Annotating Named Entities in the Climate Change Domain Using Large Language Models: An Experimental Study

Anonymous COLING 2025 submission

Abstract

This paper examines whether few-shot methods for Named Entity Recognition (NER) utilizing existing large language models (LMs) as their backbone can be used to reliably annotate named entities (NEs) in scientific texts on climate change and biodiversity. The objective is to assess whether LMs can be integrated in an end-to-end pipeline that (1) takes a nested Python list as input and (2) generates a Python list with token-level NE annotations as output, thereby reducing efforts for output post-processing. LMs capable of such NE annotation would allow for seamless enrichment of corpora with additional token-level features and the re-use of corpora that have already been tokenized and annotated with other linguistic features. Experiments are run on three LMs, two NER datasets, and ten and nine different prompt types per NER dataset (3831 prompts in total). While the results show that few-shot methods are far from being a silver bullet for NER in highly-specialised domains, improvement in LM performance is observed for some prompt designs. For the time being, few-shot methods would find better use in a human-in-the-loop scenario and tasks involving either augmentation of training data, or exploratory data analysis preceding manual NE annotation.

1 Introduction

Analysing the language of climate change is an important step in following and understanding ongoing developments in this field. One precondition to performing such an analysis is the access to richly annotated corpora with token-level morpho-syntactic and semantic features. Named entities (NEs) belong to the latter category and constitute an important part of linguistic analysis: Glaser et al. (2022) underline that linguistic choices in terms of decisions to explicitly name or leave out a certain entity or concept is an important notion in analysing political speeches. This line of thinking can easily apply to the climate change domain, too.

Widely-used tools for corpus annotation, such as CoreNLP (Manning et al., 2014), stanza (Qi et al., 2020), and spaCy[1], offer out-of-the-box named entity recognition (NER) components with pre-defined sets of NE categories, such as PERSON, ORGANIZATION, or LOCATION, to name a few. An important advantage of these tools is that each token is simultaneously annotated with a set of morpho-syntactic and semantic features. Some noticeable disadvantages include a drop in performance of the NER component when applied to texts in domains not widely represented in the training data (Volkanovska et al., 2023), and a customization option that is data-intensive.

Meanwhile, it has been reported that pre-trained large language models (LMs) perform well on NLP tasks in zero- and few-shot settings in data-poor contexts (Brown et al., 2020). This has attracted the attention of NER and annotation researchers alike. The former have been experimenting with few-shot NER for in-domain and cross-domain applications (Hu et al., 2024; Ashok and Lipton, 2023; Chen et al., 2022; Yang et al., 2022; Epure and Hennequin, 2022). In linguistic data annotation, which has largely relied on resource-intensive coding tasks and human annotators, LMs are believed to be capable of "destabilizing some of the inequalities of academic research" by allowing moderately-funded labs to perform analyses that were previously accessible to well-funded institutes only (Törnberg, 2024, p.17).

When adopting LM-based NE annotation techniques for fine-grained corpus annotation, the LM output needs to be integrated with token-level features obtained from other annotation tools. For example, tokenization techniques applied by domain-specific NER tools that use transformer-based LMs make their integration with tools that perform morpho-syntactic annotation a challenge due to

[1]https://github.com/explosion/spaCy

variations in tokenization approaches.[2] A linguistic annotation tool that has embraced large LMs in its pipeline and is making an effort to overcome this discrepancy is spaCy. By incorporating the work done by Ashok and Lipton (2023) into its new LLM-supporting package, spacy-llm,[3], spaCy allows users to define their own NE categories and integrate an LM-based component in a text annotation task. While extremely valuable, one can only utilize models that are supported by spacy-llm, which makes experimentation with some LMs out of reach.

Motivated by the experiments of Ashok and Lipton (2023), this study examines whether LMs can be plugged in an annotation pipeline as an end-to-end solution for data-poor NER tasks using custom-designed prompts and tokenized corpora. A series of few-shot NER experiments aim to (1) elicit correct annotations from three different LMs and (2) elicit LM output that requires minimum post-processing. Two NER datasets in highly-specialised domains are used to test this approach: Climate-Change-NER (Bhattacharjee et al., 2024) and BiodivNER (Abdelmageed et al., 2022).[4]

2 Related work

Jehangir et al. (2023) distinguish between three types of NER techniques: a rule-based approach, unsupervised learning, and supervised learning. A rule-based approach entails the careful crafting of domain-specific rules to extract and classify patterns representing NEs of interest. Unsupervised learning is used in data-poor contexts, but can yield results that are difficult to evaluate. Supervised learning utilizes manually annotated data to learn representations of relevant NE categories. Corpus annotation libraries, such as CoreNLP, spaCy, and stanza, have incorporated supervised learning in a modular pipeline design, allowing researchers to train their own NER component provided that they have sufficient data.

The advent of Transformer-based LMs has shifted the focus towards transfer learning and fine-tuning, methodologies that demonstrate robust results with fewer manually labelled training examples. In fine-tuning, the architecture of an LM is modified in line with the task requirements: Wang et al. (2022) present a methodology for learning an LM to understand language structure, and then test its performance on downstream tasks including NER.

The increased availability of open-source and paid text-generation and question-answering LMs, alongside prompt engineering techniques for guiding and eliciting LM responses, have fuelled the popularity of zero-shot and few-shot learning approaches. Epure and Hennequin (2022) perform zero-shot and few-shot NER using GPT-2. Prior to prompting the model, they ensure a low level of ambiguity between NE categories by merging possibly confusing NE labels into a single, unambiguous label. They also simplify the task by prompting the model to recognise one NE category at a time. Wang et al. (2023) ensure that the input sentence from which the model is expected to extract NEs is semantically similar to the example sequence in the prompt template by retrieving the k nearest neighbour of the input sequence. They also prompt the model to enclose the NE into special tokens, which should allow for span retrieval. Ashok and Lipton (2023) have presented an intuitive approach to NER, where they propose a prompt template that can easily be customized to any project using custom NE categories and definitions. Their approach has been implemented in spacy-llm's NE annotation pipeline, where users can define NE categories on the fly and annotate their data with an LM of their choice.

This study builds on existing work in the field of few-shot NER and conducts experiments using different prompt templates and a varying number of task examples. It differs from previous methods in (1) the format of the input given to the model and the requested output, and (2) the use of highly-specialised NER datasets, which, to the best of my knowledge, have not been used in a few-shot NER setting previously.

3 Data

Two NER datasets relevant for NE extraction from climate-related scientific texts are used in the experiments: Climate-Change-NER and BiodivNER. This section covers the essential information about each dataset, while Appendices A and B give a comprehensive description of each dataset's con-

[2] A "token" can be a unit at the word- or punctuation-, character-, or sub-word level. Discussing tokenization approaches is beyond the scope of this study; however, it is worth mentioning that LMs using transformer architecture (Vaswani et al., 2017) mostly rely on sub-word units.

[3] https://github.com/explosion/spacy-llm

[4] All prompts and model responses are available here: https://anonymous.4open.science/r/coling2025_submission-27A1/.

tent and an overview of relevant statistical features.

Climate-Change-NER is a publicly-available dataset[5] for NER in scientific texts on climate change, developed in an IBM Research AI[6]-led initiative, involving NASA[7] (Bhattacharjee et al., 2024) among other organisations. The dataset has 13 climate-specific NE classes, which originate from complex taxonomies used in climate-related literature. These are: climate-assets, climate-datasets, climate-greenhouse-gases, climate-hazards, climate-impacts, climate-mitigations, climate-models, climate-nature, climate-observations, climate-organisms, climate-organizations, climate-problem-origins, climate-properties. Seed keywords, such as *wildfire* and *floods*, are used to collect a total of 534 abstracts from the Semantic Scholar Academic Graph (Kinney et al., 2023), which are then manually annotated with the inside-outside-beginning (IOB) tagging scheme, with the help of a set of class-specific dictionaries (Pfitzmann, 2024). Appendix A provides definitions for each NE class, information about the distribution of NE instances per category and per data split (Table 7), descriptive statistical sentence- and token-level information (Tables 8 and 9), and the ten most and least frequent instances in each NE class, per data split (Tables 10 and 11).[8]

BiodivNER is a publicly-available dataset[9] for English-language NER in the biodiversity domain (Abdelmageed et al., 2022). The dataset has 6 biodiversity-related NE classes: organism, phenomena, matter, environment, quality, and location. The annotated corpus comprises of abstracts, tables, and metadata files collected by using a set of keywords from Semedico,[10] BEF-China,[11] and data.world[12] (Abdelmageed et al., 2021) and manually annotated with the IOB tagging scheme. The definitions of each NE class are available in Appendix B, alongside information about the distribution of NE instances per category and per data split (Table 12), descriptive statistical sentence- and token-level information (13 and 14), and the ten most and least frequent instances in each NE class, per data split (15 and 16).

Data preprocessing The NER data is used in two settings: (1) to train a custom NER component in spaCy, and (2) to design prompts for few-shot learning. Use case (1) requires span information about each NE instance, while for use case (2) each sentence needs to be saved as a Python list, with each token index and token saved as sublists (see prompt example in Appendix C). To achieve (1) and guarantee compatibility between each dataset's and spaCy's tokenization, all sentences were re-tokenized and only those that were identical to the tokenized sentences in the original datasets were taken into account. All re-tokenized sentences for Climate-Change-NER were identical; from BiodivNER, 90 re-tokenized sentences from the train file, and 11 from the development and test file each were not identical.

4 Method

This section presents the prompt design, the LMs used in the experiments, the evaluation approach, and the baseline against which the LMs' performance is compared.

4.1 Prompt design

The custom prompt template used in this study differs from the few-shot prompt design suggested by Ashok and Lipton (2023) in five ways: (1) the definition of each NE category is followed by several real-world instances of what the NE category might refer to; (2) a tokenized sentence is passed as input, which is a Python list with sublists of token indices and tokens; (3) only true NE instances are provided as examples of correct answers; (4) the LM is not prompted to emulate reasoning for its decision; (5) the LM is instructed to generate only a nested Python list as the final output. The examples under (3) are sourced exclusively from the train data split, while tokenized sentences that the LM is "tasked" to annotate are sentences from the test data split of each dataset.[13] Each prompt has three sections: (a) definitions-and-instances section, where real-world instances of the NE accompany the definition, (b) questions-answers (QA) section, where each *question* is a tokenized sentence, presented to the model

[5] https://huggingface.co/datasets/ibm/Climate-Change-NER

[6] International Business Machines Corporation

[7] National Aeronautics and Space Administration

[8] NE instances are not lowercased prior to their frequency count, in order to preserve the orthographic features as they appear in the dataset.

[9] https://zenodo.org/records/6575865

[10] A semantic search engine for the life sciences.

[11] https://bef-china.com/

[12] https://data.world/

[13] It was noticed that seven sentences of the train data split were also contained in the test data split in BiodivNER.

as a nested list of [token index + token], and each *answer* is the correct annotation output, which is a Python list containing the NE instance, the NE class, the NE start token, and the NE end token, and (c) task section, where the model is presented with the tokenized sentence in the same format as the question in section (b), asked to annotate the sentence for the named entities defined in section (a), and generate an output in a format identical to the answer format of section (b).

The prompt is not purely a text-as-a-string task, since models are expected to process input and generate output in a specific programming language.

Sections (a), (b) and (c) are universal to each prompt of the prompt types described below; prompt variants are created by applying three different selection criteria for the QA pairs of section (b) and by simplifying the task described in section (a). Rather than converge NE classes, the simplified prompts introduce clusters of NE classes, substantially reducing the NE labels. A description of prompt types and variants is given below, while an overview of the number of prompt variants for Climate-Change-NER and BiodivNER is provided in Tables 1 and 2 respectively.

Prompt version one: random k-examples A k number of random prompt examples is extracted from the train data split, where k can be 3, 4, or 5 example QA pairs, before adding the task sentence in section (c).

Prompt version two: semantically similar k-examples Each sentence of the test split of both datasets is paired with five sentences of the train data split, which have the highest similarity score with the test sentence. Semantic text similarity is calculated with the library sentence-transformers[14] (Reimers and Gurevych, 2019) and the model *sentence-transformers/stsb-distilroberta-base-v2*. The QA section of the prompt is populated with k number of semantically similar QA pairs, where k can be 3, 4, or 5.

Prompt version three: clustered NE classes In order to simplify the task at hand, clusters of NE classes within each dataset are created on the basis of the classes' perceived relatedness. Four NE class clusters are created for Climate-Change-NER and three for BiodivNER. Prompt sections (a) and (b) are populated with four QA pairs pertaining only to the cluster's classes. The NE clusters for Climate-Change-NER are: (1) climate-hazards, climate-problem-origins, climate-greenhouse-gases; (2) climate-impacts, climate-assets, climate-nature, climate-organisms; (3) climate-datasets, climate-models, climate-observations, climate-properties, and (4) climate-mitigations, climate-organisations. For BiodivNER, the three clusters are: (1) environment, location; (2) organism, matter; (3) phenomena, quality. A limitation of 60 tokens was introduced for QA pairs from BiodivNER's training data, due to the observation that the data contained tokenized sentences of over 1000 tokens, which would have substantially inflated the input should they have been included in the prompt.

An example of each prompt version is provided in Appendix C, in Figures 1, 2, and 3.

Prompt version	k=3	k=4	k=5
Random k	177	177	177
Similar k	177	177	177
NE cluster 1	0	177	0
NE cluster 2	0	177	0
NE cluster 3	0	177	0
NE cluster 4	0	177	0
Total prompts	**354**	**1062**	**354**

Table 1: Number of prompts for test sentences of Climate-Change-NER

Prompt version	k=3	k=4	k=5
Random k	229	229	229
Similar k	229	229	229
NE cluster 1	0	229	0
NE cluster 2	0	229	0
NE cluster 3	0	229	0
Total prompts	**458**	**1145**	**458**

Table 2: Number of prompts for test sentences of BiodivNER

4.2 Language models

The choice of LMs was guided by three factors: (1) whether experiments are conducted on consumer-grade equipment or on high-performance computing (HPC) units, (2) previous successful deployment in similar tasks, (3) cost. Two models of OpenAI's GPT family, gpt-4o-2024-05-13 and gpt-4o-mini[15] were run locally via an API. OpenAI's models were chosen over other proprietary models of

[14] https://sbert.net/

[15] https://platform.openai.com/docs/models

similar performance and price range because models of this family have already been successfully deployed in a similar few-shot NER setting (Ashok and Lipton, 2023). The experiment is also run on the public model *mistralai/Mistral-7B-Instruct-v0.3*,[16] which is available for the experiment via an external GPU-supported server and which is loaded in a 16-bit floating point format. The choice of the open-source model is not a claim that the open-source model's performance is comparable to that of the proprietary models, but it is a decision primarily guided resource availability.[17]

4.3 Evaluation

In addition to reporting micro F1 scores in accordance with the standard CoNLL metric (Sang and De Meulder, 2003), strict and simple span-and-category matches achieved by each LM are also reported (Chinchor and Sundheim, 1993). The former refer to a complete match in NE instance, label, start token and end token, while the latter take into account only the NE instance and label. Since the custom prompt instructs the LMs to generate a very specific output (a nested Python list), all instances in which the LMs failed to deliver the requested format are also reported.

4.4 Baseline

The performance of the three models on the BiodivNER dataset is compared against the results of BiodivBERT (Abdelmageed et al., 2023), an LM pre-trained and fine-tuned specifically for an NER task in the biodiversity domain, with a reported F1 score of 0.87. For Climate-Change-NER, the baseline is that of the model INDUSBASE (Bhattacharjee et al., 2024), an LM pre-trained and fine-tuned on relevant scientific data, with a reported F1 score of 0.64. In addition to this, two custom NER components were trained using spaCy and the model en_core_web_lg [18] as a base model. SpaCy's NER tagger is a transition-based classifier which uses convolutional neural networks (CNNs); it achieves an F1 score of 0.73 on BiodivNER's test data, and 0.43 on the Climate-Change-NER test data.

[16] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[17] A more comparable open-source LLM would have been *meta-llama/Meta-Llama-3.1-405B*; however, using an LLM of this scale hinges upon access to HPC units of a grade not available for this study.

[18] https://spacy.io/models/en_core_web_lg

5 Results and analysis

Tables 3 and 4 summarize the micro F1 scores for the test datasets of Climate-Change-NER and BiodivNER involving the prompts described in Section 4.1. Due to cost and infrastructure constraints, one iteration was performed on each test set. In the tables, *k* stands for the number of QA pairs included as task examples in the prompt template. Prior to calculating the results, each model's output was cleaned from misspelled or non-existing categories (e.g. ORGANSIM instead of ORGANISM); such annotations were given an "O" label.

Tables 5 and 6 present the strict span overlap, where a model correctly annotates the NE instance, the NE class, and the start- and end-token index, the simple span overlap, where the model correctly annotates the NE instance and the NE class, and the number of sentences for which the model failed to return output in the proper format (column *Err.*), for each type of prompt and the *k* variation per prompt.

Results are presented in bold in two scenarios: one among the simplified prompts with fewer NE classes in the definition and task sections and 4 QA example pairs, and one in the prompts that include all NE classes and a varying number and type of QA example pairs. A detailed report on the performance of each LM on each prompt version tested in this paper is available in Appendix D for Climate-Change-NER and Appendix E for BiodivNER.

Baseline comparison Even the best-performing prompt & model combination substantially lags behind the baseline NER models for the datasets, more so in the case of BiodivNER, where the baseline F1 score is 0.87 and the spaCy classifier F1 score is 0.73. This gap, however, is smaller for Climate-Change-NER, where the baseline score is 0.64, and the best-performing prompt & model combination achieve a better F1 score than spaCy's 0.43.

General analysis Looking at the F1 scores alone, gpt-4o-2024-05-13 has the best performance overall among the three LMs used in the experiment, followed by gpt-4o-mini and mistral-7B-instruct-v0.3. For both datasets, the best-performing LM also seems to respond favourably to being presented with sentences that bear semantic similarity to the sentence in the prompt's *task* section; this improvement is more noticeable in the Biodi-

Prompt type	k	gpt-4o-mini	gpt-4o-2024-05-13	Mistral-7B-Instruct-v0.3
NE class cluster 1	4	**0.3923**	**0.5277**	0.0701
NE class cluster 2	4	0.2298	0.3287	**0.0728**
NE class cluster 3	4	0.3339	0.4444	0.0589
NE class cluster 4	4	0.1744	0.3824	0.0360
Random k examples	3	0.3148	0.3889	0.1117
Random k examples	4	0.3250	0.4158	0.0965
Random k examples	5	**0.3614**	0.4416	**0.1143**
Similar k examples	3	0.3262	0.3996	0.1002
Similar k examples	4	0.3603	0.4245	0.0941
Similar k examples	5	0.3591	**0.4579**	0.1019

Table 3: F1 scores for each prompt type and each LM tested on **Climate-Change-NER**. The NE class combinations of the first four rows are available in 4.1. Prompts in the last six rows include all classes.

Prompt type	k	gpt-4o-mini	gpt-4o-2024-05-13	Mistral-7B-Instruct-v0.3
NE class cluster 1	4	0.2087	0.3429	0.0570
NE class cluster 2	4	**0.3051**	**0.4253**	**0.0931**
NE class cluster 3	4	0.2585	0.3595	0.0696
Random k examples	3	0.2650	0.3212	0.1453
Random k examples	4	0.2856	0.3277	0.1322
Random k examples	5	0.2864	0.3671	0.1452
Similar k examples	3	0.3543	0.4128	0.1782
Similar k examples	4	0.3673	**0.4836**	0.2049
Similar k examples	5	**0.3761**	0.4000	**0.2218**

Table 4: F1 scores for each prompt type and each LM tested on **BiodivNER**. For the NE class clusters of the first four rows see 4.1. Prompts in the last six rows include all NE classes.

vNER dataset relative to the Climate-Change-NER dataset. In general, the models perform better when the prompt includes more examples (*k* is either 4 or 5); the only exception to this is gpt-4o-mini for the similar-k-examples group of prompts in Climate-Change-NER and gpt-4o-2024-05-13 for the same group of prompts in BiodivNER.

Clustering NE classes into somewhat related groups so as to reduce the task complexity has a positive outcome for the models of the GPT family for the first NE class cluster of the Climate-Change-NER dataset, which also achieves the highest overall F1 score. For other NE clusters, however, the task simplification approach did not yield any improvements.

Looking at the strict span overlap, gpt-4o-2024-05-13 has the best performance for Climate-Change-NER; again, NE class cluster 1 yields the best overall result, while prompts of the similar-k-examples group where k=5 yield the second-best result. In terms of simple overlap, gpt-4o-2024-05-13 performs slightly better than gpt-4o-mini, and mistral-7B-instruct-v0.3 catches up by an impressive rate to the two GPT models.

Regarding the correctness of the format requested in the prompt, gpt-4o-mini outperforms the two other models, which have over 100 outputs each in an incorrect format for both datasets. One reason for this could be that newer generation models are exposed to more coding tasks in the training data relative to “older” models.

Error analysis To obtain a better understanding of how the examined LMs annotate named entities, qualitative error analysis is performed on the output of the highest-F1 score models for prompt types that encompass all NE classes. For Climate-Change-NER, this is the model gpt-4o-2024-05-13 with a prompt containing 5 similar QA examples, while for BiodivNER this is the same model with a prompt containing 4 similar QA examples.

Prompt type	k	gpt-4o-mini			gpt-4o-2024-05-13			Mistral-7B-Instruct-v0.3		
		Strict overlap	Simple overlap	Err.	Strict overlap	Simple overlap	Err.	Strict overlap	Simple overlap	Err.
NE cluster 1	4	**51/85 (60 %)**	**52/85 (61.18 %)**	0	**61/85 (71.76 %)**	**64/85 (75.29 %)**	10	**18 / 85 (21.18 %)**	**45 / 85 (52.94 %)**	1
NE cluster 2	4	63/175 (36 %)	66/175 (37.71 %)	2	81/176 (46.02 %)	82/176 (46.59 %)	2	32 / 172 (18.6 %)	66 / 172 (38.37 %)	1
NE cluster 3	4	96/226 (42.48 %)	104/226 (46.02 %)	0	95/214 (44.39 %)	96/214 (44.86 %)	20	29 / 219 (13.24 %)	89 / 219 (40.64 %)	1
NE cluster 4	4	15/68 (22.06 %)	23/68 (33.82 %)	1	37/68 (54.41 %)	38/68 (55.88 %)	7	8 of 68 (11.76 %)	28 / 68 (41.18 %)	0
Random k examples	3	177/555 (31.89 %)	193/555 (34.77 %)	0	195/516 (37.79 %)	199/516 (38.56 %)	10	78 / 544 (14.34 %)	192 / 544 (35.3 %)	4
Random k examples	4	196/555 (35.32 %)	215/555 (38.74 %)	0	219/536 (40.86 %)	220/536 (41.04 %)	11	70 / 539 (13 %)	189 / 539 (35.1 %)	3
Random k examples	5	**221/555 (39.82 %)**	**228/555 (41.08 %)**	0	247/548 (45.07 %)	247/548 (45.07 %)	2	85 / 535 (15.89 %)	188 / 535 (35.14 %)	5
Similar k examples	3	182/551 (33.03 %)	198/551 (35.68 %)	2	193/495 (38.99 %)	196/495 (39.6 %)	26	63 / 365 (17.26 %)	128 / 365 (35.1 %)	46
Similar k examples	4	202/555 (36.4 %)	219/555 (39.6 %)	0	205/489 (41.92 %)	209/489 (42.74 %)	22	54 / 337 (16.02 %)	120 / 337 (35.61 %)	49
Similar k examples	5	204/552 (36.95 %)	215/552 (38.95 %)	1	**215/467 (46.04 %)**	**219/467 (46.9 %)**	34	**59 / 341 (17.3 %)**	**128 / 341 (37.54 %)**	53
Total err.	n/a	n/a	n/a	6	n/a	n/a	144	n/a	n/a	163

Table 5: Overview of strict and simple span overlap achieved by each LM, and the number of outputs in an inadequate format, for **Climate-Change-NER**.

Prompt type	k	gpt-4o-mini			gpt-4o-2024-05-13			Mistral-7B-Instruct-v0.3		
		Strict overlap	Simple overlap	Err.	Strict overlap	Simple overlap	Err.	Strict overlap	Simple overlap	Err.
NE cluster 1	4	39 / 186 (20.97 %)	46 / 186 (24.73 %)	0	70 / 180 (38.89 %)	77 / 180 (42.78 %)	32	27 / 161 (16.77 %)	55 / 161 (34.16 %)	2
NE cluster 2	4	**174 / 340 (51.18 %)**	**191 / 340 (56.18 %)**	1	**287 / 569 (50.44 %)**	**329 / 569 (57.82 %)**	10	56 / 259 (21.62 %)	104 / 259 (40.15 %)	5
NE cluster 3	4	130 / 282 (46.11 %)	151 / 282 (53.54 %)	7	212 / 477 (44.44 %)	238 / 477 (49.9 %)	19	44 / 244 (18.03 %)	104 / 244 (42.62 %)	14
Random k examples	3	295 / 959 (30.76 %)	335 / 959 (34.93 %)	1	394 / 1263 (31.20 %)	436 / 1263 (34.52 %)	7	**142 / 612 (23.02 %)**	**264 / 612 (43.14 %)**	15
Random k examples	4	314 / 956 (32.85 %)	354 / 956 (37.03 %)	3	400 / 1257 (31.37 %)	465 / 1257 (37 %)	9	123 / 669 (18.39 %)	257 / 669 (38.42 %)	16
Random k examples	5	297 / 920 (32.28 %)	330 / 920 (35.87 %)	3	370 / 908 (40.75 %)	412 / 908 (45.37 %)	20	123 / 616 (19.97 %)	247 / 614 (40.23 %)	15
Similar k examples	3	390 / 959 (40.67 %)	434 / 959 (45.25 %)	1	470 / 1220 (38.52 %)	500 / 1220 (41 %)	16	**150 / 637 (23.55 %)**	300 / 637 (47.1 %)	14
Similar k examples	4	395 / 959 (41.19 %)	439 / 959 (45.78 %)	1	**570 / 1153 (49.44 %)**	**626 / 1153 (54.3 %)**	18	158 / 610 (26 %)	293 / 610 (48.03 %)	12
Similar k examples	5	**407 / 925 (44 %)**	**460 / 925 (49.73 %)**	3	413 / 1099 (37.58 %)	441 / 1099 (40.13 %)	30	174 / 615 (23.3 %)	**308 / 615 (50.1 %)**	14
Total err.	n/a	n/a	n/a	20	n/a	n/a	161	n/a	n/a	107

Table 6: Overview of strict and simple span overlap achieved by each LM, and the number of outputs in an inadequate format, for **BiodivNER**.

Climate-Change-NER Annotations in the two worst-performing NE classes were analysed: CLIMATE-ASSETS (F1 of 0.2368) and CLIMATE-OBSERVATIONS (F1 of 0.2857).[19] CLIMATE-ASSETS is defined as "objects or services of value to humans that can get destroyed or diminished by climate-hazards"; in the test dataset, some annotated examples in this category include *building*, *crop*, and *water availability*, to name a few. When annotating instances of this class, the model tends to prefer the longest-span option: it annotates the phrase *public health sector* with the label CLIMATE-ASSETS, while in the gold standard only the token *health* is annotated with the same label. The phrase *maize yield* is treated in a similar fashion. A particular difficulty seem to be tokens that can be both stand-alone NEs and belong to an NE phrase. One such case is *agricultural*, which in the gold data is mostly annotated as a stand-alone NE of the class CLIMATE-ASSETS, except in one case, when it is part of the phrase *agricultural productivity*. CLIMATE-OBSERVATIONS is defined as "climate observation tools with a name"; some examples from the gold data include *TerraSAR - X* and *lidar*. For this NE class, the model is capable of guessing the correct label, but struggles to extract the complete span. For example, the NE *Analytical Spectral Device (ASD) Field Spec Pro* is parsed in three different NEs: *Analytical Spectral Device*, *ASD*, and *Field Spec Pro*.

BiodivNER The two lowest-scoring classes in this instance are ENVIRONMENT (F1 of 0.3933) and PHENOMENA (F1 of 0.4).[20] ENVIRONMENT is defined as "natural and man-made environments organisms live in"; some examples from the gold data include *ecosystem*, *microbal commu-*

[19] A complete classification report for each category is available in Table 21.

[20] A complete classification report for each category is available in Table 28.

nities, and *arrable land*. Once again, the model prefers to extract the longest span: *forest land* and *grazing land* instead of just *land*, *open field* instead of just *field*. Interestingly, the model annotates *cropland* as environment, when it appears in the same context as *forest land* and *grazing land*; in the gold dataset, *cropland* is not annotated at all. The NE class PHENOMENA is defined as "occurring natural, biological, physical or chemical processes". The task seems to be challenging for the model, which misses instances such as *decomposition* and *pollination*. Some of the errors are due to the model extracting the NE in all-lowercase rather than preserving the original spelling, or preferring the longest span over single tokens.

6 Discussion and future work

The experiments revealed that few-shot NER methods are not a turnkey solution for highly-specialised NE annotation and should not be treated as such in real-world deployments. Unsurprisingly, newer LMs of the GPT family are better at handling input and output in unconventional formats. While LMs cannot be integrated in an end-to-end pipeline for data annotation in the context explored in this paper, they could be valuable assets in testing the definitions and labels of an existing NER dataset, as well as in pre-processing a dataset intended for an NER task before manual annotation. The first suggestion is corroborated by the fact that in BiodivNER, NE candidates that could be deemed valid NE instances, but were not annotated as such, were annotated by the LMs. This experimental setup would be an affordable way of probing NE definitions and categories prior to embarking on manual annotation. Future research in few-shot NER could consider several directions, such as experimenting with domain-specific open-source LMs, such as those of the ClimateGPT family (Thulke et al., 2024), experimenting with automatic (self-)verification strategies to improve the output of the LM, including more extensive annotation guidelines in the task description, using JSON format to structure a prompt's input and output, and calculating inter-annotator agreement between models. In the error analysis, hallucinations in the form of the model suggesting NE classes not included in the prompt were considered "O" annotations. An extended analysis could investigate these categories and the degree of additional insight they might offer. It would also be interesting to see how the models perform on evaluation metrics that allow for softer boundary rules or provide more error categories (Chinchor and Sundheim, 1993).

In an annotation scenario, it is paramount to adopt a human-in-the-loop approach, especially at a time when LMs' training data and processes are not consistently reported. In addition, at the current cost, processing a large corpus with an LLM would have limited financial justification over using an LLM to create training data for a local processing pipeline. Finally, reporting limitations of LMs' capabilities should help inform decisions towards a more responsible use of this technology, especially in high-stakes domains such as climate change and biodiversity.

For the time being, large LMs should be treated as NLP tools that could help augment NER datasets for building NER components, which would ultimately be used to annotate data that will be analysed by human experts. Nevertheless, research in LLMs is a rapidly advancing field, and it would be interesting to see how OpenAI's recently-published models o1 and o1-mini handle this task.

7 Ethical considerations

This study uses publicly available datasets. The GPT-based experiments require no special infrastructure and can be reproduced with an OpenAI API and the prompts provided in the GitHub repository.[21] The costs per dataset are: Climate-Change-NER - USD 0.6 for gpt-4o-mini and USD 18.79 for gpt-4o-2024-05-13; BiodivNER - USD 0.66 for gpt-4o-mini and USD 21.51 for gpt-4o-2024-05-13. Sending sensitive data to proprietary APIs is not recommended. Mistral-7B-Instruct-v0.3 requires GPU infrastructure.

Limitations

Some of the experiments use text generation in an LM-as-a-service setup, which makes them vulnerable to non-responsive APIs. Given that an LM may not yield the same result twice even when prompted with the same text, it is impossible to guarantee 100% reproducibility. Guardrails against bias and offensive content are recommended before real-world deployment. Another limitation is that the annotation pipeline for cannot be saved locally. An open-source LM might offer a bit more stability, but this could come at high infrastructure costs.

[21] https://anonymous.4open.science/r/coling2025_updated_repo-DB73

References

Nora Abdelmageed, Alsayed Algergawy, Sheeba Samuel, and Birgitta König-Ries. 2021. Biodivonto: towards a core ontology for biodiversity. In *European Semantic Web Conference*, pages 3–8. Springer.

Nora Abdelmageed, Felicitas Löffler, Leila Feddoul, Alsayed Algergawy, Sheeba Samuel, Jitendra Gaikwad, Anahita Kazem, and Birgitta König-Ries. 2022. Biodivnere: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10.

Nora Abdelmageed, Felicitas Löffler, and Birgitta König-Ries. 2023. Biodivbert: a pre-trained language model for the biodiversity domain. *CEUR-WS. org*, pages 62–71.

Dhananjay Ashok and Zachary Chase Lipton. 2023. Promptner: Prompting for named entity recognition. *ArXiv*, abs/2305.15444.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, et al. 2024. Indus: Effective and efficient language models for scientific applications. *arXiv preprint arXiv:2405.10725*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2022. Prompt-based metric learning for few-shot ner. *arXiv preprint arXiv:2211.04337*.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Elena V. Epure and Romain Hennequin. 2022. Probing pre-trained auto-regressive language models for named entity typing and recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1408–1417, Marseille, France. European Language Resources Association.

Luis Glaser, Ronny Patz, and Manfred Stede. 2022. Unsc-ne: A named entity extension to the un security council debates corpus. *Journal for Language Technology and Computational Linguistics*, 35(2):51–67.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Birgit Pfitzmann. 2024. Personal correspondence. Personal correspondence with Birgit Pfitzmann on 2 September 2024.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Volkanovska, Sherry Tan, Changxu Duan, Sabine Bartsch, and Wolfgang Stille. 2023. The insightsnet climate change corpus (iccc). BTW.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. *arXiv preprint arXiv:2208.11464*.

Appendix A: Additional information about Climate-Change-NER

Appendix A provides the definitions of the 13 named entities contained in the dataset Climate-Change-NER. The definitions have been obtained from the dataset card available on Hugging Face.[22] In addition to the NE definitions that have been used to build the prompts, Table 7 gives an overview of the frequency of NE instances per NE class and per data split. Tables 8 and 9 give basic statistical information about the average, median, maximum, and minimum length of sentences and tokens in each data split and in the whole dataset, alongside the total number of sentences and tokens per data split and in the complete dataset. Finally, Tables 10 and 11 list the ten most frequent and least frequent real-world instances of the 13 NE categories respectively. This data is presented per data split. For weakly represented NE classes, such as the class *CLIMATE-GREENHOUSE-GASES* in the development data split, the number of NE instances in the two tables is lower than ten, due to the fact that the total number of instances is less than ten.[23] The NE frequencies presented in this Appendix are raw counts.

Definitions of NE classes

- **CLIMATE-HAZARDS**: hazards with potential negative impact on climate, such as floods, wildfires, droughts, and heatwaves. Where a hazard is named in more detail in a text, the entire term is annotated, e.g., surface water flood or soil liquefaction;
- **CLIMATE-MITIGATIONS**: activities to reduce climate change or to better deal with the consequences;
- **CLIMATE-PROPERTIES**: properties of the climate itself (not abstract objects like models and datasets) that typically come with values and units;
- **CLIMATE-NATURE**: aspects of nature that are not alive, such as oceans, rivers, the atmosphere, winds, and snow;
- **CLIMATE-MODELS**: specific physical, mathematical, or artificial intelligence objects, nowadays always computer-executable, used to analyze and usually predict climate parameters;
- **CLIMATE-PROBLEM-ORIGINS**: problems that describe why the climate is changing. Key examples are fossil fuel and deforestation. We also mention sectors that can be cited as causes of energy use. For instance, in a text about the energy consumption by the transport sector, transport sector is annotated as problem;
- **CLIMATE-OBSERVATIONS**: climate observation tools with a name. Examples are satellites, radiospectrometers, rain gauges, wildlife cameras, and questionnaires;
- **CLIMATE-ASSETS**: objects or services of value to humans that can get destroyed or diminished by climate-hazards. Key categories are health, buildings, infrastructure, and crops or livestock;
- **CLIMATE-IMPACTS**: effects of hazards, primarily negative effects on humans. We also consider impacts on livestock as impacts, as it indirectly affects humans;
- **CLIMATE-GREENHOUSE-GASES**: gases that cause heating of the atmosphere (greenhouse gases);
- **CLIMATE-ORGANIZATIONS**: real-world organizations with climate-related interests;
- **CLIMATE-ORGANISMS**: animals, plants, and other organisms that are considered for their own sakes (in contrast to as food for humans) as climate organisms;
- **CLIMATE-DATASETS**: specific collections of climate data with a name. A climate dataset can be the result of observations or of a model, e.g., as a prediction or reanalysis. The data may be lists, tables, databases, inventories or historical records, where the data dominate over attached code.

Named eintity class	Train	Dev	Test	Total per category
CLIMATE-HAZARDS	320	50	34	404
CLIMATE-MITIGATIONS	185	30	38	253
CLIMATE-PROPERTIES	455	107	86	648
CLIMATE-NATURE	705	195	98	998
CLIMATE-MODELS	325	78	94	497
CLIMATE-PROBLEM-ORIGINS	129	19	20	168
CLIMATE-OBSERVATIONS	105	4	21	130
CLIMATE-ASSETS	248	31	50	329
CLIMATE-IMPACTS	63	16	17	96
CLIMATE-GREENHOUSE-GASES	25	2	31	58
CLIMATE-ORGANIZATIONS	112	35	30	177
CLIMATE-ORGANISMS	203	17	11	231
CLIMATE-DATASETS	154	28	25	207
Total per data split	3029	612	555	4196

Table 7: Named entity instances per category and per data split in Climate-Change-NER

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	985	32	29	115	2
development	191	33.04	30	86	1
test	177	32.63	31	97	10
total	1353	32.23	30	115	2

Table 8: Climate-Change-NER sentence features

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	31516	4.91	4	21	1
development	6311	4.89	4	18	1
test	5775	4.91	4	21	1
total	43602	4.9	4	21	1

Table 9: Climate-Change-NER token features

Most frequent NE instances per category in Climate-Change-NER			
NE Class	**TRAIN**	**DEVELOPMENT**	**TEST**
CLIMATE-HAZARDS	flood, fire, drought, fires, sea level rise, floods, landslides, pollution, storm surge, extinction	drought, pollution, forest fire, flood, water scarcity, biomass burning, forest fires, climate extremes, Ixodes scapularis, tick-borne pathogens	flood, fire, SLR, Fires, drought, soil moisture depletion, fires, hurricanes, earthquakes, storm surges
			Continued on next page

[22]https://huggingface.co/datasets/ibm/Climate-Change-NER

[23]The NE instance *CO2* is the only NE instance of this category in the development dataset, with a total count of 2.

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-MITIGATIONS	irrigation, renewable energy, mitigation, climate policy, electric vehicles, carbon tax, FMNR, nuclear power, SAT, Dam	eco-environmental management, insulation, Wind energy, mitigation, water use efficiency, water savings, climate-smart agricultural, CSA, greenhouse gas mitigation policies, Biosphere Reserve	irrigation, urban water management, greenhouse gas abatement, irrigated, urban irrigation, anaerobic digestion, climate regulations, natural gas-fired combined cycle, NGCC, Clean Air Interstate Rule
CLIMATE-PROPERTIES	temperature, streamflow, precipitation, burned area, discharge, soil moisture, altitude, Precipitation, solar radiation, albedo	temperature, precipitation, SOS, OHC, wind speed, soil moisture, SPEI, GPP, ZTD, EOS	temperature, soil moisture, ρ_{eff}, ETref, effective density, EE, thermal comfort
CLIMATE-NATURE	precipitation, vegetation, rainfall, aerosol, hydrological, forest, runoff, ocean, atmosphere, aerosols	precipitation, atmosphere, reef, aerosol, glacier, atmospheric, ocean, land surface, sea ice, SAOD	precipitation, rainfall, atmosphere, tundra, aerosols, water balance, snow, water vapor, urban vegetation, ET
CLIMATE-MODELS	SWAT, DSSAT, HBV, CCSM, MAR, WRF, NARCCAP, CMIP3, HadCM3, CERES-Rice	CMIP5, CMIP6, CMIP3, WRF, LDAS-Monde, PCR-GLOBWB, RACMO, HIRHAM, Integrated Valuation of Ecosystem Services and Tradeoffs, System of Integrated Environmental and Economic Accounting	CMIP5, STIRPAT, RegCM4, EMIL, WRF-UCM, DCSM, Whole Atmosphere Community Climate Model, WACCM, EPIC, Coupled Model Intercomparison Project
CLIMATE-PROBLEM-ORIGINS	emission, emissions, fossil fuel, urbanization, LUCC, land use change, fossil fuels, population growth, land use changes, NG	emission, land use change, urbanization, LUC, nutrient loading, toxic substances, water abstraction, population growth, corruption, food production	emission, manure, fossil fuel combustion, emissions, population growth, Manure, intestines of animals, coal, fuel wood, space heating
CLIMATE-OBSERVATIONS	NDVI, MODIS, Landsat, lidar, ALOS, PALSAR, HAMSR, ALS, GPS, ERBE	Landsat, research cruises SO234-2, SO235, RV SONNE	SAR, TerraSAR-X, Sentinel-1, campaign STABLE, Beijing Station, DMA-CPMA-SP2, SP2, Lyman–Birge–Hopfield, LBH, band systems

Continued on next page

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-ASSETS	agriculture, agricultural, crop, livestock, food security, wheat, building, maize, health, food	water resources, agricultural, food security, farmers, eco-environmental benefits, water supplies, farm, crop yields, water infrastructure, health benefits	agricultural, livestock, nutrients, water availability, crop, buildings, health, monocrops, maize, urban water systems
CLIMATE-IMPACTS	damage, damages, diseases, disease, disaster, deaths, pneumonia, mortality, disruptions, downy mildew	Lyme disease, poverty, malaria, eco-livelihood impacts, economic losses, encephalitis, babesiosis, anaplasmosis, homeless, disruption	climate catastrophe, continuous damage, killing, Burn area, plant diseases, Food insecurity, poverty, unsustainable livelihoods, loss of crop, exhaustion
CLIMATE-GREENHOUSE-GASES	CO2, carbon dioxide, methane, BC, CH4, Carbon dioxide, Non-Methane Hydrocarbons, NMHC, NOx	CO2	methane, carbon dioxide, CO2, rBC, perfluorocarbons, PFCs, decafluorobutane, C4F10, dodecafluoropentane, C5F12
CLIMATE-ORGANIZATIONS	IPCC, ECMWF, NCEP, GMEP, Geophysical Fluid Dynamics Laboratory, BRI, IIASA, GFDL, NASA, WOCE	SALKKU, IPCC, SMHI, Expert Team on Climate Change Detection and Indices, ETCCDI, National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory, Interreg IVB project AMICE, FEM, UNESCO, European Commission	IPCC, North American Regional Climate Change Assessment Program, NASA, DAMOCLES, Australian Bureau of Meteorology, CIERA, European Centre for Medium-Range Weather Forecasts, ECMWF, Climatic Research Unit, CRU
CLIMATE-ORGANISMS	species, habitat, biodiversity, plant, frogs, phytoplankton, tree, Joshua trees, diversity, butterfly	species, pine, Ecosystem diversity, rare species, snow leopard, demersal fish, biodiversity, birds, vascular plants, endemic	plant, eelgrass, biological species, habitat, Bryophytes, habitats, trees, Zostera marina L.

Continued on next page

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-DATASETS	TRMM, A2, CMORPH, UTCI, RCP8.5, SRES, TMPA, A1B, Tropical Rainfall Measuring Mission, ERA-Interim	CAMS-OPI, AR5, APHRODITE, Princeton Global Forcing, Soil-adjusted Vegetation Index, SAVI, Index-based Built-up Index, IBI, Soil Brightness Index, NDSI	RCP4.5, RCP8.5, A2, CIMIS, B2, ERA-Interim 6, ERA-Interim, CRU, Fourth Assessment Report, Climate Hazards group InfraRed Precipitation with Station data

Table 10: Ten most frequent NE instances in each category for every data split of Climate-Change-NER

Least frequent NE instances per category in Climate-Change-NER			
NE Class	**TRAIN**	**DEVELOPMENT**	**TEST**
CLIMATE-HAZARDS	Tropical Cyclones, Extratropical Transition, extratropical transition, ET, cyclone, Tropical Cyclone, Lightning, LANDSLIDE, volcano, debris flowing	shoreline instability, sea level rise, Water scarcity, water gap, fire, PM2.5 emissions, fire emissions, ozone depletion, Flood, extreme precipitation	woody plant encroachment, insect pests, ocean acidification, pests, Sea level rise, nitrate leaching, fire emissions, Drought, Inundation, flooding
CLIMATE-MITIGATIONS	National Plans to Combat Desertification, NPCD, cycling, alternative cropping systems, Urban water management, stormwater capture, strategy for adaptation, alternative management strategies, fire risk prevention, Agroforestry	soil conservation technique, residue mulch, pigeon pea hedges, wind technology, wind energy, trading scheme, fire management, mitigation strategy, water recycling, disaster preparedness	early warning systems, disaster relief strategies, building envelopes, double facades, solar chimney, passive and active solar control systems, wildlife reservoirs, clothing insulation, Paris Agreement, suppression of fire
CLIMATE-PROPERTIES	dNBR, aspect, damage, SPEI, WCI, PRECIPITATION, population, surface displacement, oxygen lines, water - vapor line	form drag, sea level changes, daily precipitation, air temperature, atmospheric lifetimes, oceanic and atmospheric concentrations, tropical lifetime, Standardized Precipitation Index, SPI, discharge	age, exposure to sun, population dynamics, sedimentation, ice - impact rates, equilibrium temperature, particle size, sea - ice cover, Precipitation, precipitations
Continued on next page			

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-NATURE	tropical island, convective, stratiform, rainclouds, raincloud, tectonic plates, tropical, stream, Rainstorms, rainstorm	ice cover, sea level, Coastal, stratosphere, oceanic emissions, Asian monsoon anticyclone, oceanic, Water Resources, watershed, coast	winds, aerosol, clouds, sea - ice retreat, dust plume, Rainfall, water - resource, hydroclimatic, soil, catchment
CLIMATE-MODELS	HadGEM2, RACMOv2, CMIP6, seasonal climate forecast system, SEAS5, Fire Events Delineation, FSU superensemble, UKMO, CPTEC, Global Metropolitan Detector	FireMIP, Community Land Model, CLM, Joint UK Land environment Simulator - Interactive Fire And Emission Algorithm For Natural Environments, JULES - INFERNO, Advanced Weather Research and Forecasting, WRF - ARW, MED - CORDEX, MikeShe, Mike11	CARMA, pSIMS, APSIM, DSSAT, INteractive Fires and Emissions algoRithm for Natural environments, UK 's Earth System Model, UKESM1, Waterdyn, Australian Water Resource Assessment, AWRA - L
CLIMATE-PROBLEM-ORIGINS	impervious areas, demographic pressure, landuse / land cover changes, land - use, Natural Gas, hydraulic fracturing, coal, natural gas, power plants, power plant	nutrient loading, toxic substances, water abstraction, population growth, corruption, food production, Inadequate timber extraction, cattle, abusive recreational practices, urban expansion	Manure, intestines of animals, coal, fuel wood, space heating, heating, gas and electricity consumption, human damage activities, urbanization, anthropogenic ignition
CLIMATE-OBSERVATIONS	Differential Interferometric of Synthetic Aperture Radar, Global Positioning System, High - Altitude MMIC Sounding Radiometer, High - Altitude Monolithic Microwave Integrated Circuit (MMIC) Sounding Radiometer, CAMEX-4, Tropical Cloud Systems and Processes, African Monsoon Multidisciplinary Analyses, GH, Synthetic Aperture Radar, TOPEX / POSEIDON Radar Altimeter	Landsat, research cruises SO234 - 2, SO235, RV SONNE	Lyman – Birge – Hopfield, LBH, band systems, Analytical Spectral Device (ASD) Field Spec Pro, Landsat 8 Operational Land Imager, OLI, SALTRACE, lidar, C- and X - band, C - band

Continued on next page

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-ASSETS	croplands, human health, FOOD SECURITY, urban areas, healthy diets, fruit, income opportunities, Legumes, forages, drinking water	national welfare, forestry, Transport infrastructure networks, infrastructure, vehicle, cassava, Smallholder, food supply, crops, Water supply	income, smallholder, high - rise, skyscrapers, built environments, wellbeing, pea, oat, soybean, agricultural productivity
CLIMATE-IMPACTS	flood damage, disruption, detrimental, personal losses, Japanese encephalitis, bovine tuberculosis, famine, homeless, food insecurity, disastrous	eco - livelihood impacts, economic losses, encephalitis, babesiosis, anaplasmosis, homeless, disruption, disruptions, flood footprint, traffic disruptions	unsustainable livelihoods, loss of crop, exhaustion, illness, destruction, calamities, Bovine tuberculosis, zoonosis, mortality, disaster
CLIMATE-GREENHOUSE-GASES	CO2, carbon dioxide, methane, BC, CH 4, Carbon dioxide, CO 2, Non - Methane Hydrocarbons, NMHC, NOx	CO 2	C4F10, dodecafluoropentane, C 5F12, tetradecafluorohexane, C6F14, hexadecafluoroheptane, C 7F16, octadecafluorooctane, C 8F18, black carbon
CLIMATE-ORGANIZATIONS	CMA, CMC, BOM, MF, KMA, eartH2Observe, NOAA, Vaisala, Jet Propulsion Laboratory, JPL	AfriCultuReS, CNRM, Centre National de Recherches Météorologiques, European Center for Medium Range Weather Forecast, ECMWF, SEAREG, Swedish Meteorological and Hydrological Institute, APEC Climate Center, APCC, ENSEMBLES	BMD, Australian Water Availability Project, AWAP, Scaling and Assimilation of Soil Moisture and Streamflow, SASMAS, ZKI, Center for Satellite - Based Crisis Information, German Aerospace Center, DLR, European Space Agency
CLIMATE-ORGANISMS	biosphere, indigenous animals, white - tailed deer, Rare plant, Trifolium repens L., Trifolium vesiculosum Savi, Clover, Loblolly Pine, Pinus taeda L., Tree	rare species, snow leopard, demersal fish, biodiversity, birds, vascular plants, endemic, organisms, plant species, habitat	plant, eelgrass, biological species, habitat, Bryophytes, habitats, trees, Zostera marina L.

Continued on next page

NE CLASS	TRAIN	DEVELOPMENT	TEST
CLIMATE-DATASETS	Precipitation Estimation from Remotely Sensed information using Artificial Neural Networks, FAOSTAT, EarthStat, WorldClim, National Lightning Detection Network, Long Range Lightning Detection Network, Canadian Lightning Detection Network, CLDN, Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks, Integrated Multi - satellitE Retrievals for Global	Global Fire Emissions Database, GFED, ERA - INTERIM, Climate Assessment and Dataset, ECA&D, ERA - Interim, UK Foresight Future Flooding Report, Climate Anomaly Monitoring System - Outgoing Longwave Radiation Precipitation Index, Asian Precipitation - Highly - Resolved Observational Data Integration Towards Evaluation, A1B	CHIRPS, E - OBS, Climate Prediction Center MORPHING, CMORPH, Tropical Rainfall Measuring Mission, TRMM, Precipitation Estimation Algorithm from Remotely - Sensed Information using an Artificial Neural Network, PERSIANN, global Satellite Mapping of Precipitation, GSMaP

Table 11: Ten least frequent NE instances in each category for every data split of Climate-Change-NER

Appendix B: Additional information about BiodivNER

Appendix B provides the definitions of the 6 named entities contained in the dataset BiodivNER, which have been obtained from the explanations of each category in Abdelmageed et al. (2022). In addition to these definitions, which have been used to design the prompts, 12 gives an overview of the frequency of NE instances per NE class and per data split. Tables 13 and 14 give basic statistical information about the average, median, maximum, and minimum length of sentences and tokens in each data split and in the whole dataset, alongside the total number of sentences and tokens per data split and in the complete dataset. Finally, Tables 15 and 16 list the ten most frequent and least frequent real-world instances of the 6 NE categories respectively. This data is presented per data split. BiodivNER is a larger dataset than Climate-Change-NER and, while some NE classes have fewer NE instances, these are not as weakly represented as some of the classes in Climate-Change-NER.

Definitions of NE classes

- **ORGANISM**: all individual life forms such as microorganisms, plants, animals, mammals, insects, fungi, bacteria etc.;
- **PHENOMENA**: occurring natural, biological, physical or chemical processes such as decomposition, colonisation, deforestation, as well as events, such as climate change;
- **MATTER**: chemical and biological compounds, and natural elements, such as carbon, sediment, sand;
- **ENVIRONMENT**: natural and man-made environments organisms live in, such as groundwater, garden, aquarium, mountain;
- **QUALITY**: data parameters measured or observed, phenotypes and traits, such as volume, age, structure, morphology;
- **LOCATION**: geographic location such as China, the United States etc.

Named entity class	Train	Development	Test	Total per category
ORGANISM	1977	164	281	2422
PHENOMENA	517	59	63	639
MATTER	471	41	292	804
ENVIRONMENT	1167	157	154	1478
QUALITY	2406	292	455	3153
LOCATION	170	20	32	222
Total per data split	6708	733	1277	8718

Table 12: Named entity instances per category and per data split in BiodivNER

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	1828	39.03	28	1053	3
development	229	39.35	28	438	3
test	229	47.74	28	2047	5
total	2286	39.94	28	2047	3

Table 13: BiodivNER sentence features

Complete dataset information Sentence lengths: mean: 39.94, median: 28.0, max: 2047, min: 3. Number of total sentences is 2286 Token lengths: mean: 5.25, median: 4, max: 99, min: 1. Number of total tokens is 91293 The data split has a total of 8718 entities.

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	71348	5.24	4	99	1
development	9012	5.12	4	65	1
test	10933	5.45	5	60	1
total	91293	5.25	4	99	1

Table 14: BiodivNER token features

Most frequent NE instances per category in BiodivNER			
NE class	**Train**	**Development**	**Test**
ORGANISM	species, tree, morphospecies, Arthropod species, Diptera, trees, plant, Species, Tree, bacteria	species, trees, tree, plant, plant species, earthworms, human, Species, flora, tree species	species, tree, plant, caterpillars, Arthropod species, morphospecies, trees, Plant, seedling, Species
ENVIRONMENT	soil, field, ecosystem, forest, Soil horizon, community, habitat, woody, land vegetation	field, ecosystem, soil, forest, community, land, soils, habitat, Vegetation, forests	field, soil, ecosystem, forest, land, habitat, community, soils, communities, Nature Reserve
LOCATION	China, country, Jiangxi Province, location, locations, Tübingen, Zhejiang Province, Mediterranean Sea, Germany, countries	China, Jiangxi Province, countries, location, Greenland, Tübingen, Netstal, Switzerland, country, Sweden	China, Lueneburg Scharnhorststr, country, New Zealand, Country, Zhejiang Province, Australia, USA, Tübingen Rümelinstr, Freiburg im Breisgau
MATTER	carbon, metal, water, woody debris, sand, wood, nitrogen, woody debris items, sediment, clay	sediment, water, oil palm, metal, carbon, maize, soya, bean, Coarse material, coarse sand	Chemical elements, Nitrogen, Carbon, carbon, woody debris, sediment, N, C, Ca, Fe
PHENOMENA	Precipitation, rainfall, climate change, sand loss, precipitation, pollination, conservation, CO2 emissions, planting, ice storm	Precipitation, conservation, rainfall, precipitation, extinction, planting, consumption, pollination, Biological activity, growth	Precipitation, rainfall, climate change, precipitation, fragmentation, growth, conservation, weather, rain events, mutualistic ant-hemipteran interactions
QUALITY	abundance, height, diameter, rainfall amount, species richness, Abundance, area, trait, soil properties, peak rainfall intensity	species description, abundance, rainfall amount, species richness, average rainfall intensity, height, biomass, density, Abundance, peak rainfall intensity	Phylogenetic biodiversity, abundance, rainfall amount, average rainfall intensity, Leaf stomata size, Abundance, species richness, area, peak rainfall intensity, biomass

Table 15: Ten most frequent NE instances in each category for every data split of BiodivNER

Least frequent NE instances per category in BiodivNER			
NE class	**Train**	**Development**	**Test**
ORGANISM	wildlife, raccoon, deer, hawk, chicken, rooster, dog, cat, squirrel, invertebrates	fishes, conifer plantations, Cunninghamia lanceolata, Pinus massoniana, Plant, caterpillars, Organism, Animal, open-habitat species, Dead wood	Cunninghamia lanceolata, Pinus massoniana, voucher specimens, fish, Herbivore, Herbivores, microbes, Animal, drought-sensitive species, rogenhoferi
ENVIRONMENT	above ground, Grasses, herbivore communities, green spaces, herb layer, lands, Tropical Agroforestry Landscapes, above-ground, harbor, sea ice cover	host, forest habitat, forest ecosystem, sites, subplots, broad-leaved forests, tree, broadleaf canopies, bare ground, dense conifer forests	nature, Island, plant communities, Park, vegetation layer, above ground, soil environment, bathyal habitat, mineral soil layers, organic soil layers
LOCATION	Laysan Island, Freiburg Germany, New York City, United States, Puerto Rico, U.S. Virgin Islands, North America, South-eastern China, Xingangshan, Tübingen Germany	China, Jiangxi Province, countries, location, Greenland, Tübingen, Netstal, Switzerland, country, Sweden	Alboran Sea, Aegean Sea, Lüneburg Germany, Leipzig, Deutscher Platz 5e Leipzig Germany, Freiburg, Tennenbacher Str, carbon, south-east China, Leipzig Deutscher Platz 5e Leipzig Germany
MATTER	terpenes, chemical elements, Soil lipid fractions, primary metabolites, nutrients, antioxidant, cellulosic, biofuels, flux, Nitrate	fine sand, below-ground carbon, biofuel, sunlight, medium silt, soil enzymes, nitrogen, phosphorus, potassium, extrafloral nectaries	Rock fragments, network, soil samples, clay, total clay, fine sand, oil palm, raindrops, CO2, stratum
PHENOMENA	Weather, death, impairment, forest, short dry season, fragmentation, tree planting, human, pressures, Forest restoration, ocean warming	intraspecific variation, trait evolution, Water consumption, summer, autumn, tree planting, climate change, Fertilization, Grazing, neighbourhood interactions	environmental change, treatment, throughfall, spring, summer, Climate Change, Rainfall, drought, mutualism, predation
QUALITY	fungal biomass, positioning, woody increment, species-level trait, trait covariations, biodiversity indices, shannon index, Altitude, forests cover, growth	mass, young, shrub species names, sulphur contents, needle area, Total needle surface area, Soil description, species, landscape heterogeneity, crown asymmetry	Species name, Capacity, landscape scales, successional age, diversity gradient, shrub position, thick layers, toxicity, pH, Microbial biomass

Table 16: Ten least frequent NE instances in each category for every data split of BiodivNER

Appendix C: Examples of prompt templates

Definition: An entity is an ORGANISM (all individual life forms such as microorganisms, plants, animals, mammals, insects, fungi, bacteria etc.), PHENOMENA (occurring natural, biological, physical or chemical processes such as decomposition, colonisation, deforestation, as well as events, such as climate change etc.), MATTER (chemical and biological compounds, and natural elements, such as carbon, sediment, sand etc.), ENVIRONMENT (natural and man-made environments organisms live in, such as groundwater, garden, aquarium, mountain etc.), QUALITY (data parameters measured or observed, phenotypes and traits, such as volume, age, structure, morphology etc.), and LOCATION (geographic location such as China, the United States etc.).
Dates, times, and adjectives are not entities.

Section (a) : Definitions and NE instances

Example 1: [[0, 'Because'], [1, 'fungal'], [2, 'pathogens'], [3, 'likely'], [4, 'have'], [5, 'similar'], [6, 'abiotic'], [7, 'requirements'], [8, 'for'], [9, 'growth'], [10, 'as'], [11, 'other'], [12, 'fungi'], [13, ','], [14, 'characterizing'], [15, 'weather'], [16, 'conditions'], [17, 'favorable'], [18, 'for'], [19, 'fungi'], [20, 'also'], [21, 'may'], [22, 'be'], [23, 'used'], [24, 'to'], [25, 'predict'], [26, 'the'], [27, 'selective'], [28, 'pressures'], [29, 'imposed'], [30, 'by'], [31, 'pathogenic'], [32, 'fungi'], [33, 'on'], [34, 'plants'], [35, 'in'], [36, 'different'], [37, 'habitats'], [38, '.']]
Answer: [['growth', 'PHENOMENA', 9, 9], ['fungi', 'ORGANISM', 12, 12], ['weather conditions', 'QUALITY', 15, 16], ['fungi', 'ORGANISM', 19, 19], ['fungi', 'ORGANISM', 32, 32], ['plants', 'ORGANISM', 34, 34], ['habitats', 'ENVIRONMENT', 37, 37]]

Example 2: [[0, '--'], [1, '-'], [2, '6'], [3, 'digit'], [4, 'metal'], [5, 'tags'], [6, 'starting'], [7, 'with'], [8, '3'], [9, 'were'], [10, 'also'], [11, 'used'], [12, 'for'], [13, 'woody'], [14, 'debris'], [15, 'items'], [16, 'CSP'], [17, 'metal'], [18, 'tag'], [19, 'number'], [20, '('], [21, 'trees'], [22, ','], [23, 'woody'], [24, 'debris'], [25, ')'], [26, '('], [27, 'TagMBa'], [28, ')'], [29, ','], [30, 'dimensionless'], [31, 'TagMBa'], [32, 'CSP'], [33, 'tree'], [34, 'individuals'], [35, 'were'], [36, 'marked'], [37, 'mostly'], [38, 'with'], [39, 'metal'], [40, 'tags'], [41, 'but'], [42, 'also'], [43, 'additional'], [44, 'tags'], [45, 'were'], [46, 'used'], [47, '.']]
Answer: [['metal', 'MATTER', 4, 4], ['woody', 'ENVIRONMENT', 13, 13], ['metal', 'MATTER', 17, 17], ['trees', 'ORGANISM', 21, 21], ['woody', 'ENVIRONMENT', 23, 23], ['tree', 'ORGANISM', 33, 33], ['metal', 'MATTER', 39, 39]]

Example 3: [[0, '('], [1, 'Phenolics'], [2, ':'], [3, 'Total'], [4, 'phenolics'], [5, 'content'], [6, 'as'], [7, 'tannic'], [8, 'acid'], [9, 'equivalent'], [10, ')'], [11, 'dimensionless'], [12, 'real'], [13, 'Secondary'], [14, 'Metabolites'], [15, 'Secondary'], [16, 'metabolites'], [17, 'are'], [18, 'organic'], [19, 'compounds'], [20, 'that'], [21, 'are'], [22, 'not'], [23, 'directly'], [24, 'involved'], [25, 'in'], [26, 'the'], [27, 'normal'], [28, 'growth'], [29, ','], [30, 'development'], [31, ','], [32, 'or'], [33, 'reproduction'], [34, 'of'], [35, 'an'], [36, 'organism'], [37, '.']]
Answer: [['tannic acid', 'MATTER', 7, 8], ['Secondary Metabolites', 'MATTER', 13, 14], ['Secondary metabolites', 'MATTER', 15, 16], ['organic compounds', 'MATTER', 18, 19], ['normal growth', 'PHENOMENA', 27, 28], ['development', 'PHENOMENA', 30, 30], ['reproduction', 'PHENOMENA', 33, 33], ['organism', 'ORGANISM', 36, 36]]

Example 4: [[0, 'Given'], [1, 'limitations'], [2, 'in'], [3, 'using'], [4, 'organ-level'], [5, 'traits'], [6, 'to'], [7, 'predict'], [8, 'ecological'], [9, 'performance'], [10, 'of'], [11, 'species'], [12, ','], [13, 'recent'], [14, 'advances'], [15, 'using'], [16, 'tolerances'], [17, 'of'], [18, 'low'], [19, 'resource'], [20, 'availability'], [21, 'as'], [22, 'plant'], [23, 'functional'], [24, 'traits'], [25, 'are'], [26, 'revealing'], [27, 'the'], [28, 'often'], [29, 'hidden'], [30, 'roles'], [31, 'these'], [32, 'factors'], [33, 'have'], [34, 'in'], [35, 'structuring'], [36, 'communities'], [37, 'and'], [38, 'are'], [39, 'becoming'], [40, 'central'], [41, 'to'], [42, 'classifying'], [43, 'plants'], [44, 'ecologically'], [45, '.']]
Answer: [['traits', 'QUALITY', 5, 5], ['species', 'ORGANISM', 11, 11], ['plant', 'ORGANISM', 22, 22], ['traits', 'QUALITY', 24, 24], ['communities', 'ENVIRONMENT', 36, 36], ['plants', 'ORGANISM', 43, 43]]

Section (b) : Question-answer, k=4, random examples

Generate ONLY a Python list with a nested list of named entities from the sentence: [[0, 'This'], [1, 'system'], [2, 'is'], [3, 'an'], [4, 'important'], [5, 'model'], [6, 'for'], [7, 'understanding'], [8, 'how'], [9, 'microbial'], [10, 'communities'], [11, 'degrade'], [12, 'plant'], [13, 'biomass'], [14, 'in'], [15, 'natural'], [16, 'systems'], [17, 'and'], [18, 'has'], [19, 'direct'], [20, 'relevancy'], [21, 'for'], [22, 'bioenergy'], [23, ','], [24, 'given'], [25, 'recent'], [26, 'interest'], [27, 'in'], [28, 'cellulosic'], [29, 'biofuels'], [30, '.']].
one-token entity: [entity, label, index of token in list, index of token of list]
multi-token entity: [entity, label, index of first token, index of last token]
DO NOT HALLUCINATE

Section (c) : Task

Figure 1: Prompt from BiodivNER with 4 randomly selected question-answer pairs.

Definition: An entity is an ORGANISM (all individual life forms such as microorganisms, plants, animals, mammals, insects, fungi, bacteria etc.), PHENOMENA (occurring natural, biological, physical or chemical processes such as decomposition, colonisation, deforestation, as well as events, such as climate change etc.), MATTER (chemical and biological compounds, and natural elements, such as carbon, sediment, sand etc.), ENVIRONMENT (natural and man-made environments organisms live in, such as groundwater, garden, aquarium, mountain etc.), QUALITY (data parameters measured or observed, phenotypes and traits, such as volume, age, structure, morphology etc.), and LOCATION (geographic location such as China, the United States etc.).
Dates, times, and adjectives are not entities.

Section (a) : Definitions and NE instances

Example 1: [[0, 'Earthworm'], [1, 'invasion'], [2, 'effects'], [3, 'on'], [4, 'soil'], [5, 'micro-organisms'], [6, 'were'], [7, 'context-dependent'], [8, ','], [9, 'such'], [10, 'as'], [11, 'depending'], [12, 'on'], [13, 'functional'], [14, 'group'], [15, 'richness'], [16, 'of'], [17, 'invasive'], [18, 'earthworms'], [19, 'and'], [20, 'soil'], [21, 'depth'], [22, '.']]
Answer: [['Earthworm invasion', 'PHENOMENA', 0, 1], ['soil', 'QUALITY', 4, 4], ['earthworms', 'ORGANISM', 18, 18], ['soil depth', 'QUALITY', 20, 21]]

Example 2: [[0, 'To'], [1, 'disentangle'], [2, 'how'], [3, 'functional'], [4, 'traits'], [5, 'explain'], [6, 'community'], [7, 'growth'], [8, 'and'], [9, 'underpin'], [10, 'biodiversity - ecosystem'], [11, 'functioning'], [12, 'relationships'], [13, ','], [14, 'we'], [15, 'should'], [16, 'elucidate'], [17, 'how'], [18, 'plant'], [19, 'traits'], [20, 'affect'], [21, 'individual'], [22, 'growth'], [23, 'across'], [24, 'species'], [25, 'richness'], [26, 'levels'], [27, ','], [28, 'because'], [29, 'the'], [30, 'role'], [31, 'of'], [32, 'functional'], [33, 'traits'], [34, 'on'], [35, 'growth'], [36, 'depends'], [37, 'on'], [38, 'the'], [39, 'ecological'], [40, 'context'], [41, 'of'], [42, 'the'], [43, 'individual'], [44, '.']]
Answer: [['functional traits', 'QUALITY', 3, 4], ['community growth', 'PHENOMENA', 6, 7], ['plant traits', 'QUALITY', 18, 19], ['growth', 'PHENOMENA', 22, 22], ['species richness', 'QUALITY', 24, 25], ['functional traits', 'QUALITY', 32, 33], ['growth', 'PHENOMENA', 35, 35]]

Example 3: [[0, 'To'], [1, 'combine'], [2, 'these'], [3, 'data'], [4, 'with'], [5, 'the'], [6, 'data'], [7, 'of'], [8, 'the'], [9, 'living'], [10, 'trees'], [11, ','], [12, 'check'], [13, 'the'], [14, 'dataset'], [15, 'on'], [16, 'metal'], [17, 'tags'], [18, 'and'], [19, 'coarse'], [20, 'woody'], [21, 'debris'], [22, 'items'], [23, '``'], [24, 'Metal'], [25, 'tags'], [26, 'for'], [27, 'coarse'], [28, 'woody'], [29, 'debris'], [30, '('], [31, 'CWD'], [32, ')'], [33, 'items'], [34, 'in'], [35, 'the'], [36, 'CSPs'], [37, 'and'], [38, 'corresponding'], [39, 'metal'], [40, 'tags'], [41, 'and'], [42, 'stem'], [43, 'ids'], [44, 'from'], [45, 'the'], [46, 'living'], [47, 'tree'], [48, 'data'], [49, '.'], [50, '``'], [51, '.']]
Answer: [['living trees', 'ORGANISM', 9, 10], ['coarse woody debris', 'MATTER', 19, 21], ['coarse woody debris', 'MATTER', 27, 29], ['living tree', 'ORGANISM', 46, 47]]

Example 4: [[0, 'Accommodating'], [1, 'Species'], [2, 'Climate-Forced'], [3, 'Dispersal'], [4, 'and'], [5, 'Uncertainties'], [6, 'in'], [7, 'Spatial'], [8, 'Conservation'], [9, 'Planning'], [10, 'Abstract'], [11, 'Spatial'], [12, 'conservation'], [13, 'prioritization'], [14, 'should'], [15, 'seek'], [16, 'to'], [17, 'anticipate'], [18, 'climate'], [19, 'change'], [20, 'impacts'], [21, 'on'], [22, 'biodiversity'], [23, 'and'], [24, 'to'], [25, 'mitigate'], [26, 'these'], [27, 'impacts'], [28, 'through'], [29, 'the'], [30, 'development'], [31, 'of'], [32, 'dynamic'], [33, 'conservation'], [34, 'plans'], [35, '.']]
Answer: [['Species', 'ORGANISM', 1, 1], ['climate change', 'PHENOMENA', 18, 19]]

Section (b) : Question-answer, k=4, examples with high similarity score

Generate ONLY a Python list with a nested list of named entities from the sentence: [[0, 'The'], [1, 'primacy'], [2, 'of'], [3, 'either'], [4, 'species'], [5, 'or'], [6, 'functional'], [7, 'group'], [8, 'richness'], [9, 'effects'], [10, 'depended'], [11, 'on'], [12, 'the'], [13, 'sequence'], [14, 'of'], [15, 'testing'], [16, 'these'], [17, 'terms'], [18, ','], [19, 'indicating'], [20, 'that'], [21, 'both'], [22, 'aspects'], [23, 'of'], [24, 'richness'], [25, 'were'], [26, 'congruent'], [27, 'and'], [28, 'complementary'], [29, 'to'], [30, 'expected'], [31, 'strong'], [32, 'effects'], [33, 'of'], [34, 'legume'], [35, 'presence'], [36, 'and'], [37, 'grass'], [38, 'presence'], [39, 'on'], [40, 'plant'], [41, 'chemical'], [42, 'composition'], [43, '.']].
one-token entity: [entity, label, index of token in list, index of token of list]
multi-token entity: [entity, label, index of first token, index of last token]
DO NOT HALLUCINATE

Section (c) : Task

Figure 2: Prompt from BiodivNER with 4 question-answer pairs, where the question has a high similarity score with the task question.

An entity is CLIMATE-HAZARDS (hazards with potential negative impact on climate, such as floods, wildfires, droughts, and heatwaves), CLIMATE-PROBLEM-ORIGINS (problems that describe why the climate is changing, such as fossil fuel, deforestation as well as sectors that can be cited as causes of energy use, such as the transport sector, animal agriculture or fuel imports), CLIMATE-GREENHOUSE-GASES (gases that cause heating of the atmosphere, such as carbon dioxide (CO2), methane (CH4), octadecafluorooctane (C 8F18)).

Section (a): Definitions and NE instances

Example 1: [[0, 'In'], [1, 'a'], [2, 'changing'], [3, 'climate'], [4, ','], [5, 'the'], [6, 'impact'], [7, 'of'], [8, 'tropical'], [9, 'cyclones'], [10, 'on'], [11, 'the'], [12, 'United'], [13, 'States'], [14, 'Atlantic'], [15, 'and'], [16, 'Gulf'], [17, 'Coasts'], [18, 'will'], [19, 'be'], [20, 'affected'], [21, 'both'], [22, 'by'], [23, 'how'], [24, 'intense'], [25, 'and'], [26, 'how'], [27, 'frequent'], [28, 'these'], [29, 'storms'], [30, 'become'], [31, '.']]
Answer: [['tropical cyclones', 'CLIMATE-HAZARDS', 8, 9]]

Example 2: [[0, 'We'], [1, 'assessed'], [2, 'the'], [3, 'uncertainties'], [4, 'around'], [5, 'oil'], [6, 'and'], [7, 'NG'], [8, 'emissions'], [9, 'by'], [10, 'using'], [11, 'measurements'], [12, 'from'], [13, 'the'], [14, 'FRAPPE'], [15, 'and'], [16, 'DISCOVER'], [17, '-'], [18, 'AQ'], [19, 'campaigns'], [20, 'over'], [21, 'the'], [22, 'Northern'], [23, 'Front'], [24, 'Range'], [25, 'Metropolitan'], [26, 'Area'], [27, '('], [28, 'NFRMA'], [29, ')'], [30, 'in'], [31, 'summer'], [32, '2014'], [33, '.']]
Answer: [['oil', 'CLIMATE-PROBLEM-ORIGINS', 5, 5], ['NG', 'CLIMATE-PROBLEM-ORIGINS', 7, 7], ['emissions', 'CLIMATE-PROBLEM-ORIGINS', 8, 8]]

Example 3: [[0, 'A'], [1, 'better'], [2, 'understanding'], [3, 'of'], [4, 'the'], [5, 'middle'], [6, 'atmosphere'], [7, 'and'], [8, 'how'], [9, 'it'], [10, 'reacts'], [11, 'to'], [12, 'the'], [13, 'current'], [14, 'increase'], [15, 'in'], [16, 'the'], [17, 'concentration'], [18, 'of'], [19, 'carbon'], [20, 'dioxide'], [21, '('], [22, 'CO2'], [23, ')'], [24, 'is'], [25, 'therefore'], [26, 'necessary'], [27, '.']]
Answer: [['carbon dioxide', 'CLIMATE-GREENHOUSE-GASES', 19, 20], ['CO2', 'CLIMATE-GREENHOUSE-GASES', 22, 22]]

Example 4: [[0, 'On'], [1, 'the'], [2, 'other'], [3, 'hand'], [4, ','], [5, 'the'], [6, 'data'], [7, '-'], [8, 'driven'], [9, 'models'], [10, 'have'], [11, 'been'], [12, 'proven'], [13, 'to'], [14, 'correct'], [15, 'this'], [16, 'bias'], [17, 'in'], [18, 'many'], [19, 'cases'], [20, ','], [21, 'unlike'], [22, 'the'], [23, 'semi'], [24, '-'], [25, 'empirical'], [26, 'hydrological'], [27, 'model'], [28, 'GR1A.'], [29, 'The'], [30, 'research'], [31, 'investigates'], [32, 'geographical'], [33, 'and'], [34, 'temporal'], [35, 'variability'], [36, 'of'], [37, 'hail'], [38, 'incidence'], [39, 'based'], [40, 'on'], [41, 'conventional'], [42, 'stations'], [43, 'reports'], [44, 'on'], [45, 'hail'], [46, 'days'], [47, 'from'], [48, '1891'], [49, 'to'], [50, '2015'], [51, '.']]
Answer: [['hail', 'CLIMATE-HAZARDS', 37, 37], ['hail', 'CLIMATE-HAZARDS', 45, 45]]

Section (b): Question-answer, k=4

Generate ONLY a Python list with a nested list of named entities from the sentence: [[0, 'Particularly'], [1, ','], [2, 'air'], [3, '–'], [4, 'surface'], [5, 'fluxes'], [6, 'of'], [7, 'methane'], [8, 'and'], [9, 'carbon'], [10, 'dioxide'], [11, 'are'], [12, 'of'], [13, 'interest'], [14, 'as'], [15, 'recent'], [16, 'observations'], [17, 'suggest'], [18, 'that'], [19, 'the'], [20, 'vast'], [21, 'stores'], [22, 'of'], [23, 'soil'], [24, 'carbon'], [25, 'found'], [26, 'in'], [27, 'the'], [28, 'Arctic'], [29, 'tundra'], [30, 'are'], [31, 'becoming'], [32, 'more'], [33, 'available'], [34, 'to'], [35, 'release'], [36, 'to'], [37, 'the'], [38, 'atmosphere'], [39, 'in'], [40, 'the'], [41, 'form'], [42, 'of'], [43, 'these'], [44, 'greenhouse'], [45, 'gases'], [46, '.']].
one-token entity: [entity, label, index of token in list, index of token of list]
multi-token entity: [entity, label, index of first token, index of last token]
DO NOT HALLUCINATE

Section (c): Task

Figure 3: Prompt from Climate-Change-NER with a cluster of NE classes.

Appendix D: Detailed performance report: Climate-Change-NER

	gpt-4o-mini				gpt-4o-2024-05-13				Mistral-7B-Instruct-v0.3			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-GREENHOUSE-GASES	0,5116	0,7097	0,5946	31	0,6757	0,8065	0,7353	31	0,2826	0,4194	0,3377	31
CLIMATE-HAZARDS	0,4898	0,7059	0,5783	34	0,4262	0,7647	0,5474	34	0,0746	0,2941	0,119	34
CLIMATE-PROBLEM-ORIGINS	0,0602	0,25	0,0971	20	0,2115	0,55	0,3056	20	0	0	0	20
micro avg	0,2914	0,6	0,3923	85	0,4133	0,7294	0,5277	85	0,0403	0,2706	0,0701	85
macro avg	0,3539	0,5552	0,4233	85	0,4378	0,7071	0,5294	85	0,1191	0,2378	0,1522	85
weighted avg	0,3967	0,6	0,471	85	0,4667	0,7294	0,559	85	0,1329	0,2706	0,1708	85
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-ASSETS	0,1318	0,34	0,1899	50	0,3881	0,52	0,4444	50	0,038	0,2	0,0639	50
CLIMATE-IMPACTS	0,12	0,3529	0,1791	17	0,1731	0,5294	0,2609	17	0,0545	0,1765	0,0833	17
CLIMATE-NATURE	0,2069	0,3711	0,2657	97	0,2265	0,4184	0,2939	98	0,0432	0,1702	0,069	94
CLIMATE-ORGANISMS	0,1724	0,4545	0,25	11	0,2609	0,5455	0,3529	11	0,1579	0,2727	0,2	11
micro avg	0,1675	0,3657	0,2298	175	0,2539	0,4659	0,3287	176	0,0453	0,186	0,0728	172
macro avg	0,1578	0,3797	0,2212	175	0,2621	0,5033	0,338	176	0,0734	0,2049	0,104	172
weighted avg	0,1748	0,3657	0,2346	175	0,2694	0,4659	0,3372	176	0,0502	0,186	0,0773	172
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-DATASETS	0,1882	0,64	0,2909	25	0,3333	0,625	0,4348	24	0,0179	0,04	0,0247	25
CLIMATE-MODELS	0,5514	0,6277	0,5871	94	0,5421	0,6304	0,5829	92	0,0947	0,2045	0,1295	88
CLIMATE-OBSERVATIONS	0,1667	0,2857	0,2105	21	0,3333	0,35	0,3415	20	0,0204	0,0952	0,0336	21
CLIMATE-PROPERTIES	0,1325	0,2326	0,1688	86	0,2963	0,3077	0,3019	78	0,018	0,0824	0,0296	85
micro avg	0,2665	0,4469	0,3339	226	0,4094	0,486	0,4444	214	0,0383	0,1279	0,0589	219
macro avg	0,2597	0,4465	0,3143	226	0,3763	0,4783	0,4153	214	0,0378	0,1055	0,0543	219
weighted avg	0,3161	0,4469	0,3601	226	0,4096	0,486	0,4413	214	0,0491	0,1279	0,0696	219
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-MITIGATIONS	0,0339	0,0526	0,0412	38	0,2222	0,4211	0,2909	38	0,0097	0,0789	0,0172	38
CLIMATE-ORGANIZATIONS	0,2889	0,4333	0,3467	30	0,3594	0,7667	0,4894	30	0,0746	0,1667	0,1031	30
micro avg	0,1442	0,2206	0,1744	68	0,2868	0,5735	0,3824	68	0,0212	0,1176	0,036	68
macro avg	0,1614	0,243	0,194	68	0,2908	0,5939	0,3901	68	0,0422	0,1228	0,0602	68
weighted avg	0,1464	0,2206	0,176	68	0,2827	0,5735	0,3785	68	0,0383	0,1176	0,0551	68

Table 17: NE cluster classes: All models

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-ASSETS	0,1304	0,12	0,125	50	0,2	0,18	0,1895	50	0,1875	0,18	0,1837	50
CLIMATE-DATASETS	0,1098	0,36	0,1682	25	0,1558	0,48	0,2353	25	0,1489	0,56	0,2353	25
CLIMATE-GREENHOUSE-GASES	0,325	0,4194	0,3662	31	0,3261	0,4839	0,3896	31	0,5854	0,7742	0,6667	31
CLIMATE-HAZARDS	0,475	0,5588	0,5135	34	0,5	0,5882	0,5405	34	0,5455	0,7059	0,6154	34
CLIMATE-IMPACTS	0,2143	0,1765	0,1935	17	0,2143	0,1765	0,1935	17	0,2	0,1765	0,1875	17
CLIMATE-MITIGATIONS	0,4091	0,2368	0,3	38	0,4	0,1579	0,2264	38	0,4348	0,2632	0,3279	38
CLIMATE-MODELS	0,4444	0,4681	0,456	94	0,3905	0,4362	0,4121	94	0,5104	0,5213	0,5158	94
CLIMATE-NATURE	0,25	0,3163	0,2793	98	0,2391	0,3367	0,2797	98	0,25	0,3776	0,3008	98
CLIMATE-OBSERVATIONS	0,4545	0,2381	0,3125	21	0,3846	0,2381	0,2941	21	0,5333	0,381	0,4444	21
CLIMATE-ORGANISMS	0,1176	0,1818	0,1429	11	0,1905	0,3636	0,25	11	0,1875	0,2727	0,2222	11
CLIMATE-ORGANIZATIONS	0,5714	0,6667	0,6154	30	0,6216	0,7667	0,6866	30	0,5122	0,7	0,5915	30
CLIMATE-PROBLEM-ORIGINS	0,1481	0,2	0,1702	20	0,2083	0,25	0,2273	20	0,2727	0,3	0,2857	20
CLIMATE-PROPERTIES	0,2456	0,3256	0,28	86	0,2336	0,2907	0,2591	86	0,1898	0,3023	0,2332	86
micro avg	0,2876	0,3477	0,3148	555	0,2947	0,3622	0,325	555	0,3162	0,4216	0,3614	555
macro avg	0,2996	0,3283	0,3017	555	0,3127	0,3653	0,3218	555	0,3506	0,4242	0,37	555
weighted avg	0,3118	0,3477	0,3212	555	0,3118	0,3622	0,3247	555	0,347	0,4216	0,3704	555

Table 18: Random k examples: gpt-4o-mini

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-ASSETS	0,2222	0,2041	0,2128	49	0,32	0,32	0,32	50	0,3125	0,3125	0,3125	48
CLIMATE-DATASETS	0,2125	0,68	0,3238	25	0,2059	0,56	0,3011	25	0,2237	0,68	0,3366	25
CLIMATE-GREENHOUSE-GASES	0,3438	0,3548	0,3492	31	0,4074	0,3548	0,3793	31	0,6316	0,7742	0,6957	31
CLIMATE-HAZARDS	0,5122	0,6176	0,56	34	0,4889	0,6471	0,557	34	0,5476	0,6765	0,6053	34
CLIMATE-IMPACTS	0,2143	0,1765	0,1935	17	0,375	0,3529	0,3636	17	0,2857	0,2353	0,2581	17
CLIMATE-MITIGATIONS	0,25	0,1053	0,1481	38	0,2273	0,1316	0,1667	38	0,3704	0,2632	0,3077	38
CLIMATE-MODELS	0,5301	0,4731	0,5	93	0,5258	0,5426	0,534	94	0,4271	0,4362	0,4316	94
CLIMATE-NATURE	0,2472	0,2245	0,2353	98	0,2475	0,2551	0,2513	98	0,2427	0,2551	0,2488	98
CLIMATE-OBSERVATIONS	0,25	0,0952	0,1379	21	0,2	0,0952	0,129	21	0,3333	0,1429	0,2	21
CLIMATE-ORGANISMS	0,1	0,2222	0,1379	9	0,2143	0,2727	0,24	11	0,1333	0,1818	0,1538	11
CLIMATE-ORGANIZATIONS	0,5833	0,7	0,6364	30	0,5938	0,6333	0,6129	30	0,5882	0,6667	0,625	30
CLIMATE-PROBLEM-ORIGINS	0,3636	0,4	0,381	20	0,3226	0,5	0,3922	20	0,3333	0,45	0,383	20
CLIMATE-PROPERTIES	0,2088	0,2209	0,2147	86	0,2717	0,2907	0,2809	86	0,1915	0,2118	0,2011	85
micro avg	0,3189	0,3339	0,3262	551	0,3455	0,3766	0,3603	555	0,3387	0,3822	0,3591	552
macro avg	0,3106	0,3442	0,31	551	0,3385	0,3812	0,3483	555	0,3555	0,4066	0,3661	552
weighted avg	0,3264	0,3339	0,3197	551	0,3483	0,3766	0,3555	555	0,3455	0,3822	0,355	552

Table 19: Similar k examples: gpt-4o-mini

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-ASSETS	0,2778	0,3409	0,3061	44	0,2973	0,234	0,2619	47	0,3023	0,2708	0,2857	48
CLIMATE-DATASETS	0,1429	0,2917	0,1918	24	0,325	0,52	0,4	25	0,2373	0,56	0,3333	25
CLIMATE-GREENHOUSE-GASES	0,5	0,5862	0,5397	29	0,7027	0,8387	0,7647	31	0,6579	0,8065	0,7246	31
CLIMATE-HAZARDS	0,6111	0,6471	0,6286	34	0,5526	0,6562	0,6	32	0,5833	0,6176	0,6	34
CLIMATE-IMPACTS	0,2609	0,4286	0,3243	14	0,2222	0,25	0,2353	16	0,3	0,3529	0,3243	17
CLIMATE-MITIGATIONS	0,5789	0,3333	0,4231	33	0,5455	0,3243	0,4068	37	0,5556	0,4054	0,4688	37
CLIMATE-MODELS	0,4324	0,5275	0,4752	91	0,4083	0,5506	0,4689	89	0,4298	0,5213	0,4712	94
CLIMATE-NATURE	0,3562	0,2737	0,3095	95	0,3529	0,3158	0,3333	95	0,4364	0,4898	0,4615	98
CLIMATE-OBSERVATIONS	0,375	0,4286	0,4	21	0,4286	0,5	0,4615	18	0,4	0,4762	0,4348	21
CLIMATE-ORGANISMS	0,2143	0,3333	0,2609	9	0,3333	0,4545	0,3846	11	0,1905	0,4	0,2581	10
CLIMATE-ORGANIZATIONS	0,7619	0,5333	0,6275	30	0,7619	0,5517	0,64	29	0,8947	0,5667	0,6939	30
CLIMATE-PROBLEM-ORIGINS	0,0625	0,0714	0,0667	14	0,2083	0,25	0,2273	20	0,2778	0,2778	0,2778	18
CLIMATE-PROPERTIES	0,3158	0,3846	0,3468	78	0,3333	0,3023	0,3171	86	0,3368	0,3765	0,3556	85
micro avg	0,3708	0,4089	0,3889	516	0,4083	0,4235	0,4158	536	0,4144	0,4726	0,4416	548
macro avg	0,3761	0,3985	0,3769	516	0,4209	0,4422	0,4232	536	0,431	0,4709	0,4377	548
weighted avg	0,3974	0,4089	0,395	516	0,4132	0,4235	0,4111	536	0,4385	0,4726	0,4471	548

Table 20: Random k examples: gpt-4o-2024-05-13

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
CLIMATE-ASSETS	0,2308	0,1915	0,2093	47	0,3438	0,2558	0,2933	43	0,225	0,25	0,2368	36
CLIMATE-DATASETS	0,3	0,5455	0,3871	22	0,1852	0,625	0,2857	16	0,3846	0,6522	0,4839	23
CLIMATE-GREENHOUSE-GASES	0,6552	0,7308	0,6909	26	0,6667	0,6667	0,6667	27	0,6667	0,48	0,5581	25
CLIMATE-HAZARDS	0,5128	0,6061	0,5556	33	0,6176	0,7241	0,6667	29	0,6061	0,6452	0,625	31
CLIMATE-IMPACTS	0,3684	0,4118	0,3889	17	0,4706	0,4706	0,4706	17	0,2727	0,3529	0,3077	17
CLIMATE-MITIGATIONS	0,6538	0,4857	0,5574	35	0,4062	0,3714	0,3881	35	0,5926	0,5333	0,5614	30
CLIMATE-MODELS	0,4607	0,5395	0,497	76	0,5444	0,5833	0,5632	84	0,5612	0,679	0,6145	81
CLIMATE-NATURE	0,3226	0,2326	0,2703	86	0,4203	0,3222	0,3648	90	0,4545	0,3409	0,3896	88
CLIMATE-OBSERVATIONS	0,3333	0,3333	0,3333	21	0,2609	0,3	0,2791	20	0,2941	0,2778	0,2857	18
CLIMATE-ORGANISMS	0,2222	0,2222	0,2222	9	0,3571	0,4545	0,4	11	0,4444	0,5	0,4706	8
CLIMATE-ORGANIZATIONS	0,65	0,4815	0,5532	27	0,9	0,4091	0,5625	22	0,8	0,5714	0,6667	28
CLIMATE-PROBLEM-ORIGINS	0,28	0,3684	0,3182	19	0,2581	0,4	0,3137	20	0,2143	0,4	0,2791	15
CLIMATE-PROPERTIES	0,2955	0,3377	0,3152	77	0,3077	0,3733	0,3373	75	0,3377	0,3881	0,3611	67
micro avg	0,3953	0,404	0,3996	495	0,4103	0,4397	0,4245	489	0,4453	0,4711	0,4579	467
macro avg	0,4066	0,422	0,4076	495	0,4414	0,4582	0,4301	489	0,4503	0,467	0,4492	467
weighted avg	0,3998	0,404	0,396	495	0,443	0,4397	0,4298	489	0,4654	0,4711	0,4605	467

Table 21: Similar k examples: gpt-4o-2024-05-13

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
CLIMATE-ASSETS	0,0667	0,1224	0,0863	49	0,0548	0,0909	0,0684	44	0,0694	0,1	0,082	50
CLIMATE-DATASETS	0,0143	0,08	0,0242	25	0,0252	0,12	0,0417	25	0,0259	0,15	0,0441	20
CLIMATE-GREENHOUSE-GASES	0,2558	0,3548	0,2973	31	0,2222	0,3226	0,2632	31	0,186	0,2581	0,2162	31
CLIMATE-HAZARDS	0,1897	0,3235	0,2391	34	0,1455	0,2353	0,1798	34	0,1017	0,1765	0,129	34
CLIMATE-IMPACTS	0	0	0	17	0	0	0	17	0	0	0	17
CLIMATE-MITIGATIONS	0,1351	0,1316	0,1333	38	0,0667	0,0789	0,0723	38	0,14	0,1892	0,1609	37
CLIMATE-MODELS	0,0811	0,1277	0,0992	94	0,0863	0,1348	0,1053	89	0,1092	0,1477	0,1256	88
CLIMATE-NATURE	0,082	0,1531	0,1068	98	0,0741	0,1443	0,0979	97	0,1043	0,2316	0,1438	95
CLIMATE-OBSERVATIONS	0,0926	0,2381	0,1333	21	0,1053	0,1905	0,1356	21	0,0952	0,1905	0,127	21
CLIMATE-ORGANISMS	0,1429	0,4444	0,2162	9	0,0741	0,2222	0,1111	9	0,0698	0,3333	0,1154	9
CLIMATE-ORGANIZATIONS	0,1471	0,1923	0,1667	26	0,1724	0,1667	0,1695	30	0,1613	0,1724	0,1667	29
CLIMATE-PROBLEM-ORIGINS	0,0317	0,1176	0,05	17	0,0172	0,0556	0,0263	18	0,0133	0,05	0,0211	20
CLIMATE-PROPERTIES	0,0762	0,0941	0,0842	85	0,0459	0,0581	0,0513	86	0,1077	0,0833	0,094	84
micro avg	0,0863	0,1581	0,1117	544	0,0761	0,1317	0,0965	539	0,0898	0,157	0,1143	535
macro avg	0,1012	0,1831	0,1259	544	0,0838	0,14	0,1017	539	0,0911	0,1602	0,1097	535
weighted avg	0,0972	0,1581	0,1179	544	0,0827	0,1317	0,0997	539	0,1019	0,157	0,1189	535

Table 22: Random k examples: Mistral-7B-Instruct-v0.3

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
CLIMATE-ASSETS	0,0889	0,2353	0,129	34	0,0541	0,1379	0,0777	29	0,0405	0,1034	0,0583	29
CLIMATE-DATASETS	0,0325	0,3333	0,0593	12	0,0303	0,1765	0,0517	17	0,0208	0,1429	0,0364	14
CLIMATE-GREENHOUSE-GASES	0,1818	0,4444	0,2581	9	0,1	0,2857	0,1481	7	0,2632	0,4545	0,3333	11
CLIMATE-HAZARDS	0,14	0,2333	0,175	30	0,1111	0,1429	0,125	28	0,1351	0,1923	0,1587	26
CLIMATE-IMPACTS	0	0	0	15	0	0	0	12	0	0	0	12
CLIMATE-MITIGATIONS	0,0465	0,0667	0,0548	30	0,1379	0,1333	0,1356	30	0,122	0,1562	0,137	32
CLIMATE-MODELS	0,0816	0,1778	0,1119	45	0,0595	0,125	0,0806	40	0,087	0,1622	0,1132	37
CLIMATE-NATURE	0,089	0,1884	0,1209	69	0,0985	0,2167	0,1354	60	0,0684	0,2167	0,104	60
CLIMATE-OBSERVATIONS	0,0741	0,2	0,1081	20	0,0577	0,1765	0,087	17	0,029	0,1	0,0449	20
CLIMATE-ORGANISMS	0,0588	0,1429	0,0833	7	0,0556	0,1667	0,0833	6	0,0833	0,1429	0,1053	7
CLIMATE-ORGANIZATIONS	0,0833	0,1053	0,093	19	0,1071	0,1429	0,1224	21	0,1053	0,1053	0,1053	19
CLIMATE-PROBLEM-ORIGINS	0,0238	0,0909	0,0377	11	0,08	0,2	0,1143	10	0,1	0,2857	0,1481	14
CLIMATE-PROPERTIES	0,0538	0,1094	0,0722	64	0,0504	0,1	0,067	60	0,0813	0,1667	0,1093	60
micro avg	0,0715	0,1671	0,1002	365	0,0689	0,1484	0,0941	337	0,0728	0,1701	0,1019	341
macro avg	0,0734	0,1791	0,1003	365	0,0725	0,1542	0,0945	337	0,0874	0,1714	0,1118	341
weighted avg	0,0757	0,1671	0,102	365	0,0763	0,1484	0,0973	337	0,0837	0,1701	0,1087	341

Table 23: Similar k examples: Mistral-7B-Instruct-v0.3

Appendix E: Detailed performance report: BiodivNER

	gpt-4o-mini				gpt-4o-2024-05-13				Mistral-7B-Instruct-v0.3			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
ENVIRONMENT	0,2	0,2078	0,2038	154	0,322	0,3851	0,3508	148	0,0361	0,1805	0,0602	133
LOCATION	0,1667	0,3438	0,2245	32	0,2381	0,4688	0,3158	32	0,0244	0,1071	0,0397	28
micro avg	0,1903	0,2312	0,2087	186	0,3	0,4	0,3429	180	0,0343	0,1677	0,057	161
macro avg	0,1833	0,2758	0,2142	186	0,2801	0,4269	0,3333	180	0,0303	0,1438	0,05	161
weighted avg	0,1943	0,2312	0,2074	186	0,3071	0,4	0,3446	180	0,0341	0,1677	0,0567	161
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
MATTER	0,0912	0,4333	0,1507	60	0,3436	0,3448	0,3442	290	0,0385	0,2549	0,0668	51
ORGANISM	0,2746	0,5571	0,3679	280	0,3788	0,6774	0,4859	279	0,0739	0,1923	0,1068	208
micro avg	0,2134	0,5353	0,3051	340	0,3658	0,5079	0,4253	569	0,0603	0,2046	0,0931	259
macro avg	0,1829	0,4952	0,2593	340	0,3612	0,5111	0,415	569	0,0562	0,2236	0,0868	259
weighted avg	0,2423	0,5353	0,3296	340	0,3609	0,5079	0,4137	569	0,067	0,2046	0,0989	259
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
PHENOMENA	0,1239	0,58	0,2042	50	0,2	0,4839	0,283	62	0,015	0,2105	0,0279	38
QUALITY	0,1958	0,4784	0,2778	232	0,3251	0,4458	0,376	415	0,0669	0,1923	0,0993	208
micro avg	0,1748	0,4965	0,2585	282	0,299	0,4507	0,3595	477	0,0424	0,1951	0,0696	246
macro avg	0,1598	0,5292	0,241	282	0,2626	0,4648	0,3295	477	0,0409	0,2014	0,0636	246
weighted avg	0,183	0,4965	0,2648	282	0,3089	0,4507	0,3639	477	0,0589	0,1951	0,0882	246

Table 24: NE cluster classes: All models

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
ENVIRONMENT	0,3097	0,2273	0,2622	154	0,3438	0,2857	0,3121	154	0,32	0,2597	0,2867	154
LOCATION	0,1176	0,4375	0,1854	32	0,1111	0,4375	0,1772	32	0,112	0,4375	0,1783	32
MATTER	0,0742	0,2833	0,1176	60	0,1111	0,4	0,1739	60	0,1053	0,3667	0,1636	60
ORGANISM	0,3213	0,4786	0,3845	280	0,342	0,5143	0,4108	280	0,3704	0,5185	0,4321	270
PHENOMENA	0,1469	0,6	0,2361	60	0,1706	0,6	0,2657	60	0,164	0,5167	0,249	60
QUALITY	0,3208	0,1823	0,2325	373	0,3046	0,1622	0,2116	370	0,3152	0,1686	0,2197	344
micro avg	0,2277	0,317	0,265	959	0,2479	0,3368	0,2856	956	0,2521	0,3315	0,2864	920
macro avg	0,2151	0,3682	0,2364	959	0,2305	0,3999	0,2586	956	0,2311	0,3779	0,2549	920
weighted avg	0,2861	0,317	0,2731	959	0,2948	0,3368	0,286	956	0,3152	0,3315	0,2864	920

Table 25: Random k examples: gpt-4o-mini

	k=3				k=4				k=5			
Category	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
ENVIRONMENT	0,3821	0,3052	0,3394	154	0,4331	0,3571	0,3915	154	0,4219	0,3649	0,3913	148
LOCATION	0,1546	0,4688	0,2326	32	0,1977	0,5312	0,2881	32	0,2043	0,6129	0,3065	31
MATTER	0,3243	0,4	0,3582	60	0,3194	0,3833	0,3485	60	0,2838	0,35	0,3134	60
ORGANISM	0,4373	0,5107	0,4712	280	0,4108	0,4607	0,4343	280	0,4369	0,5315	0,4796	254
PHENOMENA	0,2263	0,5167	0,3147	60	0,2388	0,5333	0,3299	60	0,25	0,5254	0,3388	59
QUALITY	0,4516	0,3753	0,41	373	0,4479	0,3807	0,4116	373	0,4521	0,4048	0,4272	373
micro avg	0,3745	0,4171	0,3947	959	0,379	0,415	0,3962	959	0,387	0,4443	0,4137	925
macro avg	0,3294	0,4294	0,3543	959	0,3413	0,4411	0,3673	959	0,3415	0,4649	0,3761	925
weighted avg	0,4043	0,4171	0,4014	959	0,4052	0,415	0,4018	959	0,4052	0,4443	0,4137	925

Table 26: Similar k examples: gpt-4o-mini

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
ENVIRONMENT	0,3039	0,4133	0,3503	150	0,3284	0,4371	0,375	151	0,3073	0,4286	0,358	147
LOCATION	0,186	0,5	0,2712	32	0,1633	0,5	0,2462	32	0,2027	0,5357	0,2941	28
MATTER	0,1805	0,0839	0,1146	286	0,2583	0,108	0,1523	287	0,1552	0,3273	0,2105	55
ORGANISM	0,4501	0,64	0,5285	275	0,4455	0,7111	0,5478	270	0,4672	0,6404	0,5403	267
PHENOMENA	0,199	0,619	0,3012	63	0,2151	0,5968	0,3162	62	0,2096	0,6604	0,3182	53
QUALITY	0,375	0,1838	0,2467	457	0,3088	0,1473	0,1994	455	0,373	0,1927	0,2541	358
micro avg	0,325	0,3175	0,3212	1263	0,3301	0,3254	0,3277	1257	0,3333	0,4086	0,3671	908
macro avg	0,2824	0,4067	0,3021	1263	0,2866	0,4167	0,3062	1257	0,2858	0,4642	0,3292	908
weighted avg	0,3253	0,3175	0,2938	1263	0,3206	0,3254	0,2915	1257	0,3621	0,4086	0,3574	908

Table 27: Random k examples: gpt-4o-2024-05-13

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
ENVIRONMENT	0,3795	0,4228	0,4	149	0,3688	0,4214	0,3933	140	0,3818	0,4437	0,4104	142
LOCATION	0,2653	0,65	0,3768	20	0,3115	0,5938	0,4086	32	0,2321	0,7222	0,3514	18
MATTER	0,36	0,1237	0,1841	291	0,5879	0,3794	0,4612	282	0,4179	0,0969	0,1573	289
ORGANISM	0,539	0,5961	0,5661	255	0,542	0,6102	0,5741	254	0,4881	0,5913	0,5348	208
PHENOMENA	0,2463	0,5238	0,335	63	0,2984	0,6066	0,4	61	0,246	0,5345	0,337	58
QUALITY	0,5257	0,3937	0,4502	442	0,5013	0,4948	0,498	384	0,4923	0,4167	0,4513	384
micro avg	0,4435	0,3861	0,4128	1220	0,4757	0,4918	0,4836	1153	0,4218	0,3803	0,4	1099
macro avg	0,386	0,4517	0,3854	1220	0,435	0,5177	0,4559	1153	0,3764	0,4675	0,3737	1099
weighted avg	0,4524	0,3861	0,3977	1220	0,4993	0,4918	0,4854	1153	0,4404	0,3803	0,3768	1099

Table 28: Similar k examples: gpt-4o-2024-05-13

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
ENVIRONMENT	0,0853	0,2273	0,1241	110	0,0919	0,2451	0,1337	102	0,0827	0,1858	0,1144	113
LOCATION	0,0595	0,2273	0,0943	22	0,0625	0,2381	0,099	21	0,0377	0,1053	0,0556	19
MATTER	0,1019	0,2115	0,1375	52	0,0694	0,2	0,1031	50	0,0885	0,1887	0,1205	53
ORGANISM	0,1672	0,2938	0,2131	194	0,1613	0,2535	0,1971	217	0,1929	0,2941	0,233	204
PHENOMENA	0,0505	0,25	0,084	40	0,04	0,2703	0,0697	37	0,0637	0,3333	0,107	39
QUALITY	0,1054	0,1804	0,1331	194	0,112	0,1116	0,1118	242	0,1	0,1277	0,1121	188
micro avg	0,1055	0,2337	0,1453	612	0,0994	0,1973	0,1322	669	0,1106	0,211	0,1452	616
macro avg	0,095	0,2317	0,131	612	0,0895	0,2197	0,1191	669	0,0943	0,2058	0,1238	616
weighted avg	0,1158	0,2337	0,1526	612	0,1162	0,1973	0,1394	669	0,1224	0,211	0,1512	616

Table 29: Random k examples: Mistral-7B-Instruct-v0.3

	k=3				k=4				k=5			
Category	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
ENVIRONMENT	0,1062	0,2051	0,1399	117	0,1121	0,2119	0,1466	118	0,1324	0,2458	0,1721	118
LOCATION	0,098	0,1923	0,1299	26	0,1	0,1923	0,1316	26	0,0784	0,16	0,1053	25
MATTER	0,0984	0,1132	0,1053	53	0,1639	0,1887	0,1754	53	0,25	0,26	0,2549	50
ORGANISM	0,1841	0,2576	0,2147	198	0,2069	0,2652	0,2324	181	0,2148	0,2792	0,2428	197
PHENOMENA	0,0729	0,3333	0,1197	42	0,075	0,3	0,12	40	0,0811	0,3243	0,1297	37
QUALITY	0,1871	0,2709	0,2213	203	0,2415	0,3333	0,2801	192	0,2337	0,3617	0,2839	188
micro avg	0,1408	0,2426	0,1782	639	0,1655	0,2689	0,2049	610	0,178	0,2943	0,2218	615
macro avg	0,1245	0,2287	0,1551	639	0,1499	0,2486	0,181	610	0,1651	0,2718	0,1981	615
weighted avg	0,1529	0,2426	0,1844	639	0,1825	0,2689	0,2142	610	0,1941	0,2943	0,2304	615

Table 30: Similar k examples: Mistral-7B-Instruct-v0.3