Hybrid KG-RAG Question-Answering for Extreme Weather Events

Lorenzo Augello[1], Maxim Ionov[1] and Jorge Gracia[1]

[1]Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain

Abstract

Extreme weather event information is distributed across heterogeneous sources, including specialized disaster databases like EM-DAT and textual reports like World Weather Attribution. This creates a challenging setting for question answering systems in the climatology domain, where some questions require precise access to event attributes, while others require explanatory textual evidence. This paper presents a hybrid KG-RAG system that combines SPARQL-based retrieval over an RDF knowledge graph with vector-based retrieval over textual documents. We hypothesize that combining the two will improve the overall quality of the question-answering system. To test this, we construct a domain ontology and a benchmark of factual question-answer pairs derived from both structured and textual sources. We then create a hybrid KG-RAG QA system tailored to this task.

The system is evaluated in a number of configurations: KG-only, RAG-only, an agentic system with access to both RAG and KG, and a deterministic hybrid configuration with and without a post-retrieval feedback loop. Results show that the deterministic hybrid system with feedback achieves the highest accuracy for both tested models: a larger cloud-based model and a smaller model running locally. This improves both the RAG-only and KG-only baselines, showing that the RDF knowledge graph, grounded in the domain ontology, complements textual retrieval with structured and queryable evidence. Notably, the smaller local model in the best hybrid setting outperforms the larger model in the RAG-only setting, suggesting that retrieval architecture can be as important as model scale. Overall, the results confirm our hypothesis that integrating RDF-based KG retrieval with document-level RAG provides a more effective access layer for climate extreme-event question answering than either of the two sources alone.

Keywords

RAG, RDF, SPARQL, Question Answering, Knowledge Graphs

CEUR-WS.org/Vol-4312/paper23.pdf

1. Introduction

Extreme weather events (e.g., heatwaves, droughts, floods) have growing consequences for human life, infrastructure, and the economy, making accurate information about them essential for disaster response and planning. Information about such events is available both as structured entries in specialized databases such as EM-DAT [1] and as narrative accounts in textual reports such as those from World Weather Attribution [2]. This heterogeneity poses a challenge for climate-event question-answering (QA): answering a single question might require synthesising data from both types of sources, yet existing architectures rarely support querying across the two. Addressing this gap also speaks to a broader goal of building interoperable infrastructures for accessing domain-specific scientific knowledge.

Retrieval-Augmented Generation (RAG) [3] systems are well suited to retrieving and using passages from textual documents, and are therefore useful for exploiting WWA reports. However, text-only RAG is less effective when questions require access to exact values inside complex structures, such as impact values, dates, disaster taxonomies, event-location relations, or links between events and their consequences. EM-DAT itself is organized as tabular records, where each event is a row and each attribute a column; this supports fast, exact lookup of individual values, but a table has no native way to express relations between rows, such as links between an event and its downstream consequences, or between related events across countries and years. Because of this, tabular lookup alone cannot support multi-hop reasoning or interoperable querying with external sources such as WWA reports. Representing events instead as nodes in a knowledge graph (KG), connected by explicit, typed relations,

Retrieval-Augmented Generation Enabled by Knowledge Graphs 2026, October 25–26, 2026, Bari, Italy

laugello@unizar.es (L. Augello); mionov@unizar.es (M. Ionov); jogracia@unizar.es (J. Gracia)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-04

addresses this limitation: it preserves the precision of structured attributes while making cross-event and cross-source relationships directly queryable. A climate-event QA system should therefore not be limited to either textual retrieval or structured lookup alone, but should combine both: the flexibility of RAG over explanatory documents and the precision of graph-based retrieval over semantically modelled event data.

This paper addresses this by developing and evaluating a hybrid RAG- and KG-based QA system for extreme weather events.[1] The central hypothesis is that a system with access to both textual and graph-structured sources can answer climate-event questions more effectively than systems relying on only one of them. Given this, our contributions are threefold. First, we construct an ontology-driven RDF KG for extreme weather events. Rather than treating EM-DAT rows as isolated table entries, each event is represented as a semantic entity identified by a stable URI and linked to its attributes, making information explicitly queryable and interoperable. Second, we create a benchmark for evaluating hybrid climate-event QA, combining context information coming from structured EM-DAT facts and WWA textual evidence. Third, we design and evaluate a hybrid KG-RAG architecture developed in different configurations, showing that the integration of KG-based retrieval and RAG performs better than the two systems in isolation.

The remainder of the paper is structured as follows. We present related work in the integration of KGs and LLMs in Section 2; in Section 3 we describe our source data, while in Section 4 our climate ontology and KG are presented. The benchmark generation process and its limitations are presented in Section 5, while a description of the different evaluated systems is provided in Section 6. The results are discussed in Section 7, before outlining the limitations and some potential directions for future work in Section 8.

2. Related Work

Recent work on question-answering for climate research has highlighted the need for domain-specific models, benchmarks, and retrieval mechanisms. ClimateGPT [4] and ChatClimate [5] show the importance of grounding LLMs in climate-specific knowledge, while ClimateEval [6] provides a benchmark dataset for climate-related NLP tasks. Li et al. [7] provide a work that is highly related to ours, addressing the modeling of extreme-event impacts, but focusing on extracting impact information from textual sources. This is complementary to our work, as we semantically model structured EM-DAT records as an RDF KG and align them with textual WWA evidence for downstream hybrid QA.

In KGQA, natural-language questions are typically translated into executable SPARQL queries, but this remains challenging because systems must identify the relevant entities and predicates and generate syntactically valid queries over a specific graph schema. Recent SPARQL generation approaches are particularly aligned with our work: for instance, GRASP [8] generates SPARQL queries over arbitrary RDF graphs without fine-tuning by strategically exploring the graph; SPINACH [9] introduces a real-world SPARQL dataset and an in-context learning agent that, following the ReAct framework [10], decomposes query construction into intermediate decisions and actions such as selecting a predicate or exploring a neighbouring node; FIRESPARQL [11] combines LLM-based query generation with optional RAG context and a SPARQL correction layer. Our work builds on this line of research by using SPARQL-based retrieval over an ontology-driven KG of climate extreme events, and, like SPINACH, we explore an agentic ReAct-style variant of this retrieval process alongside a deterministic alternative — a comparison we return to in Section 6.

