Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis

Petros Stavropoulos[1,2], Ioannis Lyris[1], Natalia Manola[3], Ioanna Grypari[1,3], Haris Papageorgiou[1]
[1]Institute for Language and Speech Processing, Athena R.C.
[2]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens
[3]OpenAIRE AMKE
pestavr@di.uoa.gr, {ioannis.lyris, igrypari, haris}@athenarc.gr, natalia.manola@openaire.eu

Abstract

Knowledge extraction from scientific literature is a major issue, crucial to promoting transparency, reproducibility, and innovation in the research community. In this work, we present a novel approach towards the identification, extraction and analysis of dataset and code/software mentions within scientific literature. We introduce a comprehensive dataset, synthetically generated by ChatGPT and meticulously curated, augmented, and expanded with real snippets of scientific text from full-text publications in Computer Science using a human-in-the-loop process. The dataset contains snippets highlighting mentions of the two research artifact (RA) types: dataset and code/software, along with insightful metadata including their Name, Version, License, URL as well as the intended Usage and Provenance. We also fine-tune a simple Large Language Model (LLM) using Low-Rank Adaptation (LoRA) to transform the Research Artifact Analysis (RAA) into an instruction-based Question Answering (QA) task. Ultimately, we report the improvements in performance on the test set of our dataset when compared to other base LLM models. Our method provides a significant step towards facilitating accurate, effective, and efficient extraction of datasets and software from scientific papers, contributing to the challenges of reproducibility and reusability in scientific research.

1 Introduction

Scientific research is dynamically and rapidly evolving, generating an overwhelming amount of knowledge in the form of research outputs. The vastness and the intricacies of scientific literature makes it impossible for researchers to keep up with all advancements in their respective fields, which is critical to their research. Consequently, knowledge discovery from scientific literature and research artifact analysis (RAA) have gained significant prominence as fields. In recent years, many new research artifact (RA) datasets have been constructed, with the goal of training models and creating benchmarks for the identification of a variety of RAs, both intangible (e.g. methods, tasks) and tangible (e.g. datasets, software) (Wang et al., 2022; Krüger and Schindler, 2020). Moreover, recent efforts have been directed towards identifying corresponding metadata and classifying those RAs based on their functions, such as their usage and provenance (Du et al., 2021; Schindler et al., 2021).

In this paper, our primary focus lies in introducing a novel RA dataset specifically designed for dataset and software extraction. Our RA dataset is constructed by leveraging ChatGPT [1] and the full-text of scientific publications in the field of Computer Science. It employs a human-in-the-loop manual curation process and comprises snippets of scientific text that encompass mentions of datasets and software (RA mentions). Each snippet includes a trigger keyword or keyphrase indicating the RA mention, as well as a curated list of essential metadata, such as the Name, Version, License, URL, Usage, and Provenance of the respective dataset or software.

This RA dataset stands out from conventional RA datasets commonly used in the academic literature, due to its unique formulation. Unlike conventional approaches to RAA that focus on Named RAs through Named Entity Recognition (NER) and entity linking, our methodology addresses a significant oversight. Those approaches often neglect unnamed and undocumented resources, leading to implications for open science and reproducibility, and rendering them out of scope. Our approach unifies those tasks, including both named and unnamed RA mentions. Each RA mention is systematically mapped to a corresponding RA, along with its associated metadata, as defined by the context of the sentence (Fig. 1). The primary reason for this approach is to create a dataset where all RA men-

[1] https://chat.openai.com/

tions, whether explicitly named or not, are treated with equal importance. This allows models trained on this RA dataset to effectively identify the presence of RAs even in more complex and ambiguous scenarios.

Additionally, we utilized the LoRA (Hu et al., 2021) method to power the human-in-the-loop process and fine-tune base LLM models on the constructed RA dataset. Employing a model trained on this RA dataset can streamline the detection of RAs within the body of scientific publications, consequently improving the reproducibility of experiments and fostering a comprehensive understanding of the research process. Furthermore, this approach contributes to resource reusability by establishing a collection of crucial resources, accelerating scientific progress, mitigating repetition, and encouraging cross-disciplinary collaborations.

In the subsequent sections we provide detailed insights into the methodology employed for dataset construction (Secs. 3.1, 3.2, 3.3, 4.1). We also discuss the training and evaluation of our LoRA models (Secs. 3.5, 4.2), showcasing their effectiveness in extracting RAs through comprehensive benchmarking on our dataset's test set (Sec. 6). By comparing them to other base LLMs, we demonstrate the feasibility of employing simpler LLM models for successful and reliable RAA.

Our key contributions[2] are as follows:

1. We created two novel datasets for RAA, containing both synthetic and real RA mentions. The construction of those RA datasets was aimed to address issues present in other RA datasets found in the literature, such as the lack of unnamed RA mentions or even of all named RA mentions in a given snippet.

2. We demonstrated the effective performance of fine-tuned LLMs in RAA. Specifically, we discovered that even small LLMs, like the Flan-T5 Base model, when fine-tuned on our RA datasets, excel at RAA, surpassing the performance of larger, base models.

3. We conducted a comprehensive qualitative evaluation of our novel RA datasets and the models trained on them.

[2]All data and software resources can be accessed at the following link: https://github.com/PetrosStav/Research-Artifact-Analysis-NLP-OSS-2023-Paper.

2 Related Work

RAA has gained significant attention in recent years. This extensive research has led to the introduction of many important RA datasets related to the disciplines of Computer Science, Biology, Sociology and more. At the same time, important breakthroughs have been made in the construction of new novel machine learning models aimed to achieve this task, taking advantage of a variety of different technologies like Recurrent Neural Networks (RNNs) (Zeng and Acuna, 2020; Hou et al., 2022; Schindler et al., 2020) and BERT-like architectures (Schindler et al., 2021; Färber et al., 2021).

The aforementioned RA datasets can be differentiated into two types with respect to their formulation and goal. The first type is characterized by abstract RAs in the form of tasks, processes, or materials like SemEval 2017 Task 10 (Augenstein et al., 2017), SciERC (Luan et al., 2018), SciREX (Jain et al., 2020), the methods dataset from (Färber et al., 2021) and SciRes (Zhao et al., 2019). In contrast, the second type is aimed to the identification of more strictly defined RAs, with the Rich Context Competition dataset created by the Coleridge Initiative[3] at New York University, NER Dataset Recognition (Heddes et al., 2021) and DMDD (Pan et al., 2023) being characteristic examples for dataset extraction and SoMeSci (Schindler et al., 2021), Softcite (Du et al., 2021) being characteristic examples for software extraction. Nevertheless, a mapping between those two cases is not always possible. This divides RAA into two tasks that are similar in concept but very different in practice. Furthermore, some RA datasets further encompass the collection of RA metadata. Two notable works that address the lack of such RA datasets are Softcite and SoMeSci.

