Ontology-Guided Knowledge Graph Construction from Structured Metadata and Climate Science Publications

Aayush Acharya[1], Qi Zhang[1], Eduard Dragut[1], Longin Jan Latecki[1],
{aayush.acharya,qi.zhang,edragut,latecki}@temple.edu
[1]Temple University,
Correspondence: aayush.acharya@temple.edu

Abstract

The rapid expansion of scientific literature far outpaces the speed of manual curation, creating a fundamental bottleneck in extracting structured knowledge from scientific texts. To address this, we introduce an agentic, ontology-guided framework for knowledge graph (KG) construction. Our framework utilizes specialized agents for document understanding, information extraction, ontology conformance, and entity linking. First, an LLM extracts factual knowledge in the form of triples with explicit textual evidence from research papers. Then the triplets undergo a multistage verification process to ground the triplets in the ontology. By enforcing schema constraints and requiring evidence in the form of text fragments, we ensure the data is both semantically consistent and traceable. In an agreement study over 1,000 extracted triplets, Gemini-2.5-flash achieves Cohen's κ scores of 0.640–0.700 against three human annotators, matching the human–human agreement range of 0.616–0.689. We applied this framework to 2,000 scientific publications and construct a novel, ontology-grounded KG for climate science with 133,990 nodes and 755,308 relations; of the 23,327 triplets that the publications contribute after validation, 98.96% are judged correct. The pipeline code and KG is publicly available[1].

1 Introduction

The volume and complexity of the scientific literature continue to grow at an unprecedented rate (Hanson et al., 2024; Ding et al., 2025), challenging the ability of researchers to synthesize knowledge across publications, models, and domains. Knowledge graphs (KGs) provide a structured representation of scientific knowledge that can support semantic search, retrieval-augmented generation, and scientific reasoning (Peng et al., 2023; Guan et al., 2025). From an NLP perspective, constructing such graphs from unstructured text requires robust information extraction (IE) systems capable of identifying entities, relations, and provenance while respecting domain semantics.

Recent progress in Large Language Models (LLMs) has significantly improved IE performance (Wei et al., 2024), enabling zero-shot and few-shot extraction of relational facts from scientific text (Kartchner et al., 2023). The large context window of newer LLMs allows the processing of documents without the need to distribute them into chunks. Moreover, the shift from simple text processing to "native" understanding of document layouts, tables, and visual hierarchies without needing separate OCR greatly improves IE (Ke et al., 2026). A parallel line of work has explored ontology-guided IE, where schemas constrain extraction to a fixed set of entity and relation types (van Cauter and Yakovets, 2024). Due to semantic consistency and interoperability, ontology-guided methods consistently show better extraction quality in domain specific setting (Bhattarai et al., 2024; Cotti et al., 2025).

Climate science exemplifies both the opportunity and the challenge. Recent works (Chen et al., 2025; Bulian et al., 2024) note that information about climate science is underrepresented in current LLMs. Climate models form the backbone of climate research and assessment, yet critical information about model genealogy, shared components, physical parameterizations, and experimental setups is dispersed across thousands of publications (Kuma et al., 2023). While controlled vocabularies such as Earth System Grid Federation (ESGF), a global infrastructure that standardizes and distributes climate model datasets, and Coupled Model Intercomparison Project Phase 6 (CMIP6), an international framework defining standardized climate model experiments and metadata, exist, they primarily describe datasets and metadata rather than the rich relational knowledge embed-

[1]https://github.com/aayushacharya/climatekg

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5615–5642
November 6-10, 2026 ©2026 Association for Computational Linguistics

ded in the scientific literature. To date, there is very little work done on comprehensive KG integrating these vocabularies with systematically extracted, literature-derived knowledge about climate models, which is the focus of this work.

Climate science further poses a distinctive challenge for LLM-based extraction due to its reliance on domain-specific, long-tail entities and implicit dependencies that are rarely stated explicitly. For example, in the statement:

> CESM2 *shares atmospheric component* with CAM6.

an extraction system must infer that CESM2 and CAM6 are climate models, that the relation corresponds to an ontology-defined predicate such as *hasComponent*, and that CAM6 represents an atmospheric component rather than a standalone system. Such inferences go beyond surface-level pattern matching and expose a key limitation of open extraction and general-purpose LLMs, which often lack familiarity with specialized climate terminology and its implicit semantics. This motivates the need for ontology-guided extraction that explicitly encodes domain knowledge, constrains relation semantics, and supports reasoning over long-tail scientific entities.

In this paper, we address these gaps by introducing an evidence-based, ontology-guided IE framework and using it to construct a novel climate KG. Our approach requires each extracted subject–predicate–object triplet to be accompanied by explicit textual evidence and applies automated LLM-based validation to verify both the triplet and its evidence. This design reduces human annotation effort by narrowing manual review to concise evidence spans rather than entire documents, while maintaining high quality through automated filtering. We show that LLM-based validation achieves high agreement with human annotators, enabling scalable KG construction with minimal human intervention. In summary, our contributions are:

- **Formulation.** We formalize domain-specific KG construction as an *evidence-grounded extraction task*. By first extracting triples with explicit textual evidence, we streamline complex literature curation into a more efficient verification process for domain-specific KGs.
- **Framework.** We propose an agentic, ontology-guided framework that utilizes specialized LLM agents for document understanding, schema conformance, and hybrid entity linking. The system integrates authoritative climate science vocabularies and a dedicated evidence validator, thereby ensuring that all extracted relations are traceable to the source text.
- **Experiments.** In a 1,000-triplet validation study, Gemini-2.5-flash achieves Cohen's κ scores of 0.640–0.700 against three human annotators, comparable to human–human agreement. Furthermore, 98.96% of the 23,327 triplets that survive validation are judged correct, and a stage ablation shows where that precision comes from, at a verification cost of roughly 1% of extraction.
- **Climate KG.** We introduce a novel climate KG that unifies ESGF metadata with CMIP6 controlled vocabularies, concepts from GCMD (Global Change Master Directory), a NASA-maintained controlled vocabulary that standardizes Earth science metadata to support consistent data discovery and interoperability, and GeoNames-based geolocation. It is further enriched with evidence-grounded relations extracted from 2,000 scientific publications, yielding 133,990 nodes and 755,308 relations that connect models, experiments, variables, institutions, and regions with literature-derived, text-verifiable links; 73,283 of those nodes and 298,747 of those relations come from the publication corpus rather than from the initialization sources.

2 Problem Formulation

Let $\mathcal{O} = (\mathcal{C}, \mathcal{R})$ denote a domain ontology, where $\mathcal{C}$ is the set of entity classes and $\mathcal{R}$ is the set of typed relations with explicit domain and range constraints. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be a corpus of scientific publications.

Our objective is to construct a domain-specific KG $G = (V, E)$, where each edge $r \in E$ is a triple (h, r, t) grounded in explicit textual evidence. For each document d_i, we aim to extract a set of evidence-backed triples

$$\mathcal{T}_i = \{(h, r, t, \varepsilon)\},$$

where ε is a sentence or paragraph(s) in d_i that explicitly supports the relation. The KG is updated

iteratively. Given a current KG G, the enriched KG is obtained as

$$G = G \cup \bigcup_{i=1}^{N} \mathcal{T}_i.$$

The initial G is extracted from established, trusted scientific sources, as discussed in Section 3.4.

3 Methodology

3.1 Ontology Construction

Purely open IE suffers from uncontrolled relation phrasing and weak schema alignment, resulting in inconsistent and difficult-to-curate outputs. Ontology-guided extraction mitigates these issues by constraining entity types and relations to a predefined vocabulary and by enforcing semantic consistency through domain and range validation. Recent work has shown that incorporating ontological structure into LLM-based extraction substantially improves relation consistency and downstream usability (Nie et al., 2024).

We design a two-level ontology that combines standardized scholarly metadata with domain-specific concepts, using OpenAlex-derived paper-centric entities as a stable backbone. Domain-specific classes are introduced only for screened, domain-relevant publications, ensuring scalable extraction while minimizing noise and schema drift.

3.1.1 Domain-Specific Ontology

The domain ontology is constructed with input from a domain expert by unifying authoritative climate resources: ESGF metadata, GCMD (Nagendra et al., 2001), CMIP6 Controlled Vocabularies, and GeoNames. These sources provide standardized model, experiment, variable, component, institution, and geographic concepts; additional details are provided in Appendix A.18. The resulting ontology $\mathcal{O}$ defines domain and range constraints such as $\text{dom}(\texttt{hasComponent}) = \texttt{Model}$ and also $\text{range}(\texttt{hasComponent}) = \texttt{Component}$. These constraints serve as soft structural priors during information extraction.

3.2 Scientific Publication Corpus Collection

We assemble a targeted corpus of domain-specific scientific literature to ensure coverage of authoritative and highly relevant sources. The first phase begins with a small seed set of core papers and expands through two levels of citation mining with LLM-based relevance filtering. After the first phase selected 1,007 papers, we used the top OpenAlex concepts from this set to retrieve recent, context-matched candidates and applied the same screening procedure, adding 993 papers for a 2,000-paper corpus; further details are in Appendix A.1. The small seed of core papers for climate-domain task was taken from the following sources. The sources were selected on consultation with climate experts: (1) curated references from the IPCC Sixth Assessment Report (Intergovernmental Panel on Climate Change, 2023) and (2) CMIP6 publications[1].

3.3 Ontology-Guided IE Pipeline

We propose an ontology-guided IE pipeline that integrates semantic constraints and explicit evidence grounding throughout the extraction process. The agentic system is presented in Figure 1. Now we give a brief overview of the role of each agent represented by an LLM.

Triplet and evidence extractor agent The detailed architecture for information extraction pipeline is given in Figure 1. For each document $d \in D$, we predict triplets $(h_1, r_1, t_1), \ldots, (h_N, r_N, t_N)$ with evidence spans $\epsilon_1 \ldots \epsilon_n$ supporting each triplet. For each triplet (h, r, t), we also predict its ontology classes c_h and c_t.

Example. From a single sentence the agent may predict

$$(h_1, r_1, t_2) = (\text{CESM2}, \texttt{hasComponent}, \text{CAM6}),$$

$$(h_2, r_2, t_2) = (\text{CAM6}, \texttt{appliesToRealm}, \text{atmos})$$

together with ontology-aligned type assignments

$$c_{h1} = \texttt{Source}, \quad c_{t1} = \texttt{SourceComponent},$$

$$c_{h2} = \texttt{SourceComponent}, \quad c_{t2} = \texttt{Realm}$$

and an explicit supporting evidence span $\epsilon_1, \epsilon_2 =$ *CESM2 uses CAM6 as its atmospheric component to simulate large-scale atmospheric circulation.*

The agent uses Gemini's native document understanding to extract these features. Beyond the triplets, there are other features extracted during this stage which is provided in Appendix A.2.

