Exploring LLM Capabilities in Extracting DCAT-Compatible Metadata for Data Cataloging

Lennart Busch[1][a], Daniel Tebernum[2][b] and Gissel Velarde[1][c]
[1]*IU International University of Applied Sciences, Erfurt, 99084, Germany*
[2]*Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, 44147, Germany*

Keywords: Large Language Models, DCAT, Metadata, Data Catalogs, GPT-4o, Llama 3.1, Llama 3.2, Gemini 1.5.

Abstract: Efficient data exploration is crucial as data becomes increasingly important for accelerating processes, improving forecasts and developing new business models. Data consumers often spend 25-98% of their time searching for suitable data due to the exponential growth, heterogeneity and distribution of data. Data catalogs can support and accelerate data exploration by using metadata to answer user queries. However, as metadata creation and maintenance is often a manual process, it is time-consuming and requires expertise. This study investigates whether LLMs can automate metadata maintenance of text-based data and generate high-quality DCAT-compatible metadata. We tested zero-shot and few-shot prompting strategies with LLMs from different vendors for generating metadata such as titles and keywords, along with a fine-tuned model for classification. Our results show that LLMs can generate metadata comparable to human-created content, particularly on tasks that require advanced semantic understanding. Larger models outperformed smaller ones, and fine-tuning significantly improves classification accuracy, while few-shot prompting yields better results in most cases. Although LLMs offer a faster and reliable way to create metadata, a successful application requires careful consideration of task-specific criteria and domain context.

1 INTRODUCTION

Text-based data and documents play a critical role for companies and organizations, forming the backbone of numerous business processes and serving as the primary repository for business data (Rowe, 2012). However, these documents are often difficult to locate. Studies reveal that knowledge workers can spend 25%-98% of their work time searching for information contained in documents (Deng et al., 2017), and up to 83% have had to recreate existing documents because they were unable to find them within their organization's network (M-Files, 2019). In a study for the European Commission, the consulting firm PwC estimated that the cost of such data that does not comply with the FAIR principles costs the European economy alone 10.2 billion Euro per year (PwC, 2018).

An effective solution for managing data and addressing associated challenges is the implementation of data catalogs. It is a system that facilitates the findability, accessibility, and organization of data. It serves as a centralized platform for semantically classifying and organizing data sources (Ehrlinger et al., 2021). An important aspect of a data catalog is the use of metadata. However, the creation of metadata, particularly in large quantities for data catalogs, is a time-consuming process that relies heavily on the skills of individual users (Mondal et al., 2018). This dependency can lead to situations where data catalogs degrade into inoperable "data swamps" due to poor metadata quality (Eichler et al., 2021).

Therefore, the development of automated processes for generating metadata has been an enduring area of research for many decades (Jenkins et al., 1999). The development of Transformer-based neural networks by researchers at Google in 2017 led to significant advancements in the field of natural language processing (Vaswani et al., 2017). These advanced neural networks exhibit exceptional proficiency in natural language processing tasks, ranging from text classification to abstract writing and problem-solving (Hasselaar et al., 2023).

Building on these advances and addressing the

[a] https://orcid.org/0009-0001-8952-3523
[b] https://orcid.org/0000-0002-4772-9099
[c] https://orcid.org/0000-0001-5392-9540

Busch, L., Tebernum, D., Velarde and G.
Exploring LLM Capabilities in Extracting DCAT-Compatible Metadata for Data Cataloging.
DOI: 10.5220/0013458500003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 299-309
ISBN: 978-989-758-758-0; ISSN: 2184-285X

identified challenges, this study poses the following research question: *How effective are autoregressive LLMs at generating descriptive, DCAT-compatible metadata from text-based documents with human-level quality?*

Based on this research question, three hypotheses (H) have been formulated and will be tested in this study:

H1. Autoregressive LLMs can generate descriptive, DCAT-compatible metadata with accuracy comparable to human-curated metadata, with the quality of the output improving when transitioning from zero-shot to few-shot prompting.

H2. Fine-tuning autoregressive models specifically for classification tasks further enhances accuracy and outperforms few-shot prompting in this scenario.

H3. Larger autoregressive models, such as GPT-4o, outperform smaller models, like Llama 3.2 3B, in generating high-quality and consistent metadata.

Our **contribution** lies in creating a comprehensive guide that outlines best practices for optimizing LLMs in metadata generation. It addresses gaps in the existing literature by providing practical strategies for effective LLM use, eliminating the need for extensive frameworks or overly complex adjustments. [1]

This paper is organized as follows: Section 2 offers an overview of metadata, data catalogs, DCAT, and LLMs, along with related work to provide context for our study. Section 3 outlines the methodology, detailing the datasets, metrics, and experimental design. Section 4 presents the results, followed by a discussion of their theoretical and practical implications in Section 5. Finally, Section 6 summarizes the key findings, acknowledges limitations, and suggests directions for future research.

2 BACKGROUND

Metadata can be described as "data about data" (Pomerantz, 2015; Horodyski, 2022) or, more precisely, as "[...] structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." (National Information Standards Organization, 2004, p. 1). Metadata provides information about the data itself without having to process the actual data (Sabot, 2022). High-quality metadata enhances productivity, compliance, and scalability within organizations. It e.g. facilitates rapid data retrieval and clarifies the applicability of legal regulations (Roszkiewicz, 2010). However, "[m]etadata collection is expensive so incremental collection along the workflow is required" (Jeffery, 2020, p. 128). This, in conjunction with the need for high-quality metadata to achieve the described improvements (Quimbert et al., 2020), is driving the further development of automation solutions in companies (Ochoa and Duval, 2009). Metadata can be categorized into different classes such as descriptive, administrative and structural metadata (Riley, 2017). In this paper, we will focus on descriptive metadata, such as the title, description, and theme of an information object.