Our dataset differs significantly from common standards in RA datasets through a holistic approach that prioritizes both named and unnamed RA mentions. We annotate all RA mentions within a snippet, effectively creating a connection between those sharing the same name. That approach, similar to those employed in the construction of the Softcite and SoMeSci datasets, allows us to utilize the information from all RA mentions for a specific RA, providing a more comprehensive view. Considering the array of definitions for RA mention "validity" across various RA datasets, it is essential

[3]https://coleridgeinitiative.org/

for us to demonstrate the robustness and applicability of our annotation schema.

Therefore, we compared our RA dataset with five highly regarded datasets of RA mentions, in order to highlight its novelty and differences, and to explore how our broader scope could benefit future applications. For example, we observed that many of the existing RA datasets lack annotations of unnamed RA mentions in a given snippet or even aim only for the identification of named datasets or software. The RA datasets we used for comparison were Softcite and SoMeSci for software mentions, and the Rich Context Competition dataset, NER Dataset Recognition and DMDD for dataset mentions.

The exploitation of those RA datasets has been largely demonstrated through the utilization of RNN- and BERT-based models handling NER tasks embodied by those datasets. In contrast, the non-NER configuration of our RA dataset fosters the training and deployment of alternative model types. In our work, we use LLMs, fine-tuned with the LoRA method, and tackle the RAA task as an instruction-based QA task. This approach serves a dual purpose: it enables us to evaluate those models' performance when deployed on such tasks, while simultaneously assessing the quality of our newly introduced RA dataset.

3 Synthetic Dataset

In this section we describe the construction process of the Synthetic dataset (Subsecs. 3.1, 3.2) and the conversion of the RA mentions to question-answer (QA) pairs (Subsec. 3.3). Next, we detail the splitting of the dataset into training, development, and testing sets (Subsec. 3.4), and conclude with the training of the LoRA model (Subsec. 3.5).

3.1 Dataset Creation

For the creation of our RA dataset, we strategically harnessed the capabilities of ChatGPT to generate a corpus of synthetic data imbued with mentions of datasets and software. We formulated a prompt (Tab. 6, App. A) that explained to ChatGPT the notion of RAs, highlighting aspects such as their validity, metadata, usage and provenance, along with the task of RAA. Subsequently, we supplied ChatGPT with positive examples that illustrated valid RAs, as well as negative examples of invalid RAs. Positive examples consist of snippets containing valid dataset or software mentions, while negative examples typically comprise snippets with triggers that refer to general or encyclopedic references (e.g. "most existing datasets"), which are out of scope for most of the current approaches. We then instructed ChatGPT to act as a data creator, maintaining the structure and style of the examples.

Our dataset is meticulously structured to include a comprehensive set of fields: Snippet, Type, Valid, Name, Version, License, URL, Provenance, and Usage. Within each RA mention, the snippet contained one or multiple sentences, accompanied by a trigger encapsulated within <m> and </m> tags that also specifies the RA type (dataset or software). The Name, Version, License and URL fields of the RA require a text span within the snippet; in cases where those are not present, a default value of "N/A" is assigned. The Provenance and Usage fields can take values "Yes" or "No" to indicate if the RA was created or used by the authors of the publication. It is essential to note that those values must be supported by textual evidence in the snippet. Thus, even if a RA is generally created or used in the publication, the value is marked "Yes" only if this fact is evident from the snippet itself. Two characteristic examples of a valid and an invalid RA mention instance are presented in Figs. 1 and 2 respectively.

Snippet	In their study, the authors utilized the PyTorch <m>library</m> (version 1.9.0) for deep learning experiments. PyTorch is released under the BSD-3-Clause license. For more information, visit https://pytorch.org/.
Type	Software
Valid	Yes
Name	PyTorch
Version	1.9.0
License	BSD-3-Clause
URL	https://pytorch.org/
Provenance	No
Usage	Yes

Figure 1: An example of a RA mention containing all metadata.

Snippet	We leveraged the power of the Apache Spark framework for distributed <m>data</m> processing. The code implementation is available on our project's GitHub repository.
Type	Dataset
Valid	No

Figure 2: An example of an invalid RA mention.

Subsequently, our team of human curators[4] generated additional examples using ChatGPT, by specifying a range of attributes for ChatGPT to focus on. Those included creating specific types

[4]The team of human curators comprises two MSc students in Natural Language Processing (NLP) that worked on the task.

of RAs such as software or datasets, using a manually curated set of keywords and keyphrases for triggers, including metadata, and indicating usage and provenance. Additionally, the curators were able to determine the domain of the examples, such as Computer Vision, NLP, BioInformatics, and so on, or even specific linguistic features like using complex language or mentioning RAs in several sentences. Furthermore, effort was given to generating robust negative examples, taking into account their complexity, diversity, and linguistic function within the snippet.

The curated dataset of synthetic RA mentions served as the seed for generating an augmented set of positive and negative examples through a human-in-the-loop process, resulting in the creation of an expanded corpus of high-quality synthetic RA mentions.

Moreover, we made an effort to address the complex challenge of capturing snippets with multiple RA mentions that pertain to more than one RA of the same or different type. That mirrors more accurately the true complexity and nature of the task. The trigger words were derived from a manually curated set of keywords and keyphrases, which included the names of the RAs present within the snippets. Consequently, models trained on our RA dataset are equipped to adeptly extract RAs employing various trigger detection mechanisms and are also enabled to acquire entity linking capabilities, especially in scenarios where multiple triggers (e.g., names and keyphrases) pertain to the same RA.

3.2 Synthetic Data Augmentation

In the following stage, we employed a T5 model, which had been trained on the ChatGPT paraphrase dataset (Vladimir Vorobev, 2023), to augment our synthetic data via a paraphrasing technique. This involved substituting the trigger word in each snippet with the [MASK] token, followed by running the model to generate five paraphrased renditions of the snippet. Each paraphrased snippet was then checked, to ensure the presence of a single [MASK] token within each snippet, thereby filtering out any spurious hallucinations and noise.

3.3 QA Pairs Construction

Following the creation of the gold synthetic dataset, we converted all RA mentions into QA pairs for each metadata field, transforming the RAA to an instruction-based QA task. The questions used are depicted in Tab. 9 (App. C), with the value of each metadata field serving as the answer (Fig. 3). Those QA pairs were then structured into an input-output format suitable for the LoRA training of the LLM, further details of which will be presented in Sec. 3.5.

In an effort to enrich our dataset, we introduced a "special" type of QA pairs (Fig. 4) that are generated from the unique snippets of the RA dataset, devoid of any <m> and </m> tags, by enumerating all the RAs contained within each snippet. The construction of the QA pairs for those instances adhere to the aforementioned methodology. The question is consistent across all instances, as illustrated in Tab. 9 (App. C). The answer encompasses a list of all RAs found within the snippet, classified by their Type and Name, or marked as "unnamed" in the absence of a Name. Multiple RAs are separated by the "|" symbol. Those are then subjected to a similar augmentation process, discarding paraphrased snippets that exhibited a discrepancy in the count of RAs compared to the original instances.

Snippet	Our experiments were conducted using the data processing software datapro. The <m>software</m> version used was 1.5. It is distributed under the GNU Lesser General Public License.
Question	What is the name of the software defined in the <m> and </m> tags?
Answer	datapro

Figure 3: An example of QA pair.