[1] https://cmip-publications.llnl.gov/

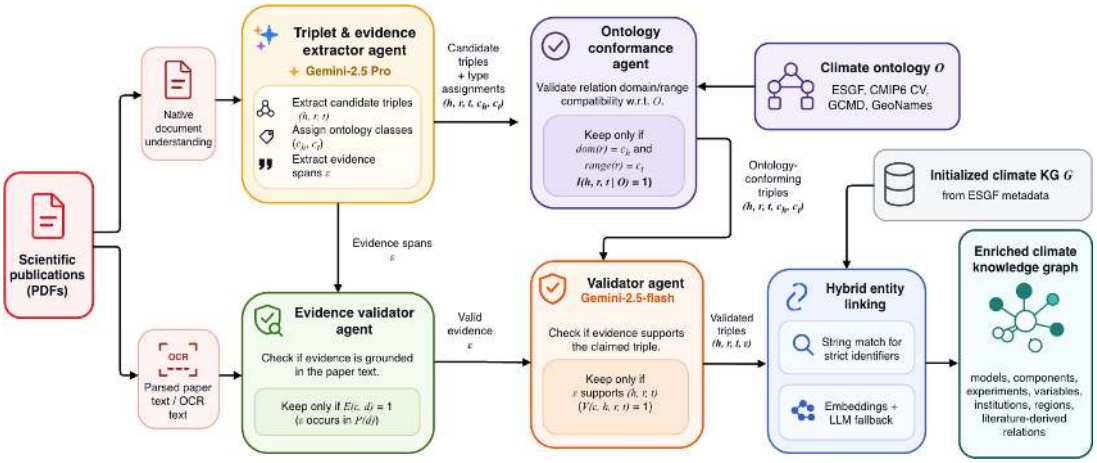


Figure 1: Ontology-guided, evidence-based KG construction pipeline. An LLM extracts candidate triples and evidence from scientific PDFs, which are validated for ontology conformance and textual grounding before final entity linking and acceptance.

Ontology conformance agent Candidate triples are first checked for ontology conformance:

$$\mathbb{I}(h, r, t \mid \mathcal{O}) = \mathbf{1}[\text{dom}(r) = c_h \wedge \text{range}(r) = c_t].$$

Only if $\mathbb{I}(h, r, t \mid \mathcal{O}) = 1$, a triplet is passed to the evidence validator agent. The check is deterministic and is specified in Appendix A.19.

Example. Assume the ontology $\mathcal{O}$ defines the relation `hasComponent` : `Source` $\rightarrow$ `SourceComponent`. Consider the following sentence in document d:

> CAM6 is the atmospheric component used within the CESM2 climate modeling framework.

From this sentence, an LLM-based extractor may plausibly generate the triple:

$$(h, r, t) = (\text{CAM6},\ \texttt{hasComponent},\ \text{CESM2}),$$

by incorrectly assigning the syntactic subject as the head entity. The inferred ontology classes are:

$$c_h = \texttt{SourceComponent}, \quad c_t = \texttt{Source}.$$

However, this assignment violates the ontology constraints for relation `hasComponent`. As a result, the ontology conformance agent returns:

$$\mathbb{I}(h, r, t \mid \mathcal{O}) = 0,$$

and the triplet is discarded before any evidence or semantic check is run on it.

Evidence validator agent Unlike other multi-agent information extraction frameworks (Lu and Wang, 2025), which enforce ontology-level constraints without explicitly verifying textual grounding, our framework introduces a dedicated evidence validator agent. It ensures that the extracted sentence/paragraph is not introduced through hallucination or paraphrasing. Given a candidate evidence span ε and the parsed content of the paper $\mathcal{P}(d)$, the agent verifies whether ε occurs verbatim or as a faithful substring within the document. Formally, we define an evidence presence function

$$\mathbb{E}(\varepsilon, d) = \begin{cases} 1, & \text{if } \varepsilon \subseteq \mathcal{P}(d), \\ 0, & \text{otherwise.} \end{cases}$$

Only triples (h, r, t, ε) satisfying $\mathbb{E}(\varepsilon, d) = 1$ are forwarded to the validator agent, so that every triple reaching semantic validation carries an evidence span that is present in the source document. Triples with missing or hallucinated evidence are discarded at this point. Conformance and evidence grounding are independent tests, so applying them in either order yields the same surviving set; we run conformance first because it is the cheaper of the two and removes the largest share of candidates.

Validator agent It receives the triples that pass both the ontology conformance agent and the evidence validator agent. It validates whether each triplet (h, r, t) of classes c_h and c_t is indeed supported by evidence ε:

$$\mathbb{V}(\varepsilon, h, r, t) = \begin{cases} 1, & \text{if } \varepsilon \text{ supports } (h, r, t), \\ 0, & \text{otherwise.} \end{cases}$$

Triples with $\mathbb{V}(\varepsilon, h, r, t) = 0$ are discarded.

3.4 Graph Initialization

We initialize the KG G from over four million ESGF dataset metadata records[2], deterministically mapping standardized model, experiment, variable, institution, and spatial attributes into ontology-aligned entities and relations. This trusted backbone grounds subsequent literature-derived enrichment by giving the entity linker canonical climate identifiers instead of relying on the LLM to generate or memorize standardized names. Additional construction details are provided in Appendix A.3.

3.5 Hybrid Entity Linking

This section describes the Entity Linking agent (Figure 7, Appendix). For each triple (h, r, t) and evidence ϵ, we consider (h, t) as candidate entities to be linked to existing nodes of $G = (V, E)$. For linking, we consider string-based similarity or embedding-based similarity, depending on the type of entity.

We use NV-Embed-V2 (Lee et al., 2025) to produce deep embeddings, since it showed consistently good performance in MTEB benchmark (Muennighoff et al., 2022; Enevoldsen et al., 2025) and was also utilized by (Pan et al., 2025b) for Climate tasks. In general, the embedding model performs well, but the embedding distance is widely inconsistent for similar items. Given a deep embedding function ψ, let $\hat{v}$ be the most similar graph node to an entity e:

$$\hat{v} = \arg\max_{v \in V} \text{sim}(\psi(e), \psi(v)),$$

If $\text{sim}(\psi(e), \psi(\hat{v})) \geq \tau$, $\hat{v}$ is linked to e. However, if $\text{sim}(\psi(e), \psi(\hat{v})) < \tau$, we retrieve the top-$k$ candidates $v_1, \ldots, v_k$ using a fallback threshold τ_f and ask an LLM to identify the best match given the provided evidence. We use LLM-based candidate selection, as shown in Figure 2, because embedding similarity can be unreliable for specialized climate terms, where distributional proximity does not always reflect ontological compatibility. Each design choice in the pipeline is settled empirically rather than by preference: the linking thresholds come from the sensitivity analysis of Appendix A.5 (Table 10), the composition of verification stages from the ablation of Table 4, and the choice of a deterministic test over an LLM call for ontology conformance and evidence grounding from the cost comparison in Appendix A.11. We empirically set the automatic linking threshold $\tau = 0.80$ and the LLM fallback threshold $\tau_f = 0.55$. These thresholds are selected using the sensitivity analysis(refer to Appendix, Table 10), where match rates rise sharply above 0.80 and bins down to 0.55 still recover useful LLM-confirmed matches. For strict identifiers such as climate model names, we use string matching rather than embedding similarity; details are provided in Appendix A.4.

It is possible that some facts are repeated by multiple papers during this stage. The handling is provided in Appendix A.7.

4 Experimental Evaluation

4.1 LLM and Human Annotator Agreement

To assess whether LLM-based validation can scale KG quality control, we compare Gemini-2.5-flash with three human annotators on 1,000 extracted triplets sampled from IPCC-linked climate-science papers (Intergovernmental Panel on Climate Change, 2023) of the corpus of Section 3.2. The sample is drawn from the raw output of the triplet and evidence extractor agent, before any verification stage, so it contains triplets that the pipeline later keeps as well as triplets it removes. Each decision is the binary judgement of whether the evidence span supports the triplet, which is the task of the validator agent in Figure 1. As shown in Table 2, Gemini-human Cohen's κ scores range from 0.640–0.700, comparable to the human-human range of 0.616–0.689; against the human majority vote, Gemini reaches $\kappa = 0.798$ and 0.977 observed agreement. This supports its use as a scalable surrogate validator for large-scale KG construction.

4.2 Evaluation metric

We evaluate extraction quality using a triplet precision, which we call *triplet correctness* metric and denote as C_{triplet}. It measures the proportion of triplets generated by the system that are judged to be correct by an evaluator.

Let T denote the set of triplets output by a given extraction stage, and let $T_{\text{valid}} \subseteq T$ denote the subset of those triplets validated as correct by the evaluator. A triplet is considered correct if the relation between the source and target entities is supported

[2]Retrieved from the LLNL ESGF index through the `esg-search` Solr API (`https://esgf-node.llnl.gov/esg-search/search`, queried through the AIMS2 proxy at `https://aims2.llnl.gov/proxy/search`) in a snapshot taken between April and May 2025. The CMIP6 controlled vocabularies are the WCRP `CMIP6_CVs` release (`https://github.com/WCRP-CMIP/CMIP6_CVs`) of the same period.

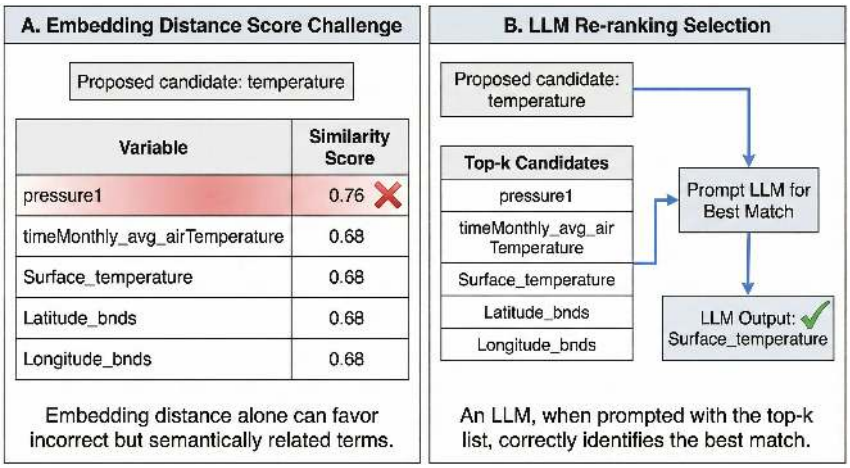


Figure 2: Hybrid entity linking pipeline showing how embedding distance fails in cases where the similarity score is low and LLM selection helps.

Table 1: Knowledge saturation over the 2,000-publication entity-linking run. Novelty is computed within each paper interval as the proportion of unique items that had not appeared in earlier papers. Relation novelty is computed over unique source–relation–target assertions.

Paper Interval	New Entities / Paper	Entity Novelty (%)	New Mentions / Paper	Mention Novelty (%)	Relation Novelty (%)
1–500	16.89	61.9	19.88	43.3	84.4
501–1,000	12.31	43.7	13.98	26.9	77.0
1,001–1,500	10.58	46.1	12.40	28.9	77.5
1,501–2,000	8.81	40.3	10.10	24.1	73.1

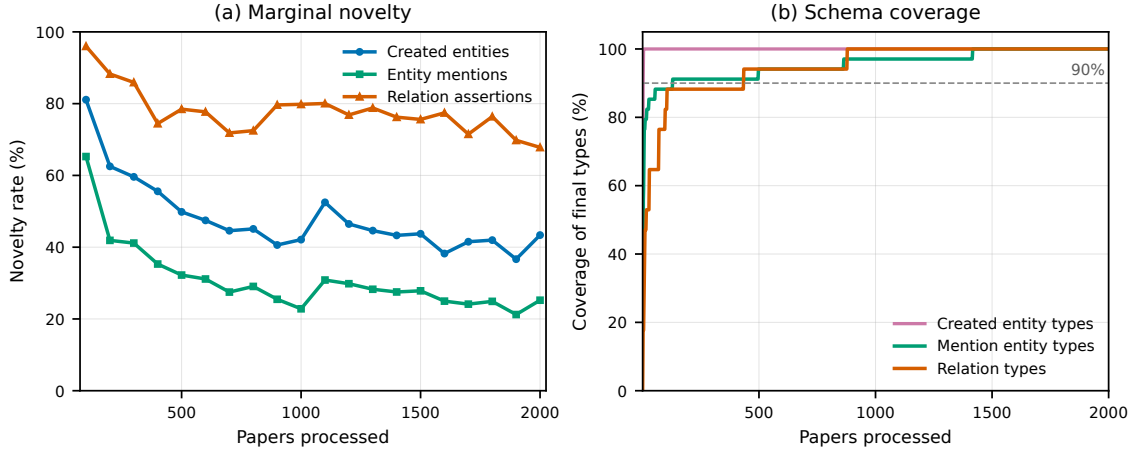


Figure 3: Knowledge saturation as papers are added to the KG. Left: novelty rates are computed over non-overlapping 100-paper windows. Right: schema coverage is the cumulative fraction of final entity or relation types observed after each paper.

by the corresponding paragraph in the source document, allowing for implicit references and anaphora resolution. The triplet correctness score is defined as:

$$C_{\text{triplet}} = \frac{|T_{\text{valid}}|}{|T|}$$

The metric is precision-like in nature, as it evaluates correctness over generated triplets without assuming an exhaustively annotated gold standard. This makes C_{triplet} well suited to the evaluation of ontology-guided information extraction from scientific publications, where the emphasis is on producing high-quality, curator-ready KG assertions rather than maximizing coverage.