Earlier approaches to integrating KGs and LLMs [12] show that explicit structured knowledge can improve language understanding and factual reasoning [13, 14], but this often implies model modification or additional pre-training. While this can be solved with traditional RAG by keeping knowledge external, a further KG integration is needed for enabling structured reasoning over chunks and entities.

Recent GraphRAG research extends this idea by combining graph-structured retrieval with LLM generation, improving context organization, multi-hop reasoning and knowledge integration [15, 16].

[1] https://github.com/sid-unizar/borrajas

Among the more specific approaches, KG2RAG [17] uses KGs to capture fact-level relationships between chunks and improve the diversity and coherence of retrieved evidence, while GFM-RAG [18] further shows the relevance of graph foundation models for RAG by using graph neural retrieval over a KG index built from documents.

Unlike GraphRAG approaches that build their graph index from extracted document chunks, our KG is constructed from curated, ontology-driven structured records, and is used alongside, rather than as a substitute for, textual retrieval over independent narrative sources.

3. Data Sources

This work combines two sources of information about extreme weather events: structured disaster records from EM-DAT and textual event descriptions from WWA. The two sources provide different but complementary forms of evidence. EM-DAT provides standardized event-level information, while WWA consists of textual content about the meteorological, social, and attribution-related aspects of selected events. The events described in WWA start in 2014, while EM-DAT contains historical events too, starting from 1900. For alignment reasons, for this paper we restricted the EM-DAT to the records from 2014 to 2026, in order to consider the same time span for both data sources.

3.1. EM-DAT

The KG component of our data that we use for the experiments is based on EM-DAT, the International Disaster Database.[2] EM-DAT presents information on disaster events in tabular form, with each row corresponding to an event or disaster episode. It contains information on more than 27,000 mass disasters from 1900 to the present, compiled from sources including UN agencies, non-governmental organizations, reinsurance companies, research institutes, and press agencies.

The dataset is publicly available and includes structured fields describing the event: identifier, country, disaster group, disaster subgroup, disaster type, disaster subtype, start and end dates, affected regions or provinces; and the impact that this event had: number of deaths, injuries, displacement, and economic damage. In our experiments, these fields provide the factual basis for precise event-level querying.

For the reasons outlined in the Introduction, instead of using EM-DAT in its tabular form, we use it as a source for constructing a KG as explained in Section 4.3. Each row is transformed into RDF triples, and the values in the structured columns are mapped to entities, properties, literals, or existing URIs. The resulting graph[3] preserves the original factual content of the table while making the information semantically explicit and queryable through SPARQL. In total, our graph has information on 4,657 climate events, from 2014 to 2026.

3.2. World Weather Attribution

The textual component of the data we use for our experiments is based on WWA[4] articles matched to corresponding events in EM-DAT. Overall, WWA contains 131 web articles, each referring to a particular extreme weather event. Among those, we manually annotate, isolate and consider only the ones for which a correspondent event could be identified in EM-DAT, consisting of 58 articles.

3.3. Alignment

For the combined dataset we collect information coming from both sources: metadata from the WWA articles provides additional information that is not encoded in the mostly numerical EM-DAT data, such as description of the event, meteorological conditions, climate attribution findings, exposure, and vulnerability. Moreover, the text from each WWA article is extracted, cleaned, and stored as

[2] https://www.emdat.be/

[3] https://github.com/sid-unizar/borrajas/blob/main/data/emdat_climate_events.ttl

[4] https://www.worldweatherattribution.org/

contextual evidence in a vector database. This textual evidence is used to build the document corpus for the RAG component. While the KG provides structured facts such as disaster type, location, dates, and impact values, the WWA text provides explanatory context that can support more narrative or attribution-oriented questions.

The two sources play distinct roles in the proposed hybrid architecture. EM-DAT is transformed into an RDF KG and queried through SPARQL; WWA webpages are indexed as textual documents and queried through vector retrieval, and they are therefore used for retrieving explanatory passages and contextual evidence.

4. Ontology Design

The *clasik*[5] ontology model shown in Figure 1 is designed to represent extreme weather events as interconnected semantic entities rather than as isolated table rows. Its goal is to make EM-DAT-style structured disaster information interoperable, queryable, and usable as the knowledge layer of an RDF-aware system.

For the ontology design and engineering, we make use of already existing ontologies (e.g., SWEET ontology for impacts,[6] GeoNames for locations,[7] AEMET for observations[8]) and merge them together to create a new ontological schema suitable to the modelling of our source EM-DAT data. To accommodate all the needs required by our data, we also develop specifically tailored classes and properties to represent as much semantic information as possible about climate extreme events.

As a future direction for making the ontology design process as methodologically sound as possible, we plan to develop explicit competency questions to validate and justify the selection, reuse, and contextual relevance of external ontologies for extreme weather representation. This would serve as an empirical demonstration of whether the reused ontologies are truly relevant, accurate, and contextual for the selected extreme weather events.

4.1. Event Representation

The central element of the *clasik* ontology is the climate event. Each EM-DAT row is transformed into an instance of the clasik:ClimateEvent class, and is associated with its attributes like in the following example which models information about Hurricane Iselle:[9]

```
clasik:ev_2014-0310-USA     a   clasik:ClimateEvent ;
    rdfs:label                  "Hurricane Iselle"@en ;
    sorelsc:hasImpact           clasik:imp_2014-0310-USA ;
    gn:locatedIn                    <https://sws.geonames.org/6252001/> ;
    clasik:hasDisSubgroup   clasik:subgroup_meteorological ;
    clasik:hasDisSubtype    clasik:subtype_trop_cyclone ;
    clasik:hasDisType           clasik:type_storm ;
    clasik:isInProvince         "Hawaii province" ;
    soteria:endDate             "2014-08-10"^^xsd:date ;
    soteria:startDate           "2014-08-08"^^xsd:date .

clasik:imp_2014-0310-USA    a   clasik:Impact ;
    clasik:deaths               1 ;
    clasik:economicDamage       66000 ;
```

[5]The name is chosen based on the name of the CLASiK research project, under the umbrella of which this work was carried out.

