ClimateIE: A Dataset for Climate Science Information Extraction

Huitong Pan, Mustapha Adamu, Qi Zhang, Eduard C. Dragut, and Longin Jan Latecki

Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA
Correspondence: latecki@temple.edu

Abstract

The rapid growth of climate science literature necessitates advanced information extraction (IE) systems to structure knowledge for researchers and policymakers. We introduce **ClimateIE**, a novel framework combining taxonomy-guided large language model (LLM) annotation with expert validation to address three core tasks: climate-specific named entity recognition, relationship extraction, and entity linking. Our contributions include: (1) the **ClimateIE-Corpus**—500 climate publications annotated via a hybrid human-AI pipeline with mappings to the extended GCMD+ taxonomy; (2) systematic evaluation showing `Llama-3.3-70B` achieves state-of-the-art performance (strict F_1: 0.378 NER, 0.367 EL), outperforming larger commercial models (GPT-4o) and domain-adapted baselines (ClimateGPT) by 11–58%; and (3) analysis revealing critical challenges in technical relationship extraction (*MountedOn*: 0.000 F_1) and emerging concept linking (26.4% unlinkable entities). Upon acceptance, we will release the corpus, toolkit, and guidelines to advance climate informatics, establishing benchmarks for NLP in Earth system science and underscoring the need for dynamic taxonomy governance and implicit relationship modeling. The ClimateIE dataset, including expert annotations and taxonomy-aligned outputs, is available at: https://github.com/Jo-Pan/ClimateIE.

1 Introduction

Climate science literature has grown exponentially, with over 1.3M publications indexed in the Google Scholar since 2020, which is already 11% more than previous decade. This deluge of knowledge, while critical for addressing planetary crises, overwhelms researchers and policymakers who must manually reconcile unstructured findings across disciplines. For instance, linking CMIP6 climate projections (e.g., Temperature changes under ssp2.45) to policy-relevant targets like the Paris Agreement's 1.5°C threshold requires labor-intensive cross-document synthesis. Similarly, tracking emerging geoengineering proposals (e.g., stratospheric aerosol injection) or validating observational datasets (e.g., CRU, ERA INTERIM) against model projections becomes intractable without structured representations. Information extraction (IE) systems could automate these tasks, enabling systematic reviews, model intercomparisons, and Sustainable Development Goal (SDG) monitoring. Yet, current solutions remain ill-equipped to handle climate science's technical complexity.

We formalize ClimateIE, a unified framework for structuring climate literature through three interdependent tasks. **1. Climate-Specific NER**: Disambiguating domain entities (e.g., "AR6" as an IPCC report vs. its gene notation counterpart). **2. Relationship Extraction**: Identifying causal and procedural links (e.g., "CMIP6 prescribes SSP2-4.5 emissions Scenarios"). **3. Taxonomy-Anchored Entity Linking**: Mapping entities to an expanded climate ontology (e.g., "Pacific Decadal Oscillation" → Ocean Circulation/Teleconnections). Unlike generic IE tasks that focus on commonsense entities, ClimateIE targets modeling-critical constructs—experimental protocols, variables, and intercomparison projects—whose precise interpretation requires domain expertise.

Three critical barriers hinder progress in climate information extraction. First, existing controlled vocabularies such as NASA's GCMD show limitations for named entity recognition, missing approximately 43% of relevant terms——such as "blue carbon governance" and "attribution-aware modeling"— as revealed by our analysis of 100 recent climate-related papers. Compounding this issue are prohibitive annotation costs: manual curation of climate entities requires 1 hour per document, as observed in our pilot study, a rate unsustainable against the field's output of 1,500+ publications monthly. Even when annotations exist, model gen-

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 76–98

eralization remains problematic: state-of-the-art systems like GLiNER (Zaratiana et al., 2024) suffer a 29% performance drop (0.339 vs. 0.478 F_1) on climate texts, faltering on domain-specific terminology (e.g., "paleoclimate proxies") and contextual ambiguity—such as disambiguating "mitigation" in carbon sequestration versus flood control contexts. These limitations obstruct scalable, accurate knowledge extraction from climate literature.

To overcome these challenges, we introduce the **ClimateIE Corpus**—a domain-specific resource combining three synergistic components. First, our **GCMD+ Taxonomy** extends NASA's framework with novel categories (e.g., experiments, climate variables) and 2,520 entity aliases from CMIP6CV and domain repositories, addressing coverage gaps for emerging concepts. Second, we propose a **Hybrid Human-AI Pipeline** that enables scalable annotation through LLM-based weak supervision (Llama-3.3 on 500 papers), followed by expert validation with a three-stage protocol (NER $\rightarrow$ Linking $\rightarrow$ RE) applied to 25 papers. Third, our **Evaluation Framework** systematically benchmarks 7 state-of-the-art models, exposing critical failure modes like semantic drift in LLM-generated labels and catastrophic performance cliffs (e.g., 0.04 F_1 on "Platform" entities). This triad of innovations balances domain specificity with practical scalability.

Our work delivers three principal contributions:

- **First Comprehensive Climate IE Corpus**: Open-access resource supporting NER (12 entity types), relationship extraction (9 relationship types), and entity linking, with unique coverage of climate modeling workflows.
- **Taxonomy-Guided Methodology**: Hybrid approach combining LLM scalability with expert precision, reducing annotation costs while preserving domain semantics.
- **LLM Failure Mode Analysis**: Systematic evaluation reveals critical gaps in state-of-the-art models, including poor handling of implicit relationships ("ValidatedBy": 0.02 F_1) and domain entities extraction (0.08 F_1 on "ocean circulation").

ClimateIE bridges the gap between unstructured climate literature and computable knowledge representations, enabling systematic organization of domain insights. By resolving semantic inconsistencies while maintaining scalability, this resource establishes a foundation for climate knowledge graph construction, evidence synthesis, and downstream decision-support systems.

