Harnessing Large Language Models as Research Assistants and in Science

Presented at: 2nd European Summer School on Artificial Intelligence (ESSAI 2024), Athens, Greece
Date of event: 24th July 2024
Presented by: Jennifer D'Souza (Junior AI Research Group Lead)
Contact: https://www.linkedin.com/in/jennifer-l-dsouza/

Open Research Knowledge Graph

Open Research Knowledge Graph

- Digital library for machine-actionable knowledge communicated in scholarly literature
- Contains structured scholarly knowledge of content beyond keywords
  - not just bibliographic metadata
- Supports multimodal interactions through human crowdsourcing, and automated text mining techniques
- Exists at various stages of the scientific publishing lifecycle: at time of creating knowledge as well as writing, submitting, publishing, and reading scientific information

Scholarly Knowledge. Structured.

IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 15, NO. 4, AUGUST 2014 1441

Situational Knowledge Representation for Traffic Observed by a Pavement Vibration Sensor Network

Markus Stocker, Mauno Rönkkö, and Mikko Kolehmainen

***Abstract*—Information systems that build on sensor networks often process data produced by measuring physical properties. These data can serve in the acquisition of knowledge for real-world situations that are of interest to information services and, ultimately, to people. Such systems face a common challenge, namely the considerable gap between the data produced by measurement and the abstract terminology used to describe real-world situations. We present and discuss the architecture of a software system that utilizes sensor data, digital signal processing, machine learning, and knowledge representation and reasoning to acquire, represent, and infer knowledge about real-world situations observable by a sensor network. We demonstrate the application of the system to vehicle detection and classification by measurement of road pavement vibration. Thus, real-world situations involve vehicles and information for their type, speed, and driving direction.**

***Index Terms*—Knowledge acquisition, knowledge representation, machine learning, sensor data, sensor networks, traffic monitoring.**

I. Introduction

We propose a software system architecture and implementation for the continuous and automated representation of knowledge for real-world situations observable by a sensor network. In this paper, we demonstrate the application of the software system to intelligent transportation systems. Thus, real-world situations involve vehicles and information for their type, speed, and driving direction.

According to Finkelstein [1], "measurement is the process of empirical, objective, assignment of numbers to properties of objects or events of the real world in such a way as to describe them." A sensor is a device that performs measurement, in that it transforms the signal of a physical property (e.g., heat) into numbers or, more generally, into data [2]. Sensor measurement is, hence, the process of recurrent application of such transformation for certain temporal and spatial locations. The result of sensor measurement is sensor data. Sensor data represent the change of the signal over time.

Despite recent advancements in sensor data management, processing, and query [2]–[4], as well as semantic description of sensors and data [5]–[7], making sense of sensor data is an ongoing challenge [8]–[10] because of the difference in the degree to which sensor data represents information about a signal and information about, or related to, a physical property [11]. In other words, it is a challenge because of the considerable gap between data produced by measurement and abstract terminology [12] used by people to describe (the properties of) real-world objects or events.

We are interested in *situations* involving real-world objects that affect a physical property, for which a signal is measured by means of sensors. In this paper, vehicles are the real-world objects and road pavement vibration is the physical property. We present the architecture of a software system that utilizes digital signal processing, machine learning, and knowledge representation and reasoning to acquire, represent, and infer knowledge about real-world situations involving vehicles. The system aims at reducing the gap between road pavement vibration measurement data and abstract terminology used to describe real-world situations involving vehicles.

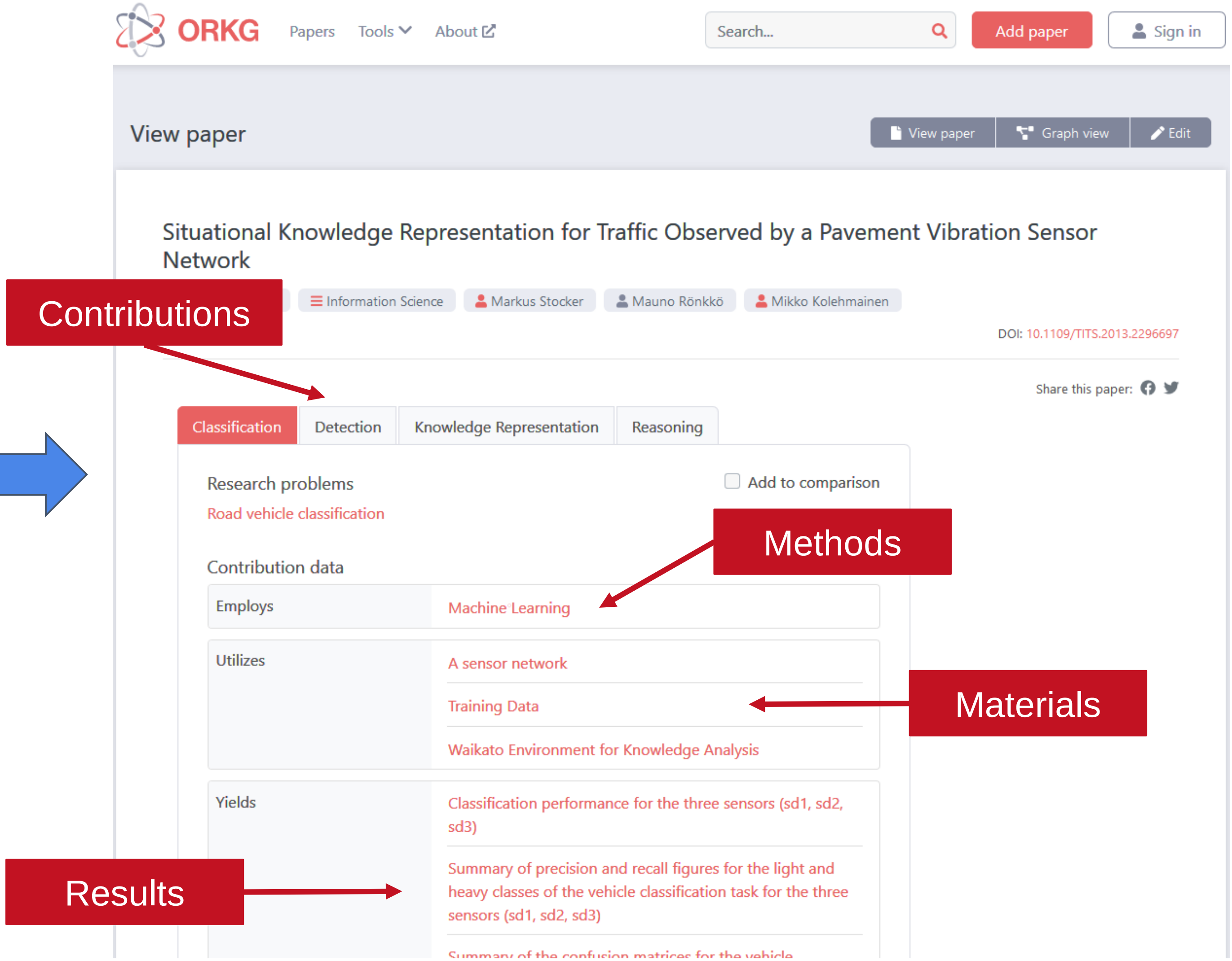
Digital signal processing techniques are iteratively applied to a sliding window over sensor data to enhance the vibration signal and to transform sensor data (time domain) into patterns (frequency domain). Machine learning is used to classify patterns. We employ multilayer perceptron (MLP) feedforward artificial neural networks [13]. Techniques in knowledge representation are utilized to formally represent domain concepts, instances, and relations. A concept of interest to our domain is the vibration sensor. The (installed) sensors are represented as instances of this concept. An instance may have a number of relations, e.g., to a spatial location. We represent sensors and observations using the Semantic Sensor Network Ontology (SSNO) [14].[1] SSNO is an "ontology for describing the capabilities of sensors, the act of sensing and the resulting observations" [15]. We employ the Situation Theory Ontology[2] (STO) [16] to represent knowledge about real-world situations, which are acquired from observations. The STO captures the key aspects of the situation theory developed by Barwise and Perry [17] and extended by Devlin [18]. The theory relates to the work on situation awareness by Endsley [19], [20] as it encompasses most of the concepts discussed in [16]. Both the SSNO and the STO serve as upper ontologies from which we extend to accommodate domain knowledge. The hybrid use of the SSNO and the STO allows for a multilevel abstraction of sensor measurement data and the use of appropriate terminology and formalization at each level.

Manuscript received April 12, 2013; revised August 16, 2013 and November 20, 2013; accepted December 22, 2013. Date of publication February 4, 2014; date of current version August 1, 2014. The infrastructure to access and collect vibration and camera data, as well as the data, are part of research funded by Tekes, the Finnish Funding Agency for Technology and Innovation (funding decision number 40075/09). The Associate Editor for this paper was P. Grisleri.

The authors are with the Department of Environmental Science, University of Eastern Finland, 70211 Kuopio, Finland (e-mail: markus.stocker@uef.fi; mauno.ronkko@uef.fi; mikko.kolehmainen@uef.fi).

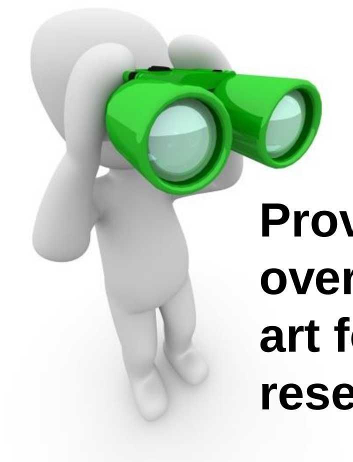
Digital Object Identifier 10.1109/TITS.2013.2296697

[1] http://purl.oclc.org/NET/ssnx/ssn
[2] http://vistology.com/ont/2008/STO/STO.owl

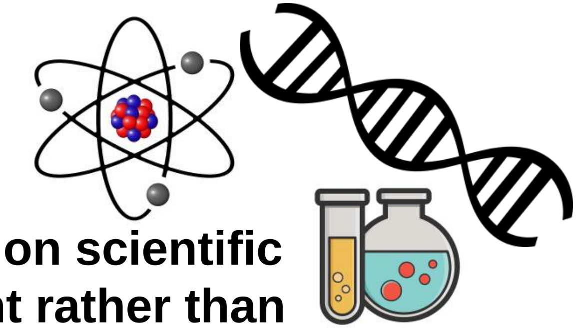


ORKG’s Objectives

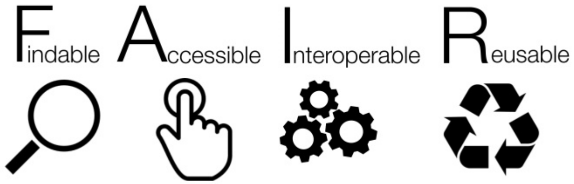
Foster collaboration



Provide overview over the state-of-the-art for specific research problems

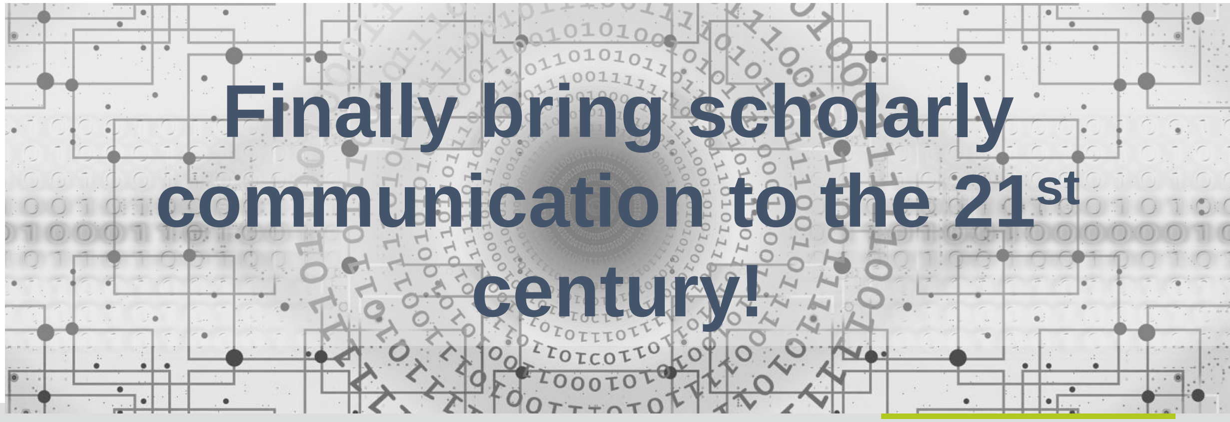


Focus on scientific content rather than document



Make research FAIR

Tackle interdisciplinary challenges such as climate change research, disease prevention, etc.



Finally bring scholarly communication to the 21st century!

Use Case 1: SARS-CoV 2 Basic Reproduction Number

Quantitative Biology > Populations and Evolution

COVID-19 e-print

Important: e-prints posted on arXiv are not peer-reviewed by arXiv; they should not be relied upon without context to guide clinical practice or health-related behavior and should not be reported in news media as established information without consulting multiple experts in the field.

[Submitted on 20 Mar 2020]

The early phase of the COVID-19 outbreak in Lombardy, Italy

Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, Trentini F, Guzzetta G, Marziano V, Barone A, Magoni M, Deandrea S, Diurno G, Lombardo M, Faccini M, Pan A, Bruno R, Pariani E, Grasselli G, Piatti A, Gramegna M, Baldanti F, Melegaro A, Merler S

In the night of February 20, 2020, the first case of novel coronavirus disease (COVID-19) was confirmed in the Lombardy Region, Italy. In the week that followed, Lombardy experienced a very rapid increase in the number of cases. We analyzed the first 5,830 laboratory-confirmed cases to provide the first epidemiological characterization of a COVID-19 outbreak in a Western Country. Epidemiological data were collected through standardized interviews of confirmed cases and their close contacts. We collected demographic backgrounds, dates of symptom onset, clinical features, respiratory tract specimen results, hospitalization, contact tracing. We provide estimates of the reproduction number and serial interval. The epidemic in Italy started much earlier than February 20, 2020. At the time of detection of the first COVID-19 case, the epidemic had already spread in most municipalities of Southern-Lombardy. The median age for of cases is 69 years (range, 1 month to 101 years). 47% of positive subjects were hospitalized. Among these, 18% required intensive care. The mean serial interval is estimated to be 6.6 days (95% CI, 0.7 to 19). We estimate the basic reproduction number at 3.1 (95% CI, 2.9 to 3.2). We estimated a decreasing trend in the net reproduction number starting around February 20, 2020. We did not observe significantly different viral loads in nasal swabs between symptomatic and asymptomatic. The transmission potential of COVID-19 is very high and the number of critical cases may become largely unsustainable for the healthcare system in a very short-time horizon. We observed a slight decrease of the reproduction number, possibly connected with an increased population awareness and early effect of interventions. Aggressive containment strategies are required to control COVID-19 spread and catastrophic outcomes for the healthcare system.

Subjects: **Populations and Evolution (q-bio.PE)**
Cite as: arXiv:2003.09320 **[q-bio.PE]**
(or arXiv:2003.09320v1 **[q-bio.PE]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

Submission history

From: Alessia Melegaro [view email]
[v1] Fri, 20 Mar 2020 15:17:36 UTC (988 KB)

Results

The epidemic in Italy started much earlier than February 20, 2020. At the time of detection of the first COVID-19 case, the epidemic had already spread in most municipalities of Southern-Lombardy. The median age for of cases is 69 years (range, 1 month to 101 years). 47% of positive subjects were hospitalized. Among these, 18% required intensive care. The mean serial interval is estimated to be 6.6 days (95% CI, 0.7 to 19). We estimate the basic reproduction number at 3.1 (95% CI, 2.9 to 3.2). We estimated a decreasing trend in the net reproduction number starting around February 20, 2020. We did not observe significantly different viral loads in nasal swabs between symptomatic and asymptomatic.

Here we provide an analysis of the first 5,830 laboratory-confirmed cases reported in Lombardy, with date of symptoms onset over the period from January 14 to March 8, 2020. Epidemiological analyses of the confirmed cases and their background demographic and exposure characteristics are presented here as well as the transmission dynamics of the infection within the Region. Also, the virological analysis on a subsample of the reported cases is included to provide preliminary assessment of the level of the viral load among symptomatic and asymptomatic cases.

ORKG Papers Tools About

Search...

Add paper

Sign in

TIB

View paper

Graph view | Edit

The early phase of the COVID-19 outbreak in Lombardy, Italy

2020 | Virology | Cereda D | Tirani M | Rovida F | Demicheli | Ajelli M | Poletti P | Trentini F | Guzzetta G | Marziano | Barone A | Magoni M | Deandrea S | Diurno G | Lombardo M | Faccini M | Pan A | Bruno R | Pariani E | Grasselli G | Piatti A | Gramegna M | Baldanti F | Melegaro A | Merler S

Published in: *arXiv.org*

Share this paper:

Contribution 1

Research problems

Add to comparison

COVID-19 reproductive number

Contribution data

95% Confidence interval	2.9-3.2
Location	Lombardy, Italy
R0 estimates (average)	3.1
Study date	2020-01-14/2020-03-08

Properties	The early pr[illegible] of the COVID-19 outbreak in Lombardy, Italy *2020 - Contribution 1*	Transmission [illegible]ential of COVID-19 in Iran *2020 - Contribution 1*	Transmission [illegible]ntial of COVID-19 in Iran *2020 - Contribution 2*	Estimating the [illegible]eration interval for COVID-19 based on symptom onset data *2020 - Contribution 1*
location	Lombardy, Italy	Iran	Iran	Singapore
Time period	Time interval	Time interval	Time interval	Time interval
has beginning	2020-01-14	2020-02-19	2020-02-19	2020-01-21
has end	2020-03-08	2020-02-29	2020-02-29	2020-02-26
Basic reproduction number	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification
Has value	3.1	3.6	3.58	1.27
Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)
Lower confidence limit	2.9	3.4	1.29	1.19
Upper confidence limit	3.2	4.2	8.46	1.36
Method*		generalized growth model	based on the calculation of the epidemic's doubling times: estimated epidemic doubling time of 1.20 (95% CI, 1.05, 1.44) days	generation interval

Example

Note this is not possible by the current predominant methods of scholarly communication that are hardly machine-actionable

Research Article

Predictability of Yen Angle & Appraisal of Various Cephalometric Parameters in the Assessment of Sagittal Relationship between Maxilla and Mandible in Angle's Class II Malocclusion

Jigar R. Doshi, Kalyani Trivedi, Tarulatha Shyagali,

Use Case 2: Apply the FAIR Semantic Publishing of Modern Scientific Digital Libraries for rTMS

- The Open Research Knowledge is an exemplar of a modern scientific digital library.

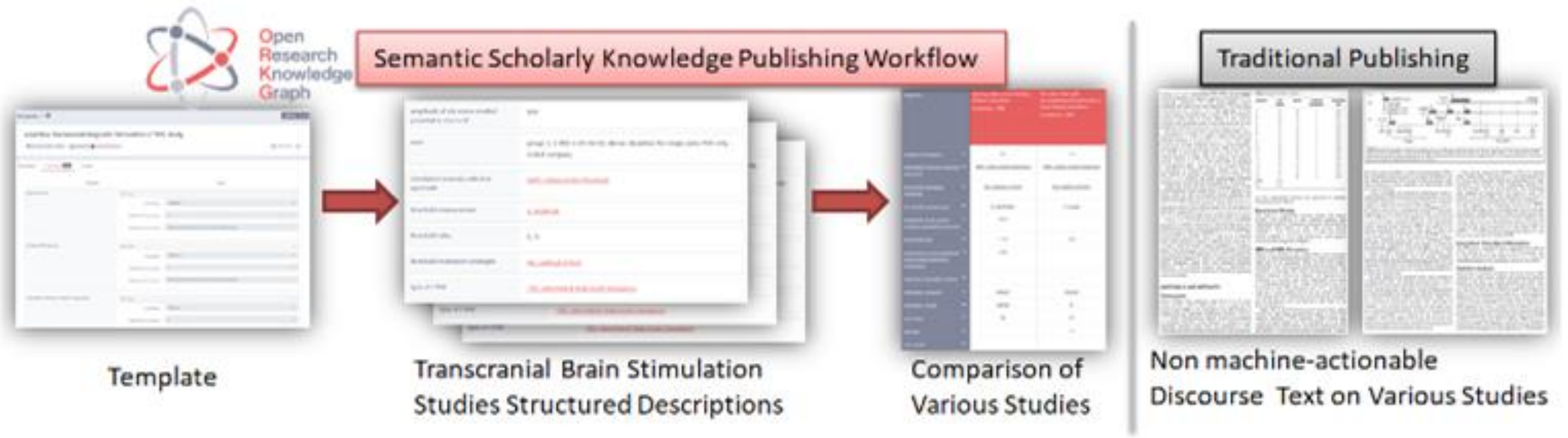
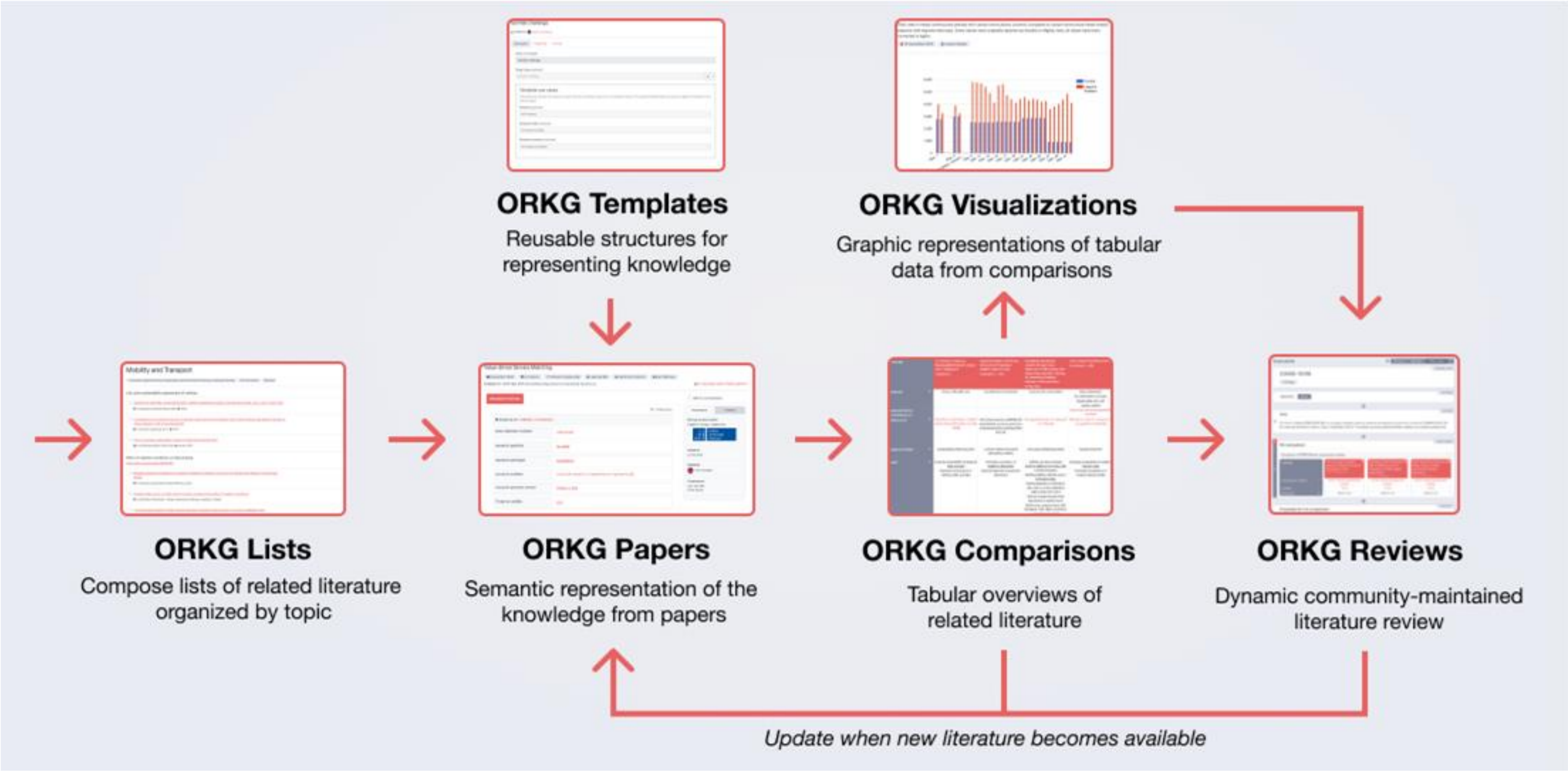


Fig. 1. Machine-actionable structured scholarly knowledge capture via semantic publishing (in red) versus traditional discourse-based non-machine-actionable publishing (in gray).

References

- Anil, S., & D'Souza, J. (2023, November). Toward Semantic Publishing in Non-invasive Brain Stimulation: A Comprehensive Analysis of rTMS Studies. In *International Conference on Asian Digital Libraries* (pp. 141-151). Singapore: Springer Nature Singapore.
- ICADL presentation: https://docs.google.com/presentation/d/1Nhonz5Eqq5FFas4Ugt8VqD8vnGEU7bMS7UcC4o4pULM/edit?usp=sharing

ORKG Platform - A FAIR Scholarly Knowledge Publishing Platform



LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

Bundesministerium
für Bildung
und Forschung

TIB

SCINEXT: Neural-SymboliC InnovatioN EXTraction

Funded by German Federal Ministry of Education and Research, BMBF
Föderkennzeichen: 01IS22070

Presented at: ESSAI Summer School
Date of event: 24th July 2024
Presented by: Jennifer D'Souza (Junior AI Research Group Lead)
PhD Student Contributors: Hamed Babaei Giglou, Mahsa Shamsabadi and Julia Evans

Leibniz
Association

SCINEXT

- The SCINEXT research group aims to setup AI services following Neural-Symbolic methods for SCholarly InnovatioN EXTraction
  - optimal mix of AI neural and symbolic approaches to automatically mine scholarly articles' contributions in a structured manner to augment and scale the Open Research Knowledge Graph (ORKG).

References
https://scinext-project.github.io/
Auer et al. (2020). Improving Access to Scientific Literature with Knowledge Graphs. *Bibliothek Forschung und Praxis*, vol. 44, no. 3, pp. 516-529 https://doi.org/10.1515/bfp-2020-2042

SCINEXT Conceptual Overview

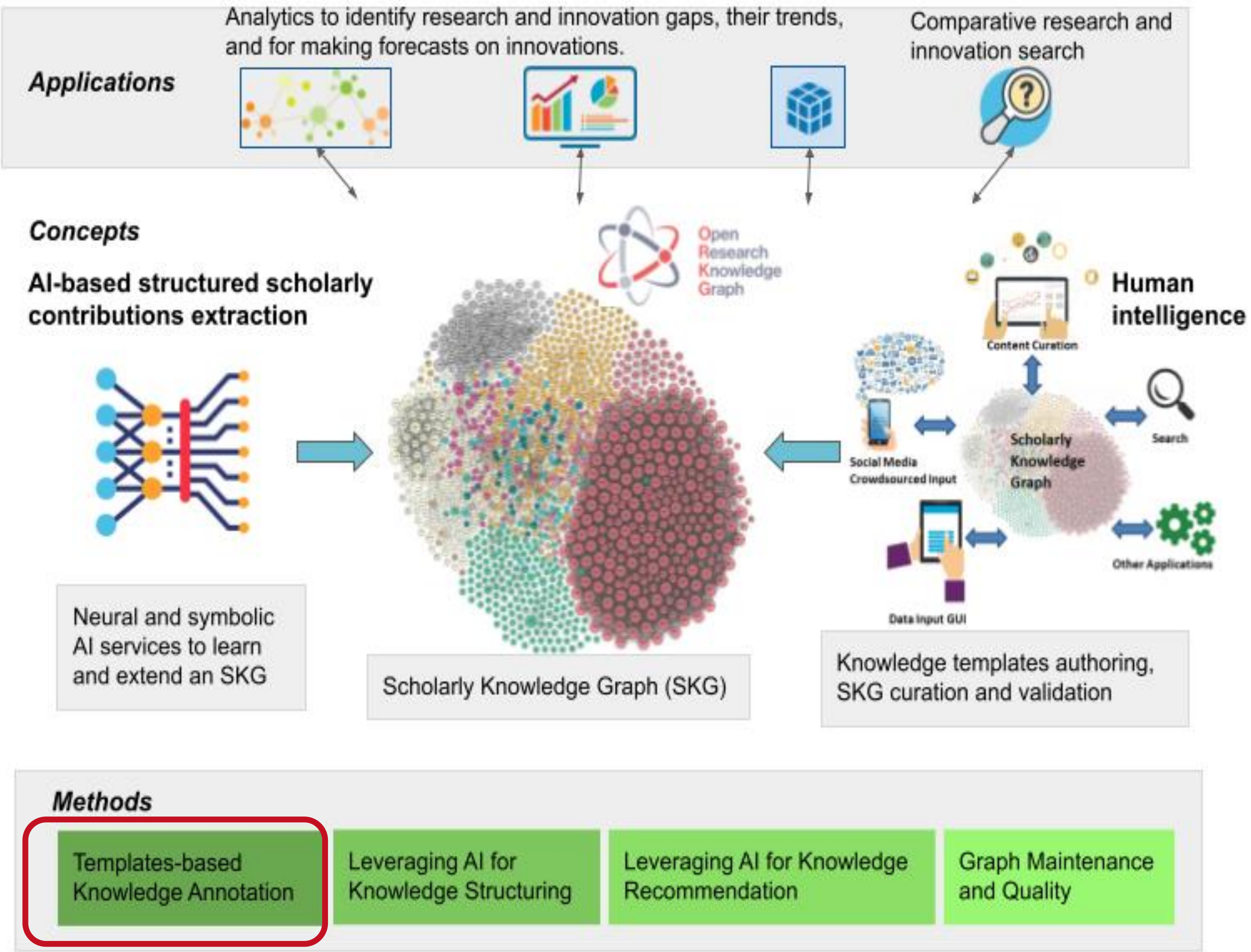


Figure 1: Conceptual view of the SCINEXT project that will synergize neural and symbolic approaches for automated scholarly contributions' knowledge extraction from scholarly articles.

The core modules of the SCINEXT project are:

- Structured **Knowledge Annotation** for the ORKG;

SCINEXT Conceptual Overview

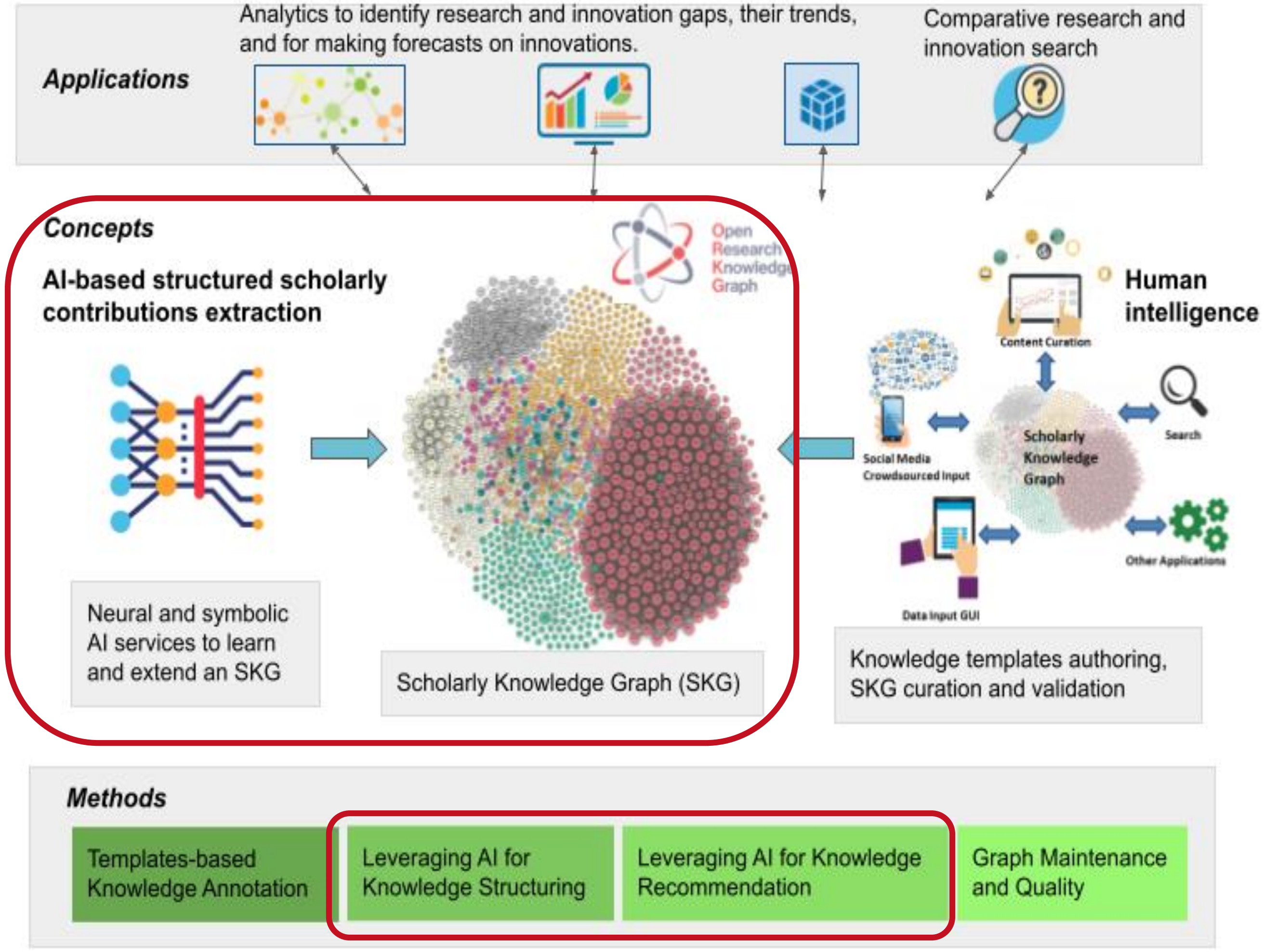


Figure 1: Conceptual view of the SCINEXT project that will synergize neural and symbolic approaches for automated scholarly contributions' knowledge extraction from scholarly articles.

