Large Language Models for Monitoring Dataset Mentions in Climate Research

Aivin V. Solatorio, Rafael Macalaba, and James Liounis

asolatorio@worldbank.org

Office of the Chief Statistician (DECCS)
The World Bank

We don't know the **landscape of data** that drives research

We don't know the **landscape of data** that drives research—and that’s a problem.

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).

Why is this an important problem?

Lack of transparency in data usage

No systematic way to track when or what datasets are mentioned in research literature.

Underutilization of valuable datasets

Particularly those from underrepresented regions or disciplines, which limits the breadth of climate analysis.

Lack of discoverability and reuse

Poor dataset citation practices obscure opportunities for data sharing, validation, and follow-up studies.

Hidden research biases

Over-reliance on certain datasets may skew findings and reinforce existing knowledge or policy blind spots.

Difficulty in assessing impact of datasets

Without tracking, it’s hard to tell which datasets shape research directions and inform policy decisions.

Limited empirical basis for data-related decisions

Funders and institutions lack data-driven insights into which datasets to support or curate further.

What should we do?

Open tools like Google Scholar and Semantic Scholar track how papers are cited, but there is **no equivalent for datasets**.

As a result, the **landscape of data use** in climate research and the broader research literature remains unclear.

A step to addressing this gap is to **develop** open-source **machine learning** (ML) frameworks to **identify and extract mentions of data** usage in the literature.

What stops us from doing this?

Absence of comprehensive labeled training data

The lack of diverse annotated examples for machine learning models hinders the development of dataset tracking tools.

Traditional natural language processing (NLP) models struggle with ambiguity and variation in dataset mentions

Traditional NLP approaches—such as rule-based extraction or classic named entity recognition—are limited in their ability to detect dataset mentions, which are often vague, inconsistently named, or embedded in complex scientific narratives.

Solving this problem requires more adaptive and semantically rich models, such as those based on **transformers or large language models (LLMs)**.

Our proposed solution

High-level framework

Data

Weakly supervised synthetic data.
Topic-agnostic labeling using LLMs for large-scale pre-fine-tuning.

Small-scale human-curated data for fine-tuning.

Model

Data-use binary classifier using a fine-tuned ModernBERT model.

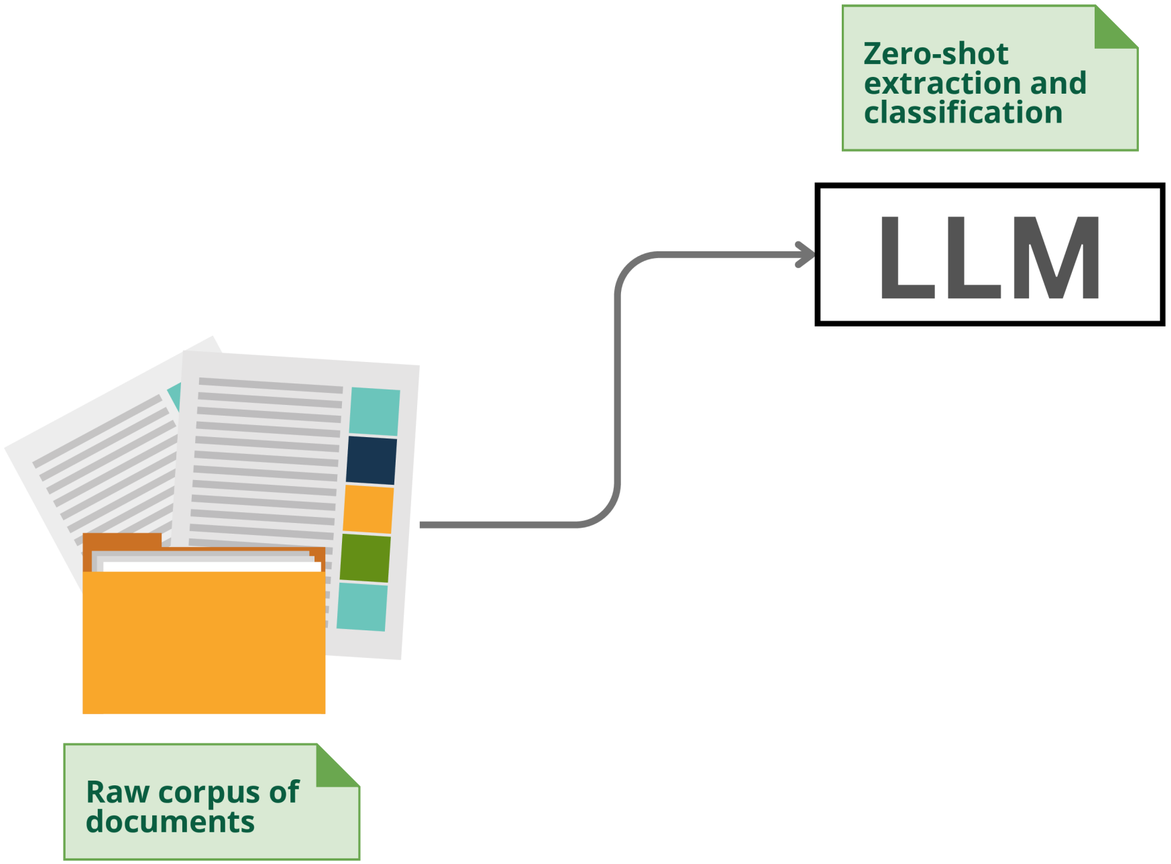
Two-stage fine-tuning of the Phi-3-mini-instruct model.

