Appendix A Models Used

Table A1 lists all models used in the experiments, while supplementary Table A2 lists all encoder models used in the SpanMarker setup experiments, including additional pretraining data information and classification according to the underlying model architecture (RoBERTa, BERT, or DistilRoBERTa).

Model	Year	Version or ID
RoBERTa (Liu et al. 2019)	2019	FacebookAI/roberta-base
INDUS (Bhattacharjee et al. 2024)	2024	nasa-impact/nasa-smd-ibm-v0.1
$\text{INDUS}_{\text{SDE}}$ (Bhattacharjee et al. 2024)	2024	nasa-impact/indus-sde-v0.2
BERT (Devlin et al. 2019)	2019	google-bert/bert-base-uncased
SciBERT (Beltagy et al. 2019)	2019	allenai/scibert_scivocab_uncased
CliSciBERT (Poleksić and Martinčić-Ipšić 2025)	2025	P0L3/cliscibert_scivocab_uncased
CliReBERT (Poleksić and Martinčić-Ipšić 2025)	2025	P0L3/clirebert_clirevocab_uncased
DistilRoBERTa (Sanh et al. 2020)	2020	distilbert/distilroberta-base
EnvironmentalBERT (Schimanski et al. 2024)	2024	ESGBERT/EnvironmentalBERT-base
ClimateBERT (Webersinke et al. 2022)	2022	climatebert/distilroberta-base-climate-f
SciClimateBERT (Poleksić and Martinčić-Ipšić 2025)	2025	P0L3/sciclimatebert
$\text{GLiNER}_{\text{SMALL v2.5}}$ (Zaratiana et al. 2024)	2024	gliner-community/gliner_small-v2.5
$\text{GLiNER}_{\text{MEDIUM v2.5}}$ (Zaratiana et al. 2024)	2024	gliner-community/gliner_medium-v2.5
GPT-5.1 (OpenAI 2025a)	2025a	gpt-5.1-2025-11-13
$\text{GPT-5.2}_{\text{PRO}}$ (OpenAI 2025b)	2025b	gpt-5.2-pro-2025-12-11
$\text{Gemini 2.5}_{\text{PRO}}$ (Comanici et al. 2025)	2025	2.5
$\text{Gemini 3.0}_{\text{PRO}}$ (Google 2025)	2025	3-pro-preview-11-2025
$\text{DeepSeek-V3.2}_{\text{NT}}$ (DeepSeek-AI et al. 2025)	2025	DeepSeek-V3.2
$\text{DeepSeek-V3.2}_{\text{THINK}}$ (DeepSeek-AI et al. 2025)	2025	DeepSeek-V3.2
$\text{Claude 4.5}_{\text{SONNET}}$ (Anthropic 2025b)	2025b	claude-sonnet-4-20250514
$\text{Claude 4.5}_{\text{OPUS}}$ (Anthropic 2025a)	2025a	claude-opus-4-20250514

Table A1: **Models used in experiments:** List of all models used in experiments, including the year of publication and their specific versions or Hugging Face IDs for open-source models.

Appendix B Entity Type Definitions

Traditionally, NER aims to assign predefined entity types, such as location, organization, and person, to text spans. However, such coarse-grained types are often insufficient to capture the domain-specific nuances required in specialised domains such as climate change research (Anonymous 2026). In this work, we start with 21 NER types initially defined in (Anonymous 2026), and expand them to 28 types. Specifically, the initial 21 types defined are: *Astronomical Object, Body of Water, Chemical, Disease, Ecosystem, Energy Source, Field of Study, Geographical Feature, Location, Mathematical Expression, Measurement Unit, Measuring Device, Meteorological Phenomenon, Natural Disaster, Natural Phenomenon, Organization, Physical Phenomenon, Quantity, Satellite, System*, and *Time Period*.

During the initial phase of this research, we iteratively discuss and edit the entity types based on quantitative feedback from the initial annotation. Specifically, across

Model	Year	Data	Domain
RoBERTa			
RoBERTa (Liu et al. 2019)	2019	BookCorpus, Wikipedia, CC-News, OpenWebText, CommonCrawl (Stories)	General
INDUS (Bhattacharjee et al. 2024)	2024	ADS, PubMed Central, PubMed Abstracts, Wikipedia, AMS/AGU, NASA CMR	Scientific / Earth Science / Biomedical / Astronomy / Astrophysics / Physics
$\text{INDUS}_{\text{SDE}}$ (Bhattacharjee et al. 2024)	2024	INDUS + SDE Data	Scientific
BERT			
BERT (Devlin et al. 2019)	2019	BookCorpus, Wikipedia	General
SciBERT (Beltagy et al. 2019)	2019	Biomedical Papers (Full-text), Computer Science Papers (Full-text)	Scientific / Biomedical / Computer Science
CliSciBERT (Poleksić and Martinčić-Ipšić 2025)	2025	SciBERT + Climate Papers (Full-text)	Scientific / Biomedical / Computer Science / Climate Science
CliReBERT (Poleksić and Martinčić-Ipšić 2025)	2025	Climate Papers (Full-text)	Scientific / Climate Science
DistilRoBERTa			
DistilRoBERTa (Sanh et al. 2020)	2020	RoBERTa	General
EnvironmentalBERT (Schimanski et al. 2024)	2024	DistilRoBERTa + Corporate News, Annual, and Sustainability Reports	Environmental / ESG
ClimateBERT (Webersinke et al. 2022)	2022	DistilRoBERTa + Climate News, Climate Papers (Abstracts) and Reports	Scientific / Climate / ESG
SciClimateBERT (Poleksić and Martinčić-Ipšić 2025)	2025	ClimateBERT + Climate Papers (Full-text)	Scientific / Climate Science

Table A2: **Overview of pretrained models:** Summary of all encoder based models used in the (SpanMarker) experiments, grouped by backbone architecture. The table lists the publication year, pretraining data sources, and the target domain.

the arbitrary set of sentences, we examine the frequency of occurrence of each entity type. For types with critically low frequency, including those with zero occurrences, we curated sets of regular expression search terms to identify sentences likely to contain the respective entities. For example, for the *Astronomical Object* type, we search for sentences containing terms such as *planet, moon, star*, and *sun*. If even targeted searches for sentences to annotate do not yield useful results, i.e. enough examples of an entity type with sufficient diversity, we merge or discard the entity type entirely, as is the case for *Astronomical Object*, which we merge with *Location*.

Conversely, we iteratively inspected the *Other* type, used as a default label for entities without a suitable type. We identified clusters of instances, performed manual inspection which introduced new types, including *Body Part, Intellectual Artefact, Person, Organism*, and Asset (borrowed from Bhattacharjee et al. (2024)), inter alia. Table B3 presents the final set of 28 NER types together with their definitions.