[6]http://sweetontology.net/sweetAll

[7]https://www.geonames.org/ontology/documentation.html

[8]http://aemet.linkeddata.es/ontology/

[9]In the remainder of this paper the following prefixes will be used for the RDF code: aemet: <http://aemet.linkeddata.es/ontology/> ; clasik: <https://example.org/clasik#> ; geo1: <http://www.w3.org/2003/01/geo/wgs84_pos#> ; gn: <http://www.geonames.org/ontology#> ; qudt: <http://qudt.org/schema/qudt/> ; rdfs: <http://www.w3.org/2000/01/rdf-schema#> ; sorelsc: <http://sweetontology.net/relaSci/> ; soteria: <https://soteria-public.pages.sintef.no/soteria_ontology_documentation#> ; xsd: <http://www.w3.org/2001/XMLSchema#> .

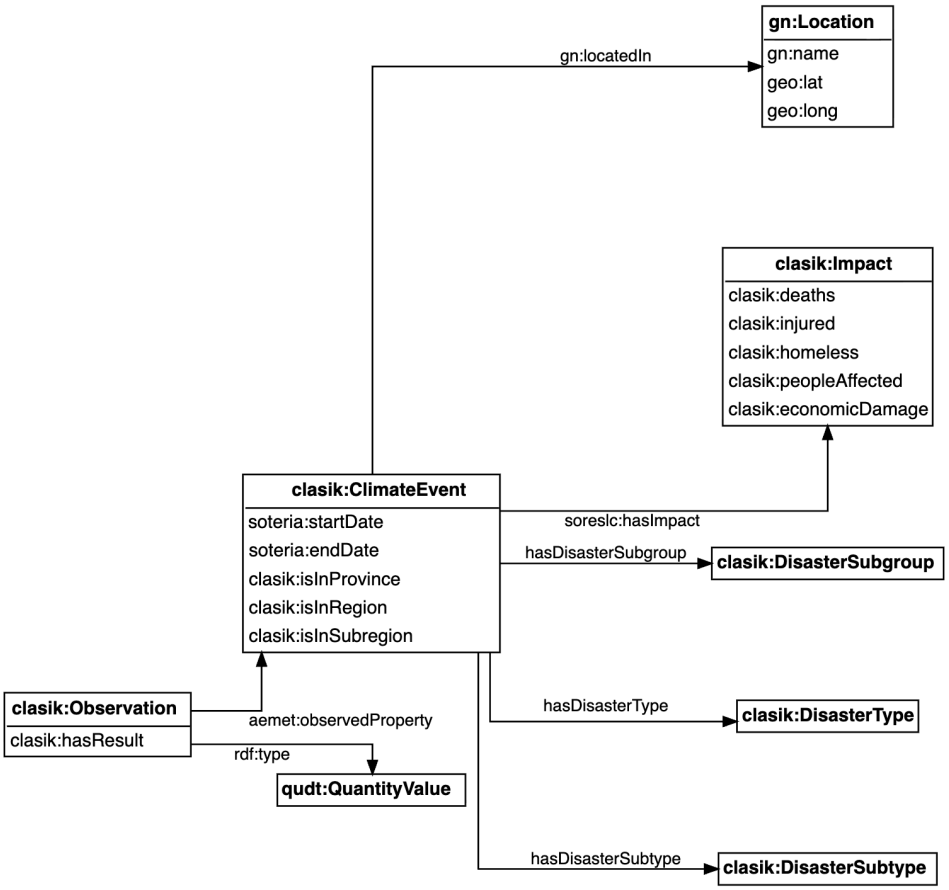


Figure 1: Extreme weather events ontology.

```
    clasik:peopleAffected       834 .

clasik:obs_2014-0310-USA    a   clasik:Observation;
    aemet:observedProperty      clasik:ev_2014-0310-USA ;
    clasik:hasResult            "95.0 Kph"@en .
```

Figure 1 shows how the climate event is central in our model. In addition to its location and date, each event is enriched with information about two main attributes: observation and impact. Each clasik:Observation is defined by a certain measurement or amount through the clasik:hasResult property, while each clasik:Impact is further specified according to a classification derived from EM-DAT. In fact, the impact of an event can be expressed as the number of deaths, injured, homeless, total people affected and economic damage (in a fixed currency).

4.2. Disaster Taxonomy

Disaster categories in EM-DAT are values from a controlled vocabulary rather than plain strings. The *clasik* model adopts this vocabulary and, following this, distinguishes between disaster subgroup, disaster type, and disaster subtype. This preserves the hierarchical structure of EM-DAT-style disaster classification, derived from the IRDR peril classification and hazard glossary,[10] and it supports queries at different levels of granularity.

For instance, an event can be linked to a general disaster type such as *flood*, while also being linked to a more specific subtype such as *riverine flood*. This allows users to query either broad categories, such as all hydrological events, or more specific categories, such as *flash flood* or *riverine flood*.

[10]https://irdrinternational.org/news/114

Table 1
Numbers of the final climate KG

Category	Count
Total RDF triples	73,781
Total number of nodes	14,233
Climate event instances	4,657
Classes	7
Data properties	15
Object properties	7

4.3. KG Construction

The KG was constructed from structured EM-DAT event data using an RDF generation pipeline implemented with rdflib.[11] The goal of the construction process was to transform the original tabular representation into an ontology-driven RDF graph, where each row is no longer treated as an isolated record but as a semantically described climate extreme event connected to its attributes.

Each event identifier provided by the source data was used to generate a stable URI for the event. For each event, we created RDF triples describing its disaster subgroup, disaster type, disaster subtype, location, date, measurements, and impact information. The structured data can therefore populate the ontology model described in Section 4, making the EM-DAT content queryable through SPARQL and reusable as a knowledge base in the RDF-aware system (see Section 6).

