Large Language Models for Scientific Information Extraction: An Empirical Study for Virology

Mahsa Shamsabadi and **Jennifer D'Souza** and **Sören Auer**
TIB Leibniz Information Centre for Science and Technology,
Hannover, Germany
{mahsa.shamsabasdi,jennifer.dsouza,auer}@tib.eu

Abstract

In this paper, we champion the use of structured and semantic content representation of discourse-based scholarly communication, inspired by tools like Wikipedia infoboxes or structured Amazon product descriptions. These representations provide users with a concise overview, aiding scientists in navigating the dense academic landscape. Our novel automated approach leverages the robust text generation capabilities of LLMs to produce structured scholarly contribution summaries, offering both a practical solution and insights into LLMs' emergent abilities.

For LLMs, the prime focus is on improving their general intelligence as conversational agents. We argue that these models can also be applied effectively in information extraction (IE), specifically in complex IE tasks within terse domains like Science. This paradigm shift replaces the traditional modular, pipelined machine learning approach with a simpler objective expressed through instructions. Our results show that finetuned FLAN-T5 with 1000x fewer parameters than the state-of-the-art GPT-davinci is competitive for the task.

1 Introduction

Scholarly communication in the digital age is facing significant challenges due to the overwhelming volume of publications (Johnson et al., 2018) thereby creating the need for efficient access to relevant knowledge. In this regard, next-generation scholarly digital libraries, such as the Open Research Knowledge Graph (ORKG) (Auer et al., 2020; Stocker et al., 2023), offer a promising solution by adopting semantic publishing principles (Shotton, 2009). The ORKG stores *scholarly contributions* in a structured and semantic way, leveraging a knowledge graph (KG) representation (Ehrlinger and Wöß, 2016; Fensel et al., 2020). The fine-grained semantic contribution representation in the ORKG utilizes property-value tuples, capturing important aspects and corresponding observations of research contributions. This representation enhances understanding and navigation of scholarly content by both humans and machines. With selected properties that apply universally to research on a specific problem, the ORKG enables intelligent exploration and assistance services, including research comparisons based on shared properties, e.g., Figure 1. Its novel information access methods provide condensed overviews of the state-of-the-art, supporting strategic reading (Renear and Palmer, 2009) in the ever-growing publication landscape.

Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy Contribution 1 - 2020	Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia Contribution 1 - 2020
Has research problem	COVID-19 reproductive number	COVID-19 reproductive number
Location	Lombardy, Italy	China
Study date	2020-02-20	2020-01-22
R0 estimates (average)	3.1	2.2
95% confidence interval	2.9-3.2	1.4-3.9

Figure 1: Two structured research contributions compared in the Open Research Knowledge Graph (papers in columns, properties in rows and values in cells).

This work, as a text mining service toward producing scalable solutions for the ORKG, for the first time, introduces a complex information extraction (IE) task. Our notion of complex IE entails joint entity and relation extraction in a single objective aligned with the structured property-value format of contributions in the ORKG. We defined the complex IE task w.r.t. a key research problem in the domain of Epidemics & Virology, i.e. estimating the basic reproduction number ($R0$) for infectious diseases. This $R0$ estimate research topic was brought to common knowledge during the recent Covid-19 pandemic by the Centers for Disease Control and Prevention (CDC) as a key informant. Important to infectious disease epidemiology, gen-

Findings of the Association for Computational Linguistics: EACL 2024, pages 374–392
March 17-22, 2024

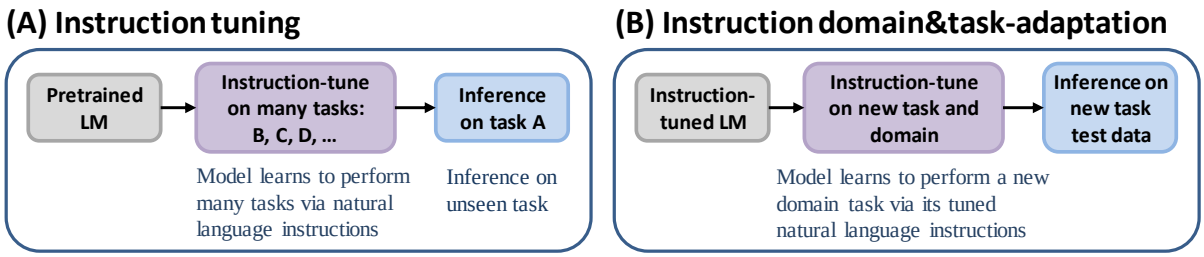

Figure 2: Comparing (A) instruction tuning with (B) instruction-tuned LLM domain- and task-tuning of this work.

erally, the $R0$ estimate represents the average number of secondary infections caused by a single infected individual (Foppa, 2017). In other words, it is an estimate of disease progression in a given population. E.g., the estimated R0 for COVID-19 has been reported between 2.5 to 5.7 (Sanche et al., 2020). It varies for different infectious diseases and populations. For researchers in Epidemics & Virology, it is interesting to be able to compare the R0 of different viruses facilitated by structured contribution data available in the ORKG. The alternative, traditional, and seemingly impossible knowledge comprehension task, would be to scour for vital information buried in unstructured text across the 44k articles by Covid-19 R0 estimate Google search.

To define our complex IE task, an expert semantic modeler created a research comparison based on structured property-value pairs for Covid-19 $R0$ estimate contributions across 30 abstracts. Consequently, six properties were modeled: *disease name*, *location*, *date*, *R0 value*, *%CI values*,[1] and *method*. The semantic modeling aimed to identify properties that were both generic enough to structure most related research on the $R0$ estimate (in the context of a research comparison) and specialized enough to reflect the vital details of the $R0$ contribution (by identifying commonalities in observations reported across 30 different abstracts). This structured format is called ORKG-R0. Thus our complex IE task focused on extracting property-value pairs for ORKG-R0 contributions in scholarly article abstracts. To address this task, a larger gold-standard corpus was annotated (details in section 3) and an LLM-based solution was optimally designed (introduced next, details in section 4).

The complex IE task introduced earlier is addressed as single-task instruction-based finetuning of an instruction-tuned Large Language Model (LLM) with the primary objective of *better aligning the LLM to our task and domain*. Our approach is characterized in Figure 2. We chose LLMs for their rich parameter spaces and ability to handle complex IE tasks with simple instruction prompts (Ouyang et al., 2022). Unlike traditional pipelined-based IE, which are prone to error propagation and require extensive manual engineering, LLMs offer flexibility, adaptability, and the ability to handle a wide range of tasks in zero- and few-shot settings through instructions (Radford et al.; Brown et al., 2020; Wei et al., 2021). By relying on instruction prompting, we can effectively address complex inter-relations without the need for an exhaustive enumeration of all possible relations or preliminary named entity recognition (NER). We finetune an LLM from the sequence-to-sequence encoder-decoder-based T5 model class (Raffel et al., 2020) to accept a research paper title and abstract and instruct it to write the ORKG-R0 structured "summary" of knowledge in the prompt as either text-based or as a structured JSON object. For the LLM, we specifically select the instruction-tuned FLAN-T5-Large model (Chung et al., 2022) with reported 780M parameters. There could have been one of two directions for this work: scaling the models or instruction fine-tuning of a moderate-sized LLM, i.e. with parameters in millions versus 1000x more in billions. We chose the latter. We believe that our choice makes model tuning more accessible within the research community while empirically proving to be nonetheless effective (experimental details in section 5). Furthermore, our choice of Google's FLAN-T5, open-sourced and easily accessible in the Transformers library, obviates any paywall that hinders access to LLMs for the research community at large. For instruction-based finetuning, we use applicable instructions from the *open-sourced instruction generalization efforts* introduced as the "Flan 2022 Collection" (Longpre et al., 2023). Our approach differs from finetuning a pretrained LM as we instead finetune an instruction-tuned LM, enabling the model to effectively follow instructions

[1] CI stands for confidence interval.

