Large Language Models for Monitoring Dataset Mentions in Climate Research

Aivin Solatorio, Rafael Macalaba, and James Liounis

AI for Data - DEC Data Group and Office of the Chief Statistician
The World Bank, Washington D.C.

ICLR International Conference On Learning Representations

Code

Paper

Motivation

- **Climate action depends on data.** Robust research for mitigation and adaptation requires diverse and accessible datasets.
- **Dataset usage is under-studied.** We lack a systematic understanding of how datasets are cited, used, or overlooked in research literature.
- **Manual tracking doesn't scale.** Tracking dataset mentions by hand is time-consuming and infeasible given the growing volume of research.
- **LLMs offer new capabilities.** Advances in large language models enable scalable, automated extraction of dataset mentions.
- **Improving data transparency.** Monitoring dataset mentions helps reveal usage patterns, gaps, and opportunities to improve data discoverability and investments in data production.

The challenge

No standard way of citing data used in the literature, making it difficult to track data usage automatically. Training data and models are scarse.

Energy Demand During a Pandemic
Evidence from Ghana and Rwanda

3 Data

This paper uses administrative data on electricity billing records from Ghana and Rwanda. The Ghana data comes from the Electricity Company of Ghana (ECG), which is the largest distributor in the country with operations in the southern and middle belts. It accounts for nearly 70% of all electricity customers in the country. We use data on billing records of the universe of electricity customers of the ECG from January 2018 to December 2020. The data identifies customer types based on the tariff applicable: residential (households), non-residential, and heavy industries. For each customer and year-month, it records the amount (kWh) of electricity consumed, the monetary value in Ghana Cedis (GHS), meter type (postpaid vs prepaid), and location (district) of the customer. In all, the data contains 42 million customer-year-month observations.

The Rwanda data comes from the Energy Utility Corporation Limited (EUCL), the main distributor, via the Rwanda Utilities Regulatory Authority (RURA). The dataset contains the billing records of the universe of electricity customers in Rwanda from January 2018 to December 2020. The data identifies customer type based on the tariff applicable: residential, non-residential (commercial, hotels, health centers, and public works (water storage and pump stations, broadcasters)), and small-and-medium industries. Also, all customers in the dataset use prepaid meters: Rwanda has a universal roll-out of prepaid meters, with large and heavy industries the only exception who are allowed to use post-paid meters. Our data exclude these customers (i.e. large and heavy industries). For each customer, we have monthly records on the amount (kWh) of electricity consumed (purchased), the monetary value in Rwandan Francs (RWF), location (community/district), and rural-urban status. In all, the data contains 21 million customer-year-month observations.

We complement the electricity data with monthly data on temperature and total precipitation from the Copernicus Climate Change Service.[9]