The KG construction followed a four-steps process: first, the EM-DAT data is downloaded as a structured table; second, each row is mapped to the ontology classes and properties, and URIs are generated for events, impacts, observations, and locations; third, the RDF triples are added to the graph; and finally, the graph is serialized as Turtle and used as the structured retrieval source of our proposed system. The final numbers obtained as a result at the end of the pipeline are presented in Table 1.

This approach is designed to be reusable. In fact, whenever a new EM-DAT dump becomes available, the same pipeline can be rerun to regenerate or update the KG. Moreover, additional sources can also be integrated into the graph: for example, event descriptions, attribution studies, geospatial identifiers, or records from other disaster databases can be connected to the corresponding EM-DAT event resources, creating links and enabling cross-source integration and comparison.

5. Benchmark Generation

Evaluating a hybrid KG-RAG system requires question-answer pairs that are grounded in the domain and that are based on the sources available to the system. Existing climate NLP benchmarks increasingly recognize the need for standardized evaluation resources. Kurfali et al. [6], for example, aggregate 13 datasets into 25 climate-related NLP tasks, including QA and information extraction, providing a broader framework for evaluating LLMs in climate applications. Some of these benchmarks, such as Pirá 2.0 MCQ [19], are built for tasks very close to ours, Closed Generative Question-Answering (CGQA). However, adapting existing benchmarks to our setting is not straightforward. Taking Pirá 2.0 as an example, it would require a lot of additional work to use it: Pirá is based on Portuguese texts, and in order to use it for our hybrid method, it would require us to also construct a knowledge graph with intersecting data rather than leveraging an existing structured resource such as EM-DAT. We are not aware of an existing benchmark built directly on EM-DAT and WWA that meets our requirements, and we therefore construct our own. We do this by prompting an LLM, namely gpt-oss:120b from OpenAI,[12] to generate question-answer pairs from grounded event-level context.

To make sure that the answers to the generated questions can be found in both sources (thus allowing a direct comparison between methods that access either of them), we first identify information that is

[11] https://rdflib.readthedocs.io/en/stable/
[12] https://openai.com/it-IT/index/introducing-gpt-oss/

Table 2
Three examples of question-answer pairs as contained in the generated benchmark.

Question	Answer
What was maximum daily precipitation recorded at Iizuka station during 6 July 2018?	338 mm
What was the estimated economic cost of the December 1, 2015 flooding in Chennai?	$ 3 billion
How many fatalities were reported in Nigeria due to the 2022 floods?	612

common to both using the prompt reported in Appendix A.1. Once we gathered the necessary context common between EM-DAT and WWA, we ask an LLM to generate question-answer pairs starting from a given shared context as reported in Appendix A.2.

By following the RAGAs methodology [20], we provide the LLM with a list of context sections gathered from the intersection between EM-DAT and WWA consisting of 58 climate events (see Section 3), and we ask it to generate two questions for each one of them. Furthermore, we also use gpt-oss:120b to answer the generated question, given the corresponding context. All context-question-answer triples are then evaluated and corrected by a domain expert in climatology, coming to a final benchmark of 88 validated question-answer-context triples, with an acceptance rate of 76%.

We are aware that the exact same base model family is utilized for both generating the synthetic benchmark ground truth answers and evaluation, which introduces potential circularity and inherited model biases.

It is important to note that the proposed benchmark is limited by design: it only contains simple domain-specific questions, among the different types of QA benchmarks such as the ones containing multi-hop [21], open domain [22], or multilingual questions [23]. In fact, the LLM was explicitly prompted to generate simple questions, as shown in Appendix A.2. Therefore our conclusions cannot immediately be generalized beyond this type of simple questions. We leave the development of a benchmark containing more complex types of questions as future work. A few examples of question-answer pairs are reported in Table 2.

6. System Configurations

We develop six separate system architectures, each corresponding to a different retrieval configuration, in order to evaluate how the configuration of information access affects performance in QA over extreme weather events. We compare two base configurations, namely RAG-only and KG-only (see Figure 2), with two hybrid workflows that combine textual and graph-based retrieval (see Figures 3 and 4). In addition, each of the hybrid workflows is tested with and without a post-retrieval feedback mechanism, proposed for the GRASP system [8].

RAG-only provides the system with access to the information contained in the WWA articles. Each document is extracted, processed, and indexed in a vector database using the *nomic-embed-text* model.[13] At inference time, the input question is used to retrieve semantically relevant passages from the vector database, and the prompt is enriched with the retrieved context. In this setting, no KG is available to the model. The system therefore relies exclusively on textual evidence from WWA documents.

KG-only gives the system access to the climate KG described in Section 4.3. Given an input question, the model receives the RDF schema of the KG as context (reported in Appendix A.3) and it generates a SPARQL query to retrieve relevant information from the KG. The query is then executed against the KG, and the returned results are used to synthesize the final answer. Following the GRASP methodology [8], the SPARQL agent can iteratively refine its query when the previous attempt does not return sufficient information, constraining the process to a maximum of ten tool calls.

Hybrid combines the two previous configurations, and it is implemented as an agentic workflow composed of multiple nodes: a router, a KG retrieval branch, a RAG retrieval branch, a synthesizer, and