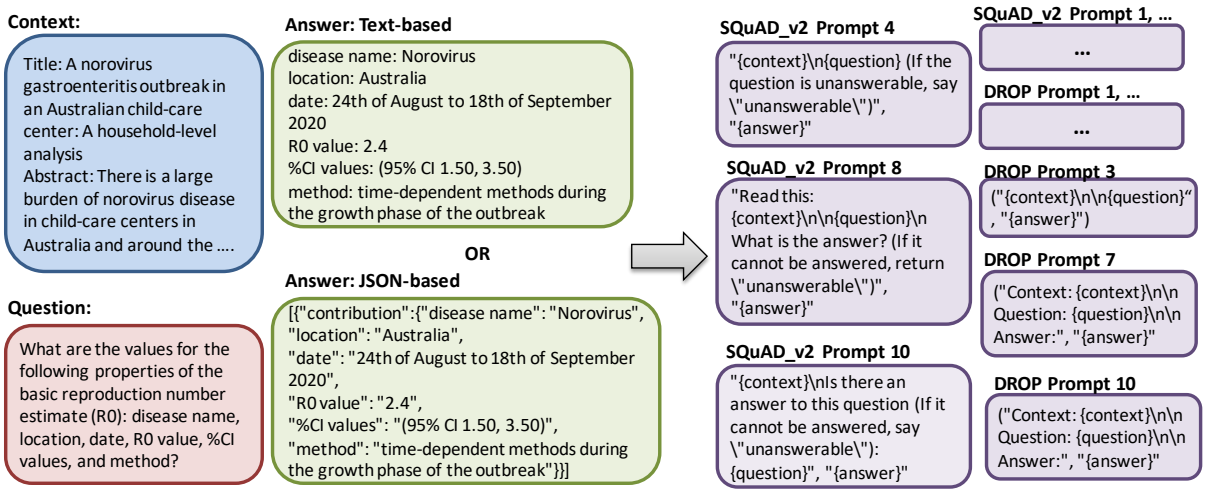

Figure 3: Multiple instruction prompts describing our complex scientific information extraction (IE) task.

it has been trained on and adapt to a new domain and complex IE task, without the need to handle variability in learning new instruction formats. Our approach is shown in Figure 3.

In this context, the central research question (**RQ**) of this work examines: *How does instruction-based finetuning enhance LLM performance in a unique domain, specifically in a complex scientific field like Virology that requires specialized expertise?* Summarily, the main contributions of our work are as follows: 1) **Corpus**: A gold-standard corpus of 1,500 annotated structured abstracts based on ORKG-R0. 2) **Methodological**: We adopt "single-task instruction-finetuning" to enhance LLMs' domain and task adaptation. It involves selecting instructions from the open-sourced FLAN collection and fine-tuning FLAN-T5 780M to respond to those instructions. Our source code is released. 3) **Methodological**: Our approach distinguishes itself in the realm of IE research by introducing an LLM-based approach that breaks away from traditional pipeline-based methods for entity and relation extraction. Instead, we propose a single-system approach utilizing a moderately-sized LLM, which holds potential for practical applications. And 4) **Results**: Our instruction-finetuned ORKG-FLAN-T5$_{R0}$ 780M outperforms pretrained T5, instruction-tuned FLAN-T5, and GPT3.5-davinci 175B on ORKG-R0 complex IE. The best model is released on HuggingFace.

2 Background: Scholarly Communication

Semantic scholarly knowledge publishing models, such as the ORKG, specifically the ORKG-R0 instance in this work, and the structured abstracts methodology (e.g., IMRAD) employed by publishers like PubMed have distinct approaches and serve different purposes in scholarly communication. This section distinguishes the two.

The ORKG (Auer, 2018) and similar semantic knowledge publishing models (Baas et al., 2020; Birkle et al., 2020; Wang et al., 2020a; Aryani et al., 2018; Manghi et al., 2019; Hendricks et al., 2020; Fricke, 2018) aim to create interconnected and machine-actionable representations of scholarly knowledge. They leverage semantic technologies, knowledge graphs (KGs), and ontologies to capture the meaning, context, and relationships between research concepts. The ORKG, for example, stores scholarly contributions as structured property-value pairs, enabling advanced exploration, comparison (Oelen et al., 2019), and analysis via visualizations (Wiens et al., 2020) of research findings. The strength of semantic knowledge publishing models lies in their ability to facilitate interdisciplinary collaborations, data integration, and automated processing of scholarly information. They enhance research transparency, enable advanced search and discovery, and support the development of novel strategic reading tools and services for researchers.

On the other hand, the structured abstracts methodology (Haynes et al., 1990; Hayward et al., 1993; Nakayama et al., 2005; Kulkarni, 1996; Hopewell et al., 2008), e.g., IMRAD (Sollaci and Pereira, 2004), focuses on organizing research articles into a specific format. IMRAD advocates for a structured abstract based on four points, viz. Introduction, Methods, Results, and Discussion, to

provide a standardized framework for reporting research. The strength of structured abstracts lies in their ability to provide a clear and consistent organization of research findings. They help readers quickly understand the key components of a study and locate specific information within the article. Structured abstracts facilitate efficient scanning and information retrieval.

In summary, semantic scholarly knowledge publishing models enhance the machine-actionability and interoperability of scholarly knowledge, enabling advanced computational exploration and analysis. They offer opportunities for interdisciplinary collaborations and innovative research tools. On the other hand, structured abstracts provide a standardized format for reporting research, facilitating efficient information retrieval.

3 Corpus

We aim to create a high-quality corpus for the complex scientific IE task introduced in this work. The corpus creation goal was to obtain gold-standard property-value structured format representation w.r.t. the six predicates in ORKG-R0 from scholarly article abstracts. These structured representations encapsulate the R0 estimate research problem for infectious diseases.

Base corpus. Our starting point was the large-scale CORD-19 dataset (Wang et al.) provided by AllenAI. This resource comprised a growing collection of publications and preprints on Covid-19, its variants, related historical coronaviruses such as SARS and MERS, as well as other infectious diseases such as H1N1 Influenza, Dengue, Monkeypox, Ebola, Zika virus, Norovirus, etc. At our download date timestamp 2022-06-02 it comprised over 800,000 total publications. The dataset covered diverse topics such as epidemiology, virology, clinical studies, public health, and more. It served as a valuable resource for researchers, policymakers, and the general public to access and analyze the latest scientific knowledge related to COVID-19. Since CORD-19 contained articles on various themes, as a next step the corpus was filtered to include only articles on the $R0$ estimate theme.

Corpus filtering. Our method for filtering the base corpus to our desired collection was simple. We implemented pattern-based heuristics using variants of the phrase "R0 estimate" and checked the publication abstract for containment. The base corpus was then reduced to 4590 instances. Post deduplication, the collection was further reduced to 3967 instances. Other than exact duplicates, there were other near-duplication patterns such as punctuation marks stripped or retained, numbers with or without decimal points served as different data instances. Near-duplicates were also filtered by clustering abstracts that were 95% similar (583 clusters from 1227 articles were created). A human annotator went through all clusters and decided on one abstract to retain while dropping all others. The resulting curated corpus contained 3024 abstracts which included a direct mention or a variant of the phrase "R0 estimate".

