Using Large Language Models to Enrich the Documentation of Datasets for Machine Learning

Joan Giner-Miguelez[a,*], Abel Gómez[a] and Jordi Cabot[b,c]

[a] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Rambla del Poblenou, 156, Barcelona, 08018, Spain

[b] Luxembourg Institute of Science and Technology, 5, Av. des Hauts-Forneaux, Esch-sur-Alzette, 4362, Luxembourg

[c] University of Luxembourg, 2 Av. de l'Universite, Esch-Belval, Esch-sur-Alzette, 4365, Luxembourg

ARTICLE INFO

Keywords:
Dataset Documentation
Responsible AI
AI Regulation
Fairness
Trustworthy AI
Large Language Models

ABSTRACT

Recent regulatory initiatives like the European AI Act and relevant voices in the Machine Learning (ML) community stress the need to describe datasets along several key dimensions for trustworthy AI, such as the provenance processes and social concerns. However, this information is typically presented as unstructured text in accompanying documentation, hampering their automated analysis and processing. In this work, we explore using large language models (LLM) and a set of prompting strategies to automatically extract these dimensions from documents and enrich the dataset description with them. Our approach could aid data publishers and practitioners in creating machine-readable documentation to improve the discoverability of their datasets, assess their compliance with current AI regulations, and improve the overall quality of ML models trained on them.

In this paper, we evaluate the approach on 12 scientific dataset papers published in two scientific journals (Nature's Scientific Data and Elsevier's Data in Brief) using two different LLMs (GPT3.5 and Flan-UL2). Results show good accuracy with our prompt extraction strategies. Concrete results vary depending on the dimensions, but overall, GPT3.5 shows slightly better accuracy (81,21%) than FLAN-UL2 (69,13%) although it is more prone to hallucinations. We have released an open-source tool implementing our approach and a replication package, including the experiments' code and results, in an open-source repository.

1. Introduction

The need for better data is a common demand in the machine learning (ML) community. Recent studies have pointed to data as one of the root causes of unintended and downstream effects along all the stages of ML applications. For instance, medical datasets imbalanced in terms of gender produce biased classifiers for computer-aided diagnosis (Larrazabal et al., 2020), language datasets gathered from Australian speakers could drop the accuracy of models trained to support users in the United States because of the different language styles (Bender and Friedman, 2018), or ML models to detect pneumonia in chest X-ray images fail to generalize to other hospitals due to specific conditions during the collection of the images (Liang et al., 2022). These examples demonstrate the importance of preserving knowledge about how the data has been collected and annotated, or the potential social impact on specific groups.

This situation has triggered the interest of regulatory agencies and the machine learning community in developing data best practices, such as dataset documentation. Recent public regulatory initiatives (such

[*] Corresponding author

jginermi@uoc.edu (J. Giner-Miguelez); agomezlla@uoc.edu (A. Gómez); jordi.cabot@list.lu (J. Cabot)

ORCID(s): 0000-0003-2335-6977 (J. Giner-Miguelez); 0000-0003-1344-8472 (A. Gómez); 0000-0003-2418-2489 (J. Cabot) as the European AI Act[1] and the AI Right of Bills[2]) and relevant scientific works have provided general guidelines for creating standard dataset documentation (Gebru et al., 2021; McMillan-Major et al., 2021; Bender and Friedman, 2018; Holland et al., 2020; Micheli et al., 2023). More recent works have proposed a structured format for these guidelines (Giner-Miguelez et al., 2023f; MLCommons, 2023), enabling them to be ingested by data search engines like Google Dataset Search (Brickley et al., 2019), increasing their discoverability. In these proposals, the authors identify which dimensions, such as the provenance of the dataset or the potential social issues, may influence how the dataset is used and the quality of the ML models trained with it.

Besides, data-sharing practices in scientific data have emerged in the last years (Feger et al., 2020; Tedersoo et al., 2021). The adoption of Data Management Plans (Bishop et al., 2023) in research institutions and the appearance of scientific data journals have motivated researchers to publish their datasets as scientific publications (e.g., data paper (Chavan and Penev, 2011)) or as accompanying technical documentation (e.g., in open data portals). Even though these documents include several of the ML community's desired dimensions (such as the methods used to collect or annotate the data), they are written in natural text and don't have a fixed structure (Thuermer et al., 2023), making them difficult to be queried and analyzed.

This paper proposes a machine-learning approach to automatically extract the demanded dimensions by the ML community from the datasets' documentation. We believe that our proposal can aid practitioners and data publishers *(i)* in creating machine-readable documentation to improve the discoverability of the data, *(ii)* in checking the compliance of their data with the emerging public AI regulations and *(iii)* helping them in evaluating the suitability of a dataset for a specific ML application. Our method is based on composing a chain of specific prompts for each dimension which will be ingested by a Large Language Model (LLM) (Ouyang et al., 2022). The prompts of the chains are designed using different prompting strategies—such as using a retriever to augment the prompts (Izacard et al., 2022)—to extract the required dimension based solely on the provided documentation while trying to avoid hallucination issues.

To validate our approach, we selected a subset of the papers published in two scientific data journals, Nature's Scientific Data[3] and Elsevier's Data in Brief[4], all describing scientific datasets. First, we manually described these papers in the specified dimensions, and then, we generated automatic descriptions of the papers with our method using two different LLMs (GPT3.5 (Ouyang et al., 2022) and FLAN-UL2 (Tay et al., 2023)). The results were then reviewed by comparing both descriptions and evaluating the accuracy and faithfulness—i.e., whether the generated answer to the input documents was truthful or not (Maynez et al., 2020; Creswell and Shanahan, 2022). Finally, we present the open-source tool (Giner-Miguelez et al., 2023a) implementing our method suited to analyze the documentation of scientific datasets. The tool ingests the dataset documentation (e.g., data papers) and is able to extract the demanded dimensions and check its level of completeness. A public demo of the tool can be found online (Giner-Miguelez et al., 2023d), and the experiment's results and data are available in an open-source repository (Giner-Miguelez et al., 2023e).

In summary, our research objectives are as follows:

- To propose an approach for automatically enriching dataset documentation for trustworthy AI.
- To explore the emerging LLM's suitability for extracting each desired dimension from raw dataset documentation.
- To propose specific prompting strategies for extracting each dimension while avoiding hallucinations.

[1] European AI Act required documentation: Annex IV: https://www.euaiact.com/annex/4

[2] https://www.whitehouse.gov/ostp/ai-bill-of-rights

[3] https://www.nature.com/sdata/

[4] https://www.sciencedirect.com/journal/data-in-brief

Table 1
Target dimensions of the extraction approach

Dimension	Subdimension	Target explanation
Uses	Design intentions	ML tasks, purposes, and gaps the dataset intends to fill
	Recommendations	Identification of the recommended and non-recommended uses
	ML Benchmarks	The ML approaches the dataset has been tested (if any)
Contributors	Authors	The authors of the dataset
	Funding	Funding information (grants, funder's type)
	Maintenance	Maintainers & policies (erratum, updates, deprecation)
Distribution	Accessibility	The links where the data can be accessed
	Licenses	Legal condition of the dataset and the models trained with it
	Deprecation policies	The deprecation plan for the dataset.
Composition	Data records	File composition and attribute identification
	Data splits	Recommended data splits to train ML models with the dataset
	Statistics	Relevant statistics pointed in the documentation
Gathering	Description & type	Description of the process and its categorization
	Team	Information about the type and demographics of the team
	Source & infrastructure	The source of the data and the infrastructure used to collect it
	Localization	Temporal and geographical localization of the data
Annotation	Description & type	Description of the process and its categorization
	Team	Information about the type and demographics of the team
	Infrastructure	The tools used to annotate the data
	Validation	Validation methods applied over the annotations
Social Concerns	Bias	Potential bias issues in data
	Sensitivity data	Potential representative or sensitivity issues in data
	Privacy	Issues concerning data privacy (p.e: anonymization)

The paper structure is as follows: in Section 2 we present the dimensions of interest, while in Section 3 we present the proposed method used to extract them. In Section 4 we present the case study on scientific data publication, and we discuss the results; while in Section 5 we present the developed tool. Finally, in Section 6 we present the related work, and Section 7 wraps up the conclusions and future work.

2. Background: Guidelines for dataset documentation

The general baseline for datasets documentation is clearly defined in the well-known paper *Datasheets for Datasets* by Gebru et al. (2021). This work gets the idea of *datasheets* from the electronic field, where every component has an associated datasheet as documentation. *Datasheets for Datasets*, together with subsequent works in the field (Gebru et al., 2021; McMillan-Major et al., 2021; Bender and Friedman, 2018; Holland et al., 2020), state a set of dimensions that need to be documented for datasets intended to be used in ML. In Table 1 we can see an overview of these dimensions, which represent the target of our extraction process.

The *Uses* dimension refers to the design intentions stated by the authors, and we focus on extracting the purposes the dataset is intended for, the gaps it is intended to fill, and its recommended and non-recommended uses. Moreover, we aim to infer the machine learning task the dataset is designed for, and the machine learning (ML) benchmarks of the dataset, if this has been tested in any ML approach. *Contributors* refers to all the participants involved in the dataset creation, the funding information, and the set of maintenance policies of the dataset. In the *Distribution* dimension, we find information about the places where the data can be accessed, the policies under the dataset is released, and the deprecation policies of the dataset. The *Composition* dimension refers to the specific format of the files, their attributes, the recommended data splits to train ML models, and the relevant statistics of the dataset.