Addressing the data problem

Synthetic pre-fine-tuning data generation

Raw corpus of documents

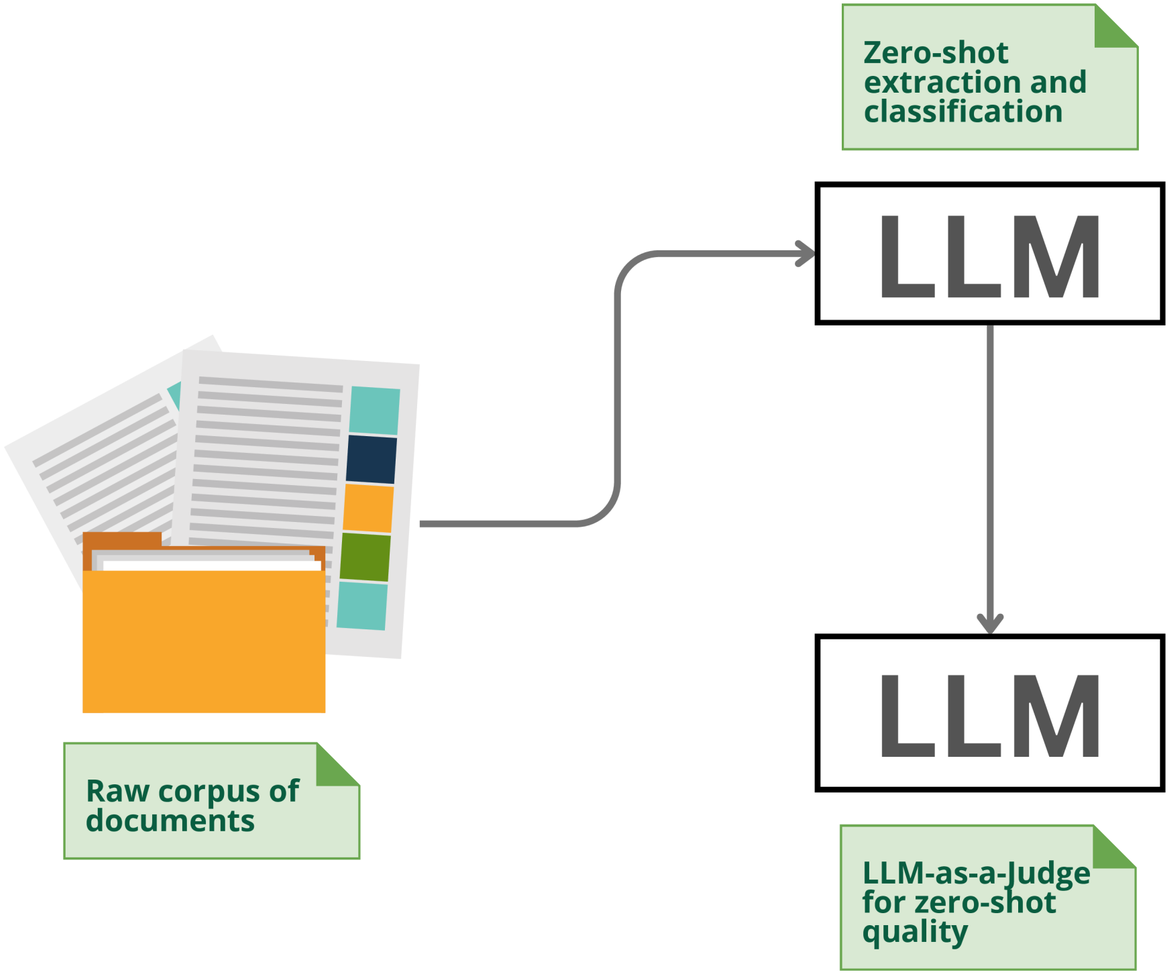
Synthetic pre-fine-tuning data generation