The ORKG-R0 model. Here we provide an explanation of ORKG-R0 as an ideal representation of a structured contribution for the research problem of "$R0$ estimate," as defined by an expert semantic modeler. The $R0$ estimate pertains to an infectious disease (*disease name*), for a specific population demographic (*location*), with validity for a specific time period (*date*). It reports a specific value ($R0$ *value*), along with a confidence interval for the statistical value (*%CI values*), and is computed by a statistical method (*method*).

Annotation exercise. To ensure a practical and realistic human annotation target, we selected a sub-sample of 1500 articles from the curated 3024 dataset. This would then serve as the gold-standard dataset for training and development purposes, as an empirical basis for future research. A team of two annotators produced the ORKG-R0 structured annotations with the corpus raw data comprising a paper title and abstract, where each instance is uniquely identified by a cord_id. The overall annotation exercise lasted 3 months. The annotation task began by distinguishing between the papers actually reporting an $R0$ value as a contribution and those that just mentioned the "R0 estimate" keyword in the abstract, but did not actually report a value as a contribution of the work. Resultingly, we found 652 articles reported an $R0$ value and thus were annotated for the ORKG-R0 structure (referred to as the "answerable" set, in short *ans*), while 850 did not (referred to as the "unanswerable" set, in short *unans*). Among the 652 articles, approximately 157 had multiple contributions for the "$R0$ estimate." Notably, a few articles stood out with 10, 11, or 16 reported contributions. The gold-standard annotated set was made available in two formats: text-based and JSON-based, which are illustrated by the green boxes in Figure 3. In the text-based format, multiple contributions were

separated using a pipe character, while in the JSON format, they were encoded as separate JSON object dictionaries. We observed that the JSON data structure is more conducive for utilization in downstream applications. Therefore, our empirical analysis regarding LLMs aimed not only to assess their ability to generate structured abstract summaries but also to evaluate their compatibility with a specific data structure. This allows for the seamless integration of their output into downstream applications.

The annotators. In our annotation process, we first developed a structured summary model for the "R0 estimate for infectious diseases" using both domain experts and a semantic modeler specializing in ontology design. Next, a PhD student populated the model using a dataset of abstracts, treating it as a form-filling task of reported facts. While the task itself is tedious in that the student needed to read all abstracts to populate the properties, the process did not entail much ambiguity in the decisions. The definition of the properties we selected are fairly straightforward and the values are to be directly extracted from the text. For discrepancies in spans for the values selected, the LLM is expected to be robust enough to arrive at the optimal extraction scenario. For any concerns on quality, our gold-standard test dataset annotations versus the LLM predictions eventually obtained can be publicly browsed at this link `https://scinext-project.github.io/#/r0-estimates`.

Our complex IE task objective. We phrased the following question to formulate our task objective w.r.t. the ORKG-R0 extraction target: *What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?* In essence, it encapsulates an IE task.

The ORKG-R0-based complex IE objective presents a unique approach compared to traditional scientific IE, particularly in biomedicine. Common biomedical IE tasks, like those in the BioCreative V chemical disease relation extraction task corpus (BC5CDR) corpus, focus on document-level entity and relation extraction, linking two elements such as a chemical and a disease with semantic interactions like "interacts" (Li et al., 2016). In contrast, the ORKG-R0 IE model aims to establish a multifaceted link among six distinct elements: infectious disease name, study location, study date, R0 estimate value, %confidence interval values, and method. This approach diverges from the semantic interaction model of BC5CDR, as it does not establish semantic relations between its elements. Instead, it aggregates these elements to form a comprehensive summary representation of a work's contribution to the research problem "$R0$ estimate for infectious diseases."

The ORKG-R0 model is characterized by the underlying principles of the ORKG platform from which it is derived, which emphasizes structured, machine-actionable models of scholarly communication beyond traditional formats like PDFs (Auer et al., 2020). The ORKG prioritizes structured representations of a work's contributions over exhaustive content coverage. In contrast, resources like the BC5CDR corpus (and other similar databases in biomedicine, e.g., BioCreative datasets (Rinaldi et al., 2016; Islamaj Doğan et al., 2019; Krallinger et al., 2017; Miranda et al., 2021)) focus on building extensive knowledge graphs of disease-chemical interactions, with annotations drawn from comprehensive scientific papers. While valuable, these annotations are different in their goal as they do not necessarily provide insights into the specific contributions of a work, such as the discovery of an interaction or the methods used for such discoveries. The ORKG-R0 IE, therefore, aligns more closely with research contribution summarization tasks than with traditional scientific IE tasks in biomedicine. Consequently, models developed for ORKG-R0 IE are unlikely to be directly applicable to conventional biomedical IE tasks.

In terms of objectives, our work is somewhat analogous to Leaderboards in artificial intelligence (AI), which annotate units or tuples comprising Task, Dataset, Metric, and Score (Kabongo et al., 2021a, 2023d,c). However, there are distinct differences in annotation scope: Leaderboards typically utilize the full text of papers, whereas our method relies solely on abstracts. Additionally, the AI community currently lacks a gold-standard dataset for Leaderboard annotations, a gap our dataset aims to fill. We propose our dataset as a pioneering resource in generating structured scientific summaries, addressing the current community need for standardized datasets in this domain.

Instructions for the LLM. Instruction tuning is a novel approach (Khashabi et al., 2020; McCann et al., 2018; Keskar et al., 2019) that improves LLMs' performance by providing explicit instructions during finetuning, guiding the model's behavior (Ouyang et al., 2022; Chung et al., 2022; Min et al., 2022) and enhancing its adaptability and ef-

fectiveness in diverse learning scenarios. Unlike traditional non-instruction tuning methods (Raffel et al., 2020; Liu et al., 2019; Aghajanyan et al., 2021; Aribandi et al., 2021) that rely solely on unlabeled data, instruction tuning incorporates specific guidance, simplifying the finetuning process and enabling better performance on new tasks and domains (Sanh et al., 2022). It became possible to generically prompt an LLM to perform different tasks with a single instruction. As such it can be considered as a template that encodes the task and its objective, in turn telling the LLM what to do with the given objective.

The "Flan 2022 Collection" was a large-scale open-sourced collection of 62 prior publicly released datasets in the NLP community clustered as 12 task types, such as reading comprehension (RC), sentiment, natural language inference (NLI), struct to text, etc. It is the most comprehensive resource facilitating open-sourced LLM development as generic multi-task models. Importantly, and of relevance to this work, FLAN was not just a super-amalgamation of datasets encapsulating different learning objectives, but also included at least 10 human-curated natural instructions per dataset that described the task for that dataset. As such, we select a set of instructions to guide the LLM for our complex IE task from the FLAN collection. Specifically, we identified the applicable instructions to our task were those designed for the SQuAD_v2 (Rajpurkar et al., 2016, 2018) and DROP (Dua et al., 2019) datasets. The general characteristic of the selected instructions is that they encode a context (in our case the paper title and abstract) and the task objective, and instruct the model to fulfill the objective. See Appendix B for further details. The purple boxes in Figure 3 show some exemplars. Examples of all instructions are in Appendix A.

Our work is positioned here, coalescing the most relevant collection of instructions that were used to instruction-finetune the T5 (2020) model class, as the strong reference point for any future open source work on single-task instruction finetuning.

4 Approach

Our approach is single-task instruction-finetuning for our novel introduced complex IE task. As such it aims to be an incremental progression of the instruction-tuning paradigm introduced as FLAN (Finetuned Language Net) (2021; 2022; 2023). Specifically Chung et al. (2022) ask: *are instruction-finetuned models better for single-task finetuning?* as a recommendation for future work. Our work then is a direct examination of this research question except for a novel task type that we also introduce for the first time in the community.

