Taxonomy-Driven Knowledge Graph Construction for Domain-Specific Scientific Applications

Anonymous ACL submission

Abstract

We present a taxonomy-driven framework for constructing domain-specific knowledge graphs (KGs) that integrates structured taxonomies, Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). A key challenge in LLM-based extraction is weak annotations: noisy or misaligned entity/relationship labels diverge from expert-curated taxonomies. For instance, state-of-the-art generalist GLiNER model achieves only 0.339 F1 on climate science entity recognition, often omitting critical concepts or hallucinating entities. Our approach addresses these issues by anchoring the extraction process to verified taxonomies, enforcing entity constraints during LLM prompting and validating outputs via RAG. Through a climate science case study using our annotated dataset of 25 publications (1,705 entity links, 3,618 relationships), we demonstrate that taxonomy-guided LLM prompting combined with RAG-based validation reduces hallucinations by 23.3% while improving F1 scores by 13.9% compared to baselines without the proposed techniques. Our contributions include: 1) a generalizable methodology for taxonomy-aligned KG construction; 2) a reproducible annotation pipeline, 3) the first benchmark dataset for climate science information retrieval; and 4) empirical insights into combining structured taxonomies with LLMs for specialized domains. Code and data will be released upon acceptance.

1 Introduction

Effective management and utilization of structured knowledge is a core challenge in domain-specific research. While scientific publications across fields, from materials science to epidemiology, routinely describe critical relationships between models, observational datasets, and analytical findings, these connections are rarely formalized or linked to standardized data sources. For instance, climate science papers might detail how green house gas emission affects the occurrence of wildfires (Touma et al., 2021), while chemistry studies could analyzes battery chemistry performance under different extreme conditions (Fan et al., 2024). Yet in both cases, these insights remain trapped in unstructured text, inaccessible to computational analysis. This lack of systematization impedes cross-study knowledge integration, slowing discovery and limiting reproducibility. Knowledge graphs (KGs) address this gap by structuring entities and relationships into semantically interconnected frameworks, enabling querying, automated reasoning, and cross-domain interoperability (Chang et al., 2023).

Although KGs have advanced research in domains like material science (Venugopal et al., 2022) and geospatial sciences (Cogan et al., 2024), constructing them in specialized fields faces two main challenges. First, existing methods overlook domain taxonomies, which are curated hierarchies of verified entities and relationships. Instead, they build KGs from scratch via LLMs. (Edge et al., 2024). While flexible, this forfeits the semantic rigor and community consensus embedded in taxonomies, leading to inconsistent representations. Second, despite LLMs' proficiency in general-purpose information extraction (Xu et al., 2024), they struggle in specialized domains: hallucinating entities, misclassifying relationships, and overlooking tail-domain concepts absent from their training data (Yu et al., 2024). For example, in climate science, models frequently conflate teleconnections (large-scale climate linkages) with generic correlations or fail to recognize emerging terms like 'Arctic amplification'. These errors compromises KG reliability for downstream tasks.

A critical bottleneck in KG construction lies in accurate named entity recognition (NER) for specialized domains. State-of-the-art generalist models like GLiNER (Zaratiana et al., 2024), which achieve competitive performance on broad-coverage benchmarks (F1: 0.478), falter in domain-specific settings—scoring only 0.339 F1 on climate science texts. This performance gap stems from two interrelated issues: 1) Domain-specific terminology—such as teleconnections, oceanic Rossby waves, and CMIP6 emission scenarios—occupies the “long tail” of knowledge underrepresented in LLM training corpora (Yu et al., 2024), and 2) LLMs lack mechanisms to disambiguate domain-relevant entities (e.g., "water" as a model variable in hydrological studies) from semantically similar generic terms (e.g., generic mentions of "water" in non-technical contexts or "signal processing" in electronics). Consequently, LLMs either omit critical concepts or misclassify them, propagating errors into downstream KG components.

To address these challenges, we propose a framework that synergizes domain taxonomies, constrained LLM extraction, and iterative validation, demonstrated through climate science KG construction. Our approach comprises three key components: **1) Taxonomy-driven KG construction**: Extraction is anchored to expert-curated taxonomies (e.g., MeSH in biomedicine, NASA’s GCMD (Nagendra et al., 2001) in climate science). By integrating RAG with LLMs, we ensure extracted entities (e.g., CMIP6 experiments) and relationships (e.g., ENSO influences Drought) align with the taxonomy’s hierarchical structure, preserving semantic consistency. **2) Constrained Entity and Relation Typing**: To reduce hallucinations, we restrict the types of named entities (NEs) and relations that LLMs can extract. This prevents irrelevant entity types, such as person names, from being included. Few-shot learning is employed to adapt the model to domain tasks, improving performance. **3) RAG-based output verification**: Unlike approaches like GraphRAG (Edge et al., 2024), which directly use model outputs for KG construction, we verify outputs using RAG against the domain taxonomy. This prevents the introduction of wrong entities and relations into the graph.

Our work advances domain-specific KG construction through the following contributions:

- **A Generalizable Taxonomy-Driven Methodology**: While demonstrated in climate science, our framework provides a blueprint for constructing KGs in any domain with structured taxonomies (e.g., Space Domain Awareness taxonomy). By anchoring extraction to expert-curated hierarchies, we ensure semantic consistency while enabling sustainable updates.
- **Hallucination-Robust LLM-RAG Integration**: We demonstrate how RAG-enhanced LLMs, constrained by taxonomic rules, reduce entity hallucination by 23% compared to baseline methods while maintaining 47% recall on tail-domain concepts.
- **A Reproducible Climate Science Benchmark**: A curated dataset of 25 publications with 1,705 entity-publication links and 3,618 expert-validated relationships.
- **Rigorous Evaluation Framework**: Ablation studies and cross-model comparisons quantify the impact of taxonomy anchoring, showing 18% F1 gains over SOTA models like GLiNER in climate science NER—a pattern generalizable to other specialized domains.

This work bridges unstructured scientific text and structured knowledge representation, offering a scalable solution not only for climate science but for any domain requiring precise, taxonomy-grounded KGs. By addressing the dual challenges of semantic consistency and domain adaptability, our framework empowers researchers to systematically organize evolving knowledge while preserving interoperability with established taxonomies.

2 Related Work

2.1 KGs & Taxonomy Integration

Domain-specific KGs have drive advances across scientific fields, from accelerating material discovery (Venugopal et al., 2022) to enabling environmental decision-making through geospatial KGs like KnowWhereGraph (Cogan et al., 2024). However, most approaches neglect existing domain taxonomies. While projects like SNOMED-CT (healthcare) and Materials Ontology provide curated hierarchies, current KG construction methods often rebuild entity structures from scratch rather than leveraging these semantic scaffolds. This oversight leads to redundant efforts and weakens interoperability. For example, biomedical KGs frequently over-represent common concepts while under-representing niche terms (Stephen et al., 2021). Our work addresses this gap by formalizing taxonomy integration as a first-class paradigm for KG construction, ensuring semantic consistency while preserving domain-specific nuance.

2.2 LLMs for Domain-Specialized Extraction

LLMs excel in general-purpose information extraction (Gabriel et al., 2024), but struggle in scientific domains, exhibiting high hallucination for tail concepts (Viviane et al., 2024) and inconsistent recognition of domain-specific entities. Recent mitigations like contrastive decoding (Derong et al., 2024) and domain-adapted models (e.g., SciLitLLM (Sihang et al., 2024)) improve precision but remain taxonomy-agnostic. Our framework advances this paradigm by hard-constraining LLMs to predefined entity/relationship types from domain taxonomies. This approach generalizes beyond climate science. In materials science, it can constrain entity recognition to the Materials Ontology while excluding irrelevant chemical classifications.

2.3 Retrieval-Augmented Generation

RAG has become a key strategy to improve LLM reliability, with applications ranging from PaperQA's provenance-aware scientific QA (Jakub et al., 2023) to G-RAG's graph-enhanced retrieval in materials science (Radeen et al., 2024). However, existing RAG systems prioritize document-level context over taxonomy alignment, risking semantic drift. For example, ATLANTIC (Sai et al., 2023) improves cross-disciplinary coherence but lacks mechanisms to validate entities against domain hierarchies. Our work introduces taxonomy-guided RAG, where retrieval candidates are filtered through domain-specific taxonomies (e.g., GCMD for climate science) before LLM processing. This dual-phase approach retrieves from both literature and taxonomies. It ensures extracted entities map to verified concepts rather than hallucinated variants.

3 Method Overview

We propose a generalizable framework for constructing domain-specific KGs that harmonizes structured taxonomies with unstructured text extraction. While demonstrated through climate science, a domain with complex terminology and rapid conceptual evolution—the methodology applies to any field with curated vocabularies (e.g., Unified Astronomy Thesaurus or GeoNames in geospatial sciences). The framework comprises three stages: **1) Taxonomy as Semantic Scaffold**: Domain taxonomies (e.g., GCMD for climate science) define entity hierarchies and relationship rules, ensuring consistency. **2) LLM-RAG Hybrid Extraction**: RAG grounds LLMs in taxonomy entities during extraction, reducing hallucinations while preserving contextual nuance. **3) Dynamic KG Assembly**: Validated entities and relationships are integrated into a graph that evolves with publications, balancing taxonomic rigor with conceptual growth.