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")
```

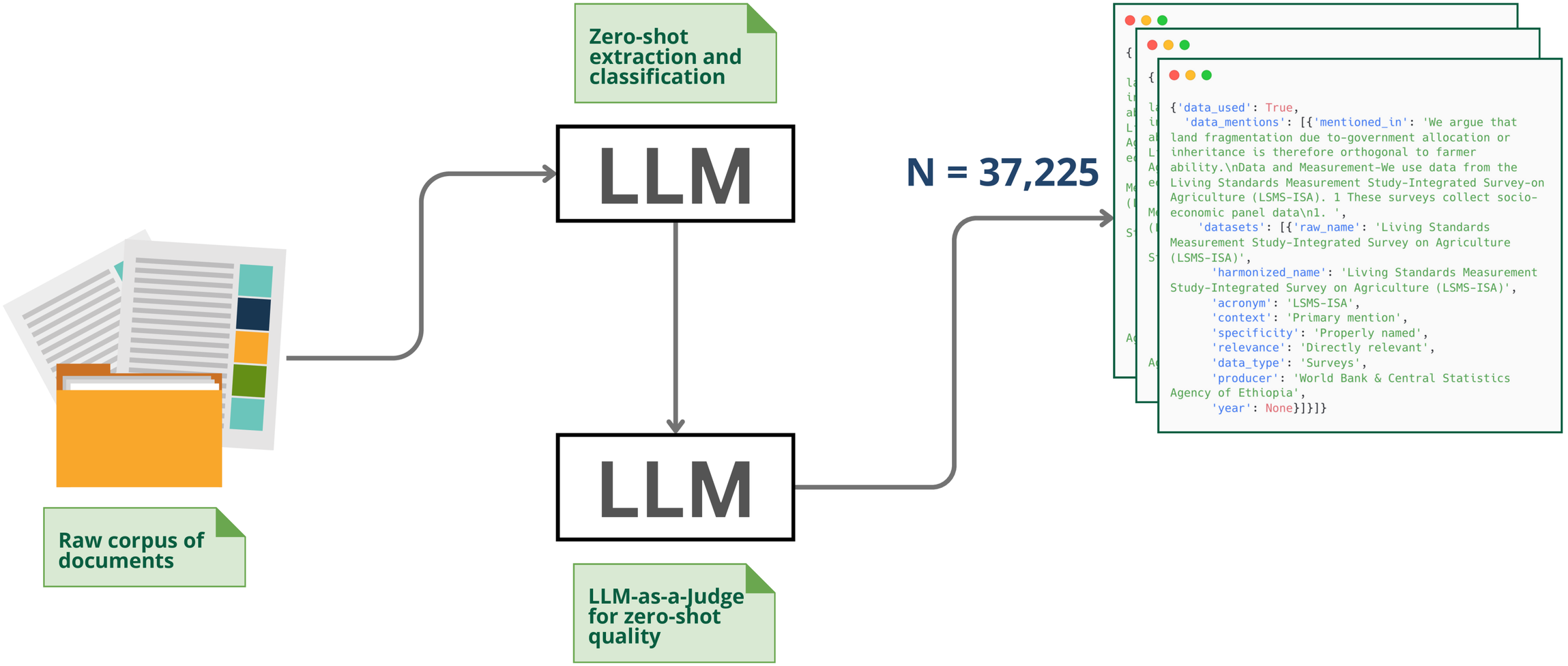
Synthetic pre-fine-tuning data generation



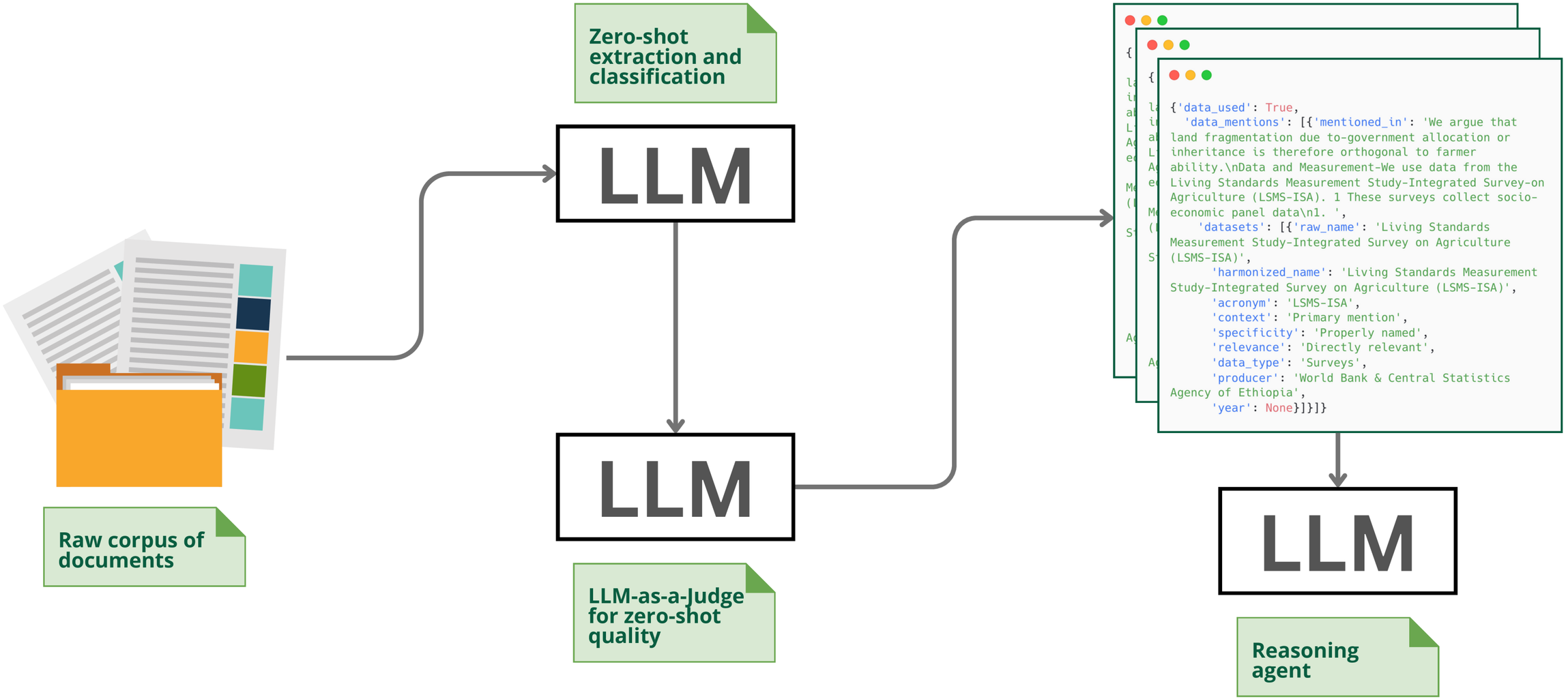
Listing 5 System prompt used to characterize the LLM-as-a-Judge agent to assess the quality of the first stage of structured data generation [1/2].

