Metadata Enhancement Using Large Language Models

Hyunju Song, Steven Bethard, Andrea K. Thomer
University of Arizona
{hyunjusong,bethard,athomer}@arizona.edu

Abstract

In the natural sciences, a common form of scholarly document is a *physical sample record*, which provides categorical and textual metadata for specimens collected and analyzed for scientific research. Physical sample archives like museums and repositories publish these records in data repositories to support reproducible science and enable the discovery of physical samples. However, the success of resource discovery in such interfaces depends on the completeness of the sample records. We investigate approaches for automatically completing the scientific metadata fields of sample records. We apply large language models in zero and few-shot settings and incorporate the hierarchical structure of the taxonomy. We show that a combination of record summarization, bottom-up taxonomy traversal, and few-shot prompting yield an F1 score as high as 0.928 on metadata completion in the Earth science domain.

1 Introduction

Not all scholarly documents are formal scientific articles. In the natural sciences (e.g. Earth science, biodiversity science, archaeology), a common form of scholarly document is a *physical sample record*. Also called *catalog records* or *specimen records*, sample records are documents written by scientists that describe samples (for instance, fossils, soil samples, sediment cores, etc.) collected and analyzed for scientific research. Since the 1980s, millions of dollars of grant funds have been dedicated to projects digitizing and sharing physical sample records through online data repositories (Nelson and Ellis, 2018; com, 2020). The specific contents of records varies by domain, but they typically contain metadata describing the type, material, and provenance of samples. Sample records are needed to find and reuse physical samples, and, when aggregated, can be treated as a scientific dataset in and of themselves. For instance, aggregated biodiversity sample records are frequently used for studies of global change biology (Heberling et al., 2021). Thus, the quality of these scholarly documents is crucial.

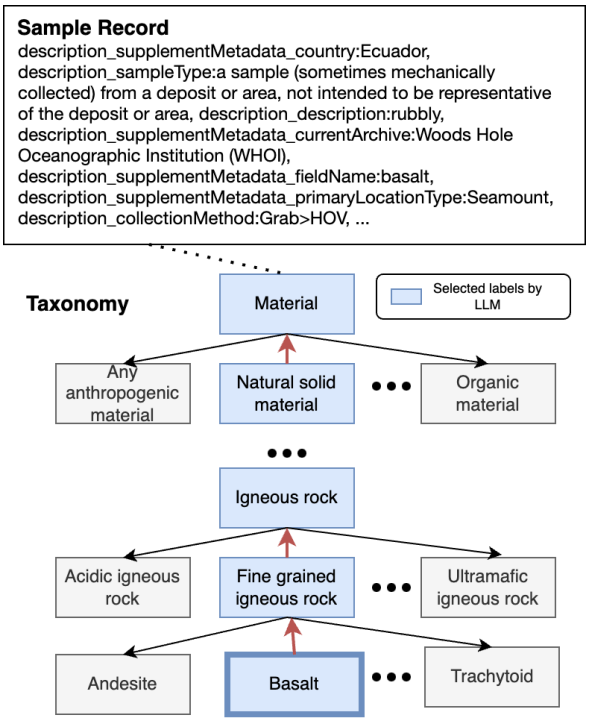


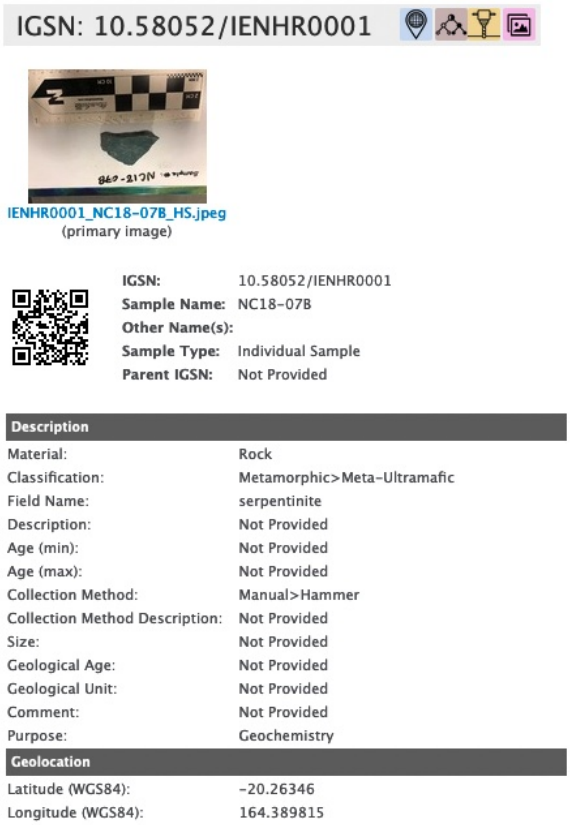
Figure 1: An illustration of our automatic metadata assignment workflow utilizing text classification with LLMs. The example record MATERIAL metadata field is predicted as result of tree traversal done on the taxonomy in a bottom-up manner.