Now, we outline our methodology. **Step 1.** Collect relevant instructions for ORKG-R0 complex IE to guide an LLM towards the desired objective. **Step 2.** Instantiate the instructions to the LLM using gold-standard structured data and a formulated question (e.g. in Appendix A). **Step 3.** Finetune the LLM with the instruction-instantiated data. Three training strategies are explored: single-instruction tuning, all-instruction tuning, and best-instruction tuning based on evaluation results.

4.1 Model

We adopt the FLAN-T5 model (Chung et al., 2022) w.r.t. its public checkpoints. This encoder-decoder sequence generation model is available in a range of sizes: Small 80M, Base 250M, Large 780M, XL 3B, and XXL 11B. We choose the Large model as a middle ground between the Small and XXL models, providing enough parameters for our complex IE task and practicality for deployment. Additionally, we find it inefficient to test extreme scale LLMs for a single task. Our choice of Flan-T5 was motivated by prior empiricism (Longpre et al., 2023) proving instruction-tuned models as more computationally efficient starting checkpoints for new tasks – FLAN-T5 required less finetuning to converge higher and faster than T5 on single downstream tasks (2023). Our model choice builds upon previous research, enhancing the T5 text-to-text sequence generation model (2020) with FLAN-T5 (2022) to improve alignment with instructions in unseen tasks and zero-shot settings. Our resulting model is called ORKG-FLAN-T5$_{R0}$.

5 Evaluations

Dataset. For evaluations, we created a 70%/10%/20% split as train/dev/test sets, respectively, of the 1500 instances. The dataset comprised 1,082 train (464 ans, 618 unans), 120 dev (53 ans, 67 unans), and 300 test (135 ans, 165 unans) instances.

Experimental setup. We used a total of 18 instructions for training, with 9 instructions each from SQuAD_v2 and DROP, specifically instantiated in appendices A.1 and A.2, respectively, suitable for our task. Among these, 2 DROP instructions were

		Highest Scores					Lowest Scores				
Model	**Format**	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**	**General -Accuracy**	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**	**General -Accuracy**
T5	text	12.46	4.56	10.37	11.99	45.00	1.37	0.52	1.21	1.37	45.00
	json	12.01	4.33	10.54	10.49	45.00	1.35	0.51	1.18	1.17	45.00
FLAN-T5	text	51.66	0.42	51.42	51.85	56.33	7.94	3.98	7.68	7.85	45.00
	json	51.64	0.41	51.39	51.74	56.33	7.66	3.82	7.41	7.39	45.00
GPT3.5	text	68.92	17.71	68.20	68.89	79.00	31.00	24.51	30.20	30.83	40.33
	json	68.44	17.26	67.72	67.92	79.00	30.33	23.92	29.57	29.29	40.33
ORKG-FLAN-T5$_{R0}$	text	78.64	28.75	78.33	78.65	86.33	71.34	27.75	70.96	71.41	81.00
	json	80.77	28.03	80.43	80.53	88.67	30.93	27.04	30.55	30.41	44.67

Table 1: Zero-shot results for T5, FLAN-T5 and GPT3.5 tested out-of-the-box to generate structured summaries versus our ORKG-FLAN-T5$_{R0}$ model. Two answer formats plus highest & lowest scores are contrasted. The general accuracy shows models' ability to distinguish between *answerable* vs. *unanswerable* contexts (details in section 3).

		Own Test Instructions							Best Test Instructions							
		Disease-Name	**Location**	**Date**	**R0-Value**	**%CI-Values**	**Method**	**Overall**	**Disease-Name**	**Location**	**Date**	**R0-Value**	**%CI-Values**	**Method**	**Overall**	
s7	Exact	54.26	56.23	29.67	52.90	32.76	34.42	43.59	56.76	55.81	30.94	53.38	33.33	37.17	44.80	s8
	Partial	54.26	59.13	46.15	57.92	62.07	44.51	54.46	56.76	58.72	47.51	58.80	63.16	47.79	55.89	
s6	Exact	54.50	52.25	33.18	52.50	36.84	33.14	43.75	58.51	53.11	35.41	53.00	37.84	33.33	45.21	s1
	Partial	56.08	55.06	48.34	60.30	63.16	40.70	54.06	60.11	55.93	49.76	61.44	64.86	41.52	55.71	
d3	Exact	57.66	55.71	35.56	53.99	18.80	32.29	42.34	58.29	55.17	35.62	56.07	22.22	32.75	43.37	s6
	Partial	59.22	57.38	52.44	58.60	56.41	41.93	54.44	59.89	57.47	52.97	61.21	58.12	42.11	55.42	

Table 2: Our top three ORKG-FLAN-T5$_{R0}$ single-task instruction-finetuned models, based on the single-instruction tuning setting in descending order of overall partial F1 for the text answer format. 1st column: models trained on SQuAD_v2 instr. 7 (s7), SQuAD_v2 instr. 6 (s6), and DROP instr. 3 (d3). Last column: best inference instructions.

		Own Test Instructions							Best Test Instructions							
		Disease-Name	**Location**	**Date**	**R0-Value**	**%CI-Values**	**Method**	**Overall**	**Disease-Name**	**Location**	**Date**	**R0-Value**	**%CI-Values**	**Method**	**Overall**	
d3	Exact	55.64	53.04	32.84	47.62	24.56	32.64	41.11	59.26	53.33	35.18	49.20	25.00	35.12	42.91	s6
	Partial	58.27	56.35	51.74	54.19	56.14	45.10	53.84	61.38	56.67	54.27	56.95	55.36	45.83	55.28	
s8	Exact	54.08	53.51	34.91	48.92	24.56	30.42	41.10	56.85	54.25	31.88	49.53	27.27	31.34	41.89	s1
	Partial	56.63	56.22	50.94	55.83	52.63	41.13	52.34	59.43	56.99	49.28	55.53	56.36	42.17	53.39	
s10	Exact	52.92	52.20	34.74	47.52	16.82	32.82	39.56	57.14	52.23	33.33	48.32	17.65	32.70	40.31	s1
	Partial	54.04	54.55	50.53	53.59	56.07	41.49	51.82	58.26	54.60	49.46	54.67	58.82	40.88	52.91	

Table 3: Our top three ORKG-FLAN-T5$_{R0}$ single-task instruction-finetuned models, based on the single-instruction tuning setting in descending order of overall partial F1 for the JSON answer. 1st column: models trained on DROP instr. 3 (d3), SQuAD_v2 instr. 8 (s8), and SQuAD_v2 instr. 10 (s10). Last column: best inference instructions.

formulated to prompt the LLM to generate a question from a given context. Although indirect to our task, we included them as they were relevant to obtaining capable models, but were excluded from testing. Thus for testing, we had 16 instructions (9 SQuAD and 7 DROP). For training, we had three main experimental settings based on the 18 training instructions. In the first setting, we trained 32 models (16 for text-format and 16 for JSON-format) by tuning FLAN-T5 with a single instruction for our task. Note here models were not trained for the indirect instruction. This setting tested the hypothesis that FLAN-T5 only needed one instruction to perform our task effectively since it already came instruction tuned. In the second setting, we trained two models: one using all 18 instructions with the full training data, and the other using a 50% random sub-sample to prevent overfitting. This resulted in four models for each answer format. The third setting followed a similar approach, training two models with best SQuAD and DROP instructions based on single instruction inference results. Overall, we trained 40 models. Model hyperparamter details are in Appendix C. In terms of compute, all experiments were run on an NVIDIA 3090 GPU. Training took 12-15 hours on smaller datasets and 30 hours on larger datasets, while inference lasted 15-30 minutes for 300 test instances.