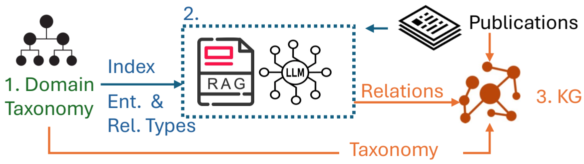


Figure 1: Overview of the proposed framework for Knowledge Graph construction

Figure 1 illustrates the proposed framework for KG construction from scientific publications. We start with a taxonomy, which provides a hierarchical classification of domain-specific named entities but lacks explicit relationships beyond hierarchical structures such as subclass relations. To enrich this taxonomy, we incorporate a broader set of relations that define interactions between entities. These relations are automatically derived from research publications, but are constrained by our RAG to predefined types of relations and entities within the taxonomy, ensuring consistency and mitigating hallucinations. The taxonomy serves as the structural foundation of the KG, anchoring entity organization, while the extracted relations add depth by capturing meaningful interactions between entities.

4 Stage 1: Taxonomy Integration

We propose a 3-step framework to transform domain taxonomies into adaptive backbones for KG construction, applicable to scientific fields requiring structured yet evolving knowledge representation. Using climate science as a case study, the process involves: aggregating domain-specific taxonomies, enhancing node definitions, and indexing for semantic alignment.

4.1 Aggregate Domain-related Taxonomies

KG construction begins by unifying domain-specific taxonomies. Starting with a core taxonomy (e.g., NASA's GCMD for climate science), we integrate: 1) Controlled vocabularies: Standardized terms from modeling protocols or experimental frameworks (e.g., CMIP6CV (Taylor et al., 2018)); 2) Data Repositories: Entity labels from observational datasets, clinical databases, or institutional repositories (e.g., obs4MIPs (Waliser et al., 2020) for climate observations; and 3) Domain-Specific Standards: Expert-curated resources tailored to niche subfields (e.g., CMIP Pub Hub[1]).

In the climate science case study, we constructed the taxonomy GCMD+ with publically available resources: GCMD, CMIP6CV, obs4MIPs and CMIP Pub Hub. Each entity in GCMD+ is assigned with a unique hierarchical path and identifier, resulting in a total of 16,360 entities, an 18% increase over the base GCMD. To enhance interoperability, we link the taxonomy to a cross-domain knowledge base, Wikidata, through Entity Matching and Metadata Integration, detailed in Appendix A.1.

Why Not General Taxonomies? Broad resources like Wikidata introduce noise through excessive granularity (e.g., redundant storm classifications by years) and irrelevant entities. Domain-specific taxonomies prioritize precision, leveraging curated hierarchies validated by practitioners.

4.2 Enhance Definitions

Taxonomy nodes often lack standardized definitions. In GCMD+, 30% of nodes lacked definitions. We address this using Llama-3.3-70B (Grattafiori et al., 2024) to generate concise descriptions using the node label, hierarchical path, and original definitions (where available). This improved definition coverage while standardizing length and clarity across the taxonomy. Additionally, removing irrelevant detail and standardized vocabulary improves indexing in later stages.

4.3 Indexing for Dynamic Alignment

All entities are embedded using NVIDIA NV-Embed-v2 (Lee et al., 2024) (4096 dimensions), a top-performing model on the MTEB benchmark (Muennighoff et al., 2022). The embeddings enable semantic search and link literature-extracted knowledge to taxonomy. This indexing ensures the taxonomy serves as a stable anchor for maintaining semantic consistency across the evolving KG.

5 Stage 2: Information Extraction via LLM-RAG Synergy

Figure 2 outlines our 3-step pipeline for taxonomy-guided information extraction: 1) prompt engineering, 2) constrained entity/relationship extraction, and 3) validation against domain taxonomies. Below we detail each stage.

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

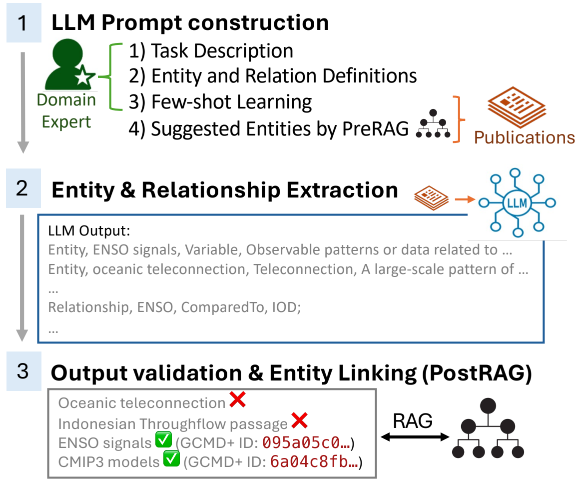


Figure 2: Stage 2: Information Extraction from publications using LLM and RAG

5.1 LLM Prompt Construction

A trivial prompt asking the LLM to extract entities and relationships from domain science literature is insufficient for ensuring accuracy, consistency, and alignment with domain knowledge. Without constraints, the model tends to hallucinate entity types, introduce ambiguous relationships, and deviate from the standardized terminology needed for structured knowledge representation. To address these challenges, we construct a domain-specific prompt framework guided by the taxonomy. The taxonomy serves as a backbone, constraining the LLM's outputs to predefined entity types and relationships, thereby reducing ambiguity and ensuring semantic coherence. We developed a 4-component prompt framework based on GraphRAG (Edge et al., 2024) (Figure 2, Step 1). The complete prompt template is provided in Appendix A.2.

Task Description : Defines the task of identifying entities from predefined domain types and extracting contextual relationships between them. This ensures outputs align with taxonomic constraints while preserving contextual nuance.

Entity & Relation Definitions: **1) Entities**: The taxonomy provides a hierarchical organization of terms, where higher-level nodes represent abstract entity types (e.g., *Teleconnection*, *Model*, and *Ocean Circulation*), while lower-level nodes correspond to specific instances. Experts select entity types from the higher-level nodes, ensuring alignment with domain interest. **2) Relationships**: Domain-critical interactions are defined by domain experts(e.g., 9 climate relationships like

ComparedTo and *MeasuredAt*).

Few-Shot Learning Few-shot learning (Yao et al., 2024; Dai et al., 2022) played a critical role in adapting the model to domain nuances. We include 10 annotated examples in the prompt to explicitly demonstrate NER and relationship extraction (RE) patterns. These examples cover all predefined types. This is particularly necessary because naive prompting leads to inconsistencies in entity classification and relationship identification.

Input with RAG Results (PreRAG) To further constrain the model and improve precision, we leveraged RAG to retrieve suggested entities using a multistep process: 1) Extract noun phrases from input text using SpaCy dependency parsing. 2) Apply pre-defined rules to filter out irrelevant phrases, such as non-climate-related terms, skip words, or phrases shorter than three characters. 3) Retrieve the most similar taxonomy nodes for each noun phrase using cosine similarity between the noun phrase embedding and node embeddings. 4) Retain candidates with similarity scores above 0.6 and append them to the input text as '*Potential Entities:*'. This process enriched the input context while maintaining strict alignment with the verified taxonomy. The 0.6 threshold balances precision and recall based on experimentation. Lower values (e.g., 0.5) caused excessive false positives, while higher values (e.g., 0.7) missed relevant entities.

5.2 Entity & Relationship Extraction

The LLM (e.g., Llama-3.3-70B-Instruct (Grattafiori et al., 2024)) processes the inputs from Section 5 to extract entities and relations from publications.

5.3 Output Validation (PostRAG)

Extracted candidates undergo rigorous validation (Figure 2, Step 3): First, each extracted entity, along with its description, is matched to domain taxonomy nodes (e.g., GCMD+ or MeSH) via cosine similarity. The entity's predicted description is leveraged to retrieve potential matches from domain taxonomy based on semantic similarity. Entities with high-similarity (0.6+) matches are accepted for inclusion in the graph.

Second, the validated entities are used to establish paper-mention-entity relationships, which are incorporated into the KG. Publications act as sources of evidence for these relationships, enhancing the KG's reliability and utility. Furthermore, only predicted relationships involving validated entities are added to the graph. Entities without sufficiently confident matches are excluded from the final graph to prevent the introduction of noise or misinformation. This process is critical for minimizing hallucinations and ensuring alignment with the domain taxonomy.

Through this structured approach, the taxonomy serves as an anchor throughout the extraction pipeline, ensuring that entity recognition, relationship extraction, and knowledge graph integration remain grounded in verified domain knowledge.

6 Stage 3: Dynamic KG Assembly & Maintenance

Our framework constructs domain-specific KGs that balance taxonomic stability with adaptability. The resulting KG (e.g., ClimatePubKG for climate science) integrates entities from domain taxonomies (e.g., GCMD+) and scholarly publications into a unified graph database (e.g., Neo4j). Each relationship inherits provenance metadata—including paper references, cited text snippets, and contextual mentions—enabling evidence-based queries. For instance, in climate science, a *MeasuredAt* relationship between ENSO signals and an oceanic location links to the source publication's methodology section.