Sample record metadata may be incomplete for a variety of reasons, including data entry issues or schema mismatches when sample records are aggregated across different repositories. For example, among the sample records that describe the resources registered in the System for Earth Sample Registration (SESAR2) data repository[1] (Fig 2), 78% of the records, totaling 3,565,478

[1] https://www.geosamples.org/

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 145–154

records, exhibited missing values in the MATERIAL metadata field. Automatic metadata assignment has consequently emerged as a critical task in the processing of sample records and other metadata aggregations. For instance, some data repositories automatically assign geocoordinates and higher geographic classifications based on text-based locality descriptions using rules-based workflows (Chapman and Wieczorek, 2020). Other approaches to inferring missing categorical fields often represent metadata fields as vectors and utilized classifiers to fill in missing values from a controlled vocabulary (Han et al., 2003; Paynter, 2005).


IGSN: 10.58052/IENHR0001

IENHR0001_NC18-07B_HS.jpeg
(primary image)

IGSN:	10.58052/IENHR0001
Sample Name:	NC18-07B
Other Name(s):	
Sample Type:	Individual Sample
Parent IGSN:	Not Provided

Description	
Material:	Rock
Classification:	Metamorphic>Meta-Ultramafic
Field Name:	serpentinite
Description:	Not Provided
Age (min):	Not Provided
Age (max):	Not Provided
Collection Method:	Manual>Hammer
Collection Method Description:	Not Provided
Size:	Not Provided
Geological Age:	Not Provided
Geological Unit:	Not Provided
Comment:	Not Provided
Purpose:	Geochemistry
Geolocation	
Latitude (WGS84):	-20.26346
Longitude (WGS84):	164.389815

Figure 2: An example of a physical sample record from SESAR2.

Often sample records contain many text fields that can be used to predict or infer missing values; thus, we approach automatic metadata assignment as a taxonomy-driven text classification problem. We propose a solution for metadata assignment based on large language models (LLMs) (fig. 1).

Our contributions are as follows:

- We compare zero and few-shot LLM formulations of the problem with competitive baselines like fine-tuning RoBERTa.
- We study different strategies to incorporate the taxonomy's hierarchy and show that a bottom-up approach is especially effective.
- We demonstrate that prompting LLMs to first summarize the sample record and then classify using the summary can sometimes outperform prompting LLMs to classify directly from the sample record.

Our evaluation focuses on sample records from the Earth science domain, though our approaches are general enough to be applied to many other domains. Our findings suggest that LLMs could provide a viable solution for digital libraries or aggregations facing missing metadata issues.

2 Related Work

Metadata Enhancement in Digital Libraries A common challenge in digital libraries or metadata aggregations is the "missing data" problem (Dushay and Hillmann, 2003). To tackle this issue, various automatic metadata assignment approaches including machine learning techniques have been employed. Paynter (2005); Tonkin and Muller (2008) used classification methods to automatically assign metadata values. Unlike these supervised learning methods, our approach employs LLMs with zero or few-shot samples to offer a more domain-agnostic solution.

LLMs for Text Classification With the advancement of larger-scale pre-trained language models trained on vast unlabeled corpora, approaches that utilize language models with instructions (zero-shot) and task-related examples directly in the prompt (few-shot) have been used. Prompting methods (Kojima et al., 2022; Wei et al., 2022) and domain-specific LLMs (Singhal et al., 2022; Taylor et al., 2022) have been investigated for text classification. Our work explores text classification with LLMs to assign metadata values.