[13] https://ollama.com/library/nomic-embed-text

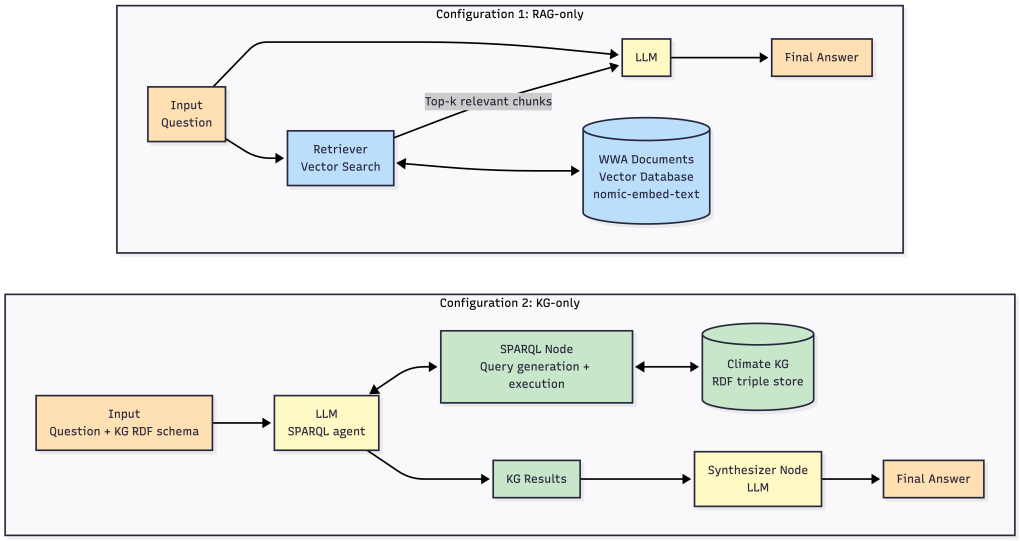


Figure 2: Visualization of the two base configurations: RAG-only in the top diagram and KG-only in the bottom diagram.

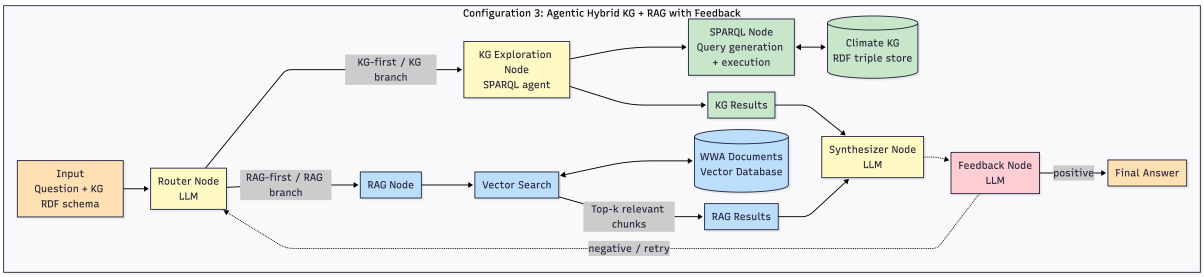


Figure 3: Visualization of the hybrid fully agentic type of configuration. The dotted arrow indicates an optional feedback loop that marks a distinction between two configurations, with and without the post-retrieval feedback.

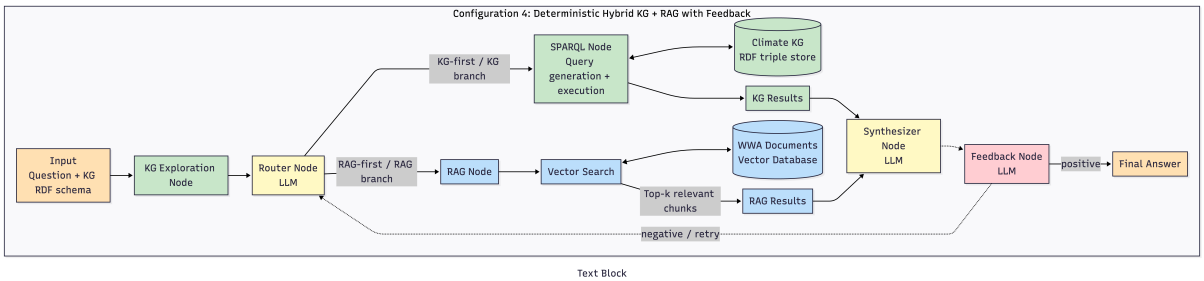


Figure 4: Visualization of the hybrid partially agentic (deterministic) type of configuration. The dotted arrow indicates an optional feedback loop that marks a distinction between two configurations, with and without the post-retrieval feedback.

a feedback component. For each input question, the router determines which knowledge source should be queried, either the climate KG or the WWA vector database. The KG branch retrieves structured evidence by generating and executing SPARQL queries, while the RAG branch retrieves textual evidence from the WWA vector database. Once the selected branch has retrieved relevant data, a synthesizer node generates the final answer. A feedback node then evaluates the generated answer and either accepts it or triggers another retrieval and synthesis iteration.

We further split the hybrid system into two configurations, which differ in the degree of control given to the LLM during KG exploration:

- A **fully agentic system** (see Figure 3), where the LLM receives the input question together

with the KG schema and decides how to interact with the available tools. Before generating the SPARQL query, the agent can call dedicated KG exploration tools, such as functions for inspecting available classes, predicates, class-property relations, or sample entities. These tool calls are not fixed in advance, but rather the LLM decides which tools to invoke, and in which order. While this approach increases flexibility, as the agent can adapt its exploration strategy to the specific question, it also increases the number of decisions delegated to the LLM. As a result, the workflow may become less predictable and more expensive in terms of LLM calls.

- A **partially agentic system**, with a **deterministic call graph** (see Figure 4), where KG exploration is no longer left to the choices of the LLM. Instead, a predefined KG exploration step is executed systematically before routing and query generation. The exploratory functions are called in a fixed and controlled way, producing a schema-oriented summary of the graph that is then provided to the next nodes. The router and SPARQL generation components remain LLM-based, but they operate over a more stable and complete representation of the KG schema, improving reproducibility and limiting unnecessary LLM-tool interaction.

The comparison between the two hybrid variants therefore allows us to assess whether greater agentic flexibility improves performance, or whether a more constrained workflow provides better reliability for climate KG-RAG QA. The two hybrid configurations in Figures 3 and 4 are each tested with and without a feedback node, giving a final total of six system configurations: *rag-only*, *kg-only*, *deterministic-feedback*, *deterministic-no-feedback*, *agentic-feedback*, and *agentic-no-feedback* (Table 3).

