Enhancing Open Data Findability: Fine-Tuning LLMs(T5) for Metadata Generation.

*Umair*Ahmed[a*], *Andrea*Polini[a]

[a]Dipartimento di Informatica, Universita di Camerino, Camerino, Italy, Emails: umair.ahmed@unicam.it, an-drea.polini@unicam.it, ORCIDs: 0000-0003-2260-2777, 0000-0002-2840-7561.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025:

Abstract. Metadata is essential for improving the discoverability and findability of datasets. According to the European Data Portal (EDP), Europe's largest open data portal, keywords and categories alone contribute to 60 percent of a dataset's visibility. Given that datasets can have highly variable and complex contexts, we propose leveraging the power of Large Language Models (LLMs), specifically T5-small and T5-large, to generate these metadata. In our study, we used EDP as our case study. We obtained 60,000 datasets from EDP and undertook thorough data cleaning and transformation. This process yielded 3,131 datasets for keyword extraction and 2,790 datasets for category extraction, ready for model fine-tuning.

The base versions of T5-small and T5-large initially struggled to produce keywords that were representative of those generated by humans, resulting in F1 scores of only 0.0455 and 0.1051, respectively. However, fine-tuning significantly improved their performance, achieving an F1 score of 0.4538 for T5-small and 0.6085 for T5-large. A similar pattern was observed in category extraction. The base T5-small and T5-large models had F1 scores of only 0.1222 and 0.3326, respectively. In contrast, the fine-tuned models produced F1 scores of 0.6284 and 0.8322, respectively. Notably, T5-large produced keywords similar to those generated by humans in over 60% of cases, and its category predictions matched human-generated ones in over 80% of cases. This highlights the potential of using large language models (LLMs) for generating human-like metadata, thereby significantly enhancing data findability and usability across various applications.

Keywords. FAIR Metadata, European Data Portal, Open Data, Large Language Models, LLM, T5, Artificial Intelligence, Annotation, Tagging, Keyword Extraction
Research paper, DOI: https://doi.org/10.59490/dgo.2025.941

1. Introduction

In the current data-driven world, there has been a significant increase in the volume and circulation of data. It has alerted stakeholders of its potential to be the new oil. Europe's data economy was estimated to be around €739 billion in 2020 (Van Loenen et al., 2021). Making this data open could lead to a more transparent and inclusive ecosystem. While its potential has already turned heads and a significant amount of data is open, its findability and discoverability across different portals still remain a challenge.

In this study we focus on the lack of quality metadata, that is one of the major causes of poor findability. Indeed, metadata definition is often overlooked by publishers when uploading a dataset. They are usually either unaware of the importance of metadata or more focused on the task of making the data open while ignoring the potential impact of having good metadata. It necessitates an alternate mechanism to fill in the metadata or recommend it to publishers, forming this study's basis. As the European Data Portal (EDP) (EU,

2025) is one of the largest open data portals, we use it as the case study for this research. There are several metadata attributes that enhance a dataset's visibility, including title, description, categories/themes, and keywords. However, according to the EDP, as shown in Table 1, keywords and categories/themes account for 60% of the overall findability score for datasets (EU, 2025). Apart from EDP, keywords and categories are considered to be the most essential elements in general findability. It motivates us to focus on the automated extraction of keywords and categories. By generating good-quality and relevant metadata, we can significantly improve the discoverability of these datasets.

Tab. 1 – European Data Portal Findability Score (EU, 2025)

Indicator	**Description**	**Metrics**	**Weight**
Keyword usage	Keywords directly support the search and thus increase the findability of the data dataset.	The system checks whether keywords are defined. The number of keywords has no impact on the score.	30
Categories	Categories help users to explore datasets thematically.	It is checked whether one or more categories are assigned to the dataset. The number of assigned categories has no impact on the score.	30
Geo search	Usage of spatial information would enable users in order to find the dataset with a geo-facetted search.	It is checked whether the property is set or not.	20
Time-based search	Usage of temporal information would enable users for a timely based facetted search.	It is checked whether the property is set or not.	20

Numerous AI methodologies exist to perform classification and concept extraction. However, this study focuses on leveraging Large Language Models (LLMs) to address the problem of missing/complete metadata, given their deep understanding of the context of texts. It mainly investigates the performance of T5-small and T5-large in generating keywords and categories for datasets within the EDP.

For this study, we extracted 60,000 datasets from EDP, each comprising metadata fields such as titles, descriptions, keywords, and categories. We filtered the data after a rigorous cleaning and transformation process to ensure consistency and relevance. The final dataset included 3,131 instances for keyword extraction and 2,790 instances for category extraction. We fine-tuned both T5 models on these datasets to generate high-quality, human-like metadata. For keywords, we evaluated model-generated outputs using ROUGE-1 and cosine similarity with a threshold of 0.5 to measure the semantic overlap with human-generated keywords. For categories, we performed a one-to-one matching between model-predicted and actual categories. We calculated precision, recall, and F1 scores to assess overall accuracy for both tasks.

The results of this study reflect the effectiveness of fine-tuning large language models (LLMs) for metadata generation. The fine-tuned Flan-T5-large model consistently outperformed T5-small and the base models in both keyword and category extraction tasks, achieving F1 scores of 0.6085 and 0.8322, respectively. These results highlight the significant potential of LLMs to improve metadata quality, address critical gaps in open data platforms, and enhance dataset findability and discoverability. While significant research is underway in the field of semantic search mechanisms due to the advent of large language models (LLMs), our proposal aims explicitly to enhance the searchability and discoverability of datasets that already exist in open data portals. The search mechanisms mostly rely on basic one-to-one matches.