```
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."`
```

Synthetic pre-fine-tuning data generation



Synthetic pre-fine-tuning data generation

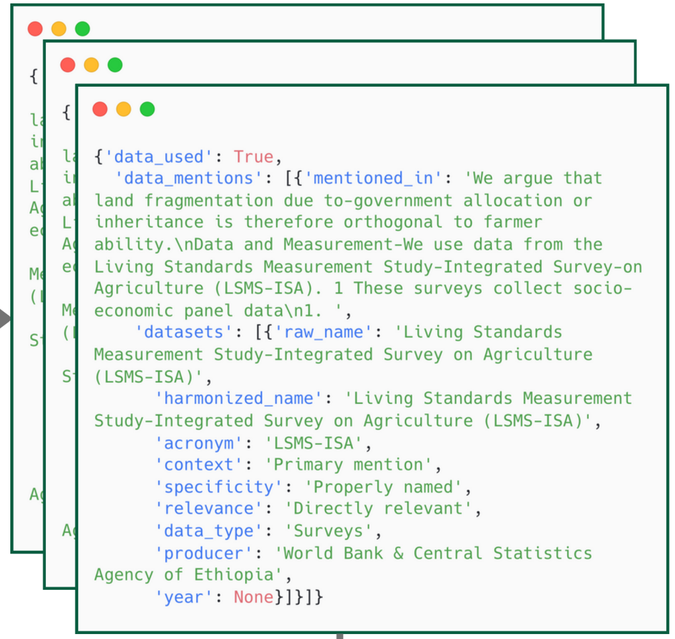


Synthetic pre-fine-tuning data generation

Raw corpus of documents

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>
```



```
{'data_used': True,
  'data_mentions': [{'mentioned_in': 'We argue that
land fragmentation due to-government allocation or
inheritance is therefore orthogonal to farmer
ability.\nData and Measurement-We use data from the
Living Standards Measurement Study-Integrated Survey-on
Agriculture (LSMS-ISA). 1 These surveys collect socio-
economic panel data\n1. ',
    'datasets': [{'raw_name': 'Living Standards
Measurement Study-Integrated Survey on Agriculture
(LSMS-ISA)',
      'harmonized_name': 'Living Standards Measurement
Study-Integrated Survey on Agriculture (LSMS-ISA)',
      'acronym': 'LSMS-ISA',
      'context': 'Primary mention',
      'specificity': 'Properly named',
      'relevance': 'Directly relevant',
      'data_type': 'Surveys',
      'producer': 'World Bank & Central Statistics
Agency of Ethiopia',
      'year': None}]}]}
```