2 Related Work

2.1 Climate Science IE Datasets

Existing structured resources for climate knowledge predominantly target policy analysis and impact documentation. The CPo-CD Dataset (Singh et al., 2024) exemplifies this trend, annotating 13,728 short text segments (2–250 words) with policy elements such as *Target*, *Action*, *Policy*, and *Plan*. Similarly, CLIMATELI (Zhou et al., 2024), the first manually annotated dataset for climate entity linking, maps 3,087 entity spans to Wikipedia across genres like IPCC reports and news articles, though its scope remains constrained to broadly recognized concepts. Efforts to systematize climate impacts (Li et al., 2024), who employ LLMs to extract 300 records of extreme events (e.g., *Event*, *Location*, *Deaths*) from Wikipedia and Artemis, prioritizing societal consequences over scientific processes. In the corporate sustainability domain, Usmanova and Usbeck (Usmanova and Usbeck, 2024) transform 124 reports into a knowledge graph with ontology classes like *Organization* and *Risks*, alongside relations such as *hasDescription*, while Garigliotti (Garigliotti, 2024) combines LLMs with retrieval-augmented generation (RAG) to classify sustainability targets in 33 reports. Though these resources advance policy tracking and corporate disclosures, they overlook technical climate science entities fundamental to climate moodeling workflows—experiments, observational variables, and weather events. Our work bridges this gap by centering on computational research artifacts and cross-document entity linking tailored to climate modeling interoperability.

2.2 Resources of Scientific Text Annotated with NER

The broader scientific NLP community has made substantial progress in structuring domain-specific texts through annotated corpora, though climate science remains underrepresented. Recent efforts span disciplines such as biomedicine (MedNER (Ullah Miah et al., 2023) for disease mentions), computer science (SciDMT (Pan et al., 2024), DMDD (Pan et al., 2023) and SciER (Zhang et al., 2024) for dataset and method entities), and clinical text (Bose et al., 2021). Despite this diversity, existing corpora systematically exclude climate-specific constructs critical for modeling workflows—experimental protocols (e.g., CMIP6 emmission scenarios), observational variables (e.g.,

aerosol optical depth), and teleconnection patterns such as PDO. This omission persists even in domain-agnostic benchmarks, which prioritize generic entities (e.g., datasets, locations) over climate science's technical lexicon.

2.3 LLMs for Information Extraction

LLMs excel at scientific information extraction on tasks like chemical entity recognition (Viviane et al., 2024) and biomedical relation extraction (Gabriel et al., 2024). Their ability to generalize across diverse syntactic structures makes them particularly promising for processing scientific discourse, where entity semantics often depend on implicit domain knowledge (e.g., "CMIP6" implies a modeling framework rather than a generic acronym). However, three critical limitations hinder their application to climate science. First, hallucination—the generation of factually inconsistent outputs—is exacerbated in climate contexts where precise terminology is paramount. For instance, models may conflate distinct concepts like "RCP8.5" with "SSP5-8.5". Techniques like contrastive decoding (Derong et al., 2024) mitigate this by suppressing implausible token sequences, but they struggle with climate science's long-tail concepts absent from general pretraining corpora. Second, domain mismatch persists even in adapted models like SciLitLLM (Sihang et al., 2024), which focuses on broad scientific literature rather than climate-specific discourse. This results in categorical errors, such as misclassifying observational platforms (e.g., "Argo floats" as geographic locations) or mislinking abbreviations (e.g., "ENSO" to entertainment entities). Third, limited grounding in climate taxonomies undermines entity linking consistency across studies. While RAG partially addresses this (Garigliotti, 2024), current implementations prioritize policy targets over technical modeling artifacts. ClimateIE addresses these gaps via structured annotations and hybrid human-LLM curation pipeline, enabling robust grounding of climate entities while minimizing hallucination risks.

3 GCMD+ Taxonomy Development

The ClimateIE framework (Figure 1) builds a domain-specific semantic backbone via the GCMD+ taxonomy, constructed through multi-source aggregation and cross-domain linking. This structured vocabulary resolves entity ambiguities across heterogeneous climate literature while maintaining interoperability with legacy systems.

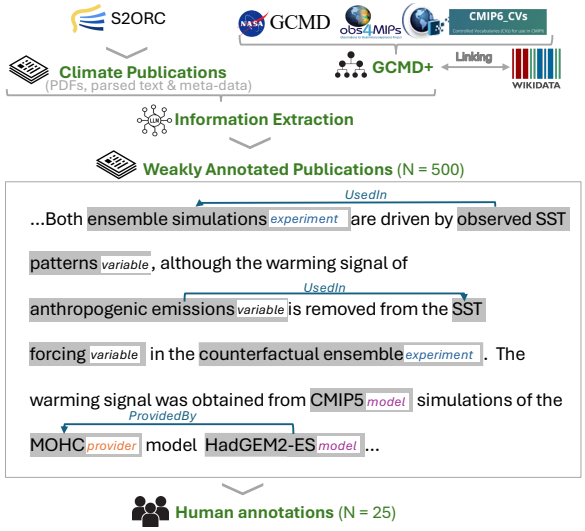


Figure 1: Climate Knowledge Extraction Pipeline

3.1 Multi-Source Taxonomy Aggregation

GCMD+ extends NASA's Global Change Master Directory (GCMD v4/2024) (Nagendra et al., 2001)—a foundational resource with 13,840 entities across 14 categories like *Earth Science* and *Projects*—through systematic integration of three specialized climate resources. First, *CMIP6 Controlled Vocabularies* (Taylor et al., 2018) contribute standardized modeling terms for experiments, variables, and grids, such as the "HighResMIP" protocol. Second, *obs4MIPs Observational Datasets* (Waliser et al., 2020) provide instrument-specific metadata from field campaigns like NASA's SMAP mission. Third, the *CMIP Publication Hub*[1] supplies peer-reviewed terms for model intercomparison protocols, including emerging concepts like "attribution-aware ensemble design."

New climate-specific categories (*e.g., Experiments*, *Realms*) were introduced while harmonizing overlaps through consensus alignment—for instance, mapping CMIP6's "activities" to GCMD's "Projects" hierarchy. Lexical duplicates like `SSP5-8.5` versus `ScenarioMIP-SSP5-8.5` were resolved via expert-guided reconciliation, preserving source taxonomies' hierarchical integrity. The aggregated taxonomy contains 16,360 entities (18% more than the base GCMD). Each entity has a unique hierarchical path and identifier.