Snippet	The CIFAR-10 dataset was used by the authors to assess the effectiveness of their image classification algorithm. This data set is freely available at https://www.cs.toronto.edu/kriz/cifar_fra.html.
Question	List all artifacts in the above snippet.
Answer	dataset : CIFAR-10 | software : unnamed

Figure 4: An example of a "special" QA pair.

As detailed in Tab. 1, prior to augmentation, the Synthetic dataset consisted of 305 unique snippets, 1616 RA mentions, and 10212 QA pairs. After the augmentation process, these figures were expanded to 4235 unique snippets, 5446 RA mentions, and 35475 QA pairs. More detailed statistics about the Synthetic dataset are depicted in Tab. 7 (App. B).

3.4 Train-Dev-Test Split

Given the meticulous process employed in the creation of the Synthetic dataset, special attention was required to split our data into training, development and testing sets. As mentioned previously and showcased in Tab. 1, although the dataset creation process yielded a total of 5446 RA mentions the original count of unique snippets was 305. For

the process of data splitting, it is imperative to select instances from those unique snippets, rather than the final instances. This was purposely done to avoid the issue of knowledge leaking from the training set to the test set.

Furthermore, it is crucial to ensure balance among the three sets in terms of the RA types, their validity (i.e. positive vs. negative instances), the inclusion of each metadata field, as well as the RA provenance and usage. To achieve this, we conducted a systematic approach that considered the distribution and characteristics of the RA mentions within each set and across the three sets as a whole, ensuring a comprehensive and fair representation of the RA mentions.

3.5 LoRA Finetuning on the Synthetic Dataset

After transforming the Synthetic gold dataset to QA pairs, we used it to fine-tune a Flan-T5 Base model (Chung et al., 2022) using the LoRA method (Hu et al., 2021). This model was chosen as it has a relative good performance-to-parameter ratio and can be used even from smaller research teams, with limited computational resources.

We trained a LoRA model on top of the Flan-T5 Base model on the QA pairs using the Huggingface PEFT library (Sourab Mangrulkar, 2022) for training and the Weights and Biases (W&B)[5] platform for logging, visualizations and the sweep hyperparameter tuning.

We trained our LoRA model on a single Quadro RXT 5000 GPU for approximately 315 hours using a W&B sweep hyperparameter tuning setting. We optimized for the best evaluation loss on the development set in each run, and implemented an early stopping mechanism with a patience of three epochs to ensure efficiency. We achieved our best model with the hyperparameters $r = 16$, $alpha_lora = 16$, $lora_dropout = 0.4$, $max_epochs = 5$.

4 Hybrid Dataset

This section explores the creation of a Hybrid dataset, focusing on the integration of RA mentions from synthetic and real snippets from scientific publications (Subsec. 4.1), and fine-tuning a LoRA model on the Hybrid dataset to enhance generalizability and performance for the RAA task (Subsec. 4.2).

[5] https://wandb.ai/site/research

4.1 Real Dataset Creation

As our Syntetic dataset was composed exclusively of synthetic RA mentions, we sought to expand it by incorporating gold-annotated RA mentions in real snippets from scientific publications (GARS dataset). This integration aimed to mitigate some of the biases in the Synthetic dataset construction, which originated from the repetitive and template-driven generated language used in the RA mentions. Such formatting is a concern as it could potentially impede the generalization capabilities and performance of models trained on the data.

Utilizing our best LoRA fine-tuned model on top of Flan-T5 Base (LoRA-Sy), we developed a tool that allows for automatic annotation of snippets. The same team of human curators employed this tool to annotate the full-text PDF files of a small collection of scientific publications within the Computer Science domain. Subsequently, this data underwent meticulous curation, expansion, and augmentation, as detailed in the previous sections. Similarly to the Synthetic dataset, the RA mentions were then converted to QA pairs to transform the RAA task into an instruction-based QA task. By combining the Synthetic dataset with the scientific publications dataset, we construct our Hybrid dataset (Tab. 1) comprising 15247 QA pairs from 2539 RA mentions spotted in 382 unique snippets, prior to augmentation and 45136 QA pairs from 7112 RA mentions spotted in 5230 unique snippets after the augmentation. More detailed statistics about the Hybrid dataset are depicted in Tab. 8 (App. B).

4.2 LoRA Finetuning on the Hybrid Dataset

We also performed fine-tuning on a LoRA model using our Hybrid dataset. The LoRA model was trained over a period of approximately 17 hours. The selection of hyperparameters was consistent with those that yielded the highest performance on the Synthetic dataset 3.5. The decision to use those particular settings was informed by the composition of the Hybrid dataset, which expanded upon the Synthetic dataset by integrating the GARS data into the training, dev, and test sets. This approach facilitated an ablation study on the test subset of the dataset, the findings of which will be explored in a subsequent section.

	Synthetic						Hybrid					
	Original			Augmented			Original			Augmented		
	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all
Unique snippets	198	225	305	2051	2301	4235	258	298	382	2350	3047	5230
RA mentions	741	875	1616	2555	2891	5446	1010	1529	2539	3017	4095	7112
QA pairs	4548	5664	10212	15559	18592	35475	5921	9326	15247	18055	25504	45136

Table 1: Statistics for the Synthetic and Hybrid datasets.

5 Dataset Analysis & Comparisons

The previously introduced Synthetic and Hybrid datasets consist of 35475 and 45136 QA pairs respectively. Each QA pair revolves around a question pertaining to a RA mention, as delineated in Tab. 9 (App. C). Although large in quantity, they encompass only 5446 and 7112 RA mentions respectively, distinguishing them from large-scale collections like DMDD with 449798 dataset mentions. However, our aim deviates from such vast RA datasets, prioritizing quality over sheer quantity due to the different annotation methodology.

A key characteristic of our dataset is that it is comprised of QA pairs that have been generated from a smaller pool of snippets. Those snippets, containing references to RAs, have undergone various rounds of annotation and paraphrasing. This process produces a multitude of differently phrased snippets, each providing unique RA mentions of the same RA while preserving the original contextual information. This depth of RA mentions is unique to our RA dataset and sets it apart from others. To further illustrate where our dataset sits within the existing research landscape, we offer a comparison of our dataset's key statistics with those from other established RA datasets in Tab. 2.

Our manual annotation process has been executed rigorously to ensure quality enhancements across several aspects:

- All dataset and software mentions have been annotated for a given snippet, regardless of whether they are named, unnamed, or invalid. That approach ensures comprehensive coverage of all potential cases of RA mentions within our RA dataset.
- We have strictly defined what constitutes a negative example in our RA dataset, following the convention presented in subsection 3.1. This approach ensures clarity by precisely defining how the snippets without any RA mentions should be identified.
- We have meticulously curated our snippets to highlight the diversity and complexity of RA mentions, as outlined in previous sections. That includes snippets with multiple RA mentions: (a) of the same or different type, (b) of new, named, or unnamed RAs, and (c) of new RAs (named or unnamed) interspersed with already seen RAs. Those intricate patterns are often stumbling blocks for models not properly trained to handle them. Through our purposeful inclusion of such diverse and complex RA mention patterns, we aim to train models on our RA dataset to effectively manage a wide range of situations, thereby enhancing their generalizability.