Data Catalogs "collect, create and maintain metadata" (Quimbert et al., 2020, p. 141). By doing so, they streamline data discovery, access, and curation (Ehrlinger et al., 2021; Jahnke and Otto, 2023) and thus support FAIR principles, (Labadie et al., 2020), data governance (Shanmugam and Seshadri, 2016) and data democratization (Eichler et al., 2022). They establish the connection between data supply and data demand (Jahnke and Otto, 2023). Data catalog solutions are highly diverse, with specialized implementations tailored to specific purposes (Zaidi et al., 2017; Jahnke and Otto, 2023). So far, limited research has been conducted on the design and requirements of data catalogs. However, the automation of metadata generation has shown to be an important feature (Petrik et al., 2023; Tebernum, 2024).

Data Catalog Vocabulary (DCAT) is the de facto standard schema for metadata management in data catalogs (Albertoni et al., 2023) and is managed by the World Wide Web Consortium (World Wide Web Consortium, 2024). It is built upon the resource-centric Dublin Core Metadata Initiative Metadata Terms (DCTERMS) and now emphasizes the description of datasets and data catalogs (Maali et al., 2010). In this study, we will incorporate the following DCAT-Properties:

- dcterms:title
- dcterms:description
- dcterms:creator
- dcterms:language
- dcterms:spatial
- dcat:issued
- dcat:keyword
- dcat:theme

Large Language Models (LLMs) are language models that employ transformer-based neural networks (Vaswani et al., 2017) and leverage deep learning algorithms (Hadi et al., 2023). These models have been pre-trained (Jurafsky and James, 2024) on large amounts of textual data (Hadi et al., 2023), enabling them to generate text and perform diverse

[1] All utilized prompts, results, and code are available in the following GitHub repository: https://github.com/l94779589/Exploring-LLM-Capabilities-in-Extracting-DCAT-Compatible-Metadata-for-Data-Cataloging

language-based tasks with exceptional, human-level performance (Hadi et al., 2023). In performing these tasks, they demonstrate strong thinking, planning, and decision-making skills, even if they have not been specifically trained for a particular task (Naveed et al., 2023).

Modern, commercial LLMs are predominantly decoder-only models, designed to generate text by predicting one token at a time based on preceding tokens in a sequence. This autoregressive architecture excels at generating fluent, contextually appropriate text and is highly adaptable across various downstream tasks (Wang et al., 2022). This study incorporates the following decoder-only models:

1. **Llama 3.1 8B & Llama 3.1 70B**, released by Meta in July 2024, supports a context window of 128,000 tokens (Meta, 2024a).
2. **Llama 3.2 3B**, released by Meta in September 2024, is optimized for mobile devices. This model supports a context window of up to 128,000 tokens and was created through knowledge distillation and structured pruning of the Llama 3.1 8B and 70B models (Meta, 2024b).
3. **Google Gemini 1.5 Flash & Pro**, introduced by Google in March 2024, consists of two variants: the smaller Flash and the larger Pro. Gemini 1.5 supports a context length of up to 2,000,000 tokens (Team et al., 2024).
4. **GPT-4o & GPT-4o Mini**, developed by OpenAI and released in July and August 2024, both support a context length of 128,000 tokens. (OpenAI Inc., 2024a,b).

Related work on LLMs has increased exponentially, with the number of research papers growing from 42 in 2018 to over 28,000 in 2024. However, work related to metadata and data catalogs remains fragmented and, in addition to new methods and frameworks, is focused on specific elements such as keywords and classification.

Despite this growth in LLM research, the field around metadata generation for data catalogs remains fragmented. Existing studies focus on isolated elements such as keywords and classification. While work such as PromptRank (Kong et al., 2023), LaFiCMIL (Sun et al., 2024), and efforts to improve FAIR-compliant ecosystems (Arnold et al., 2024) represent valuable advances. However, these initiatives primarily address specific technical challenges, methods, or frameworks, and none of them adopt a truly comprehensive and practical approach that integrates multiple DCAT metadata properties, diverse LLM model families, and size variants, while ensuring broad accessibility to a wide range of users.

3 METHOD

The method design is illustrated in Figure 1. We converted each dataset into an Excel file, processed each item in the dataset with an LLM, and stored the LLM's responses alongside the human-annotated ground truth in a separate Excel file for subsequent evaluation. For each DCAT property, at least one dataset was utilized, see Section 3.1. The same prompt design was applied across all LLMs for each DCAT property and dataset. For straightforward information extraction (e.g., dcterms:creator), we used predefined examples in few-shot prompts; for more complex, context-dependent properties, dynamic few-shot prompts were employed. Examples were retrieved from a vector database—which stores data as high-dimensional vectors for similarity searches—drawn either from the validation subset or a thematically similar dataset. The embedding model used was Google's "text-embedding-004." (Google LLC, 2024). All LLMs used the same settings: for all DCAT properties, the temperature was set to 0 for reproducibility, and a repetition penalty of 0.5 was applied specifically for dcat:keyword to encourage diversity. The LLM's role was chosen based on its task, reflecting role sensitivity in LLMs (Zheng et al., 2023).

For the dcat:theme property, a Gemini 1.5 Flash model was fine-tuned, as classification tasks require heightened domain-specific precision compared to other metadata properties (Chalkidis et al., 2019). Due to limited resources, we did not perform fine-tuning on other attributes such as abstracts and keywords, nor did we perform such procedures on additional models.

3.1 Datasets

We evaluated 9 **datasets** across all eight DCAT-properties, including (Wang et al., 2018; Cohan et al., 2018; Goyal et al., 2022; Yucheng, 2024; Hattori et al., 2022; Kim et al., 2010; Augenstein et al., 2017; Chalkidis et al., 2019; Singh, 2023b).

dcterms:title: The dataset comprises 10,874 title-and-abstract pairs sourced from the ACL Anthology Network. On average, the titles and abstracts consist of 9 and 116 words each (Wang et al., 2018). The dataset was utilized by Mishra et al. (2021), whose results will serve as a baseline for this work. Few-shot prompting was implemented using a dynamic retrieval approach. The dataset for the examples consisted of 5,714 abstract-title pairs sourced from two other ACL Anthology datasets (ACLMeeting, 2023, 2024).