Metrics. We experimented in two main settings: zero-shot evaluations and single-task finetuned model evaluations. For the latter, we used recall, precision, and F1 metrics in exact and partial match settings for each of the six ORKG-R0 extraction targets and overall. In the zero-shot evaluations, where models were not guaranteed to respond with the desired structure, we treated the task as struc-

tured summarization. To evaluate these summaries, we used standard summarization ROUGE metrics (Lin, 2004) (details in Appendix D) instead of F1 metrics, which would require complex post-processing and could lead to misinterpretation of the model's response.

5.1 Results and Discussion

Zero-shot evaluations. Table 1 results show model's capacity in generating structured summaries per ORKG-R0. Notably, our single-task instruction-finetuned ORKG-FLAN-T5$_{R0}$ model surpasses its incremental predecessors T5 and FLAN-T5 with the same parameter size of 780M, as well as GPT3.5 (with 1000x more parameters at 175B), confirming the effectiveness of instruction-tuned models for single-task finetuning. Additionally, the general accuracy of the model, which distinguishes between answerable and unanswerable contexts, is significantly improved, at nearly 89%.

Single-task finetuning of instruction-tuned LLM. From the 40 trained models, the best results were achieved in the single-instruction tuning setting, as shown in Table 2 and 3 for text and JSON answers respectively. The best partial overall F1 scores were 55.89% for text answer and 55.28% for JSON answer. Among the 6 properties, extracting $R0$ and %CI values was relatively easier with higher scores for text than JSON. Extracting the method and date proved to be the most challenging. Since our work builds upon the instruction-tuned FLAN-T5 model, it already possesses the capability to handle the instructions we use. Thus, the best inference instruction was not necessarily the same as the one the model was trained on. More results from the all-instruction and best-instruction models can be found in Appendix E.

Impact of diverse inference instructions. Figure 4 offers a look into the inference performance differences from the best ORKG-FLAN-T5$_{R0}$ model. As such the model shows better responses to the SQuAD (orange lines) versus DROP (green lines) inference templates in both text (darker lines) and JSON (lighter lines) answers.

6 Error Analysis

Based on an analysis of all the erroneous responses on the test set from our best model, we identified five main error types. They were further categorized on their impact on recall or precision. For each, mismatching (prediction, annotated label),

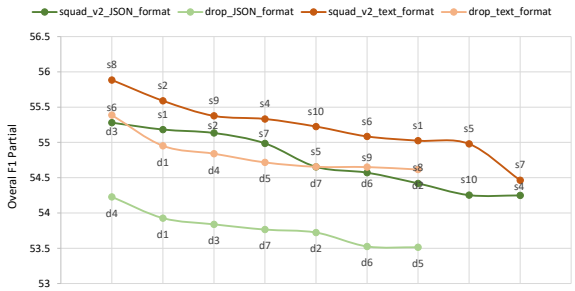

Figure 4: Performances range on inference instructions.

we assigned an error type(s) and on which properties that error had an effect. The five error types are: *Type 1* is where the model answers unanswerable questions (Type 1.1) or fails to provide answers for answerable questions (Type 1.2). *Type 2* is where the model predicts values for a property and the label had no value (Type 2.1) or does not predict a value when the label had a value (Type 2.2). *Type 3* is where the model predicts either more (Type 3.1) or fewer (Type 3.2) contributions than indicated in the label. *Type 4* were inconsistencies between predicted and label values. This may include minor typographical errors (Type 4.1), not fully addressing the label values but still providing a related value in prediction (Type 4.2), including extra related information in prediction (Type 4.3), or generating totally unrelated predicted values (Type 4.4). *Type 5.1* is an invalid predicted JSON.

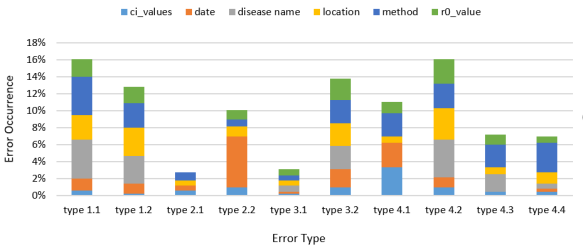

Figure 5: Our best model error types for text format.

Text Response Format. As shown in Figure 5, the most frequent errors in the text-based settings are unanswerable labels (Type 1.1) and incomplete predictions (Type 4.2). These two errors have similar distributions across properties and "method" is the most affected property overall. Type 2.2 errors significantly impact the accuracy of extracting "date" values. In contrast, Type 2.1 and Type 3.1 errors are rare, indicating the model's ability to generate property values and contributions appropriately. Typographical errors (Type 4.1) are common, particularly for "%CI values" and "date,"

suggesting that normalizing label values and using a standard can improve performance in this regard.

Figure 6: Our best model error types for JSON format.

JSON Response Format. Figure 6 shows error Type 4.2 is the primary error affecting properties, similar to text-format errors. The "method" property is the most affected overall, while "date" is particularly impacted by error Type 2.2, highlighting a common issue in JSON-based models. However, JSON models exhibit fewer errors of Type 1 (unanswerable) and instead tend to make more errors in predicting extra text (Type 2.1 and Type 3.1).

7 Conclusions and Future Directions

Searching scientific articles for the Covid-19 R0 estimate yields around 44,000 results. To navigate through this vast amount of information and stay up-to-date with the latest R0 estimates, is inundating for researchers. Next-generation digital libraries like ORKG are transforming this traditional paradigm by capturing machine-actionable data, enabling advanced computational tools such as research comparisons. LLM-powered complex IE technologies can play a crucial role in scaling scientific information extraction. We present a concrete use-case in virology, showcasing the acquisition of LLM-powered structured knowledge with the ORKG-R0 model. To facilitate reproducibility and foster future research, we have made available several resources: our dataset (`https://doi.org/10.5281/zenodo.8068441` licensed under CC BY 4.0), instructions, source code (`https://github.com/mahsaSH717/r0-estimates` licensed under MIT), and our optimally finetuned model for the ORKG-R0 IE task at `https://huggingface.co/orkg/R0_contribution_IE`. Additionally, for enhanced transparency, a selection of our human-annotated test dataset and its corresponding model predictions can be browsed online here `https://scinext-project.github.io/#/r0-estimates`.

To sum up, our work can be seen as a flavor of meta-learning that was seminally proposed by Min et al. (2022) as the meta in-context learning paradigm. We explore meta-learning through instruction-finetuning of an instruction-tuned model, and differ in that we use a zero-shot rather than a few-shot training and testing scenario. We relegate few-shot in-context model learning to future work. While this work comprehensively evaluates the T5 class of LLMs, there are other promising LLMs like PaLM (Chowdhery et al., 2022), Chinchilla (Hoffmann et al., 2022), and ChatGPT (Brown et al., 2020; Ouyang et al., 2022) that can be further investigated for NLP tasks with instructions. Exploring alternative model families is a fruitful direction for future research. Additionally, model distillation (Hinton et al., 2015; Jiao et al., 2020; Sanh et al., 2019; Wang et al., 2020b) holds potential for transferring knowledge from large teacher models to smaller, efficient student models. This approach holds promise, particularly in scenarios where single-task tuned models are desired, as we propose in this study.

Limitations

This section presents a discussion of the limitations w.r.t. the two main facets of this work: structured scholarly knowledge publishing (paragraph I) and LLM scaling experiments for single-task instruction finetuning (paragraph II).