The core modules of the SCINEXT project are:

- Structured **<u>Knowledge Annotation</u>** for the ORKG;
- **<u>Implementing AI learners</u>** over the annotated data for human-in-the-loop Scholarly Knowledge Structuring and Recommendation;

SCINEXT Conceptual Overview

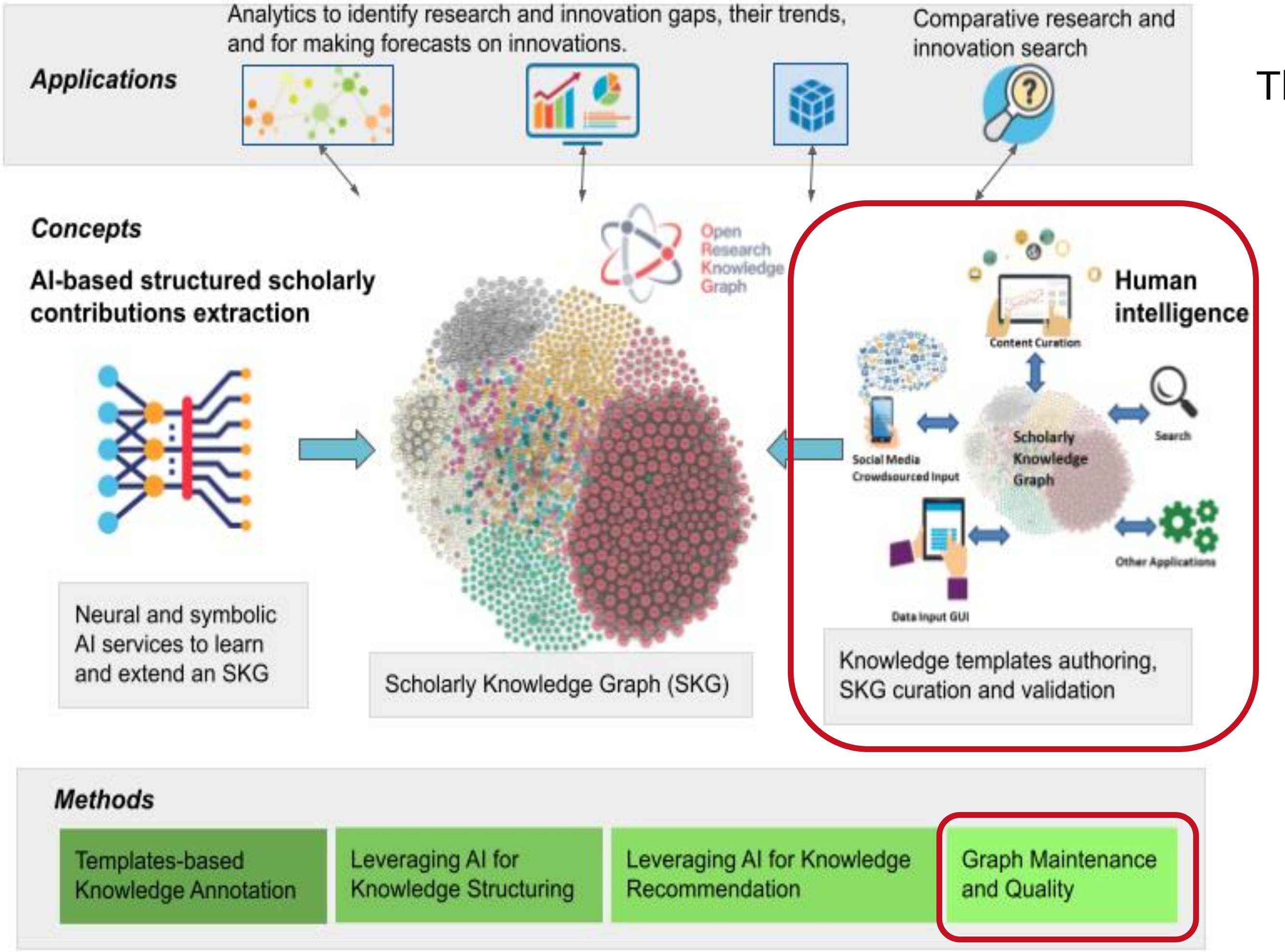
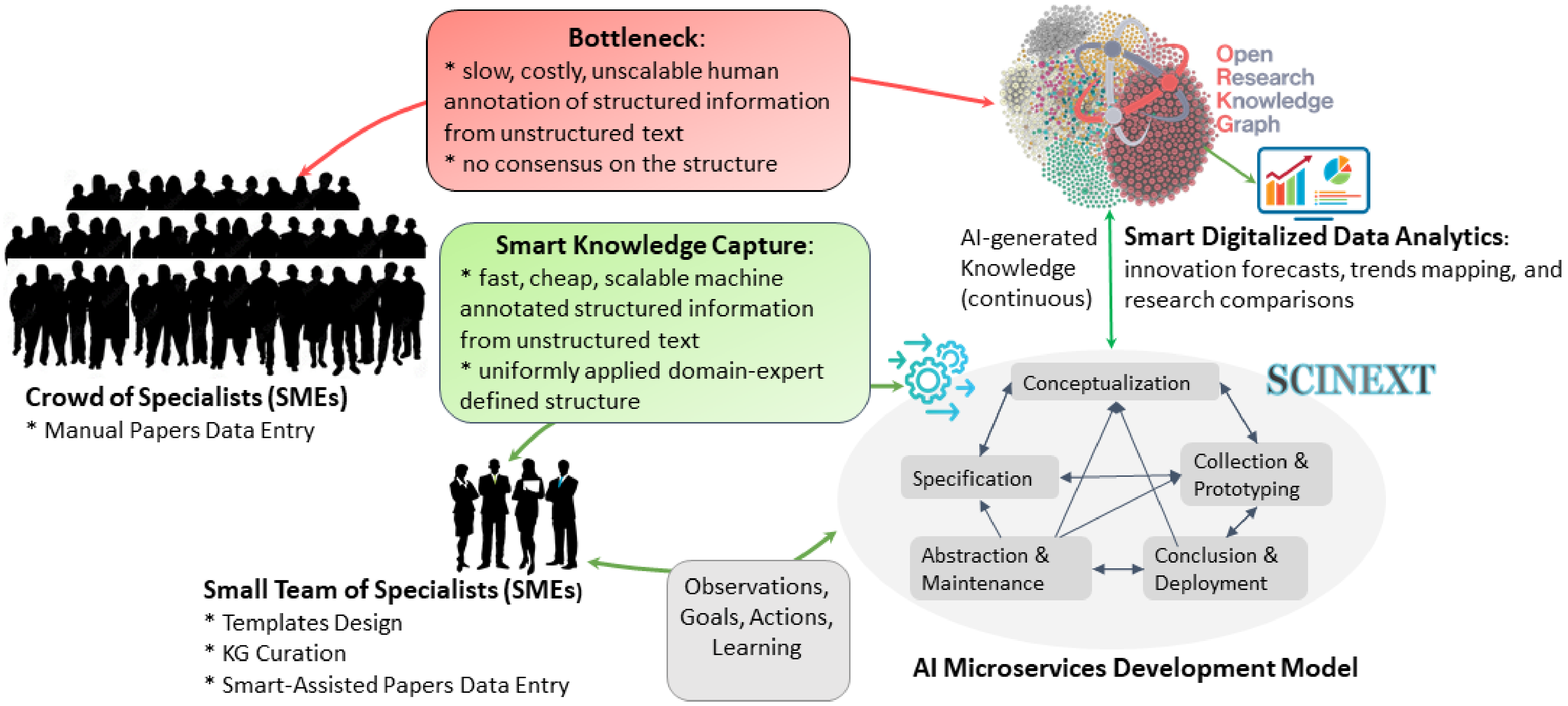


Figure 1: Conceptual view of the SCINEXT project that will synergize neural and symbolic approaches for automated scholarly contributions' knowledge extraction from scholarly articles.

The core modules of the SCINEXT project are:

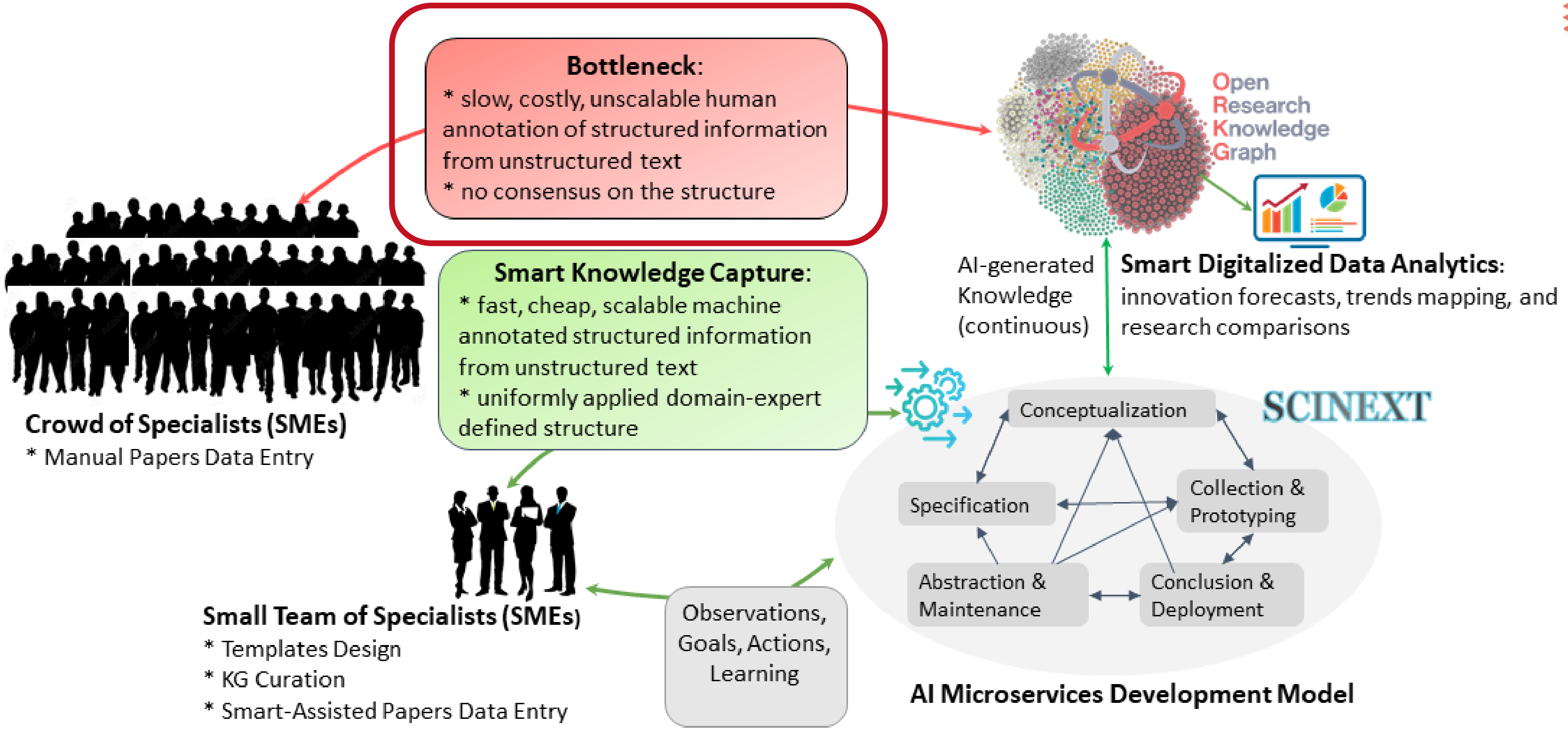
- Structured **Knowledge Annotation** for the ORKG;
- **Implementing AI learners** over the annotated data for human-in-the-loop Scholarly Knowledge Structuring and Recommendation; and
- Graph **Quality Evaluation**

Why AI-powered Research Assistance in the context of the ORKG?



Time to scale

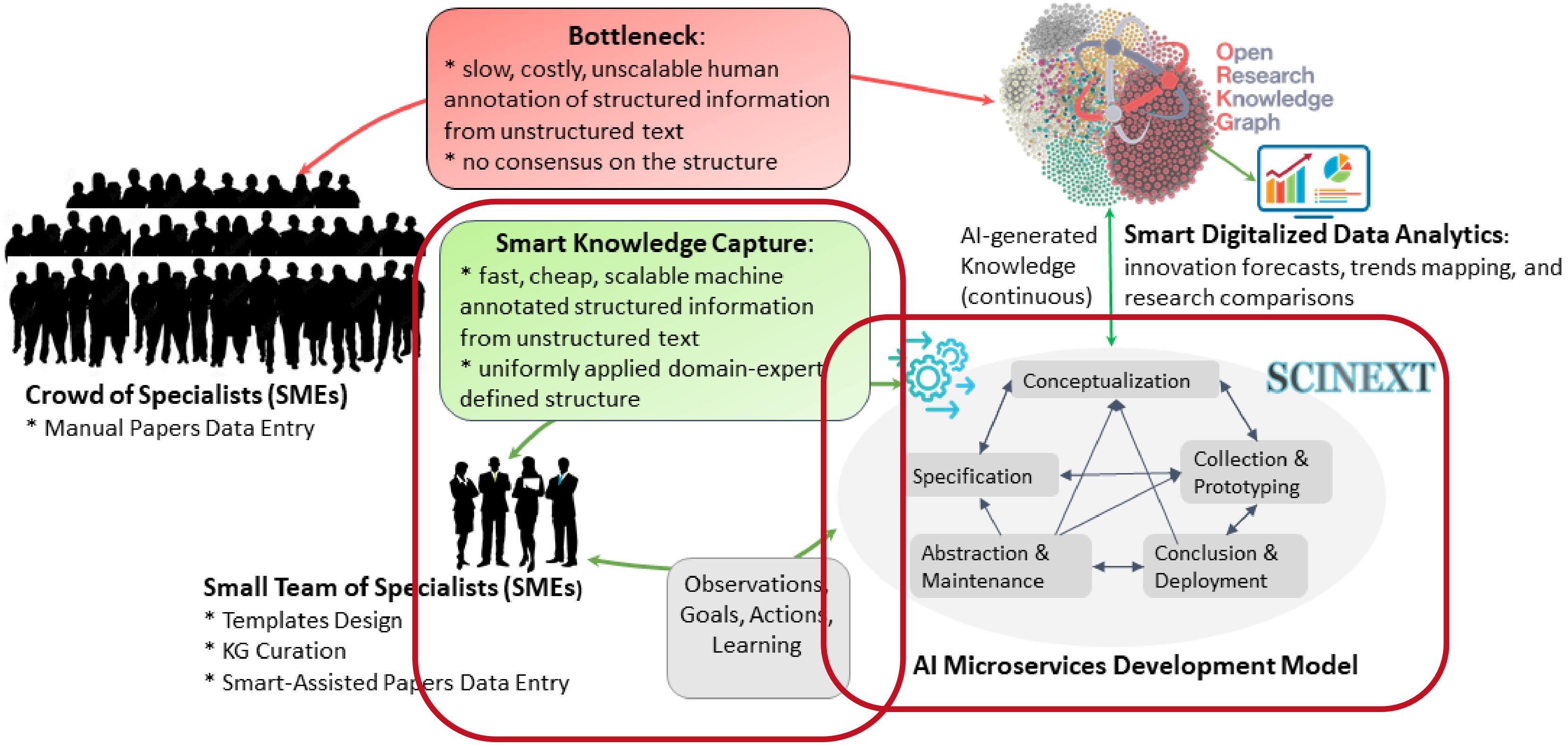
Why AI-powered Research Assistance in the context of the ORKG?



Time to scale

- if structured scholarly knowledge were created only by humans, this would create a information bottleneck owing to a slow, costly, and non-uniform process.

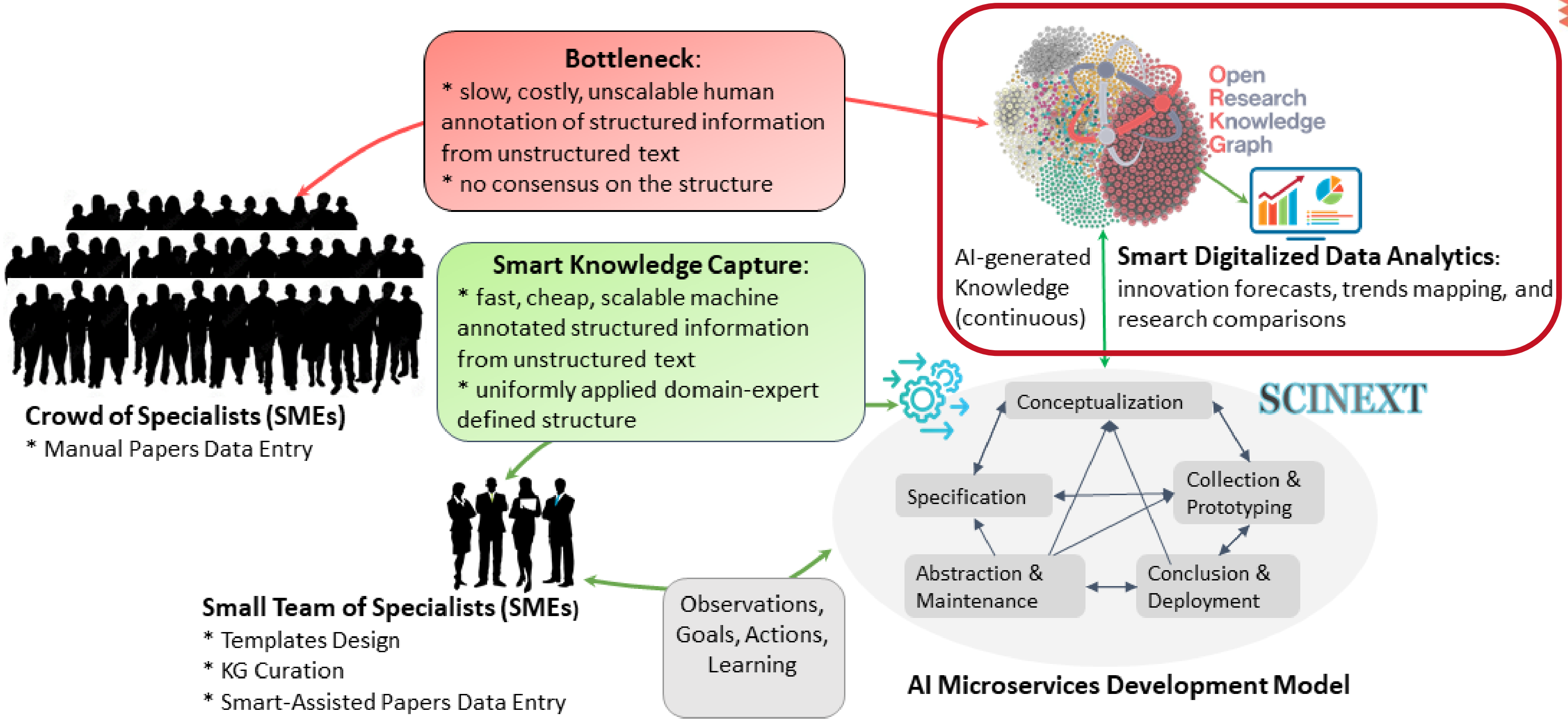
Why AI-powered Research Assistance in the context of the ORKG?



Time to scale

- if structured scholarly knowledge were created only by humans, this would create a information bottleneck owing to a slow, costly, and non-uniform process.
- instead via the proposed AI solutions from SCINEXT, acquiring structured scholarly knowledge could be significantly expedited and rely only on a small team of experts whose task would simply be to curate the knowledge

Why AI-powered Research Assistance in the context of the ORKG?



Next-generation Information Technology (IT)

- Given the large-scale structured knowledge produced by the AI services, next-generation IT solutions as visualization dashboards can be implemented providing researchers with easily comprehensible visual summaries of scholarly information, thereby assisting them in filtering for relevant research.

SCINEXT

- Reflections on some of our research:
  - A FAIR and Free Prompt-based Research Assistant | Demo paper at NLDB 2024, preprint
  - Large Language Models for Scientific Information Extraction: An Empirical Study for Virology
    - AI-powered Virology Dashboard | In: EACL 2024 Findings – paper link
  - LLMs4OL: Large Language Models for Ontology Learning
    - Babaei Giglou, H., D'Souza, J., Auer, S. (2023). LLMs4OL: Large Language Models for Ontology Learning. In: ISWC 2023. https://doi.org/10.1007/978-3-031-47240-4_22
  - LLMs4OM: Matching Ontologies with Large Language Models
    - In: ESWC 2024 Special Track on LLMs for KE | preprint
  - Large Language Models as Evaluators for Scientific Synthesis
    - In: KONVENS 2024 short paper | preprint
  - Survey on measures of quality of crowdsourced data in the Open Research Knowledge Graph (ORKG) for six different domains
    - Quality Assessment of Research Comparisons in the Open Research Knowledge Graph: a Case Study | In: JLIS 2024

References
https://scinext-project.github.io/

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
    - Research Assistant
    - An AI-powered Virology Dashboard
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
    - LLMs for evaluating Scientific Synthesis

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
    - **Research Assistant**
    - An AI-powered Virology Dashboard
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
    - LLMs for evaluating Scientific Synthesis

- Many of us use the free ChatGPT research interface (and more recently Gemini) as an assistant for our tasks. The tasks can be as varied as are our particular needs in handling natural language.

- A concrete example could be summarization of long texts. The ChatGPT agent via a prompt instruction can be fed an input text and be instructed to generate a summary of the input with as detailed specification as the number of sentences within which to restrict the summary to.
  - It could be that on reading the ChatGPT generated summary we are not happy with the result. Nevertheless, just having an application or assistant as a sounding board can greatly ease our task performance.

ChatGPT 4

You

Summarize the following text in 4 sentences: "Extracting complex knowledge from unstructured sources is a challenge: in the industrial domain, for example, troubleshooting documents may contain the description of long and articulated procedures (i.e., sequences of steps to be performed in a precise order and under specific conditions) and those natural language instructions may be represented in very different textual forms, thus making it hard for a knowledge extraction algorithm to correctly identify and structure the relevant information. Oftentimes, automatic extraction is followed by manual revision of domain experts. In any case, all machine-learning-based methods require training data which is often not readily available, therefore novel approaches are emerging to exploit interactive dialogues and language models~\cite{bellan2021process}.

Extracting procedural knowledge from human natural language instructions is a challenging task. Firstly, natural language instructions are not interpretable by machines. In the easiest case, the instructions are given as numbered lists which can easily be identified. However, complications arise when the document contains procedures in different forms: a list without numbers, an indented text or simply a full text in which the different steps are connected by conjunctions like "then", "afterwards", etc. Secondly, procedures can either be composed of only simple steps or contain other sub-procedures that are located elsewhere in the document. Thirdly, the procedures can differ substantially from one document to the other because of different authors' and editors' styles, but the goal would be to integrate information from different documents. Specifically, we investigate the potential of large language

...

Our code and dataset are publicly released \url{https://github.com/jd-coderepos/proc-tm/}."

‹ 2 / 2 ›

ChatGPT

The text discusses the challenges and methodologies involved in extracting procedural knowledge from unstructured sources, highlighting the difficulties posed by the diverse formats of natural language instructions and the limitations of machine learning algorithms in processing such information. It explores the use of large language models (LLMs), particularly GPT-4, for procedural text mining, detailing an incremental question-answering approach that incorporates both zero-shot and in-context learning scenarios to improve extraction accuracy. The research focuses on enhancing LLMs' ability to interpret and structure complex procedural information through tailored in-context learning modifications, addressing the scarcity of training data which is a significant obstacle in the field. The study contributes to the broader field of natural language processing (NLP) by offering insights into the capabilities of in-context learning to overcome data limitations, thereby advancing the development of more sophisticated information retrieval systems, with the code and dataset made publicly available.

Message ChatGPT...

Open Research Knowledge Graph + ChatGPT or Gemini

- Many of us use the free ChatGPT research interface (and more recently Gemini) as an assistant for our tasks. The tasks can be as varied as are our particular needs in handling natural language.

- A concrete example could be summarization of long texts. The ChatGPT agent via a prompt instruction can be fed an input text and be instructed to generate a summary of the input with as detailed specification as the number of sentences within which to restrict the summary to.
  - It could be that on reading the ChatGPT generated summary we are not happy with the result. Nevertheless, just having an application or assistant as a sounding board can greatly ease our task performance.

Open Research Knowledge Graph + ChatGPT or Gemini

- Many of us use the free ChatGPT research interface (and more recently Gemini) as an assistant for our tasks. The tasks can be as varied as are our particular needs in handling natural language.

- A concrete example could be summarization of long texts. The ChatGPT agent via a prompt instruction can be fed an input text and be instructed to generate a summary of the input with as detailed specification as the number of sentences within which to restrict the summary to.
    - It could be that on reading the ChatGPT generated summary we are not happy with the result. Nevertheless, just having an application or assistant as a sounding board can greatly ease our task performance.

- Given this motivation, we defined a tool that comprised specific scenarios where the free ChatGPT/Gemini chat interface can be used to assist researchers in two main ways: 1) primary objectives – designed to produce FAIR research comparisons for potential import into the ORKG; and 2) secondary objectives – addressing tasks like automatically generating blog posts, research ideas, writing assistance based on prespecified research properties pertaining to specific research problems.

ChatGPT-ORKG Objective 1: ORKG Papers

View paper

Graph view | Edit

The early phase of the COVID-19 outbreak in Lombardy, Italy

2020 | Virology | Cereda D | Tirani M | Rovida F | Demicheli | Ajelli M | Poletti P | Trentini F | Guzzetta G | Marziano | Barone A | Magoni M | Deandrea S | Diurno G | Lombardo M | Faccini M | Pan A | Bruno R | Pariani E | Grasselli G | Piatti A | Gramegna M | Baldanti F | Melegaro A | Merler S

Published in: *arXiv.org*

Share this paper:

Contribution 1

Research problems

Add to comparison

COVID-19 reproductive number

Contribution data

95% Confidence interval	2.9-3.2
Location	Lombardy, Italy
R0 estimates (average)	3.1
Study date	2020-01-14/2020-03-08

ChatGPT-ORKG Objective 1: ORKG Papers

View paper

Graph view | Edit

The early phase of the COVID-19 outbreak in Lombardy, Italy

2020 | Virology | Cereda D | Tirani M | Rovida F | Demicheli | Ajelli M | Poletti P | Trentini F | Guzzetta G | Marziano | Barone A | Magoni M | Deandrea S | Diurno G | Lombardo M | Faccini M | Pan A | Bruno R | Pariani E | Grasselli G | Piatti A | Gramegna M | Baldanti F | Melegaro A | Merler S

Published in: *arXiv.org*

Share this paper:

Contribution 1

Research problems

Add to comparison

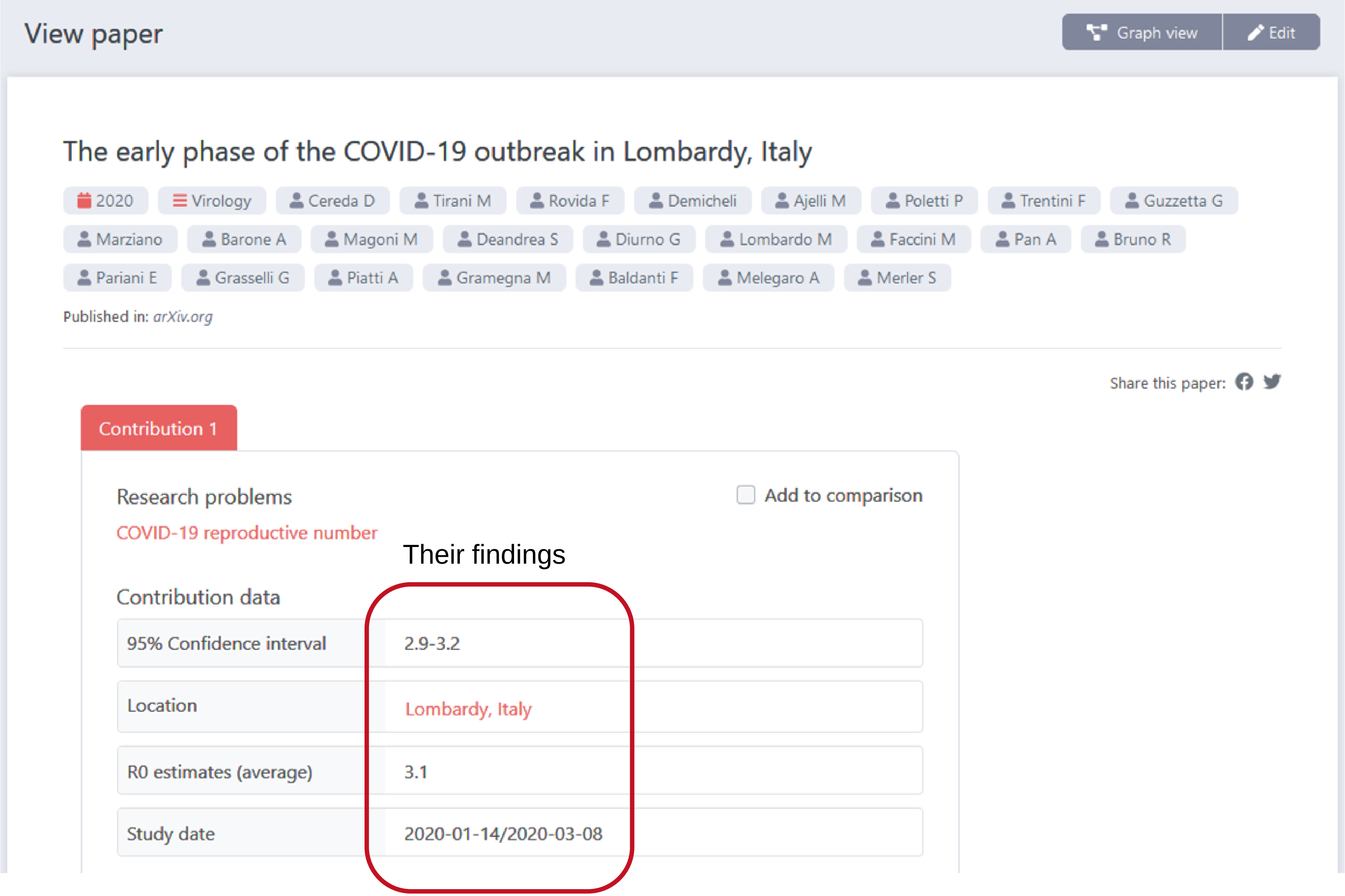
COVID-19 reproductive number

Salient properties of the research

Contribution data

95% Confidence interval	2.9-3.2
Location	Lombardy, Italy
R0 estimates (average)	3.1
Study date	2020-01-14/2020-03-08

ChatGPT-ORKG Objective 1: ORKG Papers



ChatGPT-ORKG Objective 1, Usage 1

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

ChatGPT-ORKG Objective 1, Usage 1

- **Usage 1: get recommendations for salient properties**
    - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Prompt: Can you provide a list of properties that encapsulate the “[insert research problem]” research theme? Provide your response as a Table with 2 columns: the first column contains the property name and the second column contains the property description.

Prompt Example: Can you provide a list of properties that encapsulate the "R0 estimates for infectious diseases" research theme? Provide your response as a Table with 2 columns: the first column contains the property name and the second column contains the property description.

Response: https://chat.openai.com/share/f27037ec-bccc-4238-85b5-fc693c559283

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties	**ORKG Modeled Properties**
Reproduction Number (R0)	Research Problem
Disease Transmission Dynamics	R0 estimates (average)
Mathematical Modeling	95% CI
Epidemiological Studies	Study date
Variability and Uncertainty	Location
Comparisons and Meta-Analyses	Method
Impact of Interventions	
Emerging and Reemerging Infectious Diseases	
Spatial and Temporal Patterns	
Public Health Implications	

(An arrow links "R0 estimates (average)" to "Reproduction Number (R0)".)