We demonstrate through a climate science case study: processing 300 papers from Semantic Scholar established 21K validated entity-publication links (e.g., connecting CMIP3 models to teleconnection studies). Automated pipelines continuously ingest new publications, expanding coverage while enforcing taxonomic alignment.

To balance comprehensiveness with reliability, unlinked entities (e.g., emerging terms like "subsurface salinity fronts") undergo systematic monitoring. 1) Frequency Tracking: Entities surpassing occurrence thresholds are flagged. 2) Expert Validation: Domain specialists assess candidates for taxonomy inclusion. 3) Taxonomy Extension: Approved entities are added with unique identifiers.

This process filters transient concepts while integrating validated knowledge. The KG architecture supports dual roles: a historical repository and a live research tool. In climate science, feedback loops between experts and extraction models enable real-time hypothesis testing (e.g., validating new teleconnection patterns against historical data).

By grounding KGs in taxonomies while accommodating domain evolution, our framework achieves precision at scale—critical for fields like climate science where terminology and relationships evolve rapidly. The methodology generalizes to other domains through configurable taxonomic constraints and validation rules.

7 Domain-Specific Annotation Pipeline

We demonstrate through a climate science annotation pipeline, validated by 4 domain experts. The 3-step process balances efficiency and precision through iterative refinement: **Step 1: NER**: Annotators validate LLM-generated pre-annotations (e.g., Llama-3.3 predictions) against domain-specific guidelines, tagging 12 predefined categories (Appendix A.2). Irrelevant predictions such as *person names* are filtered out, while missing domain entities (e.g., *teleconnections*) are added. This step achieved moderate inter-annotator agreement (Kappa: 0.77), reflecting challenges in consistently identifying climate science entities, particularly nuanced variables like *orbital period* and domain-specific experiments like *RCP*. **Step 2: Entity Linking (EL)**: (Kappa: 0.89) Validated entities are mapped to GCMD+ taxonomy IDs. Ambiguous cases are flagged for expert review, while unmatched entities are retained for evaluation. **Step 3: RE**: (Kappa: 0.82) Annotators verify and add relationship predictions between entities, excluding speculative or unsupported connections.

At each step, the consistency of the annotated entities and relationships was verified, and discrepancies were resolved collaboratively. Using the INCEpTION annotation tool, (Klie et al., 2018) we annotated 25 publications from Semantic Scholar, covering a wide range of climate science topics, including atmospheric processes, ocean dynamics, and climate modeling. This yielded 13,773 entity mentions (10,174 linked to GCMD+) and 3,618 validated relationships. Frequent categories include *variable* (3,953 mentions), *location* (2,767), and (climate) *model* (1,500), as detailed in Appendix A.5. By recycling step outputs as inputs (e.g., NER results inform linking), we reduced annotation effort. Annotation guidelines are in Appendix A.9.

8 Experiments

The experiments aim to evaluate the proposed framework's effectiveness and investigate the contributions of its key components, including few-shot learning, RAG, backbone models, and relationship extraction. The evaluation is conducted on three tasks: NER, EL, and RE.

8.1 Evaluation Protocol

We evaluate using 600-token chunks with 100-token overlaps, following GraphRAG (Edge et al., 2024). For NER, the strict measure requires exact matches between predicted and ground truth entity strings with matching labels (Ojha et al., 2023). The relaxed measure counts predictions as correct if they overlap with ground truth substrings, regardless of label. It retains only the longest non-overlapping substring in both ground truth and predictions (e.g., preferring 'long-latitudes' over 'latitude'). This approach evaluates the model's ability to identify unique entities while handling terminological variations common in scientific literature.

For RE, strict evaluation requires exact matches for source entity, target entity, and type, while relaxed evaluation ignores type. EL performance is assessed by comparing PostRAG entity IDs against human-annotated GCMD+ IDs.

We compute precision (P), recall (R), F1-score (F1), prediction count (#PD), and ground truth count (#GT) at both chunk and paper levels. Paper-level results are in Appendix A.6.

8.2 Backbone Model Comparison

We evaluate the proposed method using multiple backbone models to assess performance variations. **1) Scale variants**: Llama-3.3-8B-Instruct (Grattafiori et al., 2024) vs. Llama-3.3-70B-Instruct (Grattafiori et al., 2024) measure model size impact. **2) Commercial APIs**: GPT-4o (OpenAI et al., 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2024) as proprietary alternatives.

We also include generalist NER baselines, GLiNER (Zaratiana et al., 2024) and NuNER (Bogdanov et al., 2024), which rely solely on text input and label names. This setup isolates the effects of model architecture, parameter count, and domain specialization under identical taxonomy constraints and RAG configurations across experiments.

All non-API models are run on a server with two NVIDIA A100 80GB GPUs. These experiments provide insights into the trade-offs between model size, cost, and accuracy, guiding the choice of backbone models for practical deployments.

8.3 Ablation Studies

Few-Shot vs. Zero-Shot Learning To assess in-context learning, we compare the framework with few-shot examples (10-shot, 1-shot) and without (0-shot). The few-shot setup includes climate-specific examples. This evaluates its impact on NER, EL, and RE, highlighting its benefits for domain-specific extraction.

RAG Efficiency RAG's effectiveness is assessed by comparing the method with and without RAG-generated input candidates (PreRAG) to isolate its impact on entity recognition and linking. For post-processing (PostRAG), predictions are compared against annotations with linked GCMD+ IDs, while base predictions use all ground truth entities.

Isolating Relationship Extraction (NER only) To isolate the contribution of the relationship extraction stage, we conduct an ablation study comparing the full pipeline with a configuration that includes only NER and EL. This experiment quantifies the incremental performance gain achieved by relationship extraction and demonstrates its importance in building KGs. The results reveal how the omission of this stage affects the system's ability to capture entity interactions and dependencies.

9 Results and Discussion

Our proposed framework includes all components including 10-shot, PreRAG, PostRAG and Relationship Extraction. Experiments yield three key findings. First, taxonomy constraints with LLMs significantly improves climate science information extraction. Second, retrieval augmentation and few-shot learning effectively reduce hallucinations. Third, relationship extraction introduces precision-recall trade-offs requiring careful balancing.

9.1 Ablation Studies

As can be seen in Table 1 our best-performing model according to NER F1 score is Llama-3.3 across all tested LLMs. Therefore, our ablation studies are based on Llama-3.3. Key findings from ablation studies highlight the contributions of each framework component:

Few-Shot Few-shot learning consistently improves NER performance significantly, as can be seen in Table 1 by comparing Llama-3.3 with all proposed components (including 10-shot) to Llama 3.3 with 0 shot: improvement **13.9%** (0.440 → 0.501). Adding just 1 example (1-shot) boosts NER F1 by 5.8% (0.440 → 0.464). This underscores the value of minimal in-context guidance.

RAG Contribution RAG is critical for disambiguation. Removing PreRAG (suggested candidates by RAG) reduces NER F1 by 3.2% (0.501 → 0.485) (Table 1). This highlights the importance of input candidates in improving extraction accuracy and reducing hallucinations. PostRAG processing reduces false positives by **23.3%**, as evidenced by precision jumps from 0.536 to 0.661 in NER. Relaxed F1 rises to 0.525—an 5% gain over the model without PostRAG. This validates our hypothesis that taxonomic constraints mitigate LLM hallucinations while preserving recall.

Isolating Relationship Extraction While removing the relationship extraction task marginally improves NER relaxed F1 (+4.2%; 0.501→0.522) and EL F1 (+3.3%; 0.367→0.379), these gains come at the expense of losing all relationship semantics critical for KG applications. Crucially, maintaining separate NER/EL and relationship stages doubles LLM computational costs due to redundant prompt processing. Our experiments suggest practitioners may prioritize relationship extraction when domain interactions are mission-critical (e.g., climate analysis), while considering the NER/EL-only approach for resource-constrained entity-centric use cases.

Model Scale Larger models (70B vs. 8B) improve NER F1 by 33% (0.395 → 0.525), as increased model size better captures domain nuances. This aligns with findings in other specialized domains, where model scale correlates with performance on tail concepts and complex terminology.

9.2 Information Extraction Performance

Entity Extraction As Table 1 shows, Llama-3.3-70B achieves 0.501 F1 (relaxed) and 0.378 F1 (strict) on NER, outperforming generalist models like GLiNER (0.461 F1) and domain-specific baselines like ClimateGPT (0.110 F1).

Entity-type analysis with Llama-3.3 (Appendix A.5) shows performance correlates with taxonomic standardization in that well-defined categories like Teleconnection (0.61 F1) and Model (0.53 F1) outperform ambiguous types (i.e., not well-defined) like Platform (0.04 F1).