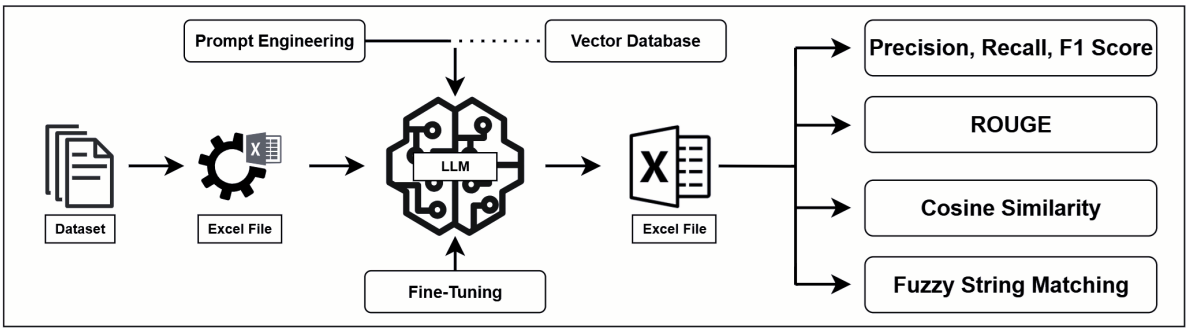

Figure 1: Method design.

dcterms:description: The dataset includes 133,000 papers and abstracts from PubMed.com and is split into training (94%), validation (3%), and test (3%) subsets (Cohan et al., 2018). The papers and abstracts have an average length of 3,016 and 203 words, respectively. The dataset was also utilized by Guo et al. (2021); their findings will serve as a benchmark for this research. For few-shot prompting, the validation subset of the dataset was employed.

dcterms:creator: The dataset consists of 1,160 scientific papers from ArXiv.com (Yucheng, 2024). It includes a total of 7,438 authors. Only the first page of each paper was presented to the LLMs. Few-shot prompting employed a predefined approach, where two examples were presented to the model.

dcterms:language: The dataset includes 3001 sentences sourced from English Wikipedia. Each sentence was translated into 101 languages by professional translators (Goyal et al., 2022). For this study, a subset of 5 sentences with their full set of 101 translations was selected. Few-shot prompting used a dynamic retrieval approach. The dataset for the examples included two other example sentences along with their 101 translations from the original dataset.

dcterms:spatial: The dataset initially contained 144 country-specific submissions to the UN (Hattori et al., 2022). After excluding non-English documents and duplicate EU submissions, 102 examples remained. To expand the dataset, we generated two additional versions of each example by translating it from English to another language and back to English using the Google Translate API, increasing the total to 306 examples. Few-shot prompting employed a predefined approach, where two examples were presented to the model.

dcat:issued: The dataset for this DCAT property is the same as the one used for "dcterms:creator". Few-shot prompting employed a predefined approach, where two examples were presented to the models.

dcat:keyword: Two datasets were utilized:

- The SemEval2010 dataset consists of 243 full scientific papers. Each paper, approximately 8,332 tokens in length, is annotated with keywords provided by the authors and professional editors (Kim et al., 2010).
- The SemEval2017 dataset consists of 493 paragraphs from ScienceDirect journal articles, averaging 178 tokens each. Keywords were annotated by both an undergraduate student and an expert (Augenstein et al., 2017).

Both datasets have been widely used in subsequent research, including the work of Kong et al. (2023), their findings will serve as a benchmark in this study.

Few-shot prompting was implemented using a dynamic retrieval approach, where three relevant examples were sourced from a third dataset called "Inspec", which comprises 2,000 abstracts of scientific journal papers (Hulth, 2003).

dcat:theme: Two datasets were utilized:

- EURLEX57K consists of 57,000 legal documents from the EU and is annotated using the EuroVoc thesaurus. While the EuroVoc thesaurus encompasses over 7,000 labels and their IDs, only 4,271 were actively assigned in this dataset. This dataset is divided into training (45,000 examples), validation (6,000 examples), and test (6,000 examples) subsets. The dataset was curated and first utilized by Chalkidis et al. (2019) and later reused by Sun et al. (2024), both results will serve as benchmarks for this study.
- The GIZ Sector Data dataset, comprises 22,674 items, divided into training (10,015 examples), validation (11,754 examples), and test (905 examples) subsets (Singh, 2023b). Each item in the test subset is annotated with labels from a set of 16 distinct labels, with an average text length of approximately 50 words per item. This dataset was utilized by Singh (2023a), whose results will serve as a benchmark.

Few-shot prompting used a dynamic retrieval approach, selecting three examples from the validation

subset of each dataset. Due to the extensive number of labels in the EUROLEX57K dataset, the workflow applied here differed from the other DCAT properties and involved four steps for few-shot prompting: (1) The LLM generated up to 10 keywords/phrases from the input text; (2) each keyword was matched against a vector database, which stored the labels individually, to find the two most relevant labels; (3) three examples were retrieved from another vector database based on the current input; and (4) the model received the combined input text, retrieved examples, and labels for final classification.

Furthermore, fine-tuning was applied to Gemini 1.5 Flash in Google AI Studio for this classification task, using the respective training subset of the datasets with the following settings:

- EURLEX57K: 5 epochs with a learning rate of 0.0006 and a batch size of 16.
- GIZ Sector Data: 8 epochs with a learning rate of 0.0003 and a batch size of 16.