Below is a list of the functions defined for the KG exploration:

- **list_classes**: returns RDF classes used in the KG.
- **list_predicates**: returns predicates present in the KG.
- **get_class_properties**: returns predicates used by instances of a given class.
- **sample_instances**: returns sample instance URIs for a given class.
- **describe_entity**: returns all predicate-object pairs for a given entity URI.
- **run_sparql**: executes a SPARQL SELECT query against the KG.

7. Results and Discussion

For the experiments we use and compare two different LLMs: a smaller one that can be run on local hardware, gemma4:12b-mlx,[14] and a bigger one using the cloud provided by ollama, gpt-oss:120b.[15] This allows us to investigate if and how the size of the model affects performance when changing the system configuration.

We evaluate accuracy as the primary metric as the benchmark provides gold-standard answers that can be directly compared with the outputs produced by the system. This was possible because the benchmark answers are short and factual, typically consisting of entities, numbers, percentages, or dates. To be able to make the comparison, the model was prompted to produce two forms of output, as shown by the system prompt at the synthesizer level in Appendix A.4: a short answer, intended to be directly comparable with the benchmark gold answer, and a longer answer, intended to provide additional explanatory context. The accuracy evaluation was therefore based on the comparison between the generated short answer and the gold answer: each prediction was evaluated with a score from 0 to 1, and was considered correct when it exactly matched the gold answer or when it constituted a close semantic match, as decided by an LLM judge, described in Appendix A.5.

We are aware that human evaluation of the LLM judge would be a valuable addition for any future study aimed at making this methodology scalable to more data and qualitatively more reliable.

The results presented in Table 3 show that the integration of the KG improves an already strong RAG baseline. RAG-only reaches high accuracy with both models, obtaining 0.871 with gpt-oss:120b and

[14] https://ollama.com/library/gemma4:12b-mlx
[15] https://ollama.com/library/gpt-oss:120b-cloud

Table 3
Results for each configuration over the 88 questions benchmark. Accuracy was computed as the mean of the question-level evaluation scores across all 88 question-answer pairs, rather than as the proportion of questions receiving a binary correct/incorrect label; therefore, the reported values are not constrained to multiples of 1/88.

Model	Configuration	Accuracy
gpt-oss:120b	KG only	0.596
	RAG only	0.871
	deterministic (no feedback)	0.854
	deterministic (feedback)	**0.919**
	agentic (no feedback)	0.774
	agentic (feedback)	0.790
gemma4:12b-mlx	KG only	0.306
	RAG only	0.838
	deterministic (no feedback)	0.790
	deterministic (feedback)	**0.903**
	agentic (no feedback)	0.713
	agentic (feedback)	0.774

0.838 with gemma4:12b-mlx. However, the best hybrid configuration further improves these scores, reaching 0.919 and 0.903, respectively. This suggests that the KG complements simple textual retrieval with structured information and hierarchical reasoning.

The feedback mechanism also plays an important role, improving accuracy in both deterministic and agentic configurations. This suggests that the feedback node helps the system revise answers that would otherwise be incomplete or insufficiently grounded, though we do not directly verify this mechanism and leave qualitative error analysis to future work.

The comparison between the agentic and deterministic hybrid settings highlights the importance of controlled KG exploration. The deterministic configuration outperforms the agentic configuration for both models. This suggests that leaving KG exploration entirely to the LLM may be less reliable. One possible explanation is that the model must decide which tools to call and how much schema information is needed, however, we do not directly measure this variance and leave a closer investigation to future work; on the other hand, making KG exploration compulsory provides a more stable and complete schema context for SPARQL generation. The lower performance of the KG-only configuration further confirms this difficulty: KG-only reaches 0.596 with gpt-oss:120b but only 0.306 with gemma4:12b-mlx, suggesting that SPARQL generation remains a challenging task and that the larger model is better able to interpret the RDF schema and generate valid queries.

The results show that model size is not the only factor determining performance. Although gpt-oss:120b generally performs better than gemma4:12b-mlx, the gap becomes small in the best deterministic-feedback configuration. This shows that a smaller local model can be close to the performance of a larger model when the workflow is well constrained and controlled.

8. Conclusions

This paper presented a hybrid KG-RAG QA system for extreme weather events, combining SPARQL-based retrieval over an RDF KG derived from EM-DAT, the international disaster database, with vector-based RAG over textual reports from WWA, an initiative that produces attribution studies about extreme weather events around the world. The system was evaluated against KG-only and RAG-only baselines, as well as in agentic and deterministic hybrid configurations with and without a post-retrieval feedback loop. The results show that the best performance is achieved by the deterministic hybrid configuration with the feedback loop, improving on both KG- and RAG-only configurations, suggesting that structured graph retrieval and textual retrieval are complementary.

At the same time, the paper has several limitations. First, the evaluation is domain-specific. Both the KG and the benchmark are built around extreme weather events, recorded in EM-DAT or described in WWA. To evaluate the general effectiveness of the methodology, future work should adapt the same architecture to other domains, using existing KGs and KGQA benchmarks [24]. This would allow us to verify whether the advantages observed here depend on the climate-event setting and are biased towards the data we used or whether they can be generalized. Second, the current benchmark is quite small and is limited to simple factual questions. Future work should extend the dataset to include additional sources, and the benchmark to include questions of different types, including multi-hop and open-ended questions. Third, the ontological framework can be further strengthened and scaled to represent more diverse data sources. This would make the KG more expressive and support more complex forms of reasoning beyond simple factual QA.

Another future direction concerns experimenting with different types of GraphRAG pipelines. Rather than relying only on SPARQL retrieval, future work should include architectures that extend RAG to RDF sources directly, e.g. by explicating the triples or indexing vectors generated from the triples. In the opposite direction, textual sources like WWA could be used to extract structured information from them, and transform it into RDF triples to enrich the KG. This would allow the system to continuously enrich the KG from unstructured sources, unifying graph-building and graph-retrieval.

Finally, future work should test a wider range of models and allocation strategies. In fact, different nodes in a workflow may have different requirements: larger models may be necessary for complex query generation, while smaller local models may be sufficient for routing or synthesis when the retrieved evidence is clear.