Error analysis highlights two key limitations. 1) Our LLMs frequently extracted acronyms (e.g., "SAM") while ignoring full names ("Southern Annular Mode"), even when both appeared in context. 2) It inconsistently handled term variants, retaining

			Relaxed						Strict					
			All NEs			PostRAG			All NEs			PostRAG		
	Model	#Params	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Proposed	Llama-3.3	70B	.536	.471	.501	.661	.436	.525	.432	.337	.378	.530	.310	.391
	Llama-3.1	8B	.385	.346	.364	.533	.314	.395	.291	.239	.262	.413	.220	.287
	DeepSeek-V3	671B	.572	.350	.435	.604	.336	.432	.472	.255	.331	.498	.244	.328
	ClimateGPT	70B	.494	.062	.110	.495	.104	.172	.305	.034	.062	.325	.061	.102
	GPT 4o	200B	.602	.323	.420	.663	.304	.417	.455	.214	.291	.510	.205	.292
Generalist	NuNER	0.35B	.727	.307	.431	-	-	-	.512	.196	.284	-	-	-
	GLiNER	0.3B	.591	.378	.461	-	-	-	.458	.269	.339	-	-	-
0-shot	Llama-3.3	70B	.469	.414	.440	.603	.386	.470	.358	.285	.317	.461	.266	.338
1-shot			.504	.431	.464	.641	.405	.497	.386	.295	.334	.485	.274	.350
NER only			.517	.456	.485	.688	.413	.516	.406	.316	.355	.535	.282	.370
No PreRAG			.539	.505	.522	.653	.468	.545	.431	.360	.392	.521	.333	.406

Table 1: NER performance for the proposed framework and ablations. Best proposed model scores are underlined.

	Model	P	R	F1	#PD
Proposed	Llama-3.3	.440	.315	.367	4,051
	Llama-3.1	.396	.247	.304	3,540
	DeepSeek-V3	.457	.272	.341	3,365
	ClimateGPT	.478	.108	.176	828
	GPT 4o	.497	.246	.330	2,779
0-shot	Llama-3.3	.427	.294	.348	3,788
1-shot		.448	.304	.362	3,840
No PreRAG		.456	.298	.360	3,692
NER only		.435	.336	.379	4,388

Table 2: Entity linking performance

		Relaxed			Strict		
	Model	P	R	F1	P	R	F1
Proposed	Llama-3.3	.066	.096	.078	.045	.066	.053
	Llama-3.1	.026	.042	.032	.016	.027	.020
	DeepSeek-V3	.075	.072	.073	.034	.032	.033
	ClimateGPT	.096	.066	.079	.000	.000	.000
	GPT 4o	.009	.001	.001	.060	.041	.049
0-shot	Llama-3.3	.037	.083	.051	.012	.028	.017
1-shot		.047	.076	.058	.031	.050	.038
No PreRAG		.064	.096	.076	.040	.061	.048

Table 3: Relationship extraction performance

"anthropogenic climate change" but omitting synonymous phrases like "climate change impacts" in the same sentences. Appendix A.3 illustrates these patterns through annotated examples.

Entity Linking Taxonomy-guided linking achieves 0.367 F1 (Table 2), with GPT-4o leading in precision (0.497) and Llama-3.3-70B in recall (0.315). The precision-recall gap reflects a trade-off: strict taxonomic alignment avoids false links but may omit novel concepts. Our dynamic update mechanism addresses this by tracking high-frequency unlinked entities for expert review.

Relationship Extraction While RE is critical for KG completeness, it remains challenging. ClimateGPT achieves the highest relaxed F1-score (0.079) but scores 0 under strict evaluation (Table 3). The performance of Llama-3.3 is more stable scoring 0.078 (relaxed) and 0.053 (strict). Similar to NER, Llama-3.3 with the proposed components performs the best. When entity matching is relaxed to allow partial alignment of source and target entities (Appendix A.7), ClimateGPT scores 0.015 F1, and Llama-3.3 scores 0.244 F1. Beyond identifying correct entity pairs, poor matching further complicates RE; even PostRAG (App.A.7) offers little help if entity matching fails.

10 Conclusion

In this work, we presented a taxonomy-driven framework for domain-specific KG construction using LLMs and RAG. Our approach addresses the challenges of extracting and organizing domain-specific knowledge from unstructured scientific literature. By grounding the KG construction process in a taxonomy (NASA's GCMD), we ensured semantic consistency and reduced hallucinations commonly associated with LLMs.

Our experiments demonstrated the effectiveness of integrating RAG with LLMs for KG construction, particularly in improving precision and reducing false positives in entity recognition and relationship extraction. The use of few-shot learning further enhanced the model's ability to adapt to the climate science domain, even with minimal training examples. Additionally, our curated dataset and annotation pipeline provide a valuable resource for future research in climate science information extraction. While demonstrated in climate science, our framework provides a blueprint for any domain with structured taxonomies. By converting unstructured text into structured, machine-readable knowledge representation, this work enables large-scale organization of specialized scientific information.

11 Limitations

Our approach faces several important constraints in constructing climate science KGs. The GCMD+ taxonomy, while comprehensive, may not fully capture emerging concepts in climate science, creating potential gaps in knowledge representation. Since our dynamic maintenance process includes climate experts in the loop, it can introduce delays in incorporating new terminology, affecting the KG's currency.

Despite taxonomic anchoring, performance varies by entity type—well-defined categories like *Teleconnection* achieve 0.61 F1 versus 0.04 F1 for ambiguous Platform entities. Acronym disambiguation (e.g., "SAM" vs. "Southern Annular Mode") remains unresolved, with 58% of errors stemming from partial term extraction.

The entity linking process presents technical challenges, particularly in our fuzzy string matching approach for Wikidata integration. Using a 60% similarity threshold involves trade-offs between coverage and accuracy, potentially missing valid matches or creating incorrect associations for complex scientific terms.

Our method's focus on English-language scientific literature introduces a language bias, potentially overlooking valuable climate knowledge in other languages. The predefined relationship types may not capture all nuanced interactions between climate science entities, particularly in interdisciplinary contexts.

These limitations suggest several directions for future research, including developing multilingual extensions, implementing more efficient computational approaches, and creating automated mechanisms for taxonomy extension that can better keep pace with advancing climate science knowledge.

References

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. *Preprint*, arXiv:2402.15343.

Rihao Chang, Yongtao Ma, Tong Hao, and Weizhi Nie. 2023. 3d shape knowledge graph for cross-domain 3d shape retrieval. *Preprint*, arXiv:2210.15136.

Shimizu Cogan, Stephe Shirly, Barua Adrita, Cai Ling, Christou Antrea, Currier Kitty, Dalal Abhilekha, Fisher Colby, K., Hitzler Pascal, Janowicz Krzysztof, Li Wenwen, Liu Zilong, Mahdavinejad Mohammad, Saeid, Mai Gengchen, Rehberger Dean, Schildhauer Mark, Shi Meilin, Norouzi Sanaz, Saki, Tian Yuanyuan, Wang Sizhe, Wang Zhangyu, Zalewski Joseph, Zhou Lu, and Zhu Rui. 2024. The knowwheregraph ontology. *arXiv preprint arXiv:2410.13948*.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *Preprint*, arXiv:2209.11755.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xu Derong, Zhang Ziheng, Zhu Zhihong, Lin Zhenxi, Liu Qidong, Wu Xian, Xu Tong, Zhao Xiangyu, Zheng Yefeng, and Chen Enhong. 2024. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. *arXiv preprint arXiv:2410.15702*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Guodong Fan, Boru Zhou, Chengwen Meng, Tengwei Pang, Xi Zhang, Mingshu Du, and Wei Zhao. 2024. Development of a comprehensive physics-based battery model and its multidimensional comparison with an equivalent-circuit model: Accuracy, complexity, and real-world performance under varying conditions. *Preprint*, arXiv:2411.12152.

Garcia Gabriel, Lino, Ribeiro Manesco João, Renato, Paiola Pedro, Henrique, Miranda Lucas, de Salvo Maria, Paola, and Papa João, Paulo. 2024. A review on scientific knowledge extraction using large language models in biomedical sciences. *arXiv preprint arXiv:2412.03531*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lála Jakub, O'Donoghue Odhran, Shtedritski Aleksandar, Cox Sam, Rodriques Samuel, G., and White Andrew, D. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

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.

Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors. 2023. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka

Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Mostafa Radeen, Baig Mirza, Nihal, Ehsan Mashaekh, Tausif, and Hasan Jakir. 2024. G-rag: Knowledge expansion in material science. *arXiv preprint arXiv:2411.14592*.

Munikoti Sai, Acharya Anurag, Wagle Sridevi, and Horawalavithana Sameera. 2023. Atlantic: Structure-aware retrieval-augmented language model for interdisciplinary science. *arXiv preprint arXiv:2311.12289*.

Li Sihang, Huang Jin, Zhuang Jiaxi, Shi Yaorui, Cai Xiaochen, Xu Mingjun, Wang Xiang, Zhang Linfeng, Ke Guolin, and Cai Hengxing. 2024. Scilitllm: How to adapt llms for scientific literature understanding. *arXiv preprint arXiv:2408.15545*.

Bonner Stephen, Kirik Ufuk, Engkvist Ola, Tang Jian, and Barrett Ian, P. 2021. Implications of topological imbalance for representation learning on biomedical knowledge graphs. *arXiv preprint arXiv:2112.06567*.

Karl E Taylor, Martin Juckes, V Balaji, Luca Cinquini, Sébastien Denvil, Paul J Durack, Mark Elkington, Eric Guilyardi, Slava Kharin, Michael Lautenschlager, et al. 2018. Cmip6 global attributes, drs, filenames, directory structure, and cv's. *PCMDI Document*.