Type	Definition
Asset	An **Asset** is an object or service of value to humans that can get destroyed or diminished by climate disasters/hazards. Key categories are health, buildings, infrastructure, and crops or livestock.
Body of Water	A **Body of Water** is a distinct volume or mass of water, whether naturally occurring (like an ocean or river) or contained within a man-made system (like a reservoir or effluent flow). This includes specific named bodies, general types, and scientifically defined water masses.
Body Part	A **Body Part** is a structural component of a living organism (plant, animal, or human), which is not the whole organism but a distinct anatomical or morphological part.
Chemical	A **Chemical** is a substance with a distinct molecular composition, including elements, compounds, ions, biological molecules, and mixtures. It also encompasses classes of substances, reagents, and terms describing their chemical state or form.
Disease	A **Disease** is an abnormal condition, disorder, or pathological state affecting an organism. This includes specific diseases, syndromes, symptoms, signs, injuries, toxicities, and observed pathological changes.
Ecosystem	An **Ecosystem** is a community of interacting organisms and their physical environment, or a term used to describe such a community. This includes specific named ecosystems, general types of habitats, and collective terms for the biological components that define them.
Energy Source	An **Energy Source** is a substance, material, natural phenomenon, or engineered system that provides or stores usable energy. This includes fuels, renewable resources, electricity, heat, and energy storage technologies.
Field of Study	A **Field of Study** is a branch of knowledge, a specific academic or scientific discipline, or a theoretical domain of inquiry.
Geographical Feature	A **Geographical Feature** is a natural or man-made feature of the Earth's solid surface or subsurface, including landforms, topographical structures, and defined regions of land.
Intellectual Artefact	An **Intellectual Artefact** is a man-made product, typically of an informational, intellectual, or representational nature, resulting from scientific research, data collection, or modeling.
Location	A **Location** is a point, area, or region in physical space, defined by natural, administrative, cultural, or geopolitical boundaries or functions. It may occur on Earth or elsewhere in the universe.
Mathematical Expression	A **Mathematical Expression** is a term, symbol, equation, or concept related to mathematics, statistics, or formal modeling. This includes formulas, variables, statistical metrics, mathematical operations, and descriptions of quantitative relationships or trends.
Measuring Device	A **Measuring Device** is an instrument, apparatus, technology, or system used to obtain quantitative measurements of a physical, biological, or chemical property.
Meteorological Phenomenon	A **Meteorological Phenomenon** is a weather- or climate-related event, process, or pattern occurring in the atmosphere and its interaction with the Earth's surface, particularly oceans. This includes specific weather events, large-scale climate oscillations, atmospheric circulation patterns, and long-term trends.
Method	A **Method** is a systematic procedure, technique, model, experiment, or approach used to conduct research, perform an analysis, collect data, or achieve a specific practical or scientific outcome.
Natural Disaster	A **Natural Disaster** (or Hazard) is a large-scale adverse event, arising from natural, environmental, or climatic processes, that has the potential to cause significant harm or disruption to ecosystems, the environment, or human well-being.
Natural Phenomenon	A **Natural Phenomenon** is an observable event or process that occurs in the natural world, driven by biological, physical, or chemical principles. It encompasses interactions, cycles, transformations, and large-scale environmental changes.
Organism	An **Organism** is an individual living being (animal, plant, fungus, microbe) or a term that refers to a species, a taxonomic group, or a collection of individuals.
Organization	An **Organization** is a formally constituted group, such as a company, institution, government agency, laboratory, or association, with a specific, collective purpose.
Other	A **Other** is a miscellaneous category for named entities that do not fit into any of the more specific, defined types. It often includes abstract concepts, generic objects, textual references, and descriptive attributes that have been isolated as entities.
Person	A **Person** is an individual or a group of human beings, identified by name, profession, or through a personal pronoun.
Physical Artefact	A **Physical Artefact** (or Object) is a tangible, material entity with physical form. This includes both man-made products, such as tools, machines, vehicles, or buildings, and natural objects, such as soil, rocks, leaves, plant samples, powders, or other collected specimens.
Physical Phenomenon	A **Physical Phenomenon** is an observable event or property that is not a substance, but rather a manifestation of energy, force, temperature, or state. It encompasses a wide range of processes from molecular interactions to large-scale environmental dynamics.
Policy or Objective	A **Policy or Objective** is a defined goal, aim, or challenge that guides actions, or a framework of principles and rules to achieve those goals. This class encompasses high-level aims, specific targets, strategic plans, and guiding principles for management and governance.
Quantity	A **Quantity** is a measurement, count, property, or value that is quantifiable. This includes numerical values with units, descriptive terms of amount or degree, statistical measures, and abstract quantifiable properties.
Satellite	A **Satellite** is a man-made object, spacecraft, or instrument platform placed into orbit around a celestial body, primarily for purposes like remote sensing, data collection, or communication.
System	A **System** is a set of interacting or interdependent components that form an integrated whole. This includes large-scale natural systems, engineered constructs, and abstract or theoretical frameworks that describe a set of interactions.
Time Period	A **Time Period** is a specific point in time, a duration, or a recurring interval. This includes dates, years, seasons, epochs, and terms describing frequency or temporal extent.

Table B3: **Entity type definitions:** A table with definitions for 28 NE types in CliReNER datasets.

Appendix C Weighted Expert Voting

To consolidate individual annotations into a unified ground truth, we use a Weighted Expert Voting (WEV) scheme as a transparent aggregation heuristic that explicitly encodes domain expertise. Annotators labelling entities within their designated expertise cluster are assigned a slightly higher weight of $w_e = 1.1$, non-expert annotators a slightly lower weight of $w_n = 0.9$, and the non-entity (O) label a baseline weight of $w_{null} = 1.0$. Note that sentences may be repeated in several expert groups; therefore, the original expertise is emphasised. Table C4 lists the six expert groups along with the number of sentences assigned to annotators in each group.

These weights are not intended to estimate annotator reliability or optimise agreement with any reference standard. Rather, they serve as weak priors designed to bias aggregation decisions conservatively in favour of domain expertise while preserving majority influence. Specifically, the relative ordering $w_e > w_{null} > w_n$ reflects the assumption that expert annotations are marginally more informative, while the constraint $2w_n > w_e$ ensures that no single expert vote can override consistent non-expert agreement. To operationalise these weights, we implement a two-stage cumulative voting strategy, designed to resolve boundary conflicts (e.g., B-Time Period vs. I-Time Period) without penalising semantic agreement. First, votes are aggregated at the semantic class level: weights for all BIO tags associated with a specific entity type are summed to determine the winning category. This prevents vote splitting, ensuring that a valid entity type does not lose due to minor disagreements regarding start/end tokens. Once the semantic class is determined, a secondary vote determines the precise BIO tag based on the highest weighted support within that type. In the event of exact weight ties, the decision is resolved deterministically by selecting the less frequent type according to the CliReNER$_{silver}$ distribution[1].

To assess whether this explicit expertise-aware aggregation introduces systematic artefacts, we compare the resulting labels with those produced by MACE (Hovy et al. 2013), a standard probabilistic model for multi-annotator annotation. The two methods show substantial agreement, achieving a Cohen's κ of 0.9076 and a raw agreement of 93.34%. This level of convergence suggests that incorporating domain expertise through a lightweight deterministic rule yields a consensus broadly consistent with

[1]We assume that rarer entity types are less likely to be selected, as identifying them may require greater cognitive effort from annotators. Although this assumption is admittedly imperfect, we consider it the least detrimental for the purposes of aggregation.