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties	ORKG Modeled Properties
Reproduction Number (R0)	Research Problem
Disease Transmission Dynamics	R0 estimates (average)
Mathematical Modeling	95% CI
Epidemiological Studies	Study date
Variability and Uncertainty	Location
Comparisons and Meta-Analyses	Method
Impact of Interventions	
Emerging and Reemerging Infectious Diseases	
Spatial and Temporal Patterns	
Public Health Implications	

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location
Method

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location
Method

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location
Method

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties		ORKG Modeled Properties
Reproduction Number (R0)		Research Problem
Disease Transmission Dynamics		R0 estimates (average)
Mathematical Modeling		95% CI
Epidemiological Studies		Study date
Variability and Uncertainty	subsumed by R0 value	Location
Comparisons and Meta-Analyses		Method
Impact of Interventions		
Emerging and Reemerging Infectious Diseases		
Spatial and Temporal Patterns		
Public Health Implications		

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location
Method

These are not necessarily indicative of the “R0 estimate” research problem. However they are a candidate for modeling new contributions defined as “public health interventions for infectious diseases”.

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location
Method

indicates that various different research works should be compared which consequently offers an overview on the R0 estimate for various populations. This is addressed via ORKG Comparisons.

ChatGPT-ORKG Objective 1, Usage 1: Discussion

- **Usage 1: get recommendations for salient properties**
  - assumptions: user knows the research problem in advance, the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Suggested Properties
Reproduction Number (R0)
Disease Transmission Dynamics
Mathematical Modeling
Epidemiological Studies
Variability and Uncertainty
Comparisons and Meta-Analyses
Impact of Interventions
Emerging and Reemerging Infectious Diseases
Spatial and Temporal Patterns
Public Health Implications

ORKG Modeled Properties
Research Problem
R0 estimates (average)
95% CI
Study date
Location

User action: analyze and distil the suggested properties to their optimal model as reflected by the ORKG Modeled Properties

ChatGPT-ORKG Objective 1, Usage 2

- **Usage 2: get recommendations for salient properties based on provided context**
  - assumptions: user knows the research problem in advance; ChatGPT is expected to offer suggestions from the provided context.

Prompt: Can you provide a list of properties that encapsulate the [“insert research problem”] research theme from the provided Context below? Provide your response as a Table with 2 columns: the first column contains the property name and the second column contains the property description.

Context: [Insert text]

Prompt Example: Can you provide a list of properties that encapsulate the “Covid 19 basic reproduction number estimate” research theme from the provided Context below? Provide your response as a Table with 2 columns: the first column contains the property name and the second column contains the property description.

Context: “The early phase of the COVID-19 outbreak in Lombardy, Italy

In the night of February 20, 2020, the first case of novel coronavirus disease (COVID-19) was confirmed in the Lombardy Region, Italy. In the week that followed, Lombardy experienced a very rapid increase in the number of cases …”

Objective 2: ORKG Comparisons

Contribution comparison 3

View | Add contribution | More

Covid-19 Reproductive Number Estimates

Method: Intelligent merge

Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy Contribution 1 - 2020	Transmission potential of COVID-19 in Iran Contribution 1 - 2020	Transmission potential of COVID-19 in Iran Contribution 2 - 2020
Has value	3.1	3.6	3.58
Location	Lombardy, Italy	Iran	Iran
Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)
Lower confidence limit	2.9	3.4	1.29
Upper confidence limit	3.2	4.2	8.46
Has beginning	2020-01-14	2020-02-19	2020-02-19
Has end	2020-03-08	2020-02-29	2020-02-29

ChatGPT-ORKG Objective 2, Usage 1

- **Usage 1: get comparisons**
  - assumptions: user knows the research problem/theme in advance and wants suggestions for properties to compare the different specified entities on the same theme; the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Prompt: Generate a property-value-based Comparison that encapsulates the "[insert research theme]" research theme [insert comparison entities]. Provide your response as a Table: the first column contains the property name and the subsequent columns contain the property value for the respective comparison entities.

Prompt Example 1: Generate a property-value-based Comparison that encapsulates the "Covid 19 basic reproduction number estimate" research theme in Italy, China, and Africa. Provide your response as a Table: the first column contains the property name and the subsequent columns contain the property value.

Response: https://chat.openai.com/share/00aa9985-f6f9-4cd3-b2c5-b68d4349a273

ChatGPT-ORKG Objective 2, Usage 1

- **Usage 1: get comparisons**
  - assumptions: user knows the research problem/theme in advance and wants suggestions for properties to compare the different specified entities on the same theme; the research problem is well-known so that ChatGPT can be expected to offer suggestions from its own knowledge.

Prompt: Generate a property-value-based Comparison that encapsulates the "[insert research theme]" research theme [insert comparison entities]. Provide your response as a Table: the first column contains the property name and the subsequent columns contain the property value for the respective comparison entities.

Prompt Example 2: Generate a property-value-based Comparison that encapsulates the "Large Language Models" research theme for T5 and GPT-3. Provide your response as a Table: the first column contains the property name and the subsequent columns contain the property value.

Response: https://chat.openai.com/share/5d477232-4211-4022-84fb-f2a2c5104da5

ChatGPT-ORKG Objective 2, Usage 2

- **Usage 2: get comparisons based on provided context**
  - assumptions: user knows the research problem/theme in advance and wants suggestions for properties to compare different works on the same theme; ChatGPT is expected to offer suggestions from the provided context.

Prompt: Generate a property-value-based Comparison that encapsulates the "[insert research theme]" research theme from the provided Contexts below. Provide your response as a Table: the first column contains the property name and the subsequent columns contain the property value.

Context 1
text text text …
Context 2
text text text …
Context 3
text text text …

The FAIR and free Prompt-based Research Assistant

Reference

Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

The FAIR and free Prompt-based Research Assistant: Motivation

- Our tool (React application) acts as a bridge between conversational AI agents (e.g ChatGPT) and next-generation digital libraries.
- It facilitates generating data for the next-generation libraries by using conversational AI agents like ChatGPT and Gemini.
- It assists end users to access relevant AI-generated data to address different kinds of tasks in research.

Reference

Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

@ChatGPT

what is research assistance?

Research assistance refers to the support provided to individuals or groups engaged in research activities. This support can come in various forms, depending on the context and needs of the researchers. Here are some common types of research assistance:

1. **Literature Review**: Assistance in searching for, identifying, and summarizing relevant literature and previous research studies to establish a solid foundation for new research.
2. **Data Collection and Analysis**: Help with designing surveys, experiments, or observational studies, collecting data, and using statistical or analytical tools to interpret the results.
3. **Technical Support**: Providing expertise in specialized software, tools, or methodologies necessary for conducting specific types of research, such as GIS mapping, laboratory techniques, or computer programming.
4. **Writing and Editing**: Aid in structuring, writing, and editing research papers, proposals, and reports to meet academic standards and publication criteria.
5. **Ethical and Legal Guidance**: Offering advice on ethical issues and regulatory compliance, particularly in fields like medical, psychological, or sociological research.
6. **Funding and Resource Allocation**: Assistance in identifying potential funding sources, writing grant proposals, and managing research budgets.
7. **Mentorship and Training**: Experienced researchers mentoring less experienced ones, providing guidance, and training on various aspects of conducting research.

Programmed as a primary task in RA to support creating ORKG comparisons as a form of literature review

Designed as a secondary task in RA to help with writing reviews, grants, blog posts

Designed as a secondary task in RA to help with suggesting research ideas to work on

Reference
https://chatgpt.com/share/c28c1770-3048-4c57-b1eb-72ca10527fbf

Research Assistant: Primary Task

- Generating FAIR research comparisons.

  1st Step

  - Obtaining salient properties for a research problem
  - Refining definitions of salient properties based on provided context

  Continuation Step

  - Comparing entities for a research problem
  - Comparing contexts for a research problem
  - Comparing contexts based on a set of salient properties for a research problem

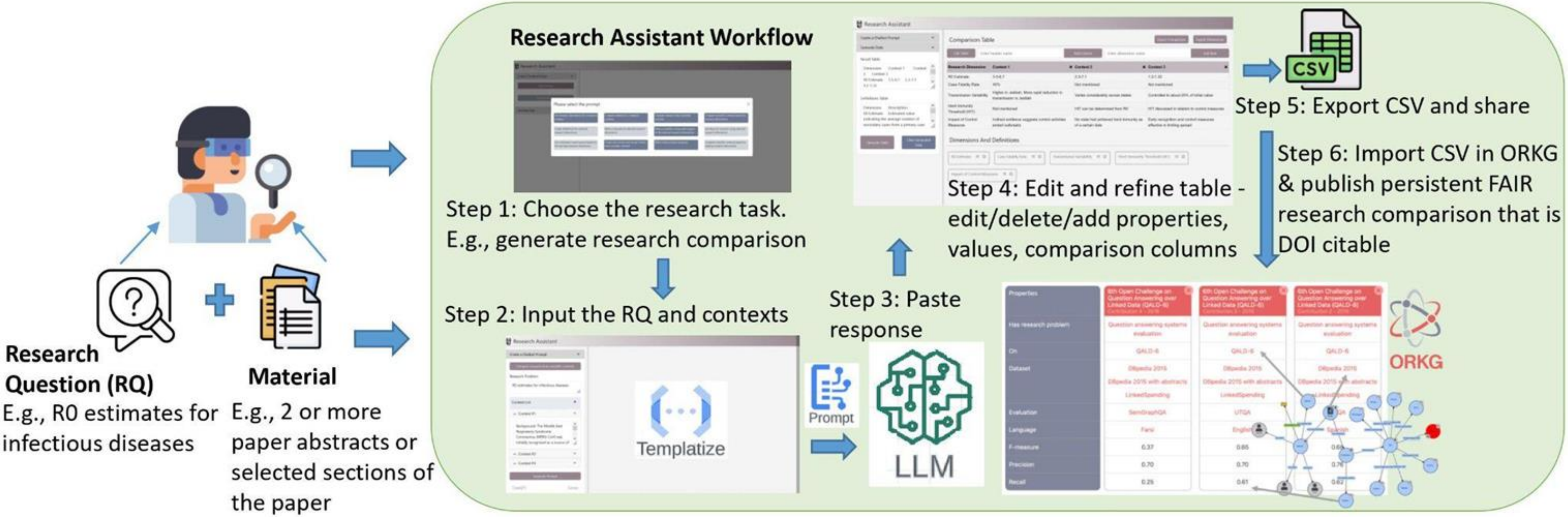
Reference
Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

Research Assistant: Secondary Tasks

- Brainstorming research ideas
  - Receiving suggestions for research ideas based on a problem and its properties
  - Obtaining user stories and criteria based on a context
- Writing grant applications
- Generating blog-posts
- Writing a preliminary review
- Consolidating keyword-based search queries with a list of synonyms

Reference
Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

Research Assistant: Workflow



Reference

Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

Research Assistant: Primary Task

- Generating FAIR research comparisons.

  1st Step

  - Obtaining salient properties for a research problem
  - Refining definitions of salient properties based on provided context

  Continuation Step

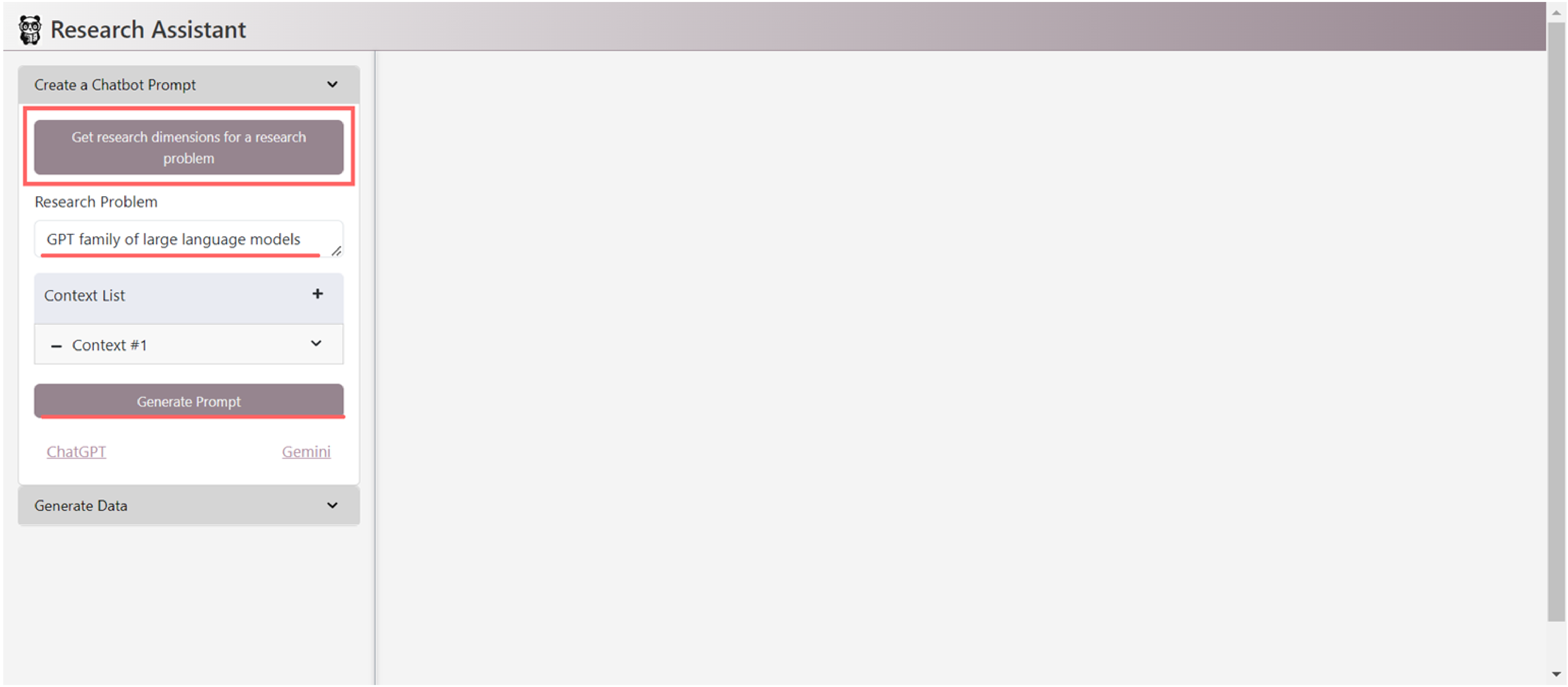
  - Comparing entities for a research problem
  - Comparing contexts for a research problem
  - Comparing contexts based on a set of salient properties for a research problem

Reference

Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

Primary Task: suggest salient properties for a research problem

Primary Task: suggest salient properties for a research problem

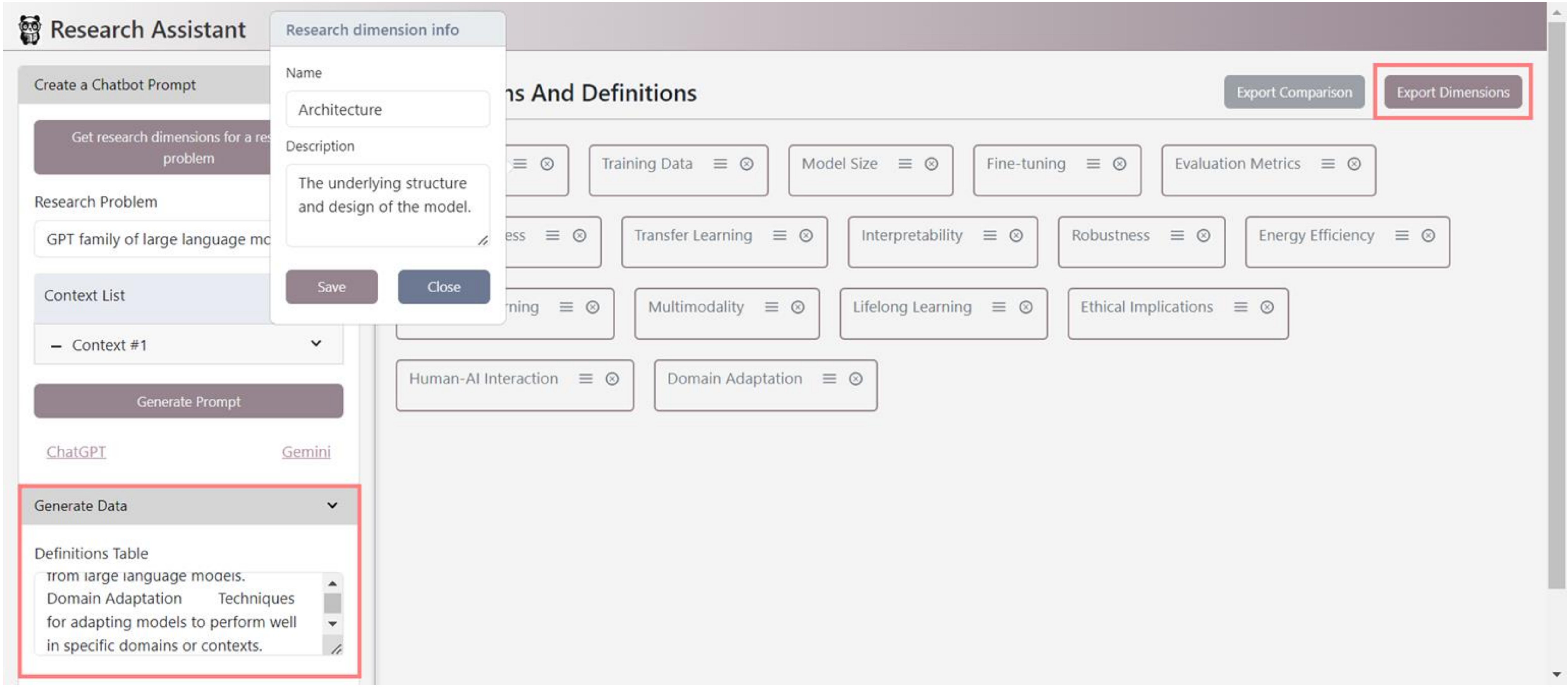


Primary Task: suggest salient properties for a research problem

Can you provide a list of research dimensions relevant to the "GPT family of large language models" research problem? Provide your response as a Table with 2 columns: the first column contains the research dimension name and the second column contains the dimension description. Note the research dimension name must be a short phrase between 1 to 3 tokens.

Research Dimension	Description
Architecture	The underlying structure and design of the model.
Training Data	The type, quality, and quantity of data used for training.
Model Size	The number of parameters and layers in the model.
Fine-tuning	Techniques for adapting pre-trained models to specific tasks or domains.
Evaluation Metrics	Metrics used to assess the performance of the model.
Bias and Fairness	Examination of biases present in the model and ensuring fairness in its application.
Transfer Learning	Studying how knowledge learned in one task can be applied to another.
Interpretability	Methods for understanding and explaining model predictions.
Robustness	Ensuring the model's performance in diverse conditions and against adversarial attacks.
Energy Efficiency	Strategies for optimizing model inference and reducing energy consumption.
Continual Learning	Techniques for the model to adapt and learn from new data over time.
Multimodality	Extending models to handle inputs and outputs of multiple modalities (e.g., text, images, audio).
Lifelong Learning	Strategies for the model to accumulate knowledge and improve over its lifetime.
Ethical Implications	Examination of the ethical considerations surrounding the use and development of large language models.
Human-AI Interaction	Studying how humans interact with and benefit from large language models.
Domain Adaptation	Techniques for adapting models to perform well in specific domains or contexts.

Primary Task: suggest salient properties for a research problem



Research Assistant: Primary Task

- Generating FAIR research comparisons.

  1st Step

  - Obtaining salient properties for a research problem
  - Refining definitions of salient properties based on provided context

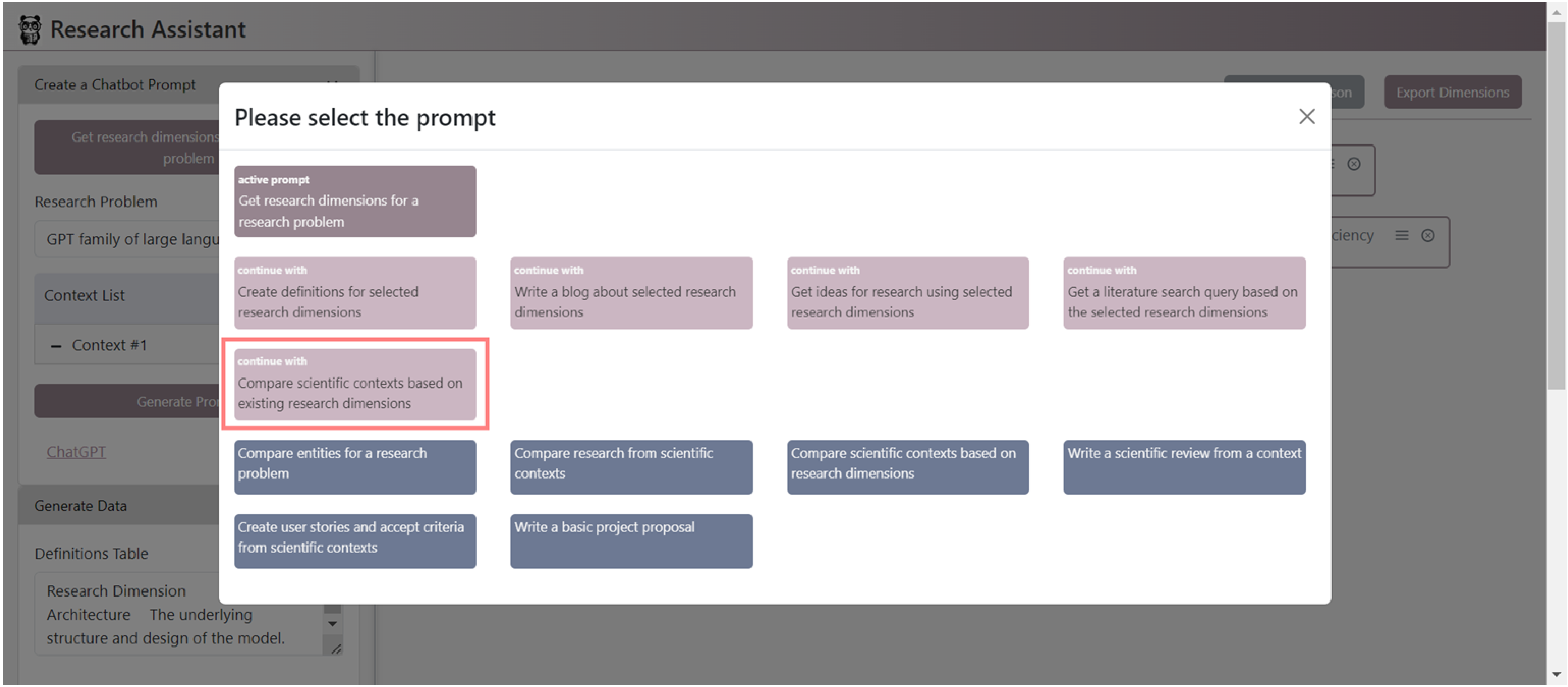
  Continuation Step

  - Comparing entities for a research problem
  - Comparing contexts for a research problem
  - Comparing contexts based on a set of salient properties for a research problem

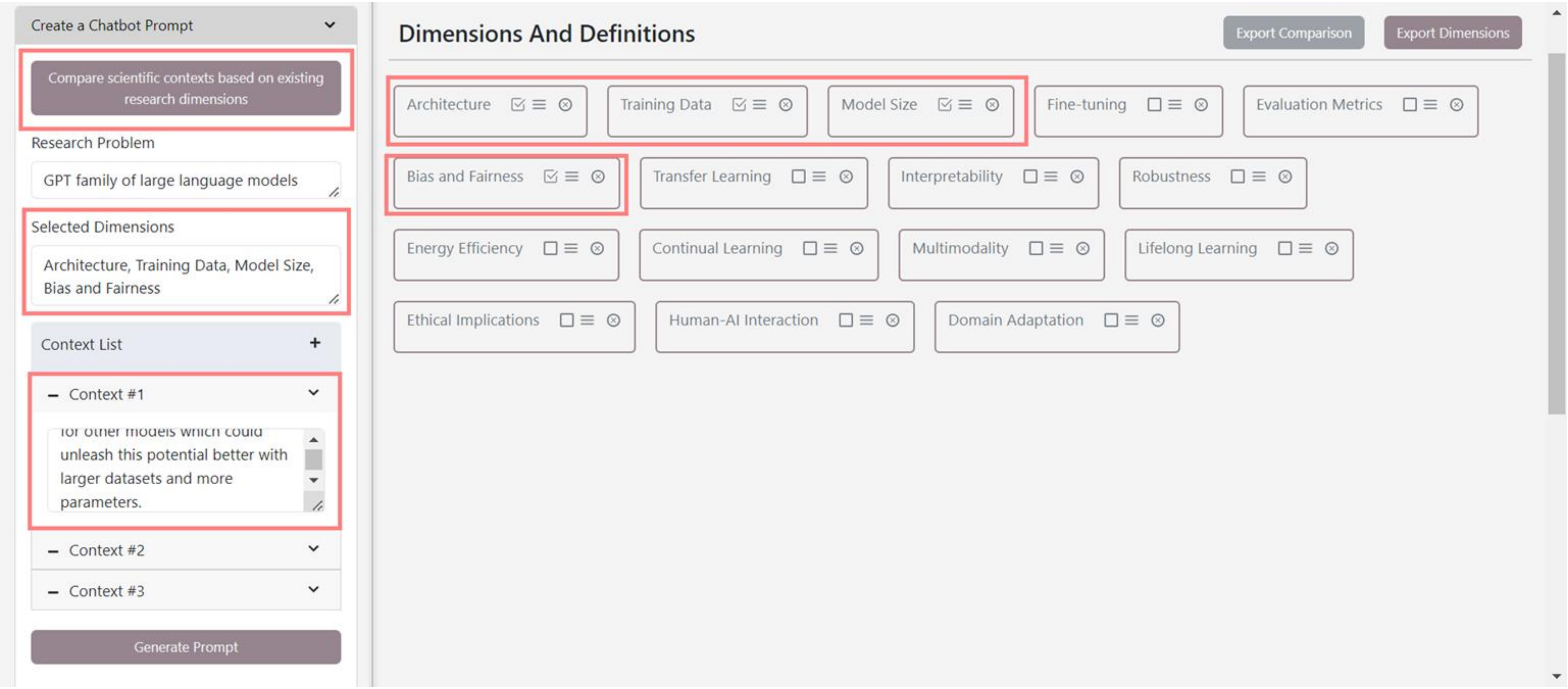
Reference

Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant

Primary Task: create a comparison from contexts based on a set of salient properties



Primary Task: create a comparison from contexts based on a set of salient properties



Primary Task: create a comparison from contexts based on a set of salient properties

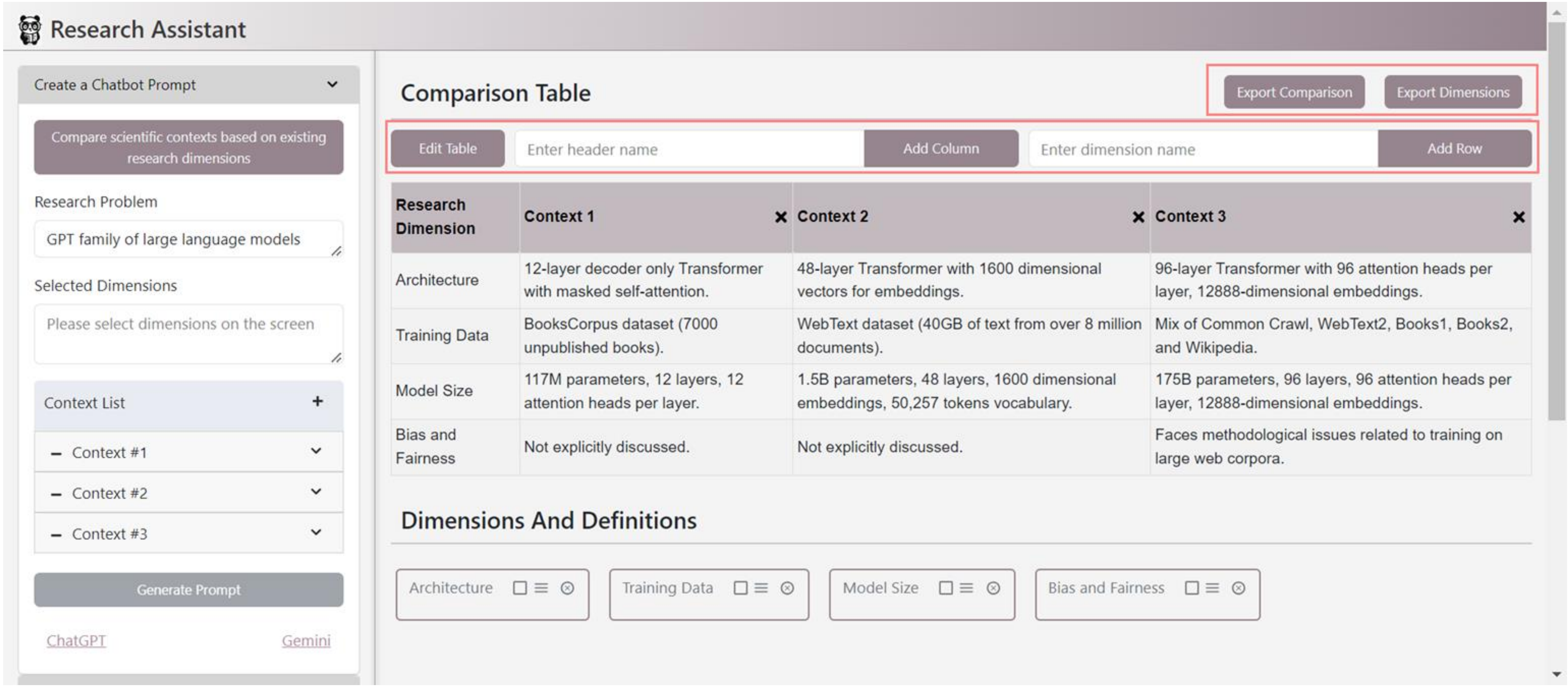
Table 1: Extracted Values Relevant to "GPT Family of Large Language Models"

Dimension	Context 1	Context 2	Context 3
Architecture	12-layer decoder only Transformer with masked self-attention.	48-layer Transformer with 1600 dimensional vectors for embeddings.	96-layer Transformer with 96 attention heads per layer, 12888-dimensional embeddings.
Training Data	BooksCorpus dataset (7000 unpublished books).	WebText dataset (40GB of text from over 8 million documents).	Mix of Common Crawl, WebText2, Books1, Books2, and Wikipedia.
Model Size	117M parameters, 12 layers, 12 attention heads per layer.	1.5B parameters, 48 layers, 1600 dimensional embeddings, 50,257 tokens vocabulary.	175B parameters, 96 layers, 96 attention heads per layer, 12888-dimensional embeddings.
Bias and Fairness	Not explicitly discussed.	Not explicitly discussed.	Faces methodological issues related to training on large web corpora.

Table 2: Research Dimension Descriptions

Dimension	Description
Architecture	The underlying structure and design of the model, including layers and attention mechanisms.
Training Data	The type, quality, and quantity of data used for training the model.
Model Size	The number of parameters and layers in the model, which impacts its capacity and performance.
Bias and Fairness	Examination of biases present in the model and ensuring fairness in its application.