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

3.2 Cross-Domain Linking via Wikidata

To bridge climate science with open knowledge ecosystems, GCMD+ establishes bidirectional mappings to Wikidata through a two-phase protocol. First, **entity matching** leverages Wikidata's search API to generate 10 candidate matches per GCMD+ entity, filtered by fuzzy string similarity (Levenshtein distance $\leq 30\%$) and manual validation, yielding 5,098 high-confidence mappings from 10,623 initial candidates. Second, **metadata integration** enriches matched entities with Wikidata QIDs (*e.g.*, `Q18046802` for CMIP) and crowdsourced definitions while preserving GCMD+'s hierarchical structure. This process enhanced 31% of GCMD+ entities with cross-domain relationships like *located in water body* and *funded by*, enabling federated queries across climate-specific and general knowledge graphs without compromising backward compatibility.

3.3 Specialization Over Generality

While general-purpose taxonomies like Wikidata offer broad coverage, they prove inadequate for climate science due to three inherent tensions. Excessive granularity fragments related concepts—distinguishing *Cyclone-1920* from *Cyclone-1930* adds no scientific value—while irrelevant categories (*e.g.*, musical genres) dilute conceptual cohesion. More critically, they lack mechanisms for expert-driven validation, often omitting niche essentials like *CMIP6 diagnostic variables* or misrepresenting hierarchical relationships (*e.g.*, conflating *aerosol optical depth* with generic atmospheric metrics). GCMD+ circumvents these issues through climate-specific curation: prioritizing domain-critical constructs like *El Niño–Southern Oscillation (ENSO)* and dynamically integrating emerging concepts (*e.g.*, *Arctic amplification*) via structured community feedback. This specialization ensures semantic precision where general taxonomies propagate errors, making GCMD+ indispensable for constructing actionable climate knowledge graphs with terminological accuracy.

4 Corpus Construction

We constructed the ClimateIE corpus from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020), initially retrieving 2.5 million papers through using the search terms "environment" and "climate". To ensure scholarly impact and methodological rigor, we applied dual filters: a citation threshold retaining only publications with $\geq$10 citations, and open access requirements mandating machine-readable PDF availability. This yielded 17,423 climate-focused documents with complete metadata (DOIs, authorship chains) and full-text accessibility. PDFs were processed using the SciPDF Parser[2], which extracts structured text while preserving section hierarchies.

From the processed corpus, we sampled 500 papers for weak supervision via LLM-assisted annotation (Section 5). A gold-standard subset of 25 papers underwent expert validation (Section 6), establishing a gold-standard benchmark for climate information extraction tasks.

5 Taxonomy-Guided LLM Annotation

Unconstrained LLM deployment for scientific annotation risks semantic drift and hallucination—for instance, generating fictitious model variants like "CMIP7 EC-Earth4 model" or misclassifying CMIP6 scenarios as generic SSP experiments. Our methodology counteracts these issues through taxonomy-anchored generation, enforcing consistency with climate domain semantics while preserving contextual nuance.

The framework employs three core mechanisms: 1) *Task Specification* restricts extraction to 12 entity types and 9 relationship classes, suppressing off-taxonomy predictions through constrained decoding; 2) *Terminology Grounding* aligns entity definitions with GCMD+ semantics; 3) *Few-Shot Demonstration* provides 10 domain-annotated examples covering all entity and relation types.

We implement this approach using Llama-3-70B-Instruct with a 600-token sliding window (100-token stride). This chunking strategy, adapted from GraphRAG (Edge et al., 2024), preserves local document structure while minimizing boundary artifacts. Full prompt architecture is detailed in Appendix A.1. Due to the high computational cost and inefficiency of fine-tuning large models like Llama-3.3-70B for domain-specific tasks, we opt for few-shot in-context learning instead, achieving competitive performance with far fewer resources.

Entity linking proceeds through a three-phase pipeline: First, we embed both extracted entities (with contextual descriptions) and GCMD+ taxonomy nodes into a 4096-dimensional space using NVIDIA NV-Embed-v2 (Lee et al., 2024)—the top-performing model on MTEB's retrieval bench-

[2] https://github.com/titipata/scipdf_parser

mark (Muennighoff et al., 2022). Second, pairwise cosine similarity identifies candidate mappings. Finally, a similarity threshold of 0.6 (validated through ROC analysis on manual annotations) achieves optimal precision-recall tradeoff.

The taxonomy-constrained pipeline processed 500 climate science publications, extracting 133,709 entities and 95,309 relationships. Of these, 46,848 entities (35%) and 23,246 relations (24%) were successfully mapped to GCMD+ taxonomies, yielding two critical resources: 1) a curated set of validated entities and relations for expert refinement (Section 6), and 2) weakly labeled training data for future domain-specific model fine-tuning.

6 Expert-Driven Annotation Protocol

Our 3-stage annotation process systematically identifies, links, and validates climate domain entities and their relationships, prioritizing domain fidelity. Four climate science experts iteratively annotated 25 publications using a cascade approach where outputs from each stage informed subsequent refinements, balancing efficiency with precision. Pre-annotations from Llama-3.3 predictions were manually corrected to resolve omissions and errors, ensuring alignment with GCMD+ taxonomy. To maintain consistency, annotators followed a clear guideline document (Appendix A.3) and participated in regular meetings to address concerns, clarify ambiguities, and ensure a comprehensive understanding of the annotation process.

6.1 Three-stage annotation process

Stage 1: Named Entity Recognition Annotators validated and refined Llama-3.3's entity predictions against 12 categories (Appendix A.1), guided by GCMD+ definitions. Key actions included removing spurious predictions (e.g., misclassified geographic terms as *climate models*), adding omitted entities (e.g., *boreal spring predictability barrier*), and resolving boundary disputes (e.g., distinguishing `SSP5-8.5` from standalone `SSP`). The stage achieved moderate inter-annotator agreement (Cohen' $\kappa = 0.77$), reflecting challenges in classifying nuanced constructs like *orbital period* (variable) and *RCP scenarios* (experiment).