In terms of data provenance, the *Gathering* dimensions refer to details about how the data has been collected. The goal of this dimension is to get a description of the process and infer its type (among a list of pre-defined types), information about the gathering team, the data source, the infrastructure used, and the localization of the process. Moreover, the *Annotation* dimensions focus on the different aspects of the creation of the dataset labels, such as the team annotating the data, the infrastructure used, or the methods used to validate the labels. Finally, the *Social Concerns* dimension covers information about the potential effects of the data on society, such as biases, representativeness (such as the example of biased diagnosis), or privacy issues of the data.

3. Technical description of our method

Our method comprises an initial preprocessing of the dataset documentation followed up by chains of specific prompts—that are ingested by an LLMs (Ouyang et al., 2022)— one for each one of the dimensions discussed in Section 2. The goals of the chains are extracting the demanded dimensions based solely on the accompanying documentation while avoiding hallucination issues. To do so, the prompts of the chains are designed using different prompting strategies depending on how such information is typically found in the documents and/or the desired type of output (categorical, descriptive,etc.).

To exemplify our method, Figure 1 shows an excerpt of the chain for extracting the tasks the dataset is designed for. The first prompt instructs the LLMs to generate an answer to a query using the context provided within the prompt. The context is given by a retriever model fed with the same query, in the form of relevant passages from the dataset documentation. Then, along the chain, we refine, validate and complement the given answer, to finally ask the LLMs to classify it into a given list of common ML tasks.

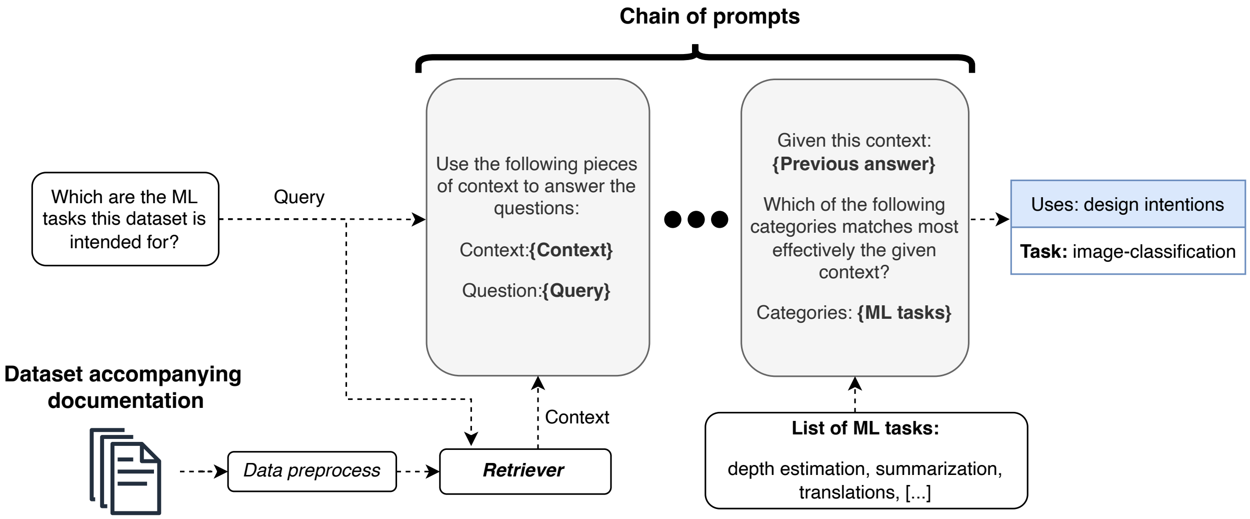

Chain of prompts
Which are the ML tasks this dataset is intended for?
Query
Use the following pieces of context to answer the questions:
Context:{Context}
Question:{Query}
Given this context:
{Previous answer}
Which of the following categories matches most effectively the given context?
Categories: {ML tasks}
Uses: design intentions
Task: image-classification
Dataset accompanying documentation
Data preprocess
Retriever
Context
List of ML tasks:
depth estimation, summarization, translations, [...]

Figure 1: Example of a chain extracting the annotation process type

The remainder of this section presents *(i)* the data preprocessing steps that must be applied to the datasets accompanying documentation to be used in our method, *(ii)* the different types of prompts that may compose the chains, and finally, *(iii)* how the prompts are assembled into the chains used to retrieve each of the analyzed dimension.

3.1. Data preprocessing

The *dataset accompanying documentation* is the source data of our extraction approach. These documents are mainly composed of natural text, and can be commonly found in a standard format, such as PDF, or published on the web in HTML. For simplicity purposes, we assume that the input of our approach is plain text file containing the main content of these documents (there are available tools to extract the running text of a PDF, such as Grobid (Romary and Lopez, 2015)).

To prepare the documents, we first split the text by passages, and then we encode it in a dense vector representation. These encoded passages are then ready to be fed to the retriever[5] together with the queries. However, in this type of documents, there is valuable information that can be found in the form of tables (for instance, the demographic's statistics of the teams). To be able to process the information in the tables too, we transform them into natural language explanations. To do so, we parse the tables from the documents, and we use an LLMs to transform them to natural text. Nevertheless, we go beyond a simple description of the table context. We need to contextualize it so that it is linked to the table description and mentions in the document. In this sense, we use a retriever to get the most relevant passages for the table in the document (by inserting the caption of the table), and we build a prompt with the parsed table and the passages as the context. This generates a new passage that is treated as any of the other passages of the document, ready to be fed to the retriever again.

Finally, the *queries* have been heuristically designed by a group of researchers by evaluating the quality of the answers of the LLMs. However, we observed inconsistencies in the vocabulary used in every documentation that has led to a worse accuracy in the LLMs answers. For instance, the gathering process is more prone to be called, "collection" or "acquisition" depending on the scientific field the dataset belongs to. To overcome this, we created a dictionary with the different terms that are inconsistent, and before executing the chains, we check which are the specific terms used by the documentation by a simple word count, getting the most popular one. Then, we tune the queries using the selected terms.

3.2. Prompt types

In general, a basic prompt is composed of a query that the LLMs aims to answer based on the knowledge acquired during the LLMs training phases. This standard behaviour is not useful in our scenario as we want to extract the relevant dimensions relying only on the dataset documentation (closed-book QA) while minimizing hallucinations. To do so, we have designed different types of prompts as we present in Figure 2, and we explain below.

3.2.1. In-context prompts

The *In-context prompts* are the basic kind of prompt that compose the chains. These prompts ask the LLMs to generate an answer from a query based solely on a given context embedded in the prompt. The context is composed of relevant passages from the dataset documentation. These passages are given by a retriever fed with the same query that performs a semantic similarity search between the query and the passages of the dataset documentation. This prompt implements a *retrieval-augmented in-context learning* strategy that allows us to mitigate content hallucination issues (Shuster et al., 2021) and has been proven useful in knowledge-intensive tasks such as question answering and text understanding (Izacard and Grave, 2021). In our use case, this type of prompt gets good results for extracting information for specific queries,

[5] In our implementation we use the FAISS library (Johnson et al., 2019) to perform semantic similarity search

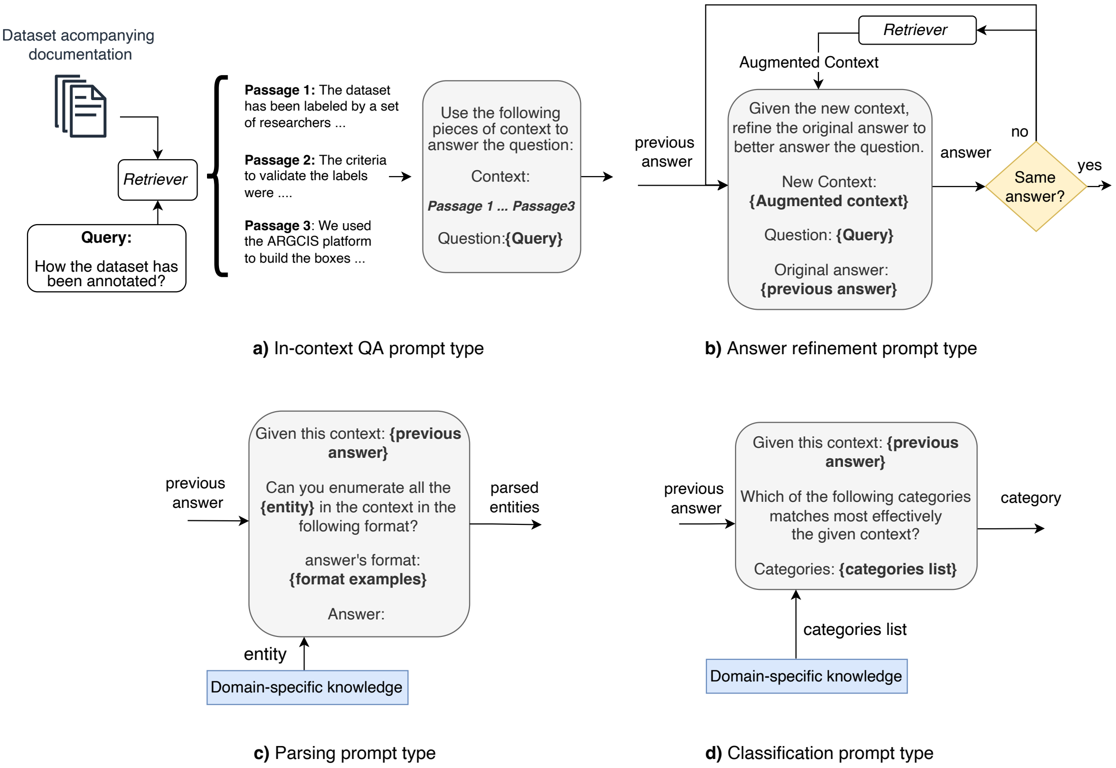
representing the most used type of prompt in the chains. In Figure 2a we can see an example of the context creation using this approach and the composition of the *In-context QA prompts* of the chains.