4.3 Results and Discussion

Table 3 shows the effect of each validation stage over 2,000 publications. The extractor produces

Table 2: Agreement among human annotators and Gemini-2.5-flash on 1,000 triplet validation decisions.

Comparison	Obs. Agreement	Cohen's κ
Ann. 1 vs. Ann. 2	0.946	0.616
Ann. 1 vs. Ann. 3	0.956	0.667
Ann. 2 vs. Ann. 3	0.960	0.689
Gemini vs. Ann. 1	0.950	0.640
Gemini vs. Ann. 2	0.952	0.645
Gemini vs. Ann. 3	0.962	0.700
Gemini vs. human majority	0.977	0.798

39,964 candidate triplets; ontology conformance, evidence validation, and final validation reduce this to 23,327 high-quality triplets with a final correctness score of 98.963%. The resulting KG contains 133,990 nodes and 755,308 relations across 55 node types and 59 relation types (Table 5). Separating the two sources of content, the deterministic initialization from ESGF, CMIP6 controlled vocabularies, GCMD, and GeoNames accounts for 60,707 nodes and 456,563 relations, while the publication corpus contributes 73,283 nodes and 298,747 relations, including 96,100 paper-to-entity mention edges and 6,530 distinct entity-to-entity relations asserted by the papers themselves.

Stage ablation. Because every candidate triplet stores the verdict of every stage, we can evaluate all eight combinations of the three verification stages on the same 39,964 triplets with the same C_{triplet} judge, without re-running any agent (Table 4). The stages act on different axes rather than on a single quality dimension. The validator agent is the precision component: removing it from the full pipeline leaves 27,900 triplets at 94.72%, so it contributes $+4.24$ points of C_{triplet}. Ontology conformance contributes almost no precision, but that is not its function: 93.4% of the 11,050 triplets it removes are judged textually correct, yet their (source type, relation, target type) signature lies outside the ontology, so they cannot be typed into the KG at all. Its role is schema compatibility rather than error correction. Evidence validation likewise leaves final precision unchanged (24,172 at 98.96% without it), and instead supplies the traceability guarantee: without it, 845 triplets whose evidence span cannot be matched against the parsed source enter the KG. Precision, schema compatibility, and traceability therefore have to be read as three separate outcomes of the pipeline. The details about what the pipeline filters remove is given in Appendix A.8. The verification stack is also inexpensive relative to what it corrects. Appendix A.11 reports cost, runtime, and ablation effect per stage.

Knowledge Saturation We further measure knowledge saturation as papers are incrementally linked into the KG. Here, created entities are new KG nodes introduced during entity linking, while mentions are paper-level entity references resolved against the KG. Novelty is computed within each paper interval as the number of unique items not seen in earlier papers divided by the total number of unique items observed in that interval. As shown in Table 1 and Figure 3, the average number of new unique created entities per paper drops from 16.89 in the first 500 papers to 8.81 in the final 500 papers, while the entity novelty rate drops from 61.9% to 40.3%. Entity mention novelty shows a similar decline, from 43.3% to 24.1%, and relation assertion novelty declines from 84.4% to 73.1%. At the schema level, all 9 created entity types are observed by paper 4, 90% of the final 34 mention entity types are observed by paper 129, and 90% of the final 17 relation types are observed by paper 434. These trends indicate that additional papers increasingly reuse the existing climate KG vocabulary and schema, while continuing to add validated assertions. The same redundancy is visible at the level of individual assertions: of the 6,530 distinct literature-derived edges, 11.0% are asserted independently by two or more papers, and 1.6% by five or more. Facts missed in one paper are therefore often recoverable from overlapping evidence in later papers, which partially mitigates per-document recall limitations at corpus scale.

Comparison with baselines Table 6 compares our framework with KARMA (Lu and Wang, 2025), the closest available agentic baseline because it is also a multi-agent framework for extracting KG-style assertions from scientific publications. We are not aware of an existing agentic KG construction system designed specifically for climate-science papers, so we adapt KARMA's prompts to the climate domain and evaluate both systems on the first 1,007 publications after final validation. Our system extracts 17,141 triplets with 99.067% correctness, while KARMA extracts 3,820 triplets with 92.277% correctness, showing that explicit evidence grounding and ontology-conformance validation improve precision while producing a much larger validated KG.

We also compare against ClimateIE (Pan et al., 2025a), a climate-specific IE benchmark rather

Table 3: Per-stage attrition over the full 2,000-publication corpus, evaluated by Gemini-2.5-flash. Each row reports the output of the cumulative pipeline through that stage.

Stage	**Input Triplets**	**Filtered**	**Output Triplets**	C_{triplet} **(%)**
Triplet and evidence extractor agent	–	–	39,964	94.05
Ontology conformance agent	39,964	11,050	28,914	94.29
Evidence validator agent	28,914	1,014	27,900	94.72
Validator agent	27,900	4,573	23,327	98.96

Table 4: Stage ablation over the same 39,964 candidate triplets. O = ontology conformance agent, E = evidence validator agent, V = validator agent. Every triplet stores the verdict of every stage, so all eight variants are evaluated post hoc with the same C_{triplet} judge.

Stages applied	**Triplets kept**	C_{triplet} **(%)**
none (extractor)	39,964	94.05
O	28,914	94.29
E	38,386	94.38
V	32,436	98.86
O+E	27,900	94.72
O+V	24,172	98.96
E+V	31,095	98.88
O+E+V (full)	**23,327**	**98.96**

Table 5: Aggregate statistics of the constructed climate KG after enrichment with 2,000 scientific publications. The split rows separate what the deterministic initialization of Section 3.4 contributes from what the extraction pipeline adds, and are computed from the graph build files, which hold two relations more than the loaded database.

Statistic	**Value**
Scientific publications	2,000
KG nodes	133,990
from initialization	60,707
added from publications	73,283
KG relations	755,308
from initialization	456,563
added from publications	298,747
Node types	55
Relation types	59
Validated publication triplets	23,327

Table 6: Comparison with KARMA over the first 1,007 scientific publications after final validation

Metric	**KARMA**[†] (Lu and Wang, 2025)	**Ours (after validator)**
Output Triplets	3,820	17,141
Valid	3,525	16,981
Invalid	295	160
C_{triplet} (%)	92.277	**99.067**

[†]Prompts changed for the climate domain.

Table 7: Comparison with ClimateIE over the 25 papers in the ClimateIE human corpus, using the ClimateIE ontology for both systems and Gemini-2.5-flash validation.

Metric	**ClimateIE** (Pan et al., 2025a)	**Ours (with ClimateIE ontology)**
Output Triplets	5,729	637
Valid	4,297	592
Invalid	1,432	45
C_{triplet} (%)	75.004	**92.936**

than an agentic KG-construction system. For this experiment, we run our pipeline on the 25-paper ClimateIE human corpus using the ClimateIE ontology for both systems, and evaluate source–relation–target triples with the same Gemini-2.5-flash validator. As shown in Table 7, our system produces fewer triplets but reaches substantially higher correctness: 92.936% compared with 75.004%. This indicates that the proposed evidence-centered agents trade some raw extraction volume for much cleaner curator-ready assertions, outperforming both the general agentic baseline and the climate-specific IE baseline in precision.

The 1,000 human-annotated triplets of Section 4.1 support the same conclusions with human labels, since they were sampled from raw extractor output and therefore cover triplets at every filter outcome. Against the majority vote of three annotators, triplets kept by the full pipeline are 99.0% correct (625 of 631), which closely tracks the 98.96% judged by the LLM evaluator.

4.4 Recall and Coverage

C_{triplet} is precision-like, so we complement it in two ways: we measure entity-level recall against external gold that we did not produce, and we decompose what the verification stages remove.

Recall against external gold. The CMIP Publication Hub[3] hosts publication metadata submit-

[3]https://cmip-publications.llnl.gov/

Table 8: Entity-level recall against CMIP Publication Hub metadata for 139 papers, computed over gold entities whose surface form is recoverable from the parsed paper text.

Entity type	Gold	Extracted	Recall (%)
Source	797	774	97.11
Experiment	196	162	82.65
Frequency	85	57	67.06
Activity	55	30	54.55
Micro avg.	**1,133**	**1,023**	**90.29**

ted through a form by the papers' own authors, who select projects, activities, experiments, frequencies, sources, realms, and variables from controlled checkbox lists. This gives an independently curated annotation that is not derived from our pipeline. We scraped the hub for DOIs, queried OpenAlex for open-access PDFs, and retained the 139 papers that both downloaded and carried entity metadata. Because a curator can tag a paper with an entity that is never written in the paper (a related model family or a nearby variant, for instance), we first check whether each gold entity occurs in the parsed papers. Only 50.78% of the curated gold entities appear in the paper text at all, which is itself evidence that exhaustive per-document gold is ill-posed in this domain. Over the entities that are recoverable from the text, our pipeline reaches 90.29% micro-averaged recall, as shown in Table 8.

5 Related work

Automatic KG construction from text is commonly treated as a pipeline of entity extraction, relation extraction, linking, and refinement, with persistent challenges around schema conformance, scalability, and factual reliability (Zhong et al., 2023). Recent LLM-based systems address these challenges by decomposing KG construction into extraction, schema alignment, canonicalization, and validation stages (Zhang and Soh, 2024).

Ontology-guided benchmarks further emphasize that extracted facts must be evaluated not only for correctness but also for conformance to a target schema (Mihindukulasooriya et al., 2023). SciER shows that entity and relation extraction over full-text scientific papers remains hard for both supervised models and LLMs, even for generic dataset, method, and task entities (Zhang et al., 2024). ClimateIE provides a climate-specific benchmark for NER, relation extraction, and entity linking (Pan et al., 2025a), and taxonomy-driven climate IE work shows that grounding extraction in curated vocabularies can reduce hallucination and improve precision (Pan et al., 2025b). KARMA (Lu and Wang, 2025) uses multiple agents for scientific KG enrichment, but it is not designed for a climate-specific system with domain ontology grounding, evidence-level validation, and entity linking into a curated climate KG. Our work builds on this line of research by combining ontology-constrained extraction, explicit evidence grounding, hybrid entity linking, and scalable validation to construct a large climate-science KG from scientific publications.