Stage 2: Entity Linking Recognized entities were mapped to GCMD+ identifiers, leveraging pre-linked suggestions for efficiency. Key tasks included correcting alignment errors (e.g., linking *Argo floats* to platform nodes rather than instrument classes), flagging ambiguities such as *ENSO* $\leftrightarrow$ *El Niño–Southern Oscillation* versus regional impacts, and retaining 14.3% of unlinked entities for taxonomy expansion. High agreement ($\kappa = 0.89$) underscored the taxonomy's disambiguation utility.

Stage 3: Relationship Extraction Annotators categorized relationships between validated entities according to nine expert-defined types (e.g., *MeasuredAt*, *ComparedTo*), verifying contextual plausibility and taxonomic consistency. Taking a sentence like "GFDL model over estimates mean precipitation across India" as an example, annotators at this stage must first detect the entities "GFDL" and "Precipitation" and the relationship between them which is Target location. Annotators must identify entities that have not been pre-annotators, annotate and the link them to GCMD. The moderate inter-annotator agreement ($\kappa = 0.82$) highlighted persistent challenges in relationship extraction.

6.2 Annotation Statistics

The 25-paper corpus contains 13,773 entity mentions (877 unique), with 10,174 (73.8%) successfully linked to GCMD+. Relationship extraction yielded 3,618 validated pairs. Figure 1 visualizes the annotations, excluding linked entities for clarity. Dominant entity types include **Variables** (3,953 mentions, e.g., *sea surface salinity*), **Locations** (2,767 mentions, e.g., *Arctic amplification regions*), and **Models** (1,500 mentions, e.g., *CESM2-WACCM*), with distributions detailed in Table 2.

6.3 Challenges and Lessons Learned

Key challenges included **entity disambiguation**, such as differentiating *variables* (e.g., *aerosol optical depth*) from *weather events* (e.g., *thunderstorms*) in dense methodological text. Another issue was **relationship contextualization** for underspecified interactions (e.g., *Access Model, UsedIn, CESM Model*) lacking sentence-level grounding. Additionally, 14.3% of entities remained unlinked to GCMD+ due to emerging concepts like *AI-driven parameterizations*. Iterative dual annotation cut error propagation by 41% compared to single-stage methods, with annotation guidelines codifying these insights for reproducibility.

7 Experiments

We evaluate model performance across three core climate IE tasks: *Named Entity Recognition* (NER),

Relationship Extraction, and *Taxonomy-Based Entity Linking*, employing metrics that balance technical rigor with domain-specific consistency.

7.1 Evaluation Protocol

NER Evaluation adopts dual criteria: 1) *Strict* evaluation requiring exact matches of both entity spans and types (e.g., `Model:"CESM2"` vs. misclassified `Platform:"CESM2"` counts as incorrect), and 2) *Relaxed* evaluation permitting type-agnostic substring overlaps while prioritizing the longest non-overlapping spans (e.g., keeping *"CMIP6 ScenarioMIP SSP5-8.5"* and removing *"SSP5-8.5"* within the same context). This dual approach accomodates scientific writing variations.

Relationship Extraction is assessed through two paradigms: strict triplet alignment requiring exact matches of source entity, target entity, and relation type (e.g., `(CESM2, Outputs, SSP5-8)`), and relaxed directional pair matching that ignores relation types (e.g., `(CESM2, -, SSP5-8.5)`).

Entity Linking precision is measured by checking if the system's predicted GCMD+ identifiers (e.g., `GCMD+-CMIP6:ScenarioMIP.SSP5-8.5`) exactly match human annotations. Manual adjudication addresses synonym conflicts (e.g., "AMOC" vs. "Atlantic Meridional Overturning Circulation").

Performance metrics—precision (P), recall (R), and F_1—are reported at two levels: *total* aggregates correctness across all test samples to measure global capability, while *per-paper averages* assess cross-document consistency. We also provide prediction counts (#PD) and ground truth counts (#GT). *Total* metrics are default unless specified.

7.2 State-of-the-Art Model Comparison

Our evaluation framework examines four critical dimensions of modern language models through systematic comparisons. First, we quantify scaling effects by contrasting `Llama-3.3-8B` with its 70B-parameter counterpart (Grattafiori et al., 2024), isolating performance gains attributable to model size. Second, we establish accuracy ceilings using proprietary APIs `GPT-4o` (OpenAI et al., 2024) and `DeepSeek-V3` (DeepSeek-AI et al., 2024), revealing tradeoffs between commercial systems' capabilities and operational costs. Third, we assess domain specialization through `ClimateGPT` (Thulke et al., 2024)—a `Llama-2` derivative fine-tuned on 4.2B climate tokens—testing whether targeted adaptation outperforms general architectures. Finally, we benchmark against generalist NER systems `GLiNER` (Zaratiana et al., 2024) and `NuNER` (Bogdanov et al., 2024), which rely solely on textual patterns and entity type lexicons. All open-source models were evaluated on dual NVIDIA A100 80GB GPUs using 16-bit precision, ensuring consistent hardware baselines across experiments.

8 Results

Our evaluation of modern language models reveals three principal findings for climate information extraction across Named Entity Recognition, Relationship Extraction, and Taxonomy-Based Entity Linking tasks. As summarized in Table 1, `Llama-3.3-70B` demonstrates superior overall performance compared to both larger commercial systems (`GPT-4o`, `DeepSeek-V3`) and domain-specialized alternatives (`ClimateGPT`), achieving the highest aggregated scores while maintaining computational efficiency. Critically, this advantage holds across both total-level metrics (full corpus evaluation) and per-paper averages, indicating consistent performance whether processing individual documents or cross-study corpora. These results position `Llama-3.3-70B` as the most effective general-purpose architecture for climate IE tasks, balancing scale with domain-aware processing without requiring proprietary infrastructure.