3.2 Evaluation Metrics

Due to the heterogeneous nature of the DCAT properties included in this study, multiple evaluation metrics were employed. However, not all metrics were applicable to every property, as certain properties cannot be meaningfully assessed using specific metrics. The following section provides a detailed description of each metric and specifies the corresponding properties to which it was applied.

Precision, Recall and F Score are metrics for assessing binary classification performance. In multiclass classification, micro F1 extend these metrics (Grandini et al., 2020). Precision measures the proportion of correctly retrieved results among all retrieved instances, while recall evaluates the ability to identify all relevant instances (Dalianis, 2018). The F1 score balances precision and recall as their harmonic mean. The micro F1 score aggregates true positives, false positives, and false negatives across all classes, reflecting overall performance with greater weight on larger classes (Grandini et al., 2020). These metrics were applied to dcterms:creator, dcterms:language, dcterms:spatial, dcat:issued, dcat:keyword and dcat:theme. This metric was selected to ensure consistency with comparison benchmarks employed in this study and due to its widespread adoption in the field of NLP research.

ROUGE evaluates the quality of automatically generated text by comparing it with a reference sequence. This approach is based on various forms of n-gram and subsequence overlap between the candidate and reference texts, assessing the similarity in terms of lexical and structural alignment. ROUGE-1 evaluates the single word overlap between a candidate summary and the reference summaries. ROUGE-2 examines two-word overlap, capturing short phrase similarities between the candidate and reference summaries. ROUGE-L measures the longest common sub sequence to assess similarity (Chin-Yew, 2004). ROUGE was applied to dcterms:title and dcterms:description. We chose ROUGE over BLEU and other frameworks for its superior recall and content overlap capture, making it more suitable for summarization tasks.

Cosine Similarity determines the cosine of the angle formed between two vectors. This angle-based approach is useful in comparing the directionality of vectors, making it a popular measure in applications like evaluating semantic similarity between high-dimensional objects (e.g., word embeddings). By normalizing the vectors, cosine similarity produces values between -1 and 1, where higher values indicate a greater similarity (Steck et al., 2024). This metric was applied to dcterms:title, dcterms:description, dcat:keyword and dcat:theme. We chose cosine similarity as a complementary metric to capture semantic meaning beyond lexical matches in micro F1 or ROUGE.

Fuzzy String Matching is a technique used to find close matches for a given sequence when an exact match does not exist. One common fuzzy matching algorithm, which was utilized in this work, is the Levenshtein distance, which calculates the "edit distance" between two strings. This distance represents the smallest number of single-character modifications needed to convert one string into another. A shorter distance indicates greater similarity between the two strings (Kalyanathaya et al., 2019). This metric was applied to dcterms:creator to handle variations and inconsistencies in names, where exact matches may not always be necessary.

4 RESULTS

Tables 1–5 present the results. The best results for each metric are highlighted in bold, while benchmark results (when available) are underlined in gray.

Analysis of performance in text generation tasks for **dcterms:title** and **dcterms:description** revealed larger differences between model families than within them, with Gemini models slightly outperforming Llama and GPT, see Table 1. Although all of the tested models achieved strong ROUGE-1 scores—either surpassing or closely matching benchmarks and indicating robust lexical matching—their

Table 1: Results: dcterms:title (dc:ti) and dcterms:description (dc:des).

Model	Prompting	ROUGE-1		ROUGE-2		ROUGE-L		Cosine-Sim.	
		dc:ti	dc:desc	dc:ti	dc:desc	dc:ti	dc:desc	dc:ti	dc:desc
T5 (Guo et al., 2021)	Fine-tuned	-	0.502	-	0.248	-	0.467	-	-
GPT-2 (Mishra et al., 2021)	Zero-shot	0.123	-	0.044	-	0.170	-	-	-
Custom (Mishra et al., 2021)	Custom	0.340	-	0.156	-	0.308	-	-	-
Llama 3.1 8B	Zero-shot	0.406	0.449	0.192	**0.195**	0.329	0.259	0.661	0.873
Llama 3.1 8B	Few-shot	0.414	0.371	0.201	0.169	0.339	0.219	0.661	0.862
Llama 3.1 70B	Zero-shot	0.404	0.428	0.187	0.167	0.323	0.244	0.664	0.853
Llama 3.1 70B	Few-shot	0.406	0.411	0.192	0.160	0.329	0.234	0.661	0.842
Llama 3.2 3B	Zero-shot	0.406	0.449	0.192	0.188	0.334	0.256	0.657	0.875
Llama 3.2 3B	Few-shot	0.406	0.409	0.195	0.163	0.336	0.233	0.653	0.836
Gemini 1.5 Flash	Zero-shot	0.416	0.476	0.197	0.187	0.341	0.269	0.667	**0.885**
Gemini 1.5 Flash	Few-shot	0.431	**0.477**	0.211	0.187	0.358	**0.271**	0.668	**0.885**
Gemini 1.5 Pro	Zero-shot	0.422	0.464	0.198	0.168	0.350	0.258	0.664	0.884
Gemini 1.5 Pro	Few-shot	**0.441**	0.465	**0.219**	0.169	**0.372**	0.259	**0.672**	**0.885**
GPT-4o Mini	Zero-shot	0.398	0.463	0.182	0.168	0.313	0.253	0.664	0.879
GPT-4o Mini	Few-shot	0.411	0.465	0.193	0.169	0.336	0.255	0.664	0.878
GPT-4o	Zero-shot	0.400	0.459	0.187	0.180	0.323	0.256	0.661	0.883
GPT-4o	Few-shot	0.400	0.460	0.195	0.182	0.344	0.258	0.664	0.881

lower ROUGE-2 and moderate ROUGE-L scores highlighted challenges with structural coherence for both DCAT properties.