The emphasis on manual annotation in our RA dataset is a result of our efforts to address the issues we have encountered in other RA datasets. Those issues include the following:

- In the case of named RAs an issue can be observed in the DMDD dataset (Pan et al., 2023), which was constructed by matching terms found in the Paper with Code repository[6]. Despite resulting in the inclusion of valid terms, this approach fails to capture all named datasets. An illustrative example can be found in the sentence "We evaluate trained translation models on wmt13 (Bojar et al., 2013) and wmt14 (Bojar et al., 2014) for en-es and en-fr, respectively." from the evaluation subset of the DMDD dataset. Here, only the "wmt14" dataset has been annotated due to the absence of a matching term for "wmt13" in the dataset construction.
- In terms of unnamed RAs, a significant number of RA datasets fail to acknowledge those RA mentions or even consider them as negative examples. The latter case is a notable characteristic of the NER Dataset Recognition dataset (Heddes et al., 2021), where unnamed dataset mentions are labeled as negative examples (Geen datasets), leading to differential metric definitions in comparison to other RA datasets. The broader issue of unnamed RAs omission can also be observed in the Rich Context Competition dataset, where phrases like "our data" would not be annotated as a RA mention. To a lesser extent, this issue is

[6] https://paperswithcode.com/

	Dataset	Instance Unit	Number of RA Mentions	Metadata Available
Dataset mentions	**Ner Dataset Recognition** (Heddes et al., 2021)	sentence	3416	-
	Rich Context Competition	paper	36597	-
	bioNerDS (Duck et al., 2013)	paper	920	-
	NLP-TDMS (Hou et al., 2019)	paper	1164	-
	TDM-Sci (Hou et al., 2021)	sentence	612	-
	SciERC (Luan et al., 2018)	abstract	770	-
	SciREX (Jain et al., 2020)	paper	10548	-
	DMDD (Pan et al., 2023)	paper	449798	-
	Synthetic Dataset (ours)	snippet	2555	URL, License, Version, Provenance, Usage
	Hybrid Dataset (ours)	snippet	3017	URL, License, Version, Provenance, Usage
Software mentions	**bioNerDS** (Duck et al., 2013)	paper	2625	-
	SoSciSoCi (Schindler et al., 2020)	method section/sentence	2385	-
	Softcite v.1 (Du et al., 2021)	paragraph	4093	URL, Version, Developer
	Softcite v.2 (Howison et al., 2023)	paragraph	5134	URL, Version, Type, Developer
	CZ Software Mentions (Istrate et al., 2022)	sentence	20.11M	Type
	SoMeSci (Schindler et al., 2021)	method section/full text/sentence	3756	URL, License, Version, Citation, Extension, Type, Provenance, Usage, Developer
	Synthetic Dataset (ours)	snippet	2891	URL, License, Version, Provenance, Usage
	Hybrid Dataset (ours)	snippet	4095	URL, License, Version, Provenance, Usage

Table 2: Comparison of dataset and software mention statistics between ours and other RA datasets.

also observed in the Softcite (Du et al., 2021) and SoMeSci (Schindler et al., 2021) datasets, which miss out on references to machine learning models in certain cases.

Finally, it is important to acknowledge that our definition of a RA mention within the context of a snippet aligns significantly with the construct employed by the SoMeSci dataset. The SoMeSci dataset, which is dedicated exclusively to software mentions, classifies those mentions into four categories: Application, Plugin, Operating System, and Programming Environment, while also annotating an extensive array of metadata. In our case, we have adopted this formalism to encompass RA mentions, without delving further into the corresponding subtypes. The metadata we have annotated includes the URL, Version, License, Provenance, and Usage of each RA mention.

6 Experimental Results on Test Set

We conducted a series of tests using the top-performing versions of the aforementioned LoRA fine-tuned models as well as the original Flan-T5 Base and XL models. Those were employed to evaluate the quality of information included within our Synthetic and Hybrid datasets.[7] Specifically, we computed the respective scores of those four models on the test sets of both RA datasets, aiming to discern the advantages of the fine-tuning process and to perform an ablation study to investigate the effect of fine-tuning a model using synthetic data versus synthetic data expanded by real data.

To ensure a balanced comparison between the original and fine-tuned models, we modified the prompts utilized by the original models (Tab. 10, App. C). Those adjusted prompts refrain from presupposing any specific training of the models to answer in a particular manner, and consequently, provide more comprehensive instructions. The F1 score measures successful identification of a valid RA mention or presence of specific metadata (Name, License, Version, URL). For Usage and Provenance metadata, it denotes successful identification of the RA's use or creation by authors. The exact match (EM) score, applicable for Name, License, Version, and URL metadata, determines exact lowercase match of the metadata text from a provided snippet, provided the model correctly identifies the presence of that metadata. The lenient match (LM) score checks if the model's answer in lowercase is within the gold truth, or vice versa. The results achieved by the four models on both RA datasets are outlined in Tabs. 3 and 4. Moreover, detailed evaluation results of named versus unnamed RA mentions can be found in App. D.

The original Flan-T5 Base and XL models achieved a high level of success in discerning the validity of RA mentions, as well as the associated metadata such as the License, Version, and URL. As anticipated, superior performance was observed on the less complicated Synthetic dataset. Furthermore, the XL model surpassed the Base model, particularly in identifying the Name, Usage, and Provenance. Those outcomes endorse the high semantic quality of both our RA datasets and showcase the efficacy of LLMs in the RAA task.

The significance of the LoRA fine-tuning procedure is evident in the scores presented in Tabs. 3 and 4. The fine-tuned models remarkably outperformed their base counterparts. In addition, the LoRA Hybrid model (LoRA-Hy) generally outperforms the LoRA-Sy model when evaluated on both the Synthetic and Hybrid datasets, indicating that the inclusion of additional real-world RA mention instances improves those models. Similar findings were observed in the model tests on the GARS dataset, as shown in Tab. 5. Notably, in the GARS

[7] Due to its substantial size, the Flan-T5 XXL model was not incorporated in this comparison.

	Flan T5 base			Flan T5 XL			LoRA-Sy			LoRA-Hy		
	Identification	Extraction		Identification	Extraction		Identification	Extraction		Identification	Extraction	
	F1	EM	LM	F1	EM	LM	F1	EM	LM	F1	EM	LM
Valid	0.841	-	-	0.870	-	-	0.967	-	-	**0.974**	-	-
Name	0.358	0.709	0.835	0.681	0.787	0.900	**0.887**	**0.917**	**0.962**	0.876	0.905	0.952
License	0.926	0.502	0.813	0.928	0.635	0.778	**0.946**	**0.700**	**0.818**	0.944	0.685	**0.818**
Version	0.677	0.620	0.816	0.942	0.687	**0.865**	0.975	0.620	0.626	**0.979**	**0.755**	0.767
URL	0.677	0.342	0.355	0.980	0.539	0.566	0.981	0.618	0.645	**0.982**	**0.632**	**0.658**
Usage	0.377	-	-	0.772	-	-	0.911	-	-	**0.914**	-	-
Provenance	0.537	-	-	0.647	-	-	0.939	-	-	**0.961**	-	-

Table 3: Experimental results on the test set of the Synthetic dataset.