In the subsequent sections, we present a detailed discussion of our study's related work, methodology, results, and implications of our study.

2. Related Works

The significance of metadata, particularly keywords and categories (tagging), in improving a resource's findability, discoverability, and accessibility is well documented. Although several studies have focused on meta-

data generation, minimal research specifically addresses AI's use for generating metadata of open datasets. Following studies have explored the problem of findability and metadata extraction in various domains and from different perspectives. Most of the explored studies employed traditional machine learning models for keyword extraction, while more recent studies also employed LLMs. One such study, RAKE, performs keyword extraction by identifying co-occurrence relationships between words in a document. Even if it is quite fast, it struggles with more semantic relations between texts (Rose et al., 2010). A more advanced method like YAKE!, an improvement on RAKE, employs additional local statistical features for unsupervised keyword extraction. It considers multiple features, including word frequency, co-occurrence, and position, to extract keywords more efficiently (Campos et al., 2020). Moreover, this particular method improves upon frequency-based methods by assigning a higher weight to words occurring often (Würsch et al., 2023). This study explores both machine learning (decision tree and kea) and deep learning (ANN and LSTM) models (Deng, 2024). Despite their effectiveness, traditional methods remain limited in extracting high-quality metadata, especially in cases where semantic understanding is essential.

While traditional methods have been quite robust, contemporary LLMs have paved the way for more semantically aware keyword extraction. For instance, LLM-TAKE utilizes theme-aware keyword extraction, leveraging LLMs to generate keywords that are contextually aware and capture contextual relations (Maragheh et al., 2023). Similarly, Lee et al. (2023) explored and employed a trained LLM, Galactica by Meta, to generate keywords for scientific literature. It showed significant improvements over traditional methods like RAKE and YAKE! (Lee et al., 2023). Purwar and Sundar (2023) use KeyBERT for keyword generation and GPT 3.5 for question answering about the context of the text (Purwar & Sundar, 2023). Ahmed et al. employed a hybrid of BERT, RAKE, TAKE, and TextRank (BRYT) to generate semantically aware keywords from dataset descriptions, which considered the best keywords from each model and reranked them to give the most optimal output (Ahmed et al., 2024)(Ahmed, 2023).

Beyond keyword extraction, category/theme tagging has also benefited from LLMs' ability to understand context, be semantically aware, and generalize across domains. Models like T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020) have been fine-tuned for domain-specific metadata generation, including multi-label classification tasks for thematic tagging. For instance, li et al. explored hierarchical transformer models for multi-label text classification. It demonstrated their capability to understand semantic knowledge at different levels to classify the text into given categories (Li et al., 2022). Recent developments also emphasize the integration of multimodal data with LLMs for combined keyword extraction and annotation. Kathiriya et al. explored multimodal LLMs to enhance metadata tagging across datasets that include diverse information (Kathiriya et al., n.d.). Dai et al. also explored the potential of LLMs, particularly GPT-3.5, to replicate human-like behavior in various tasks (Dai et al., 2023). The integration of LLMs with multimodal data, as explored by Kathiriya et al. (2023), further demonstrates the potential of AI-driven approaches to enhance the discoverability of information across diverse platforms, such as e-commerce and data portals (Kathiriya et al., n.d.). Moreover, prompt-tuning and zero-shot capabilities of LLMs, such as OpenAI's GPT models, have made it possible to extract both keywords and thematic categories as multimodal efficiently. By combining supervised fine-tuning and techniques like few-shot learning, these models can address the dual challenge of keyword extraction and thematic tagging at the same time, making them invaluable for dataset-level metadata generation. However, they are significantly more expensive and might hallucinate answers (Ouyang et al., 2022).

3. Methodology

In today's information-driven world, open datasets can potentially solve real-world problems, whether they concern the economy, climate, governance, or health. However, for this to be realized, those datasets have to be findable, accessible, and usable. This notion shaped the basis of our research question: "How can we make vast amounts of data findable and discoverable?". Unfortunately, this vast amount of data, more often than not, hide in plain sight due to poor metadata, such as missing/misrepresentative keywords or missing/misclassified categories. Our goal in this study was to select a semantically capable AI agent, train it to thoroughly understand dataset descriptions, and generate accurate, representative metadata, much like an expert human curator might do.

To achieve this, we curated a methodology (reflected in Fig. 1 and Fig. 2) that involved gathering data, cleaning and refining it, selecting the AI model, designing clear instructions for AI (prompts), fine-tuning the model to

mimic an expert human curator, and evaluating it using the collected data. In the following sections, we provide insights into each step of the methodology.