6 Conclusion

We presented an evidence-based, ontology-guided framework for constructing climate-science KGs from scientific publications. By requiring explicit evidence, validating ontology conformance, and linking entities to a trusted climate KG backbone, the framework produces high-precision, traceable relations at scale. Across 2,000 publications, the pipeline constructs a KG with 133,990 nodes, 755,308 relations, while maintaining high triplet correctness after validation.

The results show that agentic extraction benefits from domain structure: ontology constraints reduce schema drift, evidence validation filters unsupported assertions, and hybrid entity linking connects literature-derived facts to canonical climate entities. Our agreement study supports Gemini-2.5-flash as a scalable validator, and comparisons with KARMA and ClimateIE show that the proposed framework achieves substantially cleaner curator-ready assertions than both a general multi-agent scientific IE baseline and a climate-specific IE baseline. These findings demonstrate the practicality of evidence-centered extraction for constructing auditable, domain-specific scientific KGs.

Acknowledgement

We thank Ryan Berkheimer from NOAA National Climatic Data Center, and Paul Smith from WCRP Coupled Model Intercomparison Project (CMIP), for countless discussions and valuable feedback. We also thank Adamu Mustapha for his valuable contribution as our climate expert. We further thank Huitong Pan, Yasmin Mohammadi, and Sanobar Rustamova for their valuable contributions to our project. This work was supported by the U.S. National Science Foundation awards III-2107213 and ITE-2333789.

Limitations

Our pipeline performs evidence grounding, ontology checks, entity linking, and validation in addition to extraction. Two of these verification stages are deterministic and the third is a short call to a small model, so the added cost is about 1% of extraction (Appendix A.11); the pipeline is nevertheless more expensive than direct extraction, and it inherits the GPU cost of local parsing and screening. We do not compute triplet-level recall, since open-ended scientific text would require exhaustive gold triplets for each document. Instead we report entity-level recall against external curated gold restricted to what the text actually states (90.29% micro-average, Section 4.4), and we quantify how many judge-correct triplets each stage removes. These two measurements do not cover relations that no paper in the corpus states explicitly, nor triplets whose entities fall outside the CMIP metadata families. Finally, LLM validators may still over-accept plausible but unstated relations or reject relations requiring dispersed evidence or domain knowledge.

Ethical Considerations

The work uses publicly available publications, bibliographic metadata, and community resources such as ESGF/GCMD, CMIP6 controlled vocabularies, and GeoNames. Human involvement is limited to low-risk validation of evidence-supported triplets from non-sensitive scientific text. We mitigate risks from third-party LLM extraction and validation by requiring explicit textual evidence, checking evidence grounding, and releasing derived resources with appropriate attribution and source-license constraints. We declare no conflicts of interest.

References

Kriti Bhattarai, Inez Y. Oh, Zachary B. Abrams, and Albert M. Lai. 2024. Document-level Clinical Entity and Relation extraction via Knowledge Base-Guided Generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 318–327, Bangkok, Thailand. Association for Computational Linguistics.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Hübscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauß. 2024. Assessing Large Language Models on Climate Information. *arXiv preprint*. ArXiv:2310.02932 [cs].

Zhou Chen, Xiao Wang, Yuanhong Liao, Ming Lin, and Yuqi Bai. 2025. ClimateChat: Designing Data and Methods for Instruction Tuning LLMs to Answer Climate Change Queries. *arXiv preprint*. ArXiv:2506.13796 [cs].

Luca Cotti, Idilio Drago, Anisa Rula, Devis Bianchini, and Federico Cerutti. 2025. OntoLogX: Ontology-Guided Knowledge Graph Extraction from Cybersecurity Logs with Large Language Models. *arXiv preprint*. ArXiv:2510.01409 [cs] version: 1.

Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. 2025. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ziyi Guan, Jason Chun Lok Li, Zhijian Hou, Pingping Zhang, Donglai Xu, Yuzhi Zhao, Mengyang Wu, Jinpeng Chen, Thanh-Toan Nguyen, Pengfei Xian, Wenao Ma, Shengchao Qin, Graziano Chesi, and Ngai Wong. 2025. KG-RAG: Enhancing GUI agent decision-making via knowledge graph-driven retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5396–5405, Suzhou, China. Association for Computational Linguistics.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Intergovernmental Panel on Climate Change. 2023. Annex II: Models. In Intergovernmental Panel on Climate Change (IPCC), editor, *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 2087–2138. Cambridge University Press, Cambridge.

David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. 2023. Zero-shot information extraction for clinical meta-analysis using large language models. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, Toronto, Canada. Association for Computational Linguistics.

Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. 2026. Large Language Models in Document Intelligence: A Comprehensive Survey, Recent Advances, Challenges, and Future Trends. *ACM Transactions on Information Systems*, 44(1):1–64.

Peter Kuma, Frida A-M Bender, and Aiden R Jönsson. 2023. Climate model code genealogy and its relation to climate feedbacks and sensitivity. *Journal of Advances in Modeling Earth Systems*, 15(7):e2022MS003588.

Jiayin Lan, Jiaqi Li, Baoxin Wang, Ming Liu, Dayong Wu, Shijin Wang, and Bing Qin. 2025. NLP-AKG: Few-Shot Construction of NLP Academic Knowledge Graph Based on LLM. *arXiv preprint*. ArXiv:2502.14192 [cs].

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint*. ArXiv:2405.17428 [cs].

Yuxing Lu and Jinzhuo Wang. 2025. KARMA: Leveraging Multi-Agent LLMs for Automated Knowledge Graph Enrichment. *arXiv preprint*. ArXiv:2502.06472 [cs].

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. 2023. Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. In *The Semantic Web – ISWC 2023*, pages 247–265, Cham. Springer Nature Switzerland.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Kishan Nagendra, Omran A. Bukhres, Srinivasan Sikkupparbathyam, Marcelo Areal, Zina Ben-Miled, Lola M. Olsen, Chris Gokey, David Kendig, Tom Northcutt, Rosy Cordova, and Gene Major. 2001. Nasa global change master directory: an implementation of asynchronous management protocol in a heterogeneous distributed environment. *Proceedings 3rd International Symposium on Distributed Objects and Applications*, pages 136–145.

Jixuan Nie, Xia Hou, Wenfeng Song, Xuan Wang, Xinyu Zhang, Xingliang Jin, Shuozhe Zhang, and Jiaqi Shi. 2024. Knowledge Graph Efficient Construction: Embedding Chain-of-Thought into LLMs. *Proceedings of the VLDB Endowment. ISSN*, 2150:8097.

Huitong Pan, Mustapha Adamu, Qi Zhang, Eduard Dragut, and Longin Jan Latecki. 2025a. ClimateIE: A Dataset for Climate Science Information Extraction. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 76–98, Vienna, Austria. Association for Computational Linguistics.

Huitong Pan, Qi Zhang, Mustapha Adamu, Eduard Dragut, and Longin Jan Latecki. 2025b. Taxonomy-Driven Knowledge Graph Construction for Domain-Specific Scientific Applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4295–4320, Vienna, Austria. Association for Computational Linguistics.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge Graphs: Opportunities and Challenges. *arXiv preprint*. ArXiv:2303.13948 [cs].

Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. olmOCR 2: Unit Test Rewards for Document OCR. *arXiv preprint*. ArXiv:2510.19817 [cs].

Zeno van Cauter and Nikolay Yakovets. 2024. Ontology-guided Knowledge Graph Construction from Maintenance Short Texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 75–84, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. *arXiv preprint*. ArXiv:2302.10205 [cs].

Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. *arXiv preprint*. ArXiv:2404.03868 [cs].

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A Comprehensive Survey on Automatic Knowledge Graph Construction. *ACM Comput. Surv.*, 56(4):94:1–94:62.

Appendix

A Detailed Extraction Methodology

This appendix provides the specific prompt structures and decision logic used in the agents described in Figure 1 and Figure 7. The detailed extraction pipeline is provided in Figure 5.

A.1 Paper Screening Protocols

To ensure the KG is populated only with empirical climate research rather than review articles or opinion pieces, the paper screening agent (Step 2 in Figure 5) utilizes a strict boolean filtering criteria. The LLM is prompted to answer three specific questions regarding the input document. A paper is retained only if it yields a positive classification for at least one of the following criteria:

- **Model Presence:** "Does this paper contain climate models (e.g., CMIP6, ESGF-listed models)?"
- **Dataset Availability:** "Does this paper contain or reference specific observational or simulation datasets?"
- **Empirical Nature:** "Does this paper present quantitative experimental results?"

The actual prompt is provided in the Section A.23.2.

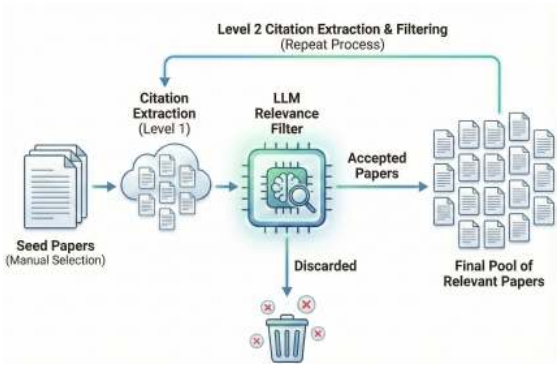


Figure 4: Pipeline for selection of relevant scientific publications. The pipeline starts with a manual seed of highly relevant scientific publications, followed by iterative citation mining and LLM-based filtering to build a high-quality corpus.

A.2 Document level features

Given a scientific paper as a PDF document, we first obtain document-level features, such as *title*, *authors*, *publication venue*, and *abstract*, from metadata provided by OpenAlex. Similar to (Lan et al., 2025), we first identify general features like *problem*, *method*, *task*, which are common to every scientific publication and domain-specific features guided by the target ontology, including entities and concepts relevant to the domain (e.g., models, datasets, variables, performance).

A.3 Graph Initialization Details

We initialize the KG from structured ESGF dataset metadata (see Section 3.4 for the exact snapshot and endpoint), where each record describes a climate simulation using attributes such as experiment identifiers, source model IDs, variables, institutions, and spatial coverage. Dataset-level attributes are deterministically converted into typed ontology entities, while semantic relationships such as `usedInExperiment`, `producesVariable`, and `coversRegion` are instantiated as graph edges. Geospatial bounding boxes are normalized and linked to GeoNames entities, and CMIP6 controlled vocabulary terms are used to enrich model genealogy and component information. The concepts like Ocean circulation, Teleconnection, Instrument and Platform are initialized from GCMD. This initialization grounds later LLM extractions in standardized identifiers; for example, an extracted mention of *precipitation* can be linked to the canonical ESGF variable *pr* rather than requiring the LLM to generate the identifier directly.

A.4 Entity Linking Categorization Strategy

As illustrated in the Entity Linking Process (Step 6), we employ a hybrid linking strategy that routes entities to either deterministic string matching or semantic embedding similarity based on the entity type. Table 9 details the categorization logic used to select the linking method.

We use string matching for rigid identifiers such as climate model names because these entities follow standardized naming conventions where small surface differences can indicate different entities. For example, `ACCESS-ESM1` and `ACCESS-ESM1.5` may have high embedding similarity but refer to distinct models, making exact or rule-based matching safer than embedding distance.

A.5 Entity Linking Threshold Sensitivity Analysis

Table 10 reports the sensitivity analysis used to select the embedding thresholds for entity linking. Scores are top-1 embedding cosine similarities, and

Table 9: Categorization strategy for routing extracted entities to the appropriate linking mechanism.