Primary Task: create a comparison from contexts based on a set of salient properties

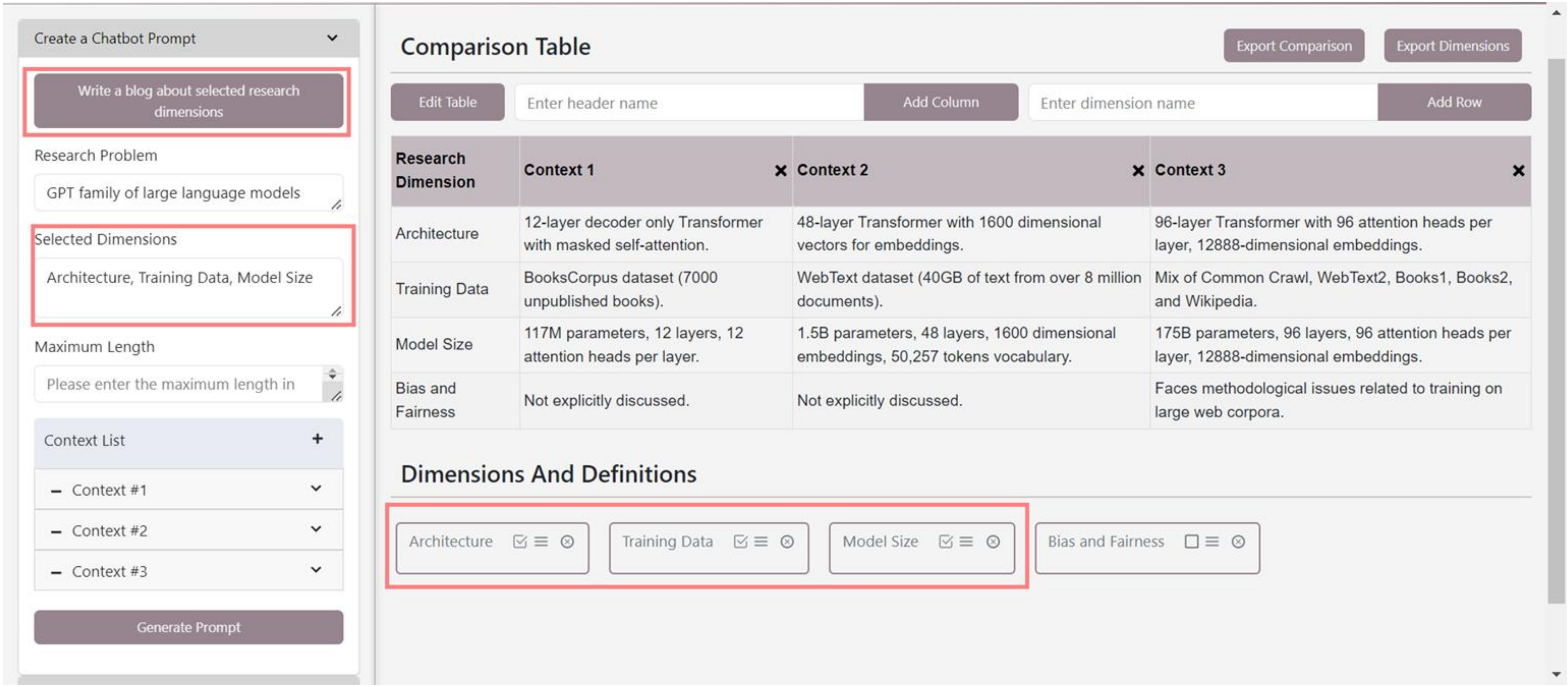


Research Dimension	Context 1	Context 2	Context 3
Architecture	12-layer decoder only Transformer with masked self-attention.	48-layer Transformer with 1600 dimensional vectors for embeddings.	96-layer Transformer with 96 attention heads per layer, 12888-dimensional embeddings.
Training Data	BooksCorpus dataset (7000 unpublished books).	WebText dataset (40GB of text from over 8 million documents).	Mix of Common Crawl, WebText2, Books1, Books2, and Wikipedia.
Model Size	117M parameters, 12 layers, 12 attention heads per layer.	1.5B parameters, 48 layers, 1600 dimensional embeddings, 50,257 tokens vocabulary.	175B parameters, 96 layers, 96 attention heads per layer, 12888-dimensional embeddings.
Bias and Fairness	Not explicitly discussed.	Not explicitly discussed.	Faces methodological issues related to training on large web corpora.

Research Assistant: Secondary Tasks

- Brainstorming research ideas
  - Receiving suggestions for research ideas based on a problem and its properties
  - Obtaining user stories and criteria based on a context
- Writing grant applications
- Generating blog-posts
- Writing a preliminary review
- Consolidating keyword-based search queries with a list of synonyms

Secondary Task: write a blog-post



Research Dimension	Context 1	Context 2	Context 3
Architecture	12-layer decoder only Transformer with masked self-attention.	48-layer Transformer with 1600 dimensional vectors for embeddings.	96-layer Transformer with 96 attention heads per layer, 12888-dimensional embeddings.
Training Data	BooksCorpus dataset (7000 unpublished books).	WebText dataset (40GB of text from over 8 million documents).	Mix of Common Crawl, WebText2, Books1, Books2, and Wikipedia.
Model Size	117M parameters, 12 layers, 12 attention heads per layer.	1.5B parameters, 48 layers, 1600 dimensional embeddings, 50,257 tokens vocabulary.	175B parameters, 96 layers, 96 attention heads per layer, 12888-dimensional embeddings.
Bias and Fairness	Not explicitly discussed.	Not explicitly discussed.	Faces methodological issues related to training on large web corpora.

Secondary Task: write a blog-post

Use the context of

....

Write a blog post focused on the research problem related to "GPT family of large language models". using only the provided defined research dimensions which are the keys in the following dictionary:

Architecture: The underlying structure and design of the model, including layers and attention mechanisms.

Training Data: The type, quality, and quantity of data used for training the model.

Model Size: The number of parameters and layers in the model, which impacts its capacity and performance.

Let's visit the result on ChatGPT: https://chatgpt.com/share/d750d275-c16d-4b22-a88d-49bee5c66ae1 (without context), https://chatgpt.com/share/bf29068c-526e-4927-9afc-6a0801b6bc3a (with context)

The FAIR and free Prompt-based Research Assistant

- Created usage scenarios in 3 different domains or research problems: “GPT family of LLMs”, “Impact of Climate Change,” and “R0 estimates in Virology”
  - https://github.com/mahsaSH717/research_assistant/tree/master/examples
- Code is publicly available, easy-to-use on your browser & lightweight installation
  - https://github.com/mahsaSH717/research_assistant
- Conclusion: Many see AI conversational agents like ChatGPT or Gemini as tools to assist with work, not replace it, already adopting them as digital secretaries or assistants (Owens, 2023). Playing into this familiar setting, engineered into RA underlying each task are a set of customised ChatGPT prompts that assists end users to access relevant AI-generated data that addresses the research task. Thus it standardizes the performance of a diverse spectrum of research tasks, in a single tool, via a modular and completely transparent workflow.

Reference

- Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. *arXiv preprint arXiv:2405.14601*. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024) | Code https://github.com/mahsaSH717/research_assistant
- Owens, B. (2023). How Nature readers are using ChatGPT. *Nature*, *615*(7950), 20. https://www.nature.com/articles/d41586-023-00500-8

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
  - Research Assistant
  - An LLM-powered Virology Dashboard
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
  - LLMs for evaluating Scientific Synthesis

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
  - Research Assistant
  - **An LLM-powered Virology Dashboard**
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
  - LLMs for evaluating Scientific Synthesis

AI-powered Virology Dashboard

- We aim to represent scholarly contributions in a structured manner, such that they are machine-actionable, for a research problem in Virology i.e. the study of the R0-estimates for infectious diseases.
  - To this end, we finetune an LLM to automatically extract the structured knowledge for new incoming papers.
- We aim to implement next-generation IT over the structured scholarly knowledge in the form of visualization dashboards that can serve as assistants to researchers in helping them easily filter for scholarly articles they seek.
  - We demonstrate this idea as a prototype web interface.

Reference

- Shamsabadi, M., D’Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Data -

Structured Scholarly Contributions about R0 estimates for Infectious Diseases

- Semantic model

Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy *Contribution 1 - 2020*	Transmission potential of COVID-19 in Iran *Contribution 1 - 2020*	Transmission potential of COVID-19 in Iran *Contribution 2 - 2020*	Estimating the generation interval for COVID-19 based on symptom onset data *Contribution 1 - 2020*
location	Lombardy, Italy	Iran	Iran	Singapore
Time period	Time interval	Time interval	Time interval	Time interval
has beginning	2020-01-14	2020-02-19	2020-02-19	2020-01-21
has end	2020-03-08	2020-02-29	2020-02-29	2020-02-26
Basic reproduction number	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification
Has value	3.1	3.6	3.58	1.27
Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)
Lower confidence limit	2.9	3.4	1.29	1.19
Upper confidence limit	3.2	4.2	8.46	1.36
Method*		generalized growth model	based on the calculation of the epidemic's doubling times: estimated epidemic doubling time of 1.20 (95% CI, 1.05, 1.44) days	generation interval

As a representation of structured scholarly knowledge, We use the R0-estimates semantic model for Covid-19. From this we obtained six properties: disease name, location, date, R0 value, %CI values, and method

Accessible here: https://orkg.org/comparison/R44930/

Data -
Structured Scholarly Contributions about R0 estimates for Infectious Diseases

- Semantic model
  - Properties: *disease name, location, date, R0 value, %CI values, and method*
- Annotate a large dataset
  - Annotated roughly 1500 papers with their structured representations from COORD-19 https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
  - Note this dataset included papers that reported r0-estimates for infectious diseases and those that may have mentioned r0-estimates but did not report it.
    - We did this because the model to be finetuned should be able to discriminate between papers for which it must create the structured representations and those that it should not, in which case it is tuned to respond “unanswerable”
- Dataset released https://zenodo.org/records/8068442

Reference

- Shamsabadi, M., D’Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

ORKG-FLAN-T5 R0 Model: Example 1

Disease name | Location | Date | R0 Value | CI Values | Method

PMID: 36483256 PMCID: PMC9723226 DOI: 10.3389/fpubh.2022.933075

Dynamic characteristics of a COVID-19 outbreak in Nanjing, Jiangsu province, China

Objectives: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) lineage B.1.617.2 (also named the Delta variant) was declared as a variant of concern by the World Health Organization (WHO). This study aimed to describe the outbreak that occurred in Nanjing city triggered by the Delta variant through the epidemiological parameters and to understand the evolving epidemiology of the Delta variant. Methods: We collected the data of all COVID-19 cases during the outbreak from 20 July 2021 to 24 August 2021 and estimated the distribution of serial interval, basic and time-dependent reproduction numbers (R0 and Rt), and household secondary attack rate (SAR). We also analyzed the cycle threshold (Ct) values of infections. Results: A total of 235 cases have been confirmed. The mean value of serial interval was estimated to be 4.79 days with the Weibull distribution. The R0 was 3.73 [95% confidence interval (CI), 2.66-5.15] as estimated by the exponential growth (EG) method. The Rt decreased from 4.36 on 20 July 2021 to below 1 on 1 August 2021 as estimated by the Bayesian approach. We estimated the household SAR as 27.35% (95% CI, 22.04-33.39%), and the median Ct value of open reading frame 1ab (ORF1ab) genes and nucleocapsid protein (N) genes as 25.25 [interquartile range (IQR), 20.53-29.50] and 23.85 (IQR, 18.70-28.70), respectively. Conclusions: The Delta variant is more aggressive and transmissible than the original virus types, so continuous non-pharmaceutical interventions are still needed.


```
{
  "contribution": {
   "disease name": "COVID-19",  "location": "Nanjing, Jiangsu province, China",  "date": "from 20 July 2021 to 24 August 2021",
   "R0 value": "3.73",  "%CI values": "[95% confidence interval (CI), 2.66-5.15]", "method": "exponential growth (EG) method"
  }
 }
```

ORKG-FLAN-T5 R0 Model: Example 2

Disease name   Location   Date   R0 Value   CI Values   Method

TIB

PMID: 32517845   PMCID: PMC7322167   doi: 10.1017/S0950268820001247

The basic reproduction number and prediction of the epidemic size of the novel coronavirus (COVID-19) in Shahroud, Iran

The aim of this study was to estimate the basic reproduction number (R0) of COVID-19 in the early stage of the epidemic and predict the expected number of new cases in Shahroud in Northeastern Iran. The R0 of COVID-19 was estimated using the serial interval distribution and the number of incidence cases. The 30-day probable incidence and cumulative incidence were predicted using the assumption that daily incidence follows a Poisson distribution determined by daily infectiousness. Data analysis was done using 'earlyR' and 'projections' packages in R software. The maximum-likelihood value of R0 was 2.7 (95% confidence interval (CI): 2.1-3.4) for the COVID-19 epidemic in the early 14 days and decreased to 1.13 (95% CI 1.03-1.25) by the end of day 42. The expected average number of new cases in Shahroud was 9.0 ± 3.8 cases/day, which means an estimated total of 271 (95% CI: 178–383) new cases for the period between 02 April to 03 May 2020. By day 67 (27 April), the effective reproduction number (Rt), which had a descending trend and was around 1, reduced to 0.70. Based on the Rt for the last 21 days (days 46–67 of the epidemic), the prediction for 27 April to 26 May is a mean daily cases of 2.9 ± 2.0 with 87 (48–136) new cases. In order to maintain R below 1, we strongly recommend enforcing and continuing the current preventive measures, restricting travel and providing screening tests for a larger proportion of the population.


```
[{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Shahroud, Iran",
    "date": "in the early 14 days",
    "R0 value": "2.7",
    "%CI values": "(95% confidence interval (CI): 2.1-3.4)",
    "method": "serial interval distribution and the number of incidence cases"
  }
},
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Shahroud, Iran",
    "date": "by the end of day 42",
    "R0 value": "1.13",
    "%CI values": "(95% CI 1.03-1.25)",
    "method": "serial interval distribution and the number of incidence cases"
  }
}]
```

Text vs. JSON Format

Text Format

disease name: primary pneumonic plague
location: Mukden, China
date: 1946
R0 value: 2.8 to 3.5
%CI values: The lower 95% confidence intervals of R(0) exceeded unity
method: statistical estimation of R(0) based on similar information

|

disease name: primary pneumonic plague
location: Madagascar
date: 1957
R0 value: 2.8 to 3.5
%CI values: The lower 95% confidence intervals of R(0) exceeded unity
method: statistical estimation of R(0) based on similar information

JSON Format

[{"contribution":{"disease name": "primary pneumonic plague",
"location": "Mukden, China",
"date": "1946",
"R0 value": "2.8 to 3.5",
"%CI values": "The lower 95% confidence intervals of R(0) exceeded unity",
"method": "statistical estimation of R(0) based on similar information"}},

{"contribution":{"disease name": "primary pneumonic plague",
"location": "Madagascar",
"date": "1957",
"R0 value": "2.8 to 3.5",
"%CI values": "The lower 95% confidence intervals of R(0) exceeded unity",
"method": "statistical estimation of R(0) based on similar information"}}]

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

- Finetuned Google's FLAN-T5 Large (780M) via instruction fine-tuning for single-task IE.

Reference

- Shamsabadi, M., D'Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

- Finetuned Google's FLAN-T5 Large (780M) via instruction fine-tuning for single-task IE.
  - Evaluations were promising against GPT (175B) with 1000x more parameters

		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

our FLAN-T5 fine-tuned model results

Reference

- Shamsabadi, M., D'Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

- Finetuned Google's FLAN-T5 Large (780M) via instruction fine-tuning for single-task IE.
    - Evaluations were promising against GPT (175B) with 1000x more parameters

		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

It outperforms GPT-3.5 with 1000x more parameters demonstrating the effectiveness of finetuning as a critical strategy to produce downstream parameter efficient models.

Reference

- Shamsabadi, M., D'Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

- Finetuned Google’s FLAN-T5 Large (780M) via instruction fine-tuning for single-task IE.
  - Evaluations were promising against GPT (175B) with 1000x more parameters
  - Qualitative observations: The LLM predictions can be compared against human annotations for a small number of papers online https://scinext-project.github.io/#/r0-estimates

Reference

- Shamsabadi, M., D’Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

[Epidemiological characteristics of COVID-19 in Gansu province]

Objective: To understand the epidemiological characteristics of COVID-19 cases in different epidemic stages in Gansu province. Methods: Epidemiological investigation was conducted to collect the information of confirmed COVID-19 cases, including demographic, epidemiological and clinical information. Results: As of 25 February 2020, a total of 91 confirmed COVID-19 cases had been reported in Gansu. The epidemic of COVID-19 in Gansu can be divided as three different stages, i.e. imported case stage, imported-case plus indigenous case stage, and indigenous case stage. A total of 63 cases were clustered cases (69.23%), 3 cases were medical staff infected with non-occupational exposure. The initial symptoms included fever (54.95%, 50/91), cough (52.75%, 48/91), or fatigue (28.57%, 26/91), the proportion of each symptom showed a decreasing trend along with the three epidemic stages, but only the differences in proportions of fever (trend χ2=2.20, P<0.05) and fatigue (trend χ2=3.18, P<0.05) among the three epidemic stages were statistically significant. The cases with critical severe symptoms accounted for 42.85% (6/14), 23.73% (14/59) and 16.67% (3/18), respectively, in three epidemic stages, showed a decreasing trend (H=6.45, P<0.05). Also, the incubation period prolonged along with the epidemic stage (F=51.65, P<0.01), but the intervals between disease onset and hospital visit (F=5.32, P<0.01), disease onset and diagnosis (F=5.25, P<0.01) became shorter along with the epidemic stage. Additionally, the basic reproduction number (R0) had decreased from 2.61 in imported case stage to 0.66 in indigenous case stage. Conclusions: The COVID-19 epidemic in Gansu was caused by the imported cases, and about 2/3 cases were clustered ones. No medical worker was observed to be infected by occupational exposure. With the progression of COVID-19 epidemic in Gansu, the change in initial symptom and incubation period suggests, the early screening cannot only depend on body temperature monitoring.

Human Annotation


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu province",
    "date": "As of 25 February 2020",
    "R0 value": "decreased from 2.61 in imported case stage
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


LLM Prediction


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu",
    "date": "-",
    "R0 value": "2.61 in imported case stage to 0.66 in indi
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


Snapshot of what the data on the website looks like

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

[Epidemiological characteristics of COVID-19 in Gansu province]

Objective: To understand the epidemiological characteristics of COVID-19 cases in different epidemic stages in Gansu province. Methods: Epidemiological investigation was conducted to collect the information of confirmed COVID-19 cases, including demographic, epidemiological and clinical information. Results: As of 25 February 2020, a total of 91 confirmed COVID-19 cases had been reported in Gansu. The epidemic of COVID-19 in Gansu can be divided as three different stages, i.e. imported case stage, imported-case plus indigenous case stage, and indigenous case stage. A total of 63 cases were clustered cases (69.23%), 3 cases were medical staff infected with non-occupational exposure. The initial symptoms included fever (54.95%, 50/91), cough (52.75%, 48/91), or fatigue (28.57%, 26/91), the proportion of each symptom showed a decreasing trend along with the three epidemic stages, but only the differences in proportions of fever (trend χ2=2.20, P<0.05) and fatigue (trend χ2=3.18, P<0.05) among the three epidemic stages were statistically significant. The cases with critical severe symptoms accounted for 42.85% (6/14), 23.73% (14/59) and 16.67% (3/18), respectively, in three epidemic stages, showed a decreasing trend (H=6.45, P<0.05). Also, the incubation period prolonged along with the epidemic stage (F=51.65, P<0.01), but the intervals between disease onset and hospital visit (F=5.32, P<0.01), disease onset and diagnosis (F=5.25, P<0.01) became shorter along with the epidemic stage. Additionally, the basic reproduction number (R0) had decreased from 2.61 in imported case stage to 0.66 in indigenous case stage. Conclusions: The COVID-19 epidemic in Gansu was caused by the imported cases, and about 2/3 cases were clustered ones. No medical worker was observed to be infected by occupational exposure. With the progression of COVID-19 epidemic in Gansu, the change in initial symptom and incubation period suggests, the early screening cannot only depend on body temperature monitoring.

Abstract Input

Human Annotation


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu province",
    "date": "As of 25 February 2020",
    "R0 value": "decreased from 2.61 in imported case stage
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


LLM Prediction


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu",
    "date": "-",
    "R0 value": "2.61 in imported case stage to 0.66 in indi
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

[Epidemiological characteristics of COVID-19 in Gansu province]

Objective: To understand the epidemiological characteristics of COVID-19 cases in different epidemic stages in Gansu province. Methods: Epidemiological investigation was conducted to collect the information of confirmed COVID-19 cases, including demographic, epidemiological and clinical information. Results: As of 25 February 2020, a total of 91 confirmed COVID-19 cases had been reported in Gansu. The epidemic of COVID-19 in Gansu can be divided as three different stages, i.e. imported case stage, imported-case plus indigenous case stage, and indigenous case stage. A total of 63 cases were clustered cases (69.23%), 3 cases were medical staff infected with non-occupational exposure. The initial symptoms included fever (54.95%, 50/91), cough (52.75%, 48/91), or fatigue (28.57%, 26/91), the proportion of each symptom showed a decreasing trend along with the three epidemic stages, but only the differences in proportions of fever (trend χ2=2.20, P<0.05) and fatigue (trend χ2=3.18, P<0.05) among the three epidemic stages were statistically significant. The cases with critical severe symptoms accounted for 42.85% (6/14), 23.73% (14/59) and 16.67% (3/18), respectively, in three epidemic stages, showed a decreasing trend (H=6.45, P<0.05). Also, the incubation period prolonged along with the epidemic stage (F=51.65, P<0.01), but the intervals between disease onset and hospital visit (F=5.32, P<0.01), disease onset and diagnosis (F=5.25, P<0.01) became shorter along with the epidemic stage. Additionally, the basic reproduction number (R0) had decreased from 2.61 in imported case stage to 0.66 in indigenous case stage. Conclusions: The COVID-19 epidemic in Gansu was caused by the imported cases, and about 2/3 cases were clustered ones. No medical worker was observed to be infected by occupational exposure. With the progression of COVID-19 epidemic in Gansu, the change in initial symptom and incubation period suggests, the early screening cannot only depend on body temperature monitoring.

Human Annotation


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu province",
    "date": "As of 25 February 2020",
    "R0 value": "decreased from 2.61 in imported case stage
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


LLM Prediction


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu",
    "date": "-",
    "R0 value": "2.61 in imported case stage to 0.66 in indi
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


Output from human annotation versus the output from the LLM

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

[Epidemiological characteristics of COVID-19 in Gansu province]

Objective: To understand the epidemiological characteristics of COVID-19 cases in different epidemic stages in Gansu province. Methods: Epidemiological investigation was conducted to collect the information of confirmed COVID-19 cases, including demographic, epidemiological and clinical information. Results: As of 25 February 2020, a total of 91 confirmed COVID-19 cases had been reported in Gansu. The epidemic of COVID-19 in Gansu can be divided as three different stages, i.e. imported case stage, imported-case plus indigenous case stage, and indigenous case stage. A total of 63 cases were clustered cases (69.23%), 3 cases were medical staff infected with non-occupational exposure. The initial symptoms included fever (54.95%, 50/91), cough (52.75%, 48/91), or fatigue (28.57%, 26/91), the proportion of each symptom showed a decreasing trend along with the three epidemic stages, but only the differences in proportions of fever (trend χ2=2.20, P<0.05) and fatigue (trend χ2=3.18, P<0.05) among the three epidemic stages were statistically significant. The cases with critical severe symptoms accounted for 42.85% (6/14), 23.73% (14/59) and 16.67% (3/18), respectively, in three epidemic stages, showed a decreasing trend (H=6.45, P<0.05). Also, the incubation period prolonged along with the epidemic stage (F=51.65, P<0.01), but the intervals between disease onset and hospital visit (F=5.32, P<0.01), disease onset and diagnosis (F=5.25, P<0.01) became shorter along with the epidemic stage. Additionally, the basic reproduction number (R0) had decreased from 2.61 in imported case stage to 0.66 in indigenous case stage. Conclusions: The COVID-19 epidemic in Gansu was caused by the imported cases, and about 2/3 cases were clustered ones. No medical worker was observed to be infected by occupational exposure. With the progression of COVID-19 epidemic in Gansu, the change in initial symptom and incubation period suggests, the early screening cannot only depend on body temperature monitoring.

Refresh button shows new abstract

Human Annotation


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu province",
    "date": "As of 25 February 2020",
    "R0 value": "decreased from 2.61 in imported case stage
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


LLM Prediction


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu",
    "date": "-",
    "R0 value": "2.61 in imported case stage to 0.66 in ind
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```

Model -

Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

[Epidemiological characteristics of COVID-19 in Gansu province]

Objective: To understand the epidemiological characteristics of COVID-19 cases in different epidemic stages in Gansu province. Methods: Epidemiological investigation was conducted to collect the information of confirmed COVID-19 cases, including demographic, epidemiological and clinical information. Results: As of 25 February 2020, a total of 91 confirmed COVID-19 cases had been reported in Gansu. The epidemic of COVID-19 in Gansu can be divided as three different stages, i.e. imported case stage, imported-case plus indigenous case stage, and indigenous case stage. A total of 63 cases were clustered cases (69.23%), 3 cases were medical staff infected with non-occupational exposure. The initial symptoms included fever (54.95%, 50/91), cough (52.75%, 48/91), or fatigue (28.57%, 26/91), the proportion of each symptom showed a decreasing trend along with the three epidemic stages, but only the differences in proportions of fever (trend χ2=2.20, P<0.05) and fatigue (trend χ2=3.18, P<0.05) among the three epidemic stages were statistically significant. The cases with critical severe symptoms accounted for 42.85% (6/14), 23.73% (14/59) and 16.67% (3/18), respectively, in three epidemic stages, showed a decreasing trend (H=6.45, P<0.05). Also, the incubation period prolonged along with the epidemic stage (F=51.65, P<0.01), but the intervals between disease onset and hospital visit (F=5.32, P<0.01), disease onset and diagnosis (F=5.25, P<0.01) became shorter along with the epidemic stage. Additionally, the basic reproduction number (R0) had decreased from 2.61 in imported case stage to 0.66 in indigenous case stage. Conclusions: The COVID-19 epidemic in Gansu was caused by the imported cases, and about 2/3 cases were clustered ones. No medical worker was observed to be infected by occupational exposure. With the progression of COVID-19 epidemic in Gansu, the change in initial symptom and incubation period suggests, the early screening cannot only depend on body temperature monitoring.

Human Annotation


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu province",
    "date": "As of 25 February 2020",
    "R0 value": "decreased from 2.61 in imported case stage
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


LLM Prediction


```
{
  "contribution": {
    "disease name": "COVID-19",
    "location": "Gansu",
    "date": "-",
    "R0 value": "2.61 in imported case stage to 0.66 in indi
    "%CI values": "-",
    "method": "Epidemiological investigation"
  }
}
```


shows why quantitative evaluations at 60% might not actually reflect that the model output is actually also correct

Model -

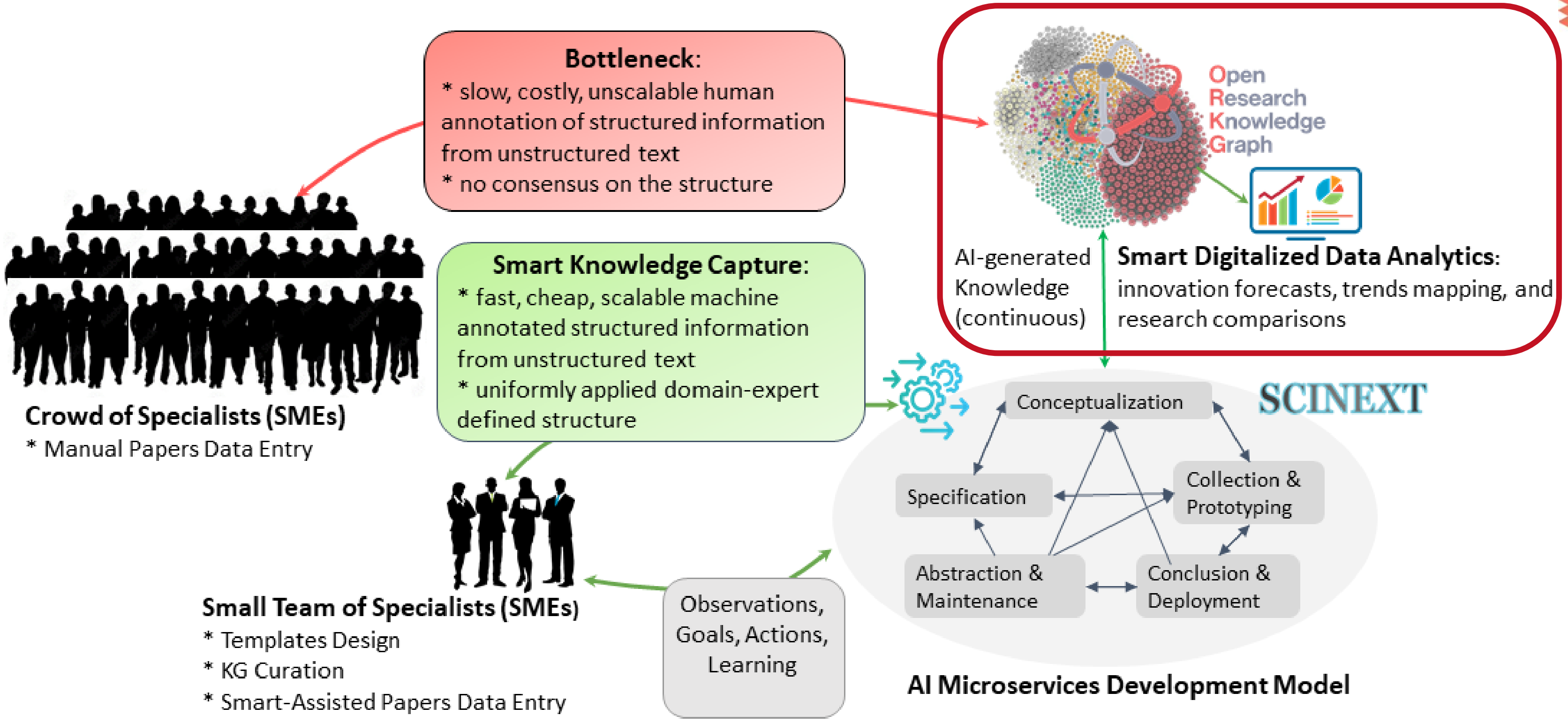
Automatic Extraction of Structured Scholarly Knowledge about R0 estimates for Infectious Diseases From Paper Abstracts

- Finetuned Google's FLAN-T5 Large (780M) via instruction fine-tuning for single-task IE.
  - Evaluations were promising against GPT (175B) with 1000x more parameters
  - Qualitative observations: The LLM predictions can be compared against human annotations for a small number of papers online https://scinext-project.github.io/#/r0-estimates
  - Model release on HuggingFace https://huggingface.co/orkg/R0_contribution_IE

Reference

- Shamsabadi, M., D'Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 374-392).

Why AI-powered Research Assistance in the context of the ORKG?

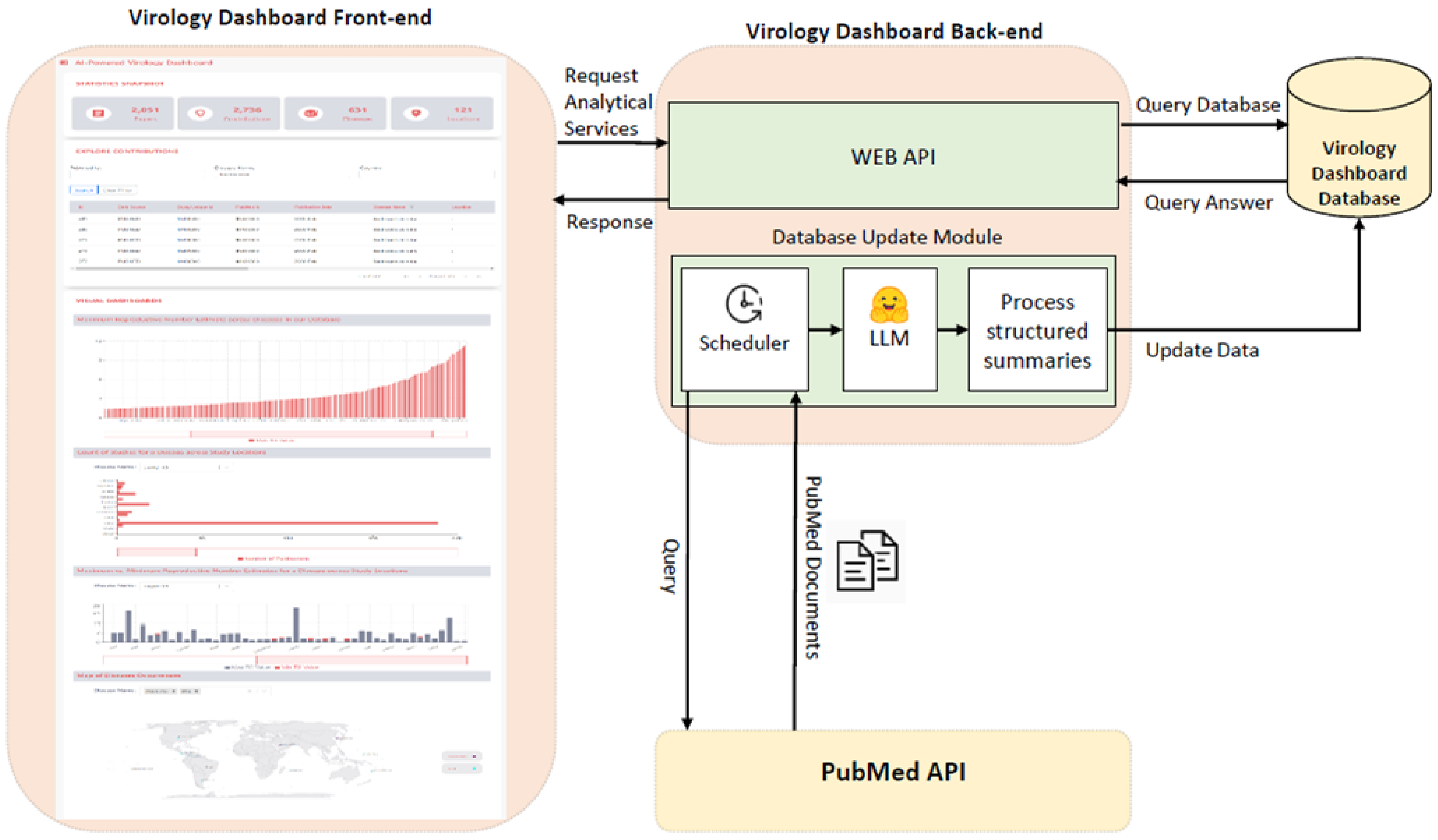


Next-generation Information Technology (IT)

- Given the large-scale structured knowledge produced by the AI services, next-generation IT solutions as visualization dashboards can be implemented providing researchers with easily comprehensible visual summaries of scholarly information, thereby assisting them in filtering for relevant research.