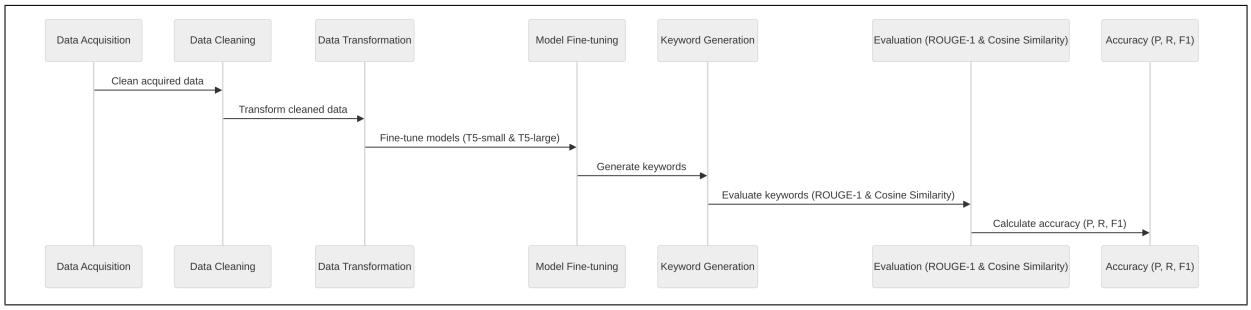

Fig. 1 – Diagram illustrating the key components of the keyword extraction

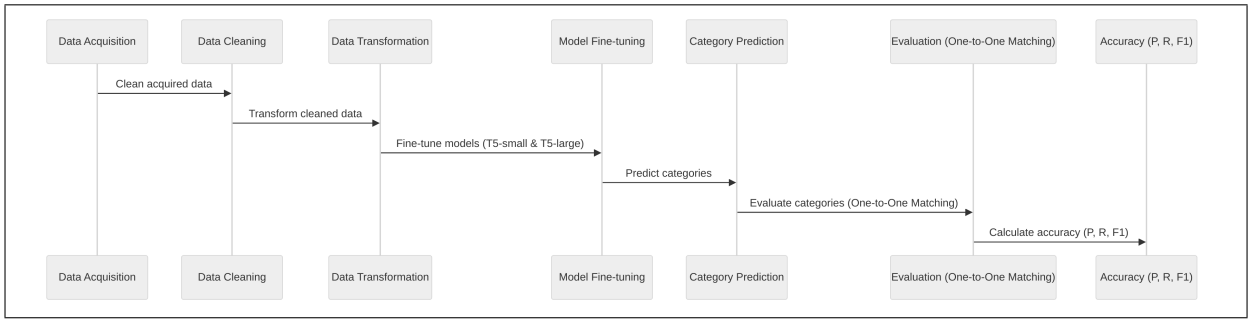

Fig. 2 – Diagram illustrating the key components of the category extraction

3.1. Data Collection and Preprocessing

The European Data Portal is one of the most significant open data initiatives, aggregating open data from EU states. We selected it as our use case and extracted 60000 datasets from it. These datasets were related to various sectors, such as environment, economy, health, and governance. Although there were significant amounts of extracted datasets, their quality was variable, which prompted the need for extensive data cleaning.

We aimed to curate two clean datasets for each task, necessitating task-specific data cleaning and transformation to ensure appropriate model fine-tuning.

3.1.1. Data Exploration and Initial Observations

Prior to data cleaning, we performed an exploratory analysis to understand it. The following observations were recorded and shaped our strategy towards data cleaning:

- **Inconsistent Metadata:** A significant amount of datasets had minimal or missing keywords and categories. Moreover, many of the keywords and categories present were not representative.
  - **keywords:** Many instances had repetitive or nonrepresentative long keywords. For instance, one of the datasets had 41 keywords and one of the keywords was 'From 25 to 34 years (current quarter)' which definitely hurts the chances of a dataset being findable
  - **Categories/Themes:** Several datasets had either missing categories or too many categories assigned to them. For instance, one of the datasets had the following six categories assigned to it: 'Economy and finance | Education, culture and sport | Government and public sector | Health | Population and society | Regions and cities'.
- **Noise in Descriptions:** Some descriptions contained garbage characters, HTML tags, just names of columns, or were too short to be considered for the AI model to extract metadata from. For instance, the following dataset description was too small to consider: 'Traffic signals in Copenhagen Municipality'

- **Multilingual Records:** Various descriptions were in different languages and had to be cleaned to keep the dataset consistent. For instance, the following description: 'Navly, ligne de navette autonome, électrique et sans conducteur au sud du quartier de la Confluence, le long de la Saône.'

These highlighted observations necessitated a more detailed cleaning of the data for efficient training and evaluation of the model.

3.1.2. Preparing the Keyword Extraction Dataset

In the keyword extraction task, we aimed to extract keywords that were relevant and concise. With that in mind, we performed the following data cleaning for it:

- **Language Filtering:** We filtered out non-English descriptions to maintain consistency in training the language models, as different languages in the prompt could confuse the model while training. For instance, a description in French could make it extract multilingual keywords.
- **Title Deduplication:** Using fuzzy matching, we eliminated datasets with titles that had more than 80% similarity to avoid redundancy. This ensured that the model was fine-tuned on a diverse range of examples.
- **HTML Tag Removal:** Some descriptions contained HTML tags, which we eliminated for better text preprocessing. It was essential because non-textual elements in the description could disrupt the model's capability.
- **Number of keywords:** We selected descriptions with 3 to 12 keywords (to cover variability), as the recommended range for optimal keyword searchability is between 5 and 10 (Pottier et al., 2023).
- **Keyword Length:** We excluded keywords with more than three words to eliminate overly complex or irrelevant terms.

After this cleaning and transformation process, we were left with **3,131 records**. These were split into training (80%), evaluation (10%), and testing (10%) sets.

3.1.3. Preparing the Category Extraction Dataset

This task required further data cleaning and transformation to make the categories consistent and meaningful for the model to learn from. These categories are essential for navigating through datasets in any data portal which made us prioritize uniformity and clarity in associations.

- **Redundant Categories:** Some datasets had the same categories with different links or texts assigned to them. We merged those categories into one. For instance: 'http://publications.europa.eu/resource/authority/data-theme/GOVE, https://data.gov.ie/Government'
- **Standard Category URL:** We removed the categories not pointing to the standard Category URL or corresponding to category list names. For instance: 'ncb3a7ce63a654bed8f47f82f30cedd96b7, ncb3a7ce63a654bed8f47f82f30cedd96b9'
- **URL to Names:** We converted URLs of categories into their category names to make the category extraction more meaningful for the user. For instance: 'http://publications.europa.eu/resource/authority/data-theme/ECON' was converted to 'Economy and finance'

This process yielded in a dataset of **2,790 records** for category extraction, similarly split into training (80%), evaluation (10%), and testing (10%) sets.

