

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

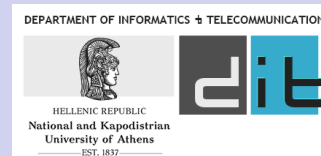
Petros Stavropoulos^{1,2}, Ioannis Lyris¹, Natalia Manola³, Ioanna Grypari^{1,3}, Haris Papageorgiou¹

¹Institute for Language and Speech Processing, Athena R.C.

²Department of Informatics and Telecommunications, National and Kapodistrian University of Athens

³OpenAIRE AMKE

Contact: petros-stavropoulos@athenarc-gr



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.

- **Importance of RAA:** Crucial for enhancing transparency, reproducibility, and innovation in scientific research.
- **Limitations of Traditional Methods:** Current methods like Named Entity Recognition fall short in comprehensively capturing research artifacts.
- **Our Novel Approach:** Expanding beyond traditional methods to include both named and unnamed research artifacts in analysis.

Our key contributions:

1. We developed **two unique RAA datasets** with synthetic and real RA 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.

	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.

