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

Contact: *petros.stavropoulos@athenarc.gr*

Paper

Github

Introduction - Background

Research Artifact Analysis (RAA) is the systematic identification, extraction, and examination of tangible **research artifacts (RAs)**, such as datasets, software and methodologies from scientific literature.

Why is RAA important?

Transparency: Clarifies research methods and resources for improved scrutiny.

Reproducibility: Enables study replication to confirm findings' reliability.

Acceleration of Innovation: Identifies and shares data and tools to advance research.

Resource Optimization: Prevents effort duplication, enhancing resource use.

Traditional RAA relied heavily on **Named Entity Recognition (NER)**.

Limitations of **NER-based RAA methods**:

- Overlook unnamed or undocumented resources.
- Fail to identify important RAs with non-standard naming.
- Prefer well-documented RAs, sidelining lesser-known artifacts.

In this work we move beyond NER to a comprehensive approach that captures both **named** and **unnamed RAs**.

Our key contributions:

1. We developed **two unique RAA datasets** with synthetic and real mentions, addressing gaps in existing literature.
2. We showed that **small, fine-tuned LLMs perform excellently on RAA**, outperforming even their larger base counterparts.
3. We performed a thorough **qualitative assessment** of the new RA datasets and models.

	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.

Instruction-based Question Answering Task

The task is to identify, extract, and analyze mentions of research artifacts (RAs) within snippets of scientific literature.

The two types of RA considered in this work are **datasets** and **code/software**.

Valid RA mentions are direct or indirect references to datasets or code/software within a scientific text, possibly accompanied by metadata: **Name**, **Version**, **License**, **URL**, **Usage**, **Provenance.**

Invalid RA mentions lack this contextual information and do not explicitly point to a dataset or software, or they may be general terms not used to indicate a specific resource.

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.

Each **key-value** from the RA mention is converted to **Question Answering pairs (QA pairs)**.

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.

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

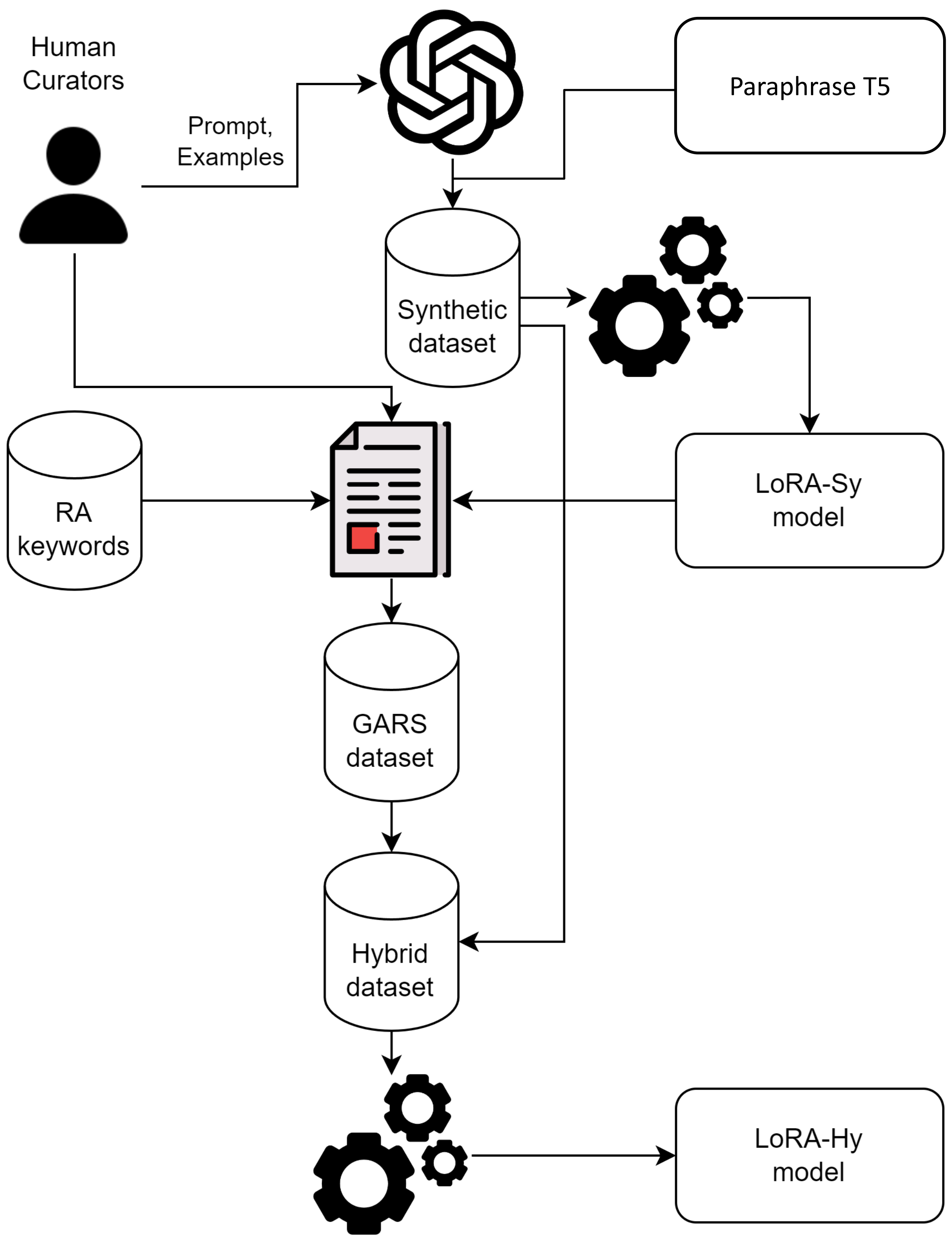
Figure 3: An example of QA pair.

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.

Datasets & Models

- Human curators generate initial data using a **ChatGPT prompt** and **RA mention examples**
- A **Synthetic dataset** is created, leveraging a **paraphrase T5 model** for **data augmentation**
- A **Flan-T5 Base** model is fine-tuned on the Synthetic dataset using **Low-Rank Adaptation (LoRA)**, resulting in the **LoRA-Sy** model
- The **GARS dataset** was developed with **human-curated RA mentions from scientific texts** to balance the Synthetic dataset biases and improve model accuracy
- A **Hybrid dataset** is formed by merging the Synthetic and GARS datasets to improve diversity and model robustness
- A **Flan-T5 Base** model is fine-tuned on the **Hybrid dataset** using **LoRA**, resulting in the **LoRA-Hy model**, which performs better due to the enriched training data



	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.

Results

	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.

EMNLP 2023 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS)
6 Dec 2023 @ EMNLP 2023 in Singapore