3.2. Model Selection: Why we chose T5

Choosing a model was a critical task, as it would affect all aspects of the study, be it accuracy, time, space, or finances. We chose T5 (Text-to-Text Transfer Transformer), designed by Google, because of its capability to understand semantic relations in the text while being a small model trainable with a modest amount of resources (Raffel et al., 2020). It puts the model in a unique position where it is not so large that it consumes

significant resources like GPT or LLaMA. Yet, it is also not too small to be unable to understand semantic relationships in the text. Moreover, it was designed to handle a wide array of tasks using a text-to-text approach. In our study, both inputs (descriptions) and outputs (keywords/categories) are in a textual format, making it ideal for these tasks given our compute resources. We employed two variants of T5, namely T5-small and T5-large. Following is a brief overview of the models' architecture.

- **T5-small:** It has 60 million parameters, with 6 encoder/decoder layers, a hidden size of 512, and 8 attention heads. This model is lightweight, resource-efficient, and suitable for low-resource environments, but it may underperform on complex tasks (Raffel et al., 2020).
- **T5-large:** It has 770 million parameters, with 24 encoder and decoder layers, a hidden size of 1024, and 16 attention heads. This model captures intricate patterns and generalizes well for complex tasks; however, it requires significant computational resources (Raffel et al., 2020).

Given various model sizes and limited resources, we wanted to explore the possibility of performing this task with the smallest model to consume minimal resources. This is indeed also needed to permit fast suggestions to the operator uploading the dataset. While T5-small provided faster training and inference, T5-large demonstrated superior performance, making it more suitable for capturing nuanced relationships in complex texts.

3.3. Fine-Tuning T5: Unlocking Task-Specific Precision

Similar to how students and researchers begin by exploring a wide array of subjects before narrowing their focus to specialize in one specific area, fine-tuning helps a language model become highly skilled at a specific task (Ohm, 2024). T5 also had been trained on a vast corpus of data, acquiring general language capabilities. However, the base versions of T5, without fine-tuning, performed inadequately on specific tasks. In our case, to generate human-like metadata, task-specific training was necessary. This led to our decision to fine-tune the model for improved performance.

We used our cleaned-out datasets to fine-tune T5 for our respective tasks. While fine-tuning, we carefully monitored the performance of models to avoid overfitting — a condition where the model starts memorizing data instead of understanding the underlying patterns, much like a student who memorizes the answers to a number of questions instead of understanding the concepts. We let it train for n epochs (iterations) until it showed any signs of overfitting, and then we stopped it. Early stoppage allowed us to determine the optimal number of epochs (training cycles). Following are the details for each task-specific fine-tuning:

3.3.1. Keywords

Flan-T5-small: As reflected in Table 2, we stopped fine-tuning at 17 epochs as the model showed signs of overfitting beyond this point. A learning rate of 4e-5 allowed the model to converge efficiently. Moreover, given the hardware resources and model, we kept the batch size at 8.

Flan-T5-large: This was the larger model, which was trained for 12 epochs before it also started overfitting. The slightly lower learning rate of 3e-5 ensured optimal convergence. It took more time due to its larger size compared to T5-small.

Model	Learning Rate	Batch Size (Train)	Batch Size (Eval)	Epochs	Weight Decay
Flan-T5-small	4e-5	8	8	17	0.02
Flan-T5-large	3e-5	8	8	12	0.02

Tab. 2 – Keyword Extraction: Finetuning configurations for T5-small and Flan-T5-large models

3.3.2. Categories

Flan-T5-small: As reflected in Table 3, we stopped fine-tuning at 8 epochs as the model showed signs of overfitting beyond this point. A learning rate of 4e-5 helped the model to converge efficiently. It took less time than the keyword extraction task because it ran for fewer epochs and had less complex task

Flan-T5-large: The larger model also stopped at 7 epochs, reducing from 12 epochs in keyword extraction because it started to reflect overfitting after that. Although it took significantly more time than T5-small to finetune but it was still less than T5-large for keyword extraction. It is reflected in Table 3.

Model	Learning Rate	Batch Size (Train)	Batch Size (Eval)	Epochs	Weight Decay
Flan-T5-small	4e-5	8	8	8	0.02
Flan-T5-large	3e-5	8	8	7	0.02

Tab. 3 – Catergory Extraction: Training configurations for T5-small and Flan-T5-large models

To perform the fine-tuning, we converted our datasets into T5-specific input prompts, as explained in the subsequent section.

3.3.3. Prompt Design

The efficiency of any LLM depends on well-curated prompts designed for a specific task (Alharbi et al., 2023; Maratsi et al., 2024) and a specific model. Conventionally, you provide an input textual prompt and labels as answers to those prompts. Following are prompts we crafted for each task:

Keyword Extraction

- **Input:** ["Generate 3 to 12 keywords for the following dataset description to improve its findability in a portal. Description: " + datasetDescription]
- **Labels:** Original Dataset Keywords

Category Extraction

- **Input:** ["From the following themes: 'Transport', 'Health', 'Government and public sector', 'Regions and cities', 'Environment', 'International issues', 'Justice, legal system and public safety', 'Energy', 'Education', 'culture and sport', 'Economy and finance', 'Population and society', 'Science and technology', predict the most relevant themes for this dataset description. Description: " + datasetDescription]
- **Labels:** Original Dataset Categories

These prompts were used to finetune T5 small and T5 large for both problems: keyword extraction and category extraction, respectively. The prompts were more concrete and simplistic yet gave enough context for T5 to be able to generate the the required output.