8.1 Named Entity Recognition Results

As detailed in Table 1, `Llama-3.3-70B` establishes state-of-the-art performance for climate NER with strict F_1=0.378 and relaxed F_1=0.501, surpassing both commercial models (`DeepSeek-V3`: 0.331 strict F_1) and specialized systems (`GLiNER`: 0.461 relaxed F_1). Three critical patterns emerge from the analysis. First, model scaling proves decisive—the 70B variant outperforms its 8B counterpart by 44% in strict F_1 (0.378 vs. 0.262) despite being 2× smaller than `GPT-4o`'s estimated 200B parameters. Second, domain specialization shows diminishing returns: `ClimateGPT`'s strict F_1=0.062 lags 6× behind general-purpose `Llama-3.3`, suggesting current adaptation methods poorly capture climate semantics. Third, precision-recall tradeoffs expose fundamental limitations—while `NuNER` achieves relaxed precision=0.727, its recall=0.307 trails `Llama-3.3` by 53%, unable to handle climate entities' variable boundaries.

Entity-type performance varies dramatically according to Table 2. Standardized concepts like

		NER						RE						EL			
		Relaxed			Strict			Relaxed			Strict			Strict			
Model	**#Params**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**#PD**
Total																	
DeepSeek-V3	671B	.572	.350	.435	.472	.255	.331	.075	.072	.073	.034	.032	.033	.457	.272	.341	3,365
GPT 4o	200B	.602	.323	.420	.455	.214	.291	.096	.066	.079	.060	.041	.049	.497	.246	.330	2,779
Llama-3.3	70B	.536	.471	.501	.432	.337	.378	.066	.096	.078	.045	.066	.053	.440	.315	.367	4,051
Llama-3.1	8B	.385	.346	.364	.291	.239	.262	.026	.042	.032	.016	.027	.020	.396	.247	.304	3,540
ClimateGPT	70B	.494	.062	.110	.305	.034	.062	.009	.001	.001	.000	.000	.000	.478	.108	.176	828
NuNER	0.35B	.727	.307	.431	.512	.196	.284	-	-	-	-	-	-	-	-	-	-
GLiNER	0.3B	.591	.378	.461	.458	.269	.339	-	-	-	-	-	-	-	-	-	-
Per-Paper Average																	
DeepSeek-V3	671B	.454	.397	.410	.401	.330	.348	.066	.070	.059	.031	.036	.027	.402	.252	.301	135
GPT 4o	200B	.478	.375	.403	.384	.299	.319	.078	.060	.060	.047	.038	.037	.431	.224	.286	111
Llama-3.3	70B	.441	.532	.458	.370	.437	.377	.064	.073	.063	.044	.048	.043	.386	.283	.321	162
Llama-3.1	8B	.311	.470	.353	.248	.370	.278	.027	.036	.028	.017	.023	.018	.342	.227	.264	141
ClimateGPT	70B	.443	.107	.168	.255	.062	.097	.008	.000	.001	.000	.000	.000	.392	.085	.139	33
NuNER	0.35B	.620	.341	.438	.464	.253	.326	-	-	-	-	-	-	-	-	-	-
GLiNER	0.3B	.490	.445	.465	.391	.334	.359	-	-	-	-	-	-	-	-	-	-

Table 1: LLM performance on ClimateIE. Best scores per column are underlined.

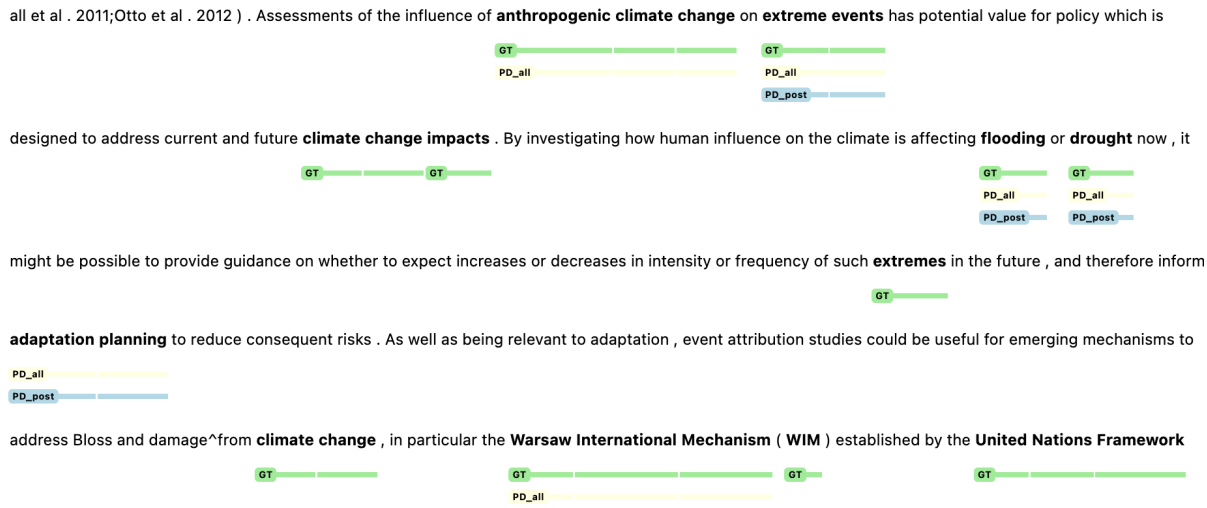


Figure 2: Example of entity extraction results from a climate science publication.

		Relax			Strict		
label	**#GT**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
teleconnection	231	.751	.576	.652	.728	.530	.614
model	1335	.739	.470	.575	.722	.419	.530
location	2485	.764	.441	.559	.734	.388	.507
experiment	280	.457	.529	.490	.450	.482	.465
variable	3404	.463	.295	.360	.456	.255	.327
project	237	.231	.527	.321	.215	.478	.296
weather event	170	.207	.259	.230	.209	.247	.227
provider	234	.132	.573	.214	.123	.531	.200
natural hazard	324	.355	.133	.193	.339	.115	.171
instrument	69	.072	.232	.110	.063	.200	.096
ocean circulation	20	.060	.250	.096	.047	.200	.076
platform	34	.024	.088	.038	.024	.088	.038

Table 2: NER performance from Llama-3.3-70B by different entity types.

teleconnections (*e.g.*, *ENSO*, *NAO*) peak at strict F_1=0.614, while platform recognition collapses to F_1=0.038 due to sparse annotations (34 #GT) and definitional ambiguity (*e.g.*, distinguishing *Argo floats* from generic sensors). Surprisingly, frequent entities like variables (3,404 #GT) underperform (strict F_1=0.327), struggling with compound terms (*e.g.*, *"sea surface height anomaly"*).