Acknowledgments

This work was supported by CHIST-ERA grant CHIST-ERA-25-SOL-02 (CLASiK project), funded by the Agencia Estatal de Investigación (AGI) under grant PCI2025-167187-2 and by the I+D+i project PID2024-159530OB-I00, funded by MCIN/AEI/10.13039/501100011033.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI ChatGPT and Anthropic Claude in order to: Grammar and spelling check and Formatting assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Delforge, V. Wathelet, R. Below, C. L. Sofia, M. Tonnelier, J. A. van Loenhout, N. Speybroeck, Em-dat: the emergency events database, International Journal of Disaster Risk Reduction 124 (2025) 105509. URL: https://www.sciencedirect.com/science/article/pii/S2212420925003334. doi:https://doi.org/10.1016/j.ijdrr.2025.105509.

[2] B. Clarke, F. Otto, R. Stuart-Smith, L. Harrington, Extreme weather impacts of climate change: an attribution perspective, Environmental Research: Climate 1 (2022).

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[4] D. Thulke, Y. Gao, P. Pelser, R. Brune, R. Jalota, F. Fok, M. Ramos, I. van Wyk, A. Nasir, H. Goldstein, T. Tragemann, K. Nguyen, A. Fowler, A. Stanco, J. Gabriel, J. Taylor, D. Moro, E. Tsymbalov, J. de Waal, E. Matusov, M. Yaghi, M. Shihadah, H. Ney, C. Dugast, J. Dotan, D. Erasmus, Climategpt: Towards ai synthesizing interdisciplinary research on climate change, 2024. URL: https://arxiv.org/abs/2401.09646. arXiv:2401.09646.

[5] S. A. Vaghefi, Q. Wang, V. Muccione, J. Ni, M. Kraus, J. Bingler, T. Schimanski, C. Colesanti-Senni, N. Webersinke, C. Huggel, M. Leippold, chatclimate: Grounding conversational ai in climate science, 2023. URL: https://arxiv.org/abs/2304.05510. arXiv:2304.05510.

[6] M. Kurfali, S. Zahra, J. Nivre, G. Messori, ClimateEval: A comprehensive benchmark for NLP tasks related to climate change, in: K. Dutia, P. Henderson, M. Leippold, C. Manning, G. Morio, V. Muccione, J. Ni, T. Schimanski, D. Stammbach, A. Singh, A. R. Su, S. A. Vaghefi (Eds.), Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 194–207. URL: https://aclanthology.org/2025.climatenlp-1.13/. doi:10.18653/v1/2025.climatenlp-1.13.

[7] N. Li, S. Zahra, M. Brito, C. Flynn, O. Görnerup, K. Worou, M. Kurfali, C. Meng, W. Thiery, J. Zscheischler, G. Messori, J. Nivre, Using LLMs to build a database of climate extreme impacts, in: D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, M. Leippold (Eds.), Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 93–110. URL: https://aclanthology.org/2024.climatenlp-1.7/. doi:10.18653/v1/2024.climatenlp-1.7.

[8] S. Walter, H. Bast, GRASP: Generic Reasoning And SPARQL Generation Across Knowledge Graphs, Springer Nature Switzerland, 2025, p. 271–289. URL: http://dx.doi.org/10.1007/978-3-032-09527-5_15. doi:10.1007/978-3-032-09527-5_15.

[9] S. Liu, S. Semnani, H. Triedman, J. Xu, I. D. Zhao, M. Lam, SPINACH: SPARQL-based information navigation for challenging real-world questions, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15977–16001. URL: https://aclanthology.org/2024.findings-emnlp.938/. doi:10.18653/v1/2024.findings-emnlp.938.

[10] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629. arXiv:2210.03629.

[11] X. Pan, V. de Boer, J. van Ossenbruggen, Firesparql: A llm-based framework for sparql query generation over scholarly knowledge graphs, 2025. URL: https://arxiv.org/abs/2508.10467. arXiv:2508.10467.

[12] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. URL: http://dx.doi.org/10.1109/TKDE.2024.3352100. doi:10.1109/tkde.2024.3352100.

[13] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-bert: Enabling language representation with knowledge graph, 2019. URL: https://arxiv.org/abs/1909.07606. arXiv:1909.07606.

[14] M. E. Peters, M. Neumann, R. L. L. IV, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, 2019. URL: https://arxiv.org/abs/1909.04164. arXiv:1909.04164.

[15] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, 2024. URL: https://arxiv.org/abs/2408.08921. arXiv:2408.08921.

[16] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, H. Chen, Y. Xiao, C. Zhou, J. Dong, Y. Chang, X. Huang, A survey of graph retrieval-augmented generation for customized large language models, 2025. URL: https://arxiv.org/abs/2501.13958. arXiv:2501.13958.

[17] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge graph-guided retrieval augmented generation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 8912–8924. URL: https://aclanthology.org/2025.naacl-long.449/. doi:10.18653/v1/2025.naacl-long.449.

[18] L. Luo, Z. Zhao, G. Haffari, D. Phung, C. Gong, S. Pan, Gfm-rag: Graph foundation model for retrieval augmented generation, 2025. URL: https://arxiv.org/abs/2502.01113. arXiv:2502.01113.

[19] P. Pirozelli, M. M. José, I. Silveira, F. Nakasato, S. M. Peres, A. A. Brandão, A. H. Costa, F. G. Cozman, Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast,

and climate change, Data Intelligence 6 (2024) 29–63.
[20] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.
[21] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. URL: https://aclanthology.org/D18-1259/. doi:10.18653/v1/D18-1259.
[22] Y. Yang, W.-t. Yih, C. Meek, WikiQA: A challenge dataset for open-domain question answering, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 2013–2018. URL: https://aclanthology.org/D15-1237/. doi:10.18653/v1/D15-1237.
[23] A. Perevalov, D. Diefenbach, R. Usbeck, A. Both, Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers, 2022. URL: https://arxiv.org/abs/2202.00120. arXiv:2202.00120.
[24] A. Perevalov, X. Yan, L. Kovriguina, L. Jiang, A. Both, R. Usbeck, Knowledge graph question answering leaderboard: A community resource to prevent a replication crisis, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2998–3007. URL: https://aclanthology.org/2022.lrec-1.321/.