3.4. Model Evaluation

3.4.1. Keywords

We compared the generated keywords from the models with the original keywords using the following similarity metrics:

- **ROUGE-1 (Grusky, 2023):** ROUGE-1 assesses the similarity between the generated keywords and reference keywords by measuring the overlap of unigrams, which are single words.

$$ROUGE-1 = \frac{Numberofoverlappingunigrams}{Totalreferenceunigrams}$$

This metric offers a clear method to evaluate how much of the original content is represented by the generated keywords. We chose ROUGE-1 because it effectively measures direct word matches and is commonly used in tasks like text summarization and text generation.

- **Cosine Similarity (Gunawan et al., 2018):** It calculates the cosine of the angle between two vectors representing the generated and original keywords.

$$CosineSimilarity = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity is useful for estimating semantic similarity between words/sentences. This metric is commonly applied in information retrieval and semantic similarity tasks.

We calculated how many original (human-like) keywords were accurately predicted by the model. If 50% or more of the original keywords matched using either of the two metrics, we conventionally considered it a match for that row. We considered using both metrics to capture two-fold matching between keywords: unigram and semantic.

3.4.2. Categories

We compared the generated categories from the models with the original categories using one-to-one matching. The categories were predefined, and the predicted category was only one of those 13 categories, which made the matching simpler.

We calculated the models' accuracy after matching the predicted and original labels in both tasks. We used precision, recall, and F1 scores to estimate the model's accuracy, as reflected below:

- **Precision:** Precision measures the proportion of correctly predicted keywords/categories (matches) out of all the keywords/categories predicted by the model (Kumar & Gupta, 2015). It reflects how many generated keywords mirror keywords a human curator might choose. It is calculated as:

$$Precision = \frac{Number of Matches}{Total Predicted Keywords/Categories} (Kumar \& Gupta, 2015)$$

- **Recall:** Recall measures the proportion of correctly predicted keywords/categories (matches) out of all the original keywords/categories in the dataset (Kumar & Gupta, 2015). It is calculated as:

$$Recall = \frac{Number of Matches}{Total Original Keywords/Categories} (Kumar \& Gupta, 2015)$$

- **F1 Score:** The F1 score is the harmonic mean of precision and recall, balancing their contributions (Kumar & Gupta, 2015). It is calculated as:

$$F1Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} (Kumar \& Gupta, 2015)$$

These metrics were calculated for all test datasets, comprehensively evaluating the model's ability to predict relevant and accurate metadata.

3.5. Machine Resources and Fine-Tuning Runtime

The fine-tuning process for both keyword and category extraction tasks was conducted on a machine with the following specifications:

- **CPU:** 8 Cores
- **RAM:** 64 GB
- **GPU:** NVIDIA A100

The fine-tuning runtimes for each model and task are summarized in Table 4.

These runtimes highlight the computational efficiency differences between Flan-T5-large and Flan-T5-small, with the latter requiring significantly less time due to its smaller size and fewer parameters.

Task	Model	Fine-Tuning Time (hh:mm:ss)
Keyword Extraction	Flan-T5-large	1:07:43
Keyword Extraction	Flan-T5-small	00:08:31
Category Extraction	Flan-T5-large	00:31:58
Category Extraction	Flan-T5-small	00:04:01

Tab. 4 – Fine-tuning runtimes for each model and task.

Results Overview

In this section, we aimed to evaluate how effectively fine-tuned LLMs can generate metadata—particularly keywords and categories—to improve the findability of the datasets. We focused on two models, namely T5-small and T5-large, and compared their performance before and after fine-tuning. This comparison reflected an apparent transformation. Fine-tuning the base models enabled them to shift from generating non-representative and unreliable metadata to generating effective metadata that closely resembled what a human curator might create. This section views these observations from the lens of metrics like precision, recall, F1 scores, and match rates.

Keyword Extraction

Keywords define the essence of any resource and allow users to understand the summary of the resource without navigating it. In this study, we employed T5-small and T5-large to generate the keywords for EDP datasets. Initially, T5-small and T5-large struggled with extracting keywords. Their base (non-finetuned) models generated keywords that were either inconsistent or irrelevant when compared to human-curated ones. They fell short in both matching algorithms, Rouge-1 and cosine similarity, with a minimum threshold of 50 percent. As reflected in (Table 5) and Fig 3, T5-small only generated keywords for 2 datasets that matched at least 50 percent with the human-labeled metadata, while T5-large only managed just three datasets at a similar threshold.

Match Percentage	T5-small (Base)	T5-small (Finetuned)	T5-large (Base)	T5-large (Finetuned)
$\geq 10\%$	36	220	82	271
$\geq 20\%$	15	206	51	258
$\geq 30\%$	9	189	25	245
$\geq 40\%$	7	148	14	220
$\geq 50\%$	2	139	3	207
$\geq 60\%$	2	126	1	195
$\geq 70\%$	0	89	0	133
$\geq 80\%$	0	62	0	115
$\geq 90\%$	0	45	0	89
$= 100\%$	0	45	0	88

Tab. 5 – Keyword Extraction: Match Percentages Across Models.

As demonstrated in (Table 6) and Fig 4, their performance was significantly weak across all key metrics. Precision, which reflects how many generated keywords were actually correct, was just 0.1093 for T5-small and 0.1919 for T5-large. Recall, which represents how well the model captured the full range of relevant keywords, was similarly low at 0.0287 for t5-small and 0.0724 for t5-large. Consequently, F1 scores, which balance precision and recall, were abysmal: 0.0455 for T5-small and 0.1051 for T5-large. These scores reflected the models' inability to understand the semantic nuances in dataset descriptions.