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions



URL: https://orkg.org/usecases/r0-estimates

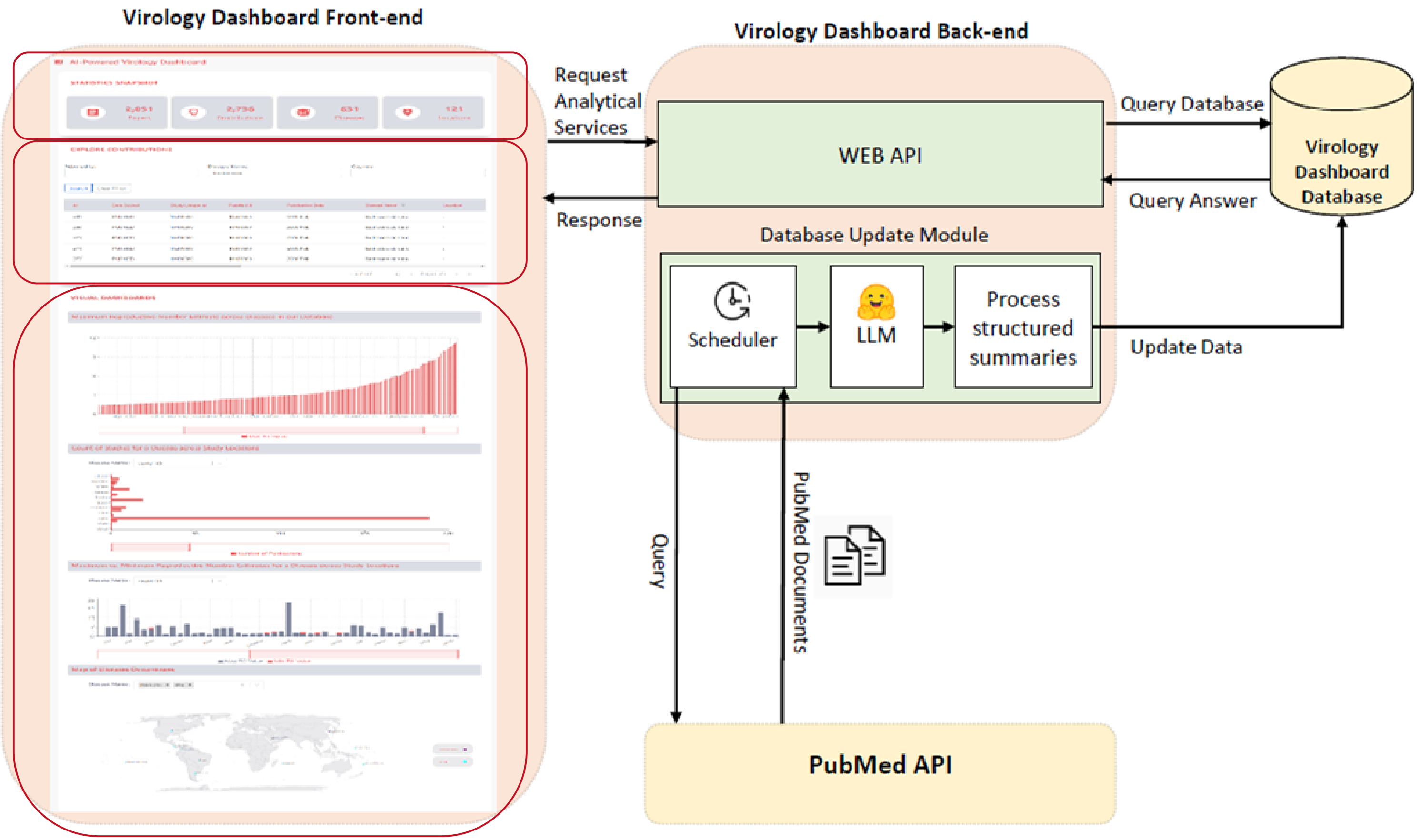
Reference

Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

1. Summary stats
2. Browse papers
3. Visualization Dashboard as assistants to researchers to filter for information



URL: https://orkg.org/usecases/r0-estimates

Reference
Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

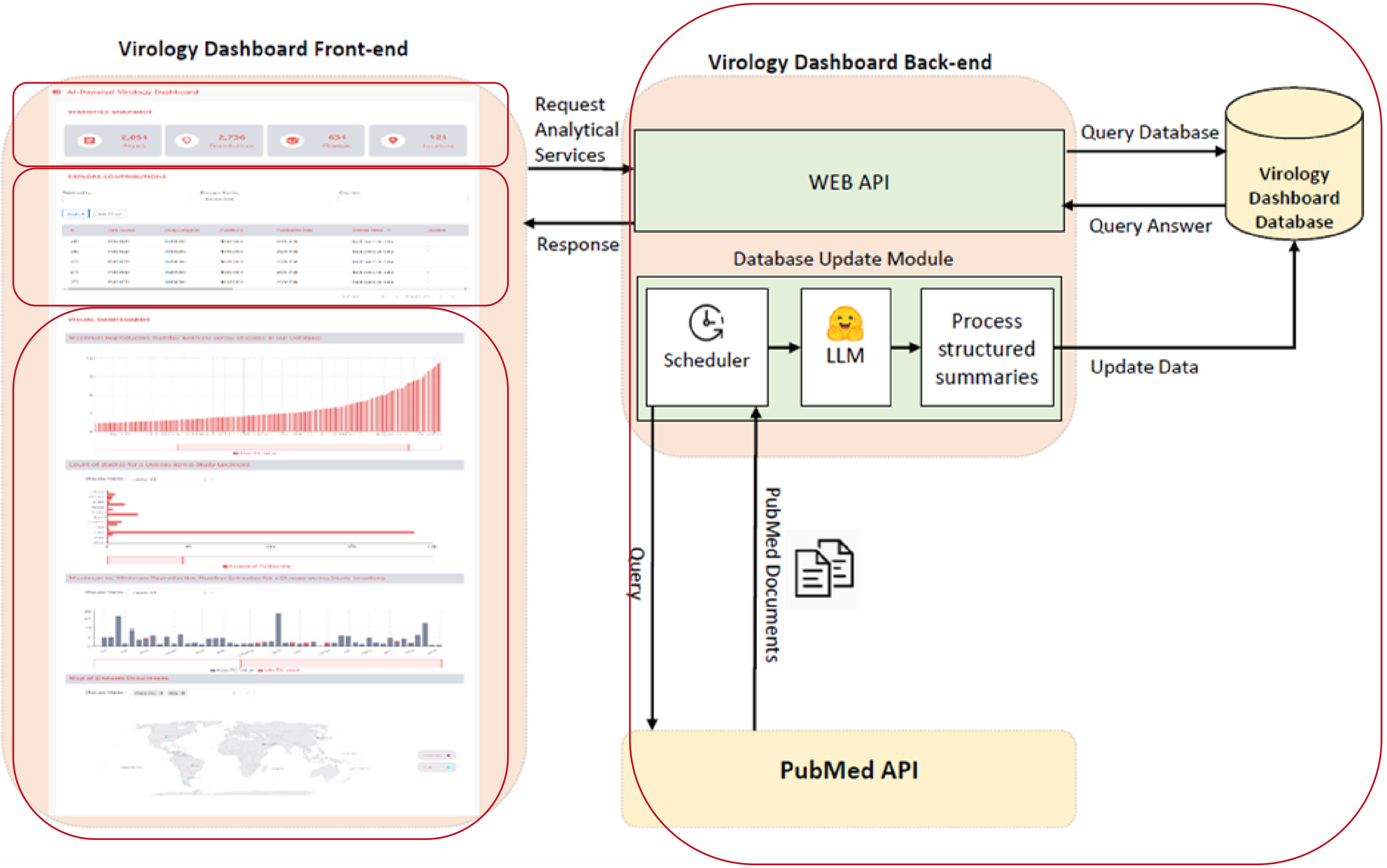
Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

1. Summary stats

2. Browse papers

3. Visualization Dashboard as assistants to researchers to filter for information



Backend workflow: fetches articles from scholarly publishers, passes them to the LLM, and adds new structured data to the database. Thousands of articles can be processed in this manner.

The scheduler is implemented to work on a weekly basis fetching new articles.

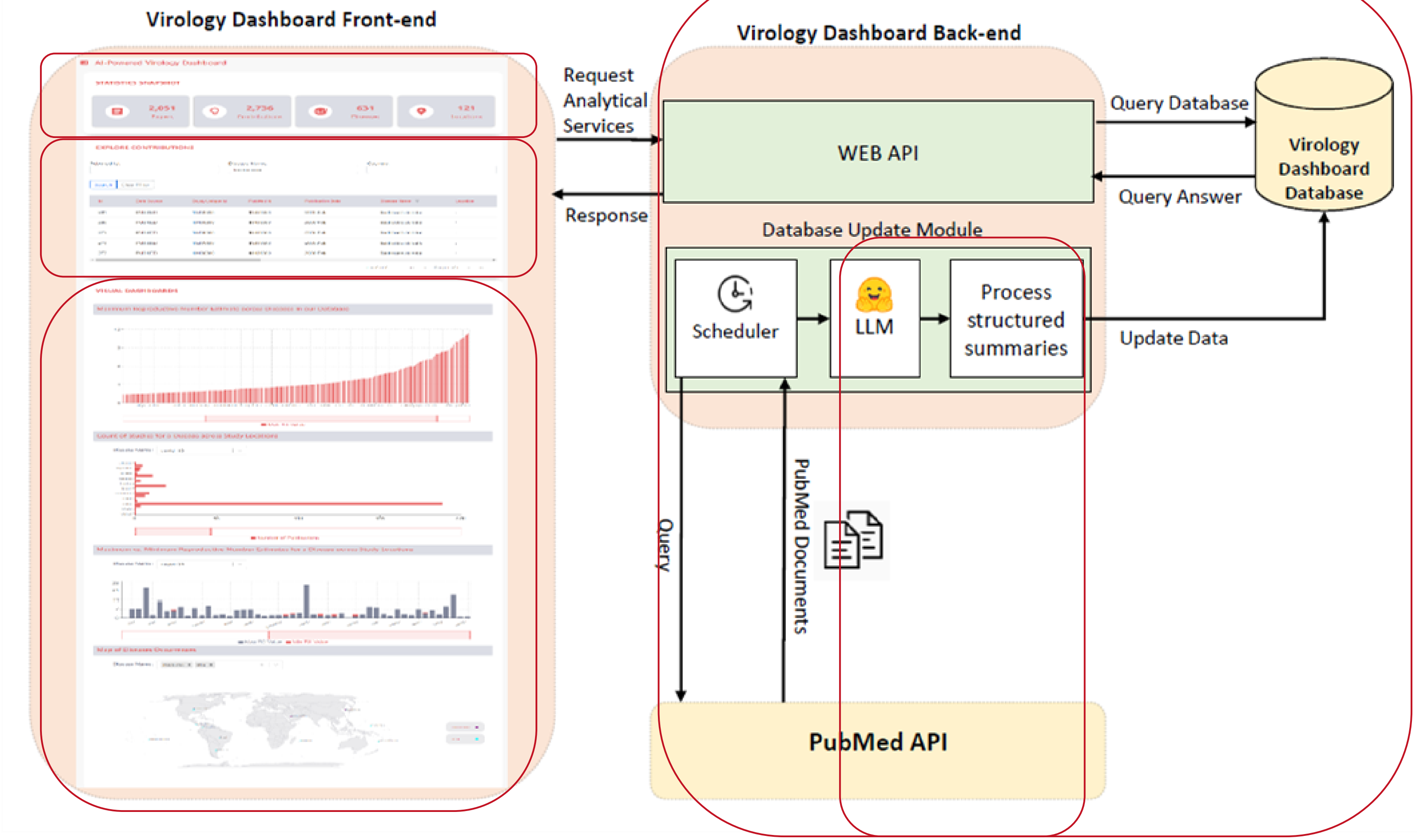
Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

1. Summary stats

2. Browse papers

3. Visualization Dashboard as assistants to researchers to filter for information



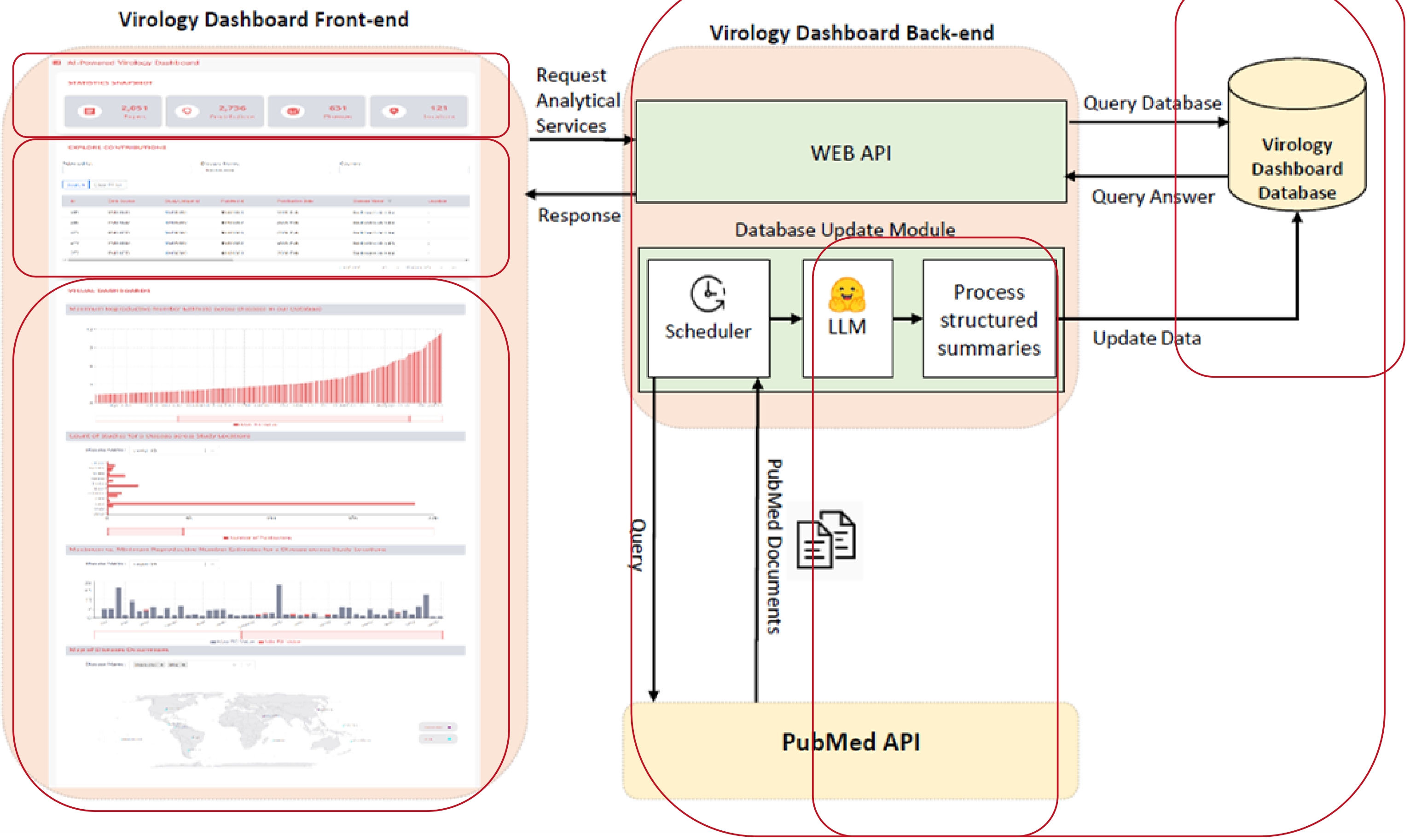
The capability of the LLM processing large-scale data in a matter of hours highlights how it alleviates the bottleneck of an otherwise costly and time-consuming human annotation task.

URL: https://orkg.org/usecases/r0-estimates

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

1. Summary stats
2. Browse papers
3. Visualization Dashboard as assistants to researchers to filter for information



The database currently has over a thousand structured summaries.

URL: https://orkg.org/usecases/r0-estimates

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

TIB

1. Summary stats
2. Browse papers
3. Visualization Dashboard as assistants to researchers to filter for information

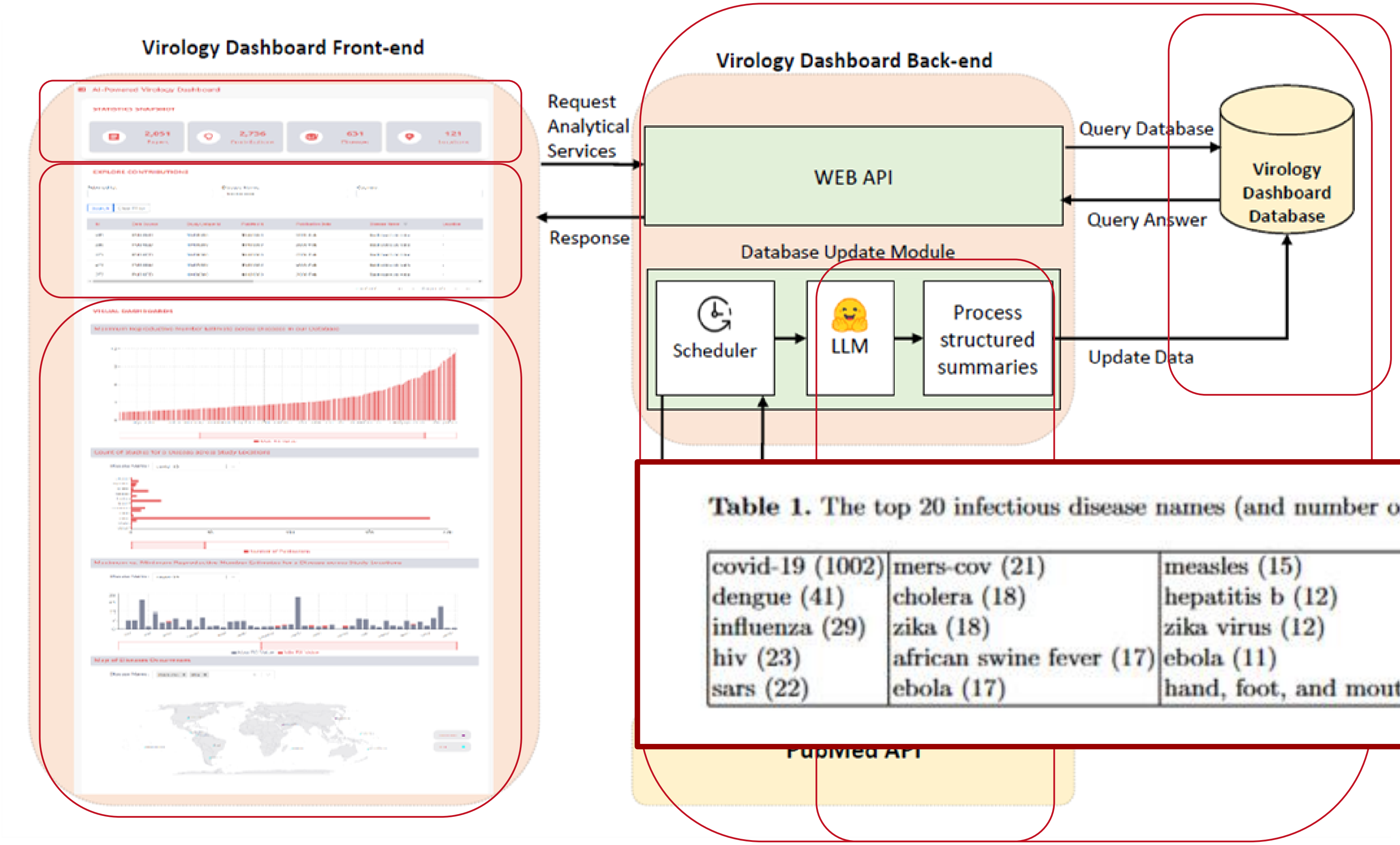


Table 1. The top 20 infectious disease names (and number of papers) in our dataset.

covid-19 (1002)	mers-cov (21)	measles (15)	hepatitis c (8)
dengue (41)	cholera (18)	hepatitis b (12)	tuberculosis (8)
influenza (29)	zika (18)	zika virus (12)	monkeypox (8)
hiv (23)	african swine fever (17)	ebola (11)	west nile virus (7)
sars (22)	ebola (17)	hand, foot, and mouth disease (8)	malaria (7)

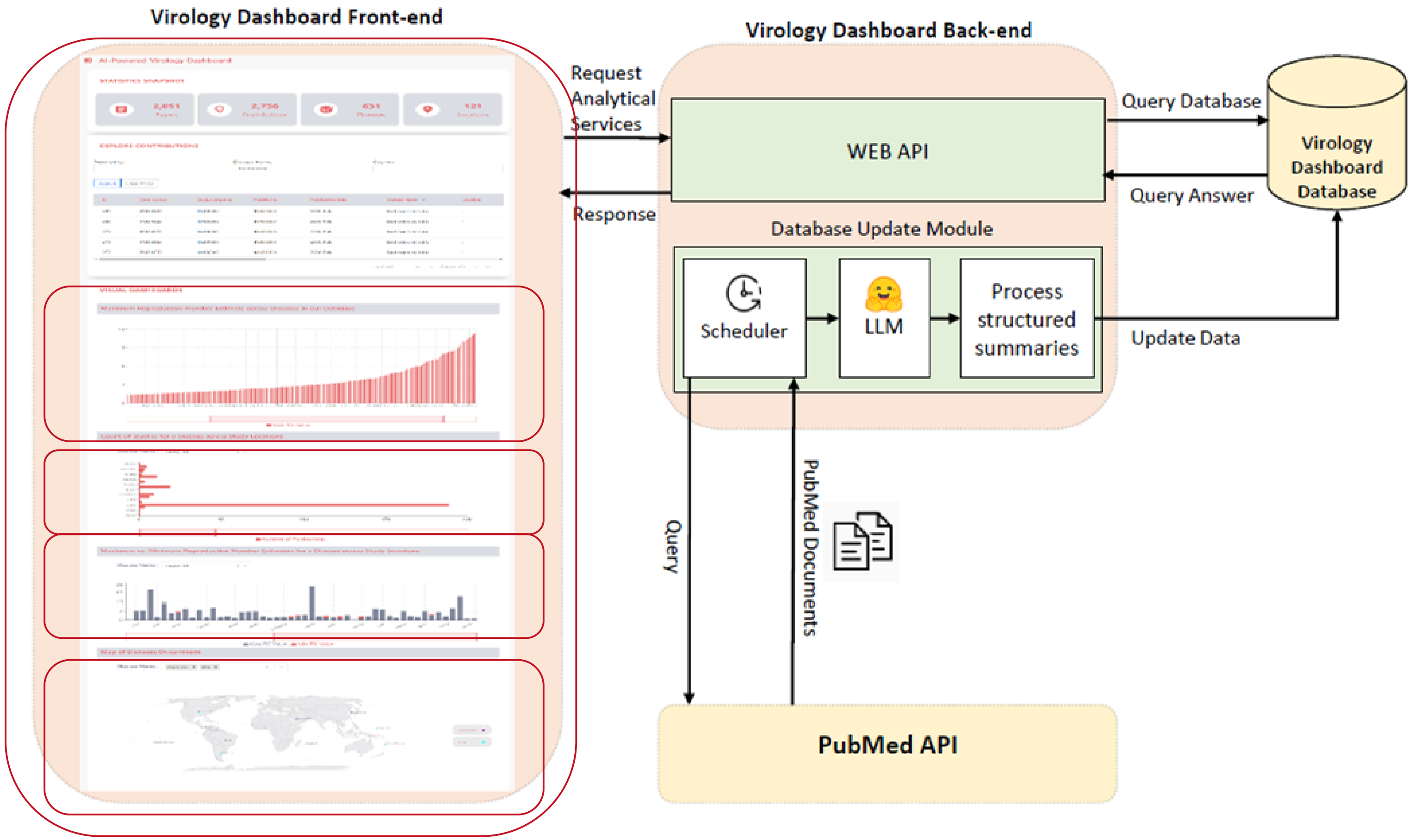
The database currently has over a thousand structured summaries.

URL: https://orkg.org/usecases/r0-estimates

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

At the outset our aim was to use the power of IT to assist researchers in their filtering for scholarly knowledge.

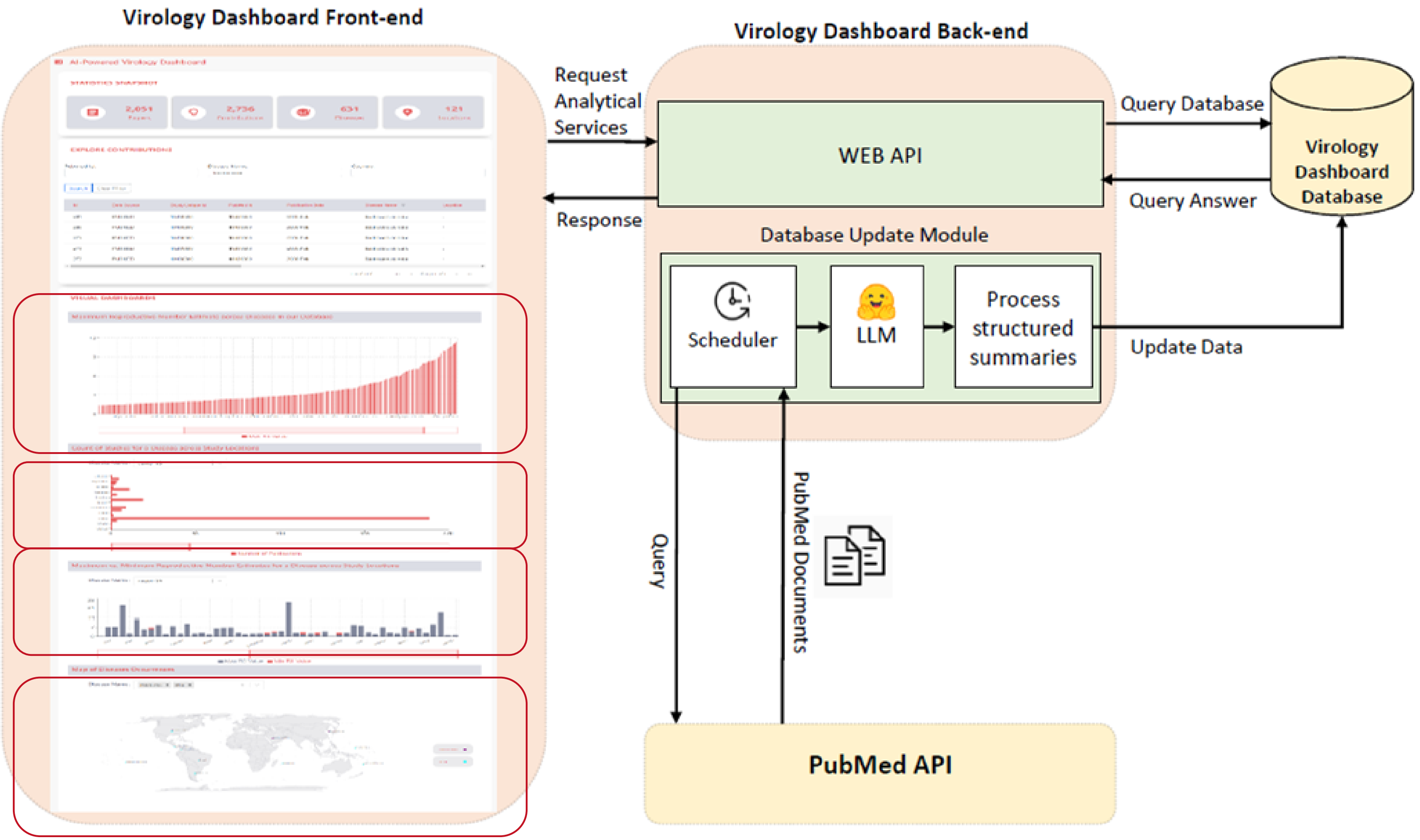


URL: https://orkg.org/usecases/r0-estimates

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

Thus each of these interfaces is designed to answer specific research questions and present the information in a visual summary.



URL: https://orkg.org/usecases/r0-estimates

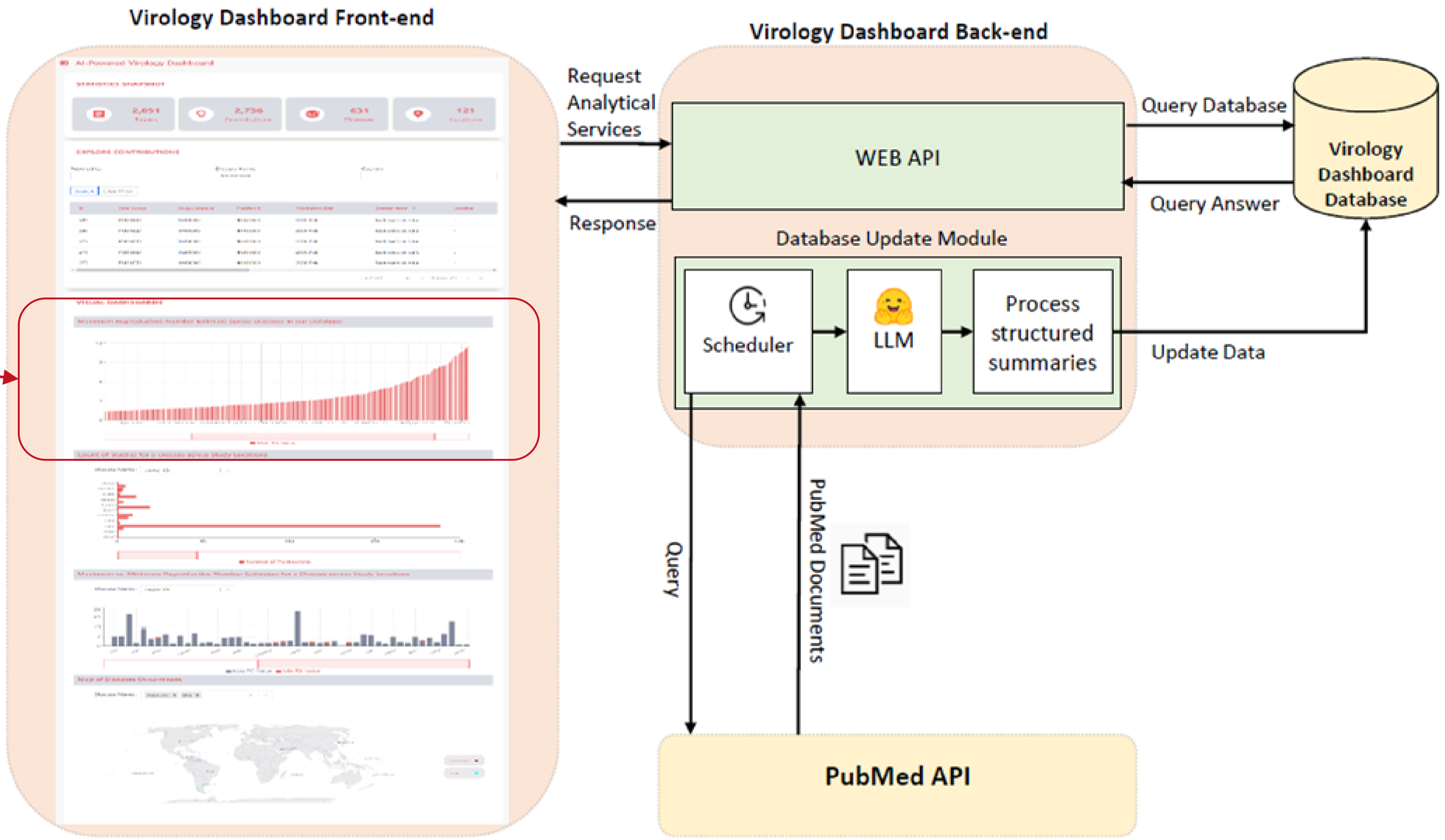
Reference

Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

RQ1: what is the max R0 reported for diseases in our database?