I. Structured Scholarly Knowledge Publishing This work proposes the ORKG-R0 model that records a fine-grained structured representation of the salient facets of a research contribution on the specific research problem of investigating the R0 number of infectious diseases. For such popular research use-cases in the community, e.g., capturing Leaderboards in the empirical AI research as Task, Dataset, Metric, and Score (Kabongo et al., 2021b, 2023a,b), as another example apart from the one we address in this work, a current limitation that such a contribution-centric fine-grained structured scholarly knowledge publishing model faces is it's *adoption and standardization*. The widespread adoption of the semantic scholarly knowledge publishing model is still in its early stages, and achieving consensus on standard formats, ontologies, and metadata remains a challenge. This lack of standardization can hinder interoperability and limit the accessibility of knowledge across different platforms and communities. To

overcome this limitation, i.e. to realize this vision of the publishing of fine-grained structured scholarly contributions to better assist researchers to stay on track with research progress many more collaborative advocacy and community-building efforts would need to be set in place. The trajectory, however, looks promising. The ORKG since its inception in 2018 currently has a knowledge base of roughly 41k structured contributions. More stats here https://orkg.org/stats. In addition, yearly paid community curation grants are run inviting researchers from various disciplines to help curate a high-quality knowledge graph (https://orkg.org/about/28/Curation_Grants). Finally, the ORKG has initiated collaborations with various conferences and journals that ask authors to submit research comparisons of their work versus related work to help expedite the peer-review process. E.g., see the last point in the Author Guidelines in the SEMANTiCS 2023 call for papers https://2023-eu.semantics.cc/page/cfp_rev_rep. To this end, the platform is integrated with content creator anonymization features to support double-blind review protocols. More information here https://orkg.org/about/22/Conferences_and_Journals.

As a second limitation of semantic publishing, the ORKG is designed to be a next-generation digital library that supports fine-grained scholarly knowledge publishing stored as a large-scale knowledge graph in the backend (Jaradeh et al., 2019). It is also amenable to be published in the Linked Open Data (LOD) Cloud https://lod-cloud.net/. Thus it follows the best practices laid out in Berners-Lee et al.'s (2001) the Semantic Web. As such the engineering of this platform entails a high degree of *technical complexity* compared with the traditional PDF-based publishing platforms. Implementing and maintaining the infrastructure required for semantic publishing models can be technically complex and resource-intensive. It requires expertise in semantic technologies, data management, and ontological engineering. Nevertheless, the ORKG platform supports the integration of widgets for its various features in other platforms. This would lower the technical entrance barrier for other publishers to also support the semantic publishing of scientific contributions.

II. Scaling Single-Task Instruction-tuning of LLMs This work has investigated the moderate-sized FLAN-T5 Large model with 780M parameters. Prior work reported: "we see that increasing model scale by an order of magnitude (i.e., 8B -> 62B or 62B -> 540B) improves performance substantially for both finetuned and non-finetuned models" (Chung et al., 2022). Borrowing insights from the earlier experiments on scaling models, potentially, a single-task finetuned model performance could be boosted if larger scale models were used. This aspect while not analyzed in this work is relegated to future work. However, a more practically viable option would not just be additional scaling investigations, but these combined with model distillation (Hinton et al., 2015).

Acknowledgements

We thank the anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was jointly supported by the German BMBF project SCINEXT (ID 01lS22070) and the DFG NFDI4DataScience initiative (ID 460234259).

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Amir Aryani, Marta Poblet, Kathryn Unsworth, Jingbo Wang, Ben Evans, Anusuriya Devaraju, Brigitte Hausstein, Claus-Peter Klas, Benjamin Zapilko, and Samuele Kaplun. 2018. A research graph dataset for connecting research data repositories using rd-switchboard. *Scientific Data*, 5:180099.

Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. 2020. Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, 44(3):516–529.

Sören Auer. 2018. Towards an open research knowledge graph.

Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. 2020. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Caroline Birkle, David A Pendlebury, Joshua Schnell, and Jonathan Adams. 2020. Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1):363–376.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2.

Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. 2020. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10.

Ivo M. Foppa. 2017. 7 - o. diekmann, j. heesterbeek, and j.a. metz (1991) and p. van den driessche and j. watmough (2002): The spread of infectious diseases in heterogeneous populations. In Ivo M. Foppa, editor, *A Historical Introduction to Mathematical Modeling of Infectious Diseases*, pages 157–194. Academic Press, Boston.

Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

R Brian Haynes, Cynthia D Mulrow, Edward J Huth, Douglas G Altman, and Martin J Gardner. 1990. More informative abstracts revisited. *Annals of internal medicine*, 113(1):69–76.

Robert SA Hayward, Mark C Wilson, Sean R Tunis, Eric B Bass, Haya R Rubin, and R Brian Haynes. 1993. More informative abstracts of articles describing clinical practice guidelines.

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Sally Hopewell, Mike Clarke, David Moher, Elizabeth Wager, Philippa Middleton, Douglas G Altman, Kenneth F Schulz, and Consort Group. 2008. Consort for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS medicine*, 5(1):e20.

Rezarta Islamaj Doğan, Sun Kim, Andrew Chatr-Aryamontri, Chih-Hsuan Wei, Donald C Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C Panyam, et al. 2019. Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. *Database*, 2019:bay147.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 243–246.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. The stm report. *An overview of scientific and scholarly publishing. 5th edition October*.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2023a. Orkg-leaderboards: A systematic workflow for mining leaderboards as a knowledge graph. *arXiv preprint arXiv:2305.11068*.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2023b. Zero-shot entailment of leaderboards for empirical ai research. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2023*.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2023c. Zero-shot entailment of leaderboards for empirical ai research. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 237–241.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021a. Automated mining of leaderboards for empirical ai research. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*, pages 453–470. Springer.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021b. Automated mining of leaderboards for empirical ai research. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*, pages 453–470. Springer.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2023d. Orkg-leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, pages 1–14.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *arXiv preprint arXiv:1904.09286*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Hemant Kulkarni. 1996. Structured abstracts: still more. *Annals of Internal Medicine*, 124(7):695–696.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Paolo Manghi, Claudio Atzori, Alessia Bardi, Jochen Shirrwagen, Harry Dimitropoulos, Sandro La Bruzzo, Ioannis Foufoulas, Aenne Löhden, Amelie Bäcker, Andrea Mannocci, Marek Horst, Miriam Baglioni, Andreas Czerniak, Katerina Kiatropoulou, Argiro Kokogiannaki, Michele De Bonis, Michele Artini, Enrico Ottonello, Antonis Lempesis, Lars Holm Nielsen, Alexandros Ioannidis, Chiara Bigarella, and Friedrich Summan. 2019. Openaire research graph dump.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, pages 11–21.

Takeo Nakayama, Nobuko Hirai, Shigeaki Yamazaki, and Mariko Naito. 2005. Adoption of structured abstracts by general medical journals and format for a structured abstract. *Journal of the Medical Library Association*, 93(2):237.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 432–447, Online. Association for Computational Linguistics.

Allard Oelen, Mohamad Yaser Jaradeh, Kheir Eddine Farfar, Markus Stocker, and Sören Auer. 2019. Comparing research contributions in a scholarly knowledge graph. In *CEUR Workshop Proceedings 2526 (2019)*, volume 2526, pages 21–26. Aachen: RWTH Aachen.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Allen H Renear and Carole L Palmer. 2009. Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325(5942):828–832.

Fabio Rinaldi, Tilia Renate Ellendorff, Sumit Madan, Simon Clematide, Adrian Van der Lek, Theo Mevissen, and Juliane Fluck. 2016. Biocreative v track 4: a shared task for the extraction of causal network information using the biological expression language. *Database*, 2016:baw067.

Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nicolas W Hengartner, and Ruian Ke. 2020. The novel coronavirus, 2019-ncov, is highly contagious and more infectious than initially estimated. *arXiv preprint arXiv:2002.03268*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

David Shotton. 2009. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2):85–94.