LLM

Reasoning agent

Synthetic pre-fine-tuning data generation

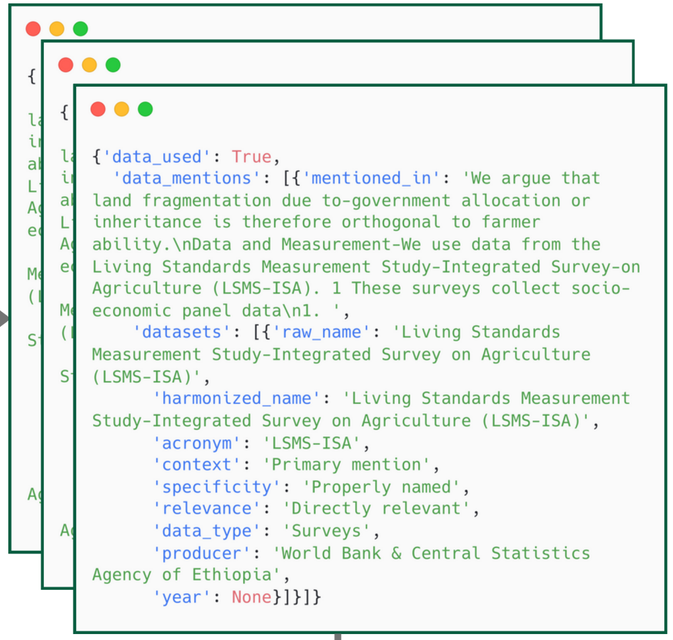
Raw corpus of documents

Zero-shot extraction and classification

LLM

LLM

LLM-as-a-Judge for zero-shot quality



LLM

N = 21,408

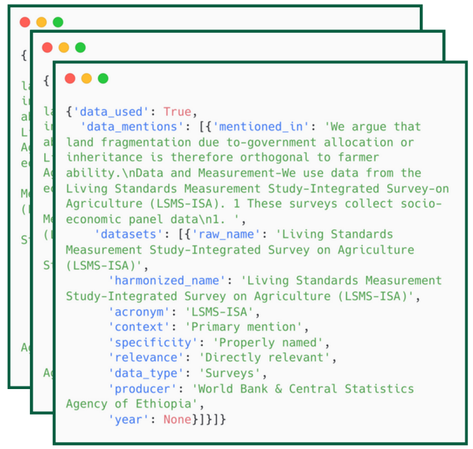
Reasoning agent

Manually curated fine-tuning data

Raw pages of documents

Curated data

N = 1,000



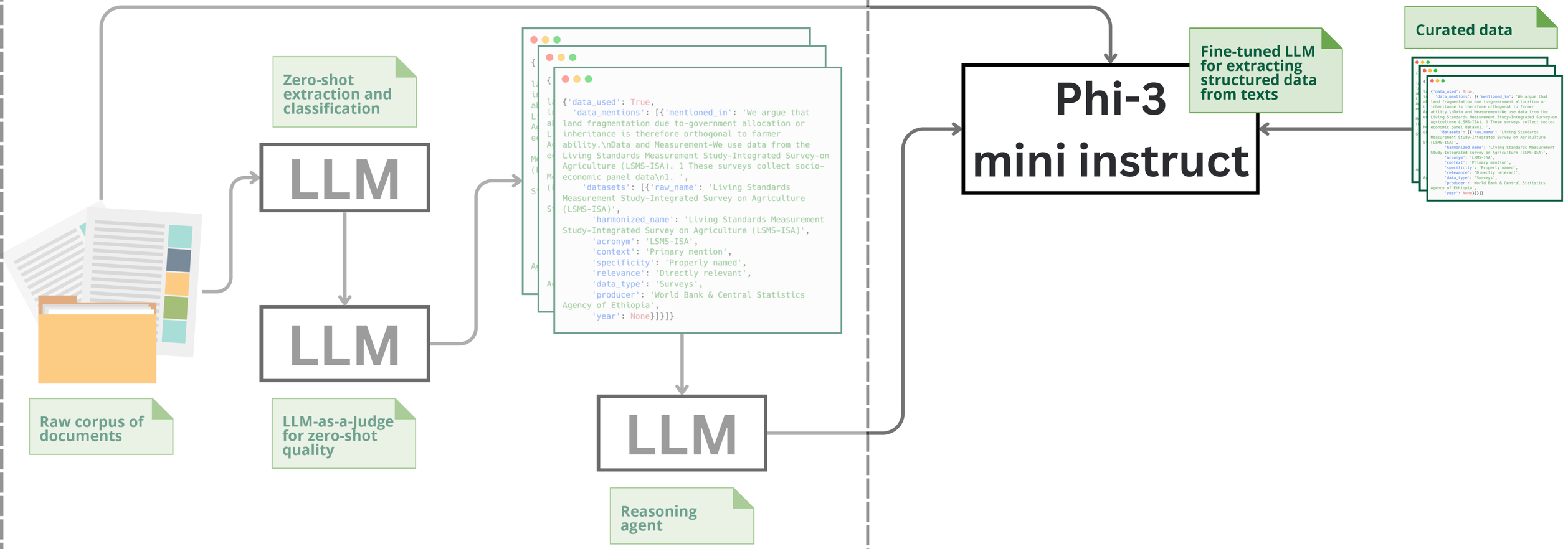
```
{'data_used': True,
  'data_mentions': [{'mentioned_in': 'We argue that
land fragmentation due to-government allocation or
inheritance is therefore orthogonal to farmer
ability.\nData and Measurement-We use data from the
Living Standards Measurement Study-Integrated Survey-on
Agriculture (LSMS-ISA). 1 These surveys collect socio-
economic panel data\n1. ',
     'datasets': [{'raw_name': 'Living Standards
Measurement Study-Integrated Survey on Agriculture
(LSMS-ISA)',
       'harmonized_name': 'Living Standards Measurement
Study-Integrated Survey on Agriculture (LSMS-ISA)',
       'acronym': 'LSMS-ISA',
       'context': 'Primary mention',
       'specificity': 'Properly named',
       'relevance': 'Directly relevant',
       'data_type': 'Surveys',
       'producer': 'World Bank & Central Statistics
Agency of Ethiopia',
       'year': None}]}]}
```