[9] https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset

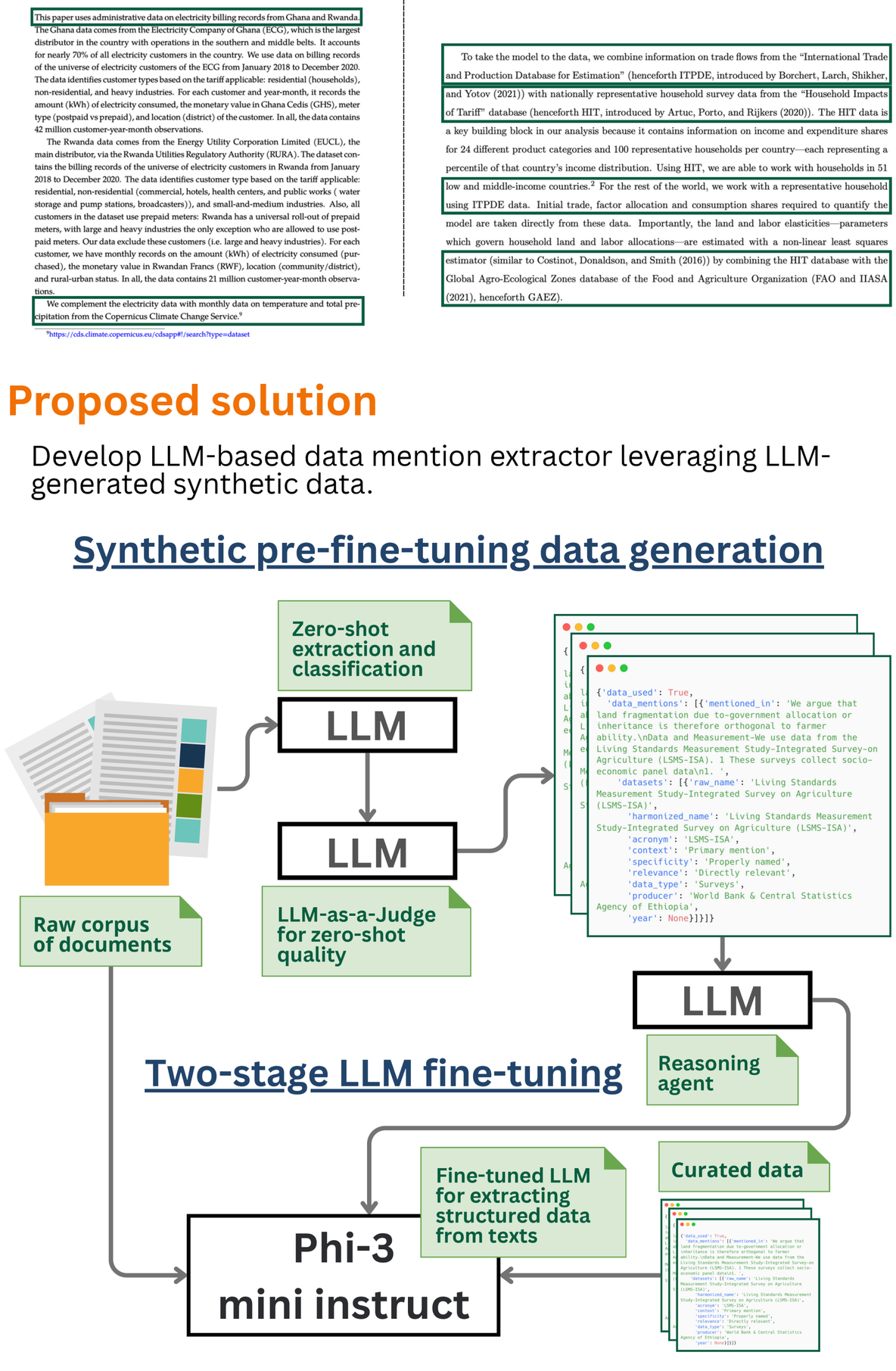
Crops, Conflict and Climate Change

To take the model to the data, we combine information on trade flows from the "International Trade and Production Database for Estimation" (henceforth ITPDE, introduced by Borchert, Larch, Shikher, and Yotov (2021)) with nationally representative household survey data from the "Household Impacts of Tariff" database (henceforth HIT, introduced by Artuc, Porto, and Rijkers (2020)). The HIT data is a key building block in our analysis because it contains information on income and expenditure shares for 24 different product categories and 100 representative households per country—each representing a percentile of that country's income distribution. Using HIT, we are able to work with households in 51 low and middle-income countries.[2] For the rest of the world, we work with a representative household using ITPDE data. Initial trade, factor allocation and consumption shares required to quantify the model are taken directly from these data. Importantly, the land and labor elasticities—parameters which govern household land and labor allocations—are estimated with a non-linear least squares estimator (similar to Costinot, Donaldson, and Smith (2016)) by combining the HIT database with the Global Agro-Ecological Zones database of the Food and Agriculture Organization (FAO and IIASA (2021), henceforth GAEZ).

