology Language (OWL) [6], which allow human experts to verify, curate, and correct both the data and their schema, to produce an ontologically-grounded knowledge graph.

A knowledge graph can be formally defined as a pair *<T, A>*, where *T* is the TBox and *A* the ABox [22]. The TBox serves as the foundational layer containing the formal descriptions of a domain's structure, encompassing specific class hierarchies, relational properties, and basic axioms or declarations [22] (e.g., a *surveillance process* is a subclass of a *planned process*, as illustrated in Figure 6). In contrast, the ABox details the properties and attributes of instantiated class objects (namely, individuals or data names) and declarations concerning their categorical placement within the TBox structure [22] (e.g., as depicted in Figure 6, the *don-record2740* is a *surveillance process* associated to the *Nipah Virus* infectious disease occurred in *India* on *2018-05-19*, with *15* infected population cases confirmed and *13* deaths). External data entities that exist beyond ontological specifications can be formulated using existing ontological paradigms through TBox strategies (by leveraging schema-level constructs) or ABox implementations (by applying individual-level representations) [25]. The class-based method characterises each data element via *subClassOf* relations with existing ontology classes, while the instance-based method characterises

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[5] https://www.w3.org/TR/REC-rdf-syntax/

them via *instanceOf* relations with existing ontology classes, as done in eKG. In particular, for the extraction of eKG we have strongly followed principles from ontology design patterns (ODPs) [52] and LOD [60], that is, whenever possible we reuse existing ODPs from online LOD repositories. Reused patterns are annotated with the OPLa framework [120] in order to facilitate class mapping and identification, which helps ground the knowledge graph to the ontology classes, ensuring semantic consistency and enhancing data integration. In particular we reuse classes and relations from IDO (Infectious Disease Ontology) [34] and include explicit alignments to them by *rdfs:subClassOf* properties. IDO is a collection of structured entities generally relevant to both the biomedical and clinical aspects of infectious diseases. The structure of IDO adheres to the Basic Formal Ontology (BFO) [100] and aims to provide a coverage of the infectious disease domain. Terms in IDO that are within the scope of other OBO Foundry ontologies [65] are derived from those respective ontologies. In this way we also map indirectly to the WHO's International Classification of Diseases, Version 10 (ICD10)[6].

Figure 6. This picture illustrates an example on the implementation of Description Logic for structured knowledge representation. The TBox comprises concepts such as *surveillance process* and *planned process*, along with assertions defining relationships between these concepts (e.g., *<surveillance process, subclass of, planned process>*). The ABox represents instances of these concepts, like *don-record2740*, and provides assertions relating these instances to other concepts (e.g., *<don-record2740, instance of, surveillance process>* and *<don-record2740, geographical region, India>*) as well as to literal values (e.g., *<don-record2740, date of infection, 2018-05-19>* and *<don-record2740, infected population cases, 15>*).



Source: Consoli et al., 2025 [31].

When creating a knowledge graph from heterogeneous sources, like in our case with eKG, there is the possibility of introducing the same real-world entity multiple times, as they may be expressed in different forms (e.g. "SARS" vs "severe acute respiratory syndrome"). *Entity linking* [119], i.e. grounding the facts in domain ontologies, is a technique used to reduce this phenomenon and an essential feature of knowledge graphs. In our case, the data resources of eKG have been linked to biomedical ontologies using the NCBO

[6] ICD10: https://icd.who.int/browse10/2019/en

BioPortal[7], and to geo-locations using GeoNames[8], ensuring consistent and accurate entity representation across datasets.

The NCBO BioPortal [101] is a comprehensive online platform designed to provide access to a vast repository of biomedical ontologies and related resources. Developed by the US National Center for Biomedical Ontology (NCBO), the BioPortal allows researchers, clinicians, and developers to browse, search, and integrate diverse ontological data to enhance biomedical research and application development. It supports a wide array of functions including ontology browsing, download, versioning, and annotation, facilitating better understanding and interoperability of complex biological and medical data. By offering these tools and resources, the BioPortal plays an important role in advancing semantic technology and data sharing in healthcare and life sciences.

GeoNames, conversely, is an open-source geographical database, regularly updated, that covers all countries worldwide and contains over eleven million place names. We have adopted it to describe geographical entities, such as cities, regions, or country names where a disease outbreak has occurred [78].

Both BioPortal and GeoNames represent invaluable resources for knowledge graph construction in a field such as public health intelligence, where data quality is essential to avoid data duplication and ensure unique identification of outbreak events. Therefore, we have searched infectious disease names and outbreak locations in eKG using the BioPortal Annotator API[9] and the GeoNames Search Webservice[10], respectively, and the discovered entities have been stored in eKG. Specifically, infectious disease names identified in the epidemiological reports have been linked to biomedical ontology entities using the *skos:related* axiom, while the *obo:RO_0001025* property, a.k.a. *located in*, has been adopted for linking the geo-locations to the equivalent GeoNames entities.

The outcome of the entity linking process is twofold. First, it allows eKG to be enriched with additional information from external domain ontologies and GeoNames. For example, if a disease outbreak in eKG is linked to a biomedical entity in the BioPortal, additional information about the disease outbreak such as its symptoms or its transmission patterns could be quickly accessed and analysed. Second, it provides a mechanism for connecting eKG to other knowledge graphs and datasets that are using standard biomedical terminologies. This is important for enabling cross-domain knowledge discovery, integration, and maximising interoperability.

The inclusion of links to external domain ontologies in the eKG entity metadata provides additional contextual information and connections to external knowledge sources, allowing for more comprehensive and diverse data analysis. This linkage allows researchers and public health professionals to draw on a wealth of existing ontological information to better understand the outbreak's characteristics and study effective response strategies. The inclusion of external links also facilitates the integration of eKG with other knowledge graphs and databases, enabling cross-disciplinary research and collaboration.

An in-depth explanation of the structure of eKG and each associated data record is provided in the "Data Records" section, which also offers an overview of the data files and formats. The disease outbreak events have been extracted from the DONs reports by the LLMs ensemble and mapped as individuals of eKG, leveraging the computational power of the JRC Big Data Analytics Platform [122]. The epidemiological

[7] https://bioportal.bioontology.org/
[8] https://www.geonames.org/
[9] https://data.bioontology.org/documentation#nav_annotator
[10] https://www.geonames.org/export/geonames-search.html

knowledge graph is made available [35] in structured RDF/XML and Turtle[11] formats, with raw extractions also given in CSV, within the Joint Research Centre Data Catalogue[12] at the following permanent location: http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601, used also as reference namespace for the eKG individuals. The knowledge graph was also released within the European Data portal [75], the official data repository for European data, at the following permanent location: https://data.europa.eu/data/datasets/89056048-7f5d-4d7c-96ad-f99d1c0f6601.

3.1.3. eKG Services & Interfaces

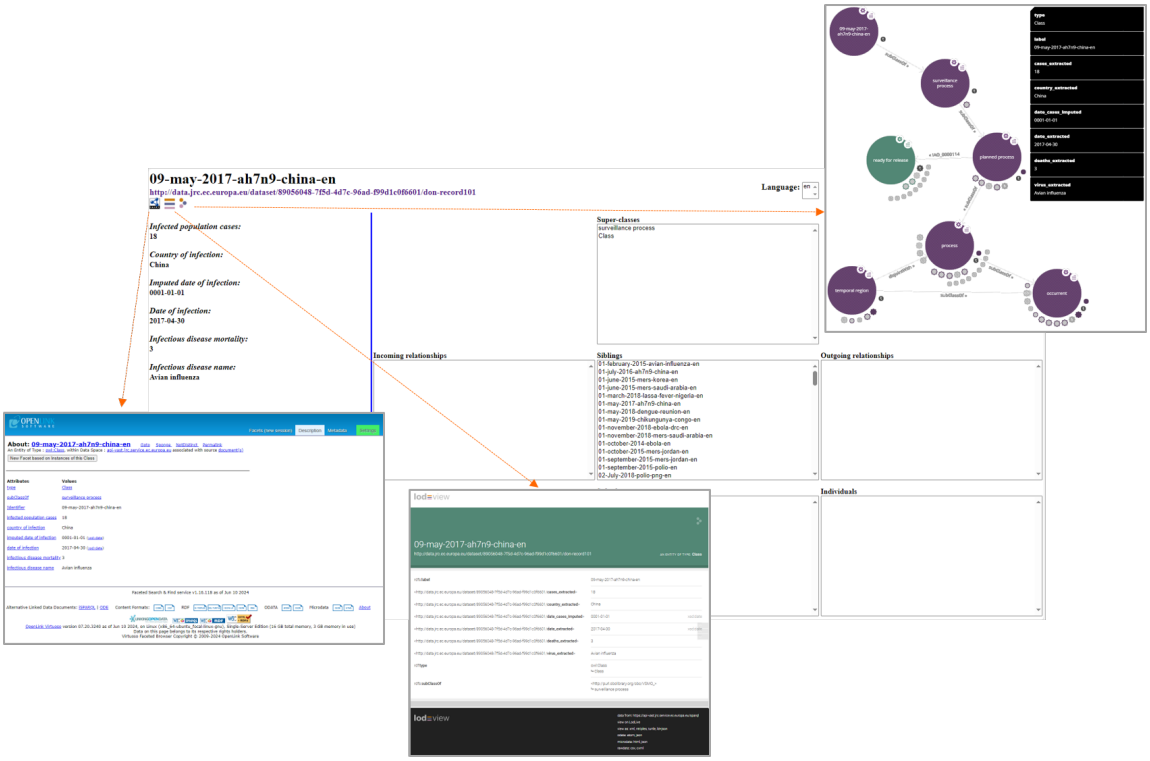
Using the described pipeline, we generated the ontologically-grounded knowledge graph of epidemiological information, comprising roughly of 46K distinct, non-reified (i.e. that are not given additional metadata or context, meaning they are expressed in a straightforward *<subject, predicate, object>* format without any further elaboration or annotation [104]) triples, representing around 2.9K unique outbreak events via a total of 3.6K generalised axioms, as of the 28^{th} February 2025. To access the data programmatically, we set up a *Virtuoso SPARQL endpoint*[13] where eKG can be queried and analytical information on target disease outbreaks, attributes, and relations can be retrieved in structured data format using the SPARQL query language [110]. On top of eKG we are currently developing a number of intuitive and user-friendly services and interfaces at European Commission premises, including content negotiation, visualisation and navigation of data, improved exploitation of the available information and knowledge discovery. These semantic services are applicable and valuable to any kind of knowledge graph based on RDF/OWL technology, enhancing their utility and accessibility across various domains. The central point of access (Figure 7) is a web interface where it is possible to have an immediate view of the main epidemiological information associated to a disease outbreak, along with quick reference pointers to other semantic applications (see also the "Usage Notes" section), namely: the *Virtuoso Faceted Browser*[14], a general purpose RDF data query facility service for faceted browsing over entity relationship types, allowing users to explore the outbreak events in a structured and intuitive way (e.g., see the left snapshot of Figure 7); *LodView* [74], a web application compliant to World Wide Web Consortium (W3C) [59] standards for IRI dereferentiation which provides a representation of our RDF disease outbreak resource through a custom intuitive HTML page[15], e.g. see the bottom snapshot of Figure 7; *LodLive* [23], a navigator of RDF resources based on a graph layout[16], e.g. see the top-right snapshot of Figure 7.

3.2. Data Records

The extracted epidemiological information from DONs is available in CSV format, while both the DONs and the knowledge graph are available in RDF/XML and Turtle formats within the Joint Research Centre Data Catalogue [35]. Data are freely available to the public and can be downloaded without any login details by selecting the appropriate URL in one of the repositories above. The raw CSV file available in the repository [35] contains the extracted details of the occurring event from each report. In particular, it contains the following information:

[11] https://www.w3.org/TR/turtle/
[12] Joint Research Centre Data Catalogue, https://data.jrc.ec.europa.eu/
[13] Virtuoso SPARQL endpoint: https://api-vast.jrc.service.ec.europa.eu/sparql/. Currently the access is password protected, with credentials available upon request to authors.
[14] Virtuoso Faceted Browser: https://api-vast.jrc.service.ec.europa.eu/fct/
[15] http://lodview.it/
[16] http://lodlive.it/

Figure 7. Snippet of eKG exploration services and interfaces.



Source: Consoli et al., 2025 [31].

— “fileid” column, presents the ID of the report, which is taken from the slug of the URL of the DON item;

— “virus_extracted” column, denotes the extracted name of the disease outbreak extracted from the current DON item;

— “country_extracted” column, depicts the involved countries as extracted from the report linked to the specific disease outbreak;

— “date_extracted” column, presents the date when the disease outbreak occurred, in the format YYYY/MM/DD;

— “date_cases_Imputed” column, shows the date reported in the fileid of the report, if present, and it can be related to the publication date of the report (YYYY/MM/DD format), although not formally defined in DONs. It is important to note that the date in the report may vary significantly from the actual date of the disease outbreak. Therefore, this information could be useful for analysis only when the “date_extracted” is missing, which means it was not possible to automatically extract the actual outbreak date from the main text of the report;

— “cases_extracted” column, shows the extracted number of confirmed cases deriving from the specific disease outbreak;

— “deaths_extracted” column, presents the related event mortality.

In our data engineering work to model the knowledge graph of epidemiological information, we adhered to standard style guidelines for identifying and describing linked data resources [53]. Specifically, within

eKG, data names are formatted in lowercase, with any spaces replaced by dashes, adhering to established conventions for naming and labeling knowledge graphs [53, 61]. Conversely, class names in the ontology definitions are formatted in uppercase, in line with common style guidelines for such terminologies [53, 61]. In this context, data names refer to the specific identifiers assigned to data elements within the ABox of the knowledge graph, while class names in the ontology definitions are part of its TBox, defining the structure and constraints of eKG [68].

The main classes of the knowledge graph are mapped to RDF/OWL, as illustrated in Figure 8. A disease outbreak event from DONs is an instance (*rdf:type*) of the *IDO:surveillance_process* class, which is a specialisation of a generic *OBI_0000011:planned_process*. A disease outbreak is associated with an *IDO_0000436:infectious_disease*, which has a specific *infectious_disease_name* that corresponds to the name of the disease extracted from the DONs text, occurring in a *GEO_000000372:geographical_region* on a *date_of _infection*, i.e., a *dcterms:date* [140]. Please note that when a date is reported in the title of a DONs, such as in the "Nipah virus - India" DON of the 31st May 2018 as an example, an additional *imputed_date_of_infection* is also associated to the disease outbreak event; however, this information should be taken with the due care, given the reporting date might be delayed with respect to the actual occurring date of the disease outbreak. In the "Nipah virus - India" DON example, indeed, the reporting *imputed_date_of_infection* would be the "2018-05-31", but the actual *date_of_infection* reported in the DON is the "2018-05-19". The *imputed_date_of_infection* could be useful for analysis only when it was not possible to automatically extract from the main text of the report an actual *date_of_infection*. A disease outbreak event is then associated with a portion of *infected_population* expressed in terms of *IDO_0000511:infected _population_cases* extracted from the DON report, which can ultimately lead to an *IDO_0000489:infectious _disease_mortality*, i.e., the extracted number of deaths from the disease outbreak report. It is worth noting that the *skos:related* triples serve to connect the "Nipah virus" disease outbreak, identified in the report, with corresponding outbreak names from various open-source vocabularies on the web through entity linking. Additionally, the triple having *obo:RO_0001025* property (which denotes the relationship *located in*) associates the location "India" with its equivalent GeoNames identifier, i.e. *geonames:1269750*, also facilitated by entity linking.

Consider the following vocabulary prefixes, in RDF/XML:

```
<rdf:RDF
    xmlns="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-99d1c0f6601/"
    xml:base="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xml="http://www.w3.org/XML/1998/namespace"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:obo="http://purl.obolibrary.org/obo/"
>
```

The RDF/XML triples of this example are encapsulated within the following OWL statement:

```
<rdf:Description rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-
                                                    f99d1c0f6601/don-record2740">
```

<rdf:type rdf:resource="http://purl.obolibrary.org/obo/VSMO_"/>
<dcterms:identifier>don-record2740</dcterms:identifierl>
<dcterms:description>Outbreak of Nipah Virus occurred in India on 2018-05-19, with confirmed 15 cases and 13 deaths.</dcterms:description>
<rdfs:label>31-may-2018-nipah-virus-india-en</rdfs:label>
<virus_extracted>Nipah Virus</virus_extracted>
<skos:related rdf:resource="http://purl.bioontology.org/ontology/MESH/D045405"/>
<skos:related rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C112359"/>
<skos:related rdf:resource="http://purl.bioontology.org/ontology/SNOMEDCT/115511007"/>
<skos:related rdf:resource="http://purl.obolibrary.org/obo/NCBITaxon_3052225"/>
<skos:related rdf:resource="http://sbmi.uth.tmc.edu/ontology/ochv#C0751673/>
<country_extracted>India</country_extracted>
<obo:RO_0001025 rdf:resource="https://sws.geonames.org/1269750/"/>
<date_extracted rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-19 </date_extracted>
<date_cases_Imputed rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-31 </date_cases_Imputed>
<cases_extracted>15</cases_extracted>
<deaths_extracted>13</deaths_extracted>
</owl:Class>

where the raw extractions metadata are linked the ontology classes of eKG as shown below:

<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/virus_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_0000436"/>
</owl:Class>

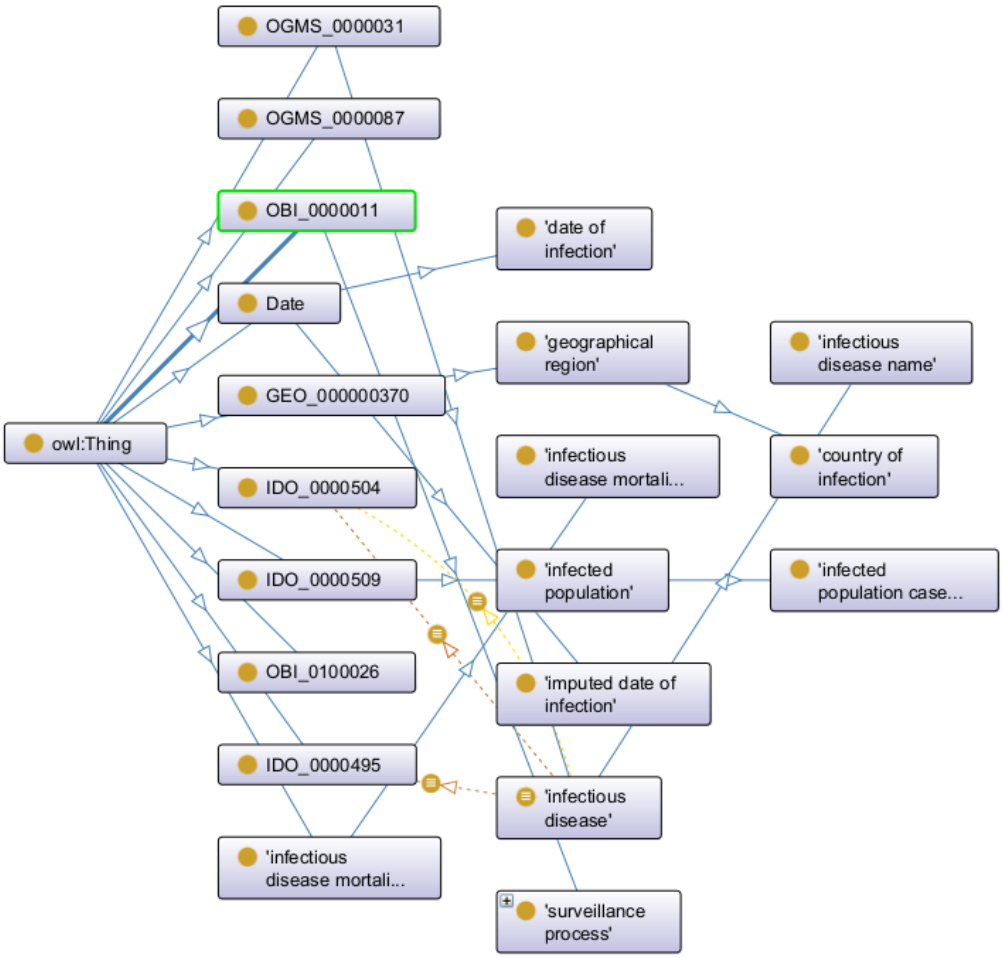
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/country_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/GEO_000000372"/>
</owl:Class>

<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/date_extracted">
<rdfs:subClassOf rdf:resource="http://purl.org/dc/terms/date"/>
</owl:Class>

<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/date_cases_Imputed">
<rdfs:subClassOf rdf:resource="http://purl.org/dc/terms/date"/>
</owl:Class>

<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/cases_extracted">

Figure 8. Main classes within eKG.



Source: Consoli et al., 2025 [31].

<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_0000511"/>
</owl:Class>

<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/deaths_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_0000489"/>
</owl:Class>

We enable direct access to the extracted knowledge graph data via custom SPARQL queries through a dedicated REST API SPARQL endpoint.

3.2.1. Data Statistics

After removing 353 entries with missing country and disease, the dataset consists of 2,556 entries, listing 144 unique pathogens/diseases and 207 unique countries. Table 2 shows the top 10 disease outbreak names, countries, and disease outbreak names - country pairs by total number of extracted events. The

events detected contain a combination of human diseases and zoonoses, with worldwide coverage. Figure 9 presents four case studies as a time series of the number of cases over time. Figure 9, Panel a, shows the numbers of MERS-Cov cases reconstructed by the LLM in Saudi Arabia, with a consistent reconstruction of the MERS-Cov cases experienced in the area[17]. Panels b) and d) of Figure 9 show the number of Ebola cases reconstructed by the LLM in the Democratic Republic of the Congo (DRC) and Guinea, respectively, illustrating the Kivu Ebola Epidemic of 2018-2020 in DRC[18] and the West Africa outbreak of 2014-2016[19]. Panel c) of Figure 9 shows instead the reconstructed number of SARS cases in China, with a qualitative reconstruction of the SARS outbreak of 2002-2004[20]. Overall, the dataset shows potential for reconstructing trends for some epidemics, particularly those involving diseases included in the International Health Regulations, as these are consistently reported and updated in the DONs. For an in-depth quantitative comparison, see the next "Technical Validation" section.