Group	Group Focus	Entity Types	#
1	Research methodology and governance	Method, Field of Study, Intellectual Artefact, Asset, and Policy or Objective	59
2	Geospatial and temporal concepts	Location, Geographical Feature, Body of Water, Time Period, and Satellite	60
3	Physical sciences and quantification	Mathematical Expression, Measuring Device, Physical Phenomenon, and Quantity	55
4	Biochemistry, ecology and pathology	Body Part, Chemical, Disease, Organism, and Ecosystem	59
5	Atmospheric science and geophysics	Energy Source, Meteorological Phenomenon, Natural Disaster, and Natural Phenomenon	75
6	General agents and tangible objects	Physical Artefact, Organization, Person, and System	67

Table C4: **Annotator group expertise and entity assignment:** Classification of the six annotator groups (pairs) by thematic focus, listing the total number of sentences each group annotated (#) and the specific entity types assigned to each pair based on their scientific backgrounds.

learned annotator-reliability models. Figure C1 presents per-entity-type Cohen's κ values in a histogram. Most types (24/28) show above-substantial agreement (Cohen's $\kappa > 0.80$) between MACE and WEV labels. Lower agreement is found for *Meteorological Phenomenon* (0.75), *Other* (0.68), *Physical Phenomenon* (0.77), and *System* (0.73). These disagreements likely result from inherent semantic ambiguity; for example, *Meteorological Phenomenon* and *System* may overlap when referring to large-scale atmospheric concepts such as *climate*.

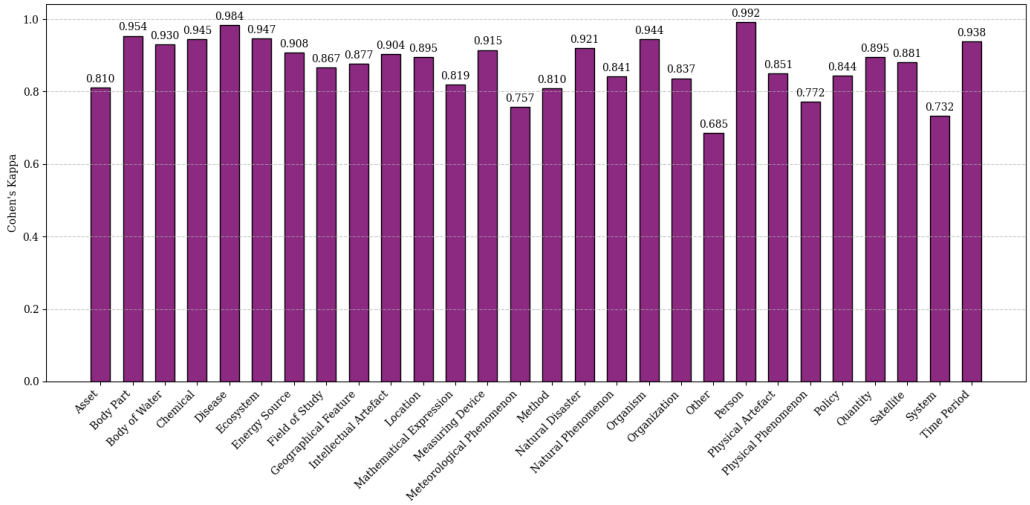

Fig. C1: **Agreement by entity type between MACE and WEV labels**: Distribution of Cohen's Kappa scores (y-axis) evaluated for each named entity type (x-axis). Bar heights indicate the reliability of annotations for specific entity types, adjusted for risk of chance agreement. Exact Kappa values are shown above each corresponding bar.

Given this alignment, we adopt WEV for final dataset construction. Compared to probabilistic aggregation, WEV provides a fully deterministic and reproducible decision process, with consensus labels directly traceable to annotator expertise.

Appendix D Hardware and Environmental Impact

All fine-tuning experiments were conducted on a single NVIDIA GeForce RTX 4070 Ti GPU (CUDA 12.8) and a 13th Gen Intel Core i7-13700F processor. For every run (13 models $\times$ 5 seeds), we monitored average GPU power ($\bar{P}$) and runtime (T) via Weights & Biases[2], deriving energy consumption as $E = \bar{P} \times T$. We report metrics averaged across seeds, where carbon footprints were quantified ($CO_2 = E \times I$) based on the local grid carbon intensity for Croatia ($I \approx 243.30$ gCO_2eq/kWh)[3]. Across the full

[2] https://wandb.ai/site/
[3] https://lowcarbonpower.org/region/Croatia

experimental suite (65 runs), the cumulative execution time was ≈ 10.77 hours. The total energy consumption for these experiments amounted to 1.4060 kWh, resulting in a net environmental footprint of 342.08 gCO_2eq (0.3421 kg).

Appendix E Fine-tuning Parameters

Parameter	Value
learning_rate	5e-5
weight_decay	0.01
model_max_length	512
marker_max_length	256
entity_max_length	14
per_device_train_batch_size	8
gradient_accumulation_steps	2
num_train_epochs	20
warmup_ratio	0.1
save_total_limit	5
logging_steps	50

(a) **SpanMarker configuration**

Parameter	Value
learning_rate	5e-6
others_lr	1e-5
others_weight_decay	0.01
focal_loss_alpha	0.75
focal_loss_gamma	2
lr_scheduler_type	linear
weight_decay	0.01
per_device_train_batch_size	8
gradient_accumulation_steps	2
num_train_epochs	20
warmup_ratio	0.1
save_total_limit	5

(b) **GLiNER configuration**

Table E5: **Hyperparameter configurations:** Detailed training settings for the two models. Subtable (a) details the SpanMarker setup including architecture constraints, while subtable (b) lists GLiNER with focal loss parameters.

Appendix F Entity Type Mapping Results

Figure F2 provides a detailed overview of entity type alignment between the CliReNER schema and the BiodivNER, Climate-Change NER, and ClimateIE schemas. Larger dot sizes and lighter colours indicate higher overlap percentages.

For Climate-Change NER, the alignment shows a shift from functional categories to stricter ontological labelling. Broad thematic types such as *climate-impacts* are divided into more specific inherent types (e.g., *Disease* and *Natural Disaster*), while *climate-observations* is systematically disambiguated into concrete instruments, methods, and platforms (e.g., *Measuring Device*, *Satellite*, and *Intellectual Artefact*). Notable ontological shifts are seen for ClimateIE. The type *project* maps primarily to *Intellectual Artefact* (68.6%), while *experiment* aligns strongly with *Policy or Objective* (82.3%).

More broadly, CliReNER enforces a clear separation between physical entities and conceptual processes (e.g., ClimateIE 's *instrument* is divided into *Measuring Device* and *Method*). Persistent low-frequency mappings to *Other*, particularly in Climate-Change NER, likely reflect residual annotation noise or genuine out-of-ontology edge

cases. To facilitate future domain-specific NER research, all harmonised datasets are publicly released on Hugging Face[4].

[4]Mapped BiodivNER: https://www.anonymised_url4review.org;
Climate-Change NER: https://www.anonymised_url4review.org;
ClimateIE: https://www.anonymised_url4review.org.

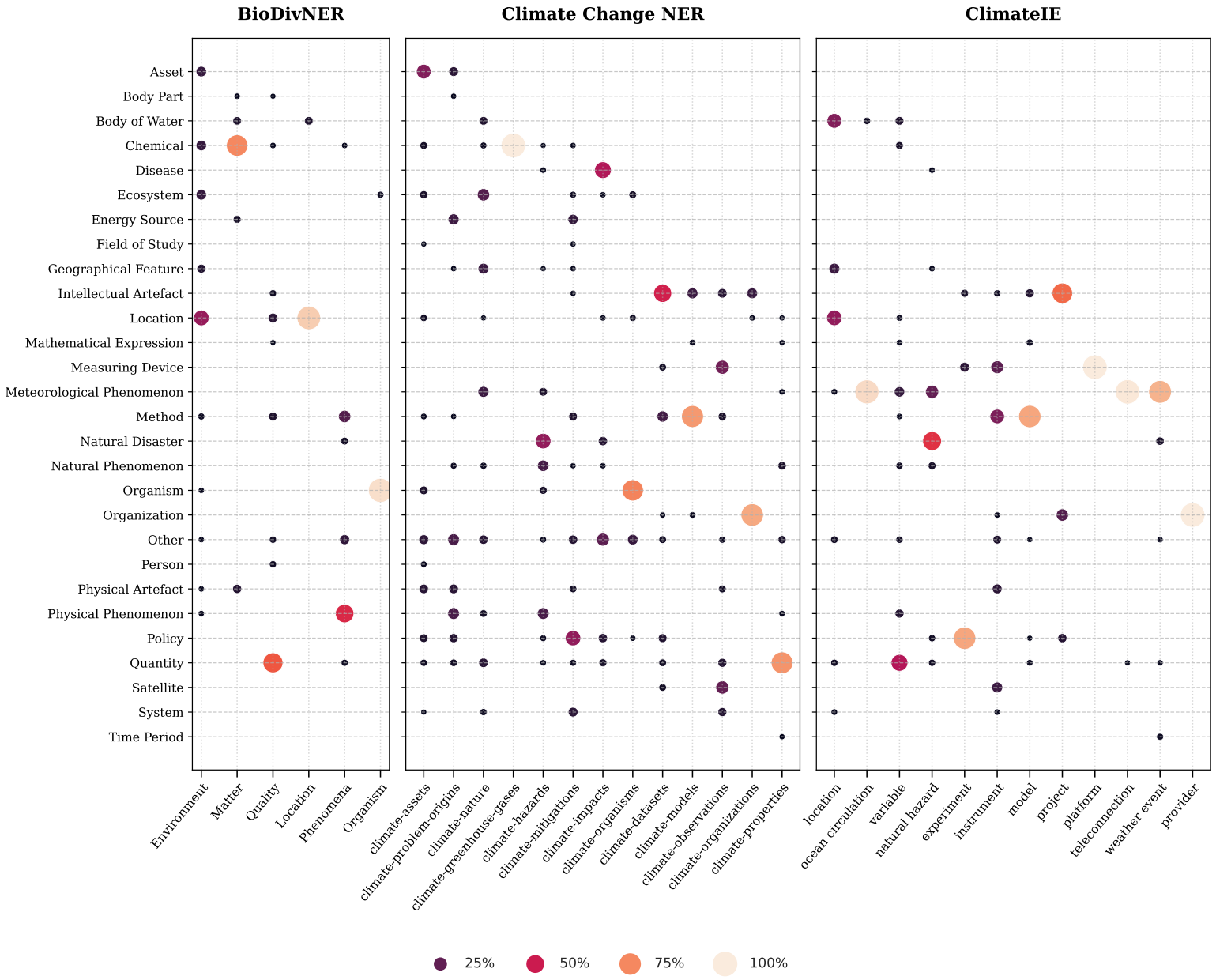

Fig. F2: **Entity alignment and distribution**: Visual mapping of original entity labels (x-axis) from BiodivNER, Climate-Change NER, and ClimateIE to the target CliReNER schema (y-axis). Dot size and color intensity represent the correspondence percentage. Mappings occurring with a frequency below 1% are omitted for clarity.

Appendix G Results Supplement

This section provides supplementary visualizations of model evaluation and domain adaptation effects. Figure G3 illustrates the relationship between lexical diversity, measured by the Unique Entity Ratio (UER), and overall *strict* F1 scores, highlighting performance decay as entity diversity increases on the CliReNER$_{gold}$ dataset.

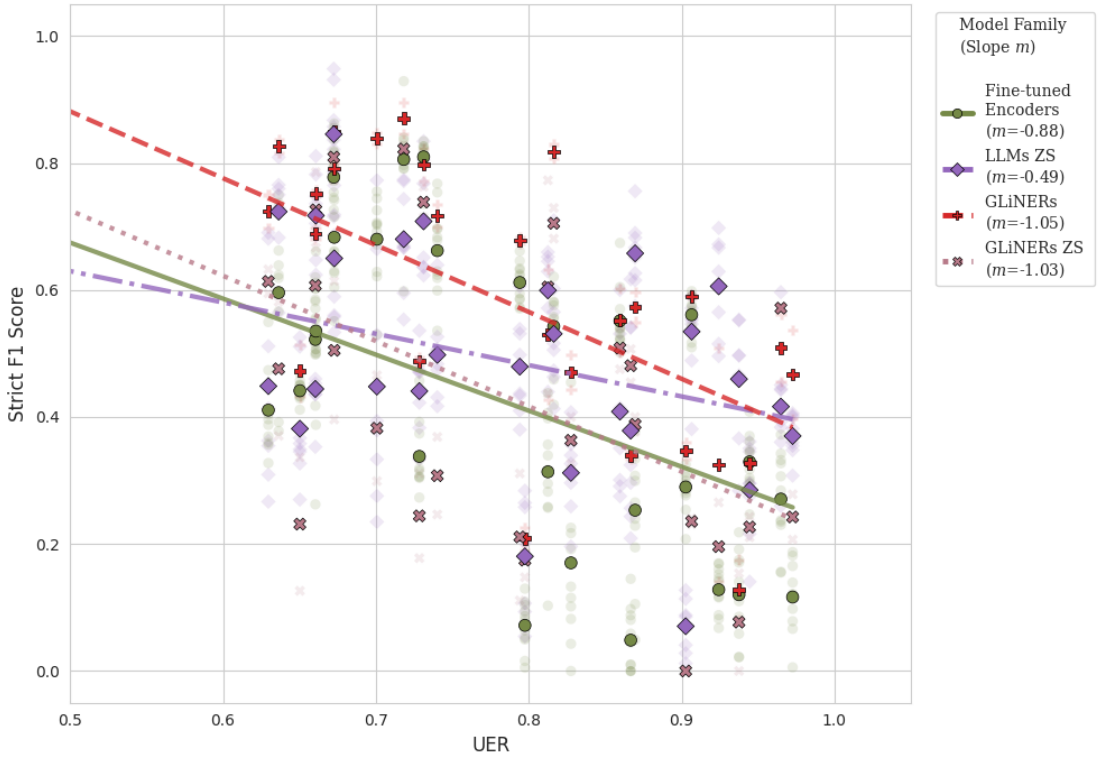

Fig. G3: **Lexical diversity vs. performance:** Impact of Unique Entity Ratio (UER) on model F1 scores. Regression slopes (m) quantify the performance decay as entity diversity increases. Solid markers denote family-level means on the CliReNER$_{gold}$ dataset, while transparent points indicate variance across individual models.

Figures G4 and G5 further present a fine-grained, entity-level analysis of domain adaptation gains, showing changes in *strict* F1 scores between domain-adapted models (INDUS, CliSciBERT) and their respective baselines (RoBERTa, SciBERT), with statistically significant differences explicitly indicated.

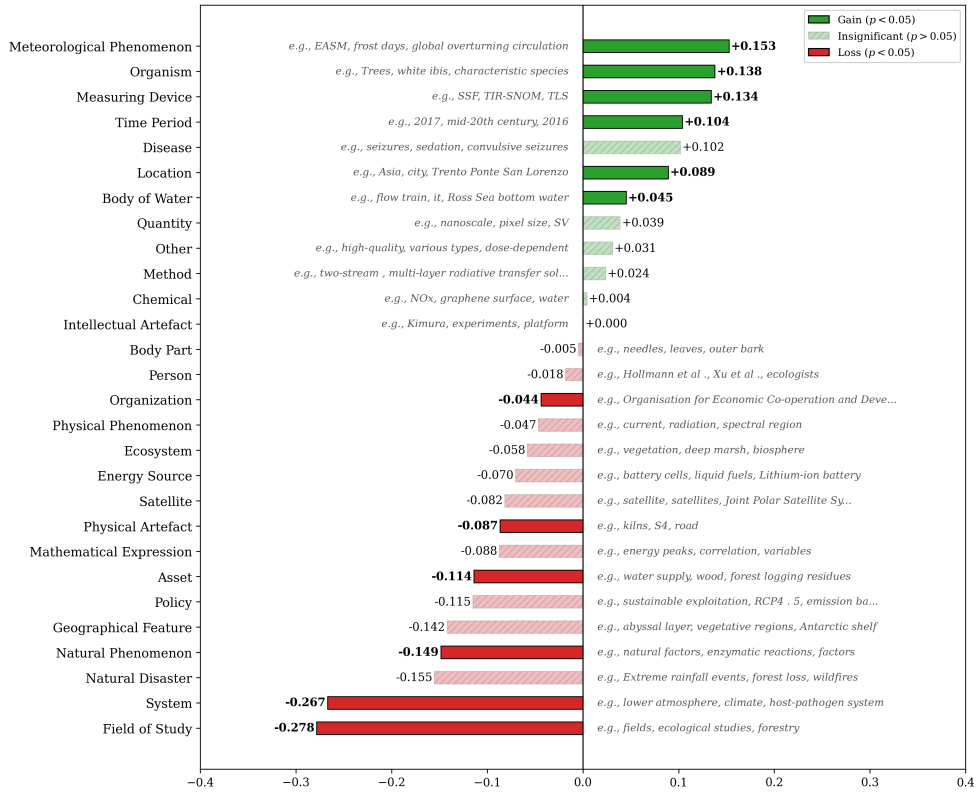

Fig. G4: **Group 1a domain adaptation gains:** Pairwise comparison of INDUS with the RoBERTa baseline. Bar magnitudes represent the change in *strict* F1 score for each entity type. Solid bars indicate statistically significant differences ($p < 0.05$), while hatched bars indicate non-significant differences. Annotations show representative entity examples selected by weighted sampling.

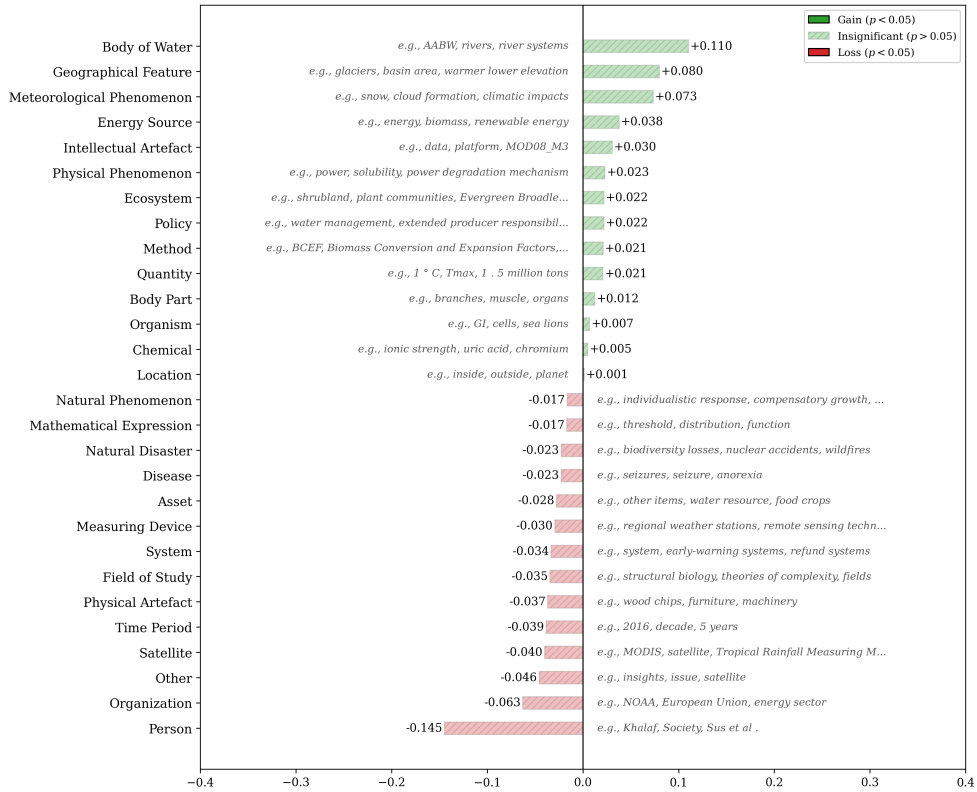

Fig. G5: **Group 1b domain adaptation gains:** Pairwise comparison of CliSciBERT with the SciBERT baseline. Bar magnitudes represent the change in *strict* F1 score for each entity type. Solid bars indicate statistically significant differences ($p < 0.05$), while hatched bars indicate non-significant differences. Annotations show representative entity examples selected by weighted sampling.

Appendix H Limitations

During the development of the entity type definitions, we relied in part on commercial LLMs such as Gemini 3.0_{PRO} and GPT-5.2_{PRO} to refine definition consistency and detailed annotation rules in the supplementary Annotation Manual. Although the annotated data itself was not exposed to these systems, their involvement in shaping the taxonomy may have introduced representational alignment with their internal semantic structures, potentially conferring an advantage in zero-shot evaluation. However, the results in Table **??** provide no consistent evidence of such an effect: while Gemini 3.0_{PRO} and Gemini 2.5_{PRO} rank among the stronger models, GPT variants (GPT-5.2_{PRO} and especially GPT-5.1) perform comparatively poorly. Overall, any potential alignment does not appear to lead to systematic performance gains.

As stated in Section 3, we use a flat NER schema that prohibits overlapping, nested, and discontinuous entities, which inevitably leads to information loss. Discontinuous cases are partially addressed through finer-grained annotation rules (Section 5, rule "5.0.3 Coordinated Modifiers" in the supplementary CliReNER Annotation Guideline), for example, by splitting coordinated expressions such as *seasonal and annual mean precipitation* into separate conceptual entities. Nested entities, however, are not modelled. To estimate the resulting loss, we re-annotated a 25% sample of the CliReNER$_{gold}$ dataset (48 sentences) using nested NER principles. The flat annotation contains 610 entities, whereas the nested version yields 651, indicating a loss of approximately 6.3%. In this sample, nested entities account for 7% of all annotations (46/651), providing a rough estimate of the information lost under the flat schema. This estimate is likely optimistic, as nested structures may be underrepresented in the sampled subset. Examples observed in the sample include *Ecosystem* in ***wetland*** *water*, *Body Part* in *structural* ***brain*** *disease*, *Meteorological Phenomenon* in ***EASM*** *index*, and *Chemical* in ***PB*** *levels*. We further observe that types such as *Method* and *Quantity* frequently participate in nested constructions, which is expected in scientific writing. Handling nested NER more systematically is therefore left for future research.

Appendix I CliReNER$_{silver}$ and CliReNER$_{gold}$ Additional Statistics

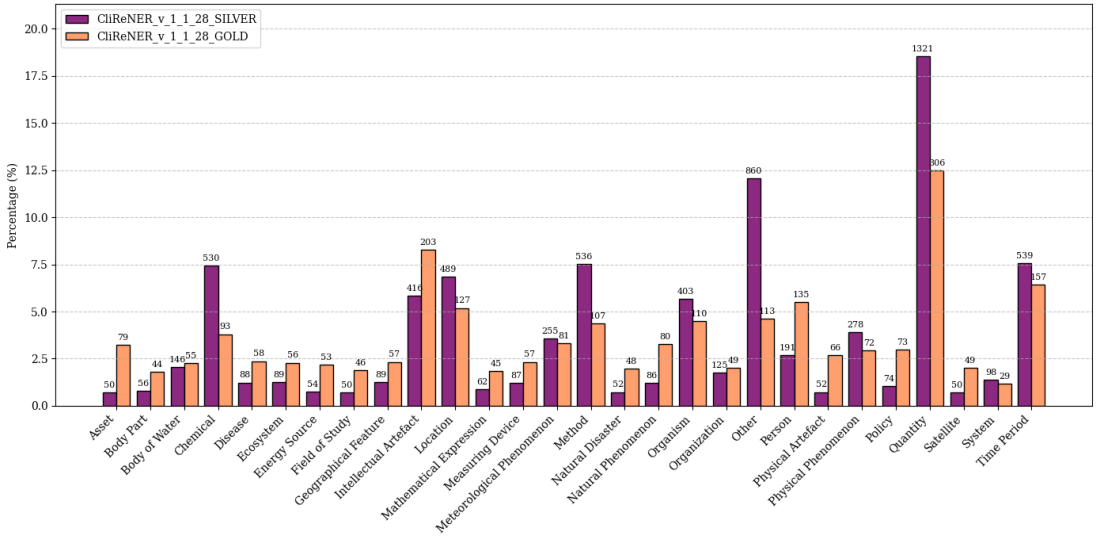

Fig. I6: **Entity type frequency:** Distribution of entity types in the full CliReNER$_{silver}$ and CliReNER$_{gold}$ datasets. The y-axis shows the percentage of entities per type, while raw entity counts are reported above each bar. Please note that for CliReNER$_{silver}$, only 10% of the dataset is used; therefore, the overall distribution might differ.

Entity Type	ΔF1 (*strict*)	ΔF1 (*exact*)	ΔF1 (*type*)	N_{silver}	N_{silver_t}	N_{gold}	UER_{silver}	UER_{gold}	UER Δ	Novelty	Len_{silver}	Len_{gold}	Len Δ
Meteorological Phenomenon	0.2583	0.0896	0.2805	255	28	81	0.4275	0.7284	-0.3009	83.0	1.59	1.69	-0.10
Person	0.2399	0.2171	0.1136	191	21	135	0.6545	0.6296	0.0248	87.1	2.10	3.67	-1.57
Ecosystem	0.2309	0.0936	0.2289	89	6	56	0.4494	0.6607	-0.2113	64.9	1.48	1.71	-0.23
Asset	0.2231	0.0635	0.2155	50	6	79	0.7600	0.7975	-0.0375	85.7	1.84	1.34	0.50
Physical Phenomenon	0.1683	0.0384	0.1548	278	38	72	0.6655	0.9444	-0.2790	91.2	1.88	1.65	0.23
Policy	0.1658	0.0448	0.1914	74	10	73	0.8784	0.9726	-0.0942	94.4	2.30	2.44	-0.14
Organization	0.1337	0.1876	0.0468	125	12	49	0.8240	0.8163	0.0077	70.0	2.57	3.27	-0.70
Intellectual Artefact	0.0868	-0.0553	0.0791	416	48	203	0.6082	0.6502	-0.0421	75.0	1.92	1.67	0.25
Method	0.0849	0.1100	0.0558	536	46	107	0.7388	0.9065	-0.1677	71.1	1.99	1.87	0.12
Other	0.0642	0.1356	0.0743	860	95	113	0.8012	0.9027	-0.1015	74.5	1.69	1.33	0.36
Measuring Device	0.0536	-0.0311	0.1656	87	9	57	0.7586	0.8596	-0.1010	71.4	2.11	2.02	0.09
Disease	0.0514	-0.0262	0.0010	88	7	58	0.6364	0.6724	-0.0361	69.2	1.69	1.64	0.05
Chemical	0.0419	0.0166	0.0316	530	52	93	0.4472	0.7312	-0.2840	52.9	1.41	1.42	-0.01
Satellite	0.0330	-0.0747	0.0738	50	7	49	0.4800	0.4286	0.0514	57.1	1.66	1.82	-0.16
Natural Phenomenon	0.0245	0.1509	0.0241	86	19	80	0.9302	0.9375	-0.0073	94.7	1.94	1.85	0.09
System	0.0238	-0.0134	0.0331	98	11	29	0.7347	0.8276	-0.0929	79.2	2.11	1.72	0.39
Organism	0.0055	-0.0434	0.0606	403	60	110	0.5558	0.7182	-0.1624	57.0	1.78	1.69	0.09
Time Period	0.0024	-0.0102	0.0394	539	68	157	0.5937	0.7006	-0.1069	50.9	1.88	1.54	0.34
Energy Source	0.0022	-0.0279	0.0009	54	4	53	0.5926	0.6604	-0.0678	68.6	2.17	1.66	0.51
Physical Artefact	-0.0122	-0.0855	-0.0132	52	10	66	0.8462	0.9242	-0.0781	90.2	1.83	1.55	0.28
Mathematical Expression	-0.0157	0.1623	-0.0161	62	9	45	0.9032	0.8667	0.0366	87.2	2.15	1.87	0.28
Location	-0.0252	-0.0461	0.0541	489	44	127	0.5010	0.7402	-0.2391	59.6	1.32	1.65	-0.33
Body of Water	-0.0294	-0.0116	0.0501	146	16	55	0.4110	0.6727	-0.2618	62.2	1.51	1.64	-0.13
Body Part	-0.0310	0.1303	-0.0632	56	5	44	0.6250	0.6364	-0.0114	67.9	1.45	1.23	0.22
Quantity	-0.0496	0.0124	-0.0724	1321	148	306	0.6654	0.7941	-0.1287	66.7	2.13	2.11	0.02
Geographical Feature	-0.0989	-0.0498	-0.1193	89	11	57	0.8876	0.9649	-0.0773	89.1	2.19	1.86	0.33
Natural Disaster	-0.1920	-0.4146	-0.1828	52	4	48	0.6731	0.8125	-0.1394	71.8	1.94	1.48	0.46
Field of Study	-0.2441	-0.2950	-0.2485	50	6	46	0.8400	0.8696	-0.0296	77.5	1.58	1.59	-0.01

Table I6: **Entity-level dataset metrics and F1 evaluation differences:** The difference in across-model average F1 scores (ΔF1) between the *silver* and *gold* evaluation sets is shown alongside key dataset statistics. The first column lists the target entity type, sorted by the *strict* evaluation gap. The following columns provide the evaluation support sizes (N), Unique Entity Ratios (UER) measuring lexical diversity, the percentage of novel entities introduced in the *gold* set (Novelty), and the average token span lengths (Len).

Appendix J Zero-Shot Prompt and Entity Instance Examples

Listing 1: **Zero-shot NER LLM prompt:** LLM prompt used for zero-shot NER model evaluation. Entity type definitions are omitted for brevity.

```
### ROLE
You are an expert Named Entity Recognition (NER) system. Your task is to extract entities from the
      user's input text and classify them according to the provided taxonomy.

### TAXONOMY & RULES
Asset - is an object or service of value to humans that can get destroyed or diminished by climate
      disasters/hazards. Key categories are health, buildings, infrastructure, and crops or
      livestock.
...
Time Period - is a specific point in time, a duration, or a recurring interval. This includes dates,
       years, seasons, epochs, and terms describing frequency or temporal extent.


### EXTRACTION PROTOCOL
Follow these steps to generate the output:

1. **Analyze Context:** Read the entire input sentence to understand the semantic meaning.
2. **Identify Candidates:** Scan for noun phrases, measurable properties, processes, and specific
      objects.
3. **Select Head Entities:**
   - Extract the *head* entity (the core noun carrying meaning).
   - Do not extract nested modifiers as separate entities unless they are distinct.
   - Example: In "surface water quality", extract "water quality" (or "quality" depending on
         definition), not just "water".
4. **Classify:** Assign the single best category from the Taxonomy based on the definitions and
      heuristics below.
5. **Resolve Overlaps:**
   - Ensure no two extracted entities share the same text spans.
   - If an overlap occurs, prefer the longer, more specific span usually, unless the Taxonomy rules
         say to prefer the head.
6. **Heuristics for Classification:**
   - **Physical Artefact:** Tangible manufactured objects.
   - **Chemical:** Substances, materials, or chemical compositions.
   - **Quantity:** Measurable properties, numbers, rates, indices, metrics (including units).
   - **Policy or Objective:** Formal plans, targets, frameworks, or barriers/challenges motivating
         action.
   - **Method:** Processes, activities, procedures, or techniques.
   - **Ecosystem:** Biological communities (use **Organism** for specific species).
   - **Location:** Places, regions, geopolitical entities.
   - **Intellectual Artefact:** Datasets, reports, models, results, theories.
   - **Person:** Authors, specific individuals.
   - **Natural/Physical Phenomenon:** Observable natural processes or physical properties (heat,
         radiation).
   - **Ambiguity Resolution:** Use the verb and sentence function. (e.g., "Industry support" -> *
         Method* (the act of supporting); "The Industry" -> *Organization* or *Group*).

### OUTPUT CONSTRAINTS
1. **Exact Match:** The `entity_text` must match the substring in the input text **exactly** (
      preserve case and punctuation) so that it can be located programmatically.
2. **Format:** Output strictly valid JSON.
3. **Schema:**
   ```json
   [
     {
       "entity_text": "extracted string",
       "category": "CategoryName",
       "reasoning": "Brief justification based on context/rule."
     }
   ]
4. If no entities are found, return an empty list [].
```
```

Entity Type	Model 1	Model 1 & Model 2	Model 2
(1)	**RoBERTa**		**INDUS**
Measuring Device	∅	For the purpose, we couple point **gauge** and satellite rainfall estimates a... ...g-to-digital converters (ADCs) and **sensors**.	...time using IASI on MetOp - A/B and **CrIS** on Suomi - NPP and JPSS are used. ...ducation in near - real time using **IASI** on MetOp - A/B and CrIS on Suomi -...
Meteorological Phenomenon	∅	...e, (1) more accurate estimation of **rainfall** at certain scales is now possible ... ...ce in California and the impact of **climate change** for drought frequency and snow ret...	Although **rainfall** magnitude remains unclear, our dat... ...inate in the Tibetan mountains and **snow cover** in Tibet provides substantial wate...
Natural Phenomenon	...information required to assess the **triggering phenomena**. ...en (N) fixation may play a role in **drought tolerance** and drought avoidance by supplying...	∅	∅
(2)	**SciBERT**		**CliSciBERT**
Body of Water	...mportant water resources for these **river systems** and regulates seasonal water suppl...	...e largest changes have occurred in **Ross Sea bottom water** (RSBW) which supplies the abyssal ... ...alysis of SST variability over the **North Pacific** poleward of 20 N. Spectral analyse...	...wing season as the tree's need for **water** increases, resulting in a higher v...
Person	...n 2010b; Wang and Key 2003, 2005b; **Warren and Eastman** 2013) that used single or few sate...	...s, botanists, soil scientists, and **field technicians** contributed to the development of ... ...Boucher, Jason Karl, Jason Taylor, **Matt Bobo**, Dave Yokel, and numerous other ec...	∅
(3)	**DistilRoBERTa**		**SciClimateBERT**
Disease	∅	.../mL, and adverse clinical signs of **sedation** and ataxia were noted at PB levels... The animal was diagnosed with **epilepsy** caused by suspected DA toxicosis b...	... ug/mL. Anorexia, weight loss, and **behavioral changes** appeared to be associated with thi... ... sedation, ataxia, weight loss and **behavioral changes** [35].
Energy Source	∅	... energy sources such as the use of **biomass** [68]. The potential of dedicated **energy crops** as suitable feedstock for producin...	... of many energy storage solutions, **lithium-ion batteries** (LIBs) are attracting more and mor... ...argets for increasing the share of **RES** in the energy mix represents a maj...
Physical Phenomenon	...lationship between temperature and **evapotranspiration** (Anderson, 1936), and increasing V...	... the lidar acquires profiles of **aerosol humidification** every 10 min compared to the 1 - ... ...aterials procedures (ASTM) for the **pyrolysis** of liquid fuels.	... influencing cloud formation and **indirect radiative forcing**.
(4)	**DistilRoBERTa**		**ClimateBERT**
Body Part	∅	...-year-old branches and needles and **wood** and outer bark, and the entire D. ... ...vestris—to a greater extent in the **branches** and needles and to a lesser extent...	∅
Organization	∅	... Institute of Chartered Surveyors (**RICS**, whose role it is to provide ' Pro... ...onautics and Space Administration (**NASA**) Tropical Rainfall Measuring Missi...	∅
Satellite	∅	...purpose, we couple point gauge and **satellite** rainfall estimates at individual g... ...t rainfall estimates obtained from **satellites** [13,14].	...l Landsat frequency by downscaling **MODIS** [19], or the semi-physical fusion ...
(5)	**DistilRoBERTa**		**EnvironmentalBERT**
Ecosystem	...and nuclear accidents on different **ecosystems**. ...e WorldClim was within 1°C outside **forest** whereas Chelsa was within 1°C insi...	...amine effects of climate change on **cloud forests** in northern Peru (drought, interan... ... examine microclimate buffering of **forest** including (1a) compare temperature...	∅
Natural Disaster	...dity either through forest loss or **drought**. ...e lower atmosphere contributing to **wildfires** (Xu et al., 2020).	...to the next possible occurrence of **calamities** caused by abiotic or biotic factor... ...annual diameter growth reaction to **drought events** were based on the following resear...	∅
Organism	...logy (deciduous or evergreen), and **life form** (tree, shrub, or herb).	The first stage describes **immature cells** after cell division (G1 stage), th... ...l stem diameter growth reaction of **trees** depending on climate in urban envi...	... in cardiac and skeletal muscle in **CSL**, creatine kinase (CK) was evaluate...

Table J7: **Model prediction examples:** Correct predictions under the *strict* evaluation strategy are compared between two models. For each model, only predictions that (i) match the CliReNER$_{gold}$ annotation and (ii) are consistently produced by at least four out of five model seeds are considered; otherwise, ∅ is reported. Within the text snippets, the target entities are denoted by **bold underline**. The first column lists the target entity type. The second and fourth columns present examples correctly predicted only by Model 1 and only by Model 2, respectively. The third column contains examples successfully identified by both models. The table presents five comparative cases discussed throughout Section 6.2: (1) RoBERTa vs. INDUS, (2) SciBERT vs. CliSciBERT, and (3, 4, 5) DistilRoBERTa vs. SciClimateBERT, ClimateBERT, and EnvironmentalBERT, respectively.

References

Anonymous A (2026) Title of the article. Journal Name XX(X):1–15. https://doi.org/xx.xxxx/xxxxxx, anonymized for double-blind review

Anthropic (2025a) Introducing Claude Opus 4.5. Blog post, URL https://www.anthropic.com/news/claude-opus-4-5, accessed: January 8, 2026

Anthropic (2025b) Introducing Claude Sonnet 4.5. Blog post, URL https://www.anthropic.com/news/claude-sonnet-4-5, accessed: January 8, 2026

Beltagy I, Lo K, Cohan A (2019) SciBERT: A pretrained language model for scientific text. In: Inui K, Jiang J, Ng V, et al (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 3615–3620, https://doi.org/10.18653/v1/D19-1371, URL https://aclanthology.org/D19-1371/

Bhattacharjee B, Trivedi A, Muraoka M, et al (2024) INDUS: Effective and efficient language models for scientific applications. In: Dernoncourt F, Preoţiuc-Pietro D, Shimorina A (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Miami, Florida, US, pp 98–112, https://doi.org/10.18653/v1/2024.emnlp-industry.9, URL https://aclanthology.org/2024.emnlp-industry.9/

Comanici G, Bieber E, Schaekermann M, et al (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. URL https://arxiv.org/abs/2507.06261, arXiv:2507.06261

DeepSeek-AI, Liu A, Mei A, et al (2025) Deepseek-v3.2: Pushing the frontier of open large language models. URL https://arxiv.org/abs/2512.02556, arXiv:2512.02556

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423/

Google (2025) Gemini 3. Blog post, URL https://blog.google/products-and-platforms/products/gemini/gemini-3/#gemini-3-deep-think, accessed: January 8, 2026

Hovy D, Berg-Kirkpatrick T, Vaswani A, et al (2013) Learning whom to trust with MACE. In: Vanderwende L, Daumé III H, Kirchhoff K (eds) Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, Georgia, pp 1120–1130, URL https://aclanthology.org/N13-1132/

Liu Y, Ott M, Goyal N, et al (2019) Roberta: A robustly optimized bert pretraining approach. URL https://arxiv.org/abs/1907.11692, arXiv:1907.11692

OpenAI (2025a) Introducing GPT-5.1 for developers. Blog post, URL https://openai.com/index/gpt-5-1-for-developers/, accessed: January 8, 2026

OpenAI (2025b) Introducing GPT-5.2. Blog post, URL https://openai.com/index/introducing-gpt-5-2/, accessed: January 8, 2026

Poleksić A, Martinčić-Ipšić S (2025) Pretraining and evaluation of bert models for climate research. Discover Applied Sciences 7(11):1278. https://doi.org/10.1007/s42452-025-07740-5, URL https://doi.org/10.1007/s42452-025-07740-5

Sanh V, Debut L, Chaumond J, et al (2020) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. URL https://arxiv.org/abs/1910.01108, arXiv:1910.01108

Schimanski T, Reding A, Reding N, et al (2024) Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. Finance Research Letters 61:104979. https://doi.org/https://doi.org/10.1016/j.frl.2024.104979, URL https://www.sciencedirect.com/science/article/pii/S1544612324000096

Webersinke N, Kraus M, Bingler J, et al (2022) ClimateBERT: A Pretrained Language Model for Climate-Related Text. In: Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges, https://doi.org/https://doi.org/10.48550/arXiv.2212.13631

Zaratiana U, Tomeh N, Holat P, et al (2024) GLiNER: Generalist model for named entity recognition using bidirectional transformer. In: Duh K, Gomez H, Bethard S (eds) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Mexico City, Mexico, pp 5364–5376, https://doi.org/10.18653/v1/2024.naacl-long.300, URL https://aclanthology.org/2024.naacl-long.300/