String Matching Path *Strict rigid identifiers*	**Embedding Similarity Path** *Semantic or contextual concepts*
Exact name matches; standardized taxonomy terms; known domain abbreviations; official model names (e.g., CESM2); institution names.	Semantic variations of variables; context-dependent entities; geographic locations; weather phenomena descriptions; abstract research concepts.

LLM decisions are used as the match/reject signal for the sampled subset. The high match rate in $[0.80, 0.85)$ supports auto-linking at $\tau = 0.80$, while nontrivial match rates in $[0.55, 0.80)$ justify using $\tau_f = 0.55$ as the lower bound for LLM fallback instead of discarding those candidates.

Table 10: Entity-linking sensitivity analysis by top-1 embedding similarity score bin.

Score Bin	Matched	Rejected	Total	Rate
$[0.50,\ 0.55)$	186	185	371	50.1%
$[0.55,\ 0.60)$	281	205	486	57.8%
$[0.60,\ 0.65)$	406	283	689	58.9%
$[0.65,\ 0.70)$	287	232	519	55.3%
$[0.70,\ 0.75)$	261	170	431	60.6%
$[0.75,\ 0.80)$	370	144	514	72.0%
$[0.80,\ 0.85)$	593	43	636	93.2%

A.6 Qualifiers and context

Many climate-science statements hold only under a qualifier: an experiment, an ensemble member, a forcing, a region, or a period. We handle these structurally rather than by attaching free-text conditions to a relation. The ontology promotes the qualifier-bearing concepts to first-class entities, so `Experiment`, `SubExperiment`, `Ensemble`, `Member`, `Forcing`, `Metric`, and `Result` are nodes in their own right, and a conditioned statement is expressed as a path through them instead of as a single qualified edge; the KG contains 35,523 `RESULT_HAS_METRIC` edges of this form. Evidence spans are stored at paragraph granularity with anaphora resolved, so a qualifier stated next to the claim stays visible to any reader auditing the triple. A statement whose qualifier the ontology cannot express, on the other hand, has no licensed type signature and is removed by the conformance agent rather than entering the KG unconditioned. This is a deliberate trade: it keeps the graph interpretable, and it is one reason the conformance stage removes textually correct triples.

A.7 Repeated and cited facts

The same fact is often asserted by several papers, and papers also restate findings that earlier, cited work introduced. Because entity linking resolves mentions to canonical KG nodes, such assertions converge on a single edge rather than on one edge per paper. Each edge accumulates the identifiers of every paper that asserted it in an `added_from_paper_id` list, and keeps each paper's evidence span, so the graph records how many independent documents support a relation and which passage each of them offers. In the 2,000-publication KG, 11.0% of the distinct literature-derived edges carry two or more paper identifiers. What the current design does not do is distinguish the paper that introduced a finding from the papers that restate it: a citation-aware provenance model is left for future work.

A.8 What triples the pipeline filters remove

Relative to raw extractor output, the pipeline discards 14,503 triplets that the judge considers correct. Their distribution across stages shows that this is mostly scope filtering rather than a precision–recall trade: 10,326 (71.2%) are removed by ontology conformance and are correct statements whose type signature lies outside the schema, so no ontology-grounded KG could load them; 835 (5.8%) fail evidence grounding, meaning their evidence span cannot be located in the parsed source; and 3,342 (23.0%) are removed by the validator agent, which as shown above rejects unconfirmed rather than only incorrect assertions. Restricting attention to triplets the schema can express, the pipeline retains 23,085 of 27,262 judge-correct triplets, a within-schema retention of 84.7%, while raising C_{triplet} from 94.05% to 98.96%. Measured on the human-annotated subset, within-schema retention is 81.9%.

A.9 Downstream Utility

The constructed KG supports multi-hop scientific queries that combine literature-derived assertions with structured climate metadata. Example questions include:

- Which variable is both produced by the REMO2009 regional climate model and measured by Vaisala sondes RS80/RS92/VS41?

- Which source component does CanESM5 have that applies to the ocean realm?
- What project did the institute that produced the CNRM-ESM2-1 source model also participate in?
- What observational dataset was used by the method that relies on "Last Glacial Maximum to control temperature change" as a metric?
- What is a result of the method that experiments on Antarctic ice core water stable isotopes?

These queries illustrate how the KG can connect papers, models, variables, instruments, methods, metrics, results, institutions, and controlled climate vocabularies through auditable paths rather than isolated keyword matches.

A.10 Other Implementation Details

We use Gemini-2.5-pro for triplet/evidence extraction and LLM-based entity-linking fallback, Gemini-2.5-flash for large-scale validation and evaluation, and Llama 3.3-70B (Grattafiori et al., 2024) for candidate paper selection.

A.11 Per-Stage Cost and Runtime

Table 11 breaks the cost of the full 2,000-paper run down by stage and places it next to the ablation effect of that stage, so that the price and the benefit of each component can be read together. Costs are estimates at the API prices in force during our runs, and they vary with PDF length and output volume.

Extraction dominates: Gemini-2.5-pro is called roughly five times per paper, which we cost conservatively at about USD 200 per 1,000 papers, and processing 1,000 papers takes about 2–3 hours. This cost is not specific to our design, since any LLM-based KG construction system must read the documents.

The verification stack that follows is close to free. Ontology conformance and evidence validation use no LLM at all: the first is a set-membership test over the licensed (source type, relation, target type) tuples with rule-based normalization, and the second is a fuzzy string match of the evidence span against the parsed document, so both run in CPU minutes at no API cost. The validator agent is the only paid verifier, at one short Gemini-2.5-flash call per surviving triplet: 27,900 calls, approximately 9M input and 0.2M output tokens, or roughly USD 3. Paper screening runs on a locally hosted Llama-3.3-70B and costs GPU time rather than API credit, as does PDF parsing with olmOCR. Entity linking runs local NV-Embed-v2 embeddings and calls an LLM only for candidates that fall in the fallback band $[\tau_f, \tau)$.

Cost across the ablation variants of Table 4 differs only in how many triplets reach the validator agent. Running it without the two deterministic filters raises its call count from 27,900 to 39,964, or roughly USD 4.7, while the extraction cost is common to every variant. Two of the three stages are deterministic and use no LLM, and the third issues one short Gemini-2.5-flash call per surviving triplet, so verification adds roughly 1% to the cost of extraction, which any LLM-based KG construction system has to pay. That 1% buys 4.91 points of C_{triplet} together with the schema and traceability guarantees.

A.12 Entity Classification and Discard Logic

Entities that fail to link to the existing KG via either string matching or embedding similarity (below threshold τ_f) undergo a secondary classification step. This ensures that the graph is enriched with novel but relevant concepts while filtering out noise. The prompt to the LLM is provided in A.23.2.

As shown in the pipeline workflow, discarded entities are logged to analyze and update exclusion patterns, creating a feedback loop that refines future classification precision.

A.13 Climate ontology

The entity and relation definitions of the climate ontology are given below.

A.14 Entity Type Definitions

- **Activity**: A coordinated modeling effort or scientific campaign.
- **ExperimentFamily**: A set of climate model experiments conducted under a common scientific objective (e.g., CMIP experiment families).
- **Experiment**: A specific climate model simulation scenario (e.g., historical, ssp585).
- **SubExperiment**: A variant of a climate model experiment with specific configurations (e.g., initialization, ensemble member, or resolution).

Table 11: Per-stage cost, runtime, and ablation effect for the full 2,000-publication run. Ablation effect reports the result of leaving that stage out of the full pipeline, which itself yields 23,327 triplets at 98.96% C_{triplet}.

Stage	Mechanism	Cost (2,000 papers)	Ablation effect
Paper screening	Llama-3.3-70B, local	no API cost, GPU time	corpus selection, not evaluated
PDF parsing	olmOCR, local	no API cost, GPU time	prerequisite for evidence grounding
Triplet + evidence extraction	Gemini-2.5-pro, ≈5 calls/paper	≈USD 400, 2–3 h per 1,000 papers	produces 39,964 triplets at 94.05%
Ontology conformance (O)	deterministic set membership	no API cost, CPU seconds	w/o: 31,095 at 98.88%, 7,768 untypeable kept
Evidence validation (E)	deterministic fuzzy match	no API cost, CPU minutes	w/o: 24,172 at 98.96%, 845 ungrounded kept
Validator agent (V)	Gemini-2.5-flash, 1 call/triplet	≈USD 3 (27,900 calls)	w/o: 27,900 at 94.72%, -4.24 points
Entity linking	NV-Embed-v2 local + LLM fallback	GPU time, few fallback calls	links accepted triplets to the KG

- **Source**: A climate model or system used to generate data (e.g., GFDL-ESM4).
- **SourceType**: The classification of the model.
- **SourceComponent**: A sub-model (e.g., atmosphere, ocean, land) within a coupled climate model; may also function as a standalone model (e.g., ACCESS-OM2).
- **PhysicalScheme**: A representation of a physical process within a climate model (e.g., cloud scheme).
- **PhysicalFeature**: A structural characteristic of a climate model's design (e.g., terrain-following grid).
- **SimulationType**: Type of simulation performed (e.g., transient, equilibrium).
- **Metric**: Quantitative measure of model performance (e.g., climate sensitivity).
- **Project**: A broader initiative under which models or experiments are conducted (e.g., CMIP6).
- **Institute**: Organization responsible for developing models or running simulations.
- **Variable**: Scientific quantities output by models (e.g., temperature, precipitation).
- **Realm**: Component of the Earth system (e.g., atmosphere, ocean).
- **Frequency**: Temporal resolution of model output (e.g., daily, monthly).
- **Resolution**: Spatial resolution or grid size of the data.
- **Ensemble**: A collection of model runs differing in initial conditions or configurations.
- **Member**: An individual member of an ensemble.
- **MIPEra**: A major generation or version of coordinated experiments (e.g., CMIP5, CMIP6).
- **RCM**: A regional climate model used for fine-resolution regional simulations.
- **Domain**: Geographical coverage of a model.
- **Continent**: A large continuous landmass (e.g., Asia, Africa).
- **Country**: A sovereign state or territory (e.g., India, USA).
- **Country_Subdivision**: Administrative units within countries (e.g., California).
- **City**: Urban locality (e.g., Paris).
- **No_Country_Region**: Areas not under country jurisdiction (e.g., open ocean).
- **Instrument**: Device used to observe environmental variables (e.g., radiometer).
- **Platform**: Physical carrier for an instrument (e.g., satellite).
- **Weather_Event**: Specific events such as storms or droughts.
- **Teleconnection**: Large-scale climate patterns (e.g., ENSO, NAO).

- **Ocean_Circulation**: Movements of ocean waters (e.g., AMOC).
- **Natural_Hazard**: Geophysical events impacting systems (e.g., tsunami, earthquake).
- **NaturalHazardType**: The specific type of natural hazard (e.g., Flood, Tornado, Wildfire).
- **ClimateZone**: Classification of climate based on temperature and precipitation patterns (Tropical, Dry, Temperate, Continental, Polar).