Table 2. The top reported disease outbreak names, countries, and disease outbreak name-country pairs by total number of extracted events (number given in parentheses). Pairs of disease outbreak name-country in bold are further illustrated in Figure 9.

Top 10 Disease Outbreaks	Top 10 Countries	Top 10 Disease Outbreak Names - Country pairs
Avian influenza (570)	China (243)	**Saudi Arabia - MERSCoV** (201)
MERSCoV (296)	Democratic Republic of Congo (226)	China - Avian influenza (172)
Ebola (211)	Saudi Arabia (225)	**Democratic Republic of Congo - Ebola** (134)
Cholera (146)	Egypt (155)	The Netherlands - Avian influenza (108)
Yellow Fever (146)	The Netherlands (142)	Egypt - Avian influenza (101)
SARS (78)	United Republic of Tanzania (89)	**China-SARS** (51)
Hemorrhagic fever (68)	Viet Nam (66)	Viet Nam - Avian influenza (41)
Dengue Fever (56)	Angola (60)	Democratic Republic of Congo - Hemorrhagic fever (31)
Meningococcal (Group C) (54)	Cameroon (51)	**Guinea - Ebola** (30)
H5N1 (51)	Guinea Bissau (49)	Cameroon - Yellow Fever (30)

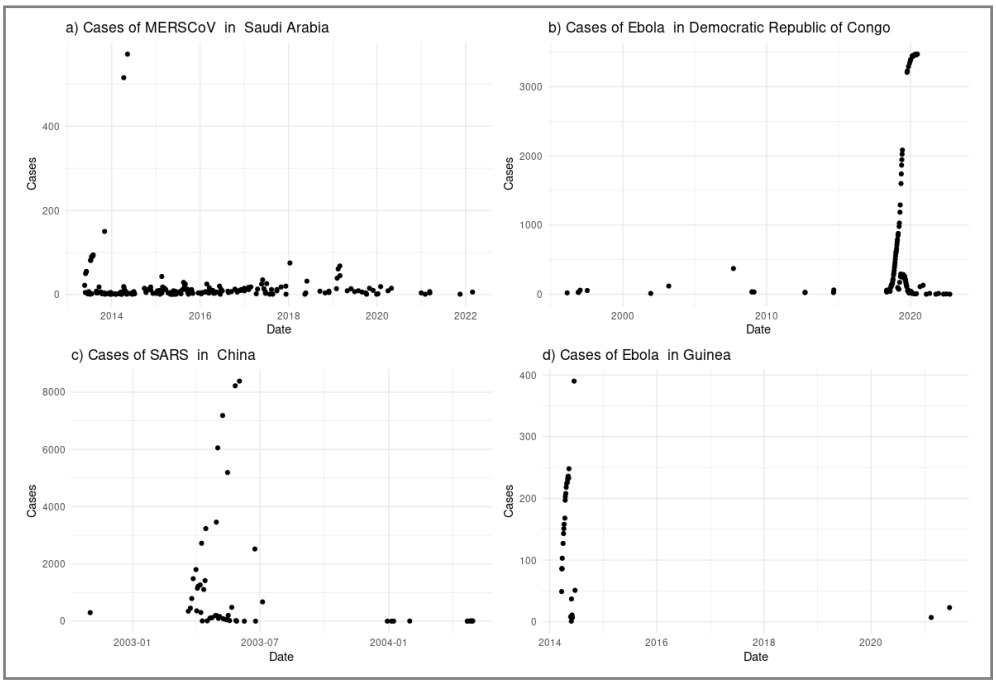
Source: Consoli et al., 2025 [31].

3.3. Validation & Usage

We compared the yearly number of MERS-Cov cases in Saudi Arabia from the reconstructed events with the number of confirmed cases reported by the WHO. Figure 10 shows the scatter plots of WHO yearly

[17] Middle East respiratory syndrome coronavirus, Kingdom of Saudi Arabia: https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON516
[18] Ebola outbreak 2018-2020, North Kivu-Ituri: https://www.who.int/emergencies/situations/Ebola-2019-drc-
[19] Ebola outbreak 2014-2016, West Africa: https://www.who.int/emergencies/situations/ebola-outbreak-2014-2016-West-Africa
[20] https://en.wikipedia.org/w/index.php?title=2002%E2%80%932004_SARS_outbreak

Figure 9. Number of extracted cases vs. time for 4 case studies: MERS-Cov in Saudi Arabia (Panel a), Ebola cases in the Democratic Republic of Congo (Panel b), SARS cases in China (Panel c), and Ebola cases in Guinea (Panel d). These four case studies correspond to the Disease Outbreak Names - Country pairs reported in bold in Table 2.



Source: Consoli et al., 2025 [31].

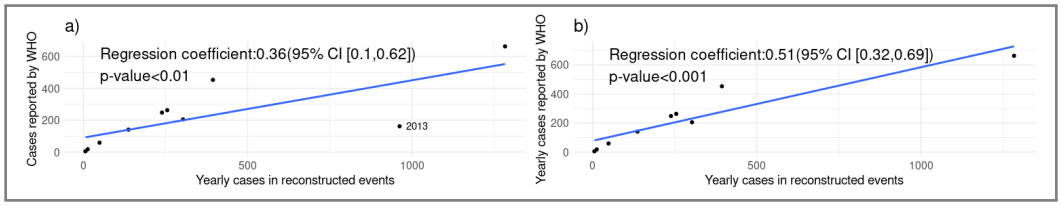
number of cases vs the number of cases from the reconstructed events, together with the best-fit regression line, the 95% CI of the regression coefficient and the corresponding p-value. Panel a shows the results for all data, while Panel b depicts the results obtained by excluding year 2013. As it can be observed, the regression coefficient between the reconstructed events and the WHO data is significant (Panel a: 95% CI of regression coefficient $[0.10, 0.62]$, p-value< 0.01) and increases considerably (Panel b: 95% CI of regression coefficient $[0.30, 0.69]$, p-value< 0.001) when data from year 2013 is not considered. Considering 2013 was the onset of the epidemic, it is plausible that WHO DONS might have included potential cases along with confirmed ones. This could result in an overestimation of the reconstructed cases compared to the confirmed ones, resulting in a lower regression coefficient if data from year 2013 are taken into consideration.

Our open-access JRC Data Catalogue[35] contains the complete eKG knowledge graph in both RDF/XML and Turtle formats, along with the raw extractions from which the knowledge graph has been derived, in CSV format. We enable the interested community to access and use the produced data and ontology under Creative Commons Attribution 4.0 International license (CC BY 4.0[21]).

All the other ontologies mentioned in our work that have been used to construct and map the eKG knowledge graph are also available under CC BY 4.0 license under their sources. These ontologies provide structured entities and terms that are relevant to the infectious disease domain, geographical entities, and formal ontological structures, which are incorporated into the eKG.

[21] CC BY 4.0: https://creativecommons.org/licenses/by/4.0/

Figure 10. WHO yearly number of cases vs the number of cases from the reconstructed events, together with regression line: including year 2013 (Panel a) and without year 2013 (Panel b). Since 2013 was the epidemic's onset, the inclusion of potential and confirmed cases by WHO DONS could lead to a lower correlation coefficient due to overestimated reconstructed cases. Correlation and p-values are reported in the top right corner of each panel.



Source: Consoli et al., 2025 [31].

To perform queries to eKG, a user can download and import the entire knowledge graph in one of the available formats into an OpenLink Virtuoso triplestore instanced on a local machine. After downloading and installing Virtuoso, users should access the server conductor interface through a web browser. They will navigate to the quad store upload section for linked data. By creating a new graph, users can configure it by specifying a name and uploading the entire RDF/XML or Turtle file to the newly created knowledge graph. Users can verify if the dataset has been successfully populated by executing queries using the Virtuoso SPARQL query interface. For convenience, in addition to the repository data, eKG is also available for users to be interrogated in an OpenLink Virtuoso instance that we have instantiated at https://api-vast.jrc.service.ec.europa.eu/sparql/, and explored in a structured and intuitive way via semantic exploration services.

As an example, consider the Nipah virus - India report, having id: *don-record2740* and namespace for the data resources: *http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601*; the RDF information related to this individual can be obtained from the endpoint with the SPARQL query:

```
SELECT *
WHERE {
<http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/don-record2740>
?p ?o}
```

Simply put, this query selects all data triples with properties ?p and objects ?o, matching the resource "Nipah virus - India" where *don-record2740* id is the subject. For each result statement, a link to its corresponding URL in the Virtuoso Faceted Browser[22] is also returned, allowing further navigation. The Faceted Browser allows the exploration of the KG by querying the endpoint with free text search patterns. This returns a list of literal property values or labels that match the text pattern specified as input. The result set can then be narrowed down by filtering for specific entity types or property values, allowing for navigation of the target sub-graph structure. For example, the *don-record2740* can be visualised with the Faceted Browser[23] as illustrated in Figure 11, where users can explore the RDF information of the Nipah virus - India report in a structured and intuitive way, providing a classification of the data in various dimensions. In Figure 12, it is also reported its LodView representation where one can exploit and navigate the DON entity, seeing for example the related outbreak information, the report categorisation as

[22] https://api-vast.jrc.service.ec.europa.eu/fct/

[23] https://api-vast.jrc.service.ec.europa.eu/describe//?url=http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601/don-record2740

*IDO:surveillance_process*and *OWL:Class*, and downloading the raw data in one of the available formats, e.g. rdf+xml or ld+json, among others. Finally, Figure 13 shows the LodLive graph representation of the RDF data of the *don-record2740* report relatively to the other outbreak events present in the KG. For instance, this allows the user to automatically expand the relations of a selected resource, calculate inverse and *owl:sameAs* relations, store images during navigation, and eventually geo-localise the browsed data as points on a map.

In the following boxes, we showcase some further examples of queries for developers and researchers that can be run on the knowledge graph of epidemiology information. To avoid repetitions, consider all the following SPARQL queries to be prefaced with the following standard headers:

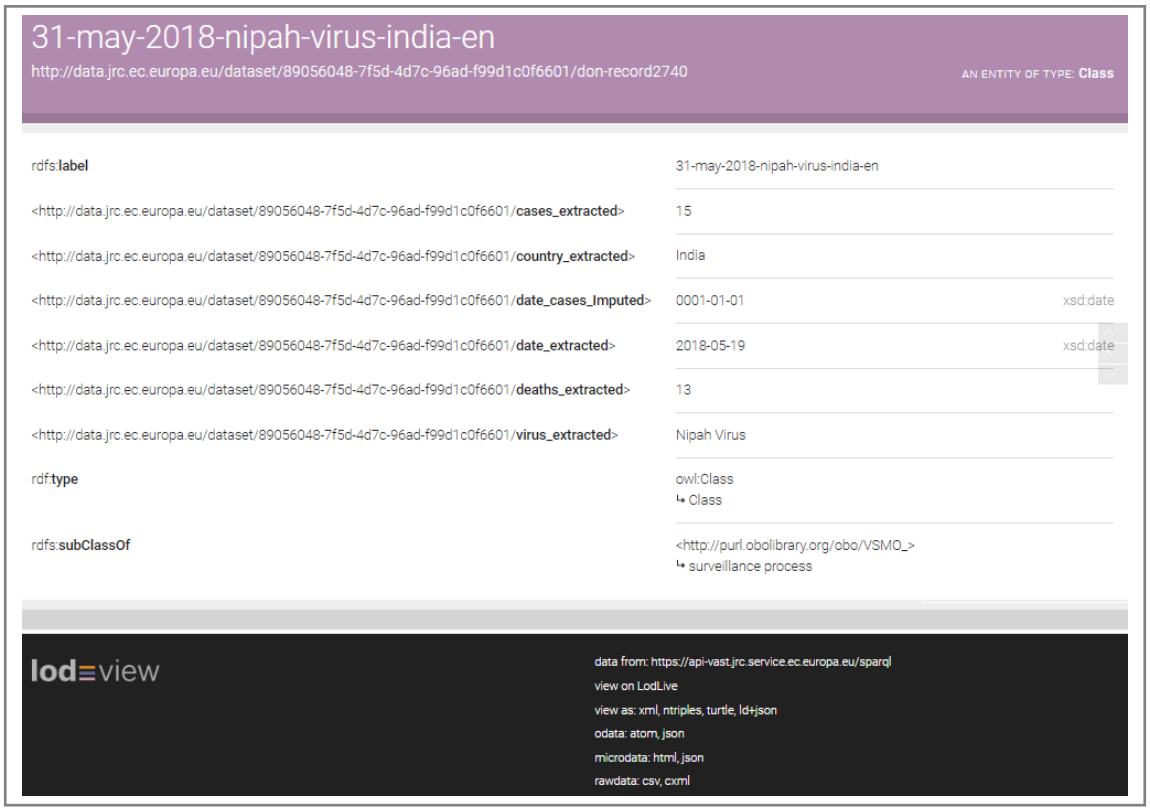
```
PREFIX data:
<http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-
                                                    f99d1c0f6601/>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterm: <http://purl.org/dc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

Figure 11. Faceted Browser showing the extracted Nipah virus - India information.

Source: Consoli et al., 2025 [31].

Figure 12. Content negotiation of the Nipah virus - India report with LodView.



Source: Consoli et al., 2025 [31].

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

Suppose one wants to know all the events related to “Nipah virus” outbreak:

```
SELECT ?event
WHERE {?event data:virus_extracted ?label .
        FILTER (?label = “Nipah Virus”)}
```

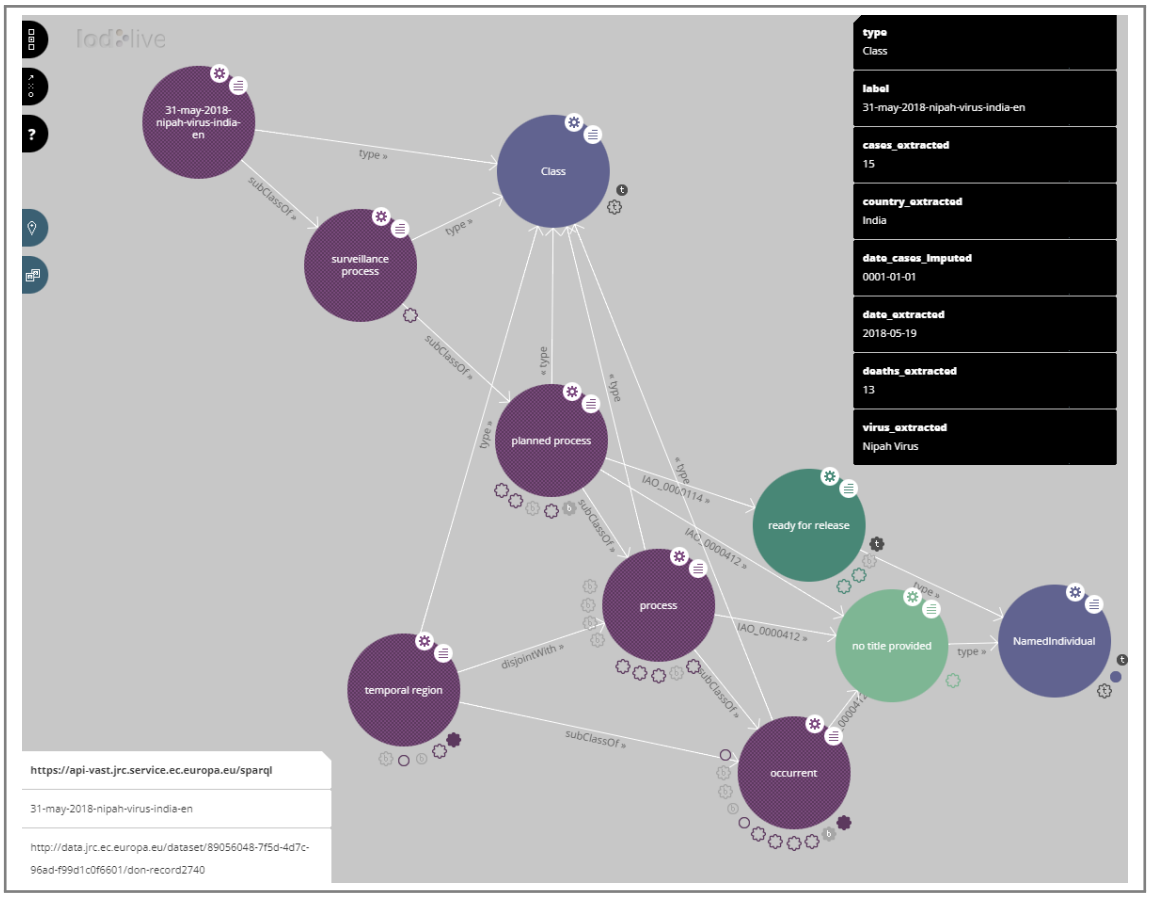
The same but with a more flexible regular expression in the SPARQL query:

```
SELECT ?event
WHERE {?event data:virus_extracted ?label .
        FILTER regex(str(?label), “nipah”, “i”)}
```

The following SPARQL query finds all the outbreaks that occurred in “Italy”:

```
SELECT ?event
WHERE {?event data:country_extracted ?label .
        FILTER (?label = “Italy”)}
```

Figure 13. Graph visualisation of Nipah virus - India in LodLive.



Source: Consoli et al., 2025 [31].

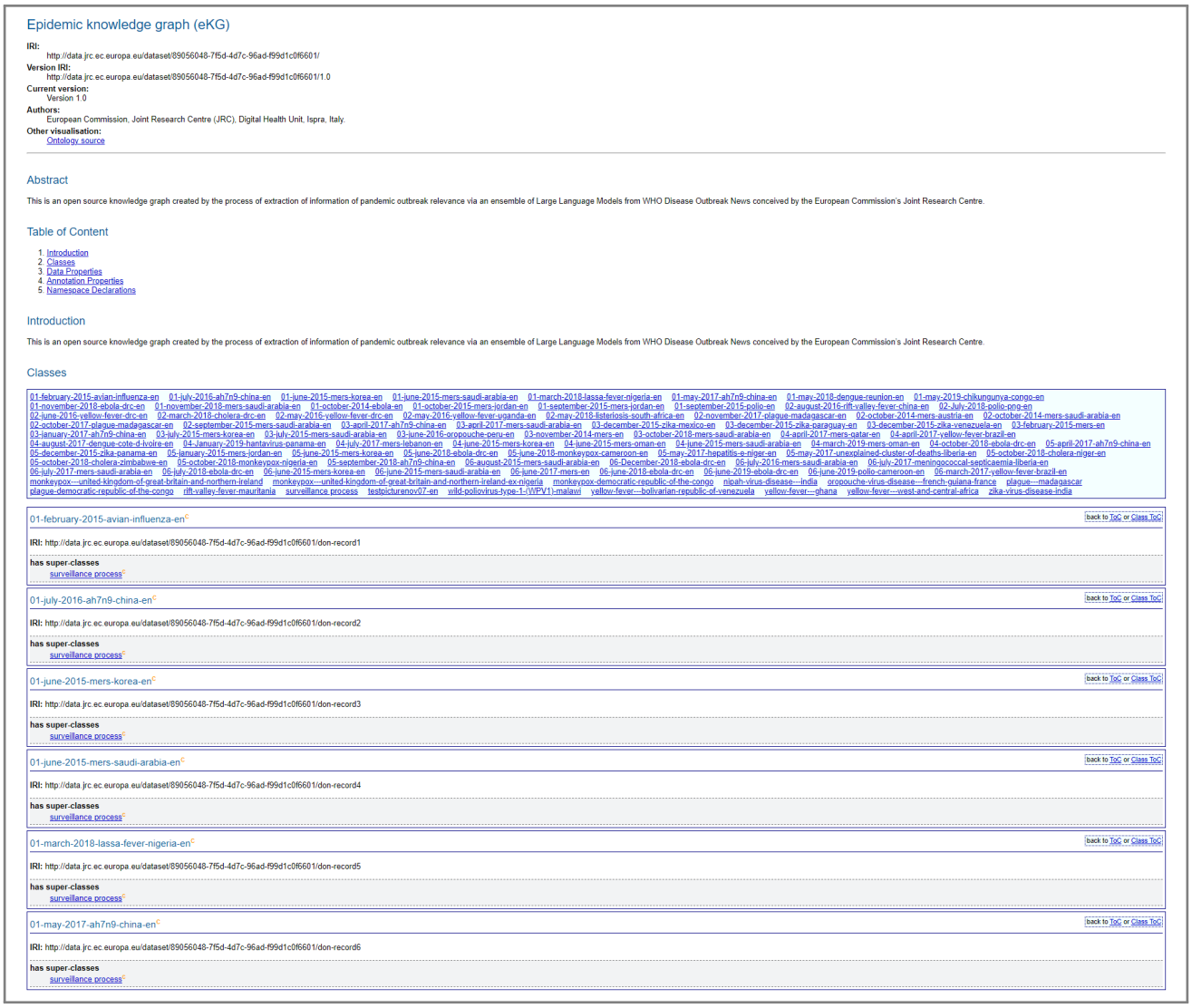
Lastly, the following SPARQL query retrieves all the outbreaks occurring in “Italy” in the year “2017”, specifying the name of the disease and the number of confirmed cases:

```
SELECT ?event ?outbreak ?cases
WHERE {?event data:country_extracted ?label .
?event data:date_extracted ?date .
        FILTER (?label = “Italy”) .
        FILTER (year(?date) = 2017) .
        ?event data:virus_extracted ?outbreak .
        ?event data:cases_extracted ?cases . }
```

These queries represent only some basic examples and there might be various other use cases that can be addressed. For more complex queries, we refer the reader to various tutorials available on the web, and to technical articles in the literature [110, 118] for a more intense deep dive into the SPARQL query language.

The knowledge graph of epidemiological information can also be browsed in a human-readable format using the *Live OWL Documentation Environment (LODE)*, which facilitates exploration for users. LODE provides an easy-to-use interface for viewing general axioms, namespace declarations, named individuals,

Figure 14. An extract from the knowledge graph representation in LODE.



Source: Consoli et al., 2025 [31].

classes, annotations, and object properties within intuitive HTML pages[24]. Figure 14 shows an extract from the KG representation in LODE. The KG can also be accessed through a holistic, force-directed graph layout using *WebVOWL*[25] as in Figure 15. Interaction with these tools allows for customisable visualisations, offering a user-friendly description of the elements of the KG and enhancing the user's ability to leverage the epidemiological information within.

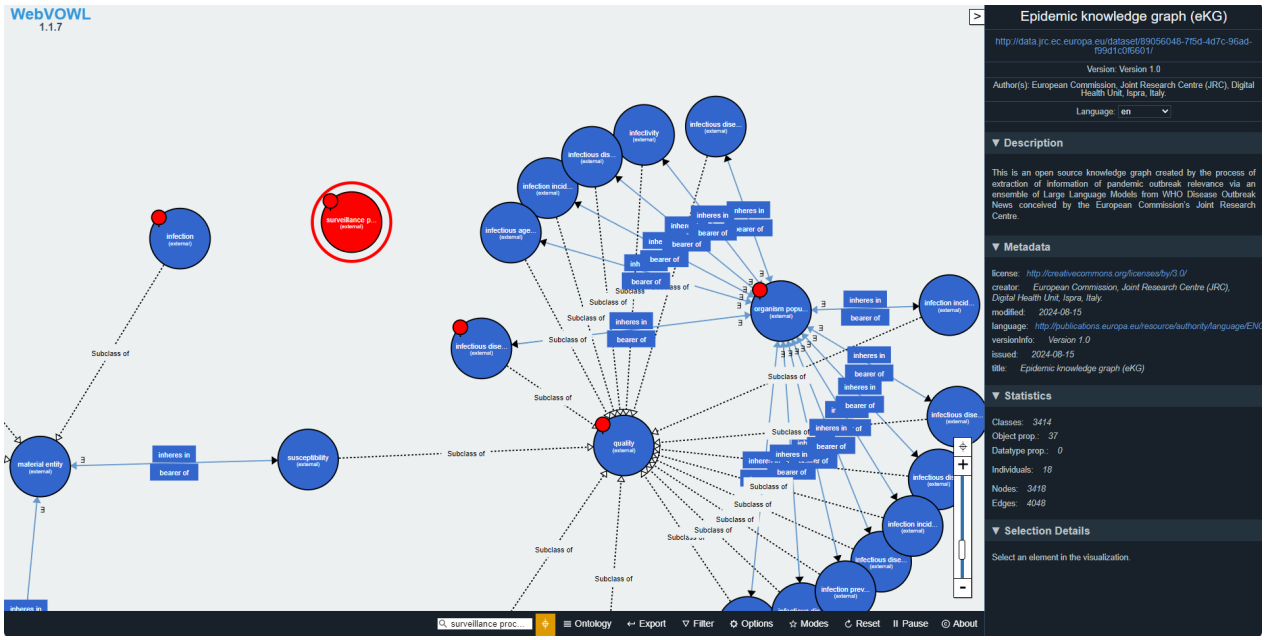
3.4. Discussion

Our work in developing eKG has resulted in a high-quality knowledge graph comprising 613 verified and reliable epidemic events, endorsed by data providers and EU institutions. We have carefully validated

[24] http://wit.istc.cnr.it/arco/lode/extract?url=https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ETOHA/ETOHA-OPEN/epidemicIE-DONs.rdf

[25] https://service.tib.eu/webvowl/#iri=https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ETOHA/ETOHA-OPEN/epidemicIE-DONs.rdf

Figure 15. A snapshot of the WebVOWL visualisation.



Source: Consoli et al., 2025 [31].

these events to ensure a consistent and accurate representation of domain knowledge. Our ongoing efforts utilise semantic technologies and knowledge graphs to improve the exploration and understanding of epidemiological data. The semantic services being developed enable users to explore this knowledge graph, providing an evolving view of extracted outbreaks from DONS via interactive visualisation dashboards. This approach enhances more timely and scalable data collection and epidemic analysis, thereby supporting better-informed public health decision-making. Moving forward, we will focus on refining the epidemiological knowledge graph model, establishing a dynamic pipeline for continuous updates, and implementing entity linking. This will align the knowledge graph with standard biomedical ontologies, facilitating metadata standardisation and improving interoperability across datasets.

4. Comparing In-Context Learning and Fine-Tuning Approaches

The challenges and opportunities provided by generative AI in infectious disease epidemiology [85] are catalysing significant initiatives, such as the "Advanced Technology for Health Intelligence and Action IT System (ATHINA)" by the EC Health Emergency Preparedness and Response Authority (HERA), the "Epidemic Intelligence from Open Sources (EIOS)" initiative [123] of WHO, and the recently launched U.S. "Biothreats Emergence, Analysis and Communications Network (BEACON)" system [17]. These initiatives aim to improve the timely detection of disease outbreaks worldwide by leveraging alternative data sources, such as curated epidemic news articles and social media, alongside advanced technologies for epidemiological surveillance, to address the delays often encountered in official statistics [105].

The application of LLMs in public health surveillance has predominantly focused so far on specific areas, such as symptoms' extraction during the COVID-19 pandemic [128, 18]. However, a gap remains in systematically exploring the applications of these models across a broader spectrum of disease outbreaks. Furthermore, while LLMs have been employed in various epidemic intelligence tasks, there is limited research on the performance obtained by different prompt engineering techniques [137, 150], such as in-context learning [19, 44] and model fine-tuning [5, 43], to optimise their performance for epidemiological data extraction.

In this chapter, we address these gaps by presenting a comprehensive study on the application of generative AI and LLMs for extracting epidemiological information from WHO DONs. Specifically, we explore and compare in-context learning and fine-tuning approaches across various popular LLMs to enhance the effectiveness of structured data extraction. Our study includes a detailed evaluation of model performance using ROUGE metrics, providing insights into the robustness and scalability of these AI techniques in operational settings. Given the vast amount of data, such as the 0.5 million articles processed daily by systems such as EIOS [123], effective and accurate processing is crucial for outbreak detection beyond the scope of WHO DONs.

To facilitate working with the high volume of unstructured data in WHO DONs, Carlson et al. [24] introduced the WHO DON Database, a retrospective collection that organises and curates structured information manually extracted from historical WHO DONs reports. According to its developers, this database compiles information from numerous WHO reports, including data up to the entire year of 2019 and comprising 3,338 records overall, converting the unstructured narrative texts into a systematically analysable data set [24]. To enhance its utility for global health research, they established a standardised framework for sharing epidemiological information, making WHO DONs a more valuable resource for researchers and public health surveillance experts.

The creation of the WHO DONs curated database involved extracting key epidemiological details from the WHO DONs, including outbreak type, geographical location, case counts, and response measures, to enable more accessible and detailed analyses [24]. This structured format allows researchers and public health experts to conduct longitudinal studies on disease emergence patterns and to assess the effectiveness of public health responses over time.

Our study leverages the WHO DONs curated database to assess, validate, and train advanced generative AI techniques, aiming to enhance the extraction and analysis of epidemiological information. The goal is to underscore the potential of AI-driven methods for structured epidemiological information extraction using the WHO DONs curated database as an established benchmark of outbreak-related data.

4.1. Methods

In recent years, prompt engineering [137, 150] has emerged as a critical technique for enhancing the performance of LLMs. It involves crafting specific input prompts that guide the model in generating desired outputs [137]. This approach leverages the background knowledge of LLMs to perform tasks with minimal task-specific data, making them highly adaptable and efficient [150]. Among the various prompt engineering strategies, in-context learning [19, 44] and chain-of-thought [139] are prominent. In-context learning involves providing the model with a few examples of the task within the input prompt, allowing it to infer the task requirements from these examples [19, 44]. This learning capability enables LLMs to generalise from limited data effectively. The chain-of-thought approach can further enhance model reasoning by structuring prompts that guide the model through a logical sequence of steps, improving its ability to solve complex reasoning and mathematical problems [139].

Given that LLMs are inherently designed to process natural language, we transformed the structured epidemiological data from WHO DONs into textual prompts that encapsulate the following epidemiological features: disease classification (*DiseaseLevel1, DiseaseLevel2*), geographical location (*Country, ISO, OutbreakEpicenter*), numbers of cases and deaths (*CasesTotal, CasesSuspected, CasesProbable, CasesConfirmed, Deaths*), and temporal information including outbreak start (*OutbreakStartYear, OutbreakStartMonth, OutbreakStartDay*), outbreak detection (*OutbreakDetectionYear, OutbreakDetectionMonth, OutbreakDetectionDay*), verification (*OutbreakVerificationYear, OutbreakVerificationMonth, OutbreakVerificationDay*), and end dates (*OutbreakEnd, OutbreakEndYear, OutbreakEndMonth, OutbreakEndDay*).

This prompt-based approach ensures compatibility with the models' instruction-tuning background and enables the application of generative and discriminative language modelling techniques to the information extraction task of disease outbreak events. By expressing structured inputs in natural language, we bridge the gap between tabular data and text-based generative AI models, thereby facilitating their deployment in this epidemiological setting.

To explore the potential of instruction-tuned LLMs for epidemiological information extraction, we selected representative popular models with different characteristics. Specifically, we used *LLaMA* [134, 135, 58] and *Mistral* [69, 86], two widely adopted general-purpose models; a LLaMA variant, namely *MedLlama*, and *Meditron*, two LLMs specialised in biomedical text; *T5* [111], a versatile encoder-decoder architecture known for its strong generalisation capabilities across natural language processing (NLP) tasks; and *Qwen* [143], a robust decoder-only model characterised by strong reasoning capabilities, particularly in complex problem-solving scenarios. In our LLMs choice, we limited to non-commercial models only, to ensure efficient utilisation of resources while maintaining the flexibility to customise and fine-tune these models for our specific epidemiological information extraction tasks. In addition, the size of the models in terms of number of parameters was also conditioned by the available hardware for this study, further influencing our selection of LLMs to ensure compatibility and optimal performance. In particular, the computations for our study leveraged the JRC Big Data Analytics Platform [122] infrastructure and were conducted specifically on a Linux cluster system featuring an Ubuntu 22.04.5 LTS operating system, powered by an Intel Xeon Platinum 8470 CPU architecture with 208 CPUs, 1TB of RAM and 8 H100 NVIDIA GPUs for accelerated processing.

4.1.1. In-Context Learning

In-context learning (iCL) [19] enables LLMs to perform a wide range of NLP tasks, such as classification, question-resolution, and structured information extraction, with minimal manual effort. In the context

of epidemiological information extraction, iCL allows models to dynamically adapt to new outbreak data formats or domain-specific terminologies by simply presenting relevant annotated examples in the input prompt. This facilitates rapid prototyping and deployment of extraction systems without the need for costly and time-consuming supervised fine-tuning.

In our analysis, we compared the iCL capability of different popular models to identify optimal iCL configurations that maximise extraction accuracy while maintaining computational efficiency. First, we processed each epidemiological report with a preliminary summarisation step. If it exceeded the context length value for the input of the LLM at hand, we splitted the text in pieces and summarised it with the following prompt:

> *Given some epidemiological TEXT in triple backtick (“‘ ”’), summarise it focusing on any aspects that are relevant to the reported disease outbreak, place and time of the infectious disease outbreak occurred, and the derived contingencies.*
> *Never invent or guess data.*
> *Write no explanations or notes .*
>
> *“‘{ TEXT }”’.*

Then, we used the following prompt to guide the iCL process to extract epidemiological information for the report:

> *Given some TEXT in triple backtick (“‘ ”’) extract disease outbreak information from the given text and format it as JSON.*
> *Use “None” for missing information.*
> *Never invent or guess data.*
>
> *{ few-shot examples }*
>
> *Now extract information from this TEXT in the same JSON format:*
>
> *“‘{ TEXT }”’.*

The prompt begins by instructing the model to extract disease outbreak information from the given piece of text, encapsulated within triple backticks (“‘), and to format the extracted information as a JSON object. This structured format ensures clear and consistent data representation, facilitating subsequent analysis and integration into epidemiological databases. Please note that to handle incomplete data, the prompt specifies using “None” for any missing information, thus maintaining data integrity and preventing the model from inventing or guessing details. Following these initial instructions, the prompt included few examples that demonstrated the desired input-output mappings. These examples served as contextual guides, allowing the model to understand the task requirements and generalise from these examples.

Finally, the prompt directed the model to apply the same extraction and formatting logic to a new text, denoted [TEXT], ensuring that the model consistently adhered to the format and rules described in the prompt. This approach leverages the model’s instruction-tuned capabilities while allowing for dynamic adaptation to specific information extraction tasks without modifying the model’s underlying parameters.

4.1.2. Fine-Tuning

The exploration of LLMs fine-tuning in our epidemiological context derives from the inherent limitations of general-purpose language models when confronted with tasks demanding strict domain-specific precision. Although LLMs leverage billions of parameters trained on vast text corpora, their learned knowledge might require refinement for specialised extraction tasks. In extracting epidemiological data from WHO DONs reports, the challenge involves identifying all the distinct structured features including disease classifications, geographical indicators, outbreak epicentres, confirmed cases and deaths. Fine-tuning addresses this by updating internal parameters to acquire domain-specific behaviours [5].

The decision to explore model fine-tuning rather than relying exclusively on in-context learning was based on the requirements of the specific task (i.e., extracting epidemiological features from WHO DONs). In-context learning conditions models on prompts containing a few examples without modifying weights, offering computational efficiency [44]. However, for applications requiring consistent structural output, parameter modification might provide some advantages. Fine-tuning explicitly trains model weights against a large amount of structured input-output pairs, encoding formatting rules and consistency requirements directly into learned representations [109]. This ensures robust structural adherence that prompt-based methods cannot reliably guarantee for outputs requiring complex, rigid structural compliance.

Our fine-tuning process relied on high-quality instruction data instances derived from the WHO DONs curated database. Structured epidemiological data were transformed into instruction-formatted input-output pairs, where inputs comprised textual reports, and outputs consisted of JSON objects with the extracted features. The instruction-input-output prompt used for the training was the following:

> *Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.*
>
> *### Instruction:*
> *Given some TEXT in triple backtick (“‘ ”’) extract disease outbreak information from the given text and format it as JSON.*
> *Return a list containing one JSON object per outbreak mentioned. Always return a list of JSON objects, even for single outbreaks.*
> *Use “None” for missing information. If no outbreak information is found, return an empty list [].*
> *Never invent or guess data.*
>
> *### Input:*
> *“‘{ TEXT }”’.*
>
> *### Output:*
> *{ [JSON objects] }.*

To ensure robustness, we stratified the WHO DONs curated database for cross-validation [33] using five folds of it. In addition, to tackle challenges related to computational resource demands, memory requirements, and training duration, we adopted Parameter-Efficient Fine-Tuning (PEFT) methodologies [43]. Among PEFT approaches, we selected Low-Rank Adaptation (LoRA) [149] for its ability to minimise trainable parameters while introducing no additional inference latency during deployment[1].

[1] Our specific LoRA configuration employed a rank of $r = 16$ and scaling factor of $\alpha = 16$, balancing parameter efficiency and adaptation capacity, and a dropout rate of 0.05 to prevent overfitting during adaptation.

Finally, in our training implementation we used the HuggingFace *Trainer* class[2] with standardised optimisation parameters [141]. We also employed the paged AdamW optimiser with learning rate $\eta = 1 \times 10^{-5}$, utilised 8-bit precision training, and implemented gradient accumulation with factor 4 to handle batch processing effectively. Automatic device mapping (device_map="auto") was utilised to optimise hardware resource allocation on the 8 H100 NVIDIA GPUs available in our cluster on the JRC Big Data Analytics Platform. The logging and evaluation strategy was adaptive, with steps adjusted according to the number of training instances for detailed tracking. Model selection was based on validation accuracy as the primary metric, with early stopping triggered after six consecutive evaluations without improvement.

The obtained fine-tuned open-source models (LoRA adapter weights) are made available via Hugging Face at https://huggingface.co/AI4PH/EpiLLaMA-3.3-70B, https://huggingface.co/AI4PH/EpiQwen2.5-7B, and https://huggingface.co/AI4PH/EpiMistral-7B.

4.2. Results

In the following we present our computational analysis comparing the effectiveness of various iCL models and fine-tuning techniques for the task of epidemiological information extraction. In our experiments, model performance was evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores [87, 10]. Although originally developed for summarisation tasks, these metrics are applicable to our text-to-graph epidemiological information extraction framework by treating each extracted information as a single-sentence summary representation. Our method could be reconducted to a text-to-graph task because it involved mapping unstructured text inputs to structured graph-like outputs, where each element of extracted epidemiological data represented nodes and relationships within a graph structure. Specifically, we serialised each extracted graph pattern—comprising nodes such as disease names, countries, and various epidemiological indicators, along with their relational edges—into a string representation. This enabled systematic comparison between model predictions and the gold-standard graph patterns annotated in the WHO DONs curated database, where each serialised graph functioned as a single-sentence summary for evaluation purposes.

4.2.1. In-Context Learning Analysis

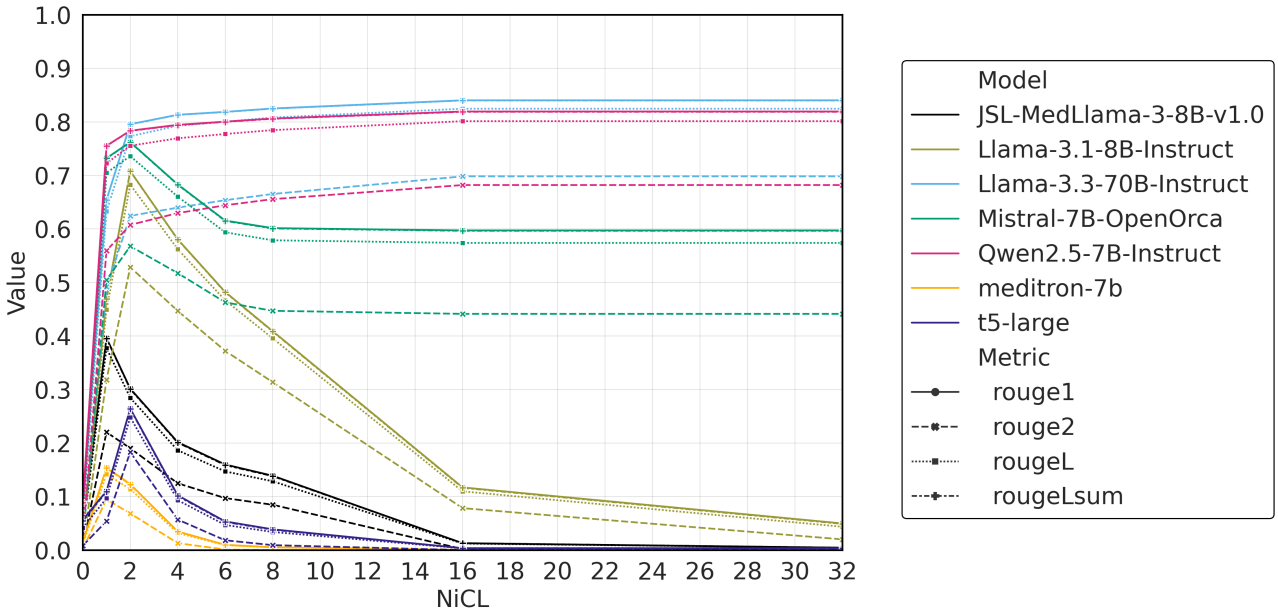
Our analysis systematically explored the impact of varying numbers of examples on extraction accuracy, ranging from 0 to 32 examples. Specifically, we evaluate eight distinct iCL settings: 0, 1, 2, 4, 6, 8, 16, and 32 examples provided in the input prompt. This progressive evaluation allows us to assess how each model's extraction capability evolves with increasing contextual guidance, thereby identifying optimal configurations that balance computational efficiency with performance. For each configuration, we computed ROUGE metrics to quantify the alignment between model predictions and gold-standard annotations from the WHO DONs curated database.

Figure 16 presents the ROUGE scores (Rouge-1, Rouge-2, Rouge-L, and Rouge-Lsum) obtained by each iCL model across all the examples. The x-axis represents the number of in-context examples (NiCL), ranging from NiCL=0 to NiCL=32, while the y-axis displays the corresponding ROUGE score values.

As shown in the figure, we can see inconsistent behaviour across models. Some LLMs exhibit a strong positive correlation between ROUGE scores and the number of in-context learning examples, whilst for others performance decreases after a few examples. For all models, the most substantial performance increase occured in the low iCL regime, moving from NiCL=0 to NiCL=4, approximately. Beyond around