Hierarchical Multi-Label Text Classification The task of assigning multiple hierarchically structured categories to a text is known as hierarchical multi-label text classification. Various methods have been explored to encode the hierarchical nature of the label space, including level-wise attention-based recurrent networks (Huang et al., 2019), graph convolutional networks (Xu et al., 2021), and LLMs with tree search strategies (Boyle et al., 2023). We also apply tree search strategies to LLMs, exploring both top-down and bottom-up

approaches, which we find to be much more successful in our data.

3 Dataset

We are working with physical sample records from the Internet of Samples (iSamples) project, which aggregates physical sample metadata records from data repositories within the domains of Earth science, bioscience, and archaeology (Davies et al., 2021). To aggregate these records, iSamples developed a common metadata schema and controlled vocabularies for the MATERIAL and OBJECT TYPE fields of each sample record (Richard et al., 2024), to which each repository's records are crosswalked. This makes it possible to easily search across a diverse range of records within the iSamples interface. The MATERIAL and OBJECT TYPE fields are used as two primary facets in the search interface.

Here, we focus on records from SESAR2, which describe specimens from the Earth science domain, encompassing rocks, fossils, fluids, and other materials registered. A substantial portion of these records contains missing values in the MATERIAL field, which negatively impacts the discoverability within both iSamples and SESAR2. However, the records contain other metadata which can be used to infer the MATERIAL type.

To frame this as a text classification dataset, for each sample record, we construct the output from the metadata field to be predicted, and construct the input from all other available metadata fields. In section 4 we explore various ways of concatenating the many metadata fields into a single text for input, but in all cases, we include the following text in the text classification prompt:

```
You are a scientist. Your task is
to analyze the description of a material
sample and determine the kind of material
that constitutes it after <<<>>> into one
of the predefined material types:
{taxonomy}
You will only respond with the material
type. Do not include the word Material
type. Do not provide explanations or
notes.
###
<<<
Description: {text}
Material type:
>>>
```

where the *taxonomy* is a list of possible labels in the taxonomy to predict and the *text* is the input constructed from metadata fields. For the output, we use a multi-label formulation that includes all appropriate labels within the taxonomy. For example, if a sample should be labeled as BASALT, then it should also be labeled as ROCK since BASALT is a type of ROCK in the iSamples taxonomy (Figure 3). For evaluation on this multi-label dataset, we use the macro F1 score.

For our training and testing purposes, we utilized the 988,426 (22%) of records that contained the MATERIAL field value. Due to the dataset's highly imbalanced nature, we randomly selected and discarded sample records such that the maximum occurrence of a label could be 10,000, while ensuring that each label appeared at least 10 times in the entire dataset. As a result, the training dataset contained 294,420 records, the development dataset contained 63,090 records, and the test dataset contained 63,091 records. For the test dataset, we focused solely on labels that appeared at least 10 times and records that was annotated up to the leaf level to estimate performance of our experiments. The number of tokens in fields of these records is listed in Appendix A.

4 Experiments

We explore three LLMs for our multi-label text classification formulation of the metadata assignment problem: Llama-2-7B-chat-hf, Mistral-7B-Instruct, and Mistral-7B-OpenOrca. Using these models, we explore the following research questions:

1. How do different record-to-text conversions affect performance?

2. Is it useful to have an LLM summarize the record text?

3. Is top-down or bottom-up search better for incorporating the label hierarchy?

4. How much can few-shot prompting improve over zero-shot prompting?

To address each of these questions, we utilize the test dataset for evaluation purposes. Our findings indicate that the bottom-up search approach is the most effective for integrating the label hierarchy, and thus, we apply this method when conducting experiments for the first two questions. The following sections detail the experiments conducted to answer each of these research questions.