However, fine-tuning transformed their performance drastically, given it enhanced the model's understanding of the particular task of keyword extraction. Once fine-tuned, the T5-small model improved its match rate to 139 rows at the 50% threshold, while T5-large matched in 207 rows. It is reflected in (Table 5) and Fig 3.

As observed in (Table 6) and Fig 4, the precision of the fine-tuned T5-large model improved to 0.6423, and its recall increased to 0.5781, resulting in an F1 score of 0.6085. This meant that the model generated keywords relevant to the dataset descriptions over 60% of the time—comparable to human experts. This improvement

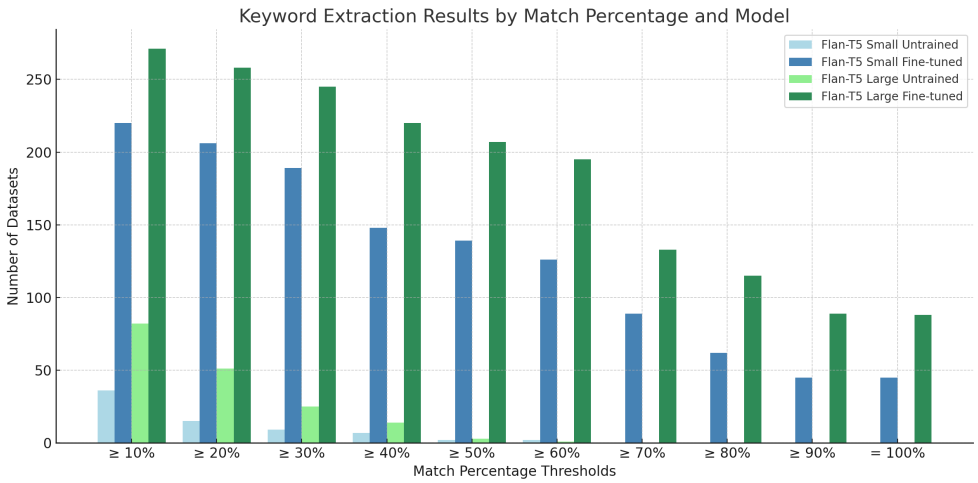

Fig. 3 – Match Percentages Across Models for Keyword Extraction.

Model	Precision	Recall	F1 Score
T5-large (Finetuned)	0.6423	0.5781	0.6085
T5-large (base)	0.1919	0.0724	0.1051
T5-small (Finetuned)	0.5227	0.4010	0.4538
T5-small (base)	0.1093	0.0287	0.0455

Tab. 6 – Keyword Extraction: Precision, Recall, and F1 Scores.

highlights how fine-tuned language models can significantly enhance metadata quality, ensuring that users can find the information they need even through basic keyword-driven searches.

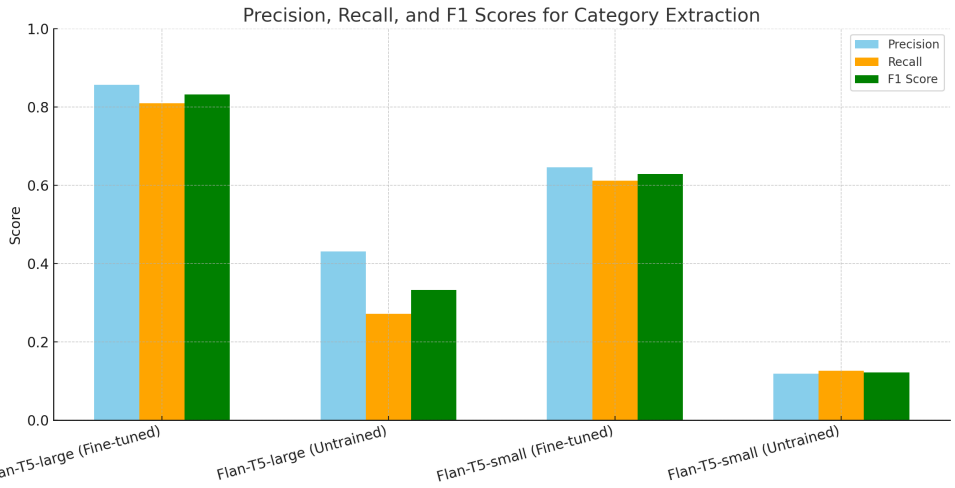

Fig. 4 – Precision, Recall, and F1 Scores for Keyword Extraction Models.

Category Extraction

Dataset themes/categories are essential for organizing datasets thematically, allowing users to navigate large data repositories more effectively. For category extraction as well, we employed T5-small and T5-large. Similar to keyword extraction, both base models performed poorly. As reflected in (Table 7) and Fig 5, T5-small had 223 rows where no category matches were found, while T5-large had 138 such rows. The number of rows with match percentages in the 75–100% range was also minimal, with only 30 matches for T5-small and 50 for T5-large. These results indicated that the models struggled to associate dataset descriptions with relevant categories due to a lack of understanding of semantic context.

Match Percentage	T5-small (Base)	T5-small (Finetuned)	T5-large (Base)	T5-large (Finetuned)
0%	223	100	138	44
$1-25\%$	5	4	11	2
$25-50\%$	20	22	79	14
$50-75\%$	1	1	1	3
$75-100\%$	30	152	50	216

Tab. 7 – Category Extraction: Match Percentages Across Models.