URL: https://orkg.org/usecases/r0-estimates

Reference

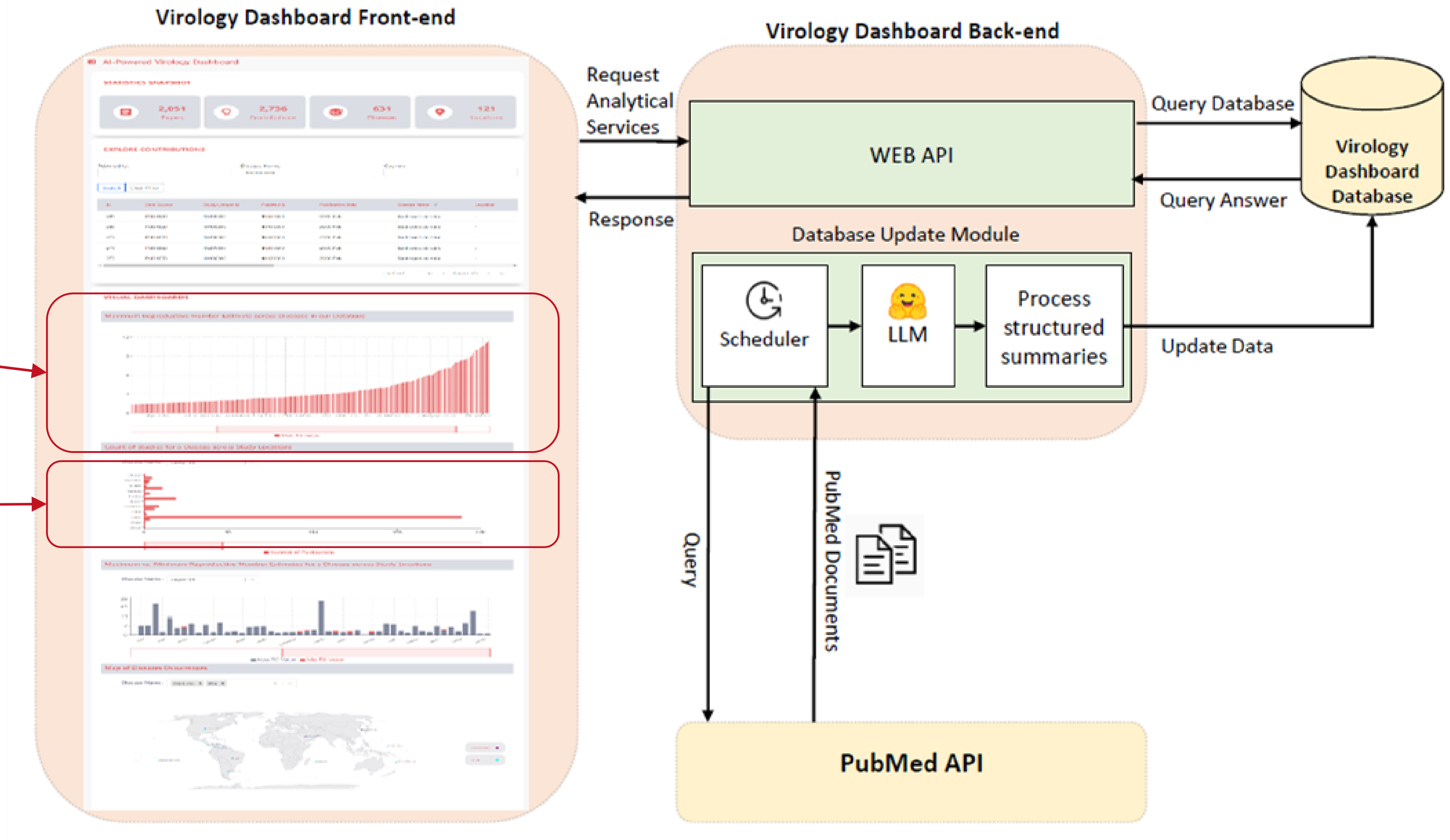
Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

RQ1: what is the max R0 reported for diseases in our database?

RQ2: For a chosen disease, how many studies have been reported across study locations?



URL: https://orkg.org/usecases/r0-estimates

Reference

Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

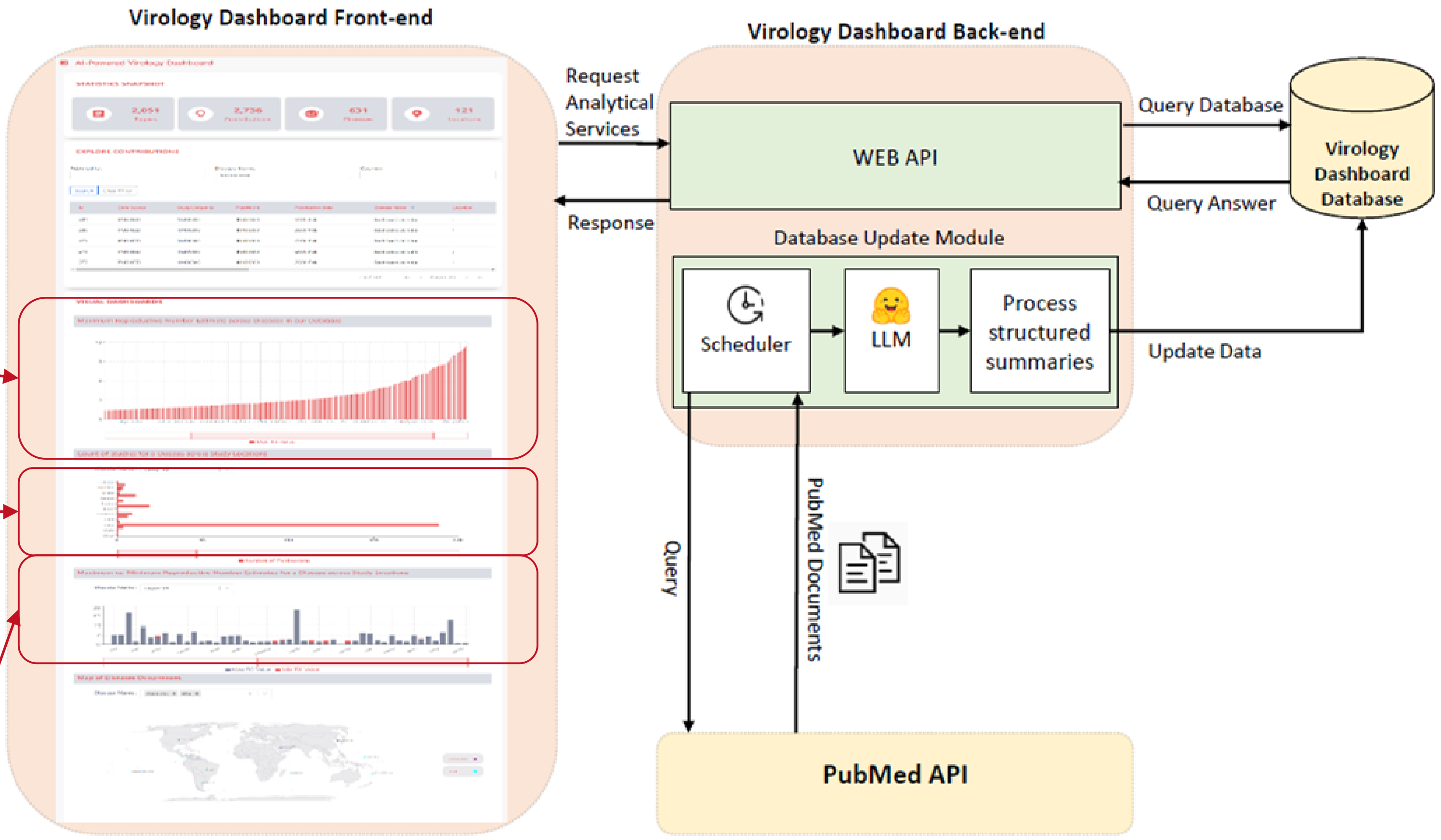
Dashboard -

Streamlining Access to Scholarly Articles via Visualization Dashboards operating over Structured Scholarly Contributions

RQ1: what is the max R0 reported for diseases in our database?

RQ2: For a chosen disease, how many studies have been reported across study locations?

RQ3: what is the min and max R0 for a disease across study locations?



URL: https://orkg.org/usecases/r0-estimates

Reference

Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. *arXiv preprint arXiv:2402.14622*.

Acknowledgements

- PhD Candidate: Mahsa Shamsabadi
- Senior Project Advisor: Sören Auer
- ORKG Team https://orkg.org/about/9/Team
- References

  - Shamsabadi, M., & D'Souza, J. (2024). A FAIR and Free Prompt-based Research Assistant. arXiv preprint arXiv:2405.14601. Forthcoming paper at: The 29th International Conference on Natural Language & Information Systems (NLDB 2024)
    - Code: https://github.com/mahsaSH717/research_assistant
  - Shamsabadi, M., D'Souza, J., & Auer, S. (2024, March). Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 374-392).
    - Datatset: https://zenodo.org/records/8068442
    - Model: https://huggingface.co/orkg/R0_contribution_IE
    - Code: https://github.com/mahsaSH717/r0-estimates
  - Shamsabadi, M., & D'Souza, J. (2024). From Keywords to Structured Summaries: Streamlining Scholarly Knowledge Access. arXiv preprint arXiv:2402.14622.
    - Tool: https://orkg.org/usecases/r0-estimates

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
    - **Research Assistant**
    - **An LLM-powered Virology Dashboard**
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
    - LLMs for evaluating Scientific Synthesis

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
  - Research Assistant
  - An LLM-powered Virology Dashboard
  - **Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task**
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
  - LLMs for evaluating Scientific Synthesis

LLMs4OL: Large Language Models for Ontology Learning

Hamed Babaei Giglou   Jennifer D'Souza   Sören Auer

TIB — Leibniz Information Centre for Science and Technology
Hannover, Germany

ISWC - **November 8, 2023**

References
https://scinext-project.github.io/
Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

Why LLMs4OL?

- Within the SCINEXT research group, we also do basic science research. Our work for LLMs4OL falls under this umbrella.

Why LLMs4OL?

- Within the SCINEXT research group, we also do basic science research. Our work for LLMs4OL falls under this umbrella.

- **Motivation for the work**
  - There are various posts in popular media claiming
    - LLMs are unfit for Science
      - Are large language models right for scientific research?
    - LLMs are unfit for Math
      - https://bdtechtalks.com/2023/03/06/chatgpt-llm-mathematics/
    - LLMs cannot really reason
      - https://languagelog.ldc.upenn.edu/nll/?p=60081

Why LLMs4OL?

TIB

- Within the SCINEXT research group, we also do basic science research. Our work for LLMs4OL falls under this umbrella.

- **Motivation for the work**
  - There are various posts in popular media claiming
    - LLMs are unfit for Science
      - Are large language models right for scientific research?
    - LLMs are unfit for Math
      - https://bdtechtalks.com/2023/03/06/chatgpt-llm-mathematics/
    - LLMs cannot really reason
      - https://languagelog.ldc.upenn.edu/nll/?p=60081

1. Select one aspect

Why LLMs4OL?

TIB

- Within the SCINEXT research group, we also do basic science research. Our work for LLMs4OL falls under this umbrella.

- **Motivation for the work**
    - There are various posts in popular media claiming
        - LLMs are unfit for Science
            - Are large language models right for scientific research?
        - LLMs are unfit for Math
            - https://bdtechtalks.com/2023/03/06/chatgpt-llm-mathematics/
        - LLMs cannot really reason
            - https://languagelog.ldc.upenn.edu/nll/?p=60081

1. Select one aspect.
2. Empirically measure the fitness of LLMs for a task, specifically Ontology Learning (OL), that contributes to the greater objective of testing the fitness of LLMs for Science.

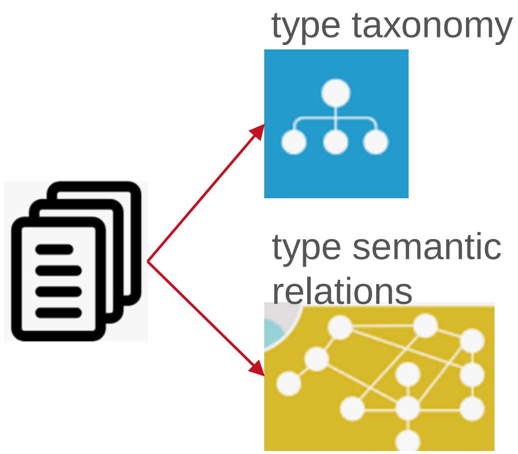
What is Ontology Learning (OL)?

- Ontology learning (OL) is the process of automatically extracting and structuring knowledge from unstructured sources, like text or databases, to create a formal representation of concepts and their relationships.
  - Ontologies were created to structure information for computers, enhancing data processing, but they also benefit humans by organizing complex information.

What is Ontology Learning (OL)?

- Ontology learning (OL) is the process of automatically extracting and structuring knowledge from unstructured sources, like text or databases, to create a formal representation of concepts and their relationships.
  - Ontologies were created to structure information for computers, enhancing data processing, but they also benefit humans by organizing complex information.

A conceptual view



structure unstructured knowledge

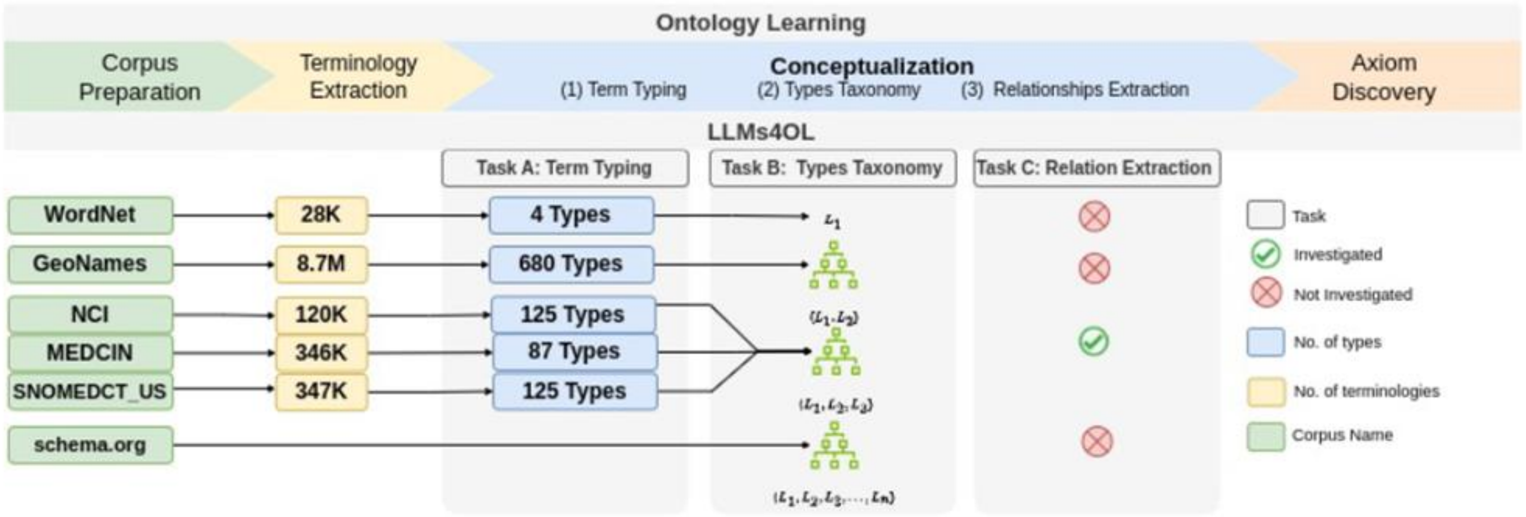
Why LLMs4OL?

- We test the hypothesis:
  - Can LLMs effectively apply their language pattern capturing capability to OL, which involves automatically extracting and structuring knowledge from natural language text?

Why LLMs4OL?

- We test the hypothesis:
  - Can LLMs effectively apply their language pattern capturing capability to OL, which involves automatically extracting and structuring knowledge from natural language text?
    - We comprehensively selected 8 diverse model families that were the state-of-the-art at the time of the research based on their architectural differences.
      - The selected LLMs for validation were: BERT (encoder-only); BLOOM, MetaAI's LLaMA, OpenAI's GPT-3, GPT-3.5, GPT-4 (all decoder-only); and BART and Google's Flan-T5 (encoder-decoder).

LLMs4OL Paradigm



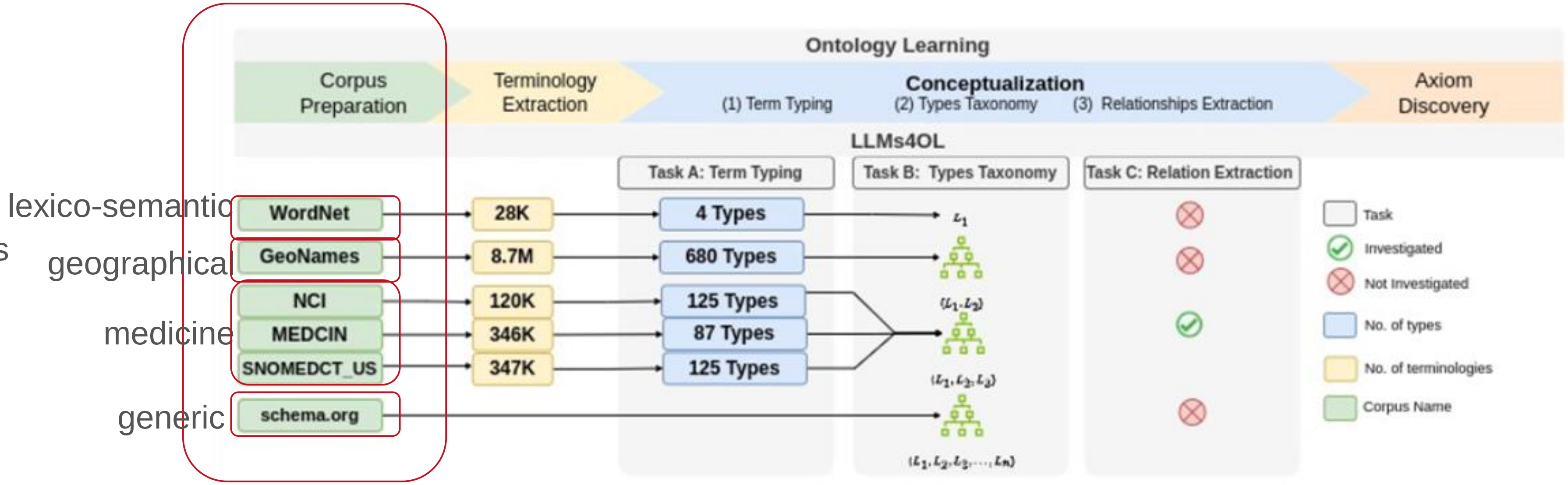
- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed on ontologies from 4 domains of knowledge.



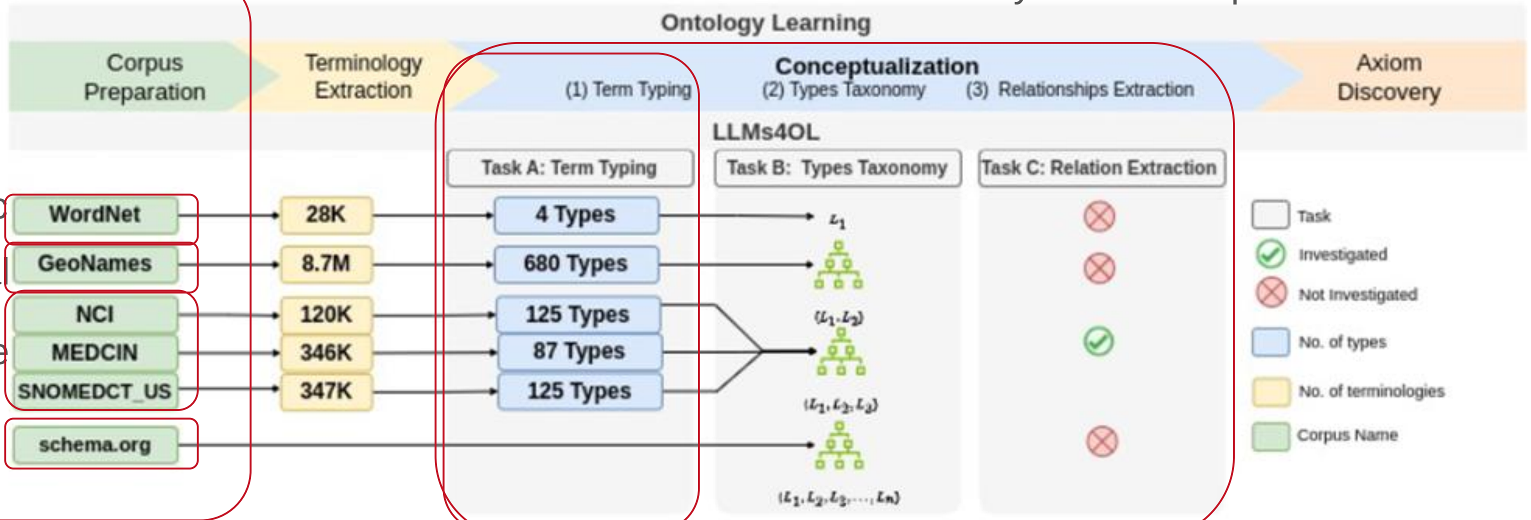
- Task A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- Task B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- Task C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



lexico-semantic

geographical

medicine

generic

The OL task is addressed on ontologies from 4 domains of knowledge.

- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

E.g., in the context of the lexico-semantic WordNet, if there is a word like "running," the LLM is queried for its expected type i.e. "Verb"

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.

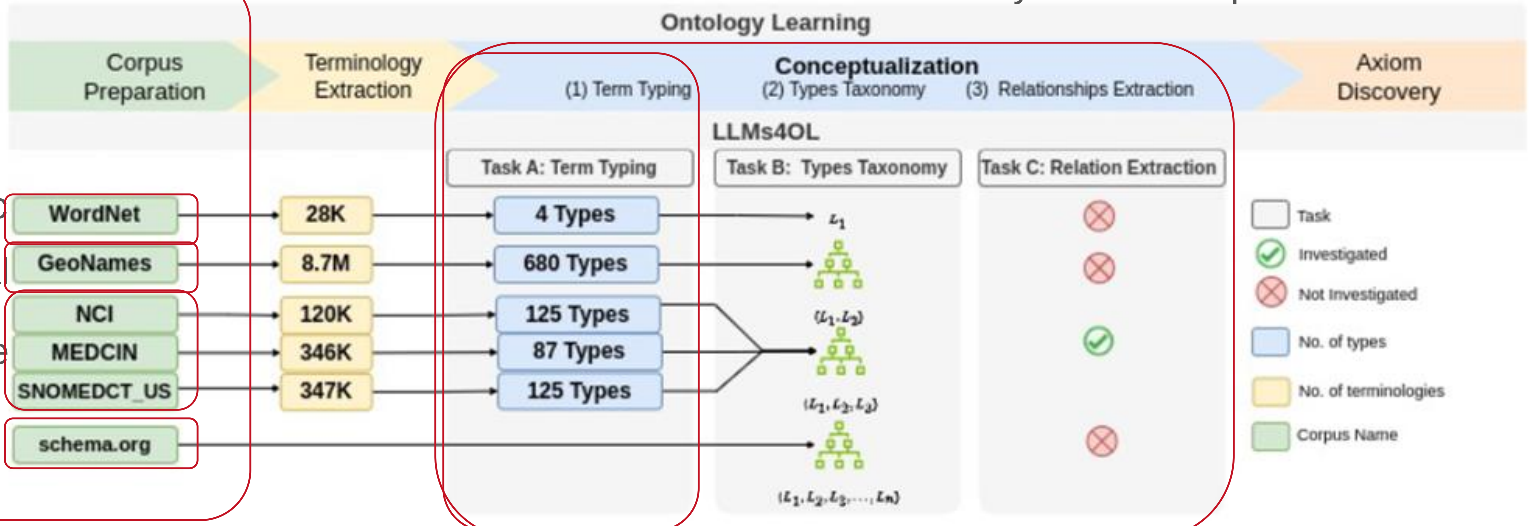
The OL task is addressed on ontologies from 4 domains of knowledge.

lexico-semantic

geographical

medicine

generic



- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

E.g., in the context of the geographical ontology GeoNames, if there is a word like "Nile," the LLM is queried for its expected type i.e. "River"

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.

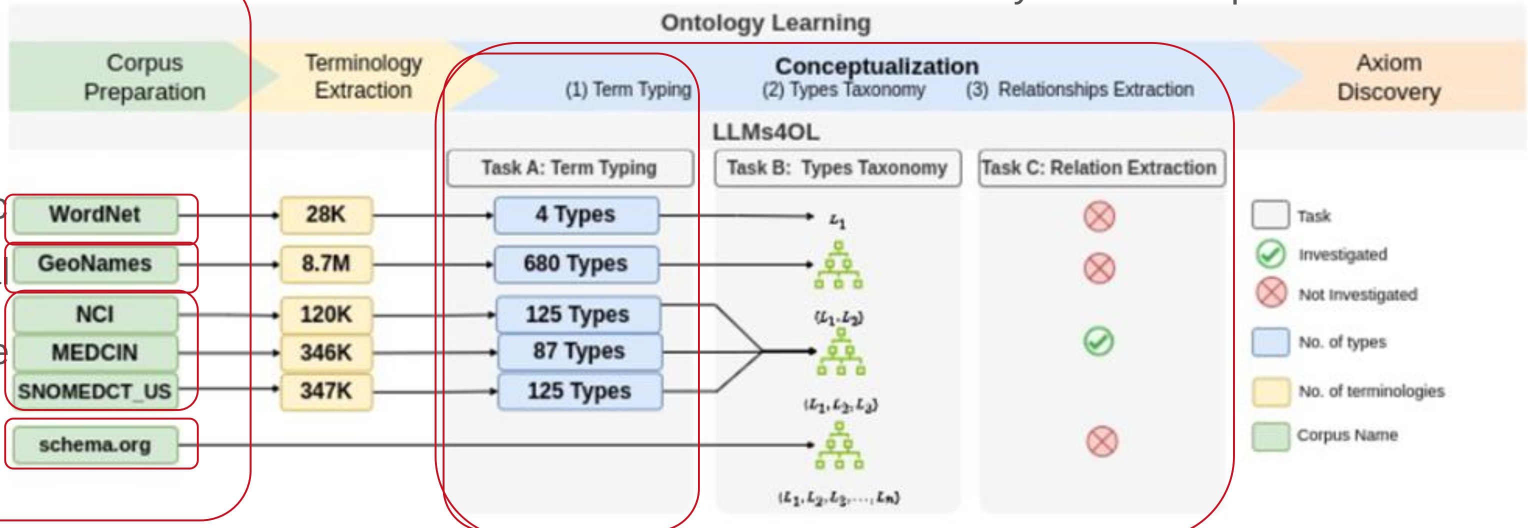
The OL task is addressed on ontologies from 4 domains of knowledge.

lexico-semantic

geographical

medicine

generic



- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type tax
- TASK C: LLMs to d

Parameter	WordNet	GeoNames	NCI	MEDCIN	SNOMEDCT_US
Train Set Size	40,559	8,078,865	96,177	277,028	278,374
Test Set Size	9,470	702,510	24,045	69,258	69,594
Types	4	680	125	87	125

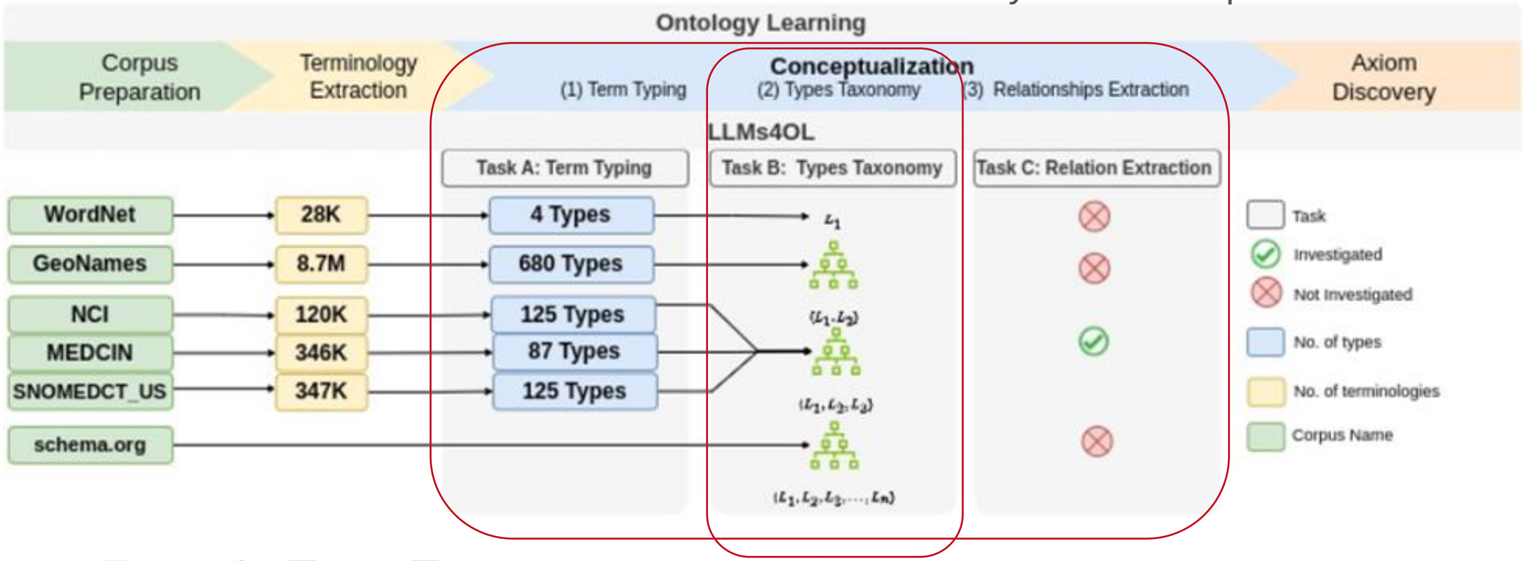
our experimental dataset statistics

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



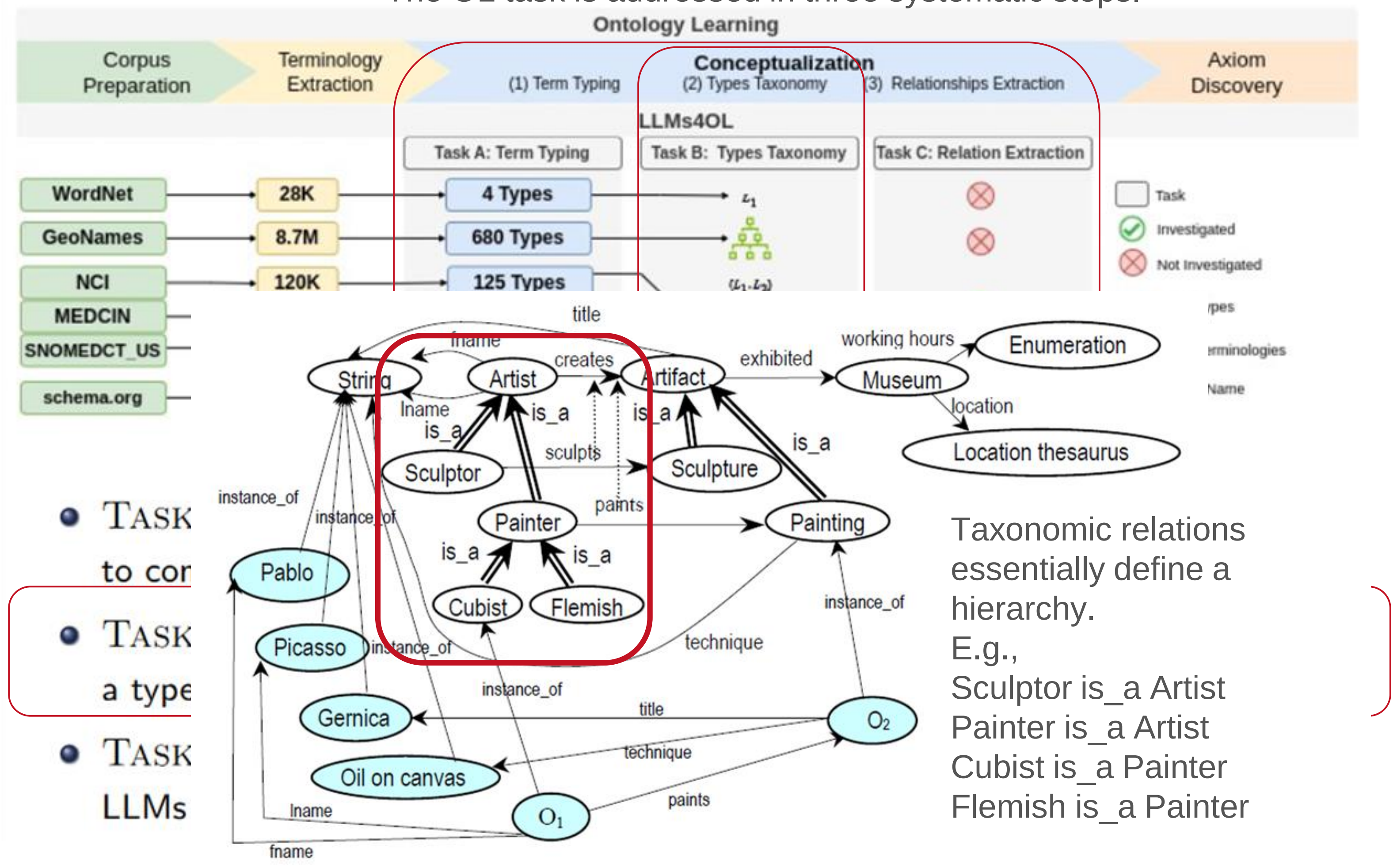
- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