3.2.2. Answer refinement prompts

This type of prompt gets a previous answer of the LLMs and performs a refinement process to capture a more complete explanation of a query. These prompts are useful when we want to capture sparse information across the document in relation of an event, such as the collection or annotation process. To do so, we implemented a *generate-then-read* prompting strategy (Yu et al., 2022): we generate a new context using the retriever, but instead of using the same query, we feed the retriever with the previous answer of the LLMs. Figure 3b depicts how the answer refinement prompts work, where the process of refinement is repeated until the new context in no longer useful to generate a refined answer.

3.2.3. Classification prompts

The *Classification* prompts are composed of previous answers of the chain (for instance, the explanation of the annotator's team), which are also fed with a list of domain-specific categories (e.g., we may know that the annotators team can be either *internal* or *external* to the authors, or may be a *crowd-worker* service). Then we ask the LLMs to classify the provided answer into one of these categories. In Figure 3d we can see the *classification prompt* template. These prompt types can be used to obtain the final answer, but also to enhance the chain by asking for more information. For instance, if the annotator team is of type "crowd-worker", we can add several steps in our chain asking for the crowd-workers labor conditions as Díaz et al. (2022) propose.


Dataset acompanying documentation
Retriever
Query:
How the dataset has been annotated?
Passage 1: The dataset has been labeled by a set of researchers ...
Passage 2: The criteria to validate the labels were ....
Passage 3: We used the ARGCIS platform to build the boxes ...
Use the following pieces of context to answer the question:
Context:
Passage 1 ... Passage3
Question:{Query}
Retriever
Augmented Context
previous answer
Given the new context, refine the original answer to better answer the question.
New Context: {Augmented context}
Question: {Query}
Original answer: {previous answer}
answer
no
Same answer?
yes
previous answer
Given this context: {previous answer}
Can you enumerate all the {entity} in the context in the following format?
answer's format: {format examples}
Answer:
parsed entities
entity
Domain-specific knowledge
previous answer
Given this context: {previous answer}
Which of the following categories matches most effectively the given context?
Categories: {categories list}
category
categories list
Domain-specific knowledge

a) In-context QA prompt type

b) Answer refinement prompt type

c) Parsing prompt type

d) Classification prompt type

Figure 2: Different prompt types used in the chains

3.2.4. Parsing prompts

The goal of the *Parsing* prompts is to identify a set of specific entities in the dataset. Once identified, we can build new queries to get further information about each one. In Figure 3c, we can see the prompt template where the prompt is composed of a previous answer, an entity to extract, and a specific answer format (as suggested in (Schick and Schütze, 2021; Ye et al., 2024)). For instance, from the explanation about the funders of the dataset, we aim to parse each funder in the explanation. Then, we can build other prompts to get specific characteristics of each one.

3.3. Chain compositions

In order to extract the desired information to cover each one of the dimensions presented in Section 2, we composed a set of chains using the different types of prompts we have presented in the previous subsection. Next, we go over the chain for each dimension, presenting its workflow and discussing the specific extracted aspects. The examples presented in the figures of this section are from the "*A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions*" dataset used in the experiment of the next section, and the complete results can be seen in the open repository[6].

3.3.1. Uses

In Figure 3, we can see en excerpt of the workflow for the *uses* chain. In this chain, we see the combination of *In-Context* and *Parsing* prompt types to get the purposes and the gaps of the dataset. However, to infer the ML tasks, we have used a combination of an *In-Context prompt* (to get an explanation regarding the tasks the dataset is intended for), with a *Classification* prompt (to classify the given answer to a specific task of a given list[7]. It could be noted that whether the task is explicitly mentioned in the documentation or not, our approach enables us to extract it in a consistent format, showing one of the advantages of using generative LLMs for this task.

Regarding the recommended and non-recommended uses, we have followed the same chain composition used with the purposes and the gaps. However, to extract the ML benchmarks of the datasets, we need to

[6] https://github.com/JoanGi/Dataset-Doc-Enrichment

[7] The list of tasks is extracted from HuggingFace https://huggingface.co/tasks)

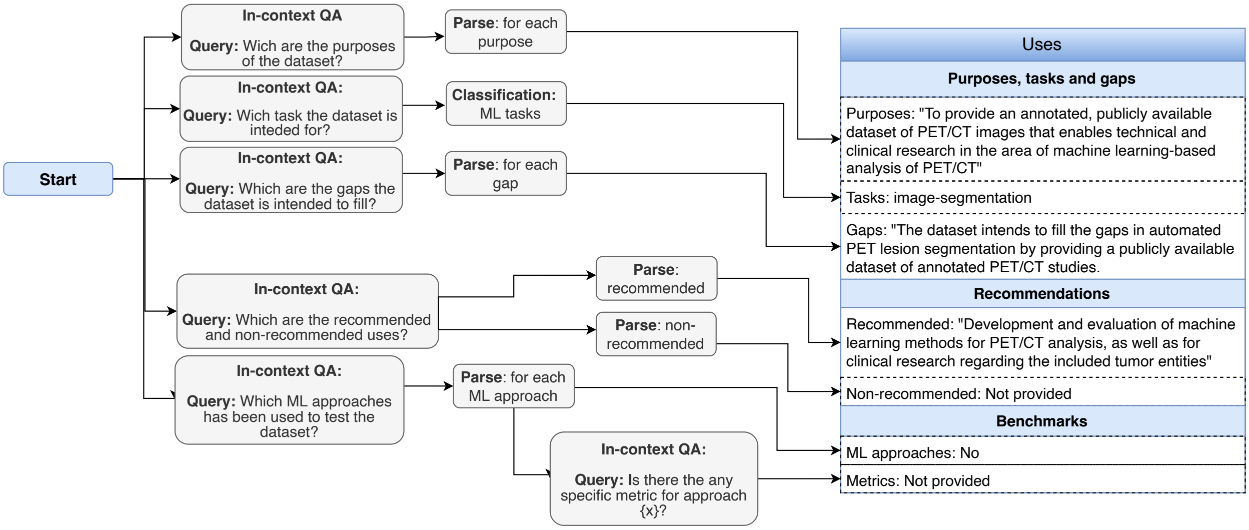

Start
In-context QA
Query: Wich are the purposes of the dataset?
Parse: for each purpose
In-context QA:
Query: Wich task the dataset is inteded for?
Classification: ML tasks
In-context QA:
Query: Which are the gaps the dataset is intended to fill?
Parse: for each gap
In-context QA:
Query: Which are the recommended and non-recommended uses?
Parse: recommended
Parse: non-recommended
In-context QA:
Query: Which ML approaches has been used to test the dataset?
Parse: for each ML approach
In-context QA:
Query: Is there the any specific metric for approach {x}?
Uses
Purposes, tasks and gaps
Purposes: "To provide an annotated, publicly available dataset of PET/CT images that enables technical and clinical research in the area of machine learning-based analysis of PET/CT"
Tasks: image-segmentation
Gaps: "The dataset intends to fill the gaps in automated PET lesion segmentation by providing a publicly available dataset of annotated PET/CT studies.
Recommendations
Recommended: "Development and evaluation of machine learning methods for PET/CT analysis, as well as for clinical research regarding the included tumor entities"
Non-recommended: Not provided
Benchmarks
ML approaches: No
Metrics: Not provided

Figure 3: Example of a chain for the *Uses* dimension

add some extra steps. After, getting the different ML approaches used to test the dataset, we parse each one, and we ask for the specific metrics for each ML approach, such as accuracy, F1, precision, and recall.

3.3.2. Contributors

The *contributor*'s dimension is composed by three subdimensions, namely: the authors, the funders and its characteristics, and the maintainers and the maintenance policies of the dataset. The chain composition for each of this subdimensions is mainly composed of an *In-context prompt* followed by a *Parsing* prompt as shown in the *uses* dimension. In Figure 4, we can see an excerpt of the workflow to obtain the funding information. Funders are extracted using the mentioned structure, but for extracting the identifier of the grants more accurately, we implemented another step of *In-context prompts* asking specifically for them.

On the other side, we also intend to classify them by their type (public, private, or mixed) and then relate they with a specific grant identifier (if this is present in the documentation). It could be noted that the type of the funder is not usually present in the documentation. However, as funders are usually well-known institutions, we used a *Basic Prompt* to ask the LLMs about them using an open-domain question and let the LLMs answer using the inherited knowledge from the pretraining phase. This is an example of how we can enhance the quality of the documentation with information that is not explicitly said in it.