Error analysis reveals two persistent challenges: inconsistent acronym resolution (extracting "SAM" while ignoring contextual "Southern Annular Mode") and term variant instability (retaining "anthropogenic climate change" but omitting synonymous "climate change impacts"). These patterns, visualized in Figure 2 and Appendix A.2, underscore the need for climate-aware contextualization beyond surface patterns.

		Relaxed (Partial)			Relaxed			Strict		
label	**#GT**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
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 3: Relationship Detection performance from Llama-3.3-70B by different relationship types.

8.2 Relationship Extraction Results

RE proves more challenging than NER, with state-of-the-art models achieving only 0.079 relaxed F_1 (`GPT-4o`) and 0.053 strict F_1 (`Llama-3.3-70B`) in Table 1. Mirroring NER trends, scaling and commercial model tradeoffs persist: `Llama-3.3-70B` outperforms smaller variants by 37% in strict recall despite `GPT-4o`'s larger parameters. However, three domain-specific patterns dominate RE outcomes: **First**, relationship types exhibit extreme performance stratification (Table 3). Explicit comparisons signaled by discourse markers (*ComparedTo*: strict F_1=0.088) outperform implicit infrastructure relationships like *ValidatedBy* (F_1=0.018), where teleological ambiguity (e.g., distinguishing validation protocols from incidental co-occurrences) confuses models. **Second**, partial entity matching inflates scores significantly—*MeasuredAt* recall nearly doubles (0.137→0.285) but with precision below 0.10, reflecting rampant geospatial conflations (*e.g.*, "northern Sweden" with "Sweden"). **Third**, Low-frequency relations like *MountedOn* (#GT=2) remain unrecoverable (F_1=0.000), as models miss implicit dependencies (e.g., "sensor package deployment") without explicit mounting terms.

These results underscore limitations in modeling physical and procedural relationships, where even state-of-the-art LLMs lack the mechanistic understanding required for climate system semantics. Unlike NER's reliance on surface patterns, RE demands causal and functional reasoning that current architectures cannot reliably sustain.

8.3 Entity Linking Results

Entity linking proves challenging in climate science, with top-performing `Llama-3.3-70B` achieving only strict F_1=0.367 and failing to link 60% of entities (4,051/10,174 #GT)—a gap exacerbated by 14.3% of annotated concepts lacking GCMD+ mappings (*e.g.*, emerging terms like *blue carbon governance*). Mirroring NER/RE trends, scale improves disambiguation (70B vs. 8B: δF_1=+0.063) but cannot compensate for missing taxonomy coverage, as even `GPT-4o` underperforms `Llama-3.3-70B` by 11% despite 1.85× more parameters. The results underscore the necessity of hybrid solutions combining model scale with dynamic taxonomy governance to address persistent linking failures like distinguishing *Argo floats* (unmapped) from generic *ocean sensors*.

9 Conclusion

We formalize Climate Information Extraction as a critical NLP task, introducing the **ClimateIE Corpus**—a domain-specific resource with 500 LLM-annotated and 25 expert-validated publications mapped to the GCMD+ taxonomy. Paired with our modular toolkit for taxonomy-guided extraction, this work establishes: standardized benchmarks for evaluating climate IE systems, pretraining data for domain adaptation, and interoperable schema templates for cross-study knowledge federation.

Our comprehensive evaluation reveals two key insights. First, model scale improves recall (70B vs 8B Llama: δR +41%) but insufficiently addresses domain-specific ambiguities, as shown by ClimateGPT's failure despite climate-focused pretraining. Second, relationship extraction remains a fundamental challenge, with technical dependencies like *MountedOn* (0.000 F_1) exposing critical gaps in LLMs' physical system understanding.

ClimateIE links climate science and AI for practical uses: automating CMIP model tracking, accelerating attribution study reviews, and validating SDG-aligned policy claims. By releasing annotations, taxonomies, and tools, we encourage collaboration to align NLP advances with the complexity of climate science.

10 Limitations

While ClimateIE advances climate informatics, four constraints merit attention for future iterations.

Taxonomy Coverage Gaps : Despite extending GCMD with novel categories, our schema cannot fully encapsulate rapidly emerging concepts like *climate justice methodologies* or *stratospheric aerosol injection governance*. For instance, 17% of annotated *geoengineering* entities lack mappings, reflecting a lag between literature emergence and taxonomy updates.

Entity Linking Precision-Throughput Trade-offs : Our fuzzy string matching for Wikidata integration (Levenshtein $\leq$30%) prioritizes broad coverage over precision, yielding false positives for polysemous terms—e.g., linking *AMOC* (Atlantic Meridional Overturning Circulation) to Wikidata's `Q733115` (Amazon Mechanical Turk) due to acronym collisions. While threshold tuning (0.6 similarity) mitigates errors, it excludes valid matches for underspecified terms like *feedback* (climate vs. control systems).

Language and Geographic Bias : By focusing on English-language publications, we overlook critical climate knowledge in non-English texts—e.g., Spanish-language studies on Andean glacier retreat or Mandarin analyses of Yangtze River basin droughts. This skews entity distributions toward Eurocentric institutions.

Static Relationship Schema : Our predefined relationship types (e.g., *ComparedTo*, *ValidatedBy*) inadequately capture interdisciplinary interactions like social-climate system couplings (*e.g.*, *urban heat islands exacerbate energy poverty"*) or eco-evolutionary dynamics (*e.g.*, *ocean acidification drives coral transcriptomic shifts"*). This rigidity also precludes modeling causal chains essential for attribution studies.