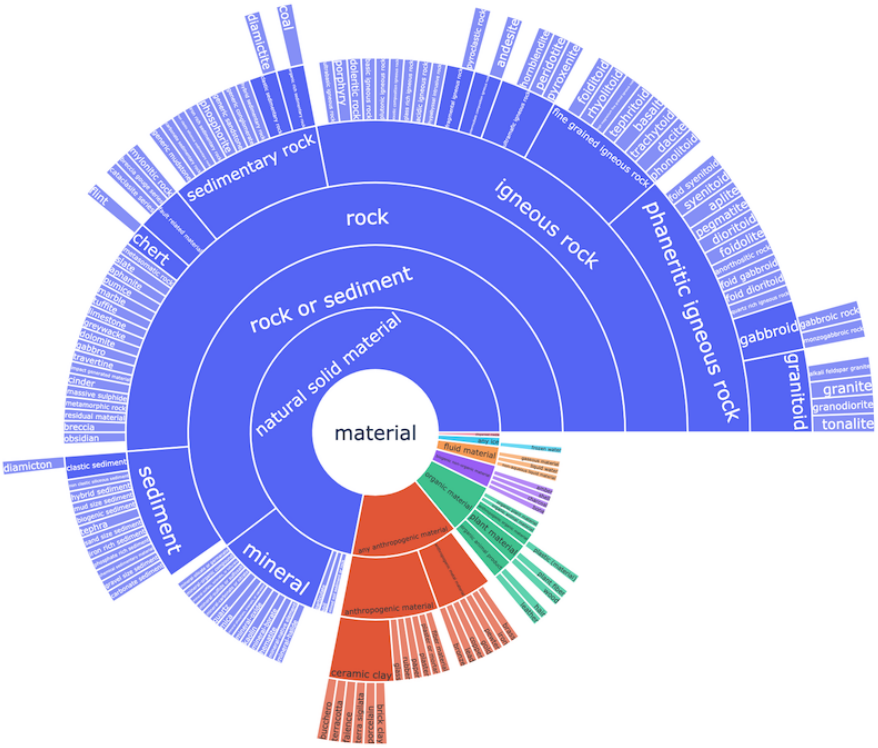


Figure 3: iSamples MATERIAL type taxonomy

Selected values	Input
Metadata field values	long cylindrical cores, Ship, Lamont-Doherty Core Repository at Columbia University (LDCR), ...
Metadata field name and value	description_sampleType is long cylindrical cores, description_supplementMetadata_platformType is Ship, description_supplementMetadata_currentArchive is Lamont-Doherty Core Repository at Columbia University (LDCR), ...
Metadata field description and value	The object type of sample is long cylindrical cores, The type of platform for the cruise is Ship, The name of institution, museum, or repository where the sample is currently stored is Lamont-Doherty Core Repository at Columbia University (LDCR), ...

Table 1: An example of different methods to generate the input for the language model.

4.1 How do different record-to-text conversions affect performance?

There are many possible ways to transform a sample record into an input for text classification. We compare the following approaches to create the input text, with examples of each shown in table 1.

Field value: We include only the values of the metadata fields.

Field name and value: For each metadata field, we include the field name, then the string " is ", then the field value.

Field description and value: The same as *Field name and value*, but instead of the field name, we include the description of the field name obtained from the SESAR collection website.

Table 2 shows the results of varying the text conversion strategy. We observe that using the metadata field description and value is the most effective across all three LLMs. This suggests that LLMs are more effective text classifiers when given a more verbose input that looks more like natural language text.

Model	Selected values	F1
Llama-7B-chat-hf	Field values	0.342
Llama-7B-chat-hf	Field name and value	0.365
Llama-7B-chat-hf	Field description and value	**0.370**
Mistral-7B-Instruct	Field values	0.377
Mistral-7B-Instruct	Field name and value	0.374
Mistral-7B-Instruct	Field description and value	**0.379**
Mistral-7B-OpenOrca	Field values	0.397
Mistral-7B-OpenOrca	Field name and value	**0.427**
Mistral-7B-OpenOrca	Field description and value	0.405

Table 2: Comparison of different methods to generate the text used as input for text classification applied to the Bottom-Up Tree Search with Zero-Shot Learning. Best performing method is in bold. Macro F1 is used for evaluation.

Model	Model Input	F1
Llama-7B-chat-hf	Original description	0.370
Llama-7B-chat-hf	Summary + Explanation	0.428
Mistral-7B-Instruct	Original description	0.379
Mistral-7B-Instruct	Summary + Explanation	0.385
Mistral-7B-OpenOrca	Original description	0.405
Mistral-7B-OpenOrca	Summary + Explanation	0.439

Table 3: Result of Zero-Shot Classification with Summarize&Explain applied to the Bottom-up Tree Search. Macro F1 is used for evaluation.