	Flan T5 base			Flan T5 XL			LoRA-Sy			LoRA-Hy		
	Identification	Extraction		Identification	Extraction		Identification	Extraction		Identification	Extraction	
	F1	EM	LM	F1	EM	LM	F1	EM	LM	F1	EM	LM
Valid	0.766	-	-	0.822	-	-	0.938	-	-	**0.960**	-	-
Name	0.375	0.613	0.771	0.602	0.698	0.830	0.832	0.820	0.907	**0.852**	**0.840**	**0.911**
License	0.948	0.502	0.813	0.953	0.635	0.778	**0.963**	**0.700**	**0.818**	0.962	0.685	**0.818**
Version	0.738	0.620	0.816	0.935	0.687	**0.865**	0.973	0.538	0.571	**0.983**	**0.755**	0.767
URL	0.723	0.330	0.352	0.968	0.495	0.527	0.973	0.538	0.571	**0.982**	**0.571**	**0.604**
Usage	0.286	-	-	0.765	-	-	0.898	-	-	**0.921**	-	-
Provenance	0.523	-	-	0.650	-	-	0.895	-	-	**0.926**	-	-

Table 4: Experimental results on the test set of the Hybrid dataset.

results, the License and Version do not have extraction scores, as such metadata were not present in those particular instances.

Interestingly, despite the fine-tuned models being built on the simpler Flan-T5 Base architecture, they significantly outperformed Flan-T5 XL model in this particular task. This suggests that a model with a relatively small number of parameters, such as the base Flan-T5 model ($220M$ parameters), can surpass a larger LLM, like the Flan-T5 XL model ($3B$ parameters), given appropriate fine-tuning. The fine-tuning procedure effectively harnesses the parameters of the base model to learn task-specific information, resulting in a model that is both precise and efficient.

7 Qualitative Analysis

In this section, we delve into an analysis of some representative examples of our models' predictions, aiming to demonstrate the impact of the LoRA fine-tuning process on the model's understanding of this specific task. We initially explore two typical instances of RA mentions, the first for a dataset and the second for a software (refer to Figs. 5 and 6, App. E), to highlight the improvements brought about by the fine-tuning of the Flat-T5 model.

Considering the dataset mention illustrated in Fig. 5 (App. E), it is evident that the correct answer is an unnamed dataset represented by "N/A". Both of our LoRA fine-tuned models yielded the correct result, in contrast to the base models. More specifically, the Flan-T5 Base model erroneously returned "100,000 reviews", while the Flan-T5 XL model inaccurately produced the general term "costumer reviews" as a dataset name. The same error is observed in the software mention depicted in Fig. 6 (App. E), where both base models failed to give correct answers. The distinction in performance can be traced back to our fine-tuning process, as it allowed our models to understand that not every name-like term in the snippet necessarily represents a specific dataset.

An examination of different scenarios reveals that the LoRA-Hy model exhibits superior performance to the LoRA-Sy model in more complex cases, like the provenance and usage question instances in Figs. 7 and 8 (App. E).

The main shortcoming of our fine-tuned models is their occasional inability to extract text-spans from the given snippet to answer questions, despite explicitly being fine-tuned for the task. This issue is demonstrated in Fig. 9 (App. E). In this instance, all models provided correct responses. However, the answers generated by the LoRA-Hy model were not exact excerpts from the original snippet text. Consequently, this deviation from the snippet text was considered an error in the evaluation process.

8 Discussion & Conclusions

In this work, we have made significant steps towards knowledge discovery from scholarly literature and RAA by advancing the identification and extraction of dataset and software mentions within scientific literature, thereby addressing pressing challenges in reproducibility and reusability of RAs.

More specifically, by leveraging the capabilities of ChatGPT in conjunction with meticulous human curation, we streamlined the extraction of dataset and software mentions. This innovative approach made it possible to transform RAA from a NER task to an instruction-based QA task. Furthermore, we investigated how LLMs could be effectively employed for this task and how the LoRA fine-tuning method can enhance such models when trained on

	Flan T5 base			Flan T5 XL			LoRA-Sy			LoRA-Hy		
	Identification	Extraction		Identification	Extraction		Identification	Extraction		Identification	Extraction	
	F1	EM	LM	F1	EM	LM	F1	EM	LM	F1	EM	LM
Valid	0.375	-	-	0.621	-	-	0.847	-	-	**0.932**	-	-
Name	0.452	0.233	0.512	0.526	0.326	0.628	0.723	0.465	0.721	**0.884**	**0.628**	**0.814**
License	0.967	-	-	**1.000**	-	-	**1.000**	-	-	**1.000**	-	-
Version	0.811	-	-	0.950	-	-	0.967	-	-	**0.976**	-	-
URL	0.832	**1.000**	**1.000**	0.956	**1.000**	**1.000**	0.983	0.500	0.500	**0.991**	**1.000**	**1.000**
Usage	0.000	-	-	0.697	-	-	0.865	-	-	**0.945**	-	-
Provenance	0.341	-	-	0.735	-	-	**0.851**	-	-	0.836	-	-

Table 5: Experimental results on the test set of the GARS dataset.

RA datasets.

Through comprehensive examples and analysis, we demonstrated the efficacy of both the RA datasets and their associated models. Our results not only confirm the feasibility of this specific approach to the RAA task but also indicate its potential as a powerful tool in future applications.

In future work, our RA datasets can be further refined and expanded by including more representative snippets drawn from a broader and more diverse assortment of scientific publications. This will facilitate the creation of new and more generalized RA datasets, helping to mitigate potential biases and incorporate knowledge from various scientific disciplines.

Limitations

In this section, we turn our attention to the limitations inherent to our work. We provide a nuanced understanding of the boundaries of our RA datasets and methods, and we identify potential areas for improvement in future work.

The RA datasets we developed for this work are confined to snippets derived from scientific publications in Computer Science. As a result, models trained on those RA datasets may struggle to effectively generalize to complex, domain-specific scenarios in other scientific fields, such as those found in Biomedical and Health Sciences, or Sociology. Additionally, the RA datasets do not make distinctions between closely associated RA types, such as materials, repositories, and datasets or software, models, and methods.

While the fine-tuned LLM models that we specifically created and tested on our RA datasets yielded commendable results in our experiments in comparison to base Flan-T5 models, an evaluation against the Flan-T5 XXL model was not possible. Such an evaluation would have provided a significant opportunity to assess their performance against an even larger model.

Despite the considerable advancements in enriching the RA dataset with real examples drawn from scientific publications, resulting in the Hybrid dataset, the representation of RA mentions from scientific publications is still relatively narrow. Consequently, some uncommon cases of RA mentions may be underrepresented or entirely absent within our RA dataset. This observation emphasizes the need for further enhancements to our RA dataset. Future work could address this by incorporating a more diverse selection of scientific publications.