A. Prompts

A.1. Information alignment

The following prompt was used to find common information to our two data sources: EM-DAT and WWA. The information regarding the common events was later used as background context to generate question-answer pairs.

```
You are an information alignment assistant for climate and extreme weather event
    records. Your task is to compare two sources about one event:
1. A structured EM-DAT-style table row.
2. A textual World Weather Attribution context.

Extract only facts that are explicitly supported by BOTH sources.

A fact is common only if:
- the structured row contains the fact,
- the WWA text contains the same fact, an equivalent label, a clear paraphrase, or
    a compatible more specific version.

Do not extract facts that appear in only one source.
Do not infer facts from general background knowledge.
Do not invent missing values.
Do not merge two events unless the overlap is strong enough to indicate that both
    sources discuss the same event.
```

A.2. Benchmark generation

The following prompt was used to generate a benchmark dataset consisting of question-answer pairs starting from given input context.

```
Your task is to generate benchmark simple question-answer pairs for evaluating a
    hybrid KG + RAG question answering system.

The input contains facts that were identified as common between two sources:
1. structured EM-DAT-style event information;
2. textual WWA study context.

Generate questions only from facts that are explicitly supported by the
common information.
Do not create questions that require facts present in only one source.
Do not invent event details, numbers, dates, places, or attribution claims.

Return only a valid JSON array.
Do not include markdown.
Each item must have this schema:
{
  "question": string,
  "answer": string,
  "answer_type": "entity" | "date" | "number" | "boolean" |
  "list" | "short_text",
  "required_sources": ["KG", "RAG"],
  "evidence_from_common_facts": string,
  "difficulty": "single_hop" | "multi_hop" | "hybrid",
  "reasoning": string
}
Rules:
- The answer must be concise.
- Prefer factual questions about shared event identity, country, disaster type,
    hazard, location, date, impacts, or common descriptions.
- For numeric answers, use only the number and unit if the unit is
  necessary.
- Avoid generic questions like then following: "In which month and year
  did the drought event begin in Brazil?".
```

A.3. RDF schema

Below is the RDF schema provided in input within the system prompt in the system configurations that had access to the climate KG.

```
Allowed prefixes:
@prefix aemet: <http://aemet.linkeddata.es/ontology/> .
@prefix clasik: <https://example.org/clasik#> .
@prefix geo1: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix qudt: <http://qudt.org/schema/qudt/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sorelsc: <http://sweetontology.net/relaSci/> .
@prefix soteria: <https://soteria-public.pages.sintef.no/
    soteria_ontology_documentation#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

The main node of the knowledge graph is an event, which is an instance of the class
     clasik:ClimateEvent.
Each event is linked to a country URI through the gn:locatedIn property, which is
    described by a string through the property gn:name.
Each event is linked to its provinces through the property clasik:isInProvince.
Each event is linked to a start date and an end date through the properties soteria
    :startDate and soteria:endDate.

Each event is related to an impact through the sorelsc:hasImpact property.
Each impact is characterized as related to deaths through the property
clasik:deaths, to people affected through the property clasik:peopleAffected, to
people injured through the property clasik:injured and to economic loss through
```

```
the property clasik:economicDamage.

Each event is characterized with a Disaster Type through the property
clasik:hasDisasterType, with a Disaster Subtype through the property
clasik:hasDisasterSubtype, and with a Disaster Subgroup with the property
clasik:hasDisasterSubgroup.

If a measurement was recorded, an observation could be linked to an event through
    the aemet:observedProperty, and the observation is a quantity
value, linked to a string through the property clasik:hasResult.
```

A.4. Synthesizer Prompt

Below is the system prompt provided to the synthesizer node of the hybrid langraph workflow.

```
You must answer using the evidence provided from the KG/SPARQL component and/or the
     RAG component.

Return only valid JSON with the following fields:

{
  "short_answer": string,
  "answer": string,
  "kg_evidence": string,
  "rag_evidence": string,
  "context": string,
  "sources_used": array of strings,
  "confidence": "high" | "medium" | "low"
}

Rules:
- The short_answer must be the shortest factual answer possible.
- If the question asks for a number, date, country, event ID, disaster type, or yes
    /no answer, short_answer must contain only that value.
- The answer field may contain one or two explanatory sentences.
- If both KG and RAG contain useful information, output both of them.
- The context field must contain the retrieved information used to answer the
    question. This should be presented exactly as retrieved in the data source.
- If only one source contains useful information, say so.
- Do not add facts not supported by the evidence.
- Reference KG-derived facts in kg_evidence.
- Reference RAG-derived facts in rag_evidence, including source URL if available.
```

A.5. LLM judge

Below is the prompt used to evaluate the correctness of the answers generated by the different system configurations. The model used for this task is gemma4:e4b,[16] run on local hardware.

```
You are evaluating the correctness of a question-answering system.
You are given:
- a question;
- a gold answer;
- a short answer generated by the system.

Your task is to assign a score from 0 to 1 to the prediction and decide whether the
     predicted short answer is correct with respect to the gold answer.

Consider the prediction correct if:
- it exactly matches the gold answer;
- it is a clear paraphrase of the gold answer;
- it gives the same entity, date, number, percentage, location, or event type;
```

[16] https://ollama.com/library/gemma4:e4b

```
- it contains the same answer with minor formatting differences, such as
    punctuation, capitalization, plural forms, or units;
- for numeric answers, the value is equivalent even if the unit or formatting
    differs.

Consider the prediction incorrect if:
- it contradicts the gold answer;
- it gives a different entity, date, number, percentage, location, or event type;
- it is too vague to verify;
- it contains additional information that changes the meaning of the answer;
- it answers a different question;
- it says that the answer is unknown when the gold answer provides a value.

Return only valid JSON with the following schema:

{
  "verdict": "score",
  "reason": "brief explanation"
}
```