4.2 Is it useful to have an LLM summarize the record text?

We also explore an alternative approach to formulate the input to the LLM, by considering the LLM as a summarizer with noisy knowledge. In this approach, given the description and value of the metadata fields of the sample record, we instruct the LLM to generate a summary with a focus on the text portion relevant to the metadata field that is to be predicted. As the sample record may contain metadata fields with domain-specific terms, we also instruct the LLM to provide explanations for these metadata fields. The prompt that is used to generate the summary is as follows:

```
You are a scientist. User will give you a
description of a material sample it
sampled from the nature. You must generate
a summarized description of the sample.
Your task is to give a brief one sentence
summary of the given description, focusing
on the parts that are helpful in determining
the type of material that constitutes it.
###
<<<
Description: {text}
>>>
```

This summary is enriched with the explanation of metadata fields that is generated by the following prompt:

```
You are a scientist. User will give you
a term that indicates {metdata field
description}. You must generate a short
description of the term. Your task is to
give a brief one sentence description of
the given {metadata field} of the sample.
###
<<<
Term:{term}
>>>
```

This enriched summary is then used as the input to the LLM text classification prompt.

Table 3 shows the results of this experiment. We see that summarizing and explaining the sample record improves performance for all models, though it benefits Llama the most. This suggests that allowing the LLM to discard redundant or irrelevant metadata fields and provide an explanation of domain specific terms can make the text classification task easier for the LLM.

4.3 Is top-down or bottom-up search better for incorporating the label hierarchy?

We explored different approaches to utilizing the hierarchical label space:

Flat Disregard the hierarchy and treat the task as a flat multi-label classification problem. We include the entire label space in the prompt, expecting the model to return multiple labels for each physical sample record.

Top-Down Tree Search Follow the approach of Boyle et al. (2023), where predictions are made starting from the root of the hierarchy and recursively making new predictions over the children at each level of the hierarchical tree. The label space at each level is restricted to the children of the current node. The recursive process continues until the leaf level is reached.

Bottom-Up Tree Search Predictions are made starting from the leaf level, predict higher-level labels only if no valid prediction is found.

Table 4 shows the results of varying the strategy for handling hierarchical labels. We see that applying Bottom-Up Tree Search is effective across all LLMs. At the same time, Top-Down Tree Search

Model	Tree Search Strategy	F1
Majority Baseline	-	0.123
Llama-7B-chat-hf	Flat	0.154
Llama-7B-chat-hf	Top-Down Tree Search	0.114
Llama-7B-chat-hf	Bottom-Up Tree Search	0.370
Mistral-7B-Instruct	Flat	0.167
Mistral-7B-Instruct	Top-Down Tree Search	0.175
Mistral-7B-Instruct	Bottom-Up Tree Search	0.379
Mistral-7B-OpenOrca	Flat	0.241
Mistral-7B-OpenOrca	Top-Down Tree Search	0.238
Mistral-7B-OpenOrca	Bottom-Up Tree Search	0.405

Table 4: Result of Zero-Shot Classification with different Tree search strategies. Flat treats the task as a multi-label classification problem without considering label hierarchy. Macro F1 is used for evaluation.

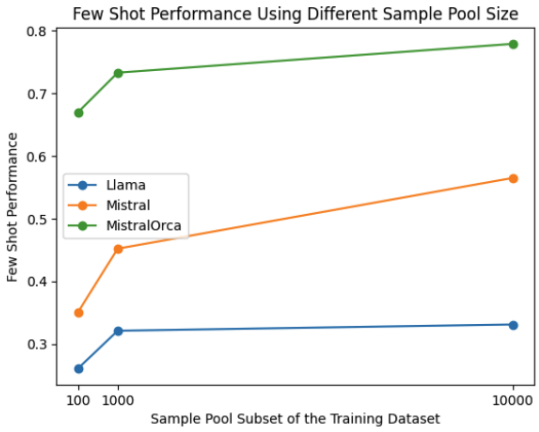


Figure 4: Sample Pool Size Impact for Few-Shot Learning with flat integration of hierarchy.

performs similarly to Flat, contradicting the previous results of Boyle et al. (2023), where using LLM-guided tree-search traversal in a top-down manner achieved state-of-the-art performance in the task of assigning diagnostic ICD codes. We suspect the failure of Top-Down Tree Search on our data is due to the terms in the top-level vocabulary, such as "any anthropogenic material" and "biogenic non-organic material", which can be difficult for the LLM to understand as they are low-frequency highly specialized terms that the LLM may not have large exposure to.