Acknowledgements

This work was supported by research grants from the European Union's H2020 IntelComp Project (https://cordis.europa.eu/project/id/101004870), European Union's HE PathOS Project (https://cordis.europa.eu/project/id/101058728) and European Union's HE TIER2 Project (https://cordis.europa.eu/project/id/101094817).

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72(7):870–884.

Geraint Duck, Goran Nenadic, Andy Brass, David L Robertson, and Robert Stevens. 2013. Bionerds: Ex-

ploring bioinformatics' database and software use through literature mining. *BMC Bioinformatics*, 14(1).

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

Jenny Heddes, Pim Meerdink, Miguel Pieters, and maarten marx. 2021. The automatic detection of dataset names in scientific articles. *Data*, 6:84.

Linlin Hou, Ji Zhang, Ou Wu, Ting Yu, Zhen Wang, Zhao Li, Jianliang Gao, Yingchun Ye, and Rujing Yao. 2022. Method and dataset entity mining in scientific literature: A cnn + bilstm model with self-attention. *Knowledge-Based Systems*, 235:107621.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

James Howison, Patrice Lopez, Caifan Du, and Hannah Cohoon. 2023. Softcite dataset version 2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Ana-Maria Istrate, Donghui Li, Dario Taraborelli, Michaela Torkar, Boris Veytsman, and Ivana Williams. 2022. A large dataset of software mentions in the biomedical literature.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Frank Krüger and David Schindler. 2020. A literature review on methods for the extraction of usage statements of software and data. *Computing in Science & Engineering*, 22:26–38.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Huitong Pan, Qi Zhang, Eduard Constantin Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. Dmdd: A large-scale dataset for dataset mentions detection. *ArXiv*, abs/2305.11779.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. Somesci- a 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4574–4583, New York, NY, USA. Association for Computing Machinery.

David Schindler, Benjamin Zapilko, and Frank Krüger. 2020. Investigating software usage in the social sciences: A knowledge graph approach. *The Semantic Web*, 12123:271 – 286.

Lysandre Debut Younes Belkada Sayak Paul Sourab Mangrulkar, Sylvain Gugger. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Yuzhuo Wang, Chengzhi Zhang, and Kai Li. 2022. A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*, 127:2479 – 2520.

Tong Zeng and Daniel Ernesto Acuna. 2020. Finding datasets in publications: the syracuse university approach.

He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5206–5215, Hong Kong, China. Association for Computational Linguistics.

A ChatGPT Prompt

During the Synthetic dataset creation phase, the following ChatGPT prompt was employed (Tab. 6), initiating an iterative and detailed annotation process. Starting with a small set of carefully curated examples, we provided the prompt to ChatGPT, which generated new synthetic RA mentions with characteristics similar to the examples. We meticulously selected certain generated instances that exhibited specific properties, contributing to the richness and diversity of the dataset.

Furthermore, a key part of the process involved closely observing and collecting keywords and keyphrases that ChatGPT associated with RA mentions. Gradually, we created a set of keywords that acted as triggers for RA mentions, along with synthetic names for RA mentions.

When we gathered a sufficient collection of snippets, we leveraged the manually curated set of keywords and keyphrases to perform exhaustive annotation of the snippets. The keywords and keyphrases were used as triggers for RA mentions, including both named and unnamed RAs. That approach ensured a balanced representation of RAs and prevented bias towards specific RA mentions.

You are DataCreatorGPT. Your task is to generate snippets that contain structured information about research artifacts extracted from scientific publications. Each snippet includes a candidate research artifact highlighted by <m>and </m>tags. For each publication snippet, you need to create the following metadata: Artifact Type: Identify the type of research artifact specified within the <m>and </m>tags. The artifact could be a dataset, software, method, etc. Valid Artifact: Determine if the artifact within the <m>and </m>tags is a valid research artifact. A valid artifact is a tangible input or output of the research publication. If the artifact is a general reference or functions as an adjective (for instance, "data" in "data analysis tool"), it is considered invalid. Name Extraction: Extract the name of the research artifact from the snippet. If no name is provided, mark it as "N/A". Version Extraction: Extract the version of the research artifact from the snippet. If no version is mentioned, mark it as "N/A". License Extraction: Extract the license of the research artifact from the snippet. If no license is indicated, mark it as "N/A". URL Extraction: Extract the URL of the research artifact from the snippet. If no URL is provided, mark it as "N/A". Provenance Classification: Determine whether the authors of the publication have created, generated, or introduced the research artifact. This determination should be clearly evident from the snippet. The response should be "Yes" or "No". Usage Classification: Determine whether the authors of the publication have used, implemented, utilized or compared/benchmarked the research artifact. This determination should be clearly evident from the snippet. The response should be "Yes" or "No".

Table 6: Prompt for ChatGPT used for the initial data creation and the human-in-the-loop process for the Synthetic dataset.

B RA Dataset Statistics

The details of the Synthetic and Hybrid datasets are summarized in Tabs. 7 and 8 respectively. Both datasets began with a specific number of unique snippets, each containing multiple mentions of datasets and software (RA mentions). These mentions may be accompanied by particular metadata (such as Valid, Name, Version, License, URL, Provenance, and Usage) relevant to the RA in question. Corresponding question-answer (QA) pairs are formulated for each RA mention and metadata field, utilizing Tab. 9 (App. C). The resulting collection of QA pairs provides a basis for fine-tuning a Large Language Model (LLM) using the LoRA method.

	Original									Augmented								
	Train			Dev			Test			Train			Dev			Test		
	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all
RA mentions	554	647	1201	98	123	221	89	105	194	1981	2247	4228	292	335	627	282	309	591
valid	476	584	1060	87	107	194	69	98	167	1694	2022	3716	258	287	545	211	295	506
w. name	401	468	869	78	90	168	58	82	140	1422	1614	3036	226	237	463	171	243	414
w. version	42	235	277	11	61	72	0	57	57	122	762	884	33	151	184	0	178	178
w. license	142	192	334	38	46	84	20	47	67	519	616	1135	119	128	247	79	139	218
w. URL	224	171	395	38	38	76	16	20	36	764	593	1357	95	60	155	28	48	76
w. provenance	158	142	300	35	10	45	29	28	57	586	499	1085	118	30	148	115	81	196
w. usage	296	469	765	57	88	145	38	74	112	1016	1631	2647	160	222	382	88	241	329
Unique snippets	148	176	240	25	25	32	25	24	33	1589	1796	3298	232	258	474	230	247	463
Special QA pairs	-	-	-	-	-	-	-	-	-	489	616	1059	64	71	124	64	84	140
All QA pairs	3419	4193	7612	620	765	1385	509	706	1215	12147	14432	27639	1840	2057	4021	1572	2103	3815

Table 7: Statistics for the Synthetic dataset.