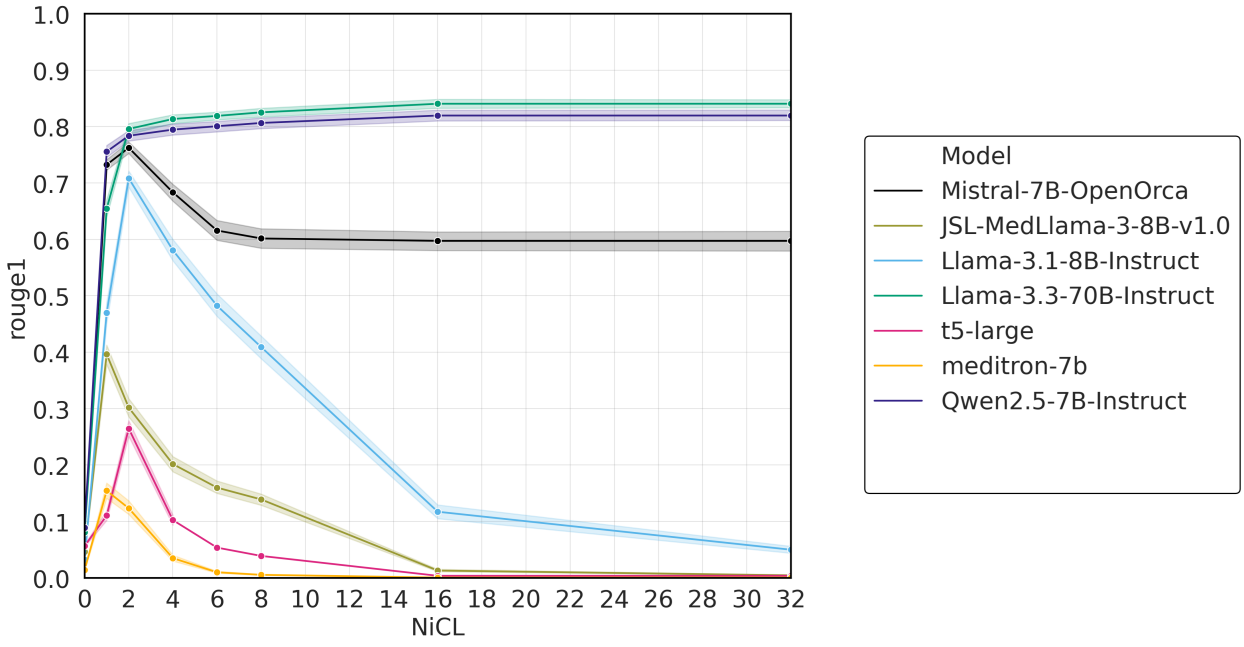
[2] https://huggingface.co/docs/transformers/en/main_classes/trainer

Figure 16. ROUGE scores obtained by each iCL model at increasing number of examples (NiCL).



Source: JRC.

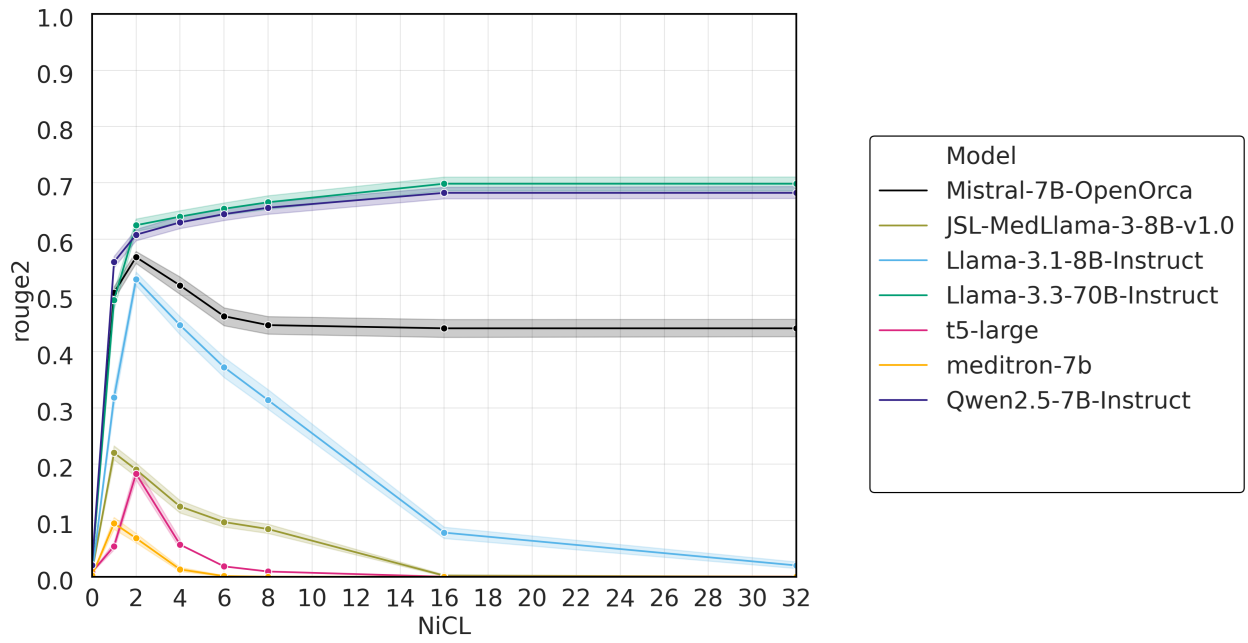
Figure 17. Rouge-1 score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL).



Source: JRC.

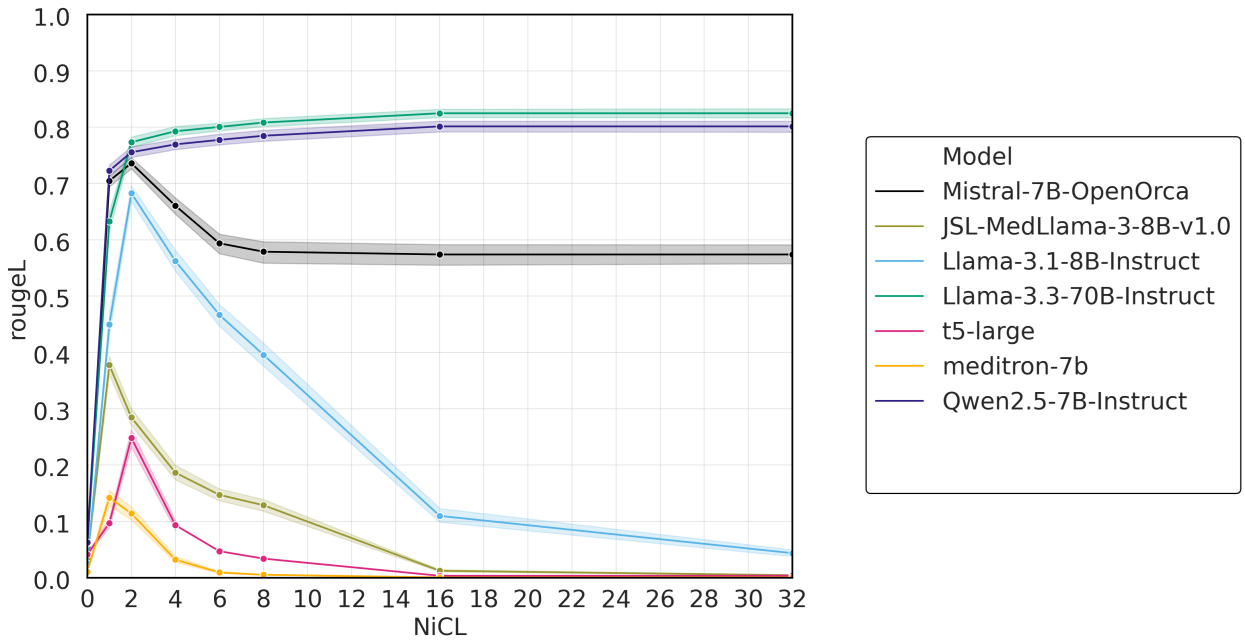
NiCL=8, the curves began to flatten, showing a clear phenomenon of *diminishing returns* [62] where additional shots yield only marginal or noisy gains. This saturation aligned with theoretical scaling laws, sug-

Figure 18. Rouge-2 score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL).



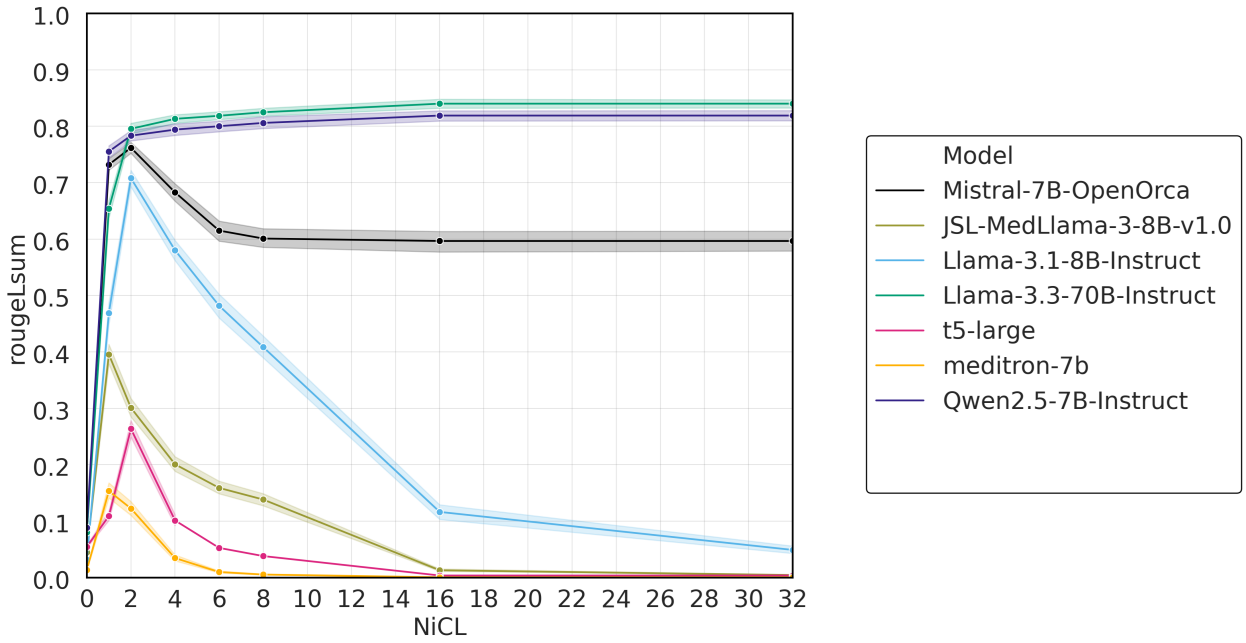
Source: JRC.

Figure 19. Rouge-L score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL).



Source: JRC.

Figure 20. Rouge-Lsum score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL).



Source: JRC.

gesting that, while iCL initially approximates a Bayes optimal learner [8], its efficacy significantly deteriorates in scenarios with many examples. In addition, while LLMs have the ability to take long contexts as input, their performance reaches often its highest value when relevant information is at the beginning or end of the input context, and can significantly decrease when relevant information is in the middle of long contexts (i.e. *context overflow*), even for explicitly long-context models [88]. In particular, while the LLaMA 3.3-70B and Qwen 2.5–7B Instruct models experienced a significant performance plateau, all other models, including the LLaMA 3.1-8B, MedLLaMA 3-8B, and Meditron-7B, consistently had lower performance, resulting in a complete failure in the extracted epidemiological information as NiCL approached larger values. The exception to this trend was the Mistral-7B-OpenOrca model, which, despite an initial decline after NiCL=4, eventually stabilised at a plateau with Rouge scores of around 0.6 as NiCL increases. Detailed per-metric analyses with 99% confidence intervals are provided in Figures 17–20. These figures provide a granular examination of each ROUGE metric's behaviour across the in-context learning configurations, with shaded bands representing 99% confidence intervals computed from the distribution of individual model predictions.

The consistent ranking of models across metrics, combined with non-overlapping confidence bands between top-tier models (LLaMA 3.3-70B, Qwen 2.5-7B Instruct) and lower-performing alternatives, confirms that the observed performance differences are statistically significant rather than artifacts of sampling variability. These detailed uncertainty quantifications reinforce our recommendation of NiCL=16 as the optimal iCL configuration for LLaMA 3.3-70B and Qwen 2.5-7B Instruct models, where performance stabilises with high confidence before the onset of diminishing returns.

4.2.2. Fine-Tuning Analysis

Building upon the insights gained from the iCL analysis, we selected the most promising LLMs to undergo parameter-efficient fine-tuning using the LoRA methodology described in the Methods section. Specifically, we chose LLaMA 3.3-70B and Qwen 2.5-7B Instruct, which demonstrated superior performance in the iCL experiments, and additionally included Mistral-7B-OpenOrca to extend our portfolio of fine-tuned models and assess whether fine-tuning could substantially improve its relatively modest iCL performance. For clarity and consistency throughout this analysis, we refer to the fine-tuned versions of these models as EpiLLaMA 3.3-70B, EpiQwen 2.5-7B, and EpiMistral-7B, respectively.

Table 3. Rouge-1 scores (mean $\pm$ standard deviation) obtained by each fine-tuned model.

Fold #	EpiMistral-7B	EpiQwen 2.5–7B	EpiLLaMA 3.3-70B
1	0.884 ± 0.051	0.922 ± 0.062	0.934 ± 0.051
2	0.899 ± 0.049	0.856 ± 0.073	0.915 ± 0.049
3	0.923 ± 0.042	0.942 ± 0.046	0.952 ± 0.042
4	0.903 ± 0.047	0.936 ± 0.054	0.950 ± 0.047
5	0.885 ± 0.039	0.932 ± 0.042	0.937 ± 0.040
Overall	0.899 ± 0.046	0.918 ± 0.055	0.937 ± 0.046

Source: JRC.

Table 4. Rouge-2 scores (mean $\pm$ standard deviation) obtained by each fine-tuned model.

Fold #	EpiMistral-7B	EpiQwen 2.5–7B	EpiLLaMA 3.3-70B
1	0.841 ± 0.062	0.875 ± 0.077	0.896 ± 0.062
2	0.848 ± 0.061	0.767 ± 0.072	0.855 ± 0.062
3	0.884 ± 0.058	0.903 ± 0.046	0.917 ± 0.059
4	0.867 ± 0.055	0.896 ± 0.063	0.921 ± 0.055
5	0.826 ± 0.049	0.877 ± 0.057	0.889 ± 0.049
Overall	0.853 ± 0.057	0.864 ± 0.063	0.896 ± 0.058

Source: JRC.

The results obtained in terms of ROUGE scores across all five folds and their aggregated overall performance are systematically presented in Tables 3 through 6. Overall, all fine-tuned models achieve consistently high performance across evaluation metrics, with scores exceeding 0.85 in all cases, indicating strong capability in extracting structured epidemiological information. EpiLLaMA 3.3-70B performs best across all metrics, demonstrating both high accuracy and strong structural coherence, while EpiQwen 2.5-7B closely follows with competitive performance despite its smaller size. EpiMistral-7B, although ranking slightly lower, still achieves robust results, confirming its effectiveness in resource-constrained scenarios. Performance remains relatively stable across folds, with some variability observed for EpiQwen

2.5-7B, but without affecting overall trends. The consistency between ROUGE-1, ROUGE-2, and ROUGE-L scores indicates that models not only capture relevant content but also preserve the structural relationships within the extracted information. Furthermore, the near-identical values of ROUGE-L and ROUGE-Lsum suggest that the extracted outputs maintain coherent structure regardless of evaluation granularity. Taken together, these results confirm the reliability and generalisability of the fine-tuned models for epidemiological information extraction.

Table 5. Rouge-L scores (mean $\pm$ standard deviation) obtained by each fine-tuned model.

Fold #	EpiMistral-7B	EpiQwen 2.5–7B	EpiLLaMA 3.3-70B
1	0.876 ± 0.052	0.914 ± 0.064	0.927 ± 0.052
2	0.891 ± 0.051	0.844 ± 0.075	0.904 ± 0.051
3	0.914 ± 0.054	0.934 ± 0.044	0.945 ± 0.043
4	0.892 ± 0.048	0.926 ± 0.055	0.940 ± 0.048
5	0.870 ± 0.041	0.921 ± 0.043	0.927 ± 0.041
Overall	0.889 ± 0.049	0.908 ± 0.056	0.928 ± 0.047

Source: JRC.

Table 6. Rouge-Lsum scores (mean $\pm$ standard deviation) obtained by each fine-tuned model.

Fold #	EpiMistral-7B	EpiQwen 2.5–7B	EpiLLaMA 3.3-70B
1	0.875 ± 0.053	0.913 ± 0.063	0.927 ± 0.052
2	0.885 ± 0.050	0.843 ± 0.072	0.905 ± 0.053
3	0.912 ± 0.044	0.933 ± 0.045	0.946 ± 0.044
4	0.889 ± 0.048	0.926 ± 0.053	0.941 ± 0.047
5	0.871 ± 0.040	0.922 ± 0.042	0.927 ± 0.049
Overall	0.887 ± 0.047	0.908 ± 0.055	0.929 ± 0.049

Source: JRC.

4.3. Discussion

This work compared in-context learning (iCL) and fine-tuning approaches for extracting epidemiological information from outbreak reports, with the aim of improving public health surveillance workflows.

The iCL experiments show that performance improves with the number of examples up to a certain point, after which it plateaus. This behaviour is consistent across models and reflects inherent limitations in leveraging longer contexts. Among the evaluated models, larger architectures such as LLaMA 3.3-70B and Qwen 2.5-7B Instruct consistently achieved the best performance, with diminishing returns observed beyond approximately NiCL=16. This suggests that moderate prompt sizes are sufficient to capture most of the achievable gains while keeping computational costs manageable. In contrast, domain-specific models did not demonstrate clear advantages over general-purpose models in the iCL setting.

Fine-tuning, however, leads to substantial and consistent performance improvements across all evaluated models. The results indicate that parameter adaptation is more effective than prompt-based learning for this task, enabling more accurate and stable extraction of structured epidemiological information. Additionally, fine-tuning reduces the performance gap between models of different sizes, showing that smaller models can achieve competitive results when properly adapted.

From an operational perspective, these findings highlight a trade-off between flexibility and performance. In-context learning offers a lightweight and adaptable solution without the need for training, while fine-tuning provides superior accuracy and reliability at the cost of additional computational resources. The choice between these approaches should therefore be guided by deployment constraints and performance requirements.

Overall, the results demonstrate that large language models can effectively support epidemiological information extraction, particularly when fine-tuned for the task. At the same time, existing surveillance systems remain essential, and AI-based approaches should be understood as complementary tools that can enhance, rather than replace, current epidemic intelligence workflows.

5. AI-Generated Health Threat and Disaster Storylines from Global News with Large Language Models and Retrieval-Augmented Generation

Natural hazards and extreme events pose escalating challenges for global disaster risk management (DRM), not only because of their increasing frequency and severity but also due to the dynamic and cascading effects they can produce. These include climate-related hazards such as hurricanes, floods, and wildfires, as well as geological hazards like earthquakes and landslides—each capable of severely disrupting human populations, infrastructure, and ecosystems. International frameworks such as the Sendai Framework for Disaster Risk Reduction and the United Nations Office for Disaster Risk Reduction (UNDRR) Hazard Information Profiles (HIPs) [66] have advanced understanding of individual hazards. However, significant knowledge gaps remain in systematically representing interactions between hazards and the socioeconomic vulnerabilities that amplify their impacts [37, 54, 155]. These challenges become particularly acute when multiple hazards occur in sequence or combination, producing compound or cascading impacts that existing disaster data systems struggle to capture [130, 131, 51]. Initiatives such as "Early Warnings for All" (EW4ALL) [113] underscore the urgency of more integrated and data-informed approaches to hazard monitoring, risk assessment, and response planning.

Recent advances in natural language processing, particularly the emergence of large language models (LLMs) [136, 19], open new opportunities to address these challenges by extracting insights from unstructured sources. In disaster contexts, where traditional datasets are often delayed, incomplete, or unavailable, textual sources such as news reports, field assessments, and social media posts can provide valuable complementary information [81, 142, 67, 63, 14]. These texts capture qualitative and quantitative details on impacts, exposed populations, local vulnerabilities, coping strategies, and cascading effects—details that fill critical blind spots in existing disaster databases [39, 121, 3, 50]. When paired with retrieval-augmented generation (RAG) [84]—a method that grounds large language model outputs in external knowledge sources—LLMs can transform this unstructured content into coherent, factually grounded narratives, enabling structured representation of events, entities, and relationships at scale. This approach supports deeper exploration of the human–hazard interface and complex risk interdependencies across spatial and temporal scales.

In this study, we present a dataset of over 3,000 global disaster events from 2014–2024, derived from the Europe Media Monitor (EMM), a multilingual media aggregation platform that curates articles from thousands of vetted news sources and includes more than 1.5 billion articles worldwide [124]. We process this corpus using a RAG pipeline combining the power of LLMs for text generation and the semantic extraction of factual knowledge from the unstructured news text. For each disaster event recorded in the Emergency Events Database (EM-DAT) [39], the system produces a structured storyline that summarises key aspects including hazard types, main drivers, direct and indirect impacts, affected sectors and populations, and response actions taken. These storylines are further transformed into Knowledge Graphs (KGs) [68] that capture multi-hazard dynamics, inter-event relationships, and interactions between physical hazards and societal responses.

In recent years, the research community has successfully used KGs for creating semantically enriched representations of data in various domains [9]. KGs are data structures that describe the key entities in a domain and their interconnections, presenting information in a format accessible and interpretable by both machines and humans [61]. The relationship between two entities is typically formalised as a triple in the format of *<subject, predicate, object>* (e.g., *<India–Pakistan heatwave, cause, 87 deaths>* or *<Ebola outbreak, occur, Congo>*). KGs are recognised for their ability to organise data in a predefined and semantically meaningful manner, offering significant support to AI systems across multiple domains [132]. In

our application context, by representing disaster events as interconnected networks of entities and relationships, the obtained dataset provides a means to examine complex risk patterns that are typically underrepresented in classical sources [39, 66].

To assess the reliability and usefulness of the extracted disaster KGs, a small sample was independently reviewed and discussed by domain experts during a validation workshop involving civil protection authorities and disaster analysts. Although this validation covered only a limited set of events, it nonetheless provided high-quality expert feedback that confirmed the relevance, plausibility, and potential value of the extracted information for supporting situational awareness and retrospective analysis. The complete dataset, along with code, extraction templates, and processing workflows, is openly available at this repository, with an accompanying dashboard for interactive exploration hosted at this interface. The data is formatted for interoperability and designed to support reuse across research, policy, and operational contexts.

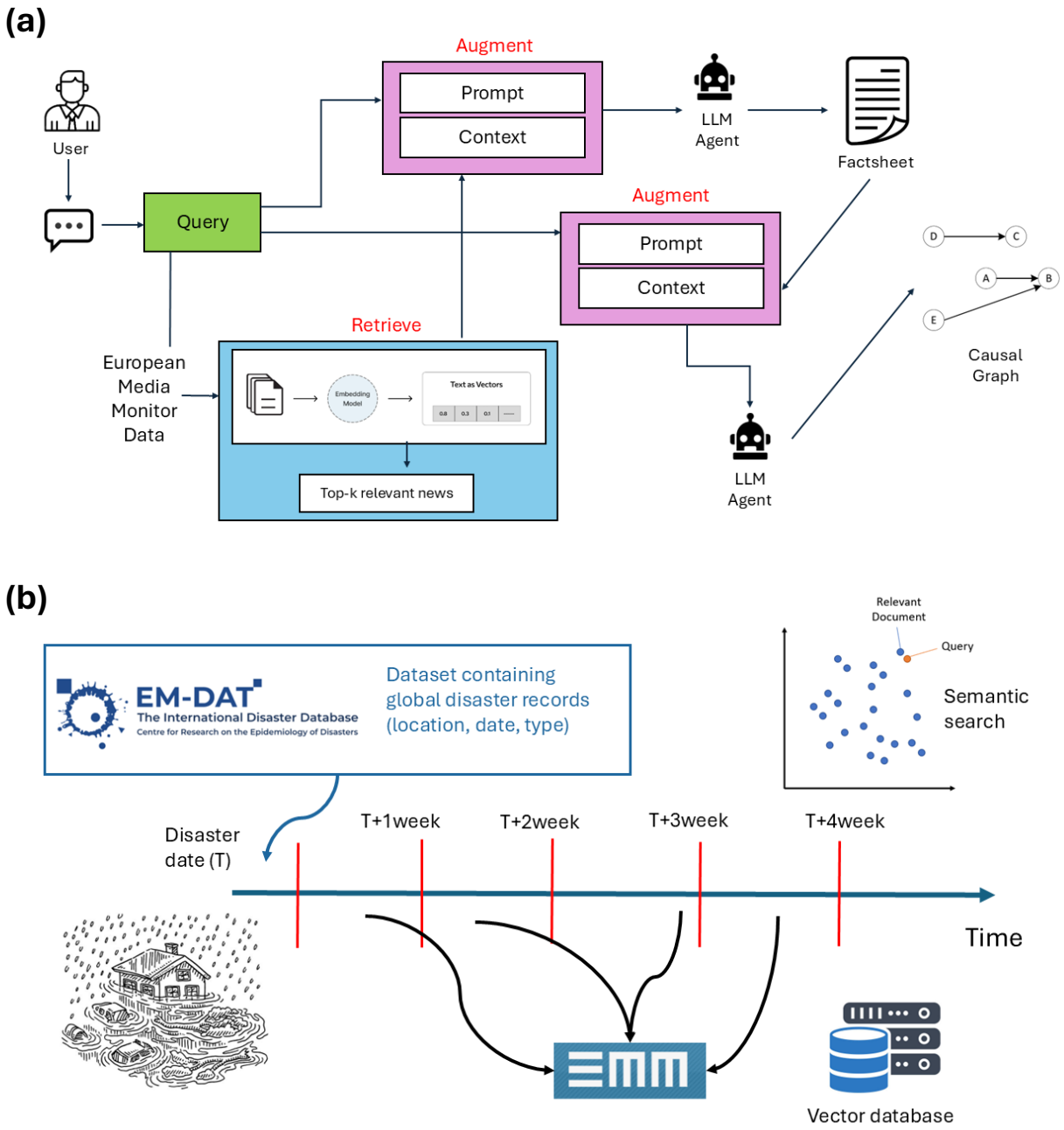
This resource complements existing disaster databases by capturing contextualised relationships between hazards, impacts, and responses that are often absent from conventional event catalogs. Alongside the dataset, we demonstrate a methodology that leverages generative AI to extract and transform information from a large corpus of news articles. This semantic parsing also supports the extraction of place names, which can be linked to geospatial coordinates. In doing so, the resource provides a layer of spatial detail often missing in disaster loss databases, which typically rely on centralised reporting or standardised templates.

Together, these contributions advance a more relational understanding of disaster risk, with applications leveraging HIPs, enhancing early warning systems, and supporting multi-hazard and multi-impact scenario modeling. By transforming diverse textual sources into an well-organised, reusable format, the dataset provides an openly available resource designed for interoperability and reuse in line with the *FAIR* (*F*indable, *A*ccessible, *I*nteroperable, and *R*eusable) data principles [60]. Beyond research, potential operational advantages include faster data validation and encoding for loss-and-damage tracking, quicker situational awareness and decision support during emergencies, and improved machine-to-machine data exchange. Future extensions could include more timely and scalable updates and scenario generation, further strengthening the resource's potential for dynamic monitoring and decision-making in disaster risk management.

5.1. Data Sources for Retrieval

We use two complementary sources to support the information retrieval pipeline: the Emergency Events Database (EM-DAT) [39], which provides a list of disaster events, and the Europe Media Monitor (EMM) [124, 125], which serves as the primary source of unstructured textual news for generating disaster narratives and knowledge graphs. EM-DAT, maintained by the Centre for Research on the Epidemiology of Disasters (CRED) at University of Louvain, is a globally recognised repository of disaster records, containing over 30,000 entries from 1900 to the present. In this study, EM-DAT is used exclusively as a reference list of events to guide downstream document retrieval. Specifically, we extract the disaster type (e.g., flood, drought), country, and event start date to construct the semantic search queries. This ensures that the retrieval process is anchored into a curated and internationally recognised catalogue of disaster occurrences, enabling broad and systematic coverage across hazard types and geographic regions. On the other hand, the EMM, developed by the European Commission's Joint Research Centre (JRC), is a large-scale, multilingual news aggregation and analysis platform. It collects up to 500,000 news reports per day from thousands of verified media sources in over 70 languages, amounting to a historical archive of

Figure 21. Pipeline for generating disaster narratives and knowledge graphs from news. (a) Schematic of the two-stage generation process. A predefined query is used to retrieve top-k relevant news articles from the EMM. These are passed to a first LLM that generates a structured storyline summarising the event's key characteristics. A second LLM then converts this storyline into a knowledge graph capturing entities and relationships. (b) Querying and filtering process. For each event, semantic search retrieves the top 20 articles for each of the four weeks following the event start. The resulting set is refined by an LLM that filters out unrelated articles to improve relevance and reduce noise.



Source: Ronco et al., 2026 [114].

approximately 1.5 billion articles to date. EMM includes automated tools for language detection, geolocation, event and entity recognition, clustering, deduplication, and semantic topic categorisation. Although EM-DAT is used here as a reference list to guide retrieval, the methodology is not limited to this catalogue. The same pipeline can be applied to user-defined events—such as a recently reported earthquake or flood—by constructing equivalent queries over the EMM corpus. In this way, the present study both enriches EM-DAT records with contextualised information and demonstrates a more generalisable workflow for transforming unstructured disaster reporting into concise, well-organised, and reusable data.

All EMM documents are embedded with the *intfloat/multilingual-e5-large* model (1024-dimensional) on a small GPU cluster (4$\times$ NVIDIA Tesla T4). Inference runs in FP16 to reduce memory footprint and increase throughput without observable loss in retrieval quality. Before embedding, each article is split with a heuristic pipeline built on LangChain's *RecursiveCharacterTextSplitter* (*chunk_size*=1500, *chunk_overlap*=300, *separators*={\n\n, \n, . , " "}, *add_start_index*=True). We then apply two post processing stages to curb noise from very short, low-informative fragments: (i) *forward* merging of chunks shorter than 300 characters into their successor whenever the resulting length stays below 1500 chars $+0.5 \times 300$ chars overflow, and (ii) a *backward* pass that guarantees the final chunk is at least 200 characters by merging it into the penultimate one if necessary. Each chunk retains provenance metadata (original start index, chunk id, length, and number of merges) to support deterministic back-mapping to the source text. To optimise semantic search efficiency, we first apply a keyword-based filter at the article level. Specifically, we restrict the corpus to articles matching 13 EMM categories related to disasters and health crises (e.g., *OHCHR-A1003-EpidemicsPandemicsInfectiousDiseases*, *CommunicableDisease Outbreaks*, *NaturalDisasters*, *Man-MadeDisasters*, *ChemicalAccidents*, *NuclearSafety*). Each category is defined by multiple keyword combinations that select relevant articles. This filtering reduces the overall article volume by roughly two orders of magnitude. This yields a total of 11,917 sources, contributing between a few hundred and up to a few million articles per year. The year 2020 is by far the most populated, with about five million articles—most likely reflecting the surge of reporting during the COVID-19 pandemic. The most prolific single source contributes approximately 310,000 articles on its own. Embeddings are stored in an OpenSearch vector index backed by FAISS using an HNSW graph (*m*=16, *ef_construction*=100) with inner-product similarity. We enable symmetric quantisation to FP16 (*encoder.name=sq*, *type=fp16*, *clip*=true) to further reduce memory while preserving recall. This setup provides low-latency kNN retrieval over the 1024-D space and serves as the semantic backbone for the downstream RAG stage described above.