3.3.3. Distribution

The *distribution* chain aims to capture aspects of the legal policies under the dataset is distributed. In that sense, using an *In-Context prompt*, we first identify the license under the dataset is released, and the links where the data can be accessed. Moreover, we try to identify if there is any third-party in charge of the license, and the specific attribution notice specified in the documentation or derived from the licenses, such as the Creative Commons licenses[8]. Complementing the license, and following recent ML domain-specific

[8] https://creativecommons.org/

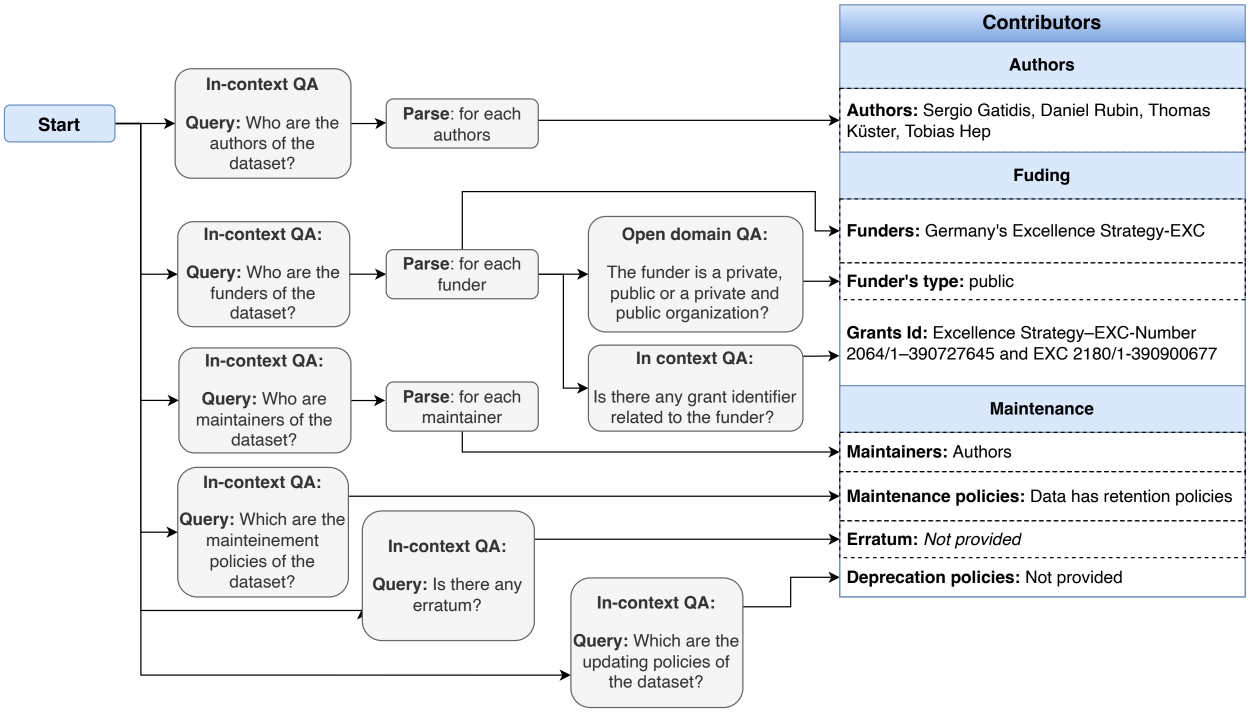

Start
In-context QA
Query: Who are the authors of the dataset?
Parse: for each authors
In-context QA:
Query: Who are the funders of the dataset?
Parse: for each funder
Open domain QA:
The funder is a private, public or a private and public organization?
In context QA:
Is there any grant identifier related to the funder?
In-context QA:
Query: Who are maintainers of the dataset?
Parse: for each maintainer
In-context QA:
Query: Which are the mainteinement policies of the dataset?
In-context QA:
Query: Is there any erratum?
In-context QA:
Query: Which are the updating policies of the dataset?
Contributors
Authors
Authors: Sergio Gatidis, Daniel Rubin, Thomas Küster, Tobias Hep
Fuding
Funders: Germany's Excellence Strategy-EXC
Funder's type: public
Grants Id: Excellence Strategy–EXC-Number 2064/1–390727645 and EXC 2180/1-390900677
Maintenance
Maintainers: Authors
Maintenance policies: Data has retention policies
Erratum: Not provided
Deprecation policies: Not provided

Figure 4: Example of a chain for the *contributors* dimension

licensing purposes[9], we try to identify if there are any mentions of the rights of the standalone data, and the rights of the models that are trained with this data. Finally, we try to identify if any deprecation policy of the dataset exists (Luccioni et al., 2022).

3.3.4. Composition

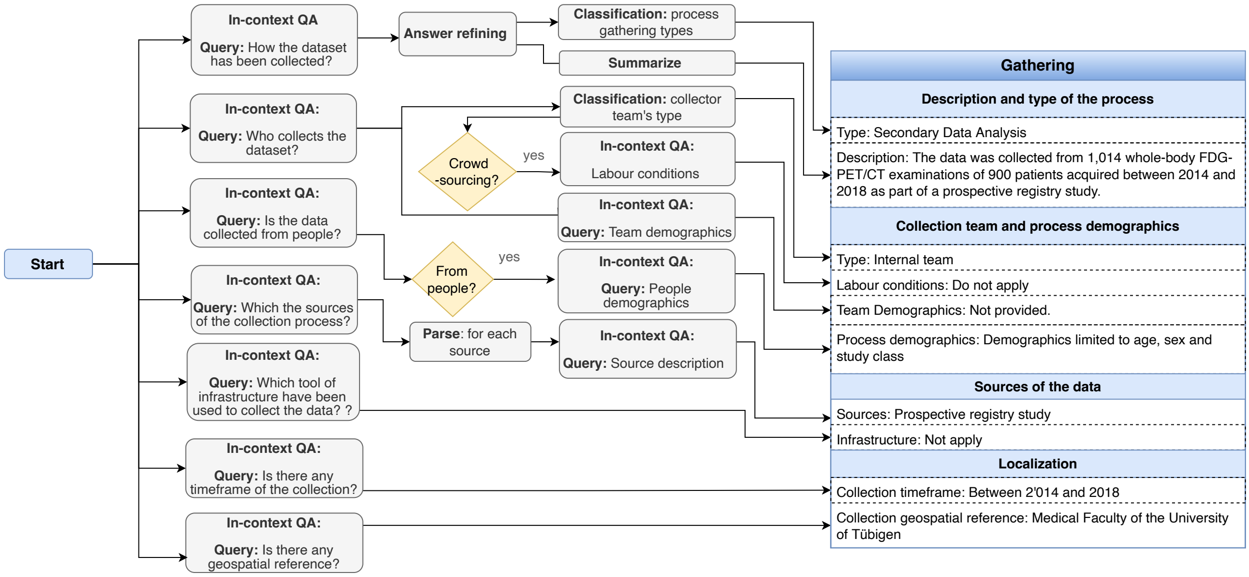
In the *composition* dimension, we try to capture aspects of the structure of the released files and its attributes. As, the data-level analysis of the dataset is usually better done using an Exploratory Data Analysis (EDA) of the data, we focus only on identifying the file structures the dataset is composed of, the files format, and a high-level description of each file and attribute. Moreover, we aim to extract if there is any recommended data split while training a ML model, and if there are any consistency rules or relevant statistics of the dataset.

3.3.5. Gathering

In the *gathering* dimension, we aim to capture relevant details about the data collection process. In Figure 5 we show an excerpt of the workflow where we aim to capture a description and infer the type of data collection process. To do so, we use an *In-context prompt*, and then we use an *Answer Refinement* prompt to refine the answer and obtain as much of the details of the process from the documentation. From the refined answer, we use a *Classification* prompt to classify the process by its type using a predefined list of types— (Giner-Miguelez et al., 2022)—and we then summarize the refined answer to extract a brief description of the process.

Afterwards, we ask about the team gathering the data. In that sense, we get a refined (combination between an in-context and a refining prompt) explanation to infer the type (*public*, *private*, or *crowdsourcing*) and see if the document provides any demographic information about them. Suppose we realize that the gatherers are a crowdsourcing service. In that case, we can add more steps to the chain asking for the company providing the service and the worker's labor condition as Díaz et al. (2022) propose.

[9] For instance, the Montreal license (Contractor et al., 2022)


Start
In-context QA
Query: How the dataset has been collected?
Answer refining
Classification: process gathering types
Summarize
In-context QA:
Query: Who collects the dataset?
Classification: collector team's type
Crowd -sourcing?
yes
In-context QA:
Labour conditions
In-context QA:
Query: Team demographics
In-context QA:
Query: Is the data collected from people?
From people?
yes
In-context QA:
Query: People demographics
In-context QA:
Query: Which the sources of the collection process?
Parse: for each source
In-context QA:
Query: Source description
In-context QA:
Query: Which tool of infrastructure have been used to collect the data? ?
In-context QA:
Query: Is there any timeframe of the collection?
In-context QA:
Query: Is there any geospatial reference?

Gathering

Description and type of the process

Type: Secondary Data Analysis

Description: The data was collected from 1,014 whole-body FDG-PET/CT examinations of 900 patients acquired between 2014 and 2018 as part of a prospective registry study.

Collection team and process demographics

Type: Internal team

Labour conditions: Do not apply

Team Demographics: Not provided.

Process demographics: Demographics limited to age, sex and study class

Sources of the data

Sources: Prospective registry study

Infrastructure: Not apply

Localization

Collection timeframe: Between 2'014 and 2018

Collection geospatial reference: Medical Faculty of the University of Tübigen

Figure 5: Example of a chain for the gathering dimension

Moreover, we intend to identify the sources of the data and the infrastructure used to collect it. To do so, we use an *In-context prompt*, adding the previously generated explanation of the gathering process to the context. In that sense, the goals are to identify the source (and any potential issue of the source) and the infrastructure used to collect the data. At last, we try to extract the geolocation and the timeframe of the process also using directly an *In-context prompt* as shown in Figure 5.