4.4 How much can few-shot prompting improve over zero-shot prompting?

We also conduct few-shot learning by including a sample of examples in the prompt to the model. We include the examples in the prompt as follows:

```
You are a scientist. Your task is
to analyze the description of a material
sample and determine the kind of material
that constitutes it after <<<>>> into one
of the predefined material types:
{taxonomy}
You will only respond with the material
type. Do not include the word Material
type. Do not provide explanations or
notes.
###
Here are some examples:
{examples}
###
<<<
Description: {text}
Material type:
>>>
```

For the selection of examples, we utilize a k-nearest neighbor (kNN) search methodology proposed by Khandelwal et al. (2019), wherein we select the k most similar examples from the training dataset for each test record. To determine the similarity between records, we represent the text of the record using a Sentence-BERT model which was trained to estimate the similarity of scientific publications (allenai-specter) (Cohan et al., 2020). The few-shot samples are ordered from least to most cosine similarity to the given test record according to the methodology proposed by Sun et al. (2023). Due to constraints on input context, Llama used 3-shot examples, while Mistral and Mistral-OpenOrca used 5-shot examples. The number of samples (k) was chosen as mentioned above, as we observed lower performance with larger k values.

Since the k examples are drawn from the training data, it is important to determine how large of a pool of training examples are needed for successful few-shot prompting. Figure 4 shows experiments with subsets of the training dataset, randomly selecting n = 100, 1000, and 10,000 records for each MATERIAL type. From the results, we see larger training dataset size enhances the performance as the probability of retrieving relevant in-context examples increases, which is consistent with prior research by Liu et al. (2021).

We compare few-shot prompting to both zero-shot prompting and the following baselines:

Most common label Predict the most common label, which is the ROCK label.

Multinomial Naive Bayes Applies a multinomial naive bayes classifier that represents the text metadata fields as TF-IDF vectors.

RoBERTA-Large-MNLI Applies a RoBERTa model that has been pre-trained on textual entailment data, is fine-tuned on the entire

Model	Model Input	Prompt	Hierarchy	F1
Most common label	-	-	Flat	0.123
Multinomial Naive Bayes	Field values	-	Flat	0.652
RoBERTA-Large-MNLI	Original description	-	Flat	0.721
kNN	Original description	-	[See text]	0.844
Llama-7B-chat-hf	Original description	Few-Shot	Flat	0.331
Llama-7B-chat-hf	Original description	Few-Shot	Bottom Up	0.504
Llama-7B-chat-hf	Summary + Explanation	Few-Shot	Bottom Up	0.501
Mistral-7B-Instruct	Original description	Few-Shot	Flat	0.565
Mistral-7B-Instruct	Original description	Few-Shot	Bottom Up	0.772
Mistral-7B-Instruct	Summary + Explanation	Few-Shot	Bottom Up	0.669
Mistral-7B-OpenOrca	Original description	Few-Shot	Flat	0.779
Mistral-7B-OpenOrca	Original description	Few-Shot	Bottom Up	0.880
Mistral-7B-OpenOrca	Summary + Explanation	Few-Shot	Bottom Up	**0.928**
Mistral-7B-OpenOrca (best zero-shot)	Summary + Explanation	Zero-Shot	Bottom Up	0.439

Table 5: Result of Few Shot Learning with Bottom-Up Tree Search and Summarize&Explain Approach with the entire training dataset used as sample pool. Best performing few-shot approach is in bold. Macro F1 is used for evaluation.

training data, and is used to classify text following Yin et al. (2019); Pàmies et al. (2023). Early stopping was used to prevent overfitting. Other hyperparameters used for fine-tuning are listed in Appendix B.

kNN Utilizing the same Sentence-BERT model employed in few-shot learning, we create embeddings representing each sample record, and label new sample records with the most common label from their 5-nearest neighbors. The most common label is determined by prioritizing the label that holds the deepest position within the taxonomy hierarchy and is most frequently observed among the neighboring labels.