However, all tested models achieved moderate cosine similarity values for dcterms:title (0.653–0.672) and high values for dcterms:description (0.836–0.885), demonstrating their ability to capture the conceptual essence even if they fall short in achieving the nuanced phrasing, consistent style, and flow of human-generated text. Few-shot prompting provided only incremental improvements, with a notable distinction between model sizes: larger models showed slight enhancements with additional context, whereas smaller models exhibited no improvement or even degraded performance, particularly for dcterms:description.

For **dcterms:creator**, all models except Llama 3.1 8B achieved precision above 0.880 for the 90% threshold for author recognition, see Table 2. The differences between the remaining models and prompting techniques were minimal. However, as the threshold increased to 100%, larger models outperformed smaller ones, with GPT-4o achieving the highest precision score of 0.907. This suggests that models with higher parameter counts are better equipped to deliver greater accuracy.

Considering **dcterms:language**, most models achieved perfect scores (1.0) for the top 10 languages and German, see Table 3. GPT models excelled across all languages, with GPT-4o achieving a precision of 0.936 even for rarer languages. By contrast, the smaller Llama models struggled with language recognition—particularly for rarer languages, likely due to limited training data and weaker generalization capabilities. However, the strong performance of Llama 3.1 70B shows that increasing model parameters can significantly enhance multilingual understanding, even though it, like its smaller counterparts, officially supports only eight languages.

For **dcterms:spatial** and **dcat:issued**, most models achieved high precision scores, approaching 1, see Table 3. Few-shot prompting offered only a marginal advantage over zero-shot prompting for these two tasks. The only outlier here was Llama 3.2 3B for date detection which showed a weak performance in

Table 2: Results: dcterms:creator with exact match (100%) and 90% fuzzy matching threshold.

Model	Prompting	Precision	
		100%	90%
Llama 3.1 8B	Zero-shot	0.671	0.709
Llama 3.1 8B	Few-shot	0.394	0.417
Llama 3.1 70B	Zero-shot	0.888	0.926
Llama 3.1 70B	Few-shot	0.881	0.922
Llama 3.2 3B	Zero-shot	0.834	0.892
Llama 3.2 3B	Few-shot	0.840	0.890
Gemini 1.5 Flash	Zero-shot	0.832	0.880
Gemini 1.5 Flash	Few-shot	0.847	0.895
Gemini 1.5 Pro	Zero-shot	0.877	0.926
Gemini 1.5 Pro	Few-shot	0.886	0.929
GPT-4o Mini	Zero-shot	0.892	0.923
GPT-4o Mini	Few-shot	0.895	0.928
GPT-4o	Zero-shot	0.906	0.936
GPT-4o	Few-shot	**0.907**	**0.936**

Table 3: Results: dcterms:language, dcterms:spatial and dcat:issued.

Model	**Prompting**	**Precision**			
		dcterms:language		dcterms:spatial	dcat:issued
		Top 10 languages + GER	Rest		
Llama 3.1 8B	Zero-shot	0.691	0.573	0.977	0.994
Llama 3.1 8B	Few-shot	0.818	0.576	0.980	0.997
Llama 3.1 70B	Zero-shot	**1**	0.869	0.984	0.983
Llama 3.1 70B	Few-shot	**1**	0.922	**0.987**	0.997
Llama 3.2 3B	Zero-shot	0.982	0.460	0.971	0.300
Llama 3.2 3B	Few-shot	0.564	0.327	0.984	0.968
Gemini 1.5 Flash	Zero-shot	**1**	0.847	0.984	0.984
Gemini 1.5 Flash	Few-shot	**1**	0.858	0.980	0.998
Gemini 1.5 Pro	Zero-shot	**1**	0.880	0.984	0.963
Gemini 1.5 Pro	Few-shot	**1**	0.858	0.977	0.992
GPT-4o Mini	Zero-shot	**1**	0.880	0.984	0.970
GPT-4o Mini	Few-shot	**1**	0.907	**0.987**	0.997
GPT-4o	Zero-shot	**1**	0.903	-	0.990
GPT-4o	Few-shot	**1**	**0.936**	-	**0.999**

zero-shot prompting but was able to close this performance gap to the other models in few-shot prompting. These results suggest that the computational requirements for this task are relatively moderate.

For **dcat:keyword**, the results are displayed in Table 4. Llama 3.2 3B presents an interesting case: despite being the smallest model in the comparison, it performs exceptionally well on F1@5 scores for both SemEval datasets, outperforming larger models like GPT-4o. However, this strong initial performance comes with a notable limitation: its accuracy declines significantly at higher F1 metrics, revealing a struggle to sustain performance as task complexity increases. While few-shot prompting provides some improvement, it fails to fully address this limitation.

The analysis highlights a clear link between model size, task complexity, and performance. Larger models consistently outperform smaller ones as tasks demand more labels, maintaining better accuracy while smaller models face increasing limitations. This trend is evident in cosine similarity scores, where larger models generate keywords more closely aligned with human annotations. The performance gap becomes more pronounced with longer input sequences, demonstrating that higher parameter counts enhance the ability to process extended inputs. Additionally, larger models benefit more significantly from few-shot prompting compared to smaller ones, while smaller models, such as Llama 3.1 8B, even exhibit a decrease in performance for certain metrics under few-shot prompting.

Considering **dcat:theme**, the results shown in Table 5 demonstrate that fine-tuned models are still better at matching document content with an appropriate label compared to foundational models. Fine-tuning adapts model parameters to specific domains and classification schemes, connecting general language understanding with the task's semantic requirements. This allows models to recognize important textual patterns, word choices, and context that would be missed without domain-specific training. If fine-tuning isn't possible due to data limitations or other constraints, larger foundational models assign labels more accurately than smaller ones. The performance gap widens with more labels available for assignment.