	Original									Augmented								
	Train			Dev			Test			Train			Dev			Test		
	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all	dataset	software	all
RA mentions	757	1126	1883	128	222	350	125	181	306	2332	3125	5457	331	507	838	354	463	817
valid	615	951	1566	108	189	297	93	149	242	1958	2712	4670	286	439	725	258	403	661
w. name	488	769	1257	88	152	240	75	120	195	1592	2199	3791	238	352	590	194	329	523
w. version	42	235	277	11	61	72	0	57	57	122	762	884	33	151	184	0	178	178
w. license	142	201	343	38	55	93	20	47	67	519	633	1152	119	131	250	79	139	218
w. URL	225	173	398	38	38	76	16	24	40	767	601	1368	95	60	155	28	63	91
w. provenance	175	235	410	36	39	75	33	53	86	620	673	1293	119	75	194	131	138	269
w. usage	427	770	1197	77	158	235	60	115	175	1262	2208	3470	186	344	530	130	332	462
Unique snippets	194	230	298	32	34	41	32	34	43	1815	2337	4027	257	369	605	278	341	598
Special QA pairs	-	-	-	-	-	-	-	-	-	575	773	1267	73	90	147	72	106	162
All QA pairs	4456	6882	11338	776	1356	2132	689	1088	1777	14082	19458	34808	2047	3141	5335	1926	2905	4993

Table 8: Statistics for the Hybrid dataset.

C Prompts for instruction-based QA

The transformation of the RAA task from RA mentions to an instruction-based QA task, characterized by QA pairs, was achieved through the utilization of specific questions. These questions, as presented in Tab. 9, were used in the training and testing of our LoRA fine-tuned models. To ensure a fair comparison with the base Flan-T5 models during evaluation on our RA datasets' test sets, necessary modifications were made to these questions, as detailed in Tab. 10. The subsequent tables provide insight into this transformation process, illustrating how each metadata field is restructured into a question, such that the answer to that question, based on the RA mention's snippet, corresponds to the metadata field value in the RA mention.

In addition to the standard metadata-related questions, a "special" type of QA pair question, which is not associated with a specific metadata field, is also included. That "special" type plays a crucial role in the conversion of unique snippets into "special" QA pairs, which enumerate all the RAs in the snippet, denoting their Type and Name.

Metadata Field	Question
Valid	Is there a valid [software/dataset] defined in the <m> and </m> tags?
Name	What is the name of the [software/dataset] defined in the <m> and </m> tags?
Version	What is the version of the [software/dataset] defined in the <m> and </m> tags?
License	What is the license of the [software/dataset] defined in the <m> and </m> tags?
URL	What is the URL of the [software/dataset] defined in the <m> and </m> tags?
Provenance	Is the [software/dataset] defined in the <m> and </m> tags introduced or created by the authors of the publication in the snippet above?
Usage	Is the [software/dataset] defined in the <m> and </m> tags used or adopted by the authors of the publication in the snippet above?
Special QA pairs	List all the artifacts in the above snippet.

Table 9: Questions to convert the RA mentions to QA pairs.

Metadata Field	Question
Valid	Is there a valid dataset|software defined in the <m> and </m> tags? Answer only using "Yes" or "No".
Name	What is the name of the [dataset/software] defined in the <m> and </m> tags? The answer must be a text span from the Snippet. If you can't answer the question then respond with "N/A".
Version	What is the version of the [dataset/software] defined in the <m> and </m> tags? The answer must be a text span from the Snippet. If you can't answer the question then respond with "N/A".
License	What is the license of the [dataset/software] defined in the <m> and </m> tags? The answer must be a text span from the Snippet. If you can't answer the question then respond with "N/A".
URL	What is the URL of the [dataset/software] defined in the <m> and </m> tags? The answer must be a text span from the Snippet. If you can't answer the question then respond with "N/A".
Provenance	Is the [dataset/software] defined in the <m> and </m> tags introduced or created by the authors of the publication in the snippet above? Answer only using "Yes" or "No".
Usage	Is the [dataset/software] defined in the <m> and </m> tags used or adopted by the authors of the publication in the snippet above? Answer only using "Yes" or "No".
Special QA pairs	List all artifacts in the above snippet. Answer with a list of artifacts in the format "artifact_type: artifact_name" separated by "|" tokens.

Table 10: Modified questions to convert the RA mentions to QA pairs.

D Evaluation results of named vs unnamed RA Mentions

In this Appendix, we present a detailed overview of the evaluation results for all models across the Synthetic, Hybrid, and GARS test sets, categorizing them by named and unnamed RA mentions. The unnamed RA Mentions do not have "Extraction" scores for the "Name" since there is no name to predict. Similarly, in the GARS test set tables, specific metrics pertaining to Licence, Version, and URL extraction do not have a score. This indicates that there is no such metadata in those particular instances.

	Flan T5 base			**Flan T5 XL**			**LoRA-Sy**			**LoRA-Hy**		
	Identification	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**	
	F1	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**
Valid	0.918	-	-	0.932	-	-	0.994	-	-	0.995	-	-
Name	0.987	0.709	0.835	0.972	0.787	0.900	0.990	0.917	0.962	0.989	0.905	0.952
License	0.947	0.502	0.813	0.944	0.635	0.778	0.958	0.700	0.818	0.961	0.685	0.818
Version	0.688	0.544	0.779	0.944	0.625	0.838	0.985	0.581	0.588	0.989	0.735	0.750
URL	0.617	0.385	0.400	0.980	0.477	0.492	0.984	0.569	0.600	0.983	0.585	0.615
Usage	0.429	-	-	0.811	-	-	0.910	-	-	0.919	-	-
Provenance	0.484	-	-	0.649	-	-	0.941	-	-	0.958	-	-

Table 11: Experimental results on the named RA mentions of the Synthetic test set.

	Flan T5 base			**Flan T5 XL**			**LoRA-Sy**			**LoRA-Hy**		
	Identification	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**	
	F1	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**
Valid	0.534	-	-	0.728	-	-	0.915	-	-	0.933	-	-
Name	0.387	-	-	0.773	-	-	0.925	-	-	0.919	-	-
License	0.883	-	-	0.895	-	-	0.919	-	-	0.907	-	-
Version	0.636	1.000	1.000	0.934	1.000	1.000	0.939	0.815	0.815	0.945	0.852	0.852
URL	0.849	0.091	0.091	0.977	0.909	1.000	0.971	0.909	0.909	0.977	0.909	0.909
Usage	0.154	-	-	0.597	-	-	0.915	-	-	0.899	-	-
Provenance	0.713	-	-	0.644	-	-	0.937	-	-	0.966	-	-

Table 12: Experimental results on the unnamed RA mentions of the Synthetic test set.