3.3.6. Annotation

This chain aims to capture details about how the dataset has been labeled, and Figure 6 shows an excerpt of its workflow. First, to get a description of the process and infer its type, the chain follows the same structure as the used in the gathering dimensions. Then, in the figure, we can see how we use an *In-Context prompt* and then an *Answer Refinement prompt* to get information about the team. Then, and following a similar structure to the one used for the gathering process, we classify the team as *internal*, *external* or a *crowd-workers service* using the obtained answers. Regarding the infrastructure, we ask for the tools and platforms used to annotate the data, parsing each one, and asking for specific details. Finally, in this dimension we also ask for the specific labels generated by the process, the validation methods applied to the labels, and the annotation guidelines shared with the annotation team.

3.3.7. Social concerns

In the *social concerns* dimension, we aim to identify the potential social issues of the data expressed in the documentation. In that sense, we identify the issues that may affect protected groups. To do so, we seek mentions of biases in the data, for instance, geographical biases (the data has been labeled only in one country), representativeness issues (the dark-skinned faces may be underrepresented in the dataset), sensitivity issues (e.g., whether the content may be offensive to a particular group of people), and privacy (e.g., whether the data may expose private data). To do so, we have built a set of in-context QA prompts asking about these issues.

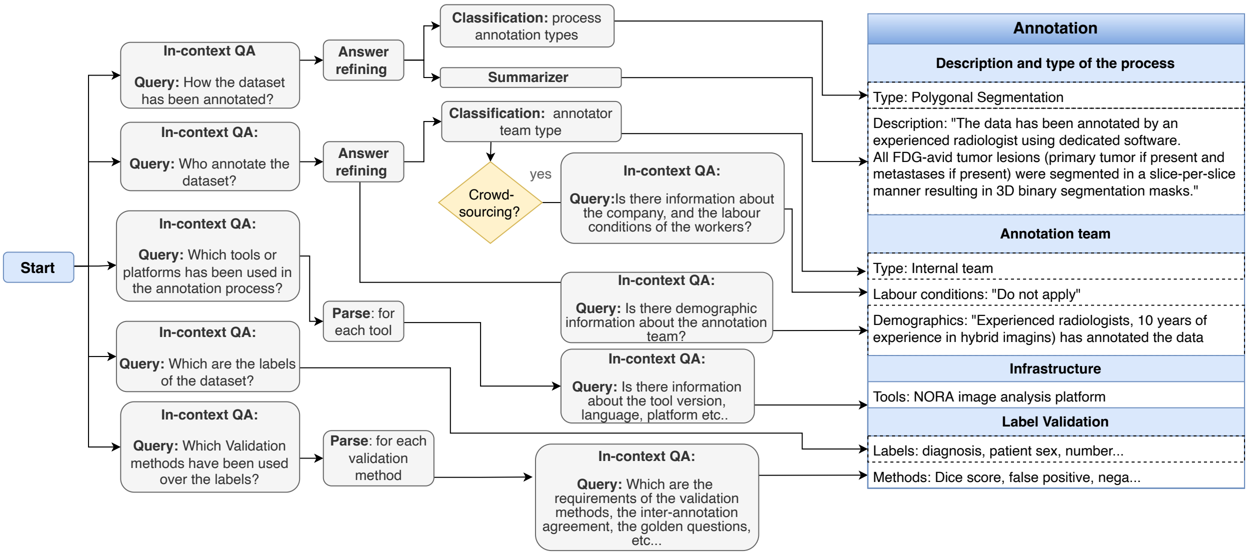

Start
In-context QA
Query: How the dataset has been annotated?
Answer refining
Classification: process annotation types
Summarizer
In-context QA:
Query: Who annotate the dataset?
Answer refining
Classification: annotator team type
Crowd-sourcing?
yes
In-context QA:
Query: Is there information about the company, and the labour conditions of the workers?
In-context QA:
Query: Which tools or platforms has been used in the annotation process?
Parse: for each tool
In-context QA:
Query: Is there demographic information about the annotation team?
In-context QA:
Query: Is there information about the tool version, language, platform etc..
In-context QA:
Query: Which are the labels of the dataset?
In-context QA:
Query: Which Validation methods have been used over the labels?
Parse: for each validation method
In-context QA:
Query: Which are the requirements of the validation methods, the inter-annotation agreement, the golden questions, etc...
Annotation
Description and type of the process
Type: Polygonal Segmentation
Description: "The data has been annotated by an experienced radiologist using dedicated software. All FDG-avid tumor lesions (primary tumor if present and metastases if present) were segmented in a slice-per-slice manner resulting in 3D binary segmentation masks."
Annotation team
Type: Internal team
Labour conditions: "Do not apply"
Demographics: "Experienced radiologists, 10 years of experience in hybrid imagins) has annotated the data
Infrastructure
Tools: NORA image analysis platform
Label Validation
Labels: diagnosis, patient sex, number...
Methods: Dice score, false positive, nega...

Figure 6: Example of a chain for the annotation dimension

#	Data paper title	Field	Journal	Year
1	A speech corpus of Quechua Collao for automatic dimensional emotion recognition (Paccotacya-Yanque et al., 2022)	Language	SData	2022
2	A patient-centric dataset of images and metadata for identifying melanomas using clinical context (Rotemberg et al., 2021)	Medical image	SData	2021
3	DeepLontar for handwritten Balinese character detection and syllable recognition on Lontar manuscript (Siahaan et al., 2022)	Language	SData	2022
4	A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions (Gatidis et al., 2022)	Medical image	SData	2022
5	An annotated image dataset for training mosquito species recognition system on human skin (Ong and Ahmad, 2022)	Biodiversity	SData	2022
6	The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria (Dobbins et al., 2022)	Medical language	SData	2022
7	A dataset for Kurdish handwritten digits and isolated characters recognition (Abdalla et al., 2023)	Language	DBrief	2023
8	A stance dataset with aspect-based sentiment from Indonesian COVID-19 vaccination-related tweets (Purwitasari et al., 2023)	Sentiment Analysis	DBrief	2023
9	An annotated dataset for event-based surveillance of antimicrobial resistance (Arınık et al., 2023)	Biodiversity	DBrief	2022
10	Human-annotated dataset for social media sentiment analysis for Albanian language (Kadriu et al., 2022)	Sentiment Analysis	DBrief	2022
11	DSAIL-Porini: Annotated camera trap image data of wildlife species from a conservancy in Kenya (Mugambi et al., 2023)	Biodiversity	DBrief	2022
12	Dataset of prostate MRI annotated for anatomical zones and cancer (Adams et al., 2022)	Medical image	DBrief	2022

Table 2
Sample of data papers used in the evaluation. SData stands for Scientific Data and DBrief for Data-in-Brief

4. Experimenting with scientific data journals

In this section, we validated our approach using a set of data papers published in scientific journals. We first describe them manually, and then we describe them using our approach. Comparing both descriptions, we present the results in Table 3 showing results for each of the dimensions presented in Section 2.

4.1. Sample selection

To validate our approach, we selected a sample of 12 data papers depicted in Table 2, from a diversity of scientific fields, published in two different scientific data journals: Nature's *Scientific Data*[1] and Elsevier's *Data in Brief*[2]. These data journals, among others, publish peer-reviewed manuscripts (data papers) describing the datasets of a wide range of scientific disciplines. They represent a notorious effort from the scientific community to assess the quality and reusability of the data, and we believe is a good test suite to evaluate our proposal.

4.2. A dataset of annotated data papers

From the selected sample of data papers, and to properly evaluate our method, we created a dataset of ground truth labels describing each of the dimensions we presented in Section 2. To create the labels, the authors have manually described each of the datasets using their data papers according to the dimension presented in Section 2. At least two authors have described each dataset, and, in case of conflicts, the participants have agreed on the final version of the description in subsequent meetings. Each row of the resultant dataset is composed of the DOI and title of the data paper, the specific dimension and sub-dimension of Section 2, and the agreed description from the authors as ground truth labels. The resultant dataset has been released (Giner-Miguelez et al., 2023e) in two different formats together with the experiment results to enable future research in this area.

4.3. Experiment setup and validation

After creating the ground truth labels, we described the data papers using our method. To test the consistency of our approach across different LLMs, we have tested it using two different LLMs: GPT3.5 and Flan-UL2. GPT3.5 (text-davinci-003) is a decoder-only model trained on the GPT family (Ouyang et al., 2022), and we have used the API service offered by the company owning the model. The Flan-UL2 is an encoder-decoder language model (Tay et al., 2023) fine-tuned using the "Flan" prompt tuning and dataset collection (Chung et al., 2022), and we have used the model deployed in Huggingface[10] for the experiment. The embeddings used in the retriever phase were created using GPT3.5 (*text-embedding-ada-002*) using the API provided by the owner, and the retriever was implemented using the open-source library FAISS (Johnson et al., 2019) for similarity search. The chains were implemented using LangChain (Chase, 2023).

Finally, once we got the results, we evaluated them by comparing both descriptions, the manual we did, and the one generated by the models. We evaluated whether the obtained results were correct or not (*accuracy*), and in the case of being incorrect, whether the results were truthful with the source document (*faithfulness* (Maynez et al., 2020)). In that sense, erroneous results that were not truthful with the source have been annotated as hallucinations. To ensure the quality of the annotations, at least two authors have evaluated each of the datasets, and the inter-annotation agreement is provided together with the results. A replication package (Giner-Miguelez et al., 2023e) has been released along with the annotated dataset, including the source documents and code used in the experiment.

4.4. Experiment results

In Table 3, we can see a summary of the results obtained using our approach on the sample datasets. They are classified according to the dimension and sub-dimensions presented in Section 2. For each one of them, we provide the average accuracy—whether the answers have been marked as correct by the annotators—and the average hallucinations—whether the answer have been marked as untruthful/unfaithful with respect to the source documentation (Maynez et al., 2020; Creswell and Shanahan, 2022)—. Finally, we provide the results of our approach using two different LLMs: GPT3.5 (text-davinci-003) and Flan-UL2. The complete results and the raw data of the extraction for each dataset can be accessed in the online repository.

As can be seen in Table 3, our method shows good overall accuracy using both models in most dimensions. GPT3.5 (81,21%) shows slightly better accuracy than FLAN-UL2 (69,13%), but is more prone to be overconfident in its answers. The accuracy varies depending on the dimensions, showing that some are more difficult to extract than others. Particularly, it is in these dimensions that are more difficult to extract, where we observed a tendency from the LLMs to give unfaithful answers. It is worth mentioning that the accuracy also reflects the ability of the approach to detect whether the information is present or not. In that sense, if the information is not present in the documentation, and the answer of the chain indicates that, this answer was annotated as correct. We have observed that the method is also good at detecting which dimensions are not covered by the documentation.

Analyzing the results specifically by each dimension, we can see that our method gets worse results in some specific dimensions. One of the reasons is the confusion of the LLMs with similar information. For instance, little information is reported about the licenses of the datasets in the documents, but, at the same time, these documents contain licensing information about the scientific publication itself which confuses the extraction process. In the same way, in datasets where the sources are not clearly described, the LLMs tend to confuse the authors' affiliation and the papers' publication dates as the geographical and temporal location of the collection process. Additionally, the validation methods of the labels are usually not present in the documentation and tend to be confused with the validation methods of the whole dataset (ML Benchmark of the Uses dimensions). We have observed that these processes are difficult to identify in this kind of documentation, and there is room to investigate new approaches to face this complexity.

[10] https://huggingface.co/google/flan-ul2

Table 3
Results of the experiment for each dimension and subdimensions. *Accuracy* are the results annotated as correct by the team, and *Unfaith* are the results annotated as not truthful with the source documentation

		GPT3 (text-davinci-003)			**FLAN-UL2**		
Dimension	**Subdimension**	Accuracy	Unfaith	Overall	Accuracy	Unfaith	Overall
Uses	Design intentions	90,91%	0%	**88,64%**	91,67%	0%	**71,46%**
	Recommendations	91,97%	0%		72,73%	0%	
	ML Benchmarks	83,33%	0%		50%	16,67%	
Contributors	Authors	100%	0%	**97,22%**	100%	0%	**86,11%**
	Funding	100%	0%		91,67%	0%	
	Maintenance	91,67%	0%		66,67%	0%	
Distribution	Accessibility	58,33%	16,67%	**55,56%**	41,67%	0%	**47,22%**
	Licenses	8,33%	41,67%		25%	8,33%	
	Deprecation policies	100%	0%		75%	0%	
Composition	Data records	91,67%	0%	**88,89%**	91,67%	0%	**80,56%**
	Data splits	83,33%	0%		100%	0%	
	Statistics	91,67%	0%		50%	0%	
Gathering	Description & type	91,67%	0%	**70,83%**	83,33%	0%	**58,33%**
	Team	100%	0%		66,67%	0%	
	Source & infrastructure	50%	16,67%		41,67%	0%	
	Localization	41,67%	8,33%		41,67%	8,33%	
Annotation	Description & type	100%	0%	**81,25%**	45,45%	0%	**63,26%**
	Team	100%	0%		90,91%	0%	
	Infrastructure	83,33%	0%		83,33%	0%	
	Validation	41,67%	25%		33,33%	0%	
Social Concerns	Bias	83,33%	0%	**86,11%**	75%	0%	**77,02%**
	Sensitivity data	91,67%	0%		72,73%	0%	
	Privacy	83,33%	0%		83,33%	0%	

Focusing on the hallucination results—whether the answers are not truthful with respect to the source documentation—we can see that these are greater in GPT-3, that tends to be more overconfident than Flan-UL2 in the answers. These unfaithful results tend to appear in the dimensions where the LLMs struggle to get the correct answer. However, after the experiment, we have analyzed these unfaithful results, classifying them as extrinsic hallucinations—when the answer is not based on the source documents—or intrinsic—where the answer is incorrect but corresponds to information contained in the documents (Ji et al., 2023). In this sense, practically all the hallucination issues we analyzed were intrinsic. Thus, we observed that using retrieval augmentation strategies—apart from reducing hallucinations issues, as Shuster et al. (2021) propose—also tends to change the type of the hallucinations to intrinsic.

5. Discussion

The results presented in the last section show that the use of LLMs along with our method exhibit good accuracy overall for extracting the demanded dimensions from raw dataset documentation. However, not all the analyzed dimensions present the same accuracy, and despite the low rate of hallucinations found in this experiment, they are still a problem to be solved. With these challenges in mind, we present a discussion about the potential application of our method in legal compliance and in data discoverability, its limitations, and the need for a mature toolkit environment.

5.1. Assessing the compliance with AI regulations

Our method has shown good accuracy in determining whether the dimensions were at least present in the documentation. Current AI regulation demands some of the dimensions we extracted in this work, for instance, information about the annotation and collection process. In this sense, our work could aid in checking dataset documentation compliance with emerging AI regulations. However, these regulations are not yet fully deployed, and new dimensions may be added in further versions. In future work, we intend to follow up on their deployment and adapt our method to the potential changes.

5.2. Automating the dataset's discoverability

Furthermore, novel initiatives to improve dataset discoverability and reuse are beginning to emerge in the ML community. Initiatives such as DescribeML (Giner-Miguelez et al., 2023f), a domain-specific language to describe datasets, or Croissant (MLCommons, 2023), a high-level format for describing machine learning datasets based on Schema.org[11], propose to create machine-readable dataset documentation that can be easily indexed by search engines such as the popular Google Dataset Search[12]. As some of the dimensions these proposals demand are already covered by of our work, our approach could easily be adapted to automatically generate these structured metadata, facilitating the discoverability of well-documented datasets.

5.3. A path to face hallucination issues

As stated before, the majority of hallucination issues in our experiments have been intrinsic. Moreover, by analyzing the dimensions where these hallucinations happened, we have been able to detect the root causes of some of them (for instance, the confusion between the legal condition of the data papers and the datasets themselves). This opens the path to work in solving these hallucinations by fine-tuning the prompts or adding specific validation steps throughout the chains as other works are starting to purpose in the ML community (Dhuliawala et al., 2023). We believe this could also help in other types of QA processes. We plan to further investigate this in the future.

5.4. Towards a toolkit for analyzing datasets using LLMs

Finally, the results of this study open a path for developing a mature tooling environment around ML datasets using LLMs. For instance, smart assistants to help data creators during the documentation process of the data, or similar tools as the one we presented but tailored to specific fields (medical, biodiversity, social science, etc.). However, using LLMs is challenging regarding computational resources and speed due to their large size. In that sense, the new set of emerging LLMs opens a path to explore the capabilities of smaller models, or fine-tuned versions, for cheaper and faster inference. We think our experiments (along with the dataset released) can be used as another benchmark for new LLMs to appear in order to analyze their trade-offs regarding different types of fine-grained data extraction tasks.

6. Tool support

To facilitate the adoption of our method, we have developed *DataDoc Analyzer* (Giner-Miguelez et al., 2023b), an open-source tool to analyze the documentation of scientific datasets. The tool implements an ML pipeline, including the proposed approach at its core, to extract the demanded dimensions from the dataset documentation and provide a completeness report. The tool is presented with two user interfaces: a demo-based one, intended for test purposes; and an API, ready to be integrated into any data pipeline.

The workflow of the tool, depicted in Figure 7, is composed of three stages. The *Data preprocess* stage and the *Dimensions extraction* stage are the stages implementing the presented approach in this work and are already discussed in Section 3. The output is the set of *extracted dimensions*. On top of this output, the tool

[11] https://schema.org/

[12] https://datasetsearch.research.google.com/

implements a *Post-processing* stage that analyzes the extracted dimensions to evaluate if the ML pipeline was able to find all the answers of the requested dimensions, providing as an output a *completeness report*.

The *Post-processing* stage performs a zero-shot classification using a distilled version of BART (Lewis et al., 2020), fine-tuned on the MultiNLI (MNLI) dataset (Williams et al., 2018). We provide the model with a set of meaningful categories for each dimension (for instance, "there is demographic information of the gathering" and "there is no demographic information of the gathering"), and we get the one with the higher probability score. Then, we compile all of the responses to create the report that allows the user to assess the overall completeness of the documentation regarding the requested dimensions. Figure 8 shows an excerpt of the extracted dimensions and the completeness report for the annotation process of the "*DeepLontar dataset*

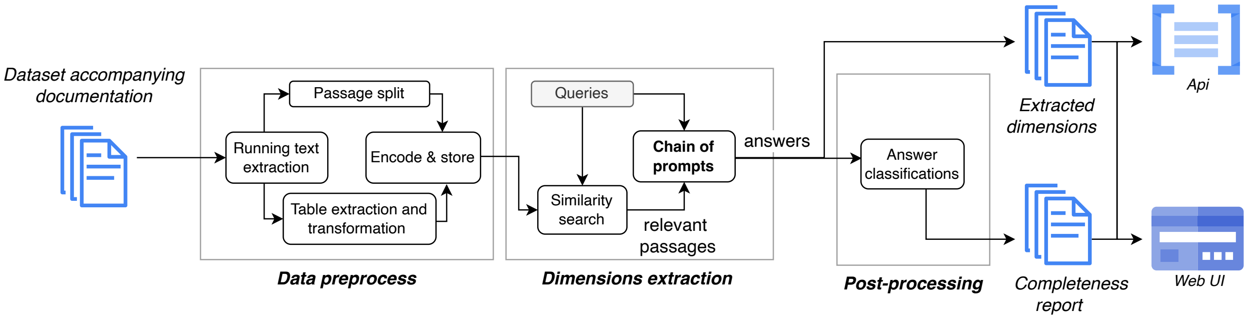
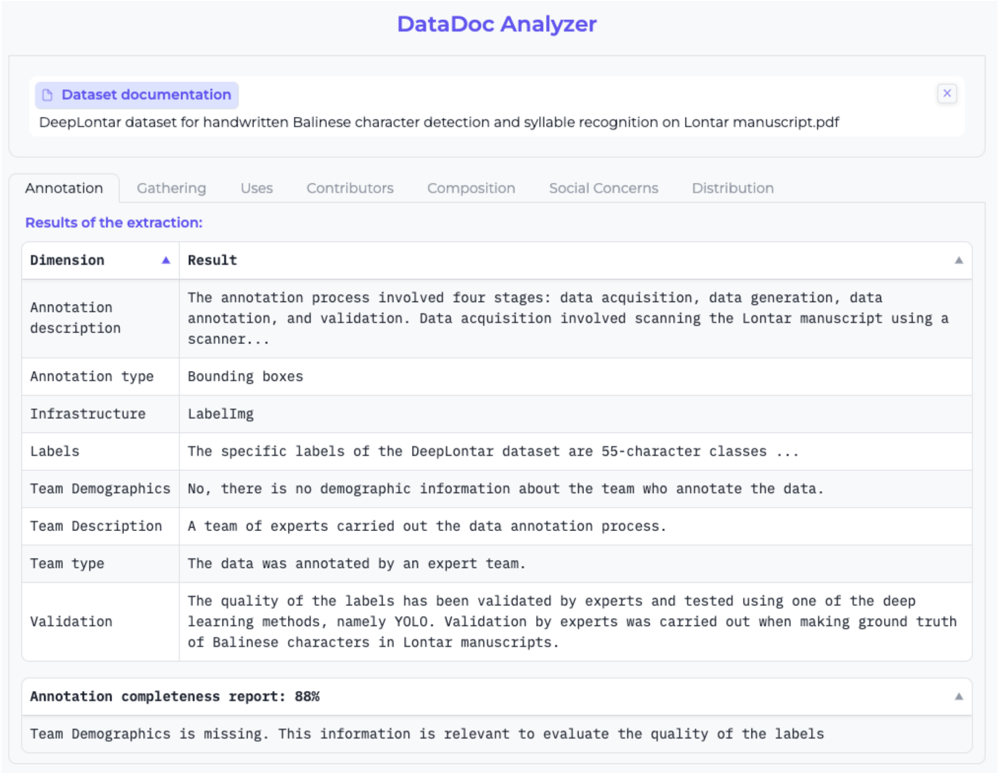


Figure 7: Workflow of the tool


DataDoc Analyzer

DeepLontar dataset for handwritten Balinese character detection and syllable recognition on Lontar manuscript.pdf

Results of the extraction:

Dimension ▲	Result ▲
Annotation description	The annotation process involved four stages: data acquisition, data generation, data annotation, and validation. Data acquisition involved scanning the Lontar manuscript using a scanner...
Annotation type	Bounding boxes
Infrastructure	LabelImg
Labels	The specific labels of the DeepLontar dataset are 55-character classes ...
Team Demographics	No, there is no demographic information about the team who annotate the data.
Team Description	A team of experts carried out the data annotation process.
Team type	The data was annotated by an expert team.
Validation	The quality of the labels has been validated by experts and tested using one of the deep learning methods, namely YOLO. Validation by experts was carried out when making ground truth of Balinese characters in Lontar manuscripts.

Annotation completeness report: 88% ▲

Team Demographics is missing. This information is relevant to evaluate the quality of the labels

Figure 8: Tool web UI of an extraction and completeness report of the annotation process of the *DeepLontar dataset for handwritten Balinese character detection and syllable recognition on Lontar manuscript*

for handwritten Balinese character detection and syllable recognition on Lontar manuscript" used in the evaluation section.

The tool is presented with two distinct user interfaces: a Web UI designed to demonstrate the tool's capabilities and suitable for analyzing a single document, and an API ready to be integrated into any data processing pipeline. The API comprises endpoints that analyze the documentation for each dimension and return the results in JSON format, ready to be ingested into any data processing pipeline. We implemented the Web UI with Graddio[13] and the API with FastAPI[14]. The tool comes with documentation and usage instructions that can be found in the repository, and a docker image (Giner-Miguelez et al., 2023c) is provided to facilitate its deployment. Regarding response time, processing unseen documents takes between 50 and 60 seconds. For already-seen documents time goes down to between 20 and 25 seconds for each dimension since the data processing phase is cached. The tool can be found in an open repository (Giner-Miguelez et al., 2023a) and a live demo can be found as a space in Huggingface (Giner-Miguelez et al., 2023d).

7. Related work

In this section, we review current extraction techniques over scientific documentation that focus on similar targets to our work (datasets and/or dimensions of interest for the ML community), and works that use similar extraction methods to ours (LLMs and prompting strategies) for other extraction domain-specific tasks over scientific documents. Following this review, we conclude that no extraction technique focuses on extracting the dimensions demanded by the ML community; however, similar extraction methods demonstrate promising capabilities in other domains, encouraging us to adapt them to our use case.

Similar extraction targets — Deep-learning methods are currently state-of-the-art in extracting information from unstructured text (Abdullah et al., 2023; Zhang et al., 2024). These methods have also been applied to technical and scientific documentation to extract insights from them, commonly focusing on sentence classification, relation extraction (RE), and the named entity recognition (NER) of scientific and technical terms of the papers (Ma et al., 2023; Rybinski et al., 2023). Brack et al. (2022) provide an overview of the existing works focusing on similar document, where only a few of them focus on extracting insights from the accompanying resources of the studies, such as the used dataset. These particular works have focused on profiling the accompanying resources of the studies to promote its search and recommendation systems (Zheng et al., 2021). The authors intend to identify the resources—such as datasets, software, or algorithms—used in a study, extracting their relations and roles with respect to the overall study. Regarding our work, these proposals show how profiling information of the accompanying resources—such as datasets—could improve their search and comparison capabilities. However, the target extraction for these assets are generic, and datasets are treated as generic asset without diving into the dimension demanded by the community.

Particularly in works focusing on profiling datasets, we find methods to detect the mentions of a dataset in a scientific publication, the tasks the dataset is designed for, and the ML metrics if tested with any ML approach (Kabongo et al., 2023; Otto et al., 2023). In addition, to promote the evaluation of these datasets, Färber et al. (2021) propose to extract the mentions to the dataset together with the research methods of the scientific papers and annotate their metadata with these mentions. Compared to ours, these works partially overlap the ML Benchmarks sub-dimension from the Uses dimension, where we also aim to extract the dataset's task, model, and metrics. Still, no other dimension is extracted by these works.

Similar extraction methods — Several works use LLMs combined with different prompting strategies to work on specific scientific extraction tasks. For instance, in the material domain, Dunn et al. (2022) explores the capabilities of LLMs to build relationships between materials by extracting material information from different scientific databases. On the other hand, Park et al. (2023) explores using LLMs to extract

[13] https://gradio.app/

[14] https://fastapi.tiangolo.com/

automatic molecular interaction from different scientific documentation. More closely to our work, Kunnath et al. (2023) explores prompting strategies for adding context to scientific citation. The authors use LLMs to contextualize the citation to provide a better search experience for researchers inspecting scientific publications. Finally, Patiny and Godin (2023) presents an ongoing work focused on analyzing scientific data, but focused on experimental data of molecules. Despite being far from our target extraction, the recent emergence of this kind of work shows the potential of LLMs and prompting strategies to extract specific information. These works could suggest other types of prompting strategies that could also be useful in our domain of interest.

In conclusion, as far as we know, there are no other works focusing on the extraction of the dimensions we target following an LLM-based approach as we do and our method could complement other existing works to improve the analysis of dataset descriptions.

8. Conclusions

In this work, we have presented an approach that explores using LLMs to automatically enrich the description of machine learning dataset by extracting key information—categorized in a set of dimensions and sub-dimensions—from its raw documentation. We have validated our approach using two different LLMs—GPT3.5 and FLAN-UL2—over a set of 12 datasets published in scientific journals. The results have shown an overall good accuracy of the approach, including a low hallucination rate. However, the accuracy vary depending on the dimension extracted. Particularly, GPT3.5 shows slightly better accuracy in the extraction while Flan-UL2 tends to hallucination less. The hallucinations detected during our experiments have been mainly intrinsic due to, in part, to the in-context retrieval augmented strategy we have used to build the prompts. Aside from the good overall results in extracting the demanded dimensions, our method has shown good accuracy in determining whether the dimensions were present or not in the documentation. Our results can be reused and replicated thanks to the release of an open-source tool, a public demo, and all the data used in our experiment during the validation.

As a further work, following the discussion in Section 5, we plan to follow up on the deployment of the new AI regulation. Because of the recent and rapid growth of AI technologies, future versions of this regulation will likely be updated, adding, for example, new dimensions to the data documentation requirements. Our plan is to adapt our method to these changes in order to assist practitioners with checking the legal compliance of their applications. On the other hand, dataset search engines (such as Google Dataset Search (Brickley et al., 2019)) have started including machine-readable dataset documentation (via the Croissant initiative (MLCommons, 2023)). We intend to adapt our method to generate indexable documentation from raw dataset documentation, such as scientific data publications, thereby assisting data publishers in increasing data discoverability. Finally, the fast-paced field around LLM is encouraging the apparition of new LLM approaches and prompting strategies. LLM with fewer computational requirements and new prompting strategies to reduce hallucinations are two promising research directions for improving our method. In that regard, we intend to investigate how these new LLM approaches and prompting strategies can help to improve the accuracy of some dimensions while reducing hallucinations and computational demands.

References

Abdalla, P. A., Qadir, A. M., Shakor, M. Y., Saeed, A. M., Jabar, A. T., Salam, A. A., and Amin, H. H. H. (2023). A vast dataset for kurdish handwritten digits and isolated characters recognition. *Data in Brief*, 47:109014.

Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., and Talpur, N. (2023). Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges. *IEEE Access*, 11:10535–10562.

Adams, L. C., Makowski, M. R., Engel, G., Rattunde, M., Busch, F., Asbach, P., Niehues, S. M., Vinayahalingam, S., van Ginneken, B., Litjens, G., et al. (2022). Dataset of prostate mri annotated for anatomical zones and cancer. *Data in Brief*, 45:108739.

Arınık, N., Van Bortel, W., Boudoua, B., Busani, L., Decoupes, R., Interdonato, R., Kafando, R., van Kleef, E., Roche, M., Syed, M. A., et al. (2023). An annotated dataset for event-based surveillance of antimicrobial resistance. *Data in Brief*, 46:108870.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bishop, B. W., Neish, P., Kim, J. H., Bats, R., Million, A., Carlson, J., Moulaison-Sandy, H., and Pham, M. T. (2023). Data management plan implementation, assessments, and evaluations: Implications and recommendations. *Data Science Journal*, 22:27.

Brack, A., Hoppe, A., Stocker, M., Auer, S., and Ewerth, R. (2022). Analysing the requirements for an open research knowledge graph: use cases, quality requirements, and construction strategies. *International Journal on Digital Libraries*, 23(1):33–55.

Brickley, D., Burgess, M., and Noy, N. (2019). Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, WWW '19, page 1365–1375, New York, NY, USA. Association for Computing Machinery.

Chase, H. (2023). LangChain main repository. Accessed: November 2023.

Chavan, V. and Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12:1–12.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Contractor, D., McDuff, D., Haines, J. K., Lee, J., Hines, C., Hecht, B., Vincent, N., and Li, H. (2022). Behavioral use licensing for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 778–788, New York, NY, USA. Association for Computing Machinery.

Creswell, A. and Shanahan, M. (2022). Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., and Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., and Denton, E. (2022). Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2342–2351, New York, NY, USA. Association for Computing Machinery.

Dobbins, N. J., Mullen, T., Uzuner, Ö., and Yetisgen, M. (2022). The leaf clinical trials corpus: a new resource for query generation from clinical trial eligibility criteria. *Scientific Data*, 9(1):490.

Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A. S., Ceder, G., Persson, K., and Jain, A. (2022). Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Feger, S. S., Wozniak, P. W., Lischke, L., and Schmidt, A. (2020). 'Yes, I comply!' motivations and practices around research data management and reuse across scientific fields. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

Färber, M., Albers, A., and Schüber, F. (2021). Identifying used methods and datasets in scientific publications. In *Proceedings of the Workshop on Scientific Document Understanding: co-located with 35th AAAI Conference on Artificial Inteligence (AAAI 2021) ; Remote, February 9, 2021.*, volume 2831 of *CEUR Workshop Proceedings*. RWTH Aachen.

Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., and Rubin, D. (2022). A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2022). Describeml: a tool for describing machine learning datasets. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, pages 22–26.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023a). Code repository of: Datadoc analyzer: A tool for analyzing the documentation of scientific datasets. `https://github.com/SOM-Research/DataDoc-Analyzer`. Accessed on 23.11.2023.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023b). Datadoc analyzer: A tool for analyzing the documentation of scientific datasets. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5046–5050.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023c). Datadoc analyzer docker image. `https://hub.docker.com/r/joangi/datadoc_analyzer`. Accessed on 23.11.2023.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023d). Datadoc analyzer public demo. `https://huggingface.co/spaces/JoanGiner/DataDoc_Analyzer`. Accessed on 23.11.2023.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023e). Repository of the replication package and data generated of this work. `https://github.com/SOM-Research/Dataset-Docs-Enrichment`. Accessed on 23.12.2023.

Giner-Miguelez, J., Gómez, A., and Cabot, J. (2023f). A domain-specific language for describing machine learning datasets. *Journal of Computer Languages*, 76:101209.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The dataset nutrition label. *Data Protection and Privacy*, 12:1–26.

Izacard, G. and Grave, É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. (2022). Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):38.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Kabongo, S., D'Souza, J., and Auer, S. (2023). Orkg-leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, pages 1–14.

Kadriu, F., Murtezaj, D., Gashi, F., Ahmedi, L., Kurti, A., and Kastrati, Z. (2022). Human-annotated dataset for social media sentiment analysis for albanian language. *Data in Brief*, 43:108436.

Kunnath, S. N., Pride, D., and Knoth, P. (2023). Prompting strategies for citation classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1127–1137, New York, NY, USA. Association for Computing Machinery.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., and Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677.

Luccioni, A. S., Corry, F., Sridharan, H., Ananny, M., Schultz, J., and Crawford, K. (2022). A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 199–212, New York, NY, USA. Association for Computing Machinery.

Ma, Y., Liu, J., Lu, W., and Cheng, Q. (2023). From "what" to "how": Extracting the procedural scientific information toward the metric-optimization in ai. *Information Processing & Management*, 60(3):103315.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

McMillan-Major, A., Osei, S., Rodriguez, J. D., Ammanamanchi, P. S., Gehrmann, S., and Jernite, Y. (2021). Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 121–135, Online. ACM.

Micheli, M., Hupont, I., Delipetrev, B., and Soler-Garrido, J. (2023). The landscape of data and AI documentation approaches in the european policy context. *Ethics and Information Technology*, 25(4):56.

MLCommons (2023). Croissant: a high-level format for machine learning datasets. `https://github.com/mlcommons/croissant`. Accessed: November 2023.

Mugambi, L., Kabi, J. N., Kiarie, G., and wa Maina, C. (2023). Dsail-porini: Annotated camera trap image data of wildlife species from a conservancy in kenya. *Data in Brief*, 46:108863.

Ong, S.-Q. and Ahmad, H. (2022). An annotated image dataset for training mosquito species recognition system on human skin. *Scientific Data*, 9(1):413.

Otto, W., Zloch, M., Gan, L., Karmakar, S., and Dietze, S. (2023). Gsap-ner: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8166–8176.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems, NeurIPS*, volume 35, pages 27730–27744. Curran Associates, Inc.

Paccotacya-Yanque, R. Y., Huanca-Anquise, C. A., Escalante-Calcina, J., Ramos-Lovón, W. R., and Cuno-Parari, Á. E. (2022). A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

Park, G., Yoon, B.-J., Luo, X., Lpez-Marrero, V., Johnstone, P., Yoo, S., and Alexander, F. (2023). Automated extraction of molecular interactions and pathway knowledge using large language model, galactica: Opportunities and challenges. In Demner-fushman, D., Ananiadou, S., and Cohen, K., editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 255–264, Toronto, Canada. Association for Computational Linguistics.

Patiny, L. and Godin, G. (2023). Automatic extraction of fair data from publications using LLM. Cambridge: Cambridge Open Engage. This content is a preprint and has not been peer-reviewed.

Purwitasari, D., Putra, C. B. P., and Raharjo, A. B. (2023). A stance dataset with aspect-based sentiment information from indonesian covid-19 vaccination-related tweets. *Data in Brief*, 47:108951.

Romary, L. and Lopez, P. (2015). GROBID - Information Extraction from Scientific Publications. *ERCIM News*, 100.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34.

Rybinski, M., Wan, S., Karimi, S., Paris, C., Jin, B., Huth, N., Thorburn, P., and Holzworth, D. (2023). Sciharvester: Searching scientific documents for numerical values. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3135–3139, New York, NY, USA. Association for Computing Machinery.

Schick, T. and Schütze, H. (2021). Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Siahaan, D., Sutramiani, N. P., Suciati, N., Duija, I. N., and Darma, I. W. A. S. (2022). Deeplontar dataset for handwritten balinese character detection and syllable recognition on lontar manuscript. *Scientific Data*, 9(1):761.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Bahri, D., Schuster, T., Zheng, S., et al. (2023). UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., et al. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific data*, 8(1):192.

Thuermer, G., Guardia, E. G., Reeves, N., Corcho, O., and Simperl, E. (2023). Data management documentation in citizen science projects: Bringing formalisation and transparency together. *Citizen Science: Theory and Practice*, 8(1):25.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ye, S., Hwang, H., Yang, S., Yun, H., Kim, Y., and Seo, M. (2024). In-context instruction learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. note: to be published.

Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., and Jiang, M. (2022). Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Zhang, H., Zhang, C., and Wang, Y. (2024). Revealing the technology development of natural language processing: A scientific entity-centric perspective. *Information Processing & Management*, 61(1):103574.

Zheng, A., Zhao, H., Luo, Z., Feng, C., Liu, X., and Ye, Y. (2021). Improving on-line scientific resource profiling by exploiting resource citation information in the literature. *Information Processing & Management*, 58(5):102638.