A.15 Relation Definitions

- **INCLUDES_EXPERIMENT**: Links an ExperimentFamily to its constituent Experiments.
- **HAS_SUB_EXPERIMENT**: Connects an Experiment to a refining SubExperiment.
- **IS_OF_TYPE**: Connects a Source or Natural_Hazard to its corresponding type.
- **APPLIES_TO_REALM**: Associates a Source, SourceComponent, or Experiment with a Realm.
- **FOCUSES_ON_REALM**: Indicates the Realm emphasized by an Activity.
- **PARTICIPATED_IN**: Connects an Institute to a Project.
- **GENERATED_BY_ACTIVITY**: Indicates that a Source was produced as part of an Activity.
- **PRODUCED_BY_INSTITUTE**: Links a Source to the responsible Institute.
- **BELONGS_TO_MIP_ERA**: Indicates association with a specific MIP Era.
- **USED_IN_EXPERIMENT**: Links a Source or RCM to an Experiment.
- **PRODUCES_VARIABLE**: Specifies Variables output by a Source or SourceComponent.
- **HAS_SUBSEQUENT_VERSION**: Links a Source or SourceComponent to a newer version.
- **INHERITED_FROM**: Indicates derivation or inheritance between Sources.
- **HAS_SOURCE_COMPONENT**: Connects a Source to its SourceComponents.
- **USES_PHYSICAL_SCHEME**: Indicates a PhysicalScheme employed by a Source.
- **HAS_PHYSICAL_FEATURE**: Links a Source to a defining PhysicalFeature.
- **HAS_SIMULATION_TYPE**: Connects a Source to a SimulationType.
- **HAS_METRIC**: Associates a Source with an evaluation Metric.
- **PART_OF_PROJECT**: Indicates membership in a Project.
- **ASSOCIATED_WITH_MEMBER**: Connects a Source to an Ensemble Member.
- **SAMPLED_AT_FREQUENCY**: Indicates output Frequency of a Source.
- **HAS_SPATIAL_RESOLUTION**: Specifies the spatial Resolution of an entity.
- **PART_OF_ENSEMBLE**: Links a Source or SubExperiment to an Ensemble.
- **USES_FORCING**: Indicates forcings used in an Experiment.
- **INCLUDES_ENSEMBLE_MEMBER**: Connects an Experiment to its Ensemble members.
- **DRIVEN_BY_SOURCE**: Indicates that an RCM is driven by a Source.
- **COVERS_DOMAIN**: Specifies the Domain covered by an RCM or Project.
- **MEASURES_VARIABLE**: Links an Instrument to the Variable it measures.
- **MOUNTED_ON**: Indicates the Platform carrying an Instrument.
- **PERFORMED_BY_INSTITUTE**: Indicates which Institute performed an Experiment.
- **COVERS_REGION**: Specifies the region spanned by an RCM.

- **IMPACTED_REGION**: Indicates the region affected by a Weather_Event.
- **INFLUENCES_REGION**: Links a Teleconnection to influenced regions.
- **OBSERVED_IN_REGION**: Indicates regions where an Ocean_Circulation is observed.
- **AFFECTED_REGION**: Indicates regions affected by a Natural_Hazard.
- **IN_CONTINENT**: Connects a Country to its Continent.
- **IN_COUNTRY**: Connects a Country_Subdivision or City to its Country.
- **HAS_CLIMATE_ZONE**: Links a No_Country_Region to a ClimateZone.

A.16 Scientific publications ontology

The scientific publications ontology is inspired from (Lan et al., 2025) and modified for climate domain.

Entity Definitions

- **Paper**: A scholarly publication that reports research contributions, methods, experiments, and results.
- **Field**: A broad research area or discipline to which a paper contributes (e.g., information retrieval, computer vision).
- **Problem**: A specific research challenge or question that the paper aims to address.
- **Method**: A technical approach, algorithm, or procedure proposed or applied to solve a problem.
- **Innovation**: A novel idea, contribution, or improvement introduced by the paper relative to prior work.
- **Task**: A well-defined computational or scientific objective used to evaluate methods (e.g., classification, forecasting).
- **Venue**: The conference, journal, or workshop where the paper is published.
- **Author**: An individual who contributed intellectually to the creation of the paper.
- **Keyword**: A salient term or phrase summarizing key topics or concepts of the paper.
- **ObservationalDataset**: A dataset used for experimentation or evaluation, typically derived from real-world or empirical observations.
- **Metric**: A quantitative measure used to evaluate the performance or quality of a method or model.
- **Result**: An experimental outcome or reported finding, often associated with a metric value.

Relation Types The scientific-publication ontology contains paper-level relations for venue, authorship, keywords, model use, metrics, results, and paper mentions; method- and task-level relations for problems, innovations, tasks, metrics, observational datasets, and results; model- and result-level relations for observational datasets, metrics, and reported results; and LLM-extracted ontology-aligned relations stored with provenance (`added_from_paper_id`).

A.17 Final KG statistic

After KG enrichment the final statistic is provided in Table 13 and Table 14.

A.18 Justification for our selection of climate domain ontology

In order to benefit from established scientific terminology and ensure interoperability with existing climate data infrastructures, our domain ontology is constructed by unifying multiple authoritative sources of climate terminology. Specifically, we integrate standardized climate model names, institutions, experiments, and variables from ESGF metadata[1] and the GCMD taxonomy defined in (Pan et al., 2025b), which together provide vetted, community-accepted identifiers and a widely adopted semantic taxonomy for Earth system science. To capture internal model structure and configuration details that are not explicitly represented in ESGF metadata, we further incorporate climate model component information from the CMIP6 Controlled Vocabulary repository[2], enabling canonical representation of model components and experiments. Finally, to provide unambiguous spatial grounding, geographic entities such as regions and countries are linked to GeoNames identifiers,

[1] https://esgf.github.io/
[2] https://github.com/WCRP-CMIP/CMIP6_CVs

supporting spatially contextualized relations (e.g., model evaluation over specific regions) and facilitating integration with external KGs that rely on standardized geographic identifiers.

A.19 Ontology Conformance Agent Specification

The ontology conformance agent is fully deterministic: it uses no LLM, no external reasoner, and no learned component. The curated ontology is compiled into a set $\mathcal{L}$ of 35 licensed tuples (source type, target type, relation), and a triple conforms if and only if it passes both of the following tests.

1. Licensing. The triple's type signature must be a member of $\mathcal{L}$:

$$(c_h, c_t, r) \in \mathcal{L}.$$

The test is a constant-time set lookup after case and whitespace normalization of the type names. For example, `(Source, SourceComponent, HAS_SOURCE_COMPONENT)` is licensed, whereas the reversed extraction `(SourceComponent, Source, HAS_SOURCE_COMPONENT)` is not, which is the case discussed in Section 3.3. Licensing is what makes relation direction and argument order enforceable without asking a model to check them.

2. Semantic boundary guards. Licensing constrains types, not the values that fill them, so a second set of rules rejects values that carry the right type name but the wrong kind of content. The guards are built from CMIP6 controlled-vocabulary lists and regular expressions, and they fire on the three types where the extractor confuses categories most often:

- `SourceComponent` must not be a realm label. A value in the CMIP6 realm vocabulary (*atmos*, *ocean*, *seaice*, *landice*, and their spelled-out variants) is rejected, because the licensed way to state that fact is `(SourceComponent, Realm, APPLIES_TO_REALM)`.

- `PhysicalScheme` must show a scheme cue. Either the value or its evidence paragraph has to contain a cue word (*scheme*, *parameterization*, *closure*, *formulation*, *algorithm*, *physics package*) or a known scheme name (*Tiedtke*, *Zhang-McFarlane*, *Arakawa-Schubert*, *Betts-Miller*, *KPP*, and others).

- `PhysicalFeature` must not be a value of another category. Ensemble member identifiers matching `r{d}i{d}p{d}f{d}`, resolutions such as "100 km" or "1.5 degrees", experiment or scenario labels such as *historical* and `ssp585`, activity and project labels such as *CMIP6* and *CORDEX*, and common variable names such as *tas* and *pr* are all rejected.

Because both tests are exact and side-effect free, the stage is reproducible, auditable, and free of API cost, and its verdict is stored per triple so that the decision can be revisited without re-running any model. Widening the ontology therefore amounts to adding tuples to $\mathcal{L}$, which is the human-in-the-loop path we suggest for the correct-but-untypeable triples reported in Section 4.4.

A.20 Per-Stage Examples

The following examples are taken verbatim from the 2,000-publication run and illustrate what each stage keeps and what it removes. Evidence spans are abbreviated.

Kept by every stage. (*severe floods*, `Natural_Hazard`) `AFFECTED_REGION` (*central and western Africa*, `No_Country_Region`), with evidence "In 2012, central and western Africa was hit by severe floods which affected 1,538,242 people and caused 340 deaths as of September of that year." The type signature is licensed by the ontology, the span occurs in the parsed document, and the validator agent confirms that the span states the relation.

Removed by ontology conformance, yet textually correct. (*MRI-CGCM3*, `Source`) `PRODUCES_VARIABLE` (*u-component of wind*, `Variable`). The paper does state this, and the judge marks it correct, but the ontology licenses `PRODUCES_VARIABLE` only from `SourceComponent` to `Variable`, because in the ESGF data model it is a model component that produces a variable. The triplet has no place in the schema and is therefore not loadable, which is the case for 93.4% of what this stage removes.

Removed by evidence validation. (*Integrated Forecasting System (IFS)*, `Source`) `HAS_PHYSICAL_FEATURE` (*137 model levels*, `PhysicalFeature`). The claim is plausible and the wording resembles the paper, but the span the extractor returned cannot be matched in the parsed document above the similarity threshold,

so the triplet loses its audit trail and is dropped. Such spans are frequently reconstructed from tables and figure captions, which the parser renders differently.

Removed by the validator agent, yet accepted by human annotators. (*FIO-ESM v2.0*, `Source`) `HAS_SOURCE_COMPONENT` (*CLM4.0*, `SourceComponent`), with evidence "The land surface model is upgraded from the Community Land Model version 3.5 (CLM3.5; Oleson et al., 2008) to CLM4.0 (Lawrence et al., 2011) with the same horizontal resolution as CAM5." The span states that the land model is upgraded to CLM4.0 without naming the parent model, so the validator cannot confirm the attachment from the span alone, while annotators who read the surrounding paper accept it. This is the conservative behaviour quantified in Section 4.4: 79% of the validator's rejections in the annotated sample are judged valid by the human majority.

A.21 Evidence Verification Against Parsed PDF and Content and Hallucination Analysis

We assess the factual reliability of LLM-extracted evidence by verifying it against the original PDF content parsed using the Olmocr v2 (Poznanski et al., 2025) parser. For each extracted triple, the associated evidence span generated by the LLM is aligned with the parsed PDF text. An evidence instance is considered *correct* if the cited text can be unambiguously located in the parsed content and faithfully supports the extracted relation; otherwise, it is marked as a hallucination.

For each extracted triple, the extractor also outputs a supporting evidence span. We compute the textual similarity between the extracted evidence span and the source document using the RapidFuzz library. RapidFuzz provides normalized string-matching scores in the range $[0, 100]$, where higher values indicate stronger lexical overlap. In our experiments, a triple is considered evidence-grounded if its similarity score exceeds a predefined threshold (70%, 75%, or 80%). Triples failing this criterion are flagged as hallucinations.

Table 12 reports results when all evidence spans are considered (including those derived from tables and figures), hallucination rates range from 7.74% at an 80% threshold to 6.14% at a 70% threshold. As expected, relaxing the similarity threshold admits more borderline matches and thus reduces the measured hallucination rate.

Table 12: Hallucination analysis of an LLM-based KG extractor agent with evidence.

Experimental setting	Threshold	Rate (%)
Evidence span	80%	7.741
Evidence span	75%	6.793
Evidence span	70%	6.137
Evidence span without table/figure interpretations	80%	1.443
Evidence span without table/figure interpretations	75%	1.035
Evidence span without table/figure interpretations	70%	0.626

However, a large portion of the mismatch is due to the way Olmocr V2 and Gemini-2.5-pro interpret tables and figures. When we do not consider extracted texts explaining table and figure portions of the document the hallucination rates drop sharply—down to $1.44\%(80\%)$, $1.04\%(75\%)$, and $0.63\%(70\%)$. This result highlights that the hallucination rate of Gemini-2.5-pro is relatively low to extract accurate text-spans from documents.