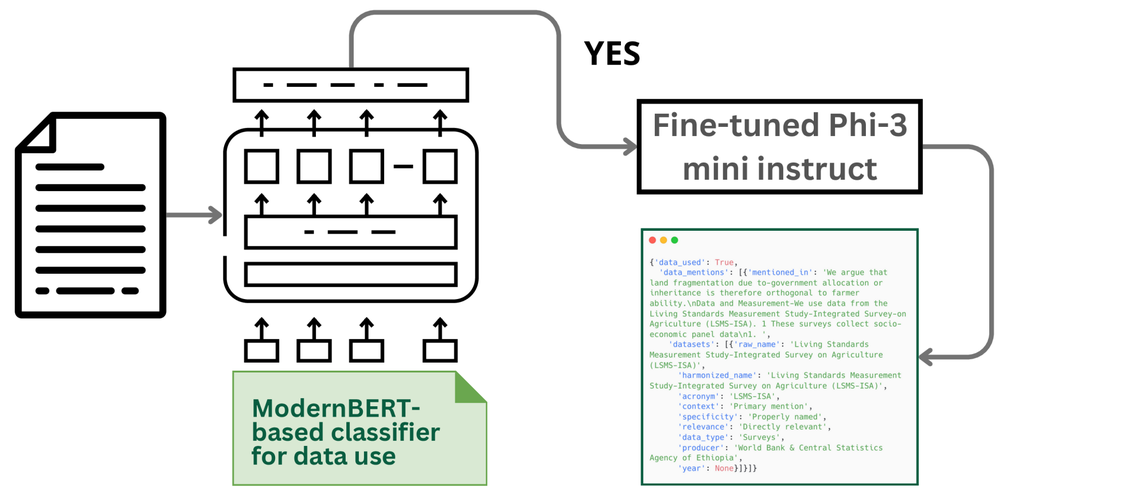
Proposed solution

Develop LLM-based data mention extractor leveraging LLM-generated synthetic data.

Synthetic pre-fine-tuning data generation



Full inference pipeline



Training parameters and Benchmarks

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

NuExtract-v1.5

NuExtract-v1.5 by NuMind 🔥

NuExtract-v1.5 is a fine-tuning of Phi-3.5-mini-instruct, trained on a private high-quality dataset for structured information extraction. It supports long documents and several languages (English, French, Spanish, German, Portuguese, and Italian). To use the model, provide an input text and a JSON template describing the information you need to extract.

Note: This model is trained to prioritize pure extraction, so in most cases all text generated by the model is present as is in the original text.

GLiNER

GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer

Urchade Zaratiana[1,2], Nadi Tomeh[2], Pierre Holat[1,2], Thierry Charnois[2]
[1] FI Group, [2] LIPN, CNRS UMR 7030, France
zaratiana@lipn.fr
https://github.com/urchade/GLiNER

Metrics and Results

Jaccard F-β score $J(S_1, S_2) = \frac{|W_1 \cap W_2|}{|W_1| + |W_2| - |W_1 \cap W_2|}$

$$\begin{cases} 1 & J(S_1, S_2) > 0.5 \\ 0 & \text{Otherwise} \end{cases}$$

W1 = Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages (EAC-I)
W2 = **Mali's** Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages (EAC-I)
J(S1, S2) = 12 / (13 + 12 - 12) = **0.923**

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

Model	Precision	Recall	F1-score
Data Use Classification Models			
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

Conclusion

Our two-stage fine-tuning approach—starting with synthetic data and followed by curated annotations—achieves state-of-the-art performance in dataset mention extraction in research papers. Pre-fine-tuning on synthetic data improves generalization and recall, while curated data increases precision. **The resulting model outperforms strong baselines like NuExtract and GLiNER, demonstrating the value of domain-adapted training.** This framework supports scalable monitoring of dataset usage, advancing transparency and data-informed decision-making in climate science.

Acknowledgment

The project has been funded by the Knowledge for Change program (KCP IV) of the World Bank and implemented by DEC Data Group (DECDG). Project name: (P503405) *Exploring Data Use in the Development Economics Literature using Large Language Models (AI and LLMs)*, KCP IV - TF0C3444. The funding received contributions from the following donors: The Sweden International Development Cooperation Agency (Sida), Agence Française de Développement (AFD), Government of Japan, and the European Union.