Smaller models benefit more significantly from few-shot prompting when the labels are few and semantically distinct compared to larger models, thereby narrowing the performance gap with larger models. For instance, Llama 3.2 3B demonstrates an over 36% gain in micro F1 on the GIZ dataset, compared to only about 9% for GPT-4o. Conversely, when labels are numerous and semantically similar, both smaller and larger models show significant performance improvements when provided with examples. On EUROLEX57K, GPT-4o achieves over a 70% gain in micro F1, while Llama 3.2 3B shows a 63% increase.

Although neither foundational nor fine-tuned decoder-only models achieve the precision of a fine-tuned encoder-only model like BERT, all tested LLMs still assign labels that closely align semantically with human annotations. This alignment is reflected in high cosine similarity scores—ranging from 0.726 for Llama 3.2.3B in zero-shot mode to 0.833 for the fine-tuned Gemini 1.5 Flash. While the micro F1 scores

Table 4: Results: dcat:keyword SemEval2010 (SE10) and SemEval2017 (SE17).

Model	Prompting	F1@5		F1@10		F1@15		Cosine-Sim.	
		SE10	SE17	SE10	SE17	SE10	SE17	SE10	SE17
T5 (Kong et al., 2023)	Custom	0.172	0.271	0.207	0.378	0.214	0.416	-	-
Llama 3.1 8B	Zero-shot	0.193	0.239	0.193	0.311	0.162	0.294	0.766	0.884
Llama 3.1 8B	Few-shot	0.138	0.245	0.151	0.329	0.129	0.306	0.764	0.888
Llama 3.1 70B	Zero-shot	0.207	0.235	0.228	0.308	0.207	0.306	0.820	0.904
Llama 3.1 70B	Few-shot	0.211	0.225	0.235	0.299	0.218	0.292	0.840	0.901
Llama 3.2 3B	Zero-shot	0.206	**0.256**	0.207	0.293	0.175	0.248	0.802	0.854
Llama 3.2 3B	Few-shot	0.211	0.243	0.221	0.320	0.190	0.290	0.811	0.877
Gemini 1.5 Flash	Zero-shot	**0.226**	0.233	**0.341**	0.322	0.243	0.349	0.836	0.907
Gemini 1.5 Flash	Few-shot	0.225	0.241	0.264	0.341	**0.266**	0.374	0.842	0.910
Gemini 1.5 Pro	Zero-shot	0.198	0.212	0.225	0.296	0.217	0.310	0.836	0.905
Gemini 1.5 Pro	Few-shot	0.219	0.219	0.252	0.297	0.240	0.289	0.834	0.892
GPT-4o Mini	Zero-shot	0.184	0.227	0.200	0.309	0.187	0.325	0.832	0.907
GPT-4o Mini	Few-shot	0.202	0.235	0.233	0.323	0.221	0.342	**0.855**	0.912
GPT-4o	Zero-shot	0.176	0.231	0.215	0.330	0.220	0.366	0.833	**0.914**
GPT-4o	Few-shot	0.197	0.244	0.236	**0.349**	0.248	**0.391**	0.843	**0.914**

of these two models differ significantly (0.208 vs. 0.547), their assigned labels demonstrate a high degree of shared semantic meaning.

5 DISCUSSION

Up to now, it was not fully understood to what extent data catalogs can leverage LLMs to automatically extract metadata. Our work provides the theoretical foundation on which the data catalog community can build to improve their practical tools and advance knowledge through future research.

Our evaluation revealed distinct performance patterns based on task complexity and model characteristics. For basic information extraction tasks, all models performed well regardless of size. However, in complex tasks requiring deep contextual understanding, larger models demonstrated clear advantages. Text generation tasks showed that all models could effectively capture semantic meaning and domain-appropriate vocabulary, though they struggled with nuanced phrasing characteristic of human writing. Classification tasks revealed a strong correlation between model size and performance, particularly when handling semantically similar labels. Notably, fine-tuned decoder-only models consistently underperformed compared to encoder-only counterparts in classification tasks.

Few-shot prompting emerged as a key differentiator, with varying effects based on model size and task complexity. Smaller models showed performance improvements with few-shot prompting on simple tasks but experienced degradation when handling complex tasks with increased context load. Larger models effectively leveraged additional context across all scenarios, particularly excelling in complex tasks.

The study proposes a tiered approach for the utilization of LLMs in data catalogs for metadata generation. For basic tasks like information extraction, smaller models offer cost-effective solutions suitable for large-scale catalogs or resource-constrained environments. However, complex tasks requiring deeper context or specialized language benefit from larger models. Furthermore, a tiered implementation strategy based on model families is proposed, leveraging each model's strengths: Gemini for coherent titles and abstracts, GPT for language identification, com-

Table 5: Results: dcat:theme EUROLEX57K (57K) and GIZ Sector Data (GIZ).

Model	Prompting	Micro F1		Cos-Sim
		57K	GIZ	57K
BERT (Chalkidis et al., 2019)	Fine-tuned	0.732	-	-
BERT (Sun et al., 2024)	Custom	0.737	-	-
BERT (Singh, 2023a)	Fine-tuned	-	0.846	-
Gemini 1.5 Flash	Fine-tuned	**0.547**	**0.794**	**0.833**
Llama 3.1 8B	Zero-shot	0.257	0.540	0.766
Llama 3.1 8B	Few-shot	0.330	0.632	0.790
Llama 3.1 70B	Zero-shot	0.271	0.587	0.753
Llama 3.1 70B	Few-shot	0.279	0.673	0.752
Llama 3.2 3B	Zero-shot	0.208	0.433	0.726
Llama 3.2 3B	Few-shot	0.341	0.589	0.780
Gemini 1.5 Flash	Zero-shot	0.268	0.575	0.760
Gemini 1.5 Flash	Few-shot	0.345	0.666	0.780
Gemini 1.5 Pro	Zero-shot	0.295	0.584	0.764
Gemini 1.5 Pro	Few-shot	0.422	0.679	0.805
GPT-4o Mini	Zero-shot	0.251	0.583	0.746
GPT-4o Mini	Few-shot	0.288	0.669	0.757
GPT-4o	Zero-shot	0.253	0.634	0.744
GPT-4o	Few-shot	0.434	0.694	0.816

plex keyword generation, and classification. Lastly, smaller Llama models for straightforward tasks such as information and feature extraction. This approach optimizes both performance and cost-effectiveness.