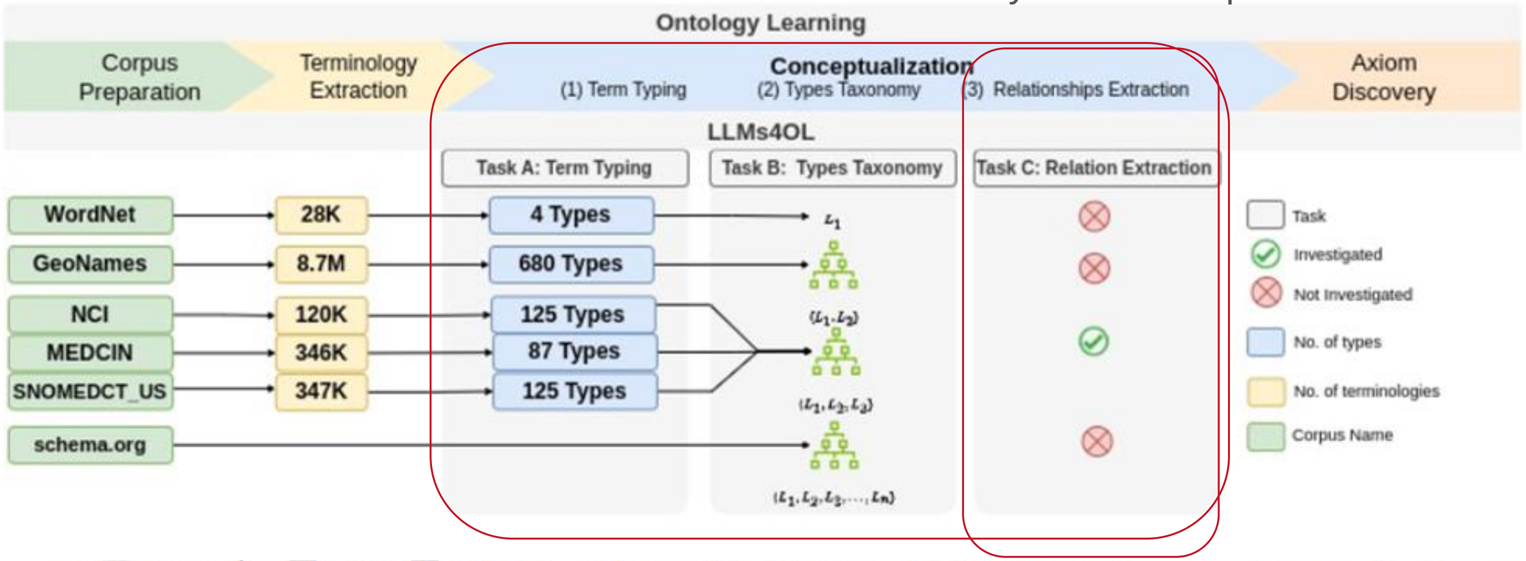
Taxonomic relations essentially define a hierarchy.
E.g.,
Sculptor is_a Artist
Painter is_a Artist
Cubist is_a Painter
Flemish is_a Painter

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



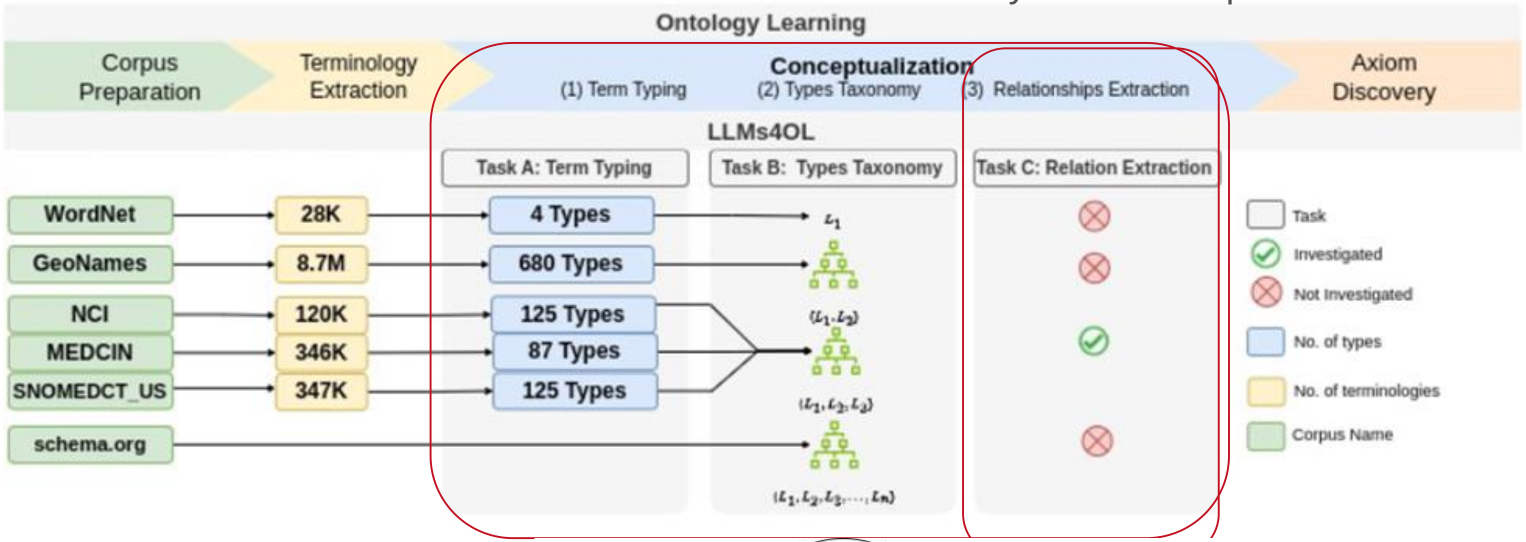
- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

Reference

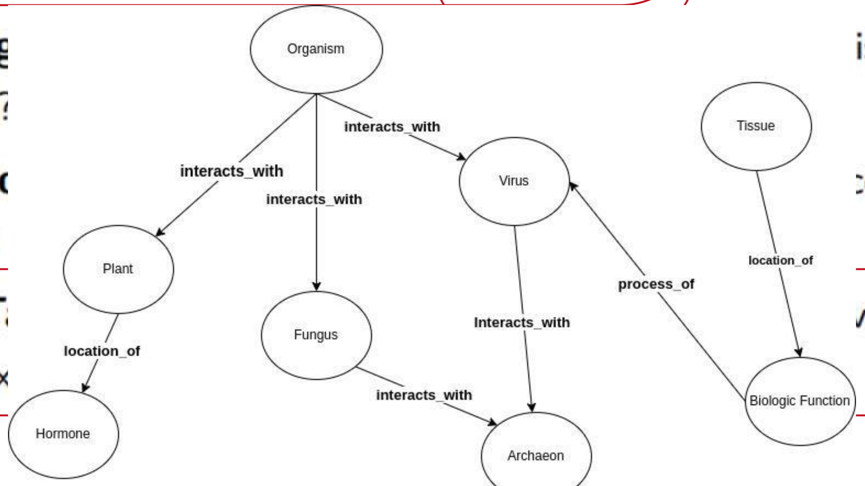
Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



- Task A: **Term Typing** … iscovery
  to construct an ontology? … cognize
- Task B: **Type Taxono**…
  a type taxonomy i.e. the
- Task C: **Type Non-T**… ve are
  LLMs to discover non-tax…

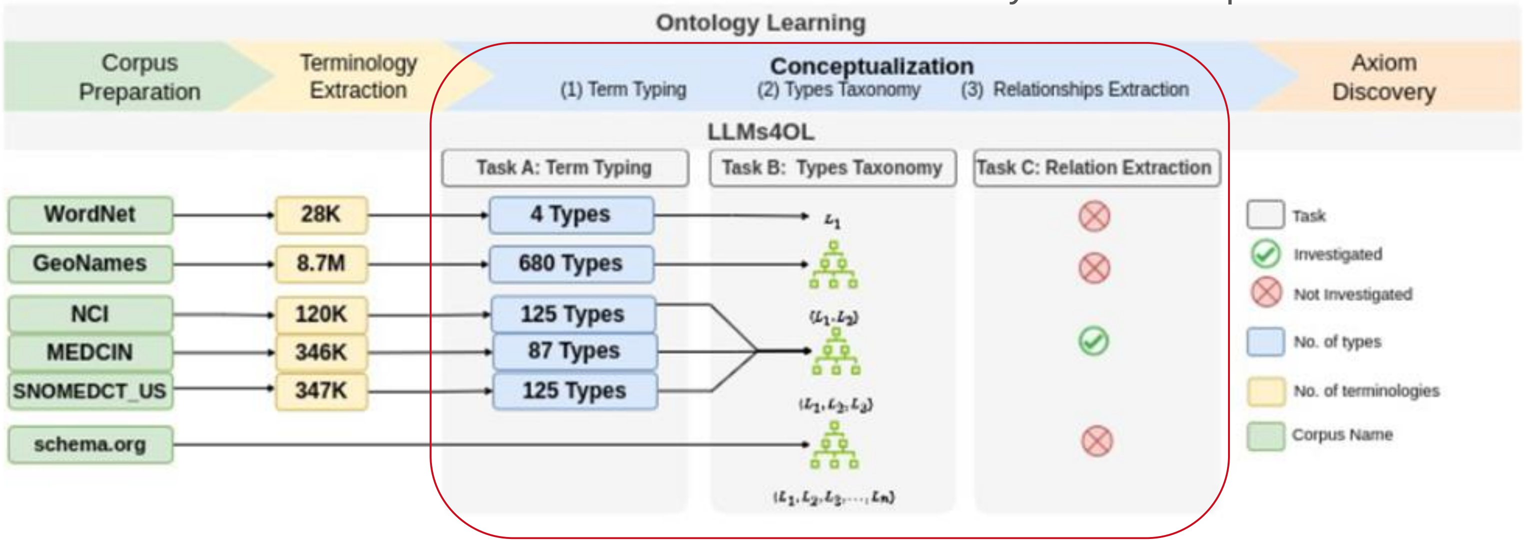


Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK I
  a type t
- TASK (
  LLMs to

Task	Parameter	GeoNames	UMLS	schema.org
Task B	Types	689	127	797
	Levels	2	3	6
	Positive/Negative Samples	680/680	254/254	2,670/2,670
	Train/Test split	272/1,088	101/407	1,086/4,727
Task C	Non-Taxonomic Relations	-	53	-
	Positive/Negative Samples	-	5,641/1,896	-
	Train/Test Split	-	1,507/6,030	-

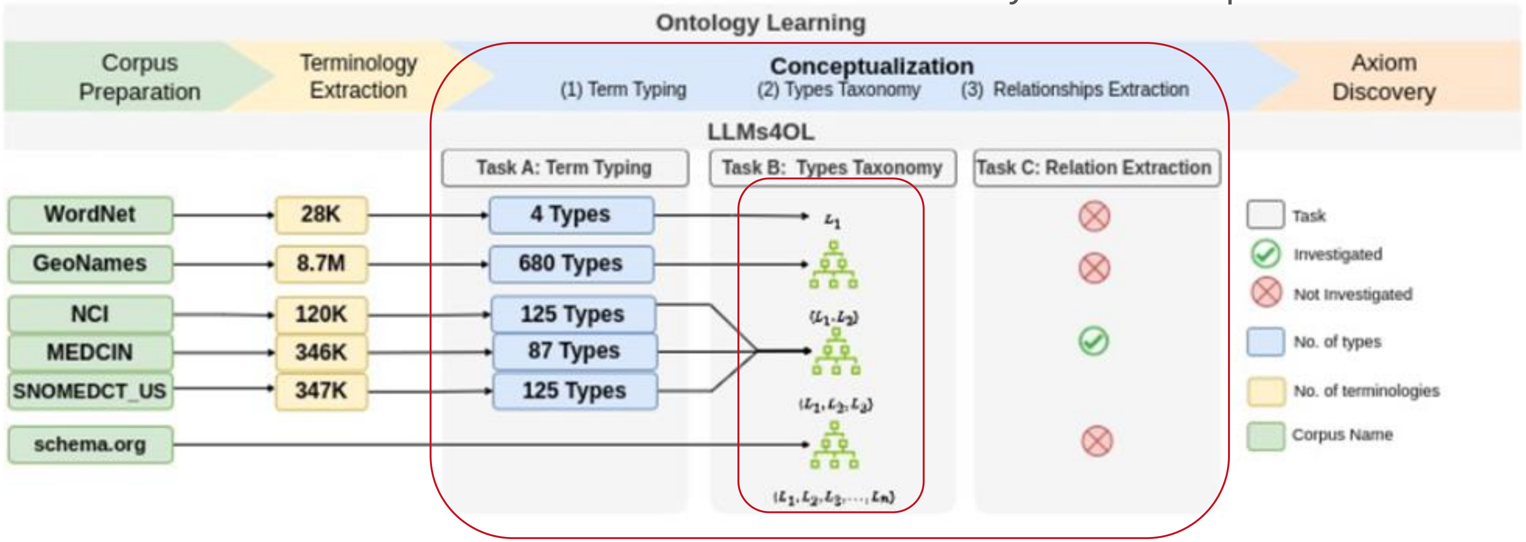
our experimental dataset statistics for Tasks B and C

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK I
  a type t
- TASK (
  LLMs to

Task	Parameter	GeoNames	UMLS	schema.org
Task B	Types	689	127	797
	Levels	2	3	6
	Positive/Negative Samples	680/680	254/254	2,670/2,670
	Train/Test split	272/1,088	101/407	1,086/4,727
Task C	Non-Taxonomic Relations	-	53	-
	Positive/Negative Samples	-	5,641/1,896	-
	Train/Test Split	-	1,507/6,030	-

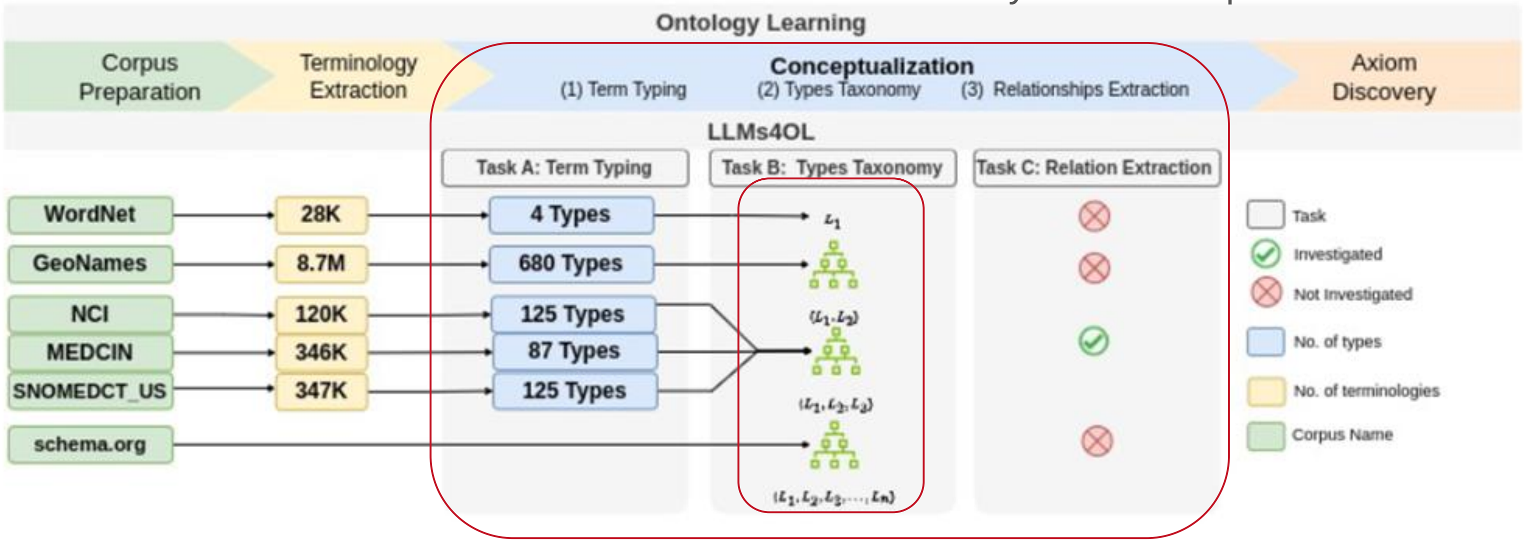
our experimental dataset statistics for Tasks B and C

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Paradigm

The OL task is addressed in three systematic steps.



- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK I
  a type t
- TASK (
  LLMs to

Task	Parameter	GeoNames	UMLS	schema.org
Task B	Types	689	127	797
	Levels	2	3	6
	Positive/Negative Samples	680/680	254/254	2,670/2,670
	Train/Test split	272/1,088	101/407	1,086/4,727
Task C	Non-Taxonomic Relations	-	53	-
	Positive/Negative Samples	-	5,641/1,896	-
	Train/Test Split	-	1,507/6,030	-

our experimental dataset statistics for Tasks B and C

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

RQ1 addresses the performance of LLMs on Task A - Term Typing

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

RQ1 addresses the performance of LLMs on Task A - Term Typing.

- The performance from LLMs is inversely proportional to the increase in domain expertise entailed by the ontologies
  - WordNet (91.7%) < GeoNames (43.1%) < UMLS (NCI, SnomedCT, MEDCIN at 16.1%, 37.7%, 29.8%)

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

RQ2 addresses the performance of LLMs on Task B - Taxonomy discovery between types.

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

RQ2 addresses the performance of LLMs on Task B - Taxonomy discovery between types.

- As seen across the three selected rows and the highest scores highlighted in purple, on average, the performance of LLMs to address Task B is higher than their performance in Task A. This shows LLMs are more effective at inferring "is-a" relations between types.

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	91.7	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	43.3
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	16.1
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	37.7	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	29.8	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	67.8	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	78.1
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	74.4	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	49.5	43.1	42.7	38.8	37.5	20.3	41.3

RQ3 addresses the performance of LLMs on Task C - Non-taxonomic relation discovery between types.

Reference
Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	91.7	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	43.3
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	16.1
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	37.7	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	29.8	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	67.8	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	78.1
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	74.4	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	49.5	43.1	42.7	38.8	37.5	20.3	41.3

RQ3 addresses the performance of LLMs on Task C - Non-taxonomic relation discovery between types.

- An open-sourced model FLAN-T5-XL showed a promising performance of 49.5% on inferring semantic relations.

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

LLMs4OL Evaluation Results

RQ1 – RQ2 – RQ3

Task	Dataset	BERT-Large	PubMedBERT	BART-Large	Flan-T5-Large	Flan-T5-XL	BLOOM-1b7	BLOOM-3b	GPT-3	GPT-3.5	LLaMA-7B	GPT-4
A	WordNet	27.9	-	2.2	31.3	52.2	79.2	79.1	37.9	**91.7**	81.4	90.1
	GeoNames	38.3	-	23.2	13.2	33.8	28.5	28.8	22.4	35.0	29.5	**43.3**
	NCI	11.1	5.9	9.9	9.0	9.8	12.4	15.6	12.7	14.7	7.7	**16.1**
	SNOMEDCT	21.1	28.5	19.8	24.3	31.6	37.0	**37.7**	24.4	25.0	13.8	27.8
	MEDCIN	8.7	15.6	12.7	13.0	18.5	28.8	**29.8**	25.7	23.9	4.9	23.7
B	GeoNames	54.5	-	55.4	59.6	52.4	36.7	48.3	53.2	**67.8**	33.5	55.4
	UMLS	48.2	33.7	49.9	55.3	64.3	38.3	37.5	51.6	70.4	32.3	**78.1**
	schema.org	44.1	-	52.9	54.8	42.7	48.6	51.3	51.0	**74.4**	33.8	74.3
C	UMLS	40.1	42.7	42.4	46.0	**49.5**	43.1	42.7	38.8	37.5	20.3	41.3

As a takeaway message, investigating the state of LLMs for Science w.r.t OL, we found that Bigger was Better!

- Given the current state of LLMs, for tasks entailing a high degree of expertise such a the Ontology learning scientific task, the more the parameters in the LLMs, in turn implying the more knowledge connections held by the LLM, the better they are suited to the task.

Reference

Babaei Giglou, H. et al. (2023, October). LLMs4OL: Large Language Models for Ontology Learning. In *International Semantic Web Conference* (pp. 408-427). Cham: Springer Nature Switzerland.

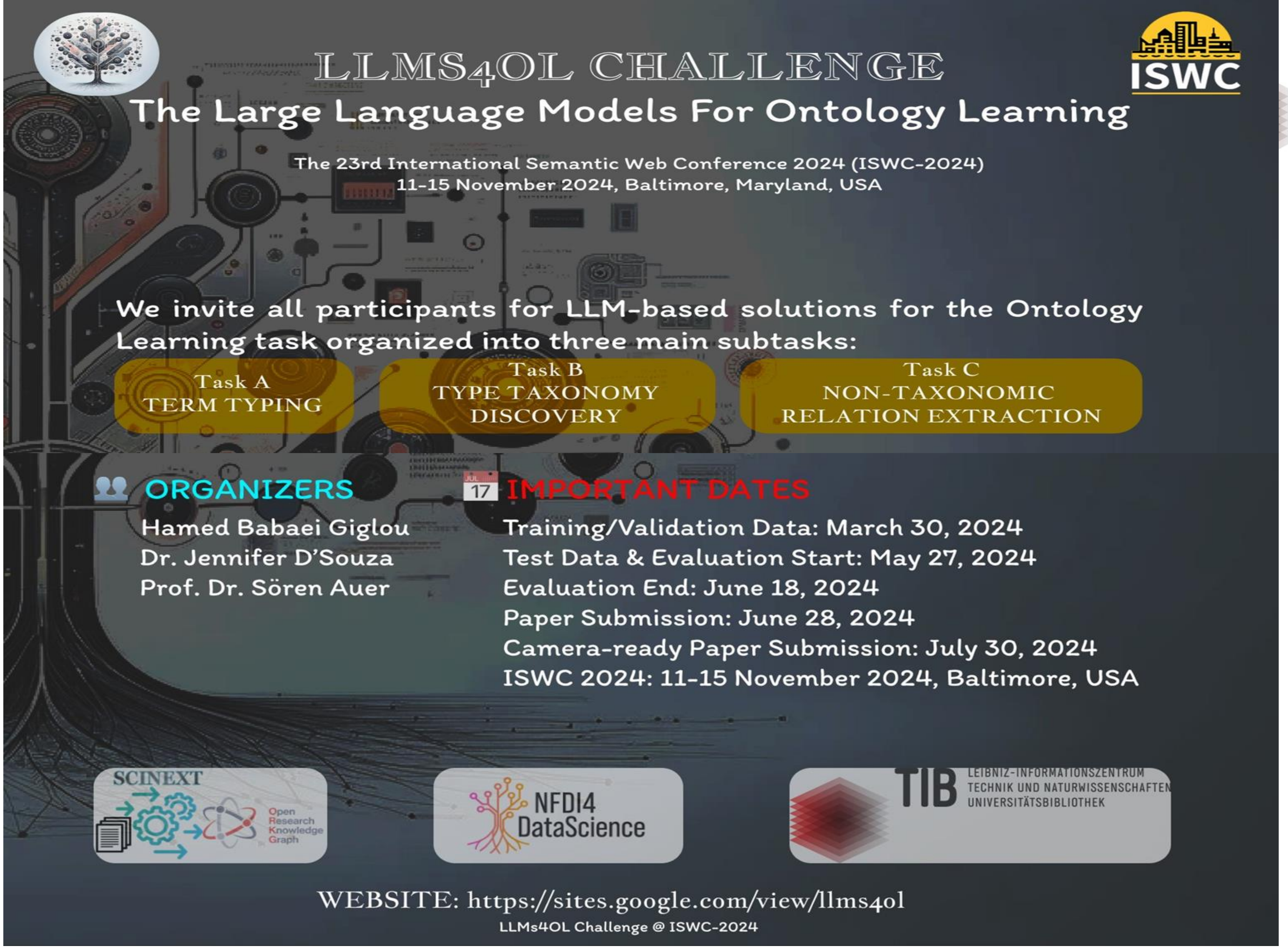
LLMs4OL: Large Language Models for Ontology Learning

LLMs4OL Paper

LLMsOL Github

ISWC 2023 - Research Track

https://github.com/HamedBabaei/LLMs4OL


LLMS4OL CHALLENGE
The Large Language Models For Ontology Learning
ISWC
The 23rd International Semantic Web Conference 2024 (ISWC-2024)
11-15 November 2024, Baltimore, Maryland, USA
We invite all participants for LLM-based solutions for the Ontology Learning task organized into three main subtasks:
Task A
TERM TYPING
Task B
TYPE TAXONOMY DISCOVERY
Task C
NON-TAXONOMIC RELATION EXTRACTION
ORGANIZERS
Hamed Babaei Giglou
Dr. Jennifer D'Souza
Prof. Dr. Sören Auer
IMPORTANT DATES
Training/Validation Data: March 30, 2024
Test Data & Evaluation Start: May 27, 2024
Evaluation End: June 18, 2024
Paper Submission: June 28, 2024
Camera-ready Paper Submission: July 30, 2024
ISWC 2024: 11-15 November 2024, Baltimore, USA
SCINEXT
Open Research Knowledge Graph
NFDI4 DataScience
TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK
WEBSITE: https://sites.google.com/view/llms4ol
LLMs4OL Challenge @ ISWC-2024

TIB

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
  - Research Assistant
  - An LLM-powered Virology Dashboard
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
  - LLMs for evaluating Scientific Synthesis

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
    - Research Assistant
    - An LLM-powered Virology Dashboard
    - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
    - **Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task**
    - LLMs for evaluating Scientific Synthesis

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

LLMs4OM

Matching Ontologies with Large Language Models

Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, Sören Auer

TIB – Leibniz Information Centre for Science and Technology

Hamed.Babaei@tib.eu

Special Track on LLMs for KE, ESWC 2024

May 26-30, Hersonissos, Greece

Leibniz Association

Ontology Matching

- **Ontology Matching (OM)** is a central task in semantic web technologies that aims in finding correspondence between the concepts/classes of two ontologies

- OM is a well-explored research area with diverse methods from traditional techniques to transformer-based methods.
    - **Traditional Word Embeddings:** GloVe
    - **BERT Variants**: BERT, RoBERTa, Sentence-BERT, Bio-ClinicalBERT
    - **LLM:** ByT5, Flan-T5, GPT-3.5, LLaMA-2

- The rapid development of LLMs calls for an in-depth exploration of their potential in OM.

Babaei Giglou, H. (2024). Ontology Matching. Open Research Knowledge Graph, ORKG: **https://orkg.org/comparison/R661569**

How to model OM using LLMs?

1. Naive approach by querying all source and target ontology to LLM and expect matching.

2. Make all the possible pairs of matching and query LLM for a given pair.

Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

How to model OM using LLMs?

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

1. Naive approach by querying all source and target ontology to LLM and expect matching.

- Exceeding LLMs input limits for large ontologies.
- Increases likelihood of erroneous or ”hallucinated” responses due to the volume of information provided.
- Challenge of getting matching scores due to the mixed outputs.

2. Make all the possible pairs of matching and query LLM for a given pair.

- Quadratic time complexity **$O(n^2)$**
  - e.g.: Source Ontology (2k), Target Ontology (3k) → 6M comparison
- It is ideal solution but highly expensive.

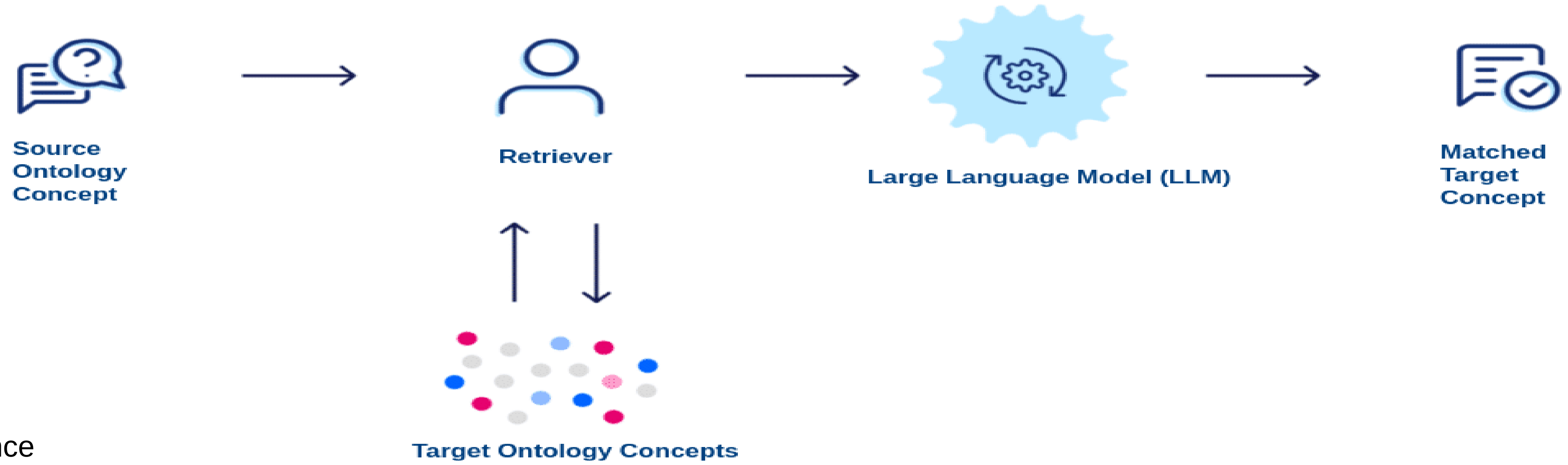
Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

Leibniz Association

Retrieval Augmented Generation Framework

- **RAG –** framework for retrieving facts from an external knowledge base for LLMs better generation process.
- Here in OM**:**
  - **Query (Q)** is **source ontology concepts**.
  - **Knowledge Base (KB)** is **target ontology concepts.**
  - **Retriever Model** finds similar concepts from **target ontology**.
  - **LLM** to finds out which retrieved concept from target ontology is **match** to the query (which is a concept from source ontology).



Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

LLMs4OM Framework

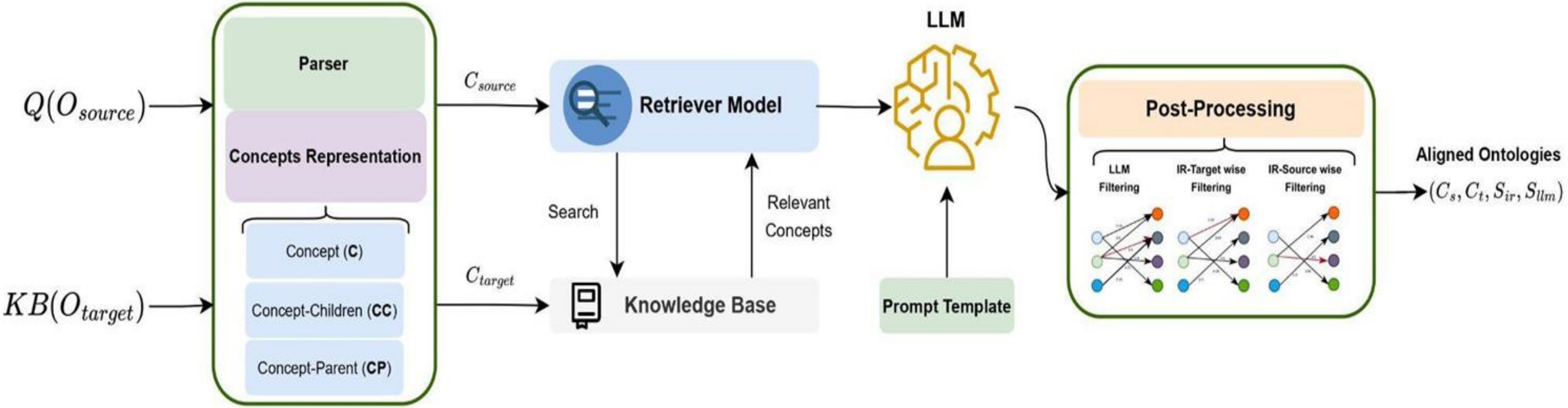
LLMs4OM is a dual-module strategy using **Retrieval Model** and **LLM.**

Components:

- Concept Representation
- Retriever Model
- LLM
- Post-processing

Advantages:

- Reducing time complexity to O(kn)
  - Source Ontology (2k), Target Ontology (3k) → 10K comparison
    - 600 times faster
- LLM-generated confidence score



LLMs4OM: Concept Representation Module

LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

What impact do the ontology representations have on improving matching efficacy?

1. **Concept (C)** – a fundamental representation.
2. **Concept-Parent (CP)** – extending beyond individual concept and including hierarchical relationship.
3. **Concept-Children (CC)** – complementing the concept-parent representation.