Using Doccano: https://github.com/doccano/doccano

Training the extractor model

Synthetic pre-fine-tuning data generation

Two-stage LLM fine-tuning

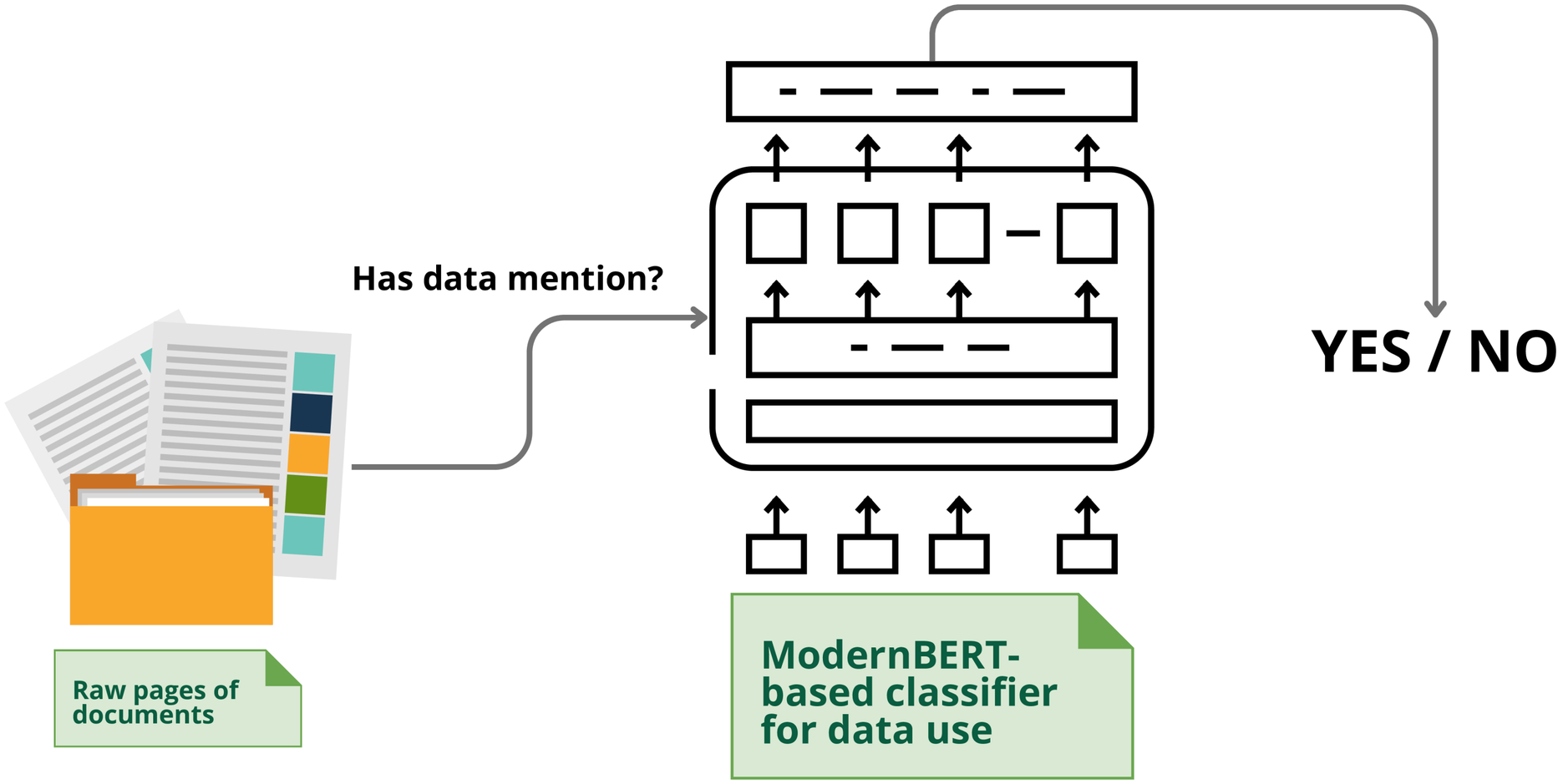


LLM fine-tuning parameters

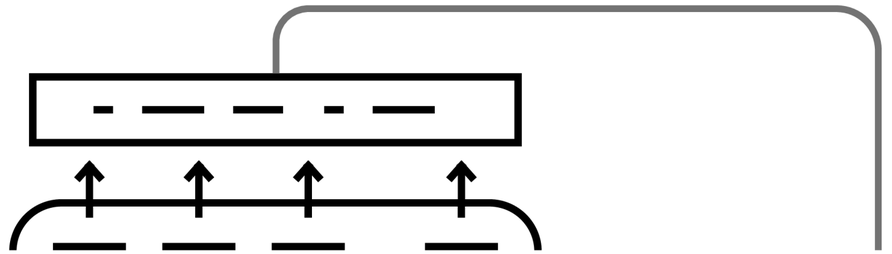
Parameter	Pre-fine-tuning on Synthetic Data	Fine-tuning on High-quality Annotated Data
Dataset Type	Large weakly supervised synthetic	Small manually annotated
Epochs	10	20
Effective Batch Size	16	2
Learning Rate	2e-4	3e-5
Warmup Ratio	1%	1% (same as pre-fine-tuning)
Scheduler	Linear with decay 0.01	Linear 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

Training the classifier model

Training a data mention binary classifier



Training a data mention binary classifier

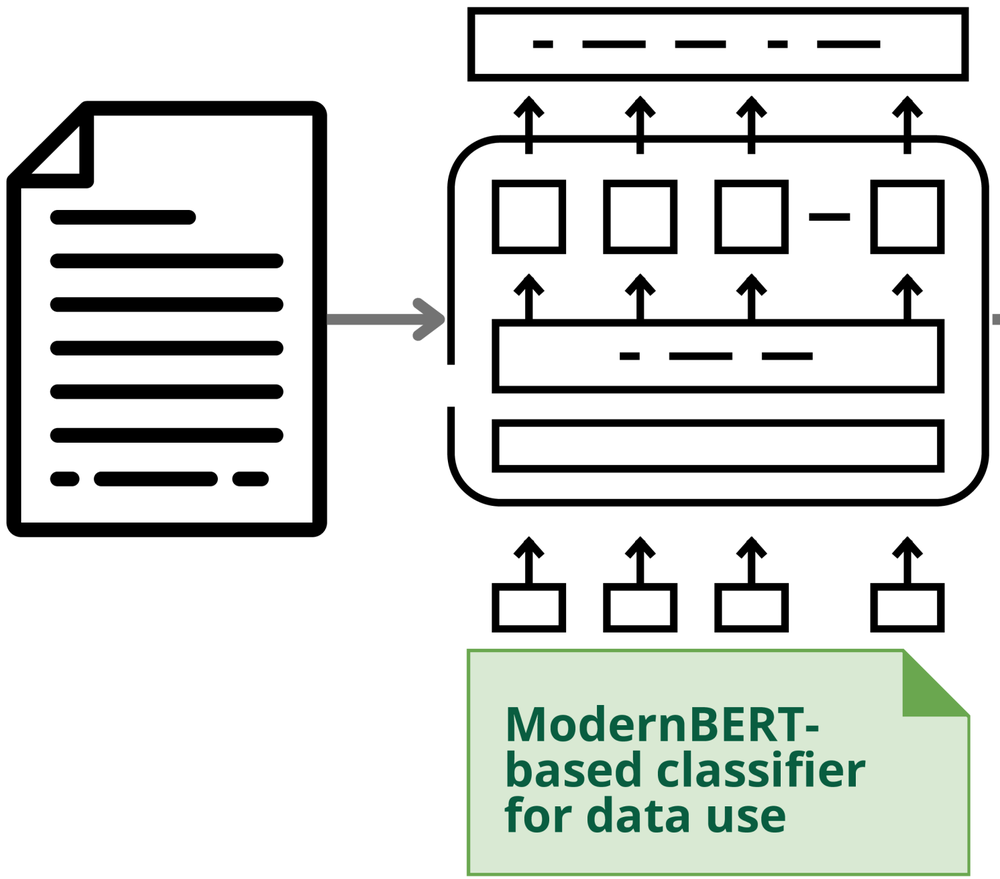


Makes the inference efficient by **not applying the extractor** model to texts that likely **do not mention data usage**.

Raw pages of documents

ModernBERT-based classifier for data use

Full inference pipeline



YES

Fine-tuned Phi-3 mini instruct

ModernBERT-based classifier for data use


```
{'data_used': True,
  'data_mentions': [{'mentioned_in': 'We argue that
land fragmentation due to-government allocation or
inheritance is therefore orthogonal to farmer
ability.\nData and Measurement-We use data from the
Living Standards Measurement Study-Integrated Survey-on
Agriculture (LSMS-ISA). 1 These surveys collect socio-
economic panel data\n1. ',
    'datasets': [{'raw_name': 'Living Standards
Measurement Study-Integrated Survey on Agriculture
(LSMS-ISA)',
      'harmonized_name': 'Living Standards Measurement
Study-Integrated Survey on Agriculture (LSMS-ISA)',
      'acronym': 'LSMS-ISA',
      'context': 'Primary mention',
      'specificity': 'Properly named',
      'relevance': 'Directly relevant',
      'data_type': 'Surveys',
      'producer': 'World Bank & Central Statistics
Agency of Ethiopia',
      'year': None}]}]}
```

Extractor Baselines

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

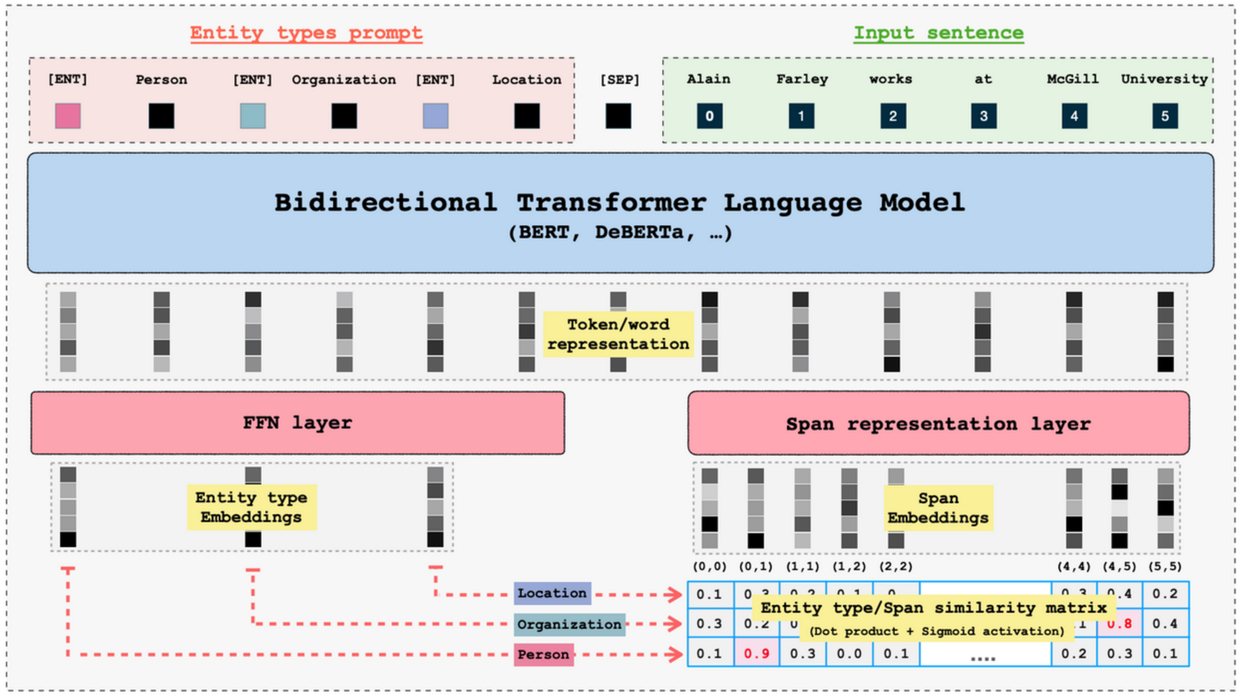


Figure 2: **Model architecture**. GLiNER employs a BiLM and takes as input entity type prompts and a sentence/text. Each entity is separated by a learned token [ENT]. The BiLM outputs representations for each token. Entity embeddings are passed into a FeedForward Network, while input word representations are passed into a span representation layer to compute embeddings for each span. Finally, we compute a matching score between entity representations and span representations (using dot product and sigmoid activation). For instance, in the figure, the span representation of (0, 1), corresponding to "Alain Farley," has a high matching score with the entity embeddings of "Person".

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.

Extraction template

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

Metrics

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**

Results

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

Key Findings

Key findings

- **Two-Stage Fine-Tuning Boosts Accuracy:** Fine-tuning Phi-3-mini on synthetic data followed by curated data achieves the highest Fβ score (71.43), demonstrating superior generalization and precision.

- **Synthetic Data Enhances Recall:** The synthetic-only model outperforms the curated-only model in recall (70.00 vs. 65.52), highlighting synthetic data's value in expanding coverage despite lower precision.

- **Curated Data Refines Precision:** Human-annotated data improves precision and plays a critical role in refining model performance after broad generalization with synthetic data.

Key findings

- **Outperforms Existing Baselines:** The two-stage fine-tuned model significantly outperforms NuExtract-v1.5 and GLiNER-large-v2.1, validating the effectiveness of domain-adapted fine-tuning.

- **Pre-fine-tuning Prevents Overfitting:** Pre-fine-tuning is essential with limited annotated data—conditioning with synthetic data improves adaptability and recall across diverse research domains.

Next Steps

- **Improve the synthetic data generation and models:** Add more diversity to the source corpus, test other prompting strategies, and explore different models.

- **Develop harmonization models/heuristics:** The same dataset may be mentioned differently within and across documents. Methods to harmonize data mentions, similar to paper citations, will be developed.

- **Scaling the extraction to a large corpus:** Apply the improved model to a large collection of documents, e.g., from Semantic Scholar, to produce the landscape of data usage.

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.

Large Language Models for Monitoring Dataset Mentions in Climate Research

Aivin V. Solatorio, Rafael Macalaba, and James Liounis

asolatorio@worldbank.org

Office of the Chief Statistician (DECCS)
The World Bank