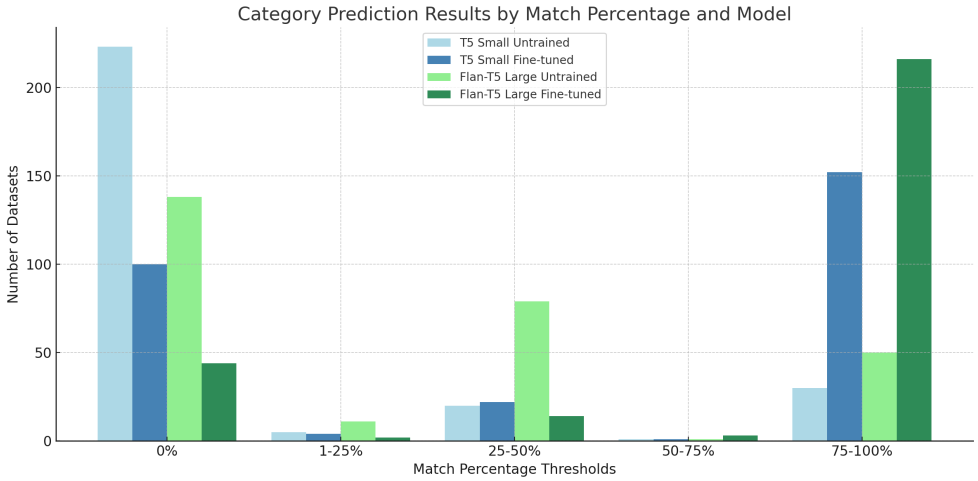

Fig. 5 – Match Percentages Across Models for Category Extraction.

Fine-tuning again led to significant improvements in match percentages. The number of rows with no matches was reduced to 100 for T5-small and 44 for T5-large. Meanwhile, rows with matches in the 75–100% range increased to 152 for T5-small and 216 for T5-large. This demonstrated a notable enhancement in the models' ability to align their predictions with human-curated categories.

Model	Precision	Recall	F1 Score
T5-large (Finetuned)	0.8566	0.8092	0.8322
T5-large (base)	0.4303	0.2710	0.3326
T5-small (Finetuned)	0.6459	0.6118	0.6284
T5-small (base)	0.1188	0.1258	0.1222

Tab. 8 – Category Extraction: Precision, Recall, and F1 Scores.

The improvement in match rates was further reflected in accuracy metrics. As observed in (Table 8) and Fig 6, the fine-tuned T5-small model achieved a precision of 0.6459, a recall of 0.6118, and an F1 score of 0.6284. The fine-tuned T5-large model demonstrated even better performance, with a precision of 0.8566, a recall of 0.8092, and an F1 score of 0.8322. These results highlight the effectiveness of fine-tuning in enabling models to interpret dataset descriptions better and predict relevant categories to enable more efficient thematic searches.

Fine-tuning consistently enhanced the performance of both models, with Flan-T5-large outperforming Flan-T5-small in both tasks. Figures 3 and 5 visually highlight the improvements in match percentages for each task, while Figures 4 and 6 emphasize the precision, recall, and F1 score advancements.

4. CONCLUSION AND FUTURE WORKS

This study explored the potential of fine-tuning Large Language Models (LLMs), specifically T5-small and Flan-T5-large, for generating keywords and categories for open datasets. While the base models struggled in both tasks, fine-tuning significantly improved their performance. Although both models demonstrated considerable improvements, T5-large consistently outperformed T5-small across all metrics.

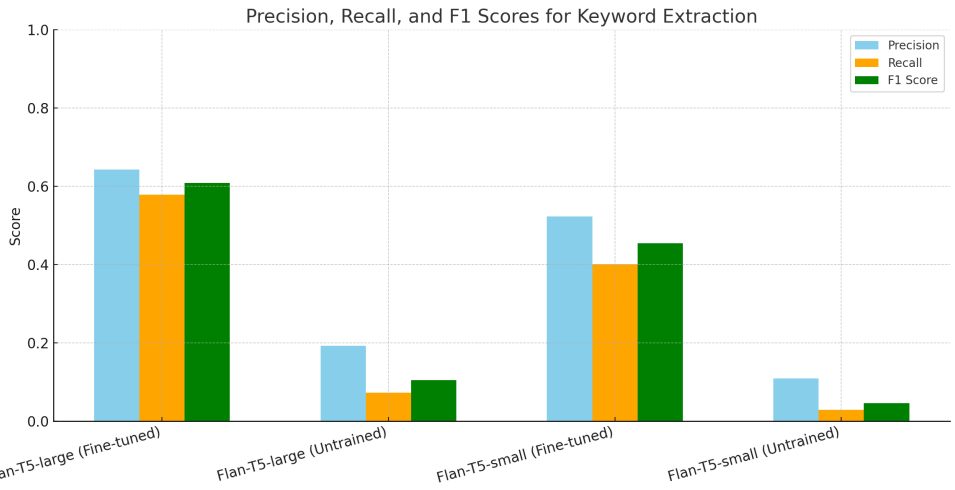

Fig. 6 – Precision, Recall, and F1 Scores for Category Extraction Models.

For keyword extraction, the base T5-small model achieved an F1 score of 0.0455, while T5-large had 0.1051. After fine-tuning, T5-small improved significantly with an F1 score of 0.4538, and T5-large achieved 0.6085, demonstrating its ability to generate human-like keywords that will improve dataset discoverability. Similarly, for category extraction, the base T5-small model had an F1 score of 0.1222, and T5-large scored 0.3326. With fine-tuning, T5-small improved its F1 score to 0.6284, while T5-large excelled with 0.8322, showcasing its capability to capture complex semantic relationships. These results highlight the potential of LLMs to generate metadata for missing or incomplete entries in open data portals, thereby improving dataset findability and discoverability.