	Flan T5 base			**Flan T5 XL**			**LoRA-Sy**			**LoRA-Hy**		
	Identification	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**	
	F1	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**
Valid	0.827	-	-	0.879	-	-	0.993	-	-	0.989	-	-
Name	0.977	0.613	0.771	0.949	0.698	0.830	0.968	0.820	0.907	0.975	0.840	0.911
License	0.959	0.502	0.813	0.964	0.635	0.778	0.972	0.700	0.818	0.973	0.685	0.818
Version	0.760	0.544	0.779	0.937	0.625	0.838	0.977	0.581	0.588	0.988	0.735	0.750
URL	0.693	0.362	0.388	0.983	0.438	0.463	0.979	0.487	0.525	0.985	0.525	0.562
Usage	0.340	-	-	0.799	-	-	0.883	-	-	0.917	-	-
Provenance	0.454	-	-	0.700	-	-	0.914	-	-	0.947	-	-

Table 13: Experimental results on the named RA mentions of the Hybrid dataset test set.

	Flan T5 base			**Flan T5 XL**			**LoRA-Sy**			**LoRA-Hy**		
	Identification	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**	
	F1	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**
Valid	0.580	-	-	0.721	-	-	0.932	-	-	0.948	-	-
Name	0.422	-	-	0.730	-	-	0.926	-	-	0.930	-	-
License	0.923	-	-	0.930	-	-	0.946	-	-	0.938	-	-
Version	0.663	1.000	1.000	0.928	1.000	1.000	0.962	0.815	0.815	0.966	0.852	0.852
URL	0.807	0.091	0.091	0.916	0.909	1.000	0.977	0.909	0.909	0.974	0.909	0.909
Usage	0.099	-	-	0.637	-	-	0.943	-	-	0.933	-	-
Provenance	0.720	-	-	0.554	-	-	0.860	-	-	0.889	-	-

Table 14: Experimental results on the unnamed RA mentions of the Hybrid dataset test set.

	Flan T5 base			**Flan T5 XL**			**LoRA-Sy**			**LoRA-Hy**		
	Identification	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**		**Identification**	**Extraction**	
	F1	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**	**F1**	**EM**	**LM**
Valid	0.333	-	-	0.696	-	-	0.989	-	-	0.989	-	-
Name	0.951	0.233	0.512	0.897	0.326	0.628	0.868	0.465	0.721	0.951	0.628	0.814
License	0.951	-	-	1.000	-	-	1.000	-	-	1.000	-	-
Version	0.822	-	-	0.938	-	-	0.951	-	-	0.964	-	-
URL	0.836	1.000	1.000	1.000	1.000	1.000	0.975	0.500	0.500	1.000	1.000	1.000
Usage	0.000	-	-	0.738	-	-	0.812	-	-	0.933	-	-
Provenance	0.222	-	-	0.947	-	-	0.941	-	-	0.947	-	-

Table 15: Experimental results on the named RA mentions of the GARS test set.

	Flan T5 base			Flan T5 XL			LoRA-Sy			LoRA-Hy		
	Identification	Extraction		Identification	Extraction		Identification	Extraction		Identification	Extraction	
	F1	EM	LM	F1	EM	LM	F1	EM	LM	F1	EM	LM
Valid	0.462	-	-	0.571	-	-	0.919	-	-	0.947	-	-
Name	0.519	-	-	0.667	-	-	0.919	-	-	0.974	-	-
License	1.000	-	-	1.000	-	-	1.000	-	-	1.000	-	-
Version	0.788	-	-	0.974	-	-	1.000	-	-	1.000	-	-
URL	0.824	-	-	0.857	-	-	1.000	-	-	0.974	-	-
Usage	0.000	-	-	0.583	-	-	0.971	-	-	0.971	-	-
Provenance	0.571	-	-	0.000	-	-	0.615	-	-	0.588	-	-

Table 16: Experimental results on the unnamed RA mentions of the GARS test set.

E Qualitative Analysis Examples

In this Appendix, we present a collection of QA pairs, along with the results produced by both the LoRA fine-tuned and Flan-T5 base models. The QA pairs have been thoughtfully selected to support the qualitative analysis elaborated in Sec. 7.

Snippet	We manually collected a remarkable dataset consisting of <m>customer reviews</m> from various e-commerce websites. This extensive dataset comprises 100,000 reviews encompassing diverse product categories. Researchers can request access to this dataset by sending an email to alex@abc.com.
Question	What is the name of the dataset defined in the <m> and </m> tags?
Answer	"N/A"
Flan-T5 Base	"100,000 reviews"
Flan-T5 XL	"customer reviews"
LoRA-Sy	"N/A"
LoRA-Hy	"N/A"

Figure 5: Example of a dataset type RA mention, where the LoRA fine-tuned models outperform the original Flan-T5 models.

Snippet	In their study, the authors incorporated various research artifacts. They employed the Apache Spark (v3.1.2) distributed computing framework and the hadoop (v3.3.1) big data processing <m>platform</m>. Apache Spark, licensed under the Apache 2.0 license, allowed efficient processing and analysis of large-scale datasets. hadoop, also released under the Apache 2.0 license, provided a robust infrastructure for distributed storage and processing. Those artifacts were instrumental in handling and analyzing massive amounts of data in their research.
Question	What is the name of the software defined in the <m> and </m> tags?
Answer	"hadoop"
Flan-T5 Base	"Apache Spark"
Flan-T5 XL	"Apache Spark (v3.1.2) distributed computing framework and the hadoop (v3.3.1) big data processing m>platform/m>"
LoRA-Sy	"hadoop"
LoRA-Hy	"hadoop"

Figure 6: Example of a software type RA mention, where the LoRA fine-tuned models outperform the original Flan-T5 models.

Snippet	The first comprehensive pre-trained multilingual model for phoneme representations, also known as <m>XPhoneBERT</m>, is presented. On the downstream TTS task, XPhoneBERT significantly enhances the strong baseline vits and confirms its effectiveness.
Question	Is the software defined in the <m> and </m> tags introduced or created by the authors of the publication in the snippet above?
Answer	"Yes"
Flan-T5 Base	"No"
Flan-T5 XL	"Yes"
LoRA-Sy	"No"
LoRA-Hy	"Yes"

Figure 7: Example of LoRA-Hy model's success in a provenance QA pair.

Snippet	To overcome the issue of accuracy, we present the <m>Sparse-Quantized Representation (SpQR)</m> approach, a new quantization and compressed format technique that delivers near-lossless compression of LLM models across model scales, while maintaining similar levels of compression as previous techniques.
Question	Is the software defined in the <m> and </m> tags used or adopted by the authors of the publication in the snippet above?
Answer	"Yes"
Flan-T5 Base	"No"
Flan-T5 XL	"No"
LoRA-Sy	"No"
LoRA-Hy	"Yes"

Figure 8: Example of LoRA-Hy model's success in a usage QA pair.

Snippet	To train <m>HeadlineSense</m>, our news headline classification model, we used the News Headlines Dataset, which consists of headlines from news articles. The dataset is widely used for text classification tasks. It is released under the Open Data Commons Attribution License (ODC-BY).
Question	What is the name of the software defined in the <m> and </m> tags?
Answer	"HeadlineSense"
Flan-T5 Base	"HeadlineSense"
Flan-T5 XL	"HeadlineSense"
LoRA-Sy	"HeadlineSense"
LoRA-Hy	"headline sense"

Figure 9: Example of a correct prediction of the LoRA-Hy model that was not a text-span from the snippet.