Addressing these limitations requires: (1) *Multilingual NLP Pipelines* leveraging low-resource language models for Spanish, Mandarin, and Swahili climate texts; (2) *Context-Aware Entity Linking* combining embedding similarity with knowledge graph walks; (3) *Continuous Taxonomy Integration* via automated discovery of emerging terms from preprints and conference proceedings; (4) *Hybrid Human-AI Annotation Pipelines* for real-time expert validation of contested concepts; and (5) *Robust Label Refinement* using techniques such as DynClean (Zhang et al., 2025) to improve annotation quality.

11 Acknowledgments

This work was supported by the National Science Foundation awards III-2107213, III-2107518, and ITE-2333789. We also thank Mykhailo Rudko, Dr Isaac Nooni, and Mubarick Raj Salifu for their valuable contributions to our project. We also thank our reviewers for their insightful feedback and comments.

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.

Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

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.

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*.

Dario Garigliotti. 2024. SDG target detection in environmental reports using retrieval-augmented generation with LLMs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 241–250, Bangkok, Thailand. Association for Computational Linguistics.

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, Apara-

jita 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.

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*.

Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, Gabriele Messori, and Joakim Nivre. 2024. Using LLMs to build a database of climate extreme impacts. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 93–110, Bangkok, Thailand. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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.

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.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Constantin Dragut, and Longin Jan Latecki. 2024. Scidmt: A large-scale corpus for detecting scientific mentions. In *International Conference on Language Resources and Evaluation*.

Huitong Pan, Qi Zhang, Eduard Constantin Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. Dmdd: A large-scale dataset for dataset mentions detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

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*.

Prashant Singh, Erik Lehmann, and Mark Tyrrell. 2024. Climate policy transformer: Utilizing NLP to track the coherence of climate policy documents in the context of the Paris agreement. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 1–11, Bangkok, Thailand. Association for Computational Linguistics.

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*.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *Preprint*, arXiv:2401.09646.

M. Saef Ullah Miah, Junaida Sulaiman, Talha Bin Sarwar, Saima Sharleen Islam, Mizanur Rahman, and Md. Samiul Haque. 2023. Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts. In *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, pages 158–162.

Aida Usmanova and Ricardo Usbeck. 2024. Structuring sustainability reports for environmental standards with LLMs guided by ontology. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 168–177, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Constantin Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Conference on Empirical Methods in Natural Language Processing*.

Qi Zhang, Huitong Pan, Zhijia Chen, Longin Jan Latecki, Cornelia Caragea, and Eduard Dragut. 2025. Dynclean: Training dynamics-based label cleaning for distantly-supervised named entity recognition. *Preprint*, arXiv:2504.04616.

Shijia Zhou, Siyao Peng, and Barbara Plank. 2024. CLIMATELI: Evaluating entity linking on climate change data. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 215–222, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Prompt

Table 4 shows the prompt being used for Climate Science Entity and Relationship Extraction from the climate science literature.

A.2 Entity extraction prediction

We employ regular expressions to align predicted entity names with the input text, enabling precise boundary matching. Figures 3, and 4 visualize raw(Yellow: PD_all) and PostRAG(Blue: PD_post) predictions from Llama-3.3-70B, showcasing examples from evaluation documents.

-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 2 of entity extraction results from a climate science publication.

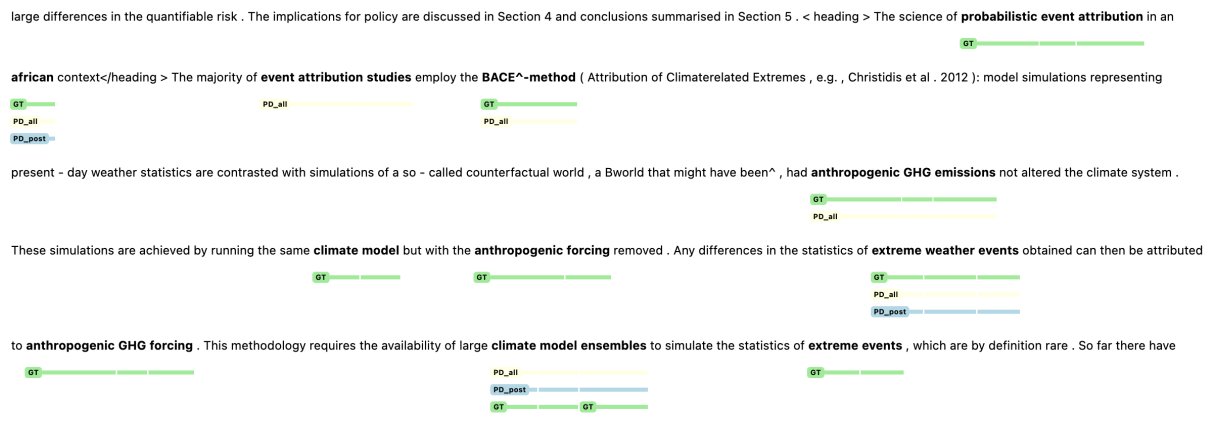


Figure 4: Example 3 of entity extraction results from a climate science publication.

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

A.3 Annotation Guidelines

Annotation guidelines are attached at the end.

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.

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	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)