Future work could further explore several avenues to enhance the scope and accuracy of metadata generation. Expanding the dataset to include more extensive and diverse examples could improve model generalizability and enable it to handle a wider range of contexts. Employing other large language models, such as GPT, LLaMA, and DeepSeek-R1, could significantly impact the study's accuracy and performance outcomes. Multilingual support would allow the models to generate metadata in multiple languages, catering to a broader audience. Additionally, integrating actual dataset samples or other metadata elements, such as titles, publishers, catalogs, or geographic information, as input features could improve prediction quality. Following keyword and category extraction, we want to generate the description using the dataset's contents. Another promising direction involves incorporating user feedback from data portal users and publishers, enabling iterative refinement of generated metadata to better align with user needs.

By addressing these areas, future studies can significantly contribute to a more accessible, interoperable, and user-friendly open data ecosystem.

Acknowledgement

Funding or Grant

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569. The opinions expressed in this document reflect only the author's view and in no way reflect the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information it contains.

Data/Software Access Statement

The data for this is placed at the following repository: https://github.com/umairahmedq/MDEusingLLM.

Contributor Statement

Umair Ahmed contributed in Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Software, Resources, Validation, Visualization, Writing – Original Draft, Writing - Review & Editing. Andrea Polini contributed in Conceptualization, Supervision, Project administration, Funding acquisition, Resources, Writing - Review & Editing.

Use of AI

During the preparation of this work, the author(s) used [Grammarly and OpenAI] in order to [refine the language and enhance clarity in communication]. After using this tool/service, the author(s) reviewed, edited, made the content their own and validated the outcome as needed, and take(s) full responsibility for the content of the publication.

Conflict Of Interest (COI)

Conflict Of Interest (COI)*: There is no conflict of interest

References

Ahmed, U. (2023). Reimagining open data ecosystems: A practical approach using ai, ci, and knowledge graphs. *BIR Workshops*, 235–249.

Ahmed, U., Alexopoulos, C., Piangerelli, M., & Polini, A. (2024). Bryt: Automated keyword extraction for open datasets. *Intelligent Systems with Applications*, *23*, 200421.

Alharbi, R., Ahmed, U., Dobriy, D., Łajewska, W., Menotti, L., Saeedizade, M. J., & Dumontier, M. (2023). Exploring the role of generative ai in constructing knowledge graphs for drug indications with medical context. *Proceedings http://ceur-ws. org ISSN*, *1613*, 0073.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289.

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*.

Deng, H. (2024). The advancements and progresses of artificial intelligence-based keyword extraction methods. *2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, 580–585.

EU. (2025). European data portal [Accessed: 2025-01-21].

Grusky, M. (2023). Rogue scores. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1914–1934.

Gunawan, D., Sembiring, C., & Budiman, M. A. (2018). The implementation of cosine similarity to calculate text relevance between two documents. *Journal of physics: conference series*, *978*, 012120.

Kathiriya, S., Mullapudi, M., & Karangara, R. (n.d.). Optimizing ecommerce listing: Llm based description and keyword generation from multimodal data.

Kumar, S., & Gupta, P. (2015). Comparative analysis of intersection algorithms on queries using precision, recall and f-score. *International Journal of Computer Applications*, *130*(7), 28–36.

Lee, W., Chun, M., Jeong, H., & Jung, H. (2023). Toward keyword generation through large language models. *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 37–40.

Li, J., Wang, C., Fang, X., Yu, K., Zhao, J., Wu, X., & Gong, J. (2022). Multi-label text classification via hierarchical transformer-cnn. *Proceedings of the 2022 14th International Conference on Machine Learning and Computing*, 120–125.

Maragheh, R. Y., Fang, C., Irugu, C. C., Parikh, P., Cho, J., Xu, J., Sukumar, S., Patel, M., Korpeoglu, E., Kumar, S., et al. (2023). Llm-take: Theme-aware keyword extraction using large language models. *2023 IEEE International Conference on Big Data (BigData)*, 4318–4324.

Maratsi, M. I., Ahmed, U., Alexopoulos, C., Charalabidis, Y., & Polini, A. (2024). Towards cross-domain linking of data: A semantic mapping of cultural heritage ontologies. *Proceedings of the 25th Annual International Conference on Digital Government Research*, 165–176.

Ohm, P. (2024). Focusing on fine-tuning: Understanding the four pathways for shaping generative ai. *Columbia Science and Technology Law Review, Forthcoming*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Pottier, P., Lagisz, M., Burke, S., Drobniak, S. M., Downing, P. A., Macartney, E. L., Martinig, A. R., Mizuno, A., Morrison, K., Pollo, P., et al. (2023). Keywords to success: A practical guide to maximise the visibility and impact of academic papers. *bioRxiv*, 2023–10.

Purwar, A., & Sundar, R. (2023). Keyword augmented retrieval: Novel framework for information retrieval integrated with speech interface. *Proceedings of the Third International Conference on AI-ML Systems*, 1–5.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1–20.

Van Loenen, B., Zuiderwijk, A., Vancau-Wenberghe, G., Lopez-Pellicer, F. J., Mulder, I., Alexopoulos, C., Magnussen, R., Saddiqa, M., De Rosnay, M. D., Crompvoets, J., et al. (2021). Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda. *JeDEM-eJournal of eDemocracy and Open Government*, *13*(2), 1–27.

Würsch, M., Kucharavy, A., David, D. P., & Mermoud, A. (2023). Llms perform poorly at concept extraction in cyber-security research literature. *arXiv preprint arXiv:2312.07110*.