5.2. Generation of Storylines and Graphs

Figure 21a provides a schematic overview of the workflow used to generate disaster narratives and corresponding knowledge graphs. For each disaster event identified through EM-DAT, we retrieve relevant English-language news articles from the EMM archive to capture both initial impact reporting and early response coverage. This pipeline condenses and organises high-quality, event-specific information for end-users in DRM and related domains. By combining LLMs with RAG, it extracts semantically relevant content from large-scale news archives and transforms it into interpretable, standardised formats suitable for decision support. EMM's continuous updates, multilingual scope, and rich metadata make it particularly well suited for retrospective disaster reconstruction and open-domain risk analysis, providing a scalable, real-world foundation for LLM-based disaster data generation.

Starting from the disaster metadata provided by EM-DAT—specifically the event's country, start date, and hazard type—we construct semantic queries to retrieve relevant articles. As explained above, EMM's document corpus is embedded into a vector database using chunk-level representations, enabling efficient similarity search across multilingual sources. For each disaster, we issue four weekly queries in the month

following the event's onset, retrieving the top $k = 20$ most relevant articles per query, yielding up to 80 candidate documents per event (Figure 21b). To reduce noise and improve specificity, we apply an LLM-based filtering step that retains only documents directly referencing the disaster in question. On average, approximately 8.7% of initially retrieved articles pass this filter, significantly enhancing contextual precision for downstream generation. The resulting set of news articles serves as the external knowledge context in a RAG framework, grounding subsequent LLM summaries in semantically relevant evidence.

The filtered document set is then provided as context for the first generation stage, where we prompt the *Meta-Llama-3-70B-Instruct* model [134] to generate a disaster storyline or fact-sheet. This open-weight model, developed by Meta AI, comprises 70 billion parameters and is fine-tuned on a broad range of instructional datasets with extensive human feedback [46]. It supports a context window of 128,000 tokens and demonstrates strong performance in producing accurate, coherent, and semantically rich summaries across multiple domains [134]. In our use case, the model is prompted in a zero-shot setting—without additional fine-tuning—to generate summaries that cover key aspects of each disaster event, including core facts, underlying drivers, observed impacts, and reported response actions (see Figure 22a for the full prompt).

In the second stage, the storyline is passed again to the Llama-3-70B-Instruct model, this time with a prompt tailored to extract the <*subject, predicate, object*> triples that define the main relationships contained in the storyline generated in the first iteration [144]. These triples are used to construct simplified and generalisable KGs centered on disaster dynamics. In this step, the model no longer relies on external retrieval but instead uses the internally generated narrative as contextual input, enabling the extraction of relations already synthesised in the previous stage. To ensure consistency across outputs, the relation types are constrained to two predefined categories: "causes" and "prevents", reflecting reinforcing and mitigating effects, respectively [147]. This convention is inspired by causal loop diagrams in systems dynamics and DRM modeling [29, 64]. Recent research has shown that LLMs are particularly effective at generating KGs from text via in-context learning (iCL) [19, 44], which involves providing examples within the prompt to influence the model's output. This approach offers a scalable alternative to traditional graph-based NLP pipelines while preserving semantic meaning across domains [107, 30, 16, 146, 7, 145, 89, 117, 70]. Examples of extracted triples include <*Heavy rainfall* $\longrightarrow$ *causes* $\longrightarrow$ *flash flooding*> or <*Early warning system* $\longrightarrow$ *prevents* $\longrightarrow$ *casualties*>. This step also uses zero-shot prompting[19] with iCL, where a formatted example is included in the prompt to guide the structure and output behavior (see Figure 22b for the full prompt).

Together, this two-stage generation process—organised narratives followed by KGs—leverages the generative capabilities of large-scale LLMs to produce interpretable, scalable outputs for retrospective disaster analysis and decision support. An example output is shown in Figure 23, including a generated storyline and its associated knowledge graph.

5.3. Data Records

The dataset contains storylines and KGs for 3,158 disaster events occurring between 2014 and 2024 across 175 countries, covering 26 distinct disaster types. Each record corresponds to a unique disaster event, following the EM-DAT structure, identified by the column *DisNo.* (Disaster Number). The dataset is a single CSV file named *DisasterStory.csv*, available at https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ETOHA/storylines/DisasterStory.csv. Regular updates are planned on a yearly basis to incorporate new disaster events and improved extraction methods. Each row in the dataset represents a single disaster and includes both standardised metadata and AI-generated analytical content. The storyline for each

Figure 22. Prompts used in the AI pipeline for disaster storyline and knowledge graph generation. (a) Prompt for generating narrative summaries of disaster events using a set of predefined fields. (b) Prompt for extracting KGs using iCL, including an example to guide the generation of subject–predicate–object triples constrained to "causes" and "prevents" relationships.

(a)

"You are an assistant tasked with providing concise and factual updates on disaster events. Based on the context provided, answer the queries with clear and actionable information. If the information is uncertain or not found, recommend verifying with official channels. For incomplete or unknown answers, respond with 'unknown'.
Complete the following factsheet on the wildfire event in Greece:
- *Key information: Quick summary with location and date.*
- *Severity: [Low, Medium, High]*
- *Key drivers: [Main causes of the disaster.]*
- *Main impacts, exposure, and vulnerability: [Economic damage, people affected, fatalities, effects on communities and infrastructure.]*
- *Likelihood of multi-hazard risks: [Chances of subsequent related hazards.]*
- *Best practices for managing this risk: [Effective prevention and mitigation strategies.]*
- *Recommendations and supportive measures for recovery: [Guidelines for aid and rebuilding.]*

Important: Use only the information provided about the event. Do not add any assumptions or external data. If specific details are missing or uncertain, indicate them as 'unknown'. The original question was: What are the latest developments on the wildfire disaster occurred in Greece on October 2024?"

(b)

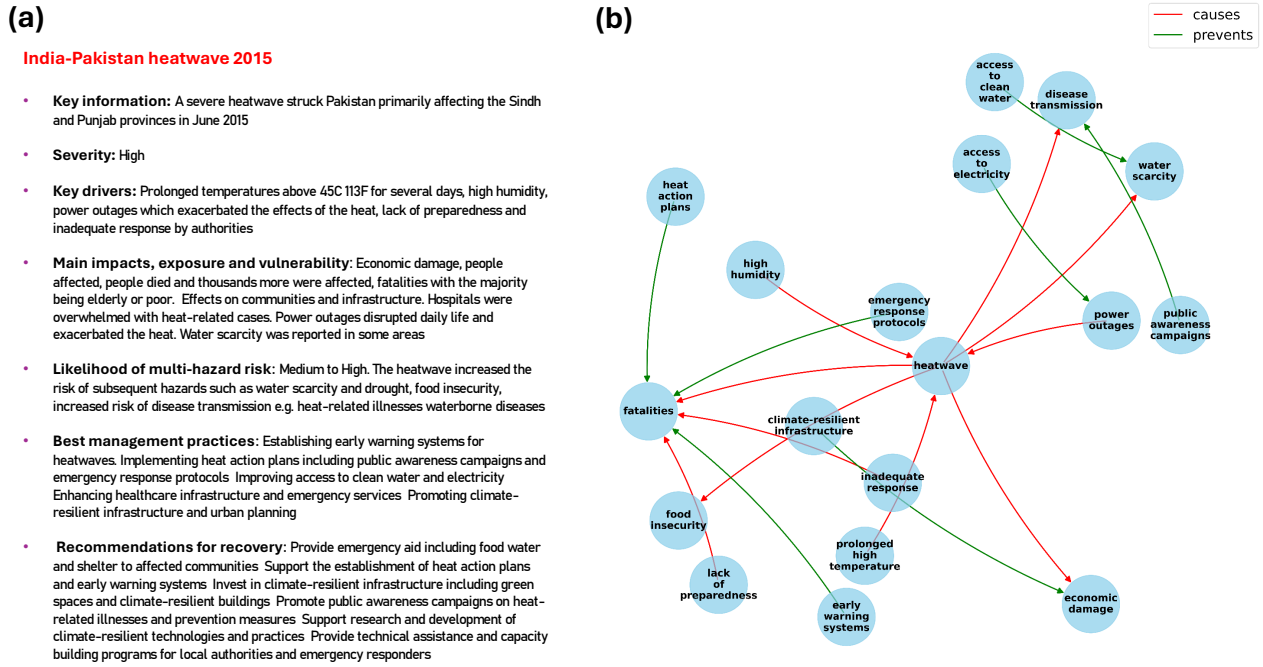
"Create a knowledge graph that captures the main causal relationships presented in the text. Follow these guidelines:
- *Only use two relationship types: 'causes' or 'prevents' (so e.g. either 'A causes B' or 'A prevents B'), no other type of relations are allowed.*
- *Minimize the number of nodes, use short node names with no more than two words if possible.*
- *Focus on drivers and impacts, specifying the type of impact or damage if mentioned explicitly (e.g. 'blackout' or 'infrastructure damage' or 'crop devastation').*
- *Do not use both a factor and its opposite (e.g., "early warning" and "lack of early warning") in the same graph. Represent the relationship with either "causes" or "prevents," not both. Avoid duplicating similar nodes.*

Example:
prompt: In the past few days, several districts of Limpopo province, north-eastern South Africa experienced heavy rain and hailstorms, causing floods and severe weather-related incidents that resulted in casualties and damage. Local authorities warned population, and many were evacuated.
graph: [["heavy rain", "causes", "flooding"], ["hailstorms", "causes", "flooding"], ["flooding", "causes", "damages"], ["flooding", "causes", "casualties"], ["early warning", "prevents", "casualties"]]
prompt:
graph:"

Source: Ronco et al., 2026 [114].

event is distributed across the following columns: *"Key Information"*, *"Severity"*, *"Key Drivers"*, *"Main Impacts, Exposure, and Vulnerability"*, *"Likelihood of Multi-Hazard Risks"*, *"Best Practices for Managing This Risk"*, and *"Recommendations and Supportive Measures for Recovery"*. These columns contain the narrative summaries generated using the Meta-Llama-3-70B-Instruct model, synthesising content primarily from English-language outputs, though the original news sources may be in multiple languages. Each record also includes a *"llama graph"* column, which encodes a list of *<subject, predicate, object>* triples extracted from the source articles using LLM-based information extraction. These KG triples represent key relationships (e.g., *<strong winds, cause, infrastructure damage>*), facilitating reasoning and modelling. The 3,158 records correspond to approximately 50% of all EM-DAT disaster events recorded during the same 2014–2024 period. This partial coverage reflects limitations in the retrieval pipeline, which may fail

Figure 23. Example outputs for the 2015 India–Pakistan heatwave. (a) Disaster storyline generated by Meta-Llama-3-70B-Instruct, organised into predefined thematic sections. (b) Corresponding knowledge graph extracted from the narrative using a second LLM call with iCL, showing causal and preventive relationships constrained to the predicates "causes" and "prevents".



Source: Ronco et al., 2026 [114].

when relevant news is not present in EMM sources, when semantic search does not surface appropriate articles, or when the LLM-based filtering step excludes non-specific or noisy content. To enhance accessibility, an interactive dashboard is available[1] via HuggingFace Spaces at https://huggingface.co/spaces/roncmic/crisesStorylinesRAG, allowing users to explore the dataset by selecting disaster type, country, and event code (*DisNo.*). The platform displays both the storyline and its associated KG in an intuitive and user-friendly interface designed for researchers, practitioners.

5.4. Validation & Usage

While multiple steps in our pipeline—from news retrieval to storyline generation—warrant validation, we focused on the knowledge graphs as a condensed yet information-rich representation of each disaster event. These graphs, built on top of storylines and grounded in retrieved news, offer a fast and interpretable format that captures key relationships between hazards, drivers, impacts, and responses. They thus serve as an effective entry point for expert assessment of the overall quality and factual consistency of the AI-generated outputs. Validating AI-generated content in the DRM domain is inherently challenging. It requires not only technical expertise in hazards and risk modeling but also deep contextual knowledge of sectoral impacts, response phases, and multi-hazard dynamics. As a result, reliable evaluation of such outputs must rely on domain experts with practical and theoretical experience in DRM—resources that are both highly specialised and often difficult to engage.

[1] Only for private users of the HuggignFace *jrc-ai* group.

To facilitate this evaluation, we leveraged the opportunity provided by the "AI for Preparedness"" workshop, hosted by the European Commission in Brussels from June 16 to 18, 2025. The event convened around 30 DRM professionals from national civil protection agencies and related response institutions across Member and Participating States of the Union Civil Protection Mechanism. This rare gathering of operational experts created a unique window for direct validation of our outputs. Experts accessed the knowledge graphs through our Hugging Face dashboard and viewed both the storyline and its corresponding graph to ensure context-based evaluation. Participants reviewed a sample of 34 KGs, selected to reflect high-impact, thematically diverse events and aligned with each expert's domain of expertise. Using a predefined rubric, they rated each graph according to the following criteria: "Fully Correct" (all key nodes and causal/preventive links are accurate), "Mostly Correct" (minor issues but generally sound), "Partially Incorrect" (some key elements wrong or missing), "Incorrect" (major inaccuracies), or "Difficult to Judge" (insufficient context or ambiguity). Experts were given approximately 40 minutes to complete the task, relying on their own knowledge and any external sources they deemed necessary. Evaluations were submitted directly via the dashboard and automatically logged. While the sample represents a small portion of the full dataset, the focused input of this specialised group provides meaningful first insights into the quality of the AI-generated graphs and helps guide future improvements.

Table 7. Distribution of expert assessments by disaster type, following EM-DAT categories [39]. Values represent the percentage of all individual expert ratings in each category. The final column indicates the number of unique disaster graphs evaluated for each type.

Disaster Type	Difficult to Judge	Fully Correct	Incorrect	Mostly Correct	Partially Incorrect	N
Drought	0.0%	0.0%	5.0%	87.4%	7.5%	7
Earthquake	0.0%	0.0%	0.0%	0.0%	100.0%	2
Epidemic	0.0%	70.0%	0.0%	15.2%	14.9%	2
Explosion (Industrial)	0.0%	0.0%	0.0%	24.4%	75.6%	2
Flood	2.3%	36.2%	2.1%	37.3%	22.2%	14
Storm	0.0%	0.0%	0.0%	94.3%	5.7%	3
Volcanic activity	0.0%	0.0%	0.0%	28.3%	71.7%	1
Water	0.0%	0.0%	0.0%	100.0%	0.0%	1
Wildfire	0.0%	0.0%	32.7%	0.0%	67.3%	2

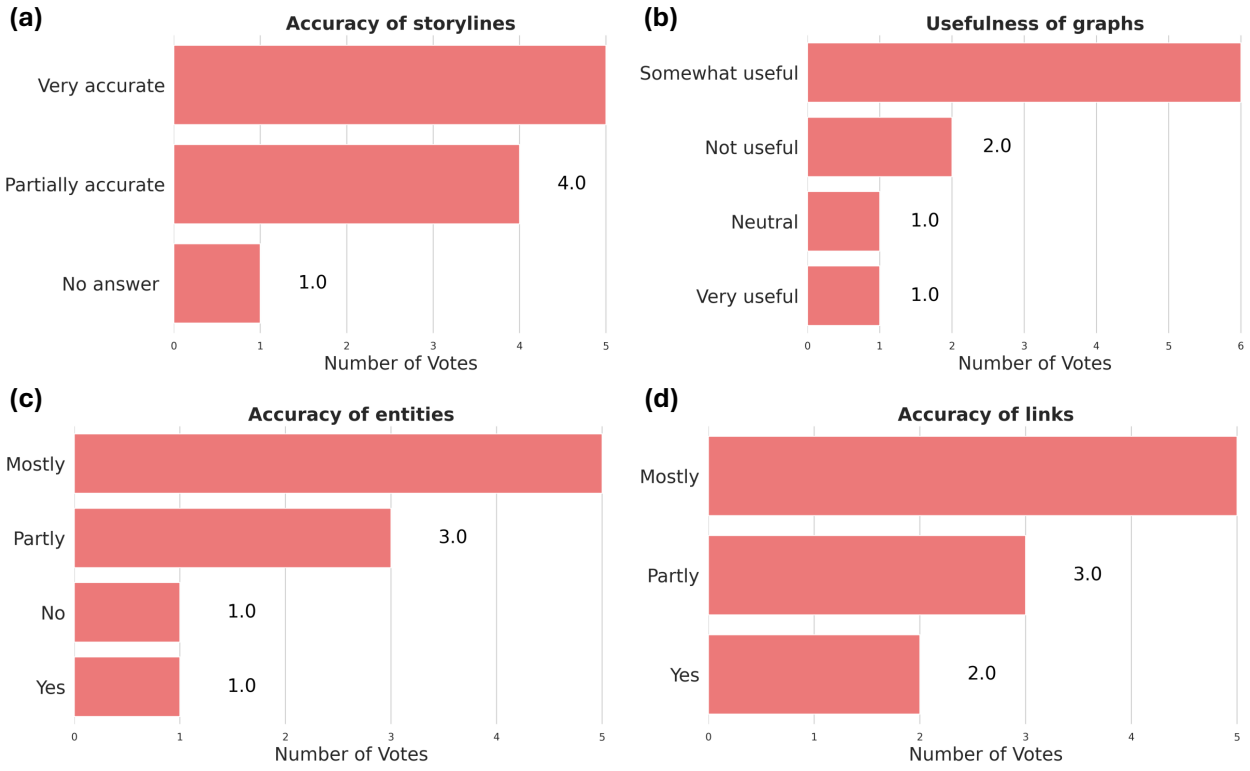
Source: Ronco et al., 2026 [114].

To summarise the validation outcomes, we aggregated expert evaluations across disaster types. Since multiple experts could assess the same event, each event received a set of ratings, allowing us to compute the relative distribution of evaluation categories per disaster type. Table 7 reports the percentage of assessments falling into each rating class—Fully Correct, Mostly Correct, Partially Incorrect, Incorrect, and Difficult to Judge—along with the total number of evaluated graphs per hazard type. These aggregated results provide an overview of perceived accuracy across hazards and help identify areas where the AI-generated outputs performed more reliably or raised interpretation challenges. This analysis not only highlights strengths in our current approach but also pinpoints specific areas needing refinement to enhance the reliability and utility of the AI-generated content in the DRM domain.

In addition to direct graph evaluation, we designed a short EU survey to gather broader, overall feedback on the quality and usefulness of the AI-generated disaster storylines. Ten participants completed the survey, representing diverse roles including Civil Protection Managers, DRM Practitioners, Researchers, and Policy Makers. We report results for four targeted questions, each assessed after participants interacted with the HuggingFace dashboard. These included: (1) "How accurate did you find the AI-generated storylines on disaster events?" (2) "How useful did you find the AI-generated knowledge graphs for under-

standing disaster narratives?" (3) "Are the entities in the graph accurately identified and represented?" and (4) "Are the relationships between entities accurately extracted and represented?". Figure 24 summarises responses in four histograms. Main findings include that half of respondents rated the storylines as "Very accurate," more than half found the graphs "Somewhat useful," and approximately half considered both the entities and links to be "Mostly accurate". These results provide encouraging evidence of perceived quality and relevance from an operational DRM perspective, even if based on a relatively small group of experts. Since all data are openly available and we provide an interactive HuggingFace dashboard to facilitate exploration, our approach is intended as a fully open-source model. This setup allows end users to engage directly with the storylines and graphs, potentially enabling larger-scale validation and refinement through broader community interaction.

Figure 24. Results of the survey assessing the AI-generated narratives and graphs. (a) Accuracy of Storylines (options: "Very accurate," "Partially accurate," and "Inaccurate.") (b) Usefulness of the Graphs (options: "Very useful," "Somewhat useful," "Neutral," and "Not useful.") (c) Accuracy of Entities (options: "Yes, all accurately identified and represented," "Mostly, but a few inaccuracies are present," "Partly, some entities are correctly identified while others are not," "No, many entities are inaccurately identified or missing," and "Not sure / I don't have enough information to assess.") (d) Accuracy of Links/Relationships (options: "Yes, all accurately identified and represented," "Mostly, but a few inaccuracies are present," "Partly, some links are correctly identified while others are not," "No, many links are inaccurately identified or missing," and "Not sure / I don't have enough information to assess.")



Source: Ronco et al., 2026 [114].

This dataset can support a range of applications in DRM, including retrospective analysis, multi-hazard risk assessment, scenario generation, and simulation of cascading effects. By transforming large volumes of unstructured news content into structured storylines and KGs, it bridges textual information and formal risk models. The data can complement existing hazard inventories such as EM-DAT, offering narrative and relational detail often absent from conventional disaster databases. Relationships linking drivers, impacts,

and response measures can inform vulnerability assessments, early warning strategies, and integration into system dynamics or Bayesian network models. Practitioners may use the interactive dashboard to explore past events, gaining context-specific insights into secondary effects, compounding drivers, and effective interventions. Researchers in climate adaptation, humanitarian coordination, or environmental security may benefit from the narratives, which capture contextual nuance such as infrastructure, governance, or pre-existing vulnerabilities shaping disaster outcomes. The dataset also provides a unique resource for AI research. To our knowledge, it is the first large-scale collection of disaster-related KGs paired with multi-sentence narratives, enabling benchmarking in text-to-graph and graph-to-text generation, causal reasoning, complex factual summarisation, and multimodal representation learning. It can serve as a testbed for evaluating alignment, coherence, and factual accuracy in generative models grounded in graph-based representations of factual information.