A.22 AI use disclosure

We used AI assistants, including GPT-5.4, only for limited manuscript-support tasks: improving wording/clarity and helping design or refine the pipeline figure. The scientific ideas, methodology, experiments, analysis, and final claims were developed and verified by the authors. No AI system was used to generate experimental results or make autonomous research decisions.

A.23 Annotation guidelines and instructions

A.23.1 Annotator Qualifications and Selection

All 3 human experts hold PhD in closely related field. They were recruited through a third-party online professional services platform. The experts compensation was according to current standards. They were recruited for the specific task of annotation and were preinformed about their task and the compensation they would receive. The participants were informed how the annotation was going to be used. The annotations began only after both parties agreed on the terms.

A.23.2 Annotation Process Overview

Annotators are tasked with validating extracted knowledge graph triples with respect to their supporting textual evidence. Given the relatively lightweight nature of this task, annotators are not

required to perform open-ended information extraction or domain discovery. Instead, they are provided only with the ontology definitions, including entity types and relation descriptions, to guide consistent judgments.

For each candidate triple, annotators are shown: (i) the structured triple, (ii) the corresponding evidence paragraph extracted from the source document, and (iii) the relevant ontological definitions. Annotators are instructed to verify whether the evidence explicitly supports the triple and whether the triple conforms to the ontology's domain and range constraints. No additional domain expertise beyond the provided definitions is assumed, and annotators are not asked to infer unstated facts or resolve ambiguities beyond the given text.

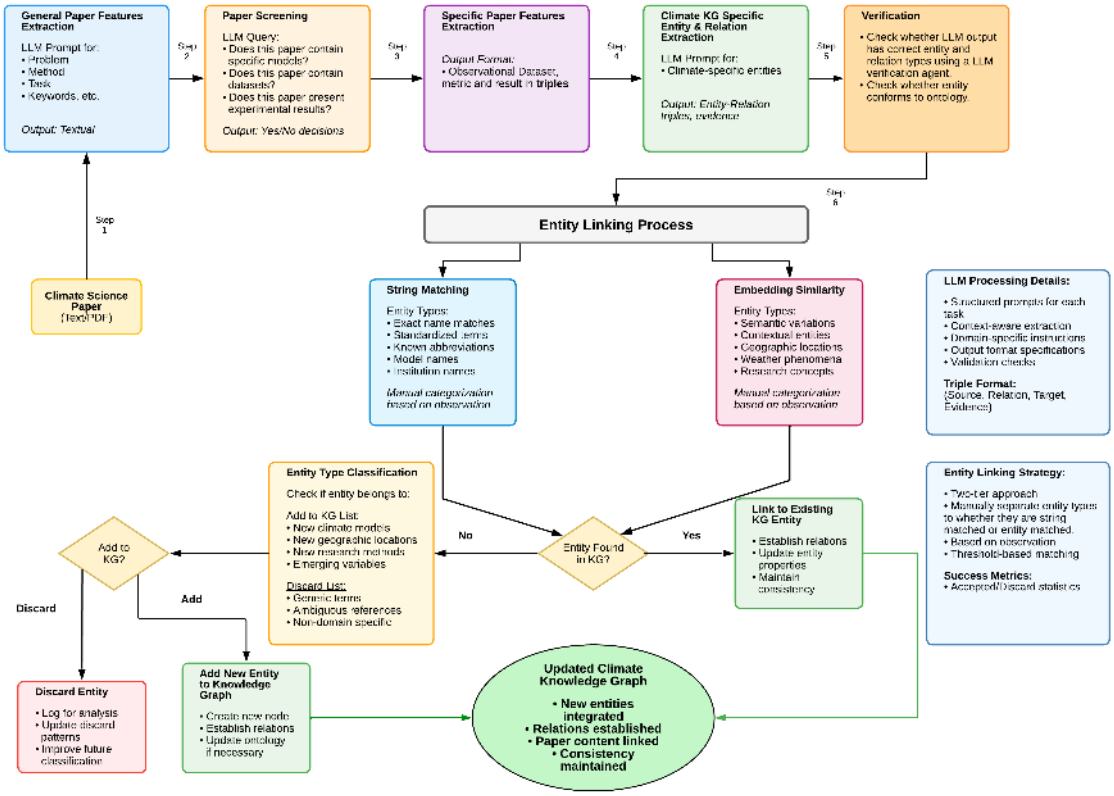


Figure 5: Overview of the ontology-guided, evidence-based KG construction pipeline. The workflow proceeds in three phases: (1) **Paper Screening and Feature Extraction**, where documents are filtered for empirical relevance (Step 1–3); (2) **Extraction and Verification**, where an LLM extracts entity-relation triples supported by explicit textual evidence and verifies ontology conformance (Step 4–5); and (3) **Hybrid Entity Linking**, which routes entities to either deterministic string matching or semantic embedding similarity (Step 6) to determine whether to discard the entity or integrate it into the Updated Climate KG.

ESGF (Earth System Grid Federation) Dataset metadata

Dataset title	Experiment id	Source id	Variable	North degrees	South degrees	East degrees	West degrees	...
D1	esm-hist	CNRM-ESM2-1	clic	37.27	-32.90	40.58	-33.99	...
D2	omip1	CNRM-ESM2-1	ua400	42.24	-45.76	60.28	-24.64	...

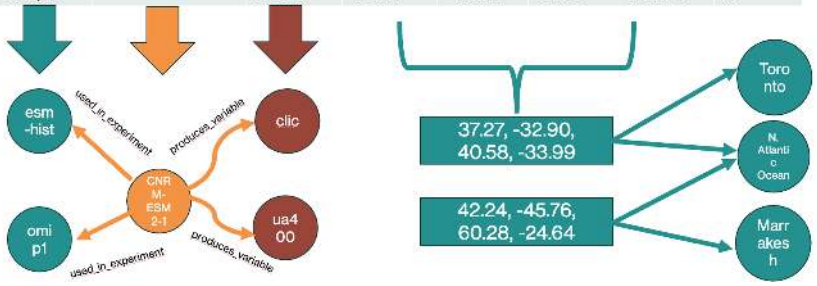


Figure 6: A small subsection of structured ESGF metadata converted to KG based on ontology specifications mapped with geolocation data from Geonames.

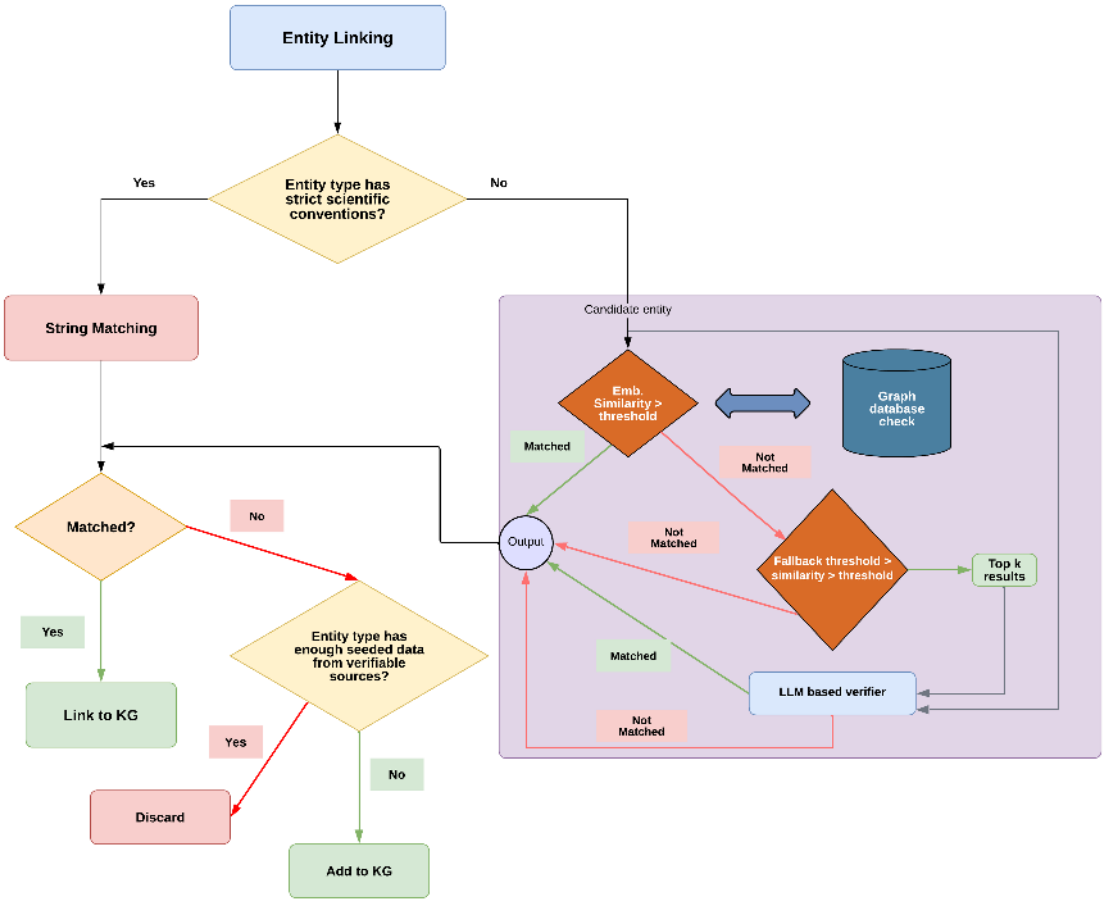


Figure 7: Decision flow for entity linking. Entities are linked via exact string matching when strict naming conventions apply; otherwise, embedding-based similarity with adaptive thresholds and LLM verification is used to link, add, or discard entities from the KG.

Table 13: Distribution of entity types in the constructed KG.

Entity Type	Count	Entity Type	Count
City	30,062	Ensemble	578
Result	16,207	Experiment	481
Author	10,437	Source	394
EvaluationMetric	8,430	Venue	373
No_Country_Region	6,965	Country	252
PhysicalScheme	6,568	Forcing	204
Metric	6,144	Institute	132
Variable	6,056	SubExperiment	63
Member	5,929	NaturalHazardType	48
Country_Subdivision	3,893	RCM	42
PhysicalFeature	3,236	Domain	39
Instrument	3,045	Project	37
ObservationalDataset	2,496	Teleconnection	27
SourceComponent	2,440	Activity	26
Keyword	2,344	Ocean_Circulation	20
Platform	2,326	GridLabel	17
Innovation	2,000	Frequency	16
Method	2,000	SourceType	16
Paper	2,000	Resolution	15
Problem	2,000	Realm	14
Task	2,000	ExperimentFamily	9
Weather_Event	1,645	Continent	7
Field	1,280	ClimateZone	5
SimulationType	840	MIPEra	4
Natural_Hazard	828	**Total**	**133,990**

Table 14: Distribution of scientific paper relation types in the constructed KG.

Relation Type	Count
PAPER_MENTIONS	96,100
RESULT_HAS_METRIC	35,522
PAPER_HAS_RESULT	29,879
METHOD_HAS_RESULT	29,879
PAPER_HAS_KEYWORD	22,644
PAPER_AUTHORED_BY	14,853
PAPER_USES_METRIC	8,720
METHOD_USES_METRIC	8,720
TASK_USES_METRIC	8,720
MODEL_USES_METRIC	4,797
MODEL_HAS_RESULT	4,362
PAPER_USES_MODEL	2,800
PAPER_EXPERIMENTS_ON_OBSERVATIONAL_DATASET	2,749
METHOD_EXPERIMENTS_ON_OBSERVATIONAL_DATASET	2,749
TASK_EXPERIMENTS_ON_OBSERVATIONAL_DATASET	2,749
PAPER_BELONGS_TO_FIELD	2,000
PAPER_SOLVES_PROBLEM	2,000
PAPER_APPLIES_METHOD	2,000
PAPER_HAS_INNOVATION	2,000
PAPER_WORKS_ON_TASK	2,000
PAPER_PUBLISHED_IN_VENUE	2,000
METHOD_SOLVES_PROBLEM	2,000
TASK_FACES_PROBLEM	2,000
METHOD_WORKS_ON_TASK	2,000
METHOD_HAS_INNOVATION	2,000
MODEL_EXPERIMENTS_ON_OBSERVATIONAL_DATASET	876

Table 15: Distribution of climate relation types in the constructed KG.

Relation Type	**Count**
COVERS_REGION	354,837
IN_COUNTRY	33,923
PRODUCES_VARIABLE	31,024
ASSOCIATED_WITH_MEMBER	11,903
HAS_SPATIAL_RESOLUTION	6,708
APPLIES_TO_REALM	4,031
INCLUDES_ENSEMBLE_MEMBER	3,069
PERFORMED_BY_INSTITUTE	2,762
USED_IN_EXPERIMENT	2,019
PART_OF_PROJECT	1,048
USES_FORCING	986
SAMPLED_AT_FREQUENCY	828
HAS_SOURCE_COMPONENT	734
GENERATED_BY_ACTIVITY	623
MEASURES_VARIABLE	607
IS_OF_TYPE	542
PRODUCED_BY_INSTITUTE	503
HAS_METRIC	465
MOUNTED_ON	422
INCLUDES_EXPERIMENT	328
BELONGS_TO_MIP_ERA	318
HAS_PHYSICAL_FEATURE	298
PARTICIPATED_IN	213
IN_CONTINENT	210
FOCUSES_ON_REALM	170
DRIVEN_BY_SOURCE	139
COVERS_DOMAIN	125
HAS_SUBSEQUENT_VERSION	79
HAS_SIMULATION_TYPE	74
USES_PHYSICAL_SCHEME	73
HAS_SUB_EXPERIMENT	72
PART_OF_ENSEMBLE	53
INHERITED_FROM	3

Prompts

Prompt titles below use the same agent names as Figure 1 and Section 3.3. The document-level prompts (field, keywords, problem, method, task, innovation, model, observation and metric) belong to the native document understanding step, and the ontology conformance and evidence validator agents use no prompt, since both are deterministic.

Title: *Triplet and evidence extractor agent prompt*
Input: *PDF in native format(not parsed), ontology, entity_definitions, relation_definitions*
Output: *Structured format: List of source entity class, source entity instance, relation, target entity class, target entity instance*
Role Description: You are an expert in extracting information from scientific papers. Your job is to answer the user's query based on the pdf file you have access to.
Task: You are given a scientific paper (or paragraph from it) related to climate modeling or Earth system science. The KG is primarily composed of ESGF metadata. Your task is to extract (source, relation, target) triplets with type that match the ontology below. For each extracted triplet, provide all supporting paragraphs from the text and explicitly resolve any anaphoras (e.g., "it", "this model") into their full entity names. The ontology is given below in the form of (source type, target type, relation):

Ontology: `{ontology}`

Sample ontology:
`('SourceComponent','Variable','PRODUCES_VARIABLE')`
`('SourceComponent','Realm','APPLIES_TO_REALM')`
`('Source','PhysicalScheme','USES_PHYSICAL_SCHEME')`

Entity type definitions: `{entity_definitions}`

Sample definitions:
`Source``: A climate model or system used to generate data (e.g., GFDL-ESM4).`
`SourceComponent``: A sub-model (e.g., atmosphere, ocean, land) within a coupled climate model. It may also function as a standalone climate model when used independently (e.g., ACCESS-OM2 as the ocean component in ACCESS-CM2).`
`Realm``: Component of the Earth system (e.g., atmosphere, ocean).`
`Variable``: Scientific quantities output by models (e.g., temperature, precipitation).`
`PhysicalScheme``: A representation of a physical process within a climate model (e.g., cloud scheme).`

Relation definitions: `{relation_defintions}`

Sample definitions:
`PRODUCES_VARIABLE``: Specifies that a Source or SourceComponent outputs a particular Variable.`
`APPLIES_TO_REALM``: Indicates that a Source, SourceComponent, or Experiment is associated with a specific Realm (e.g., atmosphere, ocean, land surface).`
`USES_PHYSICAL_SCHEME``: Indicates that a Source employs a particular PhysicalScheme (e.g., radiation scheme, convection scheme).`

ONLY PROVIDE THE TRIPLETS IF THEY ARE PRESENT IN THE PAPER.

Title: *Candidate paper selection prompt (corpus collection)*
Input: *Title and Abstract*
Output: `yes` *or* `no`
Task: Based on the title and abstract of the paper, analyze whether the paper likely contains climate model simulation data or results. Answer with a simple "yes" or "no".

Title: *Hybrid entity linking: LLM fallback prompt*
Input: *key_word_phrase, entity_type, instances (candidate instance list), paragraph_context*
Output: *Exactly one instance string from the list, or* `no match` *(no explanation)*
Role Description: You are an expert at entity linking under schema constraints. You must pick the most likely instance match from a provided candidate list.
Task: Given the following key word/phrase of an entity type extracted from paragraph context, output which of the instances it most likely refers to. If none of the instances match, output `no match`. Do not provide an explanation.

Key Word/Phrase: `{key_word_phrase}`
Entity type: `{entity_type}`
Instances: `{instances}`
Paragraph context: `{paragraph}`

OUTPUT RULES: Output **only** one of the given instances **exactly as written**, or `no match`. No punctuation, no extra text.

Title: *Field extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *One short phrase describing the research field (no extra text)*
Role Description: You are an expert at identifying the primary research field of scientific papers from their content.
Task: Summarize the research field of the paper in **one phrase**. Example: *climate change*

Title: *Keywords extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *3–5 keywords separated by commas (single line, no bullet points)*
Role Description: You are an expert at extracting concise topical keywords that summarize a scientific paper.
Task: Extract **3–5 keywords** summarizing the content, topic, and field of the paper. Each keyword must be separated by a comma. Example: *climate change, global warming, greenhouse gases*

Title: *Problem extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *One sentence (max 50 words)*
Role Description: You are an expert at identifying the main problem statement a scientific paper addresses.
Task: Describe the problem the paper focuses on solving (e.g., shortcomings of existing methods, challenges in the task, factors affecting performance). Summarize in **one sentence (max 50 words)**. Example: *Existing climate models struggle to accurately simulate extreme precipitation events due to coarse spatial resolution and inadequate representation of convective processes.*

Title: *Method extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *One sentence (max 50 words)*
Role Description: You are an expert at summarizing proposed methods from scientific papers.
Task: Summarize the method proposed in the paper to solve the problem. Summarize in **one sentence (max 50 words)**. Example: *We propose a bias-correction method using quantile mapping to improve the accuracy of downscaled temperature projections from regional climate models.*

Title: *Task extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *One sentence (max 50 words)*
Role Description: You are an expert at identifying the main task a scientific paper focuses on.
Task: Summarize the task the paper focuses on solving. Summarize in **one sentence (max 50 words)**. Example: *The task is to project future temperature and precipitation patterns over South Asia using downscaled CMIP6 climate model outputs.*

Title: *Innovation / contribution extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *One sentence (max 50 words)*
Role Description: You are an expert at identifying the key innovation or contribution of a scientific paper.
Task: Describe the innovation or contribution of the paper in **one sentence (max 50 words)**. Example: *This paper introduces a novel bias-correction technique that significantly improves the accuracy of regional climate model projections for extreme precipitation events in South Asia.*

Title: *Paper screening agent prompt (observation + metric)*
Input: *PDF in native format (not parsed)*
Output: `yes` *or* `no` *(single token)*
Role Description: You are an expert at screening climate-science papers for evaluation content.
Task: Does this paper have **(1) an observational dataset** which the model produces or is evaluated upon, **and (2) at least one quantitative evaluation metric**? Answer with `yes` only if it fulfills **both** conditions, otherwise answer with `no`. Example: `yes` or `no`

Title: *Evaluator prompt ($C_{triplet}$ judge)*
Input: *(Source, Relation, Target) triplet and evidence*
Output: `valid` *or* `invalid`
Task: You are given a paragraph related to climate modeling or Earth system science, and a (source instance, relation, target instance) triplet extracted from it. The KG is primarily composed of ESGF metadata. Your task is to verify if the triplet supports the paragraph from the paper. Assume anaphoras are there in the paragraph. Use your own knowledge to insert the source or target to the context of the paragraph. Answer with 'valid' or 'invalid'. Do not provide an explanation.

Extracted Triplet:
Source: `{source}`
Relation: `{relation}`
Target: `{target}`
Paragraph context: `{paragraph}`

Title: *Model extraction prompt*
Input: *PDF in native format (not parsed)*
Output: *Comma-separated list of climate model abbreviated names, or* `no`
Role Description: You are an expert at identifying climate model names mentioned in scientific papers.
Task: Name the climate model(s) this paper talks about. If nothing is mentioned, answer `no`. If present, each model name must be a **single abbreviated form** (no expansions in parentheses). Correct example: `ACCESS-CM2`. Wrong example: `ACCESS-CM2 (Australian Community Climate and Earth System Simulator Climate Model Version 2)`.

Title: *Observation and metric extraction prompt (table-only)*
Input: *PDF in native format (not parsed)*
Output: *Structured list of entries: Model, ObservationalDataset, Metric, ResultValue(s)*
Role Description: You are an expert at extracting evaluation details from scientific paper tables with high precision.
Task: Extract the **observational dataset**, **evaluation metrics**, and **results** for **each climate model only from the table(s) in the paper**.

- Extract **only quantitative** metric values (numbers, percentages, error values, correlations, etc.), not descriptive statements.
- If a required field is missing in the tables, use `no`.

ONLY USE TABLE CONTENT. DO NOT INFER FROM CAPTIONS OR BODY TEXT.

Title: *Validator agent prompt*
Input: *Paragraph from a scientific paper and a candidate (source, source type, relation, target, target type)*
Output: `valid` *or* `invalid`
Role Description: You are an expert in climate modeling and Earth system science, with strong familiarity with ESGF metadata and related controlled vocabularies.
Task: Determine whether the given paragraph reasonably supports the extracted (**source instance**, **relation**, **target instance**) triplet. Each instance is accompanied by its semantic type (e.g., Source, Experiment, Variable, Institution, Region).

- The source or target may not be mentioned verbatim but may be inferred from context.
- Domain knowledge may be used to resolve abbreviations or implicit mentions.
- The paragraph may contain anaphora or implicit references.

The triplet should be considered **valid** if the relationship is explicitly stated or clearly implied by the paragraph. It should be considered **invalid** only if the relationship is contradicted or unsupported by the text.

Respond with *only* one of the following labels:

- `valid`
- `invalid`

Do not provide explanations or additional text.

Extracted Triplet:

- **Source:** {source}
- **Source Type:** {source_type}
- **Relation:** {relation}
- **Target:** {target}
- **Target Type:** {target_type}

Paragraph context:
{paragraph}