Table 5 compares few-shot prompting to the best zero-shot prompting model and these baselines. The best few-shot setting is the same as the best zero-shot setting: Mistral-7B-OpenOrca, summarize and explain the input, and use the bottom up search strategy. But the best few-shot setting substantially outperforms the best zero-shot setting, 0.928 F1 to 0.439 F1. It also outperforms all the baselines, including those that are fine-tuned over the entire training data. The closest baseline is the k-nearest neighbors, achieving 0.844 F1, which indicates that selecting relevant training examples for few-shot prompting is driving a large part of the performance, but the knowledge embedded in the LLM provides extra power beyond those few-shot examples.

One difference that can be observed between the zero-shot and few-shot models is that the Summarize and Explain approach did not help Llama and hurt Mistral, while that approach benefited both models in the zero-shot setting (see table 3). This suggests that almost any summary helps without any training examples in the prompt, but as soon as the prompt contains relevant training examples, only the best summaries are helpful. A deeper analysis of the quality of summaries coming out of each model would help to confirm or deny this hypothesis.

5 Analysis

Given that sample records that were registered in the same institution often share similar characteristics, we investigate whether this introduces bias into our few-shot learning. Out of the 102 unique institutions in the test dataset, we filtered out the records of 50 institutions from the training dataset. We then used this dataset to create three models:

RoBERTA-Large-MNLI Applies the same approach as RoBERTA-Large-MNLI above, but fine-tuned on only the filtered training data.

kNN Applies the same approach as kNN above,

Model	Unseen	Seen	Both
RoBERTA-Large-MNLI	0.477	0.513	0.497
kNN	0.168	0.861	0.331
Mistral-7B-OpenOrca-Best	0.426	0.915	0.555
Mistral-7B-OpenOrca-Robust	0.491	0.878	0.620

Table 6: Different few shot approaches. Unseen stands for test records that come from institutions that have not been seen during training, and seen stands for test records that come from institutions that have been seen during training. Macro F1 is used for evaluation.

but neighbors are drawn only from the filtered training data.

Mistral-7B-OpenOrca-Best Applies the same approach as the best performing Mistral-7B-OpenOrca model but with few shot examples drawn only from the filtered training data.

Table 6 shows that all three models see a drop in performance when evaluated on institutions not present in the training data. The k-nearest neighbor model suffers the most severe drop, but the Mistral-7B-OpenOrca-Best also suffers a large drop, yielding 0.426 F1 on sample records from unseen institutions, lower than the 0.477 F1 of the RoBERTA model. These results suggest that few-shot learning will be substantially more effective when there exist sample records from the same institution in the training data that is used for example retrieval.

In an attempt to mitigate some of this bias, we conducted experiments by reducing the number of few-shot samples to k=3 and placing greater emphasis on the analysis of the given test record's description by adding the following to the prompt:

```
Answer what is the material type of the
sample by going through each sentence of
the given description and analyzing.
Choose from {taxonomy} the material type
that constitutes the given description.
```

This model, Mistral-7B-OpenOrca-Robust, was able to reduce the existing bias as we can see improvement in test records that originate from unseen institutions. We plan to further investigate to find a way to improve the robustness most effectively in future research.

6 Conclusion

Physical sample records, a form of scholarly document created by natural scientists, play a crucial role in ensuring the reproducibility and reusability of sample-based scientific knowledge; additionally, they are often used as scientific datasets in and of themselves. However, they are often incomplete, thereby impacting their usability and the usability of the data repositories that store them. In this research, we explore various methodologies aimed at leveraging recent developments in LLMs to address this issue.

To transform a sample record into an input suitable for text classification by LLMs, we find that maximizing verbosity by using descriptions of metadata field names and their corresponding values is beneficial. Additionally, we demonstrate that harnessing the LLM's inherent knowledge and summarization capabilities enhances the comprehensibility of inputs for text classification tasks.

Our results also reveal the effectiveness of integrating taxonomy hierarchies through bottom-up tree search, which we expect to be particularly beneficial for domain-specific taxonomies with which the LLM has limited exposure. Furthermore, incorporating few-shot examples into the prompt leads to substantial improvements. Notably, the best few-shot learning performance surpasses that of a fine-tuned RoBERTA textual entailment classifier, indicating a promising approach for embedding training data with minimal computational resources, applicable across diverse domains of sample records.

Overall, our experimental results suggest that LLMs hold potential as a solution for digital libraries and aggregations grappling with metadata quality issues, particularly in domain-specific cases. Our approach offers promise for addressing metadata quality challenges stemming from incomplete records across various domains.