Users should be aware of several limitations. All storylines and KGs are derived from publicly available news sources, which vary in factual reliability, regional coverage, and thematic emphasis. This introduces potential geographic and content biases, including underrepresentation of events in low-income regions or slow-onset hazards. The driving factors identified reflect reported narratives and may not correspond to underlying systemic drivers. The zero-shot LLM-based generation process, while scalable, can produce incomplete or imprecise representations, especially for complex or under-reported events. KGs are simplified by design, capturing only two relation types, and are not intended as comprehensive causal models. The dataset is seeded on EM-DAT and thus inherits its known incompleteness; some disasters are missing [39]. In addition, the completeness and accuracy of EM-DAT records may vary across hazard categories, with some event types being more comprehensively documented than others; therefore, the quality and coverage of the generated storylines and knowledge graphs may partially reflect these underlying differences in the source inventory. Furthermore, although the pipeline has been designed in EM-DAT only, it is worth noting that the methodology is fully generalisable, allowing users to generate storylines and graphs from alternative disaster inventories, news sources, or user-defined sets of events. For high-stakes applications, outputs should be triangulated with additional sources or reviewed by domain experts.

This release is fully open: code, models, prompts, and processed outputs are publicly available, ensuring methodological transparency and enabling adaptation to new sources or contexts. Beyond providing openly accessible data, this work introduces a novel methodology that combines large-scale news retrieval, embedding-based retrieval augmentation, and LLM-driven extraction of disaster knowledge graphs and storylines. While the current dataset represents a first demonstration rather than a definitive disaster knowledge repository, it establishes the feasibility of generating structured, interpretable representations from unstructured news. Future work may include more extensive validation, integration of additional data modalities (e.g., time series, official statistics), multilingual coverage, probabilistic reasoning modules (e.g., Bayesian networks), and regular updates to capture emerging events. Overall, the dataset offers a scalable, interpretable, and reproducible foundation for multi-hazard awareness, disaster risk management research, and AI-driven analysis.

The dataset containing both storylines and KGs is available at https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ETOHA/storylines/DisasterStory.csv. The dataset is released under Creative Commons Attribution 4.0 International (CC BY 4.0) license[2]. As an open resource, the dataset is designed to be extensible.

[2] Creative Commons Attribution 4.0 International (CC BY 4.0) license, https://creativecommons.org/licenses/by/4.0/

6. Open issues and Critical Analysis on AI for Public Health: Challenges, Limitations, and Ethical Imperatives

6.1. Technical and Data-Related Limitations

The application of generative AI in high-stakes fields like health threats surveillance, preparedness and response, is not without significant technical challenges. A primary concern is the phenomenon of "hallucinations," which refers to the models' tendency to generate nonsensical or factually incorrect responses [102]. This issue stems from a lack of true conceptual comprehension; LLMs are fundamentally statistical models that operate on word associations rather than an understanding of the underlying domain [41]. Training on low-quality or unvalidated data, such as public web corpora, can also predispose models to hallucinate [20]. The failure of Meta's scientific LLM, Galactica, which was suspended shortly after its launch due to criticisms of its accuracy, serves as a prominent example of this vulnerability [41].

Another major limitation is the "black box" problem. Generative AI models transform their training data into mathematical representations, making it impossible to trace the origin of a given output [41]. Unlike traditional clinical resources that provide explicit references, LLMs often fabricate sources when prompted to cite their information, which makes it exceedingly difficult for healthcare professionals to trust and validate the output in a medical context [36, 15]. This opacity undermines the transparency and explainability essential for ethical deployment and accountability [77].

Furthermore, LLMs face challenges with slow adaptation and inconsistent outputs. Medical knowledge evolves rapidly, with new research and treatment guidelines emerging constantly. LLMs can struggle to efficiently integrate these new discoveries, potentially rendering them obsolete or even dangerous in a clinical environment where up-to-date information is paramount [21]. Additionally, the intrinsic randomness in their design can lead to inconsistent responses to the same prompt across different sessions, introducing volatility into their performance and complicating their real-world application [41].

Beyond these challenges, an additional limitation concerns scalability and computational requirements. Processing large volumes of incoming textual data, particularly when using large language models or ensemble approaches, can be computationally intensive and may limit the system's ability to scale efficiently. This can have direct implications for timeliness, as delays in processing and information extraction may reduce the effectiveness of the system in timely epidemic surveillance scenarios.

Finally, it should be noted that while unstructured data sources can substantially enhance the timeliness and breadth of epidemic intelligence, their effective operational use should be complemented by primary clinical, laboratory and surveillance data to validate emerging signals and strengthen the reliability of public health decision-making. In addition, while the underutilisation of unstructured data sources is often framed as a technical challenge, institutional and operational factors also play an important role. In many Member States, surveillance systems have historically been designed around mandatory reporting frameworks based on predefined and structured variables, reducing incentives to systematically collect and analyse free-text information. Consequently, the adoption of AI-based approaches for unstructured data analysis should be evaluated through carefully designed pilot studies in operational settings. Regional-scale pilots could help assess not only the added epidemiological value of extracting information from clinical narratives, reports, and other textual sources, but also the organisational changes, computational infrastructure, human expertise, and resource requirements needed for sustainable deployment. Such evaluations would provide important evidence for future integration into national and European surveillance workflows.

6.2. Ethical and Societal Considerations

The use of generative AI in epidemiology introduces profound ethical concerns. A fundamental challenge is data equity and algorithmic bias [77]. AI models require vast datasets for training, yet biomedical and public health surveillance data have historically excluded underrepresented populations, including women, minorities, and low-income communities [41]. When trained on such biased data, these models can produce algorithms that fail to accurately model disease spread or healthcare responses for marginalised groups, thereby reinforcing and exacerbating existing health disparities [102, 77]. The use of secondary data from digital platforms in "digital epidemiology" can also introduce new, systematic biases that are nearly impossible to correct for a priori [94].

The risk of misinformation and public distrust is another critical concern [97]. The COVID-19 pandemic demonstrated how quickly and significantly misinformation can impact public health efforts [77]. Generative AI models are highly vulnerable to repeating and elaborating on false medical information, as evidenced by a study from Mount Sinai that showed that chatbots would confidently generate explanations for non-existent conditions when presented with a fabricated medical term [98]. This vulnerability can lead to a significant erosion of trust in both public health institutions and the technology itself [77, 97].

A key finding from this research is that practical, engineering-level safeguards can make a significant difference in mitigating this risk. The Mount Sinai study found that a simple, one-line warning added to the prompt—reminding the AI that the information might be inaccurate—cut the rate of hallucinations by nearly half [98]. This demonstrates that the solution is not to abandon AI but to engineer tools that are designed to spot dubious input, respond with caution, and ensure human oversight remains central to the process [98].

The future will not be a simple matter of technology adoption; it will require a re-framing of the challenges and a fundamental rethinking of how we integrate these powerful generative AI tools into our existing epidemiology and public health surveillance systems [94, 40, 77].

The path forward hinges on a multi-faceted approach centered on building trust. This requires:

1. *Trust by Design:* Adopting an approach where ethical considerations are engineered into AI systems from the outset [56]. This includes using privacy-preserving techniques like federated learning, employing data anonymisation and encryption, and designing for transparency and explainability to address privacy and accountability concerns [77].
2. *A Human-Centric Model:* Reaffirming that AI systems should serve as collaborative tools, not replacements for human experts. All critical outputs must be subject to a clear human-in-the-loop validation process [94, 30]. The most crucial skills for a future with generative AI will be critical interrogation and ethical judgment, as the focus shifts from finding the right answer to asking the right questions of the technology [94, 20].
3. *Global Collaboration:* Ensuring that AI tools are made widely available to resource-limited regions to prevent deepening global health disparities [77]. This requires not only making the technology accessible but also ensuring that local public health surveillance agencies have the expertise and resources to use them effectively.

While challenges are significant and require careful navigation, the potential for generative AI to enhance our global understanding and management of disease outbreaks is transformative. By prioritising ethical deployment and maintaining human oversight as the ultimate arbiter of public health decisions, we can exploit the power of this technology to build a more resilient, equitable, and data-driven public health surveillance ecosystem.

6.3. Privacy, Security, and Regulatory Compliance

The data-intensive nature of generative AI systems introduces acute privacy and security risks. Models can unintentionally memorise and "regurgitate" sensitive personal data they were trained on, leading to data leakage [56]. The use of AI-powered surveillance for infectious diseases also raises serious concerns, as it may rely on personal and location data from mobile devices and social media [77]. This could lead to governmental or corporate misuse of data, resulting in increased surveillance and restrictions on personal freedoms [77].

Several mitigation strategies can address these challenges. Institutions can opt for local hosting of AI models to maintain greater control over data usage and compliance [127]. Privacy-preserving techniques like federated learning allow multiple entities, such as hospitals, to collaboratively train a model by sharing model updates without sharing the underlying patient data [77]. Additionally, developers can use data anonymisation, encryption, and other security measures to reduce the risk of re-identification and prevent unauthorised access [56].

Table 8 summarises the key challenges and mitigation strategies.

Table 8. Key Challenges and Mitigation Strategies

Challenge	Description/Implication	Mitigation Strategy
Hallucinations	The tendency of models to generate nonsensical or factually incorrect responses due to a lack of conceptual understanding	Use of higher-quality training data; human-in-the-loop validation; built-in warnings [41]
The "Black Box" Problem	The lack of transparency in how outputs are generated, making it impossible to trace the origin of information	Designing systems with version-controlled, auditable, and explainable outputs; human oversight for final validation [71]
Algorithmic Bias/-Data Equity	Models trained on historically biased data may produce algorithms that fail to accurately represent or serve underrepresented populations	Prioritising the collection of diverse and representative data; using advanced data augmentation techniques; generating synthetic data to fill gaps [77]
Privacy & Surveillance Risks	The potential for models to memorise and leak sensitive data and for AI-powered surveillance to be misused	Implementing privacy-preserving techniques like federated learning; using data anonymisation and encryption; opting for local hosting of models [77]
Misinformation	The vulnerability of models to repeating and amplifying false medical information, which can erode public trust	Engineering tools with built-in safeguards to spot dubious input; ensuring human oversight remains central; promoting digital literacy [97]

Source: JRC.

7. Conclusions

This report demonstrates that generative AI, and in particular Large Language Models (LLMs), has significant potential to augment existing public health surveillance systems. In particular, this study highlights the effectiveness of LLMs in extracting and interpreting critical information from diverse sources, including the Epidemic Intelligence from Open Sources (EIOS) system, WHO Disease Outbreak News, and global media reports. The fine-tuned models, particularly EpiLLaMA 3.3-70B and EpiQwen 2.5-7B, demonstrate AI's ability to synthesise vast amounts of unstructured data into actionable insights with remarkable accuracy, achieving ROUGE scores exceeding 0.90 in epidemiological information extraction tasks. The development of the epidemiology knowledge graph (eKG) enables outbreak data to be not only accessible but also semantically interoperable and computationally tractable, reducing the time from outbreak detection to actionable public health response. Rather than replacing established health threats surveillance, preparedness and response infrastructures, these technologies can complement and potentially enhance current capabilities. While improvements in timeliness and scalability were not directly benchmarked in this study, the fine-tuned LLMs approach points in this direction, as it avoids repeated inference steps required by ensemble methods and can be deployed in local environments, potentially enabling more efficient and scalable processing.

Despite the promising results, significant challenges remain, particularly concerning data privacy, algorithmic bias, and the integration of AI with existing health infrastructures. The technical limitations identified—including the "black box" problem, susceptibility to hallucinations, and slow knowledge adaptation—require ongoing research attention and engineering solutions. The ethical dimensions are equally critical: algorithms trained on historically biased data risk perpetuating health disparities, while the potential for surveillance overreach threatens fundamental rights to privacy and autonomy.

The path forward demands a collaborative, multidisciplinary approach that brings together AI researchers, epidemiologists, public health surveillance practitioners, policymakers, and affected communities. Human-in-the-loop validation must remain central to any operational deployment, with AI serving as a powerful augmentation tool rather than a replacement for human expertise and judgment. As demonstrated e.g. by the Mount Sinai study we referred to [98], simple engineering safeguards—such as built-in warnings about potential inaccuracies—can dramatically reduce error rates, proving that practical solutions exist for many of the challenges we face.

Additionally, continued research is necessary to bridge knowledge gaps, particularly in exploring multimodal data fusion approaches that combine text, images, and time-series data, and developing explainable AI methods that increase transparency without sacrificing performance. Future work should also prioritise global equity, ensuring that AI-driven public health surveillance tools are accessible and beneficial to resource-limited regions, thereby preventing the exacerbation of existing global health disparities.

As digital innovation continues to reshape public health landscapes, it is imperative to leverage AI's full potential to build resilient and adaptive health systems. By embracing AI-driven methodologies, public health surveillance policies can be more proactive, data-driven, and responsive to emerging health threats. The integration of AI into surveillance systems, coupled with enhanced data infrastructure and skilled human oversight, positions public health institutions to detect outbreaks earlier, understand their dynamics more comprehensively, and coordinate responses more effectively. This report serves as an initial step toward realising such a vision, advocating for strategic policy adaptations that align with the evolving digital health ecosystem.

References

[1] Abbood, A., Ullrich, A., Busche, R., Ghozzi, S., 'EventEpi-A natural language processing framework for event-based surveillance,' *PLoS Computational Biology*, vol. 16, no. 11, 2020, doi:10.1371/journal.pcbi.1008277.

[2] Abendroth Dias, K., Arias Cabarcos, P., Bacco, F., Bassani, E., Bertoletti, A., et al., *Generative AI Outlook Report - Exploring the Intersection of Technology, Society and Policy*, Publications Office of the European Union, Luxembourg, 2025, doi:10.2760/1109679, URL https://data.europa.eu/doi/10.2760/1109679, jRC142598.

[3] Alencar, P. H. L., Sodoge, J., Paton, E. N., de Brito, M. M., 'Flash droughts and their impacts—using newspaper articles to assess the perceived consequences of rapidly emerging droughts,' *Environmental Research Letters*, vol. 19, no. 7, p. 074048, 2024, doi:10.1088/1748-9326/ad58fa.

[4] Alocci, D., Mariethoz, J., Horlacher, O., Bolleman, J. T., Campbell, M. P., Lisacek, F., 'Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search,' *PLOS ONE*, vol. 10, no. 12, pp. 1–17, 2015, doi:https://doi.org/10.1371/journal.pone.0144578.

[5] Anisuzzaman, D., Malins, J. G., Friedman, P. A., Attia, Z. I., 'Fine-tuning large language models for specialized use cases,' *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, 2025, doi:10.1016/j.mcpdig.2024.11.005.

[6] Antoniou, G., van Harmelen, F., *Web Ontology Language: OWL*, pp. 67–92, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, doi:10.1007/978-3-540-24750-0_4.

[7] Antonucci, A., Piqué, G., Zaffalon, M., 'Zero-shot causal graph extrapolation from text via llms,' *arXiv preprint*, vol. arXiv:2312.14670 [cs.AI], 2023, URL https://doi.org/10.48550/arXiv.2312.14670, xAI4Sci Workshop @ AAAI24.

[8] Arora, A., Jurafsky, D., Potts, C., Goodman, N., 'Bayesian scaling laws for in-context learning,' in: 'Proceedings of the 2nd Conference on Language Modeling (COLM),' Montreal, Canada, Oct. 2025, URL https://arxiv.org/abs/2410.16531.

[9] Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M. E., 'Towards a Knowledge Graph for Science,' in: 'Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics,' WIMS '18, Association for Computing Machinery, New York, NY, USA, 2018.

[10] Auriemma Citarella, A., Barbella, M., Ciobanu, M. G., De Marco, F., Di Biasi, L., Tortora, G., 'Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques,' *Journal of Computational Science*, vol. 87, p. 102571, 2025, doi:https://doi.org/10.1016/j.jocs.2025.102571.

[11] Azamfirei, R., Kudchadkar, S. R., Fackler, J., 'Large language models and the perils of their hallucinations,' *Critical Care*, vol. 27, no. 1, 2023, doi:https://doi.org/10.1186/s13054-023-04393-x.

[12] Baader, F., Horrocks, I., Lutz, C., Sattler, U., *An Introduction to Description Logic*, Cambridge University Press, 2017, doi:https://doi.org/10.1017/9781139025355, URL https://doi.org/10.1017/9781139025355.

[13] Babanejaddehaki, G., An, A., Papagelis, M., 'Disease outbreak detection and forecasting: A review of methods and data sources,' *ACM Transactions On Computing For Healthcare*, vol. 6, no. 2, Feb. 2025, doi:10.1145/3708549, URL https://doi.org/10.1145/3708549.

[14] Balashankar, A., et al., 'Predicting food crises using news streams,' *Science Advances*, vol. 9, p. eabm3449, 2023, doi:10.1126/sciadv.abm3449.

[15] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., Shah, N. H., 'Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review,' *JAMA*, vol. 333, no. 4, 2025, doi:10.1001/jama.2024.21700.