ENVO-SWEET	C	CP	CC
http://purl.obolibrary.org/obo/ENVO_00000109	woodland area	woodland area vegetated area	woodland area forested area
http://sweetontology.net/matrPlant/Scrub	Scrub	Scrub Vegetation	Scrub

Scrub: A general term for vegetation dominated by shrubs, i.e. low, woody plants, which typically forms an intermediate community between grass or heath and high forest.

Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

Leibniz Association

LLMs4OM: Retriever Model & LLM Modules

Which Retriever Model?

Which LLM?

Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

LLMs4OM: Retriever Model & LLM Modules

Which Retriever Model?

- **4 Retriever Models:**
  - TFIDF
  - SPECTER2 – a scientific specific variant of BERT
  - Sentence-BERT
  - OpenAI ada-text-embedding

Which LLM?

- **8 LLMs:**
  - Falcon (7B)
  - LLaMA-2 (7B)
  - MPT (7B)
  - GPT-3.5 (174B)
  - Mistral (7B)
  - Vicuna (7B)
  - Mamba (3B)
  - BioMistral (7B) – for biomedical domain only.

Prompt based classification by calculating generation probabilities for ”yes” and ”no” classes using label words such as yes/true/right for the ”yes” class and no/false/wrong for the ”no” class.

Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

LLMs4OM: LLM Prompt Templates

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
{source_concept}
Second concept:
{target_concept}
Answer:

Concept (C)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
{source_concept}
Parents: {source_concept_parents}
Second concept:
{target_concept}
Parents: {target_concept_parents}
Answer:

Concept-Parent (CP)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
{source_concept}
Children: {source_concept_children}
Second concept:
{target_concept}
Children: {target_concept_children}
Answer:

Concept-Children (CC)

LLMs4OM: LLM Prompt Templates - Negative Example

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Second concept:
Vascular Endothelium
Answer:

Concept (C)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Parents: organ system
Second concept:
Vascular Endothelium
Parents: Endothelium, Blood Vessel Tissue
Answer:

Concept-Parent (CP)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Children: vascular system
Second concept:
Vascular Endothelium
Children: Arterial System Endothelium, Venule Endothelium, Lymphatic Vessel Endothelium
Answer:

Concept-Children (CC)

LLMs4OM: LLM Prompt Templates - Negative Example


```
Classify if two concepts refer to the same real world entity or not (answer only yes or no).
### First concept:
cardiovascular system
### Second concept:
Vascular Endothelium
### Answer:
```

Concept (C)

```
Classify if two concepts refer to the same real world entity or not (answer only yes or no).
### First concept:
cardiovascular system
Parents: organ system
### Second concept:
Vascular Endothelium
Parents: Endothelium, Blood Vessel Tissue
### Answer:
```

Concept-Parent (CP)

```
Classify if two concepts refer to the same real world entity or not (answer only yes or no).
### First concept:
cardiovascular system
Children: vascular system
### Second concept:
Vascular Endothelium
Children: Arterial System Endothelium, Venule Endothelium, Lymphatic Vessel Endothelium
### Answer:
```

Concept-Children (CC)

LLMs4OM: LLM Prompt Templates - Negative Example

Concept (C)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Second concept:
Vascular Endothelium
Answer:

Concept-Parent (CP)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Parents: organ system
Second concept:
Vascular Endothelium
Parents: Endothelium, Blood Vessel Tissue
Answer:

Concept-Children (CC)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Children: vascular system
Second concept:
Vascular Endothelium
Children: Arterial System Endothelium, Venule Endothelium, Lymphatic Vessel Endothelium
Answer:

LLMs4OM: LLM Prompt Templates - Positive Example

Concept (C)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
dilatation
Second concept:
aneurysm
Answer:

Concept-Parent (CP)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
dilatation
Parents: **abnormal vascular morphology**
Second concept:
aneurysm
Parents: **abnormal cardiovascular system morphology**
Answer:

Concept-Children (CC)

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
dilatation
Children: dilatation of an abdominal artery, **aortic aneurysm**, dilatation of the ventricular cavity
Second concept:
aneurysm
Children: intestinal microaneurysm, gastric microaneurysm, **aortic aneurysm**
Answer:

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

1 ontology pair in the Anatomy domain

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

8 ontology pairs in Biodiversity

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

2 ontology pairs in the Phenotype research domain

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

2 ontology pairs, i.e. nell-dbpedia and yago-wikidata, with commonsense knowledge graphs

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

5 ontology pairs in biomedicine and machine learning research

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

Evaluation Datasets – OAEI Tracks & Tasks

Ontology Alignment Evaluation Initiative (OAEI)

https://oaei.ontologymatching.org/

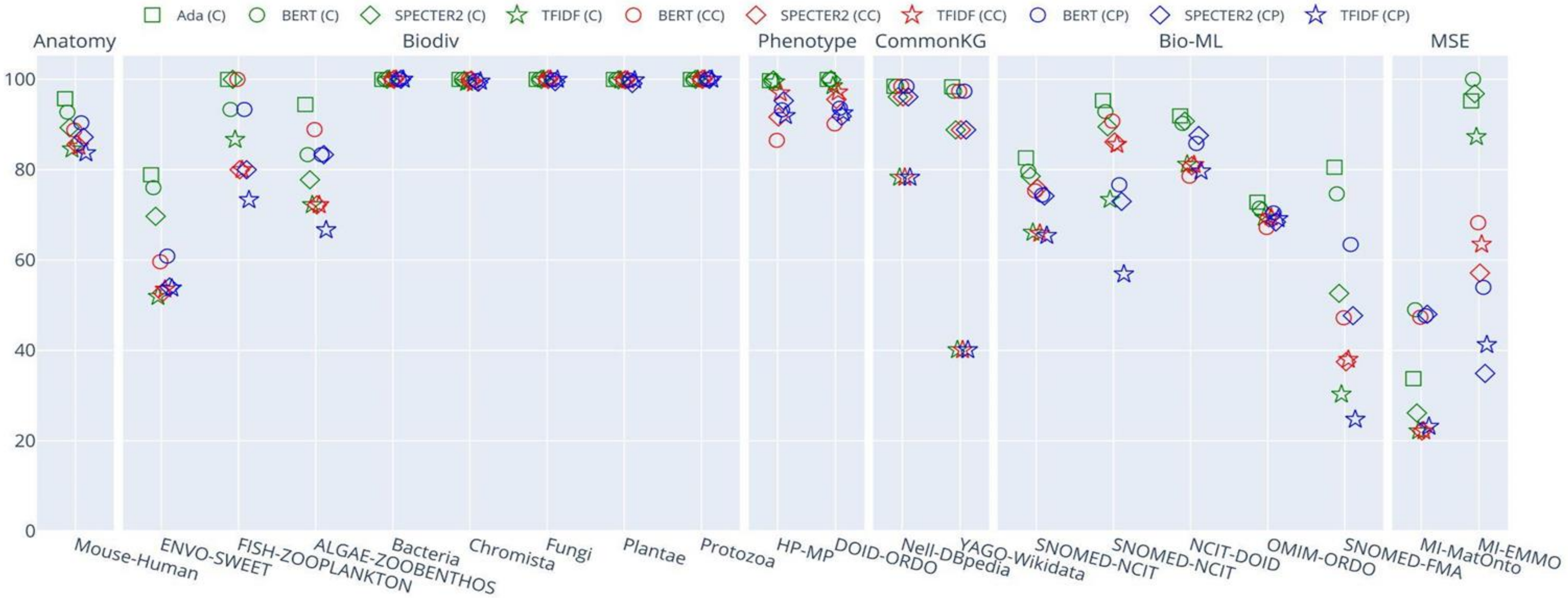
6 domains of knowledge or track defined in OAEI and 20 total paired ontologies for evaluation of the OM task.

2 ontology pairs in material science

Track	Task	Labels		Children		Parents		Alig
		Source	Target	Source	Target	Source	Target	
anatomy	mouse-human	2737	3298	482	673	1687	3297	1516
biodiv	envo-sweet	6566	4525	2695	1256	6109	4514	805
	fish-zooplankton	145	56	145	56	34	7	15
	algae-zoobenthos (Macro)	108	128	108	123	24	27	18
	taxrefld-ncbi (Bacteria)	312	326	137	151	311	325	175
	taxrefld-ncbi (Chromista)	2290	2344	933	966	2289	2343	1405
	taxrefld-ncbi (Fungi)	12732	13149	2716	3138	12731	13148	10162
	taxrefld-ncbi (Plantae)	26302	27013	7324	8003	26301	27012	19914
	taxrefld-ncbi (Protozoa)	501	538	147	184	500	537	357
phenotype	doid-ordo	15511	13504	4506	961	13116	13497	1237
	hp-mp	40703	37965	16789	16324	40531	37513	696
commonkg	nell-dbpedia	134	137	0	0	0	0	129
	yago-wikidata	304	304	0	0	0	0	304
bio-ml	ncit-doid.disease	15762	8465	2440	2074	7880	8464	4686
	omim-ordo.disease	9648	9275	519	1026	4215	9270	3721
	snomed-fma.body	34418	88955	8373	28636	13459	88950	7256
	snomed-ncit.neoplas	22971	20247	1302	2706	2693	8560	3804
	snomed-ncit.pharm	29500	22136	1300	2284	3527	19030	5803
mse	MI-EMMO	545	903	64	232	536	704	63
	MI-MatOnto	545	825	64	114	536	793	302

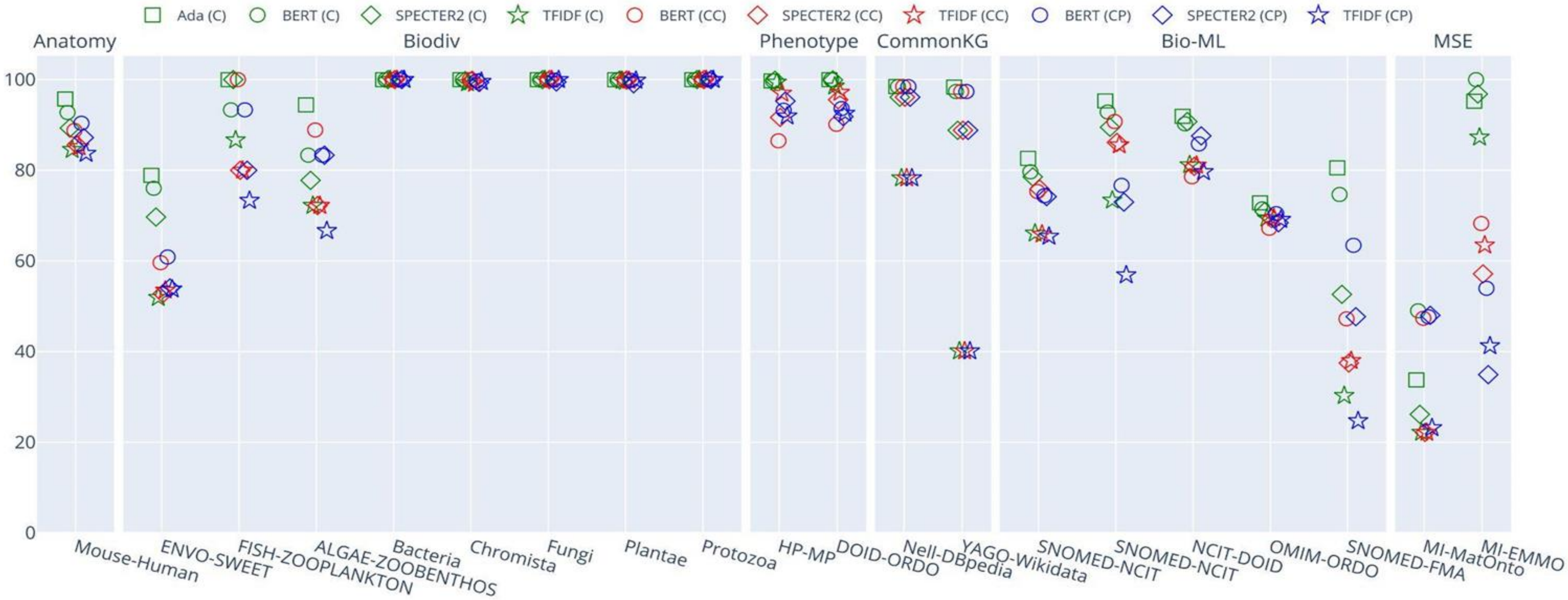
Results - Retriever Models (Recall Analysis)

- **What impact do the three concept representations, respectively have on improving matching efficacy?**
  - **Concept (C)** superiority!
  - **Bio-ML** track tasks are more sensitive to the concept representation.



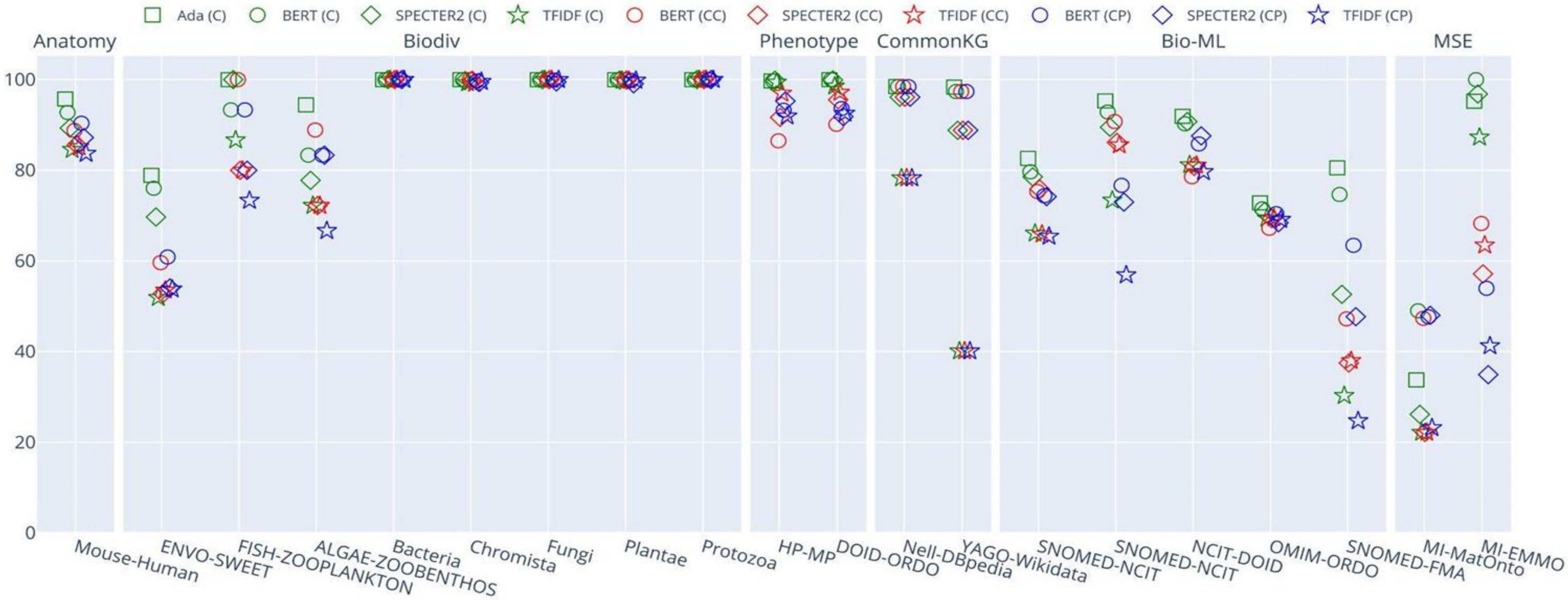
Results - Retriever Models (Recall Analysis)

- **Which retriever performs best per task?**
  - For most of the tasks **Sentence-BERT (MSE and Phenotype)** and **OpenAI ada-text-embedding** works the best.
  - Challenging task: **MI-MatOnto** task, sentence-BERT achieves a 49% recall



Results - Retriever Models (Recall Analysis)

- **How does recall vary in the retrieval module across our different retrieval techniques employed?**
  - **Averaged Recalls:** 82.09% (k=5), 84.66% (k=10), 86.82% (k=20)
  - **OpenAI ada-text-embedding 90.88% > Sentence-BERT 86.09%** > SPECTER2 82.10% > TFIDF 75.15%



Results - LLMs4OM Framework (summary of best performers)

- **What impact do the three concept representations, respectively have on improving matching efficacy?**

Track	Tasks	Prec	Rec	F1	Best Model	OAEI
ANATOMY	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10
BIODIV	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70
PHENOTYPE	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80
COMMONKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00
BIO-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90

Results - LLMs4OM Framework (summary of best performers)

- **What impact do the three concept representations, respectively have on improving matching efficacy?**
  - **Concept (C)** representation excels in **6 tasks**. While, **Concept-Parent (CP)** outperforms in **9 tasks**.

Track	Tasks	Prec	Rec	F1	Best Model	OAEI
Anatomy	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10
Biodiv	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70
Phenotype	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80
CommonKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00
Bio-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90

Results - LLMs4OM Framework (summary of best performers)

- **What impact do the three concept representations, respectively have on improving matching efficacy?**
  - **Concept (C)** representation excels in **6 tasks**. While, **Concept-Parent (CP)** outperforms in **9 tasks**.
  - So, the inclusion of information from taxonomy is valuable for LLMs in enhancing their understanding of concepts.

Track	Tasks	Prec	Rec	F1	Best Model	OAEI
Anatomy	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10
Biodiv	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70
Phenotype	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80
CommonKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00
Bio-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90

Results - LLMs4OM Framework (summary of best performers)

- **Which LLM performs best per track?**

Track	Tasks	Prec	Rec	F1	Best Model	OAEI	
ANATOMY	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10	
BIODIV	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40	
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80	
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00	OLaLa
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80	LogMapLt
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30	LogMapLt
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10	OLaLa
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60	OLaLa
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70	OLaLa
PHENOTYPE	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50	AML
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80	LogMap
COMMONKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00	
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00	
BIO-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80	
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50	
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50	
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10	
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20	
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80	Matcha
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90	

Results - LLMs4OM Framework (summary of best performers)

- **Which LLM performs best per track?**
  - Bests: **GPT-3.5 > Mistral-7B** > LLaMA-2-7B > MPT-7B
  - **OpenAI ada-text-embedding** performed better than BERT when combined with LLM.

Track	Tasks	Prec	Rec	F1	Best Model	OAEI	
ANATOMY	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10	
BIODIV	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40	
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80	
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00	OLaLa
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80	LogMapLt
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30	LogMapLt
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10	OLaLa
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60	OLaLa
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70	OLaLa
PHENOTYPE	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50	AML
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80	LogMap
COMMONKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00	
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00	
BIO-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80	
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50	
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50	
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10	
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20	
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80	Matcha
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90	

Results - LLMs4OM Framework (summary of best performers)

- **Which LLM performs best per track?**
  - Bests: **GPT-3.5 > Mistral-7B** > LLaMA-2-7B > MPT-7B
  - **OpenAI ada-text-embedding** performed better than BERT when combined with LLM.
  - LLMs4OM stood out in **9 tasks** and performed the better performance w.r.t state-of-the-art.

Track	Tasks	Prec	Rec	F1	Best Model	OAEI	
ANATOMY	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10	
BIODIV	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40	
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80	
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00	OLaLa
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80	LogMapLt
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30	LogMapLt
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10	OLaLa
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60	OLaLa
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70	OLaLa
PHENOTYPE	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50	AML
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80	LogMap
COMMONKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00	
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00	
BIO-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80	
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50	
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50	
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10	
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20	
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80	Matcha
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90	

Low performance of LLMs4OM on the Bio-ML track

- Despite strong retriever performance in candidate retrieval, LLMs' overall performance remains low in Bio-ML track..

Bio-ML Track Tasks	*BioMistral-7B*	*GPT-3.5 +Ada*
NCIT-ORDO	69.04	83.01
OMIM-ORDO	57.84	64.14
SNOMED-FMA	**33.98**	25.64
SNOMED-NCIT(neoplas)	46.24	49.47
SNOMED-NCIT(pharm)	62.00	68.02

Knowledge that this work contributes toward future advancements

- The complementary integration of the Retrieval Augmented Generation (RAG) paradigm and Large Language Models (LLMs) to effectively support downstream tasks in semantic web.
- A well-structured Python-based framework to support all steps of dataset, representations, and model integration for OM.

Future Challenges:

- New LLMs are rapidly released, so experimental investigations need to be continuously updated.
- It is hard to know how much of the evaluation data is being used in training those LLMs.

Reference

Giglou, H. B., D'Souza, J., & Auer, S. (2024). LLMs4OM: Matching Ontologies with Large Language Models. *arXiv preprint arXiv:2404.10317*.

Thank You!

LLMs4OM Paper

LLMs4OM Github

https://github.com/HamedBabaei/LLMs4OM

Plan for the Talk

- A walkthrough of various applications of LLMs for Science:
  - Research Assistant
  - An LLM-powered Virology Dashboard
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Learning (OL) task
  - Empirical Evaluation of Various LLMs for Science w.r.t. the Ontology Matching (OM) task
  - **LLMs for evaluating Scientific Synthesis**

LLMs as Evaluators for Scientific Synthesis: Background

- LLMs offer substantial benefits in streamlining machine learning model development, particularly in evaluation processes.
  - They reduce the dependency on human-generated ground truth data and the necessity for human evaluators in two key ways:
    - by facilitating the generation of synthetic ground truth data and
    - by serving as evaluators for model predictions themselves.
  - This approach not only speeds up the evaluation process but also broadens the scope of evaluation criteria to include factors such as diversity and coverage, enhancing the efficiency and comprehensiveness of model assessments.

Reference

- Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., ... & Hou, L. (2024). Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, *36*.

LLMs as Evaluators for Scientific Synthesis: Contributions

- Our work investigated the use of LLMs as evaluators to streamline the evaluation process, moving away from traditional reliance on human evaluators and human-generated ground truth data.
  - It specifically examined the effectiveness of LLMs in synthesizing scientific abstracts seen generally as a multi-document summarization task.
    - Let’s visit an example synthesis task on ORKG Ask https://ask.orkg.org/
  - The main focus of the research was to assess how two state-of-the-art LLMs—the proprietary GPT-4 Turbo and the open-source Mistral-7B—perform in evaluating scientific syntheses.
    - Leveraging LLMs meant better versatility in evaluation considerations, which meant that the evaluations tested varied dimensions of syntheses quality, viz. comprehensiveness, trustworthiness, and utility.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

LLMs as Evaluators: Related Work

- Several recent works have compared LLMs' text evaluations to human evaluations on multiple tasks, and report that LLMs produce results similar to human judgements
  - One work finds only minor variations in results depending on task instructions and hyperparameters, whereas they find a high degree of variation in performance of different LLMs and the quality characteristics being assessed (Chiang and Lee, 2023b).
    - In evaluating the quality of story fragments by *grammaticality*, *cohesiveness*, *likability*, and *relevance*, they find only a weak correlation between humans and LLMs on *grammaticality*, but a moderate correlation on *relevance*.
  - Another work found ChatGPT's performance sensitive to prompt instructions (Wang et al., 2023). They also showed that ChatGPT evaluations correlate especially well with human evaluations for creative tasks like story generation (Wang et al., 2023).
  - Another work demonstrated that requiring LLMs to provide a justification for their ratings "significantly improved the correlation between the LLMs' ratings and human ratings" (Chiang and Lee, 2023a).

References

- Cheng-Han Chiang and Hung-yi Lee. 2023b. Can Large Language Models Be an Alternative to Human Evaluations? In Proceedings of the 61st ACL (Volume 1: Long Papers), pages 15607–15631, Toronto, Canada.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In Proceedings of the 4th New Frontiers in Summarization Workshop, pages 1–11, Singapore. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023a. A Closer Look into Using Large Language Models for Automatic Evaluation. In Findings of EMNLP 2023, pages 8928–8942.

LLMs as Evaluators: Related Work

- Several recent works have compared LLMs' text evaluations to human evaluations on multiple tasks, and report that LLMs produce results similar to human judgements
  - Closer to our work, only one work has investigated the task of text summarization evaluation.
    - They evaluated single-document news article summaries on the aspects of *coherence*, *consistency*, *fluency*, and *relevance*; their results exceed the correlation with human judgements of most automatic approaches, including ROUGE.

References

- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

LLMs as Evaluators for Scientific Synthesis: Motivation

- The accurate evaluation of scientific syntheses is a critical task in research, ensuring the integrity and reliability of the synthesized information.
  - While recent advancements have demonstrated the efficacy of LLMs in generating such syntheses (Pride et al., 2023), also known as the CORE-GPT work, their potential in evaluating them remains relatively unexplored.
- Motivated by the success of LLMs in other text evaluation tasks, our work seeks to investigate the suitability of LLMs for the task of assessing the quality of scientific syntheses.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024.
- David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In Linking Theory and Practice of Digital Libraries, pages 146–159. Springer Nature Switzerland.

LLMs as Evaluators for Scientific Synthesis: Task Setup

TIB

- Dataset of Scientific Syntheses: CORE-GPT dataset
  - This dataset comprises 100 research questions spanning 20 diverse domains, each accompanied by the titles and abstracts of five related works and an answer to the research question generated by GPT-4 by synthesizing the provided abstracts.
  - Additionally, human ratings from two annotators, on a scale of 0 to 10, are available on the quality of each synthesis in three dimensions, viz. *comprehensive*, *trust*, and *utility*.
- Two LLM Evaluators: GPT-4 Turbo and Mistral 7B
  - Evaluation Prompt: It contains two lines of task instruction, explanation of the quality aspects (as defined for the CORE-GPT dataset annotators) and the rating scale, response format instructions, and finally the answer to be evaluated with its question and abstracts. The response is requested in JSON format, with a numeric rating between 0 and 10 for each aspect as well as a rationale for each rating.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024.
- David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In Linking Theory and Practice of Digital Libraries, pages 146–159. Springer Nature Switzerland.

LLMs as Evaluators for Scientific Synthesis: Task Setup

- Evaluation Prompt

A Prompt

A.1 Main Evaluation Prompt

Evaluate the quality of the following question and answer pair. The answer should succinctly address the question while referencing only content from the provided abstracts.

Consider the following characteristics while assessing the answer. For each characteristic rate the quality from 0 (poor) to 10 (excellent). Provide a short rationale for each rating.

1. Comprehensive: How comprehensively is the question answered?

2. Trust: How trustworthy is the answer?

3. Utility: How useful is the answer?

Return your response in JSON format: {Comprehensive : {'rating' : '', 'rationale' : ''}, Trust : {'rating' : '', 'rationale' : ''}, Utility : {'rating' : '', 'rationale' : ''}}

Question: <content>

Answer: <content>

Abstracts: <content>

A.2 Additional Mistral System Prompt

Follow the user instructions to the letter. The output should be only what the user requests, do not respond with anything else. Provide the answer in JSON with no additional text.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024

LLMs as Evaluators for Scientific Synthesis: Evaluation Output

B Example GPT-4 Turbo Output

GPT-4 Turbo

COMPREHENSIVE: rating: 7, rationale: The answer provides a broad overview of how living in a low-income neighborhood can affect social networks and well-being, referencing various studies that discuss mental health, social inclusion, and subjective well-being. However, it does not fully elaborate on the specific mechanisms or provide detailed insights from each abstract, such as the role of men in community well-being from abstract 3 or the impact of natural capital from abstract 4.
TRUST: rating: 8, rationale: The answer is trustworthy as it directly references findings from the provided abstracts, indicating a reliance on research-based evidence. However, the answer could be more trustworthy if it included more specific data or findings from the abstracts to support the claims.
UTILITY: rating: 7, rationale: The answer is useful as it gives an overview of the potential impacts of living in a low-income neighborhood on residents' social networks and well-being. It could be more useful if it provided actionable insights or recommendations based on the findings of the studies.

Figure 1: Example output from GPT-4 Turbo.

C Example Mistral Output

Mistral

COMPREHENSIVE: rating: 9, rationale: The answer provides specific examples and references to studies that directly address the question of how social media influences the formation and maintenance of relationships among young adults in the United States, specifically focusing on LGBTQ+ individuals with disabilities and young adults who have moved abroad.
TRUST: rating: 10, rationale: The answer is based on empirical research and provides clear citations to support the claims made.
UTILITY: rating: 10, rationale: The answer is directly relevant to the question and provides useful information for those interested in understanding how social media impacts relationship formation and maintenance among young adults in the United States.

The given question and answer pair is of high quality. The question is clear, specific, and comprehensive, while the answer is well-researched, trustworthy, and useful. It directly addresses the question by providing examples from empirical studies that explore how social media influences the formation and maintenance of relationships among young adults in the United States,

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024

LLMs as Evaluators for Scientific Synthesis: Discussion

- An overview of LLM performance was obtained by reviewing one synthesis from each domain evaluated by both GPT-4 and Mistral. Qualitatively, both models demonstrated credible and logically consistent ratings and rationales. GPT-4 provided more detailed rationales compared to Mistral, with slightly lower ratings overall.
- Correlation

	A1	A2	GPT-4	Mistral
A1				
ρ	-	**0.710**	0.248	0.015
p-value	-	**0.001**	0.305	0.951
A2				
ρ	**0.710**	-	0.058	-0.038
p-value	**0.001**	-	0.814	0.878
GPT-4				
ρ	0.248	0.058	-	**0.786**
p-value	0.305	0.814	-	**0.000**
Mistral				
ρ	0.015	-0.038	**0.786**	-
p-value	0.951	0.878	**0.000**	-

Table 1: Spearman's ρ calculated for the combined mean of *Comprehensive*, *Trust*, and *Utility* scores. Statistically significant results are in bold.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024.
- David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In Linking Theory and Practice of Digital Libraries, pages 146–159. Springer Nature Switzerland.

LLMs as Evaluators for Scientific Synthesis: Discussion

- An overview of LLM performance was obtained by reviewing one synthesis from each domain evaluated by both GPT-4 and Mistral. Qualitatively, both models demonstrated credible and logically consistent ratings and rationales. GPT-4 provided more detailed rationales compared to Mistral, with slightly lower ratings overall.
- Correlation

	A1	A2	GPT-4	Mistral
A1				
ρ	-	**0.710**	0.248	0.015
p-value	-	**0.001**	0.305	0.951
A2				
ρ	**0.710**	-	0.058	-0.038
p-value	**0.001**	-	0.814	0.878
GPT-4				
ρ	0.248	0.058	-	**0.786**
p-value	0.305	0.814	-	**0.000**
Mistral				
ρ	0.015	-0.038	**0.786**	-
p-value	0.951	0.878	**0.000**	-

Table 1: Spearman's ρ calculated for the combined mean of *Comprehensive*, *Trust*, and *Utility* scores. Statistically significant results are in bold.

Human annotators exhibited
a strong positive correlation (0.710)

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024.
- David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In Linking Theory and Practice of Digital Libraries, pages 146–159. Springer Nature Switzerland.

LLMs as Evaluators for Scientific Synthesis: Discussion

- An overview of LLM performance was obtained by reviewing one synthesis from each domain evaluated by both GPT-4 and Mistral. Qualitatively, both models demonstrated credible and logically consistent ratings and rationales. GPT-4 provided more detailed rationales compared to Mistral, with slightly lower ratings overall.
- Correlation

	A1	A2	GPT-4	Mistral
A1				
ρ	-	**0.710**	0.248	0.015
p-value	-	**0.001**	0.305	0.951
A2				
ρ	**0.710**	-	0.058	-0.038
p-value	**0.001**	-	0.814	0.878
GPT-4				
ρ	0.248	0.058	-	**0.786**
p-value	0.305	0.814	-	**0.000**
Mistral				
ρ	0.015	-0.038	**0.786**	-
p-value	0.951	0.878	**0.000**	-

Table 1: Spearman's ρ calculated for the combined mean of *Comprehensive*, *Trust*, and *Utility* scores. Statistically significant results are in bold.

as did GPT-4 Turbo and Mistral (0.786)

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024

LLMs as Evaluators for Scientific Synthesis: Discussion

- An overview of LLM performance was obtained by reviewing one synthesis from each domain evaluated by both GPT-4 and Mistral. Qualitatively, both models demonstrated credible and logically consistent ratings and rationales. GPT-4 provided more detailed rationales compared to Mistral, with slightly lower ratings overall.
- Correlation

	A1	A2	GPT-4	Mistral
A1				
ρ	-	**0.710**	0.248	0.015
p-value	-	**0.001**	0.305	0.951
A2				
ρ	**0.710**	-	0.058	-0.038
p-value	**0.001**	-	0.814	0.878
GPT-4				
ρ	0.248	0.058	-	**0.786**
p-value	0.305	0.814	-	**0.000**
Mistral				
ρ	0.015	-0.038	**0.786**	-
p-value	0.951	0.878	**0.000**	-

Table 1: Spearman's ρ calculated for the combined mean of *Comprehensive*, *Trust*, and *Utility* scores. Statistically significant results are in bold.

- Correlations between annotators and LLMs were weak or very weak, with p-values indicating insufficient evidence for genuine association.
- These findings suggest LLMs cannot directly replicate human performance in evaluating scientific syntheses.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024

LLMs as Evaluators for Scientific Synthesis: Conclusion

- