Large Language Models for Monitoring Dataset Mentions in Climate Research

Aivin V. Solatorio *, **Rafael Macalaba** †, **& James Liounis** ‡
Development Data Group and Office of the Chief Statistician
The World Bank
1818 H Street N.W.,
Washington, 20433
District of Columbia, USA
{asolatorio,rmacalaba,jliounis}@worldbank.org

Abstract

Effective climate change research relies on diverse datasets to inform mitigation and adaptation strategies and policies. However, the ways these datasets are cited, used, and distributed remain poorly understood. This paper presents a machine learning framework that automates the detection and classification of dataset mentions in climate research papers. Leveraging large language models (LLMs), we generate a weakly supervised dataset through zero-shot extraction, quality assessment via an LLM-as-a-Judge, and refinement by a reasoning agent. The Phi-3.5-mini instruct model is pre-fine-tuned on this dataset, followed by fine-tuning on a smaller manually annotated subset to specialize in extracting data mentions. At inference, a ModernBERT-based classifier filters for dataset mentions, optimizing computational efficiency. Evaluated on a held-out manually annotated sample, our fine-tuned model outperforms NuExtract-v1.5 and GLiNER-large-v2.1 in dataset extraction accuracy. As a framework for monitoring dataset mentions in research papers, this approach enhances transparency, identifies data gaps, and enables researchers, funders, and policymakers to improve data discoverability and usage, leading to more informed decision-making.

1 Introduction

Climate change research depends on vast amounts of data to monitor environmental shifts, model future scenarios, and inform policy decisions (Meehl et al., 2007; Camarillo-Naranjo et al., 2019; Hasegawa et al., 2022). Yet, despite the growing reliance on data-driven approaches, a critical gap persists: we lack a systematic way to track when and what datasets are mentioned in scientific literature. This oversight makes it difficult to assess how data informs research, whether certain datasets disproportionately shape findings, and where gaps in data utilization might exist (Georgeson et al., 2017). Without these insights, biases in climate science may go undetected, and key datasets—particularly those from underrepresented regions or disciplines—could remain underutilized, ultimately limiting the effectiveness of climate action Parsons et al. (2024).

Understanding patterns of data usage is essential given the interdisciplinary nature of climate science, as in climate-related projects (Vine & Sathaye, 1999). Research in this field integrates data from atmospheric studies, oceanography, agriculture, public health, socioeconomics, and more. The datasets used by researchers directly influence the scope of their analyses and the solutions they propose. By systematically detecting dataset mentions in research papers, we can uncover trends in data reliance, identify imbalances in research focus, and highlight underused datasets that could enrich our understanding of climate challenges

*GitHub/HF: @avsolatorio, avsolatorio@gmail.com
†GitHub: @rafmacalaba, rafael.macalaba@yahoo.com
‡GitHub: @jamesliounis, liounisjames@gmail.com

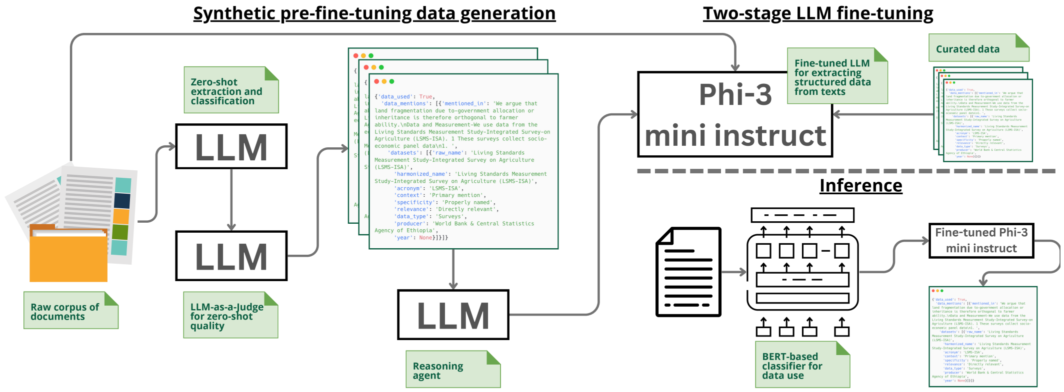

Figure 1: Diagram for the proposed data use extraction and classification framework. The figure shows the high-level process for generating the synthetic pre-fine-tuning data, the two-stage fine-tuning of the Phi-3.5-mini instruct model, and the inference stage where ModernBERT is used to distinguish which texts likely have mentions of data and require fine-tuned LLM processing.

(Osuteye et al., 2016). This process can also expose blind spots in scientific discourse, where reliance on a narrow set of datasets may inadvertently reinforce existing knowledge gaps or policy blind spots.

Beyond research biases, disparities in data accessibility remain a pressing issue (O'Malley et al., 2009). Some research communities benefit from extensive, well-curated datasets, while others—especially in low-income regions—face barriers to accessing critical climate data. The ability to track dataset mentions in scientific literature can shed light on these disparities, providing empirical evidence to support more equitable data-sharing initiatives (Stacy et al., 2024). It can also help funders and policymakers identify underutilized but valuable datasets, ensuring that resources are directed toward improving data availability where it is needed most. This, in turn, strengthens global efforts to address climate change by promoting a more inclusive and representative use of data.

Moreover, tracking dataset mentions enhances research transparency and reproducibility—essential factors for trust in scientific findings Hussain et al. (2023). Policymakers and practitioners who rely on climate studies for high-stakes decisions often lack visibility into the underlying data sources. A clearer understanding of dataset usage patterns can improve trust in scientific conclusions, enhance data discoverability, and support targeted investments in high-impact datasets (Stacy et al., 2024). Establishing structured methods for monitoring dataset citations could also encourage better data documentation and citation practices within the research community.

Advancements in artificial intelligence (AI) and natural language processing (NLP) provide new opportunities to systematically extract and classify dataset mentions in scientific literature (Potok, 2022; Hussain et al., 2023; Younes & Scherp, 2023). This paper presents insights from an ongoing study exploring a machine learning (ML) framework that combines large language models (LLMs) with specialized classifiers to detect and categorize dataset references in climate research. Given the lack of labeled training data, we generate a weakly supervised dataset using a sequence of LLM-based extraction methods applied to research papers, followed by manual annotation of a subset. Our results demonstrate the feasibility of this approach, offering a first step toward analyzing dataset citations and their role in climate research.

2 Methods

This section outlines our methodology for monitoring data usage in climate research papers. We address the lack of training datasets for extracting and classifying dataset mentions and then develop and fine-tune models specialized for classification and extraction, outlined in Figure 1.

2.1 Data

To implement and test the proposed framework, we compile a collection of climate-related papers. This collection includes papers identified in Sietsma et al. (2024) (subsequently referred to as the One Earth corpus) and climate-related research from the World Bank's Policy Research Working Papers (PRWP). The PRWP series is a diverse collection of research papers addressing various topics in socio-economic development, with climate change being one of the themes. We use the World Bank's Documents and Reports platform to filter documents related to climate change using a set of tags that pertain to climate change. Since the process requires full-text access, we utilize the Semantic Scholar Paper title search API to identify the open-access papers and retrieve their PDF links. Applying this process to both collections resulted in 2,123 papers and 582 papers with PDFs for the One Earth corpus and the PRWP, respectively.

2.2 Weakly Supervised Data Generation

The absence of publicly available datasets tailored to classify dataset mentions by their purpose and citation quality presents a significant bottleneck in developing reliable machine learning models for this task. While the Coleridge Initiative's Show US the Data dataset provides examples of dataset mentions, it suffers from two key limitations: (1) it lacks information about the context and purpose of dataset usage, and (2) it primarily focuses on a limited subset of datasets, introducing a systematic bias against the broader range of datasets mentioned in climate research (Potok, 2022).

We address this gap by harnessing the extractive, classification, and pseudo-reasoning capabilities of LLMs. Our method leverages LLMs to generate a weakly supervised synthetic fine-tuning dataset derived from our corpus, from which we strategically sample for manual curation, enhancing efficiency while ensuring data quality.

Zero-shot extraction and classification. To create a weakly supervised pre-fine-tuning dataset, each page of a research paper in our corpus is processed by an LLM that extracts structured information. The output indicates whether a dataset is referenced in the text and provides details such as the dataset name and its classification based on the quality of the name and the usage context.

LLM-as-a-Judge for quality. A manual review of the extracted outputs revealed instances where non-dataset entities were misclassified as mentions of data. We integrate an LLM-as-a-Judge mechanism to address this, where a second LLM evaluates the extracted information for accuracy, ensuring more reliable dataset classification.

Autonomous reasoning for filtering. Further manual checks showed that the LLM judge often misclassifies reports, frameworks, and organizations as datasets, failing to filter them out. To improve classification accuracy, we introduced a reasoning agent that autonomously develops and executes a structured evaluation strategy, including a "devil's advocate" argument to reassess its conclusions. This step-by-step self-evaluation process reduces reliance on implicit assumptions and enhances classification rigor. The LLM retains autonomy in reclassifying previously validated datasets, provided it justifies its decisions. As a result, the reasoning agent filtered out approximately 42% of the 37,225 mentions initially shortlisted by the LLM judge, identifying only 21,408 as likely valid.

All LLM-based methods use the OpenAI's gpt-4o-mini (2024-07-18) model. The prompts for these processes are provided in A.5.1, A.5.2, A.5.3 for the zero-shot, LLM-as-a-Judge, and reasoning agent methods, respectively.

2.3 Fine-tuning Data

We sampled 1,000 pages from the output of the previous method and manually annotated them to remove any remaining false positives from the weakly supervised approach. This

Table 1: Performance of Classification and Extraction Models for Data Use

Data Use Classification Models			
Model	**Precision**	**Recall**	**F1-score**
BERT-uncased	**100.0**	50.0	67.0
ModernBERT-base	**100.0**	**100.0**	**100.0**
Data Use Extraction Models			
Model [Training data]	**Precision**	**Recall**	**Fβ-score**
Phi-3-mini [Synthetic and curated]	**69.45**	**80.65**	**71.43**
Phi-3-mini [Synthetic only]	60.00	70.00	61.76
Phi-3-mini [Curated only]	55.68	65.52	57.58
GLiNER-large-v2.1	62.50	71.43	64.10
NuExtract-v1.5	20.97	46.43	23.55

curated dataset serves as a foundational resource for training and evaluating downstream models specialized in extracting and classifying dataset mentions offline.

2.4 Fine-tuning Models

Detecting Data Use. We fine-tune both BERT (Devlin et al., 2019) and ModernBERT (Warner et al., 2024) models for detecting dataset mentions in texts and compare their performance. The goal is to offload the filtering stage to a more efficient encoder-based model, reducing computational overhead when processing large volumes of text. Since dataset mentions are typically sparse within documents, only texts identified as containing data mentions by the encoder models are passed to the LLM for further processing.

Extracting and classifying data mentions. We fine-tune the Phi-3.5-mini instruct model (Abdin et al., 2024) using 16-rank LoRA (Hu et al., 2021) designed to extract structured information about dataset mentions in text. We first use the large weakly supervised dataset to pre-fine-tune the model for 3 epochs (LR=2e-4). We then take the smaller manually annotated dataset to fine-tune over 20 epochs (LR=3e-5), with checkpoints created at the end of each epoch. This fine-tuning process enables the model to “specialize” in identifying key attributes of dataset mentions, such as their contextual usage and the specificity of their naming (e.g., whether the dataset is explicitly named, descriptively referenced, or vaguely mentioned). To evaluate our approach, we compare performance against NuExtract-v1.5 (Cripwell et al.) and GLiNER-large-v2.1 (Zaratiana et al., 2023), two state-of-the-art models for named entity recognition and zero-shot structured data extraction, serving as baselines.

3 Results

To measure the extraction accuracy, we use the Jaccard Similarity-based (Equation 1) Fβ Score, a metric introduced in the Coleridge Initiative’s data extraction Kaggle competition (Gupta, 2021). This method evaluates the overlap between predicted and ground-truth dataset mentions, allowing for partial matches rather than requiring exact string matches—an important consideration given the variability in how datasets are referenced in text and how the models select which snippets constitute dataset names.

$$J(S_1, S_2) = \frac{|W_1 \cap W_2|}{|W_1| + |W_2| - |W_1 \cap W_2|} \tag{1}$$

In Equation equation 1, $J(S_1, S_2)$ represents the Jaccard similarity between two strings S_1 and S_2. The sets W_1 and W_2 contain the unique words obtained from tokenizing S_1 and S_2, respectively. The term $|W_1 \cap W_2|$ denotes the number of words common to both sets, while $|W_1|$ and $|W_2|$ represent the total number of unique words in each set. The denominator, $|W_1|+|W_2|-|W_1 \cap W_2|$, ensures that words shared between both sets are not double-counted, yielding a similarity score between 0 and 1. In this context, a Jaccard score greater than 0.5 is considered a match. Based on this classification, the precision, recall, and Fβ-score are computed to evaluate performance.

Our results show that fine-tuning the Phi-3-mini instruct model significantly improves extraction accuracy, outperforming NuExtract-v1.5 and GLiNER-large-v2.1, Table 1. Among the different fine-tuning strategies:

- The Phi-3-mini model trained on both synthetic and curated data achieves the highest Fβ score (71.43), demonstrating the effectiveness of a two-stage fine-tuning approach: pre-fine-tuning with a larger synthetic dataset followed by refinement on a smaller, manually curated sample. This combination improves generalization while enhancing precision.
- The synthetic-only model outperforms the curated-only model, achieving higher recall (70.00 vs. 65.52), suggesting that synthetic data, despite lacking human-verified precision, contributes to broader coverage and generalization. However, the curated-only model offers better precision, reinforcing the importance of human-verified refinements.
- The baseline NuExtract-v1.5 model performs significantly worse, while GLiNER-large-v2.1 achieves comparable results to the LLM pre-fine-tuned exclusively on synthetic data. However, both underperform relative to the two-stage fine-tuned LLM, underscoring the advantages of domain-adapted fine-tuning.

These findings highlight the benefits of leveraging synthetic data for broad generalization while using curated data to refine accuracy. The 9.67-point improvement in Fβ score with curated fine-tuning demonstrates its critical role in enhancing dataset mention extraction. This also suggests that further optimizing the pre-fine-tuned LLM with topic-specific annotated data—such as curated dataset mentions in disaster management, refugee and forced displacement, or labor markets—could enhance its adaptability to specialized domains.

Our ablation study further confirms that pre-fine-tuning is crucial when only a small volume of annotated data is available. Conditioning the model with synthetic data mitigates overfitting to limited human annotations and enhances recall for unseen dataset mentions. Notably, the fine-tuned model achieves a high recall, which is desirable if we aim to discover as much data as mentioned in research papers as possible. The same table also reports the performance of the ModernBERT and the BERT models for classifying whether a dataset is likely to be mentioned in a text. We find that ModernBERT shows better classification performance. During training, we observed that the BERT-based model struggled to learn, likely due to its limited context size of 512 tokens compared to the 2048 tokens we used in the ModernBERT model. Additionally, we provide examples of extracted data alongside empirical dataset mentions in Annex Table 3.

4 Conclusion

This study presents a machine learning framework for detecting, classifying, and extracting dataset mentions in climate research, addressing a critical gap in understanding how data informs scientific inquiry. By leveraging LLM-based weak supervision, fine-tuning on manually annotated data, and employing an efficient inference strategy, our approach improves dataset extraction accuracy while optimizing computational resources. The fine-tuned model outperforms NuExtract-v1.5 and GLiNER-large-v2.1, demonstrating its effectiveness in extracting dataset references. Beyond the technical aspects, this work lays the foundation for greater transparency, accessibility, and equity in climate data usage. Future efforts will focus on scaling the framework, refining extraction accuracy, and broadening its applicability to ensure more comprehensive insights into dataset usage in climate research and beyond.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, August 2024. URL http://arxiv.org/abs/2404.14219. arXiv:2404.14219 [cs].

Juan Mariano Camarillo-Naranjo, José Ignacio Álvarez Francoso, Natalia Limones-Rodríguez, María Fernanda Pita-López, and Mónica Aguilar-Alba. The global climate monitor system: from climate data-handling to knowledge dissemination. *International Journal of Digital Earth*, 12(4):394–414, April 2019. ISSN 1753-8947. doi: 10.1080/17538947.2018.1429502. URL https://doi.org/10.1080/17538947.2018.1429502. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17538947.2018.1429502.

L. Cripwell et al. NuExtract 1.5 - Multilingual, Infinite context, still small, and better than GPT-4o! - NuMind. URL https://www.numind.ai/blog/nuextract-1-5---multilingual-infinite-context-still-small-and-better-than-gpt-4o,.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Lucien Georgeson, Mark Maslin, and Martyn Poessinouw. Global disparity in the supply of commercial weather and climate information services. *Science Advances*, 3(5):e1602632, May 2017. doi: 10.1126/sciadv.1602632. URL https://www.science.org/doi/full/10.1126/sciadv.1602632. Publisher: American Association for the Advancement of Science.

Pashupati Gupta. Evaluation metric implementation and demo, 2021. URL https://kaggle.com/code/pashupatigupta/evaluation-metric-implementation-and-demo.

Toshihiro Hasegawa, Hitomi Wakatsuki, Hui Ju, Shalika Vyas, Gerald C. Nelson, Aidan Farrell, Delphine Deryng, Francisco Meza, and David Makowski. A global dataset for the projected impacts of climate change on four major crops. *Scientific Data*, 9(1):58, February 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01150-7. URL https://www.nature.com/articles/s41597-022-01150-7. Publisher: Nature Publishing Group.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685. arXiv:2106.09685 [cs].

Tayyaba Hussain, Muhammad Usman Akram, and Anum Abdul Salam. A Novel Data Extraction Framework Using Natural Language Processing (DEFNLP) Techniques, July 2023. URL https://papers.ssrn.com/abstract=4524837.

Gerald A. Meehl, Curt Covey, Thomas Delworth, Mojib Latif, Bryant McAvaney, John F. B. Mitchell, Ronald J. Stouffer, and Karl E. Taylor. THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. September 2007. doi: 10.1175/BAMS-88-9-1383. URL https://journals.ametsoc.org/view/journals/bams/88/9/bams-88-9-1383.xml. Section: Bulletin of the American Meteorological Society.

Robin O'Malley, Anne S. Marsh, and Christine Negra. Closing the Environmental Data Gap. *Issues in Science and Technology*, 25(3):69–74, 2009. ISSN 0748-5492. URL https://www.jstor.org/stable/43314948. Publisher: University of Texas at Dallas.

E. Osuteye, C. Johnson, and D. Brown. The data gap: An analysis of data availability on disaster losses in sub-Saharan African Cities, July 2016. URL https://www.urbanark.org/data-gap-analysis-data-availability-disaster-losses-sub-saharan-african-cities. Issue: 11 Num Pages: 41 Number: 11 Pages: 1-41 Publisher: Urban Africa Risk Knowledge (Urban ARK).

Ella Sandrine Parsons, Ashley Jowell, Erika Veidis, Michele Barry, and Sonoo Thadaney Israni. Climate change and inequality. *Pediatric Research*, pp. 1–8, June 2024. ISSN 1530-0447. doi: 10.1038/s41390-024-03153-z. URL https://www.nature.com/articles/s41390-024-03153-z. Publisher: Nature Publishing Group.

Nancy Potok. Show us the data. *Harvard Data Science Review*, 4(2), 2022.

Anne J. Sietsma, Emily Theokritoff, Robbert Biesbroek, Iván Villaverde Canosa, Adelle Thomas, Max Callaghan, Jan C. Minx, and James D. Ford. Machine learning evidence map reveals global differences in adaptation action. *One Earth*, 7(2):280–292, February 2024. ISSN 2590-3322. doi: 10.1016/j.oneear.2023.12.011. URL https://www.sciencedirect.com/science/article/pii/S2590332223005663.

Brian William Stacy, Lucas Kitzmüller, Xiaoyu Wang, Daniel Gerszon Mahler, and Umar Serajuddin. Missing Evidence : Tracking Academic Data Use around the World. Policy Research Working Paper Series WPS 10673, World Bank Group, January 2024. URL http://documents.worldbank.org/curated/en/099503001172493286.

Edward Vine and Jayant Sathaye. The Monitoring, Evaluation, Reporting and Verification of Climate Change Projects. *Mitigation and Adaptation Strategies for Global Change*, 4 (1):43–60, March 1999. ISSN 1573-1596. doi: 10.1023/A:1009651316596. URL https://doi.org/10.1023/A:1009651316596.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024. URL http://arxiv.org/abs/2412.13663. arXiv:2412.13663 [cs].

Yousef Younes and Ansgar Scherp. Question Answering Versus Named Entity Recognition for Extracting Unknown Datasets. *IEEE Access*, 11:92775–92787, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3309148. URL https://ieeexplore.ieee.org/abstract/document/10231147. Conference Name: IEEE Access.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, November 2023. URL http://arxiv.org/abs/2311.08526. arXiv:2311.08526 [cs].

Acknowledgments

This work is supported by the "KCP IV - Exploring Data Use in the Development Economics Literature using Large Language Models (AI and LLMs)" project funded by the Knowledge for Change Program (KCP) of the World Bank - RA-P503405-RESE-TF0C3444.

Disclaimer and Disclosure of AI Use

The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

This work used AI tools at various stages, including generating synthetic data and reasoning using the gpt-4o-mini model (API) and open source AI models. In addition, Microsoft Co-Pilot and ChatGPT were employed to enhance the manuscript’s readability.

A Appendix

A.1 Data Sources

The datasets/corpus used in this study are derived from:

- **One Earth corpus source:** The dataset where the One Earth corpus was derived from was obtained from Zenodo (`https://zenodo.org/records/7893023`). The dataset was introduced in (Sietsma et al., 2024), which provides context and methodological details regarding its creation.
- **PRWP corpus source:** To identify Policy Research Working Papers (PRWPs) relevant to climate change, a structured filtering approach was applied using the *World Bank Documents and Reports* portal (source).
  - The query parameters in the source URL indicate the applied filters, ensuring that only documents meeting specific criteria were selected. The filtering criteria included:
    - **Document Type:** Policy Research Working Papers (PRWP)
    - **Query:** climate change
    - **Language:** English
    - **Selected Topics:**
      - Climate Change and Agriculture
      - Adaptation to Climate Change
      - Climate Change and Environment
      - Climate Change Impacts
      - Climate Change Mitigation and Greenhouse Gases
      - Climate Change and Health
      - Climate Change Economics
      - Investment and Investment Climate
      - Climate Change Policy and Regulation
      - Climate and Meteorology
      - Science of Climate Change
      - Social Aspects of Climate Change

The process for building the corpus involved ensuring a PDF is available for the paper. This requires the following approach: metadata retrieval, validating for open access, and downloading of the PDFs.

We retrieve the metadata via Semantic Scholar:

- The selected paper titles were queried in *Semantic Scholar* using the Paper Title Search API (`https://api.semanticscholar.org/graph/v1/paper/search/match`). This allowed the retrieval of structured metadata, including authorship details, publication year, abstracts, and citation counts, and, importantly, a flag indicating if the resource is open access.
- If available, PDFs of the identified papers were downloaded for further analysis.

A.2 Training parameters

Parameter	Pre-fine-tuning on Synthetic Data	Fine-tuning on High-quality Annotated Data
Dataset Type	Large weakly supervised synthetic	Small manually annotated
Epochs	10	20
Effective Batch Size	16	2
Learning Rate	2e-4	3e-5
Warmup Ratio	1%	1% (same as pre-fine-tuning)
Scheduler	Linear with decay 0.01	Linear with decay 0.01 (same)
Checkpoint Frequency	Every 100 steps	Every 50 steps and at end of each epoch
Model Selection Criterion	Best validation loss	Best evaluation loss

Table 2: Training configurations for pre-fine-tuning and fine-tuning phases.

A.3 Templates and classifications

Below is the JSON template used for extracting data mentions using the NuExtract-v1.5 model.

Listing 1 JSON template for the NuExtract-v1.5 model.

```
NU_TEMPLATE = {
  "data_mentions": [
    {
      "mentioned_in": "",
      "datasets": [
        {
          "raw_name": "",
          "acronym": ""
        }
      ]
    }
  ]
}
```

A.4 Listing of empirical and extracted data mentions

Table 3: Empirical and Predicted Datasets

Empirical	**Predicted**
Africa Rainfall and Temperature Evaluation System (ARTES) Soil data Hydrology data from the University of Colorado	Africa Rainfall and Temperature Evaluation System (ARTES) Soil data from FAO Data concerning hydrology from the University of Colorado
India's quinquennial labor force survey 30-year agricultural wage series for Indian districts Wholesale crop price data	Domestic crop price data Crop price data
Balanced-panel of 2,382 households	Baseline survey
Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages (EAC-I) Fourth-General Census of Population and Housing (2009) Meteorological data	Mali's Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages (EAC-I) Fourth General Census of Population and Housing (2009)
Household survey data Republic of Uganda 2005	-
Shock modules	Shock modules
2005 SAM for Ghana	2005 SAM for Ghana
Agroalimentary and Fisheries Information Service (SIAP) Coupled Model Intercomparison Project Phase 3 (CMIP3) Income and Expenditure Household Survey (ENIGH) Count of Population and Housing 2005 2007 Agricultural Census	Agroalimentary and Fisheries Information Service (SIAP) National Weather Service (SMN) National Water Commission (CONAGUA) Income and Expenditure Household Survey (ENIGH) Count of Population and Housing 2005 Summary Statistics of the 2007 Agricultural Census (INEGI)
Africa Rainfall and Temperature Evaluation System (ARTES) Soil data Hydrology data from the University of Colorado	Africa Rainfall and Temperature Evaluation System (ARTES) Soil data Hydrology data
Climate data from the 18 meteorological stations of highest quality in Bolivia from May 1948 to May 2008	Climate data from the 18 meteorological stations of highest quality in Bolivia
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	Toxic Release Inventory (TRI)

A.5 Prompts

A.5.1 Zero-shot extraction and classification prompt

Listing 2 System prompt used to extract the initial structured data containing likely data mentions from a given text [1/3].

```
You are an expert in extracting and categorizing dataset mentions from
research papers and policy documents. Your task is to **identify and
extract all valid dataset mentions**, ensuring they are correctly
classified based on naming specificity, context, and relevance.

### **What Qualifies as a Dataset?**
A dataset is a structured collection of data used for empirical
research, analysis, or policy-making. Examples include:
- **Surveys & Census Data** (e.g., LSMS, DHS, national census records)
- **Indicators & Indexes** (e.g., HDI, GFSI, WDI, ND-GAIN, EPI)
- **Geospatial & Environmental Data** (e.g., OpenStreetMap, Sentinel-2
imagery)
- **Economic & Trade Data** (e.g., UN Comtrade, Balance of Payments
Statistics)
- **Health & Public Safety Data** (e.g., epidemiological surveillance,
crime reports)
- **Time-Series & Energy Data** (e.g., climate projections, electricity
demand records)
- **Transport & Mobility Data** (e.g., road accident statistics, smart
city traffic flow)
- **Other emerging dataset types** as identified in the text.

**Important:**
If the dataset does not fit into the examples above, infer the **most
appropriate category** from the context and **create a new
`"data_type"` if necessary**.

### **What Should NOT Be Extracted?**
Do **not** extract mentions that do not clearly refer to a dataset,
including, but not limited to:
1. **Organizations & Institutions** (e.g., WHO, IMF, UNDP, "World Bank
data" unless it explicitly refers to a dataset)
2. **Reports & Policy Documents** (e.g., "Fiscal Monitor by the IMF",
"IEA Energy Report"; only extract if the dataset itself is referenced)
3. **Generic Mentions of Data** (e.g., "various sources", "survey
results from multiple institutions")
4. **Economic Models & Policy Frameworks** (e.g., "GDP growth
projections", "macroeconomic forecasts")
5. **Legislation & Agreements** (e.g., "Paris Agreement", "General Data
Protection Regulation")
```

Listing 3 [Continued] System prompt used to extract the initial structured data containing likely data mentions from a given text [2/3].

```
### **Rules for Extraction**
1. **Extract All Structured Data Mentions**
   - If the dataset is explicitly named (e.g., "Global Fishing Watch"),
   label it as `"properly_named"`.
   - If the dataset is described but not explicitly named (e.g.,
   "electricity usage data from Albania"), label it as
   `"descriptive_but_unnamed"`.
   - If the dataset mention is too generic (e.g., "electricity usage
   data"), label it as `"vague_generic"`.

2. **Ensure `"data_type"` Is Always Assigned**
   - **Use an existing category if applicable.**
   - **If no suitable category exists, create a new `"data_type"` based
   on context.**

3. **Classify `"context"` Correctly**
   - `"primary"`: The dataset is used for direct analysis in the
   document.
   - `"supporting"`: The dataset is referenced to validate or compare
   findings.
   - `"background"`: The dataset is mentioned as general context or
   prior research.

   **Examples:**
   - `"The LSMS-ISA data is analyzed to assess the impact of
   agricultural practices on productivity."` → `"primary"`
   - `"Our results align with previous studies that used LSMS-ISA."` →
   `"supporting"`
   - `"LSMS-ISA is widely recognized as a reliable data source for
   agricultural research."` → `"background"`

4. **Capture Full Sentence Context**
   - The `"mentioned_in"` field must always include the **full
   sentence** where the dataset is referenced.
   - If a dataset is mistakenly extracted from an unrelated sentence,
   correct it.
```

Listing 4 [Continued] System prompt used to extract the initial structured data containing likely data mentions from a given text [3/3].

```
### **Extraction Schema**
Each extracted dataset should have the following fields:
- `raw_name`: Exact dataset name from the text (**no paraphrasing**).
- `harmonized_name`: If properly named, use directly; if referenced in
multiple ways, standardize using the most precise form in the text,
otherwise, set this to None.
- `acronym`: Extract if explicitly mentioned.
- `mentioned_in`: **Full sentence** where the dataset appears (**no
paraphrasing**).
- `context`: **primary / supporting / background**
- `specificity`: **properly_named / descriptive_but_unnamed /
vague_generic**
- `relevance`: **directly_relevant / indirectly_relevant /
not_relevant**
- `producer`: **Extract only if explicitly mentioned; otherwise, set to
`None`.**
- `data_type`: **Assign based on existing categories, but create new
ones if necessary.**

### **Handling New or Unlisted Data Types**
- If a dataset does not fit into existing categories, **infer an
appropriate name** for its `"data_type"` based on context.
- Use a **general but informative label** for new data types (e.g.,
`"Climate Risk Data"`, `"Social Media Analytics"`).

### **Important: Do NOT Skip Unnamed Datasets**
If a dataset is described but lacks a proper name, extract it under
`"descriptive_but_unnamed"` or `"vague_generic"`, which ever is
appropriate.
If `"producer"` is not mentioned, set it to `None` rather than
inferring.
```

A.5.2 LLM-as-a-Judge prompt

Listing 5 System prompt used to characterize the LLM-as-a-Judge agent to assess the quality of the first stage of structured data generation [1/2].

```
You are an expert in dataset validation. Your task is to assess whether
each dataset mention is **valid, invalid, or requires clarification**,
ensuring correctness and consistency based on the dataset's **empirical
context**.

---

### **Dataset Validation Criteria**
A dataset is **valid** if:
1. **It is structured**|collected systematically for research, policy,
or administrative purposes.
2. **It is reproducible**|meaning it consists of collected records
rather than being derived purely from computations or models.

**Always Valid Datasets:**
- Government statistical and geospatial datasets (e.g., census,
official land records).
- Official surveys, administrative records, economic transaction data,
and scientific research datasets.

**Invalid Datasets:**
Set as invalid all `"raw_name"` that belong under the following
classes.
- Derived indicators or computational constructs (e.g., "wealth score",
"mine dummy", "district total production").
- Standalone statistical metrics without a clear underlying dataset
(e.g., "average income growth rate" without source data).
- General organizations, reports, or methodologies (e.g., "World Bank",
"UNDP Report", "machine learning model").

**Uncertain Cases:**
- If a dataset is **vaguely named but potentially valid**, set it as
valid but return: `"Potentially valid|needs dataset name
confirmation."`
- If a dataset reference is **too generic** (e.g., `"time-varying data
on production"`), set it as valid but return: `"Needs
clarification|dataset name is too generic."`
```

Listing 6 [Continued] System prompt used to characterize the LLM-as-a-Judge agent to assess the quality of the first stage of structured data generation [2/2]

```
---

### **Key Validation Rules**
1. **Consistency Check:**
   - If a `"raw_name"` has been marked **valid earlier**, it **must
   remain valid** unless its meaning significantly differs in a new
   context.

2. **Context-Aware Inference:**
   - If certain details are missing such as the **Year**, **Producer**,
   or **Data Type**, try to extract them from the `mentioned_in` field
   if available and correctly relate to the data.

3. **Data Type Classification (Flexible & Adaptive):**
   - Infer the most appropriate `"data_type"` dynamically from context.
   - Possible types: **Surveys, geospatial data, administrative
   records, financial reports, research datasets, climate observations,
   etc.**
   - If **no predefined category fits**, create a **new `"data_type"`
   that best describes the dataset.**

4. **Producer Identification:**
   - If the **producer (organization/institution) is explicitly
   mentioned**, extract it.
   - If not mentioned, **do not infer|set `"producer": None"`
   instead.**

---

### **JudgeResponseFormat Schema**
Each dataset assessment must conform strictly to the
JudgeResponseFormat schema."
```

A.5.3 Reasoning agent prompt

Listing 7 System prompt used to characterize the reasoning agent.

```
Your task is to review a structured user input that may mention a
dataset in a text. Please take your time.

Carefully analyze what the text in the `mentioned_in` field explicitly
means and in what context the `raw_name` is discussed. Never infer,
imply, or assume, so you must exclusively rely on the text as facts. If
there are multiple datasets, do the assessment individually.

Plan a strategy to ensure you can maximize the chances of correctly
judging and classifying whether the provided input:
- Clearly, the `raw_name` falls under the concept of a data/dataset and
not by extension or implicitly.
- Whether the raw_name is actually in the `mentioned_in`.
- Whether the harmonized_name (if present) is actually in the
`mentioned_in`. If not found, remove it from the output.
- The `raw_name` is `properly_named` (e.g., DHS, LSMS, etc.),
`descriptive_but_unnamed` (administrative school records in Ghana for
2020) , or `vague_generic` (a survey data). Any of these are valid data
mentions. To be sure, elaborate how you interpret these classes and use
that for classifying.
- The context concerning usage of the dataset is mentioned: is it
`primary`, `supporting`, or `background`.

Then, write down your strategy.

After you write down your strategy, synthesize it to develop a rubric
of what qualifies as a dataset, which you must use to base your
judgment.

Incorporate a devil's advocate review as part of your strategy. If the
review shows inconsistency, update accordingly. Do not reason based on
assumption, inference, or implicit thinking.  Relationships do not
count as a dataset; for example, the producer is not a dataset.

Execute the strategy, **step by step**, and write an analysis of how
you interpret the `raw_name` in the context of the `mentioned_in`.

If your analysis results in the `raw_name` being a dataset, set the
`valid` field to `true`, otherwise, set it to `false`. In both cases,
return the result of your analysis focusing on the `raw_name` in the
`reason` field. If it is invalid, set the `specificity` and `context`
to null.

ALWAYS WRITE A DEVIL'S ADVOCATE REVIEW AFTER THE ANALYSIS BEFORE
CONCLUDING.

After you write your analysis, your output must repeat the input with
the `specificity`, `context`, `valid` and `invalid_reason` values
replaced accordingly in the same level as the corresponding `raw_name`.
IMPORTANT: the final output must be between these tags
<OUTPUTDATA>```json<the output must be here>```</OUTPUTDATA>
```

B Reasoning agent example

Listing 8 A candidate data mention extracted from a page in a research paper.


```
{
  "mentioned_in": "In Global warming of 1.5 \nC. An IPCC\nSpecial
  Report on the Impacts of Global Warming of 1.5C above Pre-Industrial
  Levels and Related Global Greenhouse\nGas Emission Pathways, in the
  Context of Strengthening the Global Response to the Threat of Climate
  Change,\nSustainable Development, and Efforts to Eradicate Poverty;
  The Intergovernmental Panel on Climate Change:\nGeneva, Switzerland,
  2018. 9.",
  "datasets": [
    {
      "raw_name": "IPCC Special Report on the Impacts of Global Warming
      of 1.5C",
      "harmonized_name": "IPCC Special Report on the Impacts of Global
      Warming of 1.5C",
      "acronym": "IPCC",
      "producer": "Intergovernmental Panel on Climate Change",
      "year": "2018"
    }
  ],
  "source": "b71b859da04440fe5f61613da6b223db9a74cf9c",
  "page": 11
}
```

Listing 9 Reasoning agent output [1/3].

```
### Strategy

1. **Identify the Dataset**: Determine if the `raw_name` explicitly
refers to a dataset or if it is merely a title of a report or
publication. A dataset should be a collection of data points or
observations, while a report may summarize findings without being a
dataset itself.

2. **Check for Mention**: Verify if the `raw_name` is explicitly
mentioned in the `mentioned_in` field. This includes checking for exact
matches or clear references.

3. **Harmonized Name Check**: If a `harmonized_name` is provided,
ensure it is also explicitly mentioned in the `mentioned_in` field. If
not, it will be excluded from the output.

4. **Classify the Naming**: Classify the `raw_name` into one of the
three categories:
   - **Properly Named**: Recognizable acronyms or titles (e.g., IPCC).
   - **Descriptive but Unnamed**: Descriptions of datasets without
   specific titles (e.g., administrative school records).
   - **Vague Generic**: General terms that do not specify a dataset
   (e.g., survey data).

5. **Context of Usage**: Determine the context in which the dataset is
used:
   - **Primary**: The main focus of the research or report.
   - **Supporting**: Provides additional information but is not the
   main focus.
   - **Background**: Contextual information that supports the main
   findings but is not directly analyzed.

6. **Devil's Advocate Review**: After the initial analysis, review the
findings critically to ensure there are no inconsistencies or
overlooked details.
```

Listing 10 [Continued] Reasoning agent output [2/3].

```
### Rubric for Dataset Qualification
- A dataset must be a collection of data points or observations.
- The `raw_name` must be explicitly mentioned in the `mentioned_in`
field.
- The `harmonized_name` must also be explicitly mentioned if present.
- Classifications must be based on clear definitions without
assumptions.
- The context of usage must be clearly identified.

### Analysis of the Input

1. **Dataset Identification**: The `raw_name` "IPCC Special Report on
the Impacts of Global Warming of 1.5C" appears to be the title of a
report rather than a dataset. It does not represent a collection of
data points but rather summarizes findings related to global warming.

2. **Mention Check**: The `raw_name` is indeed mentioned in the
`mentioned_in` field, but it is in the context of a report rather than
a dataset.

3. **Harmonized Name Check**: The `harmonized_name` is identical to the
`raw_name` and is also mentioned in the `mentioned_in` field. However,
since the `raw_name` is not a dataset, the harmonized name's presence
does not change the classification.

4. **Classification**: The `raw_name` is classified as **properly
named** since it is a recognized title. However, it does not qualify as
a dataset.

5. **Context of Usage**: The context of the `raw_name` is primarily as
a report on climate change impacts, which does not fit into the
categories of primary, supporting, or background in terms of dataset
usage.

### Devil's Advocate Review
Upon reviewing the analysis, it is clear that while the `raw_name` is
properly named, it does not meet the criteria for being classified as a
dataset. The context of its mention reinforces that it is a report
summarizing findings rather than a dataset. Therefore, the initial
conclusion stands.
```

Listing 11 [Continued] Reasoning agent output [3/3].

````
### Final Output
```json
{
  "mentioned_in": "In Global warming of 1.5 \nC. An IPCC\nSpecial
  Report on the Impacts of Global Warming of 1.5C above Pre-Industrial
  Levels and Related Global Greenhouse\nGas Emission Pathways, in the
  Context of Strengthening the Global Response to the Threat of Climate
  Change,\nSustainable Development, and Efforts to Eradicate Poverty;
  The Intergovernmental Panel on Climate Change:\nGeneva, Switzerland,
  2018. 9.",
  "datasets": [
    {
      "raw_name": "IPCC Special Report on the Impacts of Global Warming
      of 1.5C",
      "harmonized_name": "IPCC Special Report on the Impacts of Global
      Warming of 1.5C",
      "acronym": "IPCC",
      "producer": "Intergovernmental Panel on Climate Change",
      "year": "2018",
      "specificity": null,
      "context": null,
      "valid": false,
      "invalid_reason": "The raw_name is a report title and does not
      represent a dataset."
    }
  ],
  "source": "b71b859da04440fe5f61613da6b223db9a74cf9c",
  "page": 11
}
```
````