While LLMs demonstrate strong capabilities, human oversight remains essential, particularly in sensitive or regulated domains. Organizations should implement hybrid workflows that combine AI efficiency with expert validation to ensure accuracy and compliance. While concerns such as hallucination are significantly mitigated in structured metadata extraction tasks—owing to input constraints that anchor outputs to source content (e.g., title generation from abstracts)—prudent validation remains advisable in scenarios with ambiguous inputs. These structured tasks inherently limit the model's creative latitude, reducing hallucination risks compared to open-ended generation, where outputs lack grounding in predefined data.

The emergence of mobile-optimized models and high-performing models like Llama 3.2 3B expands accessibility, enabling real-time metadata generation in resource-limited settings. These advancements also support the development of local, private AI-driven data catalogs and other applications. The key recommendation is maintaining flexibility in implementation - organizations should regularly evaluate their systems, considering both performance and cost-efficiency, while staying informed about emerging models that could optimize their metadata management processes.

6 CONCLUSION

We return to the three hypotheses raised:

H1. Autoregressive LLMs can generate DCAT-compatible metadata comparable to human-curated metadata in many scenarios, though their suitability varies by use case. They excel in generating semantically aligned titles and abstracts, reliably perform information extraction, and achieve moderate to high accuracy in classification and label creation while effectively capturing overall semantic meaning. Few-shot prompting improves precision and recall, benefiting smaller models on simpler tasks by bridging the gap to larger models, while enhancing complex reasoning for larger models.

H2. Domain-specific fine-tuning significantly improves classification performance, especially with numerous semantically similar labels. While few-shot prompting helps provide examples, it cannot fully capture subtle contextual relationships. Fine-tuning adapts model parameters to specific domains, making them more sensitive to particular word choices and contextual hints.

H3. Larger LLMs often outperform smaller ones, particularly in tasks requiring complex semantic reasoning and contextual interpretation. However, this performance gap varies by task—for simpler, constrained assignments, smaller models can match larger ones' accuracy while being more resource-efficient. The relationship between model size and performance remains task-dependent.

We recognize the following limitations. The standardized use of identical prompts across model families may have unintentionally favored specific architectures, given LLMs' sensitivity to prompt design. Furthermore, the subjectivity of human-generated metadata complicates evaluation, as ground truth often reflects personal interpretation rather than absolute correctness, making alternative valid annotations hard to classify as errors. External validity concerns focus on the generalizability of findings, as the research may have limited transferability to other domains where metadata generation dynamics differ across content types or industries.

Future work should apply this methodology to other domains, such as news media or social platforms, to validate LLMs' metadata generation capabilities. Consistent performance would confirm robustness across industries, while variations would reveal domain-specific challenges, like specialized terminology and structural differences, warranting further investigation for broader optimization. Further research should incorporate expert validation to complement quantitative metrics with real-world qualitative assessments, ensuring alignment with domain standards and usability needs.

REFERENCES

ACLMeeting (2023). Acl2023-papers. Online dataset. https://huggingface.co/datasets/ACLMeeting/ACL2023-papers.

ACLMeeting (2024). Acl2024-papers. Online dataset. https://huggingface.co/datasets/ACLMeeting/ACL2024-papers/tree/main/data.

Albertoni, R., Browning, D., Cox, S., Gonzalez-Beltran, A. N., Perego, A., and Winstanley, P. (2023). The w3c data catalog vocabulary, version 2: Rationale, design principles, and uptake. *arXiv preprint arXiv:2303.08883*.

Arnold, B. T., Theissen-Lipp, J., Collarana, D., Lange, C., Geisler, S., Curry, E., and Decker, S. (2024). Towards enabling fair dataspaces using large language models. *arXiv preprint*, arXiv:2403.15451.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). Semeval 2017 task 10: Scienceie-

extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., and Androutsopoulos, I. (2019). Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*.

Chin-Yew, L. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Dalianis, H. (2018). *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing : Imprint: Springer, Cham, 1st ed. 2018 edition.

Deng, D., Fernandez, R. C., Abedjan, Z., Wang, S., Stonebraker, M., Elmagarmid, A. K., Ilyas, I. F., Madden, S., Ouzzani, M., and Tang, N. (2017). The data civilizer system. In *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org.

Ehrlinger, L., Schrott, J., Melichar, M., Kirchmayr, N., and Wöß, W. (2021). Data Catalogs: A Systematic Literature Review and Guidelines to Implementation. In Kotsis, G., Tjoa, A. M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., and Khan, S., editors, *Database and Expert Systems Applications - DEXA 2021 Workshops*, volume 1479, pages 148–158. Springer International Publishing, Cham.

Eichler, R., Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. (2021). Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges. *Business Information Systems*, pages 269–279.

Eichler, R., Gröger, C., and Hoos, E. (2022). Data shopping — how an enterprise data marketplace supports data democratization in companies. *Lecture Notes in Business Information Processing*, 452:19–26. © 2022, The Author(s), under exclusive license to Springer Nature Switzerland AG.

Google LLC (2024). Text embeddings api documentation. Online documentation. Retrieved November 13, 2024, from https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., and Yang, Y. (2021). Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., and Mirjalili, S. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1:1–26. Preprint.

Hasselaar, E., Silva, A., Zahidi, S., Decety, N., Daugherty, P., Espinosa, H., Horn, A., Ryan, M., Nanan, C., O'Reilly, K., and Yosef, L. (2023). Jobs of Tomorrow large Language Models and Jobs – A Business Toolkit. White Paper, World Economic Forum.