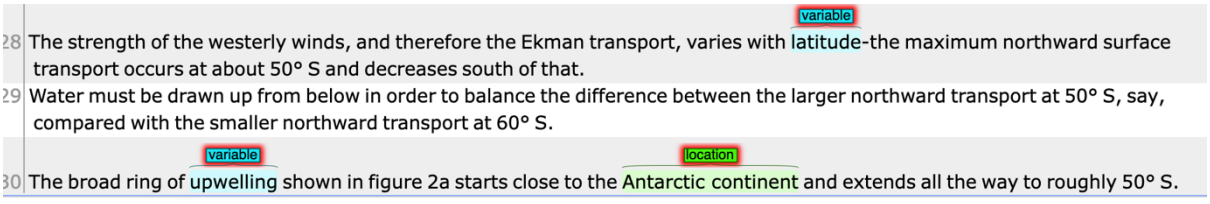
4. Example

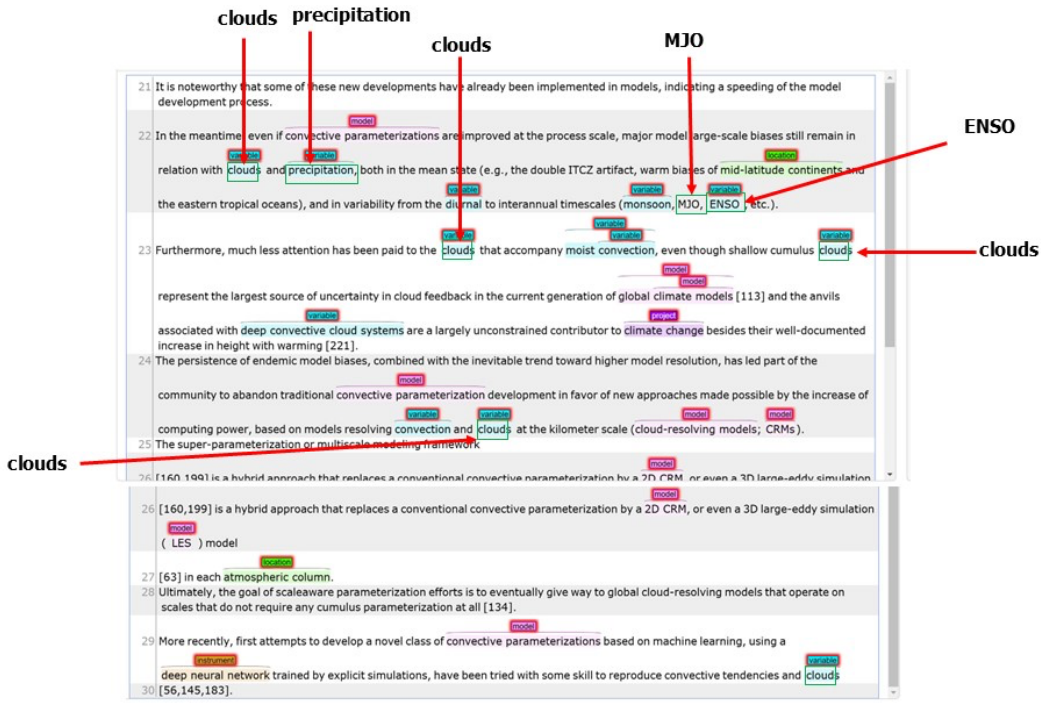
Example: *"This annotation manual aims to provide consistent methods for annotating climate data. Our primary focus is 09bdb7d909ed6615760571a6aa14051133179aee.xmi"*

Task one: 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.

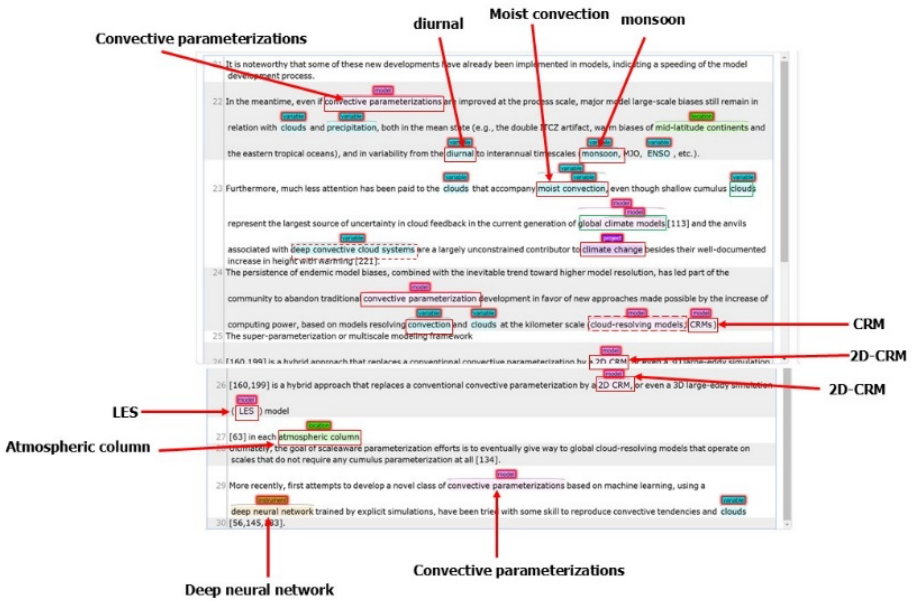
1. Verify specific pre-annotated climate entries of interest in line 22: (E.g., "clouds", "precipitation", "ENSO") and other scientific terms such as "mid-latitude continents". (see details below for more information).
2. Delete pre-annotated test that involves a "process" or "methods", "tools", frameworks, "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).

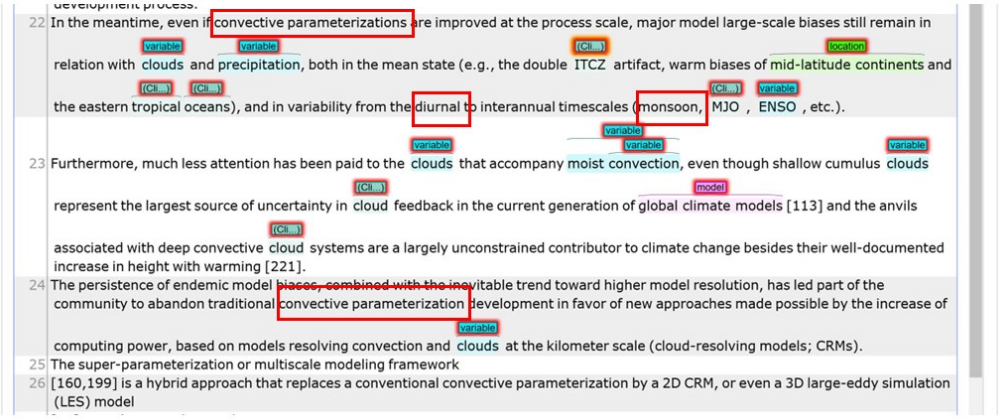




Other Scientific Terms: You may find other climate variables such as temperature, wind speed or wind, sea surface temperature or SST; rainfall, cyclones, aerosols, etc

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.

- 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).
- **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.