Luciana B Sollaci and Mauricio G Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the medical library association*, 92(3):364.

Markus Stocker, Allard Oelen, Mohamad Yaser Jaradeh, Muhammad Haris, Omar Arab Oghli, Golsa Heidari, Hassan Hussein, Anna-Lena Lorenz, Salomon Kabenamualu, Kheir Eddine Farfar, et al. 2023. Fair scientific information with the open research knowledge graph. *FAIR Connect*, 1(1):19–21.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020a. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 5776–5788.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Vitalis Wiens, Markus Stocker, and Sören Auer. 2020. Towards customizable chart visualizations of tabular data using knowledge graphs. In *Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, November 30–December 1, 2020, Proceedings 22*, pages 71–80. Springer.

A Instructions: Qualitative Examples

In this section, we elicit each of the instructions that were considered in this work as formulated in the FLAN 2022 Collection for the SQuAD_v2 and DROP datasets.

A.1 The Stanford Question Answering Dataset (SQuAD_v2)

Instruction 1:

title: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

context: The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27 ...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction:

{title}:\n\n{context}\n\n Please answer a question about this article. If the question is unanswerable, say "unanswerable". {question}

Instruction 2:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27 ...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Read this and answer the question. If the question is unanswerable, say "unanswerable".\n\n{context}\n\n{question}

Instruction 3:

This instruction is omitted in this work.

Instruction: (What is a question about this article? If the question is unanswerable, say "unanswerable"),\n{context}\n{question}

Instruction 4:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27 ...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n{question} (If the question is unanswerable, say "unanswerable")

Instruction 5:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n Try to answer this question if possible (otherwise reply "unanswerable"):{question}

Instruction 6:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on

the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n If it is possible to answer this question, answer it for me (else, reply "unanswerable"): {question}

Instruction 7:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n \n Answer this question, if possible (if impossible, reply "unanswerable"): {question}

Instruction 8:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Read this: {context}\n \n {question} \n What is the answer? (If it cannot be answered, return "unanswerable")

Instruction 9:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Read this: {context}\n Now answer this question, if there is an answer (If it cannot be answered, return "unanswerable"): {question}

Instruction 10:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n Is there an answer to this question (If it cannot be answered, say "unanswerable"): {question}

A.2 Discrete Reasoning over Paragraphs (DROP) Dataset

Instruction 1:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Answer based on context: \n \n{context}\n \n {question}

Instruction 2:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n \n Answer this question based on the article: {question}

Instruction 3:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n \n {question}

Instruction 4:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n Answer this question: {question}

Instruction 5:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Read this article and answer this question {context}\n {question}

Instruction 6:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the

onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n \n Based on the above article, answer a question. {question}

Instruction 7:

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Context: {context}\n \n Question: {question}\n \n Answer:

Instruction 8:

This instruction is omitted in this work.

Instruction: Write an article that answers the following question: {question}

Instruction 9:

Note single-instruction finetuned models were not trained on this instruction. This instruction was only used in the all-instruction training setting.

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: Write a question about the following article: {context}

Instruction 10:

Note single-instruction finetuned models were not trained on this instruction. This instruction was only used in the all-instruction training setting.

context: Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 22 February 2020

The novel coronavirus disease (COVID-19) poses a serious threat to global public health and economics. Serial interval (SI), time between the onset of symptoms of a primary case and a secondary case, is a key epidemiological parameter. We estimated SI of COVID-19 in Shenzhen, China based on 27...

question: What are the values for the following properties of the basic reproduction number estimate (R0): disease name, location, date, R0 value, %CI values, and method?

Instruction: {context}\n \n Ask a question about this article.

B ORKG-R0 for the FLAN Collection

In this section, we discuss the relation of our complex IE task formulated as ORKG-R0 to the task types already in the FLAN collection (2021; 2023) as a new candidate for inclusion. As mentioned earlier, FLAN has 12 task type clusters of 63 datasets. Two of which are reading comprehension (RC) and struct-to-text, among others. In this respect, our task could either be considered part of the RC task or as a new task type i.e. text-to-struct. In an RC task, e.g. SQuAD (2016), a context passage is provided along with a question to test comprehension. Our complex IE task is similar, where given a scholarly paper's title and abstract as context, the machine must generate a structured summary by understanding the context and assigning applicable extracted values for the ORKG-R0 properties. Furthermore, the model must also create ORKG-R0 clusters for abstracts reporting multiple contributions.[2] Otherwise, it could be intro-

[2]Note, there is a subtle difference between RC and the related question-answering (QA) task type. In QA, complex

duced into the FLAN collection as a new task type called text-to-struct. As such, for instance, the WebNLG (Gardent et al., 2017) or DART (Nan et al., 2021) datasets in the struct-to-text cluster, seek to convert structured data in RDF to text. Notably, our task is its direct inverse which seeks to obtain structured property-value tuples which can easily be represented in RDF syntax.

C Our Experimental Hyperparamters

We had different training experimental settings to train on different datasets with different sizes (single-instruction model tuning, all-instructions model tuning, all-instructions model tuning with 50% subsampled training data, best-instructions model tuning, and best-instructions model tuning with 50% subsampled training data).

The hyperparameters are: batch size and number of training epochs, which differ based on each dataset group mentioned above. the batch size was either 32 or 16 and the number of epochs were one of 10, 15, 20, and 30 values. In all settings we used early stopping which stops the training if the "Overall Partial F1" score dose not improve at least 0.1% after completing 10 consecutive training epochs. For all settings we used AdafactorSchedule and Adafactor optimizer (Shazeer and Stern, 2018) with scale_parameter=True, relative_step=True, warmup_init=True, lr=None, which is one of the combinations working well according to the community for T5 finetuning.

The evaluations were done on each epoch on the dev set and we kept two best (the one maximizing the "Overall Partial F1" score) and last checkpoints in each model training process to then use for inference on test set.

D ROUGE Evaluation Metrics

The ROUGE metrics (Lin, 2004) are commonly used for evaluating the quality of text summarization systems. ROUGE-1 measures the overlap of unigram (single word) units between the generated summary and the reference summary. ROUGE-2 extends this to measure the overlap of bigram (two consecutive word) units. ROUGE-L calculates the longest common subsequence between the generated and reference summaries, which takes into account the order of words. ROUGE-LSum is an extension of ROUGE-L that considers multiple reference summaries by treating them as a single summary. These metrics provide a quantitative assessment of the similarity between the generated and reference summaries, helping researchers and developers evaluate and compare the effectiveness of different summarization approaches. They have become widely used benchmarks in the field of automatic summarization.

E Additional Results

Finally, in this last appendix section, we show the highest and lowest results obtained from the two other experimental settings discussed in the main paper. I.e. all-instruction model finetuning, in two subsettings: with all the training data and with a 50% random subsample of the training data. These results are presented in Table 4 and Table 5, respectively, for the text format and JSON format responses. And furthermore, results are shown for the best-instruction finetuning setting in two subsettings: with all the training data and with a 50% random subsample of the training data. These results are presented in Table 6 and Table 7, respectively, for the text format and JSON format responses.

IE would require breaking down the RC extraction target into multiple questions, such as the disease name or the reported location, etc., unlike in RC.