[16] Bertolini, L., Hulsman, R., Consoli, S., Puertas Gallardo, A., Ceresa, M., 'On constructing biomedical text-to-graph systems with large language models,' in: 'CEUR Workshop Proceedings,' vol. 3747, May 26–30 2024, third International Workshop on Knowledge Graph Generation from Text, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece, JRC137500.

[17] Bhadelia, N., 'An Improved Alert System for Emerging Infectious Diseases,' *JAMA*, vol. 333, no. 2, pp. 115–116, 2025, doi:10.1001/jama.2024.22023.

[18] Bizel-Bizellot, G., Galmiche, S., Lelandais, B., Charmet, T., Coudeville, L., Fontanet, A., Zimmer, C., 'Extracting circumstances of covid-19 transmission from free text with large language models,' *Nature Communications*, vol. 16, no. 1, p. 5836, 2025, doi:10.1038/s41467-025-60762-w, URL https://doi.org/10.1038/s41467-025-60762-w.

[19] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al., 'Language models are few-shot learners,' in: 'Advances in Neural Information Processing Systems,' vol. 2020-December, 2020.

[20] Brownstein, J. S., Rader, B., Astley, C. M., Tian, H., 'Advances in Artificial Intelligence for Infectious-Disease Surveillance,' *New England Journal of Medicine*, vol. 388, no. 17, p. 1597 – 1607, 2023, doi:10.1056/NEJMra2119215.

[21] Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., Bressem, K. K., 'Current applications and challenges in large language models for patient care: a systematic review,' *Communications Medicine*, vol. 5, no. 1, p. 26, 2025, doi:10.1038/s43856-024-00717-2.

[22] Callahan, T. J., Tripodi, I. J., Stefanski, A. L., Cappelletti, L., Taneja, S. B., Wyrwa, J. M., Casiraghi, E., Matentzoglu, N. A., Reese, J., Silverstein, J. C., Hoyt, C. T., Boyce, R. D., Malec, S. A., Unni, D. R., Joachimiak, M. P., Robinson, P. N., Mungall, C. J., Cavalleri, E., Fontana, T., Valentini, G., Mesiti, M., Gillenwater, L. A., Santangelo, B., Vasilevsky, N. A., Hoehndorf, R., Bennett, T. D., Ryan, P. B., Hripcsak, G., Kahn, M. G., Bada, M., Baumgartner, W. A., Hunter, L. E., 'An open source knowledge graph ecosystem for the life sciences,' *Scientific Data*, vol. 11, no. 1, 2024, doi:https://doi.org/10.1038/s41597-024-03171-w.

[23] Camarda, D. V., Mazzini, S., Antonuccio, A., 'LodLive, exploring the Web of Data,' in: 'ACM International Conference Proceeding Series,' p. 197 – 200, 2012, doi:10.1145/2362499.2362532.

[24] Carlson, C. J., Boyce, M. R., Dunne, M., Graeden, E., Lin, J., Abdellatif, Y. O., Palys, M. A., Pavez, M., Phelan, A. L., Katz, R., 'The World Health Organization's Disease Outbreak News: A retrospective database,' *PLOS Global Public Health*, vol. 3, no. 1, pp. 1–13, 01 2023, doi:10.1371/journal.pgph.0001083, URL https://doi.org/10.1371/journal.pgph.0001083.

[25] Cavalleri, E., Cabri, A., Soto-Gomez, M., Bonfitto, S., Perlasca, P., Gliozzo, J., Callahan, T. J., Reese, J., Robinson, P. N., Casiraghi, E., Valentini, G., Mesiti, M., 'An ontology-based knowledge graph for representing interactions involving RNA molecules,' *Scientific Data*, vol. 11, no. 1, 2024, doi:https://doi.org/10.1038/s41597-024-03673-7.

[26] CDC, 'Natural Language Processing for Cancer Surveillance,' https://www.cdc.gov/national-program-cancer-registries/data-modernization/natural-language-processing.html, 2024, accessed: September 24, 2025.

[27] Chang, Y.-C., Chiu, Y.-W., Chuang, T.-W., 'Linguistic Pattern–Infused Dual-Channel Bidirectional Long Short-term Memory With Attention for Dengue Case Summary Generation From the Program for Monitoring Emerging Diseases–Mail Database: Algorithm Development Study,' *JMIR Public Health Surveill*, vol. 8, no. 7, p. e34583, 2022, doi:10.2196/34583, URL https://doi.org/10.2196/34583.

[28] Chowell, G., Luo, R., Sun, K., Roosa, K., Tariq, A., Viboud, C., 'Real-time forecasting of epidemic trajectories using computational dynamic ensembles,' *Epidemics*, vol. 30, p. 100379, 2020, doi:https://doi.org/10.1016/j.epidem.2019.100379, URL https://www.sciencedirect.com/science/article/pii/S1755436519301112.

[29] Coletta, V. R., Pagano, A., Pluchinotta, I., Fratino, U., Scrieciu, A., Nanu, F., Giordano, R., 'Causal loop diagrams for supporting nature based solutions participatory design and performance assessment,' *Journal of Environmental Management*, vol. 280, p. 111668, 2021, doi:10.1016/j.jenvman.2020.111668.

[30] Consoli, S., Coletti, P., Markov, P. V., Orfei, L., Biazzo, I., Schuh, L., Stefanovitch, N., Bertolini, L., Ceresa, M., Stilianakis, N. I., 'An epidemiological knowledge graph extracted from the world health organization's disease outbreak news,' *Scientific Data*, vol. 12, no. 1, p. 970, Jun. 2025, doi:10.1038/s41597-025-05276-2, URL https://doi.org/10.1038/s41597-025-05276-2.

[31] Consoli, S., Coletti, P., Markov, P. V., et al., 'An epidemiological knowledge graph extracted from the world health organization's disease outbreak news,' *Scientific Data*, vol. 12, p. 970, 2025, doi:10.1038/s41597-025-05276-2.

[32] Consoli, S., Markov, P., Stilianakis, N. I., Bertolini, L., Gallardo, A. P., Ceresa, M., 'Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models,' in: 'Proceedings of Ninth International Congress on Information and Communication Technology (ICICT 2024),' vol. 1011 LNNS, p. 241 – 252, Lecture Notes in Networks and Systems, 2024.

[33] Consoli, S., Reforgiato Recupero, D., Petkovic, M. (editors), *Data Science for Healthcare - Methodologies and Applications*, Springer Nature, 2019, doi:10.1007/978-3-030-05249-2, URL https://doi.org/10.1007/978-3-030-05249-2.

[34] Cowell, L. G., Smith, B., *Infectious Disease Ontology*, Springer Berlin Heidelberg, 2010, doi:10.1007/978-1-4419-1327-2_19.

[35] [dataset] European Commission, Joint Research Centre (JRC), 'Epidemic Information Extraction from WHO Disease Outbreak News, European Data Portal, Joint Research Centre Data Catalogue,' 2024, doi:https://doi.org/10.2905/89056048-7f5d-4d7c-96ad-f99d1c0f6601, PID: http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601.

[36] Dave, T., Athaluri, S. A., Singh, S., 'ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations,' *Frontiers in Artificial Intelligence*, vol. 6, 2023, doi:10.3389/frai.2023.1169595, URL https://doi.org/10.3389/frai.2023.1169595.

[37] De Angeli, S., Malamud, B., Rossi, L., Taylor, F., Trasforini, E., Rudari, R., 'A multi-hazard framework for spatial-temporal impact analysis,' *International Journal of Disaster Risk Reduction*, vol. 73, p. 102829, 2022, doi:10.1016/j.ijdrr.2022.102829.

[38] Deka, P., Jurek-Loughrey, A., P, D., 'Evidence extraction to validate medical claims in fake news detection,' in: A. Traina, H. Wang, Y. Zhang, S. Siuly, R. Zhou, L. Chen (editors), 'Health Information Science,' pp. 3–15, Springer Nature Switzerland, Cham, 2022.

[39] Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., van Loenhout, J. A. F., Speybroeck, N., 'EM-DAT: the Emergency Events Database,' *International Journal of Disaster Risk Reduction*, vol. 124, p. 105509, 2025, doi:10.1016/j.ijdrr.2025.105509.

[40] Deloitte, 'Generative AI to Reshape the Future of Health Care,' https://www.deloitte.com/us/en/Industries/life-sciences-health-care/articles/generative-ai-in-healthcare.html, 2024, accessed: September 24, 2025.

[41] Deng, J., Zubair, A., Park, Y.-J., 'Limitations of large language models in medical applications,' *Postgraduate Medical Journal*, vol. 99, no. 1178, pp. 1298–1299, 2023, doi:10.1093/postmj/qgad069.

[42] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,' in: 'ACL-HLT 2019 Conference Proceedings,' vol. 1, pp. 4171–4186, 2019.

[43] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., Sun, M., 'Parameter-efficient fine-tuning of large-scale pre-trained language models,' *Nature Machine Intelligence*, vol. 5, no. 3, p. 220 – 235, 2023, doi:10.1038/s42256-023-00626-4.

[44] Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., et al., 'A survey on in-context learning,' 2023.

[45] Du, H., Zhao, Y., Zhao, J., Xu, S., Lin, X., Chen, Y., Gardner, L. M., Yang, H. F., 'Advancing real-time infectious disease forecasting using large language models,' *Nature Computational Science*, vol. 5, no. 6, pp. 467–480, Jun. 2025, doi:10.1038/s43588-025-00798-6, URL https://doi.org/10.1038/s43588-025-00798-6.

[46] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 'The Llama 3 Herd of Models,' 2024, URL https://arxiv.org/abs/2407.21783.

[47] Elastic, 'Traditional AI vs. generative AI: What's the difference?' https://www.elastic.co/blog/traditional-ai-vs-generative-ai, 2024, accessed: November, 2025.

[48] Espinosa, L., Salathé, M., 'Use of large language models as a scalable approach to understanding public health discourse,' *PLOS Digital Health*, vol. 3, no. 10, 2024, doi:10.1371/journal.pdig.0000631.

[49] Espinosa, L., Wijermans, A., Orchard, F., Hohle, M., Czernichow, T., Coletti, P., Hermans, L., Faes, C., Kissling, E., Mollet, T., 'Epitweetr: Early warning of public health threats using Twitter data,' *Eurosurveillance*, vol. 27, no. 39, 2022, doi:10.2807/1560-7917.ES.2022.27.39.2200177.

[50] Firmansyah, H. B., Fernandez-Marquez, J. L., Cerquides, J., Lorini, V., Bono, C. A., Pernici, B., 'Enhancing disaster response with automated text information extraction from social media images,' in: '2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService),' pp. 71–78, 2023, doi:10.1109/BigDataService58306.2023.00017.

[51] Gallina, V., Torresan, S., Critto, A., Sperotto, A., Glade, T., Marcomini, A., 'A review of multi-risk methodologies for natural hazards: Consequences and challenges for a climate change impact assessment,' *Journal of Environmental Management*, vol. 168, pp. 123–132, 2016, doi:10.1016/j.jenvman.2015.11.011.

[52] Gangemi, A., 'Ontology design patterns for semantic web content,' *Lecture Notes in Computer Science*, vol. 3729 LNCS, pp. 262–276, 2005.

[53] Gangemi, A., Presutti, V., 'Ontology design patterns,' in: 'Handbook on Ontologies,' pp. 221–243, International Handbooks on Information Systems, 2009.

[54] Gill, J. C., Malamud, B. D., 'Reviewing and visualizing the interactions of natural hazards,' *Reviews of Geophysics*, vol. 52, pp. 680–722, 2014, doi:10.1002/2013RG000445.

[55] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., 'Detecting influenza epidemics using search engine query data,' *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009, doi:10.1038/nature07634, URL https://doi.org/10.1038/nature07634.

[56] Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., Sikdar, B., 'Privacy and Security Concerns in Generative AI: A Comprehensive Survey,' 2024, doi:10.1109/ACCESS.2024.3381611.

[57] Gong, X., Jiang, J., Duan, Z., Lu, H., 'A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology,' *BMC Bioinformatics*, vol. 19, 2018, doi:https://doi.org/10.1186/s12859-018-2064-y.

[58] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 'The Llama 3 Herd of Models,' 2024, URL https://arxiv.org/abs/2407.21783.

[59] Griset, P., Schafer, V., 'Hosting the world wide web consortium for Europe: From CERN to INRIA,' *History and Technology*, vol. 27, no. 3, p. 353 – 370, 2011, doi:10.1080/07341512.2011.604177.

[60] Heath, T., Bizer, C., 'Linked Data: Evolving the Web into a Global Data Space,' *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, no. 1, pp. 1–121, 2011.

[61] Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A., 'Knowledge graphs,' *ACM Computing Surveys*, vol. 54, no. 4, 2021, doi:10.1145/3447772.

[62] Hong, Z., Ajith, A., Pauloski, J. G., Duede, E., Chard, K., Foster, I., 'The diminishing returns of masked language models to science,' in: 'Proceedings of the Annual Meeting of the Association for Computational Linguistics,' p. 1270 – 1283, 2023, doi:10.18653/v1/2023.findings-acl.82.

[63] Hou, J., Xu, S., 'Near-real-time seismic human fatality information retrieval from social media with few-shot large-language models,' in: 'Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems,' pp. 1141–1147, 2022.

[64] Inam, A., Adamowski, J., Halbe, J., Prasher, S., 'Using causal loop diagrams for the initialization of stakeholder engagement in soil salinity management in agricultural watersheds in developing countries: A case study in the rechna doab watershed, pakistan,' *Journal of Environmental Management*, vol. 152, pp. 251–267, 2015, doi:10.1016/j.jenvman.2015.01.052.

[65] Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., Stoeckert, C. J., Vasilevsky, N. A., Walls, R. L., Zheng, J., Mungall, C. J., Peters, B., 'OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies,' *Database*, vol. 2021, 2021, doi:10.1093/database/baab069.

[66] Jacot des Combes, H., Murray, V., Kirsch-Wood, J., Stevance, A.-S., Elten, M., Hinwood, A., Abdelhakim, D., Fakhruddin, B., Douris, J., Shaw, B., et al., 'Hazard definition and classification review: Technical report (2025),' *United Nations Office for Disaster Risk Reduction*, 2025, doi:10.24948/2025.05.

[67] Jeba, S. M., Aurpa, T. T., Adib, M. R. S., 'From facebook posts to news headlines: Using transformer models to predict post-disaster impact on mass media content,' *Social Network Analysis and Mining*, vol. 14, no. 1, p. 200, 2024.

[68] Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P. S., 'A Survey on Knowledge Graphs: Representation, Acquisition, and Applications,' *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, 2022, doi:10.1109/TNNLS.2021.3070843.

[69] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., Sayed, W. E., 'Mistral 7b,' 2023.

[70] Jiralerspong, T., Chen, X., More, Y., Shah, V., Bengio, Y., 'Efficient causal graph discovery using large language models,' *arXiv preprint*, vol. arXiv:2402.01207 [cs.LG], 2024, URL https://doi.org/10.48550/arXiv.2402.01207.

[71] John Snow Labs, 'How Generative AI Transforms Clinical Data Extraction,' https://www.johnsnowlabs.com/how-can-generative-ai-improve-clinical-data-extraction/, 2024, accessed: September 24, 2025.

[72] Kamran, A. B., Naveed, H., 'GOntoSim: a semantic similarity measure based on LCA and common descendants,' *Scientific Reports*, vol. 12, no. 1, 2022, doi:https://doi.org/10.1038/s41598-022-07624-3.

[73] Keller, M., Freifeld, C. C., Brownstein, J. S., 'Automated vocabulary discovery for geo-parsing online epidemic intelligence,' *BMC Bioinformatics*, vol. 10, no. 1, p. 385, 2009, doi:10.1186/1471-2105-10-385, URL https://doi.org/10.1186/1471-2105-10-385.

[74] Kirillovich, A., Nikolaev, K., 'Adapting the LodView RDF Browser for Navigation over the Multilingual Linguistic Linked Open Data Cloud,' in: '2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2022,' p. 143 – 149, 2022, doi:10.1109/SETIT54465.2022.9875628.

[75] Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M., 'Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP,' *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11685 LNCS, p. 192 – 204, 2019, doi:10.1007/978-3-030-27325-5_15.

[76] Korst, J., Pronk, V., Barbieri, M., Consoli, S., 'Introduction to classification algorithms and their performance analysis using medical examples,' in: S. Consoli, D. R. Recupero, M. Petkovic (editors), 'Data Science for Healthcare: Methodologies and Applications,' pp. 39–73, Springer Nature, 2019.

[77] Kraemer, M. U. G., Tsui, J. L. H., Chang, S. Y., et al., 'Artificial intelligence for modelling infectious disease epidemics,' *Nature*, vol. 638, pp. 623–635, 2025, doi:10.1038/s41586-024-08564-w, URL https://www.nature.com/articles/s41586-024-08564-w.

[78] Krauer, F., Schmid, B. V., 'Mapping the plague through natural language processing,' *Epidemics*, vol. 41, 2022, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142124550&doi=10.1016%2fj.epidem.2022.100656&partnerID=40&md5=b9aabb6e86aa82dfe003734670b707dd.

[79] Kwok, K. O., Huynh, T., Wei, W. I., Wong, S. Y., Riley, S., Tang, A., 'Utilizing large language models in infectious disease transmission modelling for public health preparedness,' *Computational and Structural Biotechnology Journal*, vol. 23, pp. 3254–3257, 2024, doi:https://doi.org/10.1016/j.csbj.2024.08.006, URL https://www.sciencedirect.com/science/article/pii/S2001037024002654.

[80] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J., 'Biobert: A pre-trained biomedical language representation model for biomedical text mining,' *Bioinformatics*, vol. 36, no. 4, p. 1234 – 1240, 2020.

[81] Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q., Li, J., 'Harnessing large language models for disaster management: A survey,' *arXiv preprint arXiv:2501.06932*, 2025, doi:10.48550/arXiv.2501.06932.

[82] Lejeune, G., Brixtel, R., Doucet, A., Lucas, N., 'Multilingual event extraction for epidemic detection,' *Artificial Intelligence in Medicine*, vol. 65, no. 2, pp. 131–143, 2015, doi:https://doi.org/10.1016/j.artmed.2015.06.005, URL https://www.sciencedirect.com/science/article/pii/S0933365715000846, intelligent healthcare informatics in big data era.

[83] Leuba, S. I., Yaesoubi, R., Antillon, M., Cohen, T., Zimmer, C., 'Tracking and predicting U.S. influenza activity with a real-time surveillance network,' *PLoS Computational Biology*, vol. 16, no. 11, 2020, doi:10.1371/journal.pcbi.1008180.

[84] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., et al., 'Retrieval-augmented generation for knowledge-intensive nlp tasks,' *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[85] Lewnard, J. A., Reingold, A. L., 'Emerging challenges and opportunities in infectious disease epidemiology,' *American Journal of Epidemiology*, vol. 188, no. 5, p. 873 – 882, 2019, doi:10.1093/aje/kwy264.

[86] Lian, W., Goodson, B., Wang, G., Pentland, E., Cook, A., Vong, C., "Teknium", 'MistralOrca: Mistral-7B Model Instruct-tuned on Filtered OpenOrcaV1 GPT-4 Dataset,' 2023, huggingFace repository.

[87] Lin, C.-Y., 'ROUGE: A package for automatic evaluation of summaries,' in: 'Text Summarization Branches Out,' Association for Computational Linguistics, Barcelona, Spain, Jul. 2004, URL https://www.aclweb.org/anthology/W04-1013.

[88] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P., 'Lost in the Middle: How Language Models Use Long Contexts,' *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi:10.1162/tacl_a_00638, URL https://aclanthology.org/2024.tacl-1.9/.

[89] Long, S., Schuster, T., Piché, A., 'Can large language models build causal graphs?' *arXiv preprint*, vol. arXiv:2303.05279 [cs.CL], 2023, URL https://doi.org/10.48550/arXiv.2303.05279, peer reviewed and accepted for presentation at the Causal Machine Learning for Real-World Impact Workshop (CML4Impact) at NeuRIPs2022.

[90] Lugo-Robles, R., Garges, E., Olsen, C., Brett-Major, D., 'Identifying nontraditional epidemic disease risk factors associated with major health events from world health organization and world bank open data,' *American Journal of Tropical Medicine and Hygiene*, vol. 105, no. 4, pp. 896–902, 2021.

[91] Madoff, L. C., Woodall, J. P., 'The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail,' *Archives of Medical Research*, vol. 36, no. 6, pp. 724–730, 2005.

[92] Maynard, D., Bontcheva, K., Augenstein, I., 'Natural language processing for the Semantic Web,' *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, no. 2, pp. 1–196, 2017.

[93] McDonald, D. J., Bien, J., Green, A., Hu, A. J., DeFries, N., Hyun, S., et al., 'Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction?' *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 51, 2021, doi:10.1073/pnas.2111453118.

[94] Mesquita, S., Perfeito, L., Paolotti, D., Gonçalves-Sá, J., 'Epidemiological methods in transition: Minimizing biases in classical and digital approaches,' *PLOS Digital Health*, vol. 4, no. 1, p. e0000670, 2025, doi:10.1371/journal.pdig.0000670, URL https://doi.org/10.1371/journal.pdig.0000670.

[95] Miller, G. A., 'WordNet: A lexical database for English,' *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[96] Mondor, L., Brownstein, J. S., Chan, E., Madoff, L. C., Pollack, M. P., Buckeridge, D. L., Brewer, T. F., 'Timeliness of nongovernmental versus governmental global outbreak communications,' *Emerging Infectious Diseases*, vol. 18, no. 7, p. 1184 – 1187, 2012.

[97] Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., Bauer, M., 'Artificial intelligence and increasing misinformation,' *The British Journal of Psychiatry*, vol. 224, no. 2, p. 33–35, 2024, doi:10.1192/bjp.2023.136, URL https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/artificial-intelligence-and-increasing-misinformation/DCCE0EB214E3D375A3006AA69FFB210D.

[98] Mount Sinai, 'AI Chatbots Can Run With Medical Misinformation, Study Finds,' https://www.mountsinai.org/about/newsroom/2025/ai-chatbots-can-run-with-medical-misinformation-study-finds-highlighting-the-need-for-stronger-safeguards, 2025, accessed: September 24, 2025.