Danielle Touma, Samantha Stevenson, Flavio Lehner, and Sloan Coats. 2021. Human-driven greenhouse gas and aerosol emissions cause distinct regional impacts on extreme fire weather. In *AGU Fall Meeting Abstracts*, volume 2021, pages A51E–01.

Vineeth Venugopal, Sumit Pai, and Elsa Olivetti. 2022. Matkg: The largest knowledge graph in materials science – entities, relations, and link prediction through graph representation learning. *Preprint*, arXiv:2210.17340.

da Silva Viviane, Torres, Rademaker Alexandre, Lionti Krystelle, Giro Ronaldo, Lima Geisa, Fiorini Sandro, Archanjo Marcelo, Carvalho Breno, W., Neumann Rodrigo, Souza Anaximandro, Souza João, Pedro, Valnisio Gabriela, de, Paz Carmen, Nilda, Cerqueira Renato, and Steiner Mathias. 2024. Automated, llm enabled extraction of synthesis details for reticular materials from scientific literature. *arXiv preprint arXiv:2411.03484*.

D. Waliser, P. J. Gleckler, R. Ferraro, K. E. Taylor, S. Ames, J. Biard, M. G. Bosilovich, O. Brown, H. Chepfer, L. Cinquini, P. J. Durack, V. Eyring, P.-P. Mathieu, T. Lee, S. Pinnock, G. L. Potter, M. Rixen, R. Saunders, J. Schulz, J.-N. Thépaut, and M. Tuma. 2020. Observations for model intercomparison project (obs4mips): status for cmip6. *Geoscientific Model Development*, 13(7):2945–2958.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Preprint*, arXiv:2312.17617.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. *Preprint*, arXiv:2403.18167.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Linking with WikiData

To enhance interoperability, we link the taxonomy to a cross-domain knowledge base, Wikidata in two phases:

Entity Matching: Retrieve 10 Wikidata candidates per taxonomy entity, filtering matches via fuzzy string alignment (70% threshold). In climate science, this yields 5,098 validated mappings from 10,623 candidates. **Metadata Integration**: Matched entities were enriched with Wikidata IDs, definitions, and relationships (e.g., broader/narrower terms), enhancing cross-domain interoperability. This step added semantic granularity to 31% of GCMD+ entities while maintaining alignment with the original taxonomy structure.

A.2 Prompt

Table 4 shows the prompt being used for Climate Science Entity and Relationship Extraction from the climate science literature. Table 5 shows the prompt template for refining the node definitions.

A.3 Entity extraction prediction

We employ regular expressions to align predicted entity names with the input text, enabling precise boundary matching. Figures 3, 4, and 5 visualize raw(Yellow: PD$_all)andPostRAG(Blue : PD_post)predictionsfromLlama - 3.3 - 70B, showcasingexamplesfromevaluationdocuments.$

A.4 Model selection choice

Fine-tuning large models such as Llama-3.3-70B was not explored due to its high computational cost and inefficiency for domain-specific tasks. Instead, we rely on in-context learning with few-shot examples and RAG to achieve competitive performance with significantly lower resource requirements.

A.5 NER performance per entity type

Entity-type analysis with Llama-3.3 (Table 6) reveals performance correlates with taxonomic standardization.

-Goal-
Given a text document with a preliminary list of potential entities, verify, and identify all entities of the specified types within the text. Note that the initial list may contain missing or incorrect entities. Additionally, determine and label the relationships among the verified entities.

-Entity Types-
A project refers to the scientific program, field campaign, or project from which the data were collected.
A location is a place on Earth, a location within Earth, a vertical location, or a location outside of the Earth.
A model is a sophisticated computer simulation that integrate physical, chemical, biological, and dynamical processes to represent and predict Earth's climate system.
An experiment is a structured simulation designed to test specific hypotheses, investigate climate processes, or assess the impact of various forcings on the climate system.
A platform refers to a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further study of its characteristics.
A instrument is a device used to measure, observe, or calculate.
A provider is an organization, an academic institution or a commercial company.
A variable is a quantity or a characteristic that can be measured or observed in climate experiments.
A weather event is a meteorological occurrence that impacts Earth's atmosphere and surface over short timescales.
A natural hazard is a phenomenon with the potential to cause significant harm to life, property, and the environment.
A teleconnection is a large-scale pattern of climate variability that links weather and climate phenomena across vast distances.
An ocean circulation is the large-scale movement of water masses in Earth's oceans, driven by wind, density differences, and the Coriolis effect, which regulates Earth's climate.

-Relationship Types and Definitions-
ComparedTo: The source entity is compared to the target entity. Outputs: A climate model, experiment, or project (source entity) outputs data (target entity).
RunBy: Experiments or scenarios (source entity) are run by a climate model (target entity).
ProvidedBy: A dataset, instrument, or model (source entity) is created or managed by an organization (target entity).
ValidatedBy: The accuracy or reliability of model simulations (source entity) is confirmed by datasets or analyses (target entity).
UsedIn: An entity, such as a model, simulation tool, experiment, or instrument (source entity), is utilized within a project (target entity).
MeasuredAt: A variable or parameter (source entity) is quantified or recorded at a geographic location (target entity).
MountedOn: An instrument or measurement device (source entity) is physically attached or installed on a platform (target entity).
TargetsLocation: An experiment, project, model, weather event, natural hazard, teleconnection, or ocean circulation (source entity) is designed to study, simulate, or focus on a specific geographic location (target entity).

-Steps-
1. Identify all entities. For each identified entity, extract the following information:
- entity name: Name of the entity
- entity type: One of the following types: [project, location, model, experiment, platform, instrument, provider, variable]
Format each entity as ("entity"<|><entity name><|><entity type><|><entity description>)
2. From the entities identified from step 1, identify all pairs of (source entity, target entity) that are *clearly related* to each other. For each pair of related entities, extract the following information:
- source entity: name of the source entity
- target entity: name of the target entity
- relationship type: One of the following relationship types: ComparedTo, Outputs, RunBy, ProvidedBy, ValidatedBy, UsedIn, MeasuredAt, MountedOn, TargetsLocation
Format each relationship as ("relationship"<|><source entity><|><target entity><|><relationship type>)
3. Return output in English as a single list of all the entities and relationships identified in steps 1 and 2. Use **** as the list delimiter. Do not output any code or steps for solving the question.
4. When finished, output <|COMPLETE|>

######################
-Examples-
{formatted examples}
######################
-Real Data-
######################
Text: {input text}
Potential Entities: {potential entities}
######################
Output:

Table 4: Prompt Template for Climate Science Entity and Relationship Extraction

Figure 3: Example 1 of entity extraction results from a climate science publication.

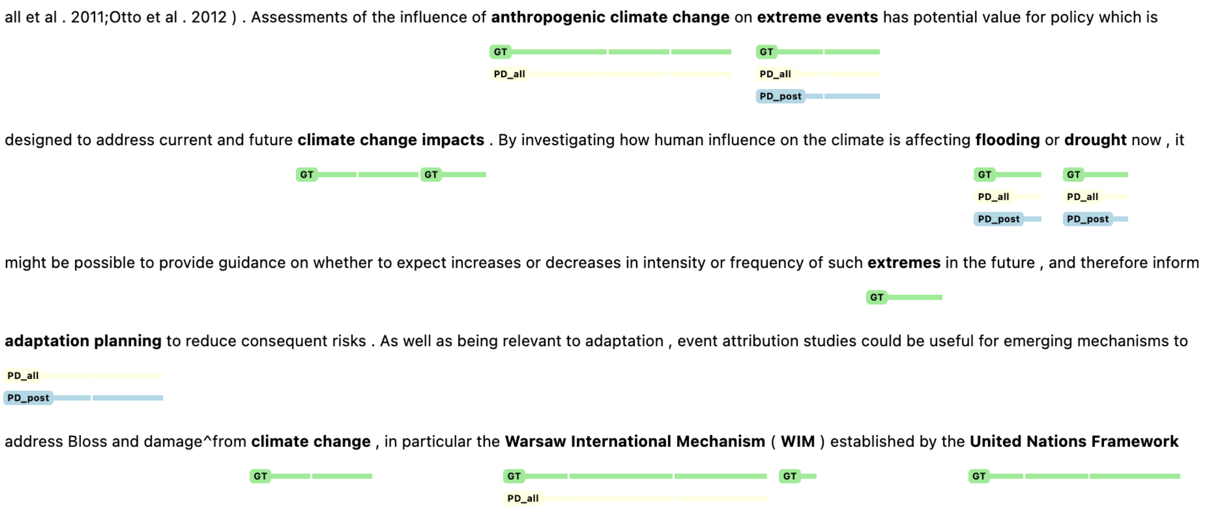


Figure 4: Example 2 of entity extraction results from a climate science publication.

Given the following metadata about an entity in a climate science ontology, which may include the entity's name, ontology path, and a definition (which may be missing), please develop an edited definition suitable for a named entity recognition (NER) task in climate science literature. The definition should be concise, clear, and limited to 150 tokens. Ensure it is precise and emphasizes the entity's unique aspects, avoiding overly general descriptions that could apply to multiple entities. Do not explain; only provide the edited definition.
Metadata: { }
Edited Definition:

Table 5: Prompt Template for Refining Definitions

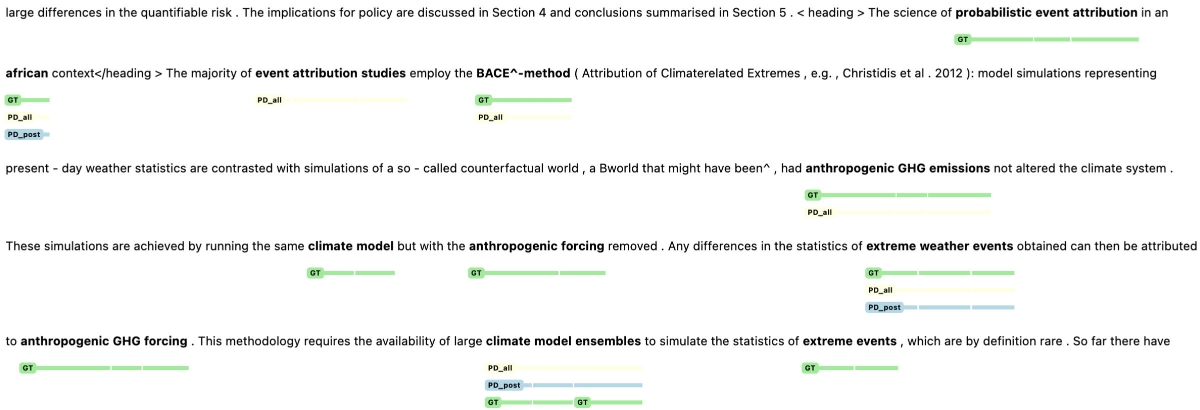


Figure 5: Example 3 of entity extraction results from a climate science publication.

A.6 NER performance on paper level

Table 7 shows paper-level performance metrics averaged across 25 papers. The results align with chunk-level evaluation, suggesting our method maintains consistent performance across different granularities of text processing.

A.7 Relationship Performance (Relaxed)

When entity matching allows partial alignment between source and target entities, the results are presented in Table 8.

A.8 Relationship performance by tag

Table 9 details relationship extraction performance across types for Llama-3.3-70B, evaluated under relaxed and strict criteria. Performance is restricted as exact boundary matching is challenging.

High-Frequency Relationships: *MountedOn* (1,842 instances) achieves poor relaxed F1 (0.058), with strict performance limited by NER's boundary matching challenges. *ComparedTo* (922 instances) shows balanced precision/recall (relaxed F1: 0.088), but struggles with implicit comparisons (e.g., "IOD differs from ENSO" vs. indirect references).

Low-Frequency Challenges: Rare types like *ValidatedBy* (2 instances) and *UsedIn* (14 instances) suffer from data sparsity, yielding near-zero F1.

A.9 Annotation Guidelines

Annotation guidelines are attached at the end.

	All NEs					PostRAG				
Label	**P**	**R**	**F1**	**#PD**	**#GT**	**P**	**R**	**F1**	**#PD**	**#GT**
tele	.73	.53	.61	180	247	.70	.50	.58	148	208
model	.72	.42	.53	870	1500	.65	.46	.54	609	861
loc	.73	.39	.51	1462	2767	.77	.33	.46	947	2233
exp	.45	.48	.47	329	307	.67	.50	.57	216	288
var	.46	.26	.33	2212	3953	.55	.25	.34	1329	2979
proj	.21	.48	.30	549	247	.12	.36	.18	380	131
wea	.21	.25	.23	215	182	.17	.15	.16	141	158
prov	.12	.53	.20	1029	239	.37	.45	.41	174	141
haz	.34	.11	.17	121	358	.33	.10	.15	76	258
instr	.06	.20	.10	221	70	.05	.09	.07	60	32
circ	.05	.20	.08	85	20	.02	.06	.02	63	18
plat	.02	.09	.04	125	34	.00	.00	.00	36	14

Table 6: NER performance from Llama-3.3 by type, comparing All vs PostRAG results. Entity types include Teleconnection (tele), Model (model), Location (loc), Experiment (exp), Variable (var), Project (proj), Weather Event (wea), Provider (prov), Natural Hazard (haz), Instrument (instr), Ocean Circulation (circ), and Platform (plat). Best scores per column are underlined.

		Relaxed						Strict					
		All NEs			PostRAG			All NEs			PostRAG		
	Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Proposed	Llama-3.3	.441	.532	.458	.528	.431	.469	.370	.437	.377	.443	.347	.383
	Llama-3.1	.311	.470	.353	.414	.385	.392	.248	.370	.278	.334	.304	.311
	DeepSeek-V3	.454	.397	.410	.472	.325	.377	.401	.330	.348	.420	.271	.322
	ClimateGPT	.443	.107	.168	.405	.096	.154	.255	.062	.097	.229	.053	.085
	GPT 4o	.478	.375	.403	.530	.301	.377	.384	.299	.319	.430	.237	.298
	NuNER	.620	.341	.438	-	-	-	.464	.253	.326	-	-	-
	GLiNER	.490	.445	.465	-	-	-	.391	.334	.359	-	-	-
0-shot	Llama-3.3	.385	.485	.410	.468	.391	.420	.306	.393	.327	.363	.307	.327
1-shot		.426	.516	.443	.512	.411	.451	.344	.404	.350	.412	.325	.358
No PreRAG		.426	.509	.439	.545	.392	.449	.340	.394	.342	.425	.291	.339
NER only		.438	.556	.468	.510	.450	.471	.365	.454	.385	.423	.361	.383

Table 7: Paper-Level Evaluation of NER performance for the proposed framework and ablation studies, with the best proposed scores underlined.

		Relaxed (Partial)			Relaxed (PostRAG)			Strict (PostRAG)		
	Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Proposed	Llama-3.3	.206	.301	.244	.060	.052	.056	.039	.034	.036
	Llama-3.1	.174	.284	.216	.042	.034	.038	.026	.022	.024
	DeepSeek-V3	.294	.282	.288	.059	.041	.049	.026	.018	.022
	ClimateGPT	.313	.216	.256	.090	.036	.052	.065	.026	.037
	GPT 4o	.132	.008	.015	.000	.000	.000	.000	.000	.000
0-shot	Llama-3.3	.198	.450	.275	.040	.051	.045	.013	.017	.015
1-shot		.205	.335	.255	.050	.050	.050	.031	.031	.031
No PreRAG		.192	.288	.230	.070	.053	.060	.044	.033	.038

Table 8: Relationship Performance with PostRAG and more relaxed metrics that allow partial match of source and target entities.

		Relaxed (Partial)			Relaxed			Strict		
label	**#GT**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
ComparedTo	922	.149	.104	.122	.107	.075	.088	.107	.075	.088
MeasuredAt	263	.094	.285	.141	.045	.137	.068	.045	.137	.068
TargetsLocation	1842	.163	.137	.149	.064	.054	.058	.064	.054	.058
Outputs	465	.137	.095	.112	.056	.039	.046	.056	.039	.046
UsedIn	242	.036	.140	.057	.020	.079	.032	.020	.079	.032
RunBy	35	.014	.057	.022	.014	.057	.022	.014	.057	.022
ProvidedBy	31	.012	.226	.023	.010	.194	.020	.010	.194	.020
ValidatedBy	14	.010	.143	.018	.010	.143	.018	.010	.143	.018
MountedOn	2	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table 9: Relationship Detection Performance from Llama-3.3-70B by different relationship types.

Annotation Guideline

STAGE ONE: Named Entity Recognition

1. Introduction

Purpose of the Manual:
This manual provides detailed instructions for annotating climate-related text or terms extracted from scientific literature. It aims to ensure consistency and accuracy in labelling climate entities, data, and models.

Intended Audience:
The guidelines are designed for annotators, including researchers, climate analysts, scientists, and students, who are familiar with climate science terminology and concepts.

Scope of Annotations:
The annotations focus on specific climate entities, including but not limited to:

- **Earth Systems**: Land, ocean, atmosphere, and biosphere entities.
- **Climate Data**: Specific datasets and measurements.
- **Climate Models**: Global and regional climate models.

2. Definitions and Examples of Key Climate Entities

2.1 Earth Systems

Land:
Refers to a specific region or unit of land that can be described and modeled geographically within the framework of a climate model. **Examples**:

- Continents/Regions: Africa, Ethiopia, United Kingdom (UK), high/mid-latitudes, tropics (tropical regions).
- Land Features: Groundwater, river flow, runoff, streamflow, land cover, land use.
- Specific Landmarks: Amazon Rainforest, Himalayas, United States Midwest (Corn Belt), Antarctica.

Atmosphere:
Refers to the layer of gases surrounding the Earth, which plays a vital role in shaping climate and weather patterns and can be modeled geographically within the framework of a climate model.
Examples:

- Atmospheric Layers: Troposphere, mesosphere.
- Climate Phenomena: Temperature, precipitation, wind, evapotranspiration, clouds.
- Weather Systems: Hadley Cells, Ferrel Cells, Trade Winds, Jet Streams, Monsoons, Intertropical Convergence Zone (ITCZ), El Niño-Southern Oscillation (ENSO), Tornadoes, Thunderstorms.

Oceans:
Refers to the large bodies of saltwater that cover about 71% of the Earth's surface and can be modeled geographically within the framework of a climate model. **Examples**:

- Oceans/Seas: Pacific Ocean, Indian Ocean, Atlantic Ocean.
- Oceanic Features: Gulf Stream, Kuroshio Current, Thermohaline Circulation.
- Climate-Related Ocean Phenomena: Ocean acidification, marine heatwaves, coral reefs, upwelling zones, sea ice, continental shelves.

2.2 Climate Data

Refers to detailed, quantitative measurements or simulations of variables that describe various components of the Earth's climate system. **Examples**:

- Datasets: CRU (Climate Research Unit), GPCC (Global Precipitation Climatology Centre), ERA5 (ECMWF Reanalysis 5th Generation).
- Climate Indices: HadCRUT, MERRA-2, GSMP3.

1309 2.3 Climate Models

Refers to computational models used to simulate the Earth's climate system. **Examples**:

2.4 Global Climate Models (GCMs): CCSM4, CNRM-CM5, HadGEM2-ES.

2.5 Regional Climate Models (RCMs): MICRO, ACCESS-ESM1.5.

3. Key Tags or Labels

Guidelines for Tagging:

- Ensure the correct spelling and usage of tags. For example, use "Variables" consistently, not "Variable>" or other variations.
- Review definitions carefully and apply tags or values strictly based on the provided examples and their accurate definitions.
- If uncertain about the definition of an entity, verify its classification (e.g., variable, teleconnection) before tagging.

Tag	Definition and examples
Variable	represents a specific measurable element or attribute of the climate system that is studied or monitored (e.g., cloud cover, temperature (i.e., surface air, ocean, or groundwater), precipitation, wind speed, vapor pressure, geopotential height, humidity (relative, specific) etc.
Project	refers to a coordinated effort or initiative aimed at investigating specific aspects of climate. Projects often involve multiple stakeholders and produce datasets, models, or assessments (e.g., Coupled Model Intercomparison Project Phase 6 (CMIP6))
Location	refers to the geographic region or coordinates being studied or monitored. This can be global, regional, or local. Examples includes West Africa, Central Africa, East Africa, or Southern Africa; tropics or polar regions; high or mid latitudes regions, specific sites (such as the Amazon, Congo Rainforest or Sahara Desert etc).
Model	refers to computational tool used to simulate and predict climate processes and interactions in the Earth system (e.g., HadGEM3, WRF etc)
Provider	refers to the organization or agency responsible for creating, maintaining, or distributing climate data or tools (e.g., NASA (e.g., GISS for climate models, MERRA datasets); ECMWF (e.g., ERA5 reanalysis datasets); NOAA (e.g., NCEP datasets and climate services).
Instrument	refers to the device or tool used to measure climate variables. Instruments can be ground-based, airborne, or spaceborne. Examples includes Radiosondes (balloons for atmospheric measurements); Satellites (e.g., MODIS, GOES, or Sentinel); Rain gauges and anemometers for ground-level data.
Event	An event is an occurrence or phenomenon in the Earth's system that varies in temporal scale, ranging from short-term weather events lasting minutes to days to long-term climate events spanning decades or more. Examples include remote teleconnection such as ENSO, IOD, etc, droughts, floods, etc
Weather event	Weather events are meteorological occurrences that impact Earth's atmosphere and surface over short timescales (hours to days). Common Weather Events; Rainfall (e.g., Drizzle, showers, or steady rain), Snowfall (e.g., Light , or heavy); Thunderstorms (e.g., storms with lightning, thunder, heavy rain, and hail), Wind Events (e.g., breezes, gusts, and strong winds), Cloud Cover (e.g., Clear skies, partly cloudy, overcast), Temperature Changes (Heatwaves or cold snaps), Fog and Mist, Frost, Dew etc.

Natural Hazard	1310 Natural hazards are phenomena with the potential to cause significant harm to life, property, and the environment. Teleconnection refers to large-scale patterns of climate variability that link weather and climate phenomena across vast geographic areas, influencing atmospheric conditions over long distances. Typical examples of hazards can be broadly classified into geophysical (e.g., earthquakes, volcanic eruptions, tsunamis, landslides), meteorological (e.g., cyclones or hurricanes or typhons, tornadoes, heatwaves), hydrological (e.g., floods, flash floods, drought, avalanches), biological (pandemics, plagues, animal borne diseases), and climatological (e.g., wildfires, frost, cold wave) categories.
Ocean circulation	Ocean circulation is the large-scale movement of water masses in the Earth's oceans, driven by wind, density differences, and the Coriolis effect, regulating Earth's climate. Key examples of ocean circulation, categorized into surface currents (Gulf Stream, Kuroshio Current, California Current, Canary Current, Equatorial Currents), deep ocean currents (North Atlantic Deep Water (NADW), Antarctic Bottom Water (AABW), Mediterranean Outflow Water, Indian Ocean Overturning), Global Ocean Circulation Systems (the Global Conveyor Belt, the Atlantic Meridional Overturning Circulation (AMOC).
Teleconnection	Teleconnection is a large-scale patterns of climate variability that link weather and climate phenomena across vast distances. Examples includes El Niño-Southern Oscillation (ENSO; (El Niño or La Niña), North Atlantic Oscillation (NAO), Arctic Oscillation (AO), Pacific Decadal Oscillation (PDO), Indian Ocean Dipole (IOD), Madden-Julian Oscillation (MJO), Atlantic Multi-Decadal Oscillation (AMO), Southern Annular Mode (SAM), Rossby Waves, Walker Circulation, Monsoonal Systems (i.e., Asian Monsoon and West African Monsoon)

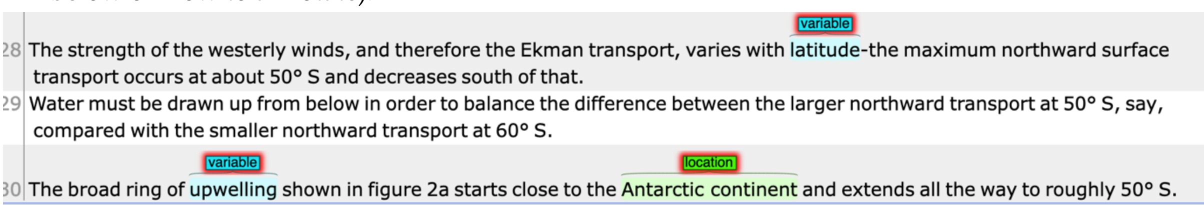
4. Example

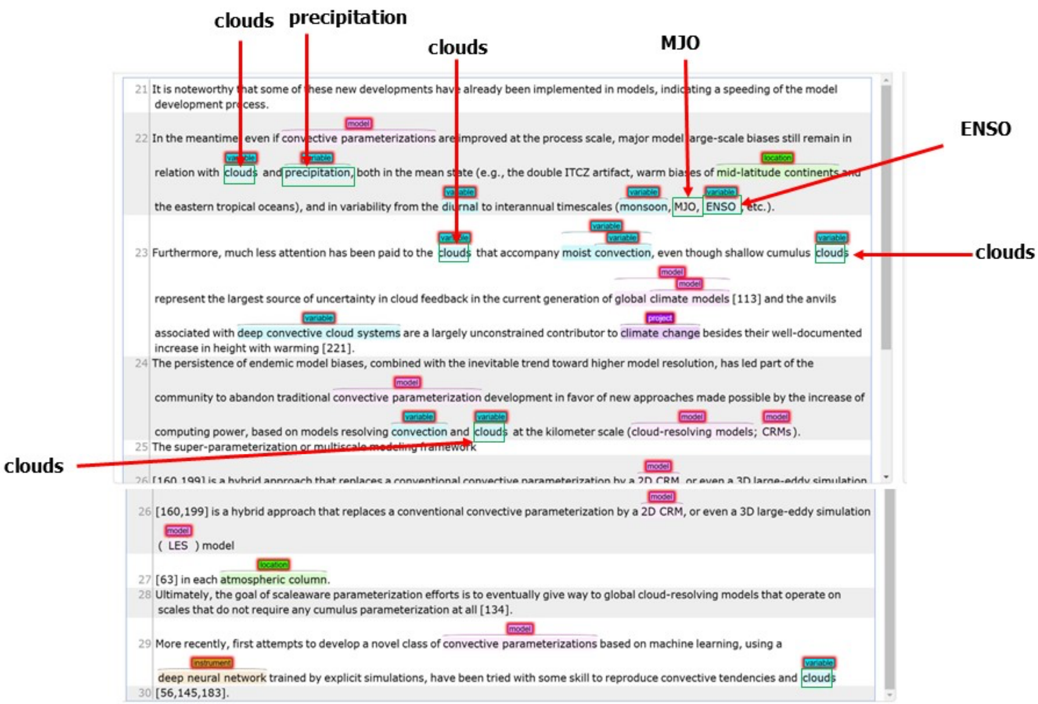
Example: *"This annotation manual aims to provide consistent methods for annotating climate data. Our primary focus is 09bdb7d909ed6615760571a6aa14051133179aee.xmi"*

<u>Task one</u>: see the scientific literature with serial number above.

Role of the annotator: The annotator is expected is to read each sentence carefully. Then, you are required to perform these tasks concurrently.

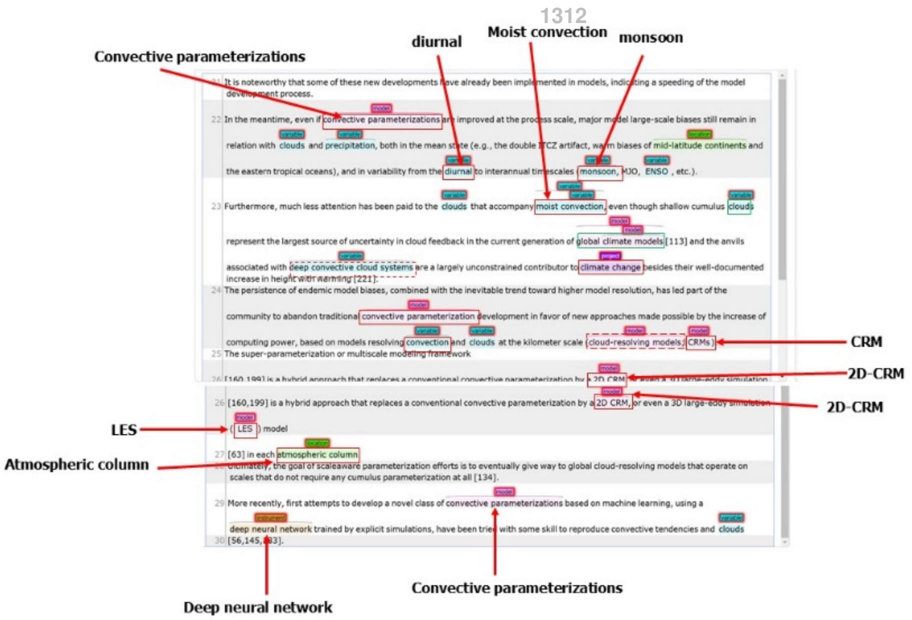
1. Verify specific pre-annotated climate entries of interest in line 22: (E.g., "clouds", "precipitation", "ENSO") and other scientific terms such as "mid-latitude continents". (see details below for more information).
2. Delete pre-annotated test that involves a "process" or "methods", "tools", frameworks, "instrument of measurements", "units of measurement", "temporal, threshold or range of values" (e.g., convective parameterisation, diurnal, monsoon (see details below for more information).
3. Annotate missing but relevant "un-annotated" text of interest (E.g., Westerly Winds) (see details below on how to annotate).

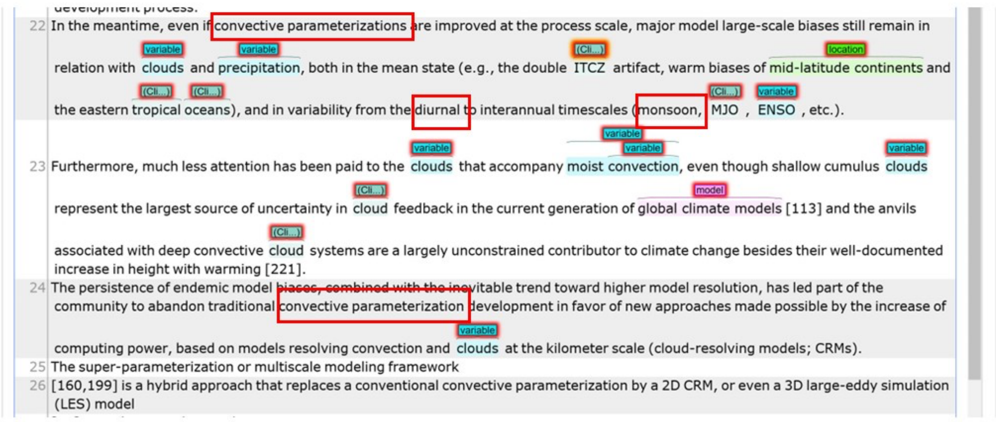




Other Scientific Terms: You may find other climate variables such as temperature, wind speed or wind, sea surface temperature or SST; rainfall, cyclones, aerosols, etc

Delete wrongly pre-annotated climate entities. These may include but not limited to methods, materials, processes, units of measurements, threshold, or range of values, etc
Units of Measurement: (e.g., Celsius for temperature, mm for rainfall, km/h for wind speed).
Thresholds and Ranges: Values or thresholds or ranges. E.g., 10°C for temperature or mm for precipitation."
Standardization: standardizing annotations across climate entities. For example, temperature (delete prefix "minimum or min", "maximum or max", "nighttime", "daytime" for temperature annotations to ensure consistency (e.g. minimum temperature to temperature).
Other Scientific Terms: Phrases that are a scientific term but do not fall into any of the above classes E.g. diurnal, interannual,





STAGE TWO: Entity Linking

1. Tag Selection Guidelines

- **Allowed Tags**: Only the following values should be selected as tags. Do not type any tags manually; only select from the provided list: project, location, model, experiment, platform, instrument, provider, variable, weather event, natural hazard, teleconnection, ocean circulation
- **Spelling and Formatting**:
    - Ensure all tags are in **lowercase**.
    - Do not use uppercase letters or modify the spellings in any way.
    - If you encounter any foreign or unrecognized tags, do not use them.

2. Annotation Setup

- Open **two tables** simultaneously:
    1. **Annotation Table**: The document or interface where you are performing the annotations.
    2. **Knowledge Base Table**: A reference table or database containing entity identifiers and their corresponding information.

1313 - Use the knowledge base to search for and verify the correct identifiers for each entity. Make sure to check if the definitions and the path match the semantic meaning.

3. Task Description

- **Objective**: Link each entity in the text to its corresponding identifier in the knowledge base.
- **Steps**:
    1. Identify the entity in the text.
    2. Double check the tag from the allowed list (e.g., location, variable, etc.).
    3. Search the knowledge base to find the correct identifier for the entity.
    4. Link the entity to its identifier in the annotation table.

4. Quality Assurance

- Double-check the spelling and formatting of tags.
- Ensure that all entities are linked to the correct identifiers in the knowledge base.
- If an entity cannot be found in the knowledge base, flag it for review rather than making an assumption.

STAGE THREE: Relationship

1. Relationship Types and Definitions

Below are the relationship types to be annotated, along with their definitions and examples. Ensure that you correctly identify the **source entity** and **target entity** for each relationship.

1. **ComparedTo**
    - **Definition**: The source entity is compared to the target entity.
    - **Example**: A climate model, experiment, or project (source entity) outputs data (target entity).
    - **Template**: [Source Entity] ComparedTo [Target Entity]
2. **RunBy**
    - **Definition**: Experiments or scenarios (source entity) are run by a climate model (target entity).
    - **Example**: An experiment (source entity) is executed by a climate model (target entity).
    - **Template**: [Source Entity] RunBy [Target Entity]
3. **ProvidedBy**
    - **Definition**: A dataset, instrument, or model (source entity) is created or managed by an organization (target entity).
    - **Example**: A dataset (source entity) is provided by a research organization (target entity).
    - **Template**: [Source Entity] ProvidedBy [Target Entity]
4. **ValidatedBy**
    - **Definition**: The accuracy or reliability of model simulations (source entity) is confirmed by datasets or analyses (target entity).
    - **Example**: A climate model simulation (source entity) is validated by observational data (target entity).
    - **Template**: [Source Entity] ValidatedBy [Target Entity]
5. **UsedIn**
    - **Definition**: An entity, such as a model, simulation tool, experiment, or instrument (source entity), is utilized within a project (target entity).
    - **Example**: A climate model (source entity) is used in a research project (target entity).
    - **Template**: [Source Entity] UsedIn [Target Entity]
6. **MeasuredAt**

- **Definition**: A variable or parameter (source entity) is quantified or recorded at a geographic location (target entity). 1314
- **Example**: Temperature data (source entity) is measured at a specific weather station (target entity).
- **Template**: [Source Entity] MeasuredAt [Target Entity]

7. **MountedOn**
   - **Definition**: An instrument or measurement device (source entity) is physically attached or installed on a platform (target entity).
   - **Example**: A weather sensor (source entity) is mounted on a satellite (target entity).
   - **Template**: [Source Entity] MountedOn [Target Entity]
8. **TargetsLocation**
   - **Definition**: An experiment, project, model, weather event, natural hazard, teleconnection, or ocean circulation (source entity) is designed to study, simulate, or focus on a specific geographic location (target entity).
   - **Example**: A climate model (source entity) targets the Amazon Rainforest (target entity).
   - **Template**: [Source Entity] TargetsLocation [Target Entity]

2. Annotation Instructions

1. **Identify Entities**:
   - Clearly identify the **source entity** and **target entity** in the text.
   - Ensure that both entities are correctly tagged (e.g., model, location, variable, etc.) before annotating the relationship.
2. **Select Relationship Type**:
   - Choose the most appropriate relationship type from the list above based on the context.
   - Refer to the definitions and examples to ensure accuracy.
3. **Annotate the Relationship**:
   - Use the provided templates to annotate the relationship between the source and target entities.
   - Double-check that the relationship type aligns with the context of the text.
4. **Verify Consistency**:
   - Ensure that the relationship annotation is consistent with the definitions and examples provided.
   - If unsure, consult the knowledge base or flag the relationship for review.