			All Data									Data From Random Selection of Templates						
	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall
Top 2 Highest	s1	Exact	54.24	52.12	21.51	47.84	13.59	33.96	37.26	d7	Exact	54.88	51.69	33.48	49.84	33.06	33.43	42.76
		Partial	54.80	53.94	38.71	55.22	54.37	44.65	50.35		Partial	55.41	54.49	48.46	56.26	57.85	40.47	52.38
	d6	Exact	53.52	51.81	21.51	47.84	13.59	33.23	36.96	d1	Exact	54.69	51.70	29.60	50.16	36.67	32.14	42.53
		Partial	54.08	53.61	37.63	55.29	54.37	43.89	49.89		Partial	55.23	55.11	43.95	56.50	58.33	39.29	51.66
Top 2 Lowest	d4	Exact	53.22	51.65	20.32	46.86	13.46	33.02	36.47	s6	Exact	56.02	47.00	27.56	45.98	36.07	31.06	40.63
		Partial	53.78	53.45	36.36	54.62	53.85	42.99	49.25		Partial	56.51	50.13	40.94	51.31	55.74	38.15	48.93
	s8	Exact	53.22	52.25	19.35	46.71	13.59	33.64	36.51	d4	Exact	52.58	47.67	27.23	47.13	36.07	32.57	40.56
		Partial	53.78	53.45	34.41	54.19	54.37	44.24	49.15		Partial	53.09	50.96	41.70	52.19	55.74	38.29	48.79

Table 4: Top two highest and lowest inference results by ORKG-FLAN-T5$_{R0}$ all-instructions and all-instructions with 50% subsampled finetuned models, in descending order of overall partial F1 for the text answer. template column: inference instructions. SQuAD_v2 instr. 1 (s1), DROP instr. 6 (d6), DROP instr. 4 (d4), SQuAD_v2 instr. 8 (s8), DROP instr. 7 (d7), DROP instr. 1 (d1), SQuAD_v2 instr. 6 (s6), and DROP instr. 4 (d4).

			All Data									Data From Random Selection of Templates						
	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall
Top 2 Highest	s5	Exact	51.25	48.94	29.03	41.97	13.59	27.04	35.38	d4	Exact	56.27	47.76	31.02	49.33	22.64	32.91	40.06
		Partial	53.48	50.15	44.09	49.89	54.37	35.85	48.06		Partial	56.82	50.75	45.99	56.19	54.72	42.41	51.27
	d2	Exact	50.14	48.94	26.88	41.97	13.59	27.67	34.93	d6	Exact	56.27	47.76	31.02	50.00	22.64	32.38	40.08
		Partial	52.37	50.15	44.09	49.68	54.37	36.48	47.95		Partial	56.82	50.75	45.99	56.32	54.72	41.90	51.20
Top 2 Lowest	s1	Exact	50.70	47.13	25.81	42.11	13.59	25.79	34.25	s8	Exact	54.55	47.06	32.46	49.01	22.43	32.50	39.72
		Partial	52.92	48.34	44.09	49.02	54.37	33.96	47.21		Partial	55.10	50.00	46.07	55.14	50.47	41.88	49.88
	d1	Exact	50.42	47.42	25.95	41.58	13.73	25.95	34.24	s9	Exact	54.14	47.34	32.46	49.83	20.75	31.97	39.47
		Partial	52.66	49.24	44.32	47.83	52.94	34.81	47.06		Partial	54.70	50.30	46.07	55.75	49.06	41.38	49.64

Table 5: Top two highest and lowest inference results by ORKG-FLAN-T5$_{R0}$ all-instructions and all-instructions with 50% subsampled finetuned models, in descending order of overall partial F1 for the JSON answer. template column: inference instructions. SQuAD_v2 instr. 5 (s5), DROP instr. 2 (d2), SQuAD_v2 instr. 1 (s1), DROP instr. 1 (d1), DROP instr. 4 (d4), DROP instr. 6 (d6), SQuAD_v2 instr. 8 (s8), SQuAD_v2 instr. 9 (s9).

			All Data									Data From Random Selection of Templates						
	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall
Top 2 Highest	s2	Exact	49.21	54.85	30.00	49.20	22.81	32.35	39.79	s6	Exact	48.04	47.15	24.88	41.59	19.42	23.18	34.16
		Partial	50.26	57.06	51.00	54.35	52.63	44.12	51.73		Partial	48.53	49.86	38.28	49.12	54.37	38.27	46.62
	d6	Exact	49.10	53.66	31.84	49.22	23.42	31.70	39.89	d3	Exact	47.62	46.19	26.92	41.92	18.35	21.47	33.87
		Partial	50.65	55.83	47.76	54.55	54.05	43.23	51.15		Partial	48.10	48.82	41.35	48.28	55.05	36.13	46.48
Top 2 Lowest	s9	Exact	48.04	52.05	32.16	49.21	23.42	32.56	39.65	s8	Exact	47.39	46.35	21.72	41.18	17.24	21.47	32.60
		Partial	49.61	54.25	47.24	54.43	54.05	43.60	50.66		Partial	47.87	48.96	34.39	48.57	51.72	35.08	44.50
	s1	Exact	47.92	51.37	30.00	47.80	23.01	32.56	38.84	s9	Exact	46.90	44.44	22.33	40.00	16.39	21.88	32.07
		Partial	49.48	53.55	45.00	53.28	53.10	43.60	49.80		Partial	47.36	46.97	34.42	45.99	47.54	35.11	43.01

Table 6: Top two highest and lowest inference results by ORKG-FLAN-T5$_{R0}$ best-instructions and best-instructions with 50% subsampled finetuned models, in descending order of overall partial F1 for the text answer. template column: inference instructions. SQuAD_v2 instr. 2 (s2), DROP instr. 6 (d6), SQuAD_v2 instr. 9 (s9), SQuAD_v2 instr. 1 (s1), SQuAD_v2 instr. 6 (s6), DROP instr. 3 (d3), SQuAD_v2 instr. 8 (s8), and SQuAD_v2 instr. 9 (s9).

			All Data									Data From Random Selection of Templates						
	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall	Template	Match Type	Disease -Name	Location	Date	R0 -Value	CI -% Values	Method	Overall
Top 2 Highest	s1	Exact	49.28	47.85	32.82	46.25	28.30	27.94	38.77	s2	Exact	47.03	50.54	32.65	42.48	27.87	25.22	37.68
		Partial	51.00	50.31	46.15	50.67	52.83	36.19	47.90		Partial	48.58	52.72	44.90	50.30	57.38	35.19	48.31
	s9	Exact	47.29	48.02	31.96	47.21	26.67	27.67	38.16	s1	Exact	49.75	49.60	33.20	39.77	25.20	24.93	37.12
		Partial	49.00	50.46	45.36	51.79	51.43	37.11	47.58		Partial	50.25	51.19	45.06	47.13	55.12	35.13	47.45
Top 2 Lowest	d3	Exact	49.13	46.91	30.77	44.74	24.76	27.56	37.33	d2	Exact	46.80	48.83	32.13	39.55	23.26	24.93	35.94
		Partial	50.29	48.77	44.10	49.33	51.43	37.18	46.88		Partial	48.28	50.39	44.18	46.83	51.16	35.46	46.13
	d1	Exact	48.26	46.58	29.32	45.03	27.18	28.03	37.43	d6	Exact	46.42	48.70	33.33	39.66	23.44	25.07	36.13
		Partial	49.42	48.45	42.93	48.77	52.43	37.58	46.63		Partial	47.90	50.26	45.24	46.72	50.00	35.65	46.05

Table 7: Top two highest and lowest inference results by ORKG-FLAN-T5$_{R0}$ best-instructions and best-instructions with 50% subsampled finetuned models, in descending order of overall partial F1 for the JSON answer. template column: inference instructions. SQuAD_v2 instr. 1 (s1), SQuAD_v2 instr. 9 (s9), DROP instr. 3 (d3), DROP instr. 1 (d1), SQuAD_v2 instr. 2 (s2), SQuAD_v2 instr. 1 (s1), DROP instr. 2 (d2), and DROP instr. 6 (d6).