7 Limitations

In this paper, our exploration was confined to a single data repository and taxonomy. To assess the applicability of the identified methods across different contexts and validate their effectiveness, further experimentation involving diverse data repositories is necessary.

Additionally, our study utilized a non-domain-specific LLM with a modest size of 7 billion parameters. Considering recent advancements in LLMs, leveraging larger-scale models or domain-specific models such as Galactica (Taylor et al., 2022), or using fine-tuned LLMs could lead to further improvements in performance.

8 Acknowledgements

This work was funded by NSF grant #2004562.

References

2020. *Biological Collections: Ensuring Critical Research and Education for the 21st Century*. National Academies Press, Washington, D.C.

Joseph S Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q O'Neil. 2023. Automated clinical coding using off-the-shelf large language models. *arXiv preprint arXiv:2310.06552*.

Arthur Chapman and John Wieczorek. 2020. Georeferencing Best Practices. Publisher: [object Object].

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Neil Davies, John Deck, Eric C Kansa, Sarah Whitcher Kansa, John Kunze, Christopher Meyer, Thomas Orrell, Sarah Ramdeen, Rebecca Snyder, Dave Vieglais, et al. 2021. Internet of samples (isamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, 10(5):giab028.

Naomi Dushay and Diane I Hillmann. 2003. Analyzing metadata for effective use and re-use. In *International Conference On Dublin Core And Metadata Applications*, pages pp–161.

Hui Han, C Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward A Fox. 2003. Automatic document metadata extraction using support vector machines. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, pages 37–48. IEEE.

J. Mason Heberling, Joseph T. Miller, Daniel Noesgaard, Scott B. Weingart, and Dmitry Schigel. 2021. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6). ISBN: 9782018093113 Publisher: National Academy of Sciences Section: Biological Sciences.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1051–1060.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Gil Nelson and Shari Ellis. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763):20170391. Publisher: Royal Society.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor González-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296.

Gordon W Paynter. 2005. Developing practical automatic metadata assignment and evaluation tools for internet resources. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 291–300.

Stephen M Richard, David Vieglais, Andrea K Thomer, Hyunju Song, Neil Davies, John Deck, Quan Gan, Eric C Kansa, Sarah Kansa, John Kunze, Kerstin Lehnert, Danny Mandel, Christopher Meyer, Rebecca Snyder, Ramona L Walls, Yuxoan Zhou, and Hong Cui. 2024. A metadata schema for documenting material samples from multiple domains | www.semantic-web-journal.net.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Emma Tonkin and Henk L Muller. 2008. Semi automated metadata extraction for preprints archives. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 157–166.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Linli Xu, Sijie Teng, Ruoyu Zhao, Junliang Guo, Chi Xiao, Deqiang Jiang, and Bo Ren. 2021. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2459–2468.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

A Average number of tokens in the dataset

Metadata field	Avg. # Token
supplementMetadata_province	1.26
sampleType	2.32
supplementMetadata_city	1.62
supplementMetadata_primaryLocationType	1.52
collectionMethodDescr	2.89
supplementMetadata_purpose	4.52
supplementMetadata_country	1.77
supplementMetadata_geologicalAge	1.49
supplementMetadata_geoUnit	2.82
supplementMetadata_primaryLocationType	1.52
supplementMetadata_locality	2.9
supplementMetadata_localityDescription	4.98
description	10.58
supplementMetadata_locationDescription	64.94
supplementMetadata_platformType	1.19
supplementMetadata_platformDescr	2.56
collectionMethodDescr	2.89
supplementMetadata_sampleComment	11.56
supplementMetadata_county	0.793
supplementMetadata_classificationComment	1.25
supplementMetadata_originalArchive	10.13
supplementMetadata_currentArchive	8.42
supplementMetadata_sampleComment	11.56
supplementMetadata_fieldName	2.55
supplementMetadata_cruiseFieldPrgrm	3.23
supplementMetadata_publicationUrl_description	8.38

Table 7: Average number of tokens of metadata fields in the dataset.

B Hyperparameters for finetuning

Hyperparameter	Value
Batch size	16, 32, 64
Learning rate	1e-5, 2e-5, 5e-05
Weight decay	0.01
Epochs	3

Table 8: Hyperparameters used for fine-tuning a RoBERTA-Large-MNLI. Grid search was used for choosing the optimal hyperparameter values.