Hattori, T., Takahashi, K., and Tamura, K. (2022). IGES NDC Database.

Horodyski, J. (2022). *Metadata matters*. Taylor and Francis, Boca Raton.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

Jahnke, N. and Otto, B. (2023). Data catalogs in the enterprise: applications and integration. *Datenbank-Spektrum*, 23(2):89–96.

Jeffery, K. (2020). Data curation and preservation. In *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges*, pages 123–139. Springer.

Jenkins, C., Jackson, M., Burden, P., and Wallis, J. (1999). Automatic rdf metadata generation for resource discovery. *Computer Networks*, 31(11–16):1305–1320.

Jurafsky, D. and James, M. (2024). *Speech and Language Processing*. https://web.stanford.edu/, 3rd ed edition.

Kalyanathaya, K. P., Akila, D., and Suseendren, G. (2019). A fuzzy approach to approximate string matching for text retrieval in nlp. *J. Comput. Inf. Syst. USA*, 15(3):26–32.

Kim, S. N., Medelyan, O., Kan, M.-Y., and Baldwin, T. (2010). SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 21–26, USA. Association for Computational Linguistics.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., and Bai, X. (2023). Promptrank: Unsupervised keyphrase extraction using prompt. *arXiv preprint arXiv:2305.04490*.

Labadie, C., Legner, C., Eurich, M., and Fadler, M. (2020). Fair enough? enhancing the usage of enterprise data with data catalogs. In *2020 IEEE 22nd Conference on Business Informatics (CBI)*, volume 1, pages 201–210. IEEE.

M-Files (2019). The 2019 intelligent information management benchmark report. White paper, M-Files. Accessed: July 7, 2024.

Maali, F., Cyganiak, R., and Peristeras, V. (2010). Enabling interoperability of government data catalogues. In *Electronic Government: 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne, Switzerland, August 29-September 2, 2010. Proceedings 9*, pages 339–350. Springer.

Meta (2024a). Introducing Llama 3.1: Our most capable models to date.

Meta (2024b). Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

Mishra, P., Diwan, C., Srinivasa, S., and Srinivasaraghavan, G. (2021). Automatic title generation for text with pre-trained transformer language model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 17–24. IEEE.

Mondal, T., Meenach, N., and Baba, M. S. (2018). Metadata creation methods: A study. *Metadata Creation Methods: A Study*, 7(2):177–182.

National Information Standards Organization (2004). Understanding metadata. Online booklet. Retrieved March 25, 2025, from https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Ochoa, X. and Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International journal on digital libraries*, 10:67–91.

OpenAI Inc. (2024a). Gpt-4o system card. Online report. Retrieved November 9, 2024, from https://cdn.openai.com/gpt-4o-system-card.pdf.

OpenAI Inc. (2024b). Hello gpt-4o. Web article. from https://openai.com/index/hello-gpt-4o/.

Petrik, D., Untermann, A., and Baars, H. (2023). Functional requirements for enterprise data catalogs: a systematic literature review. In *International Conference on Software Business*, pages 3–18. Springer Nature Switzerland Cham.

Pomerantz, J. (2015). *Metadata*. The MIT Press essential knowledge series. The MIT Press, Cambridge, Massachusetts ; London, England.

PwC (2018). *Cost-benefit analysis for FAIR research data: cost of not having FAIR research data.* Publications Office, LU.

Quimbert, E., Jeffery, K., Martens, C., Martin, P., and Zhao, Z. (2020). Data cataloguing. In *Towards interoperable research infrastructures for environmental and earth sciences: A reference model guided approach for common challenges*, pages 140–161. Springer.

Riley, J. (2017). Understanding metadata. *Washington DC, United States: National Information Standards Organization*, 23:7–10.

Roszkiewicz, R. (2010). Enterprise metadata management: How consolidation simplifies control. *Journal of Digital Asset Management*, 6:291–297.

Rowe, N. (2012). Handling paper in a digital age. Accessed: July 7, 2024.

Sabot, F. (2022). On the importance of metadata when sharing and opening data. *BMC Genomic Data*, 23(1):79.

Shanmugam, S. and Seshadri, G. (2016). Aspects of data cataloguing for enterprise data platforms. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 134–139. IEEE.

Singh, P. (2023a). mpnet-multilabel-sector-classifier. Online model repository. https://huggingface.co/ppsingh/mpnet-multilabel-sector-classifier.

Singh, P. (2023b). sector_data. Online dataset. https://huggingface.co/datasets/GIZ/sector_data.

Steck, H., Ekanadham, C., and Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, pages 887–890.

Sun, T., Pian, W., Daoudi, N., Allix, K., F. Bissyandé, T., and Klein, J. (2024). Laficmil: Rethinking large file classification from the perspective of correlated multiple instance learning. In *International Conference on Applications of Natural Language to Information Systems*, pages 62–77. Springer.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tebernum, D. (2024). *A Design Theory for Data Catalogs*. Phd thesis, TU Dortmund.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Q., Zhou, Z., Huang, L., Whitehead, S., Zhang, B., Ji, H., and Knight, K. (2018). Paper abstract writing through editing mechanism. *arXiv preprint arXiv:1805.06064*.

Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H. W., Beltagy, I., Launay, J., and Raffel, C. (2022). What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR.

World Wide Web Consortium (2024). Data Catalog Vocabulary (DCAT) - Version 3.

Yucheng, L. (2024). arxiv_latest. Online dataset. https://huggingface.co/datasets/RealTimeData/arxiv.

Zaidi, E., De Simoni, G., Edjlali, R., and Duncan, A. D. (2017). Data catalogs are the new black in data management and analytics. *Gartner, Consultancy Report*.

Zheng, M., Pei, J., and Jurgens, D. (2023). Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*, 8.