[99] Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A., 'Orca: Progressive Learning from Complex Explanation Traces of GPT-4,' 2023.

[100] Neil Otte, J., Beverley, J., Ruttenberg, A., 'BFO: Basic Formal Ontology,' *Applied Ontology*, vol. 17, no. 1, p. 17 – 43, 2022, doi:10.3233/AO-220262.

[101] Ochs, C., Perl, Y., Geller, J., Arabandi, S., Tudorache, T., Musen, M. A., 'An empirical analysis of ontology reuse in BioPortal,' *Journal of Biomedical Informatics*, vol. 71, p. 165 – 177, 2017, doi: https://doi.org/10.1016/j.jbi.2017.05.021.

[102] Omar, M., Brin, D., Glicksberg, B., Klang, E., 'Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review,' *American Journal of Infection Control*, vol. 52, no. 9, pp. 992–1001, 2024, doi:https://doi.org/10.1016/j.ajic.2024.03.016, URL https://www.sciencedirect.com/science/article/pii/S0196655324001597.

[103] Oppenheim, B., Gallivan, M., Madhav, N. K., Brown, N., Serhiyenko, V., Wolfe, N. D., Ayscue, P., 'Assessing global preparedness for the next pandemic: Development and application of an Epidemic Preparedness Index,' *BMJ Global Health*, vol. 4, no. 1, 2019.

[104] Orlandi, F., Graux, D., O'sullivan, D., 'Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF∗,' in: 'Proceedings - 2021 IEEE 15th International Conference on Semantic Computing, ICSC 2021,' p. 233 – 240, 2021, doi:https://doi.org/10.1109/ICSC50631.2021.00049.

[105] Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., et al., 'Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience,' *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 17–21, 2014.

[106] Peng, C., Xia, F., Naseriparsa, M., Osborne, F., 'Knowledge Graphs: Opportunities and Challenges,' *Artificial Intelligence Review*, vol. 56, no. 11, 2023, doi:10.1007/s10462-023-10465-9.

[107] Peng, C., Xia, F., Naseriparsa, M., et al., 'Knowledge graphs: Opportunities and challenges,' *Artificial Intelligence Review*, vol. 56, pp. 13071–13102, 2023, doi:10.1007/s10462-023-10465-9.

[108] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., Weinstein, R. A., 'Using internet searches for influenza surveillance,' *Clinical infectious diseases*, vol. 47, no. 11, pp. 1443–1448, 2008.

[109] Pratap, S., Aranha, A. R., Kumar, D., Malhotra, G., Iyer, A. P. N., S.S., S., 'The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques,' *Natural Language Processing Journal*, vol. 11, p. 100144, 2025, doi:https://doi.org/10.1016/j.nlp.2025.100144.

[110] Pérez, J., Arenas, M., Gutierrez, C., 'Semantics and complexity of SPARQL,' *ACM Transactions on Database Systems*, vol. 34, no. 3, 2009.

[111] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,' 2023, URL https://arxiv.org/abs/1910.10683.

[112] Reimers, N., Gurevych, I., 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,' in: 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,' Association for Computational Linguistics, 11 2019.

[113] Rokhideh, M., Fearnley, C., Budimir, M., 'Multi-hazard early warning systems in the sendai framework for disaster risk reduction: Achievements, gaps, and future directions,' *International Journal of Disaster Risk Science*, vol. 16, pp. 103–116, 2025, doi:10.1007/s13753-025-00622-9.

[114] Ronco, M., Bandelli, L., Bertolini, L., Consoli, S., Delforge, D., Spadaro, A., Verile, M., Corbane, C., 'Disaster storylines and knowledge graphs from global news with large language models and retrieval-augmented generation,' *Scientific Data*, vol. 13, no. 1, p. 689, 2026, doi:10.1038/s41597-026-07036-2.

[115] Sagi, O., Rokach, L., 'Ensemble learning: A survey,' *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, 2018, doi:10.1002/widm.1249.

[116] Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., et al., 'Digital epidemiology,' *PLoS Computational Biology*, vol. 8, no. 7, p. e1002616, 2012.

[117] Samarajeewa, C., De Silva, D., Osipov, E., Alahakoon, D., Manic, M., 'Causal reasoning in large language models using causal graph retrieval augmented generation,' in: '2024 16th International Conference on Human System Interaction (HSI),' pp. 1–6, 2024, doi:10.1109/HSI61632.2024.10613566.

[118] Schmidt, M., Meier, M., Lausen, G., 'Foundations of SPARQL query optimization,' in: 'ACM International Conference Proceeding Series,' p. 4 – 33, 2010.

[119] Shen, W., Wang, J., Han, J., 'Entity linking with a knowledge base: Issues, techniques, and solutions,' *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, p. 443 – 460, 2015, doi:https://doi.org/10.1109/TKDE.2014.2327028.

[120] Shimizu, C., Hirt, Q., Hitzler, P., 'A protégé plug-in for annotating OWL ontologies with OPLa,' *Lecture Notes in Computer Science*, vol. 11155 LNCS, pp. 23–27, 2018.

[121] Sodoge, J., Kuhlicke, C., Mahecha, M. D., de Brito, M. M., 'Text mining uncovers the unique dynamics of socio-economic impacts of the 2018–2022 multi-year drought in germany,' *Natural Hazards and Earth System Sciences*, vol. 24, pp. 1757–1777, 2024, doi:10.5194/nhess-24-1757-2024.

[122] Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., Vasilev, V., 'A versatile data-intensive computing platform for information retrieval from big geospatial data,' *Future Generation Computer Systems*, vol. 81, p. 30 – 40, 2018, doi:10.1016/j.future.2017.11.007.

[123] Spagnolo, L. and Abdelmalik, P. and Doherty, B. and Fabbri, M. and Linge, J. and Marin Ferrer, M. and Moussas, C. and Osato, F. and Petrillo, M. and Poljansek, K. and Schnitzler, J. and Vernaccini, L., 'Integration of the Epidemic Intelligence from Open Sources (EIOS) and the INFORM suite,' JRC Technical Report EUR 30353 EN, Publications Office of the European Union, Luxembourg, 2020, doi:10.2760/958918, JRC121515.

[124] Steinberger, R., Atkinson, M., Garcia, D. T., Van der Linge, J., Macmillan, C., Tanev, H., Verile, M., Wagner, G., et al., 'EMM: Supporting the analyst by turning multilingual text into structured data,' in: 'Transparenz aus Verantwortung: neue Herausforderungen für die digitale Datenanalyse,' Erich Schmidt Verlag, 2017.

[125] Steinberger, R., Pouliquen, B., van der Goot, E., 'An introduction to the Europe Media Monitor family of applications,' 2013, URL https://arxiv.org/abs/1309.5290, arXiv:1309.5290 [cs.CL].

[126] Sutskever, I., Vinyals, O., Le, Q. V., 'Sequence to sequence learning with neural networks,' in: 'Advances in neural information processing systems,' pp. 3104–3112, 2014.

[127] Templin, T., Perez, M. W., Sylvia, S., Leek, J., Sinnott-Armstrong, N., 'Addressing 6 challenges in generative AI for digital health,' *PLOS Digital Health*, 2024, doi:10.1371/journal.pdig.0000503, URL https://doi.org/10.1371/journal.pdig.0000503.

[128] Thakkar, V., Silverman, G. M., Kc, A., Ingraham, N. E., Jones, E. K., King, S., Melton, G. B., Zhang, R., Tignanelli, C. J., 'A comparative analysis of large language models versus traditional information extraction methods for real-world evidence of patient symptomatology in acute and post-acute sequelae of sars-cov-2,' *PLOS ONE*, vol. 20, no. 5, pp. 1–18, 05 2025, doi:10.1371/journal.pone.0323535, URL https://doi.org/10.1371/journal.pone.0323535.

[129] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., Ting, D. S. W., 'Large language models in medicine,' *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023, doi:10.1038/s41591-023-02448-8, URL https://doi.org/10.1038/s41591-023-02448-8.

[130] Thomas, D. S. K., Jang, S., Scandlyn, J., 'The chasms conceptual model of cascading disasters and social vulnerability: The covid-19 case example,' *International Journal of Disaster Risk Reduction*, vol. 51, p. 101828, 2020, doi:10.1016/j.ijdrr.2020.101828.

[131] Tilloy, A., Malamud, B., Winter, H., Joly-Laugel, A., 'A review of quantification methodologies for multi-hazard interrelationships,' *Earth-Science Reviews*, vol. 196, p. 102881, 2019, doi:10.1016/j.earscirev.2019.102881.

[132] Tiwari, S., Ortíz-Rodriguez, F., Abbés, S. B., Usip, P. U., Hantach, R., *Semantic AI in Knowledge Graphs*, Taylor & Francis, Boca Raton, US, 2023, doi:10.1201/9781003313267.

[133] Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., Nelson, N. P., 'An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics,' *International Journal of Medical Informatics*, vol. 80, no. 1, pp. 56–66, 2011, doi:https://doi.org/10.1016/j.ijmedinf.2010.10.015.

[134] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al., 'LLaMA: Open and Efficient Foundation Language Models,' 2023.

[135] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al., 'Llama 2: Open Foundation and Fine-Tuned Chat Models,' 2023.

[136] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., 'Attention is all you need,' in: 'Advances in Neural Information Processing Systems,' pp. 5999–6009, 2017.

[137] Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., Li, J., 'Prompt engineering in consistency and reliability with the evidence-based guideline for llms,' *npj Digital Medicine*, vol. 7, no. 1, 2024, doi:10.1038/s41746-024-01029-4.

[138] Warsame, A., Murray, J., Gimma, A., Checchi, F., 'The practice of evaluating epidemic response in humanitarian and low-income settings: a systematic review,' *BMC Medicine*, vol. 18, no. 1, 2020.

[139] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al., 'Chain-of-thought prompting elicits reasoning in large language models,' 2023.

[140] Weibel, S. L., Koch, T., 'The Dublin core metadata initiative: Mission, current activities, and future directions,' *D-Lib Magazine*, vol. 6, no. 12, 2000, doi:10.1045/december2000-weibel.

[141] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al., 'Transformers: State-of-the-Art Natural Language Processing,' in: Q. Liu, D. Schlangen (editors), 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,' pp. 38–45, Association for Computational Linguistics, Online, Oct. 2020, doi:10.18653/v1/2020.emnlp-demos.6.

[142] Xu, F., Ma, J., Li, N., Cheng, J. C. P., 'Large language model applications in disaster management: An interdisciplinary review,' *International Journal of Disaster Risk Reduction*, vol. 127, p. 105642, 2025, doi:10.1016/j.ijdrr.2025.105642.

[143] Yang, A., et al., 'Qwen2.5 technical report,' 2025, URL https://arxiv.org/abs/2412.15115.

[144] Yang, J., Han, S. C., Poon, J., 'A survey on extraction of causal relations from natural language text,' *Knowl. Inf. Syst.*, vol. 64, no. 5, p. 1161–1186, May 2022, doi:10.1007/s10115-022-01665-w, URL https://doi.org/10.1007/s10115-022-01665-w.

[145] Yang, R., Yang, B., Ouyang, S., She, T., Feng, A., Jiang, Y., Lecue, F., Lu, J., Li, I., 'Graphusion: Leveraging large language models for scientific knowledge graph fusion and construction in nlp education,' *arXiv preprint*, vol. arXiv:2407.10794 [cs.CL], 2024, URL https://doi.org/10.48550/arXiv.2407.10794, 24 pages, 11 figures, 13 tables. arXiv admin note: substantial text overlap with arXiv:2402.14293.

[146] Yang, R., Zhu, J., Man, J., Fang, L., Zhou, Y., 'Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement,' *arXiv preprint*, vol. arXiv:2310.08279 [cs.CL], 2023, URL https://doi.org/10.48550/arXiv.2310.08279, new version.

[147] Yerkhassym, A., Pak, A. A., Akhmetov, I., Yelenov, A., Gelbukh, A., 'On causality problem in natural language processing field,' *Computacion y Sistemas*, vol. 26, no. 4, p. 1549 – 1556, 2022, doi: 10.13053/CyS-26-4-4434.

[148] You, J., Expert, P., Costelloe, C., 'Using text mining to track outbreak trends in global surveillance of emerging diseases: Promed-mail,' *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 184, no. 4, pp. 1245–1259, 07 2021.

[149] Yu, Y., Yang, C.-H. H., Kolehmainen, J., Shivakumar, P. G., Gu, Y., Ren, S. R. R., Luo, Q., Gourav, A., Chen, I.-F., Liu, Y.-C., Dinh, T., Filimonov, A. G. D., Ghosh, S., Stolcke, A., Rastow, A., Bulyko, I., 'Low-Rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition,' in: '2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),' p. 1–8, IEEE, Dec. 2023, doi:10.1109/asru57964.2023.10389632.

[150] Zaghir, J., Naguib, M., Bjelogrlic, M., Névéol, A., Tannier, X., Lovis, C., 'Prompt Engineering Paradigms for Medical Applications: Scoping Review,' *Journal of Medical Internet Research*, vol. 26, 2024, doi: 10.2196/60501.

[151] Zhang, K., Meng, X., Yan, X., Ji, J., Liu, J., Xu, H., Zhang, H., Liu, D., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Shao, C., Wang, W., Li, J., Zheng, M.-Q., Yang, Y., Tang, Y.-D., 'Revolutionizing Health Care: The Transformative Impact of Large Language Models in Medicine,' *Journal of Medical Internet Research*, vol. 27, 2025, doi:10.2196/59069.

[152] Zhou, X., Zhou, J., Wang, C., Xie, Q., Ding, K., Mao, C., Liu, Y., Cao, Z., Chu, H., Chen, X., Xu, H., Larson, H. J., Luo, Y., 'PH-LLM: Public Health Large Language Models for Infoveillance,' *medRxiv*, vol. Preprint, 2025, doi:10.1101/2025.02.08.25321587, URL https://pubmed.ncbi.nlm.nih.gov/39990576/.

[153] Zhou, Y., Li, C., Huang, G., Guo, Q., Li, H., Wei, X., 'A Short-Text Similarity Model Combining Semantic and Syntactic Information,' *Electronics*, vol. 12, no. 14, 2023, doi:https://doi.org/10.3390/electronics12143126.

[154] Zhou, Z.-H., *Ensemble methods: Foundations and algorithms*, Chapman & Hall/CRC, 2012, doi:10.1201/b12207.

[155] Šakić Trogrlić, R., Reiter, K., Ciurean, R. L., Gottardo, S., Torresan, S., Daloz, A. S., Ma, L., Padrón Fumero, N., Tatman, S., Hochrainer-Stigler, S., de Ruiter, M. C., Schlumberger, J., Harris, R., Garcia-Gonzalez, S., García-Vaquero, M., Febles Arévalo, T. L., Hernandez-Martin, R., Mendoza-Jimenez, J., Ferrario, D. M., Geurts, D., Stuparu, D., Tiggeloven, T., Duncan, M. J., Ward, P. J., 'Challenges in assessing and managing multi-hazard risks: A european stakeholders perspective,' *Environmental Science & Policy*, vol. 157, p. 103774, 2024, doi:10.1016/j.envsci.2024.103774.

List of Figures

Figure 1. Comparison of the models for the extraction of the pandemic name. . . . 19
Figure 2. Comparison of the models for the extraction of the country name. . . . 20
Figure 3. Comparison of the models for the extraction of the pandemic date. . . . 21
Figure 4. Comparison of the models for the extraction of the number of cases engender by the pandemic. . . . 22
Figure 5. Schematic overview of the pipeline to extract and publish relevant epidemiological information from the WHO's Disease Outbreak News. . . . 24
Figure 6. This picture illustrates an example on the implementation of Description Logic for structured knowledge representation. The TBox comprises concepts such as *surveillance process* and *planned process*, along with assertions defining relationships between these concepts (e.g., <*surveillance process, subclass of, planned process*>). The ABox represents instances of these concepts, like *don-record2740*, and provides assertions relating these instances to other concepts (e.g., <*don-record2740, instance of, surveillance process*> and <*don-record2740, geographical region, India*>) as well as to literal values (e.g., <*don-record2740, date of infection, 2018-05-19*> and <*don-record2740, infected population cases, 15*>). . . . 28
Figure 7. Snippet of eKG exploration services and interfaces. . . . 31
Figure 8. Main classes within eKG. . . . 34
Figure 9. Number of extracted cases vs. time for 4 case studies: MERS-Cov in Saudi Arabia (Panel a), Ebola cases in the Democratic Republic of Congo (Panel b), SARS cases in China (Panel c), and Ebola cases in Guinea (Panel d). These four case studies correspond to the Disease Outbreak Names - Country pairs reported in bold in Table 2. . . . 36
Figure 10. WHO yearly number of cases vs the number of cases from the reconstructed events, together with regression line: including year 2013 (Panel a) and without year 2013 (Panel b). Since 2013 was the epidemic's onset, the inclusion of potential and confirmed cases by WHO DONS could lead to a lower correlation coefficient due to overestimated reconstructed cases. Correlation and p-values are reported in the top right corner of each panel. . . . 37
Figure 11. Faceted Browser showing the extracted Nipah virus - India information. . . . 38
Figure 12. Content negotiation of the Nipah virus - India report with LodView. . . . 39
Figure 13. Graph visualisation of Nipah virus - India in LodLive. . . . 40
Figure 14. An extract from the knowledge graph representation in LODE. . . . 41
Figure 15. A snapshot of the WebVOWL visualisation. . . . 42
Figure 16. ROUGE scores obtained by each iCL model at increasing number of examples (NiCL). . . . 48
Figure 17. Rouge-1 score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL). . . . 48
Figure 18. Rouge-2 score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL). . . . 49
Figure 19. Rouge-L score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL). . . . 49
Figure 20. Rouge-Lsum score and standard deviation bands at 99% confidence interval obtained by each iCL model at increasing number of examples (NiCL). . . . 50

Figure 21. Pipeline for generating disaster narratives and knowledge graphs from news. (a) Schematic of the two-stage generation process. A predefined query is used to retrieve top-k relevant news articles from the EMM. These are passed to a first LLM that generates a structured storyline summarising the event's key characteristics. A second LLM then converts this storyline into a knowledge graph capturing entities and relationships. (b) Querying and filtering process. For each event, semantic search retrieves the top 20 articles for each of the four weeks following the event start. The resulting set is refined by an LLM that filters out unrelated articles to improve relevance and reduce noise. . . . 56

Figure 22. Prompts used in the AI pipeline for disaster storyline and knowledge graph generation. (a) Prompt for generating narrative summaries of disaster events using a set of predefined fields. (b) Prompt for extracting KGs using iCL, including an example to guide the generation of subject–predicate–object triples constrained to "causes" and "prevents" relationships. 59

Figure 23. Example outputs for the 2015 India–Pakistan heatwave. (a) Disaster storyline generated by Meta-Llama-3-70B-Instruct, organised into predefined thematic sections. (b) Corresponding knowledge graph extracted from the narrative using a second LLM call with iCL, showing causal and preventive relationships constrained to the predicates "causes" and "prevents". 60

Figure 24. Results of the survey assessing the AI-generated narratives and graphs. (a) Accuracy of Storylines (options: "Very accurate," "Partially accurate," and "Inaccurate.") (b) Usefulness of the Graphs (options: "Very useful," "Somewhat useful," "Neutral," and "Not useful.") (c) Accuracy of Entities (options: "Yes, all accurately identified and represented," "Mostly, but a few inaccuracies are present," "Partly, some entities are correctly identified while others are not," "No, many entities are inaccurately identified or missing," and "Not sure / I don't have enough information to assess.") (d) Accuracy of Links/Relationships (options: "Yes, all accurately identified and represented," "Mostly, but a few inaccuracies are present," "Partly, some links are correctly identified while others are not," "No, many links are inaccurately identified or missing," and "Not sure / I don't have enough information to assess.") 62

List of Tables

Table 1. Traditional AI vs. Generative AI 11
Table 2. The top reported disease outbreak names, countries, and disease outbreak name-country pairs by total number of extracted events (number given in parentheses). Pairs of disease outbreak name-country in bold are further illustrated in Figure 9. 35
Table 3. Rouge-1 scores (mean $\pm$ standard deviation) obtained by each fine-tuned model. 51
Table 4. Rouge-2 scores (mean $\pm$ standard deviation) obtained by each fine-tuned model. 51
Table 5. Rouge-L scores (mean $\pm$ standard deviation) obtained by each fine-tuned model. 52
Table 6. Rouge-Lsum scores (mean $\pm$ standard deviation) obtained by each fine-tuned model. 52
Table 7. Distribution of expert assessments by disaster type, following EM-DAT categories [39]. Values represent the percentage of all individual expert ratings in each category. The final column indicates the number of unique disaster graphs evaluated for each type. 61
Table 8. Key Challenges and Mitigation Strategies 66

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/write-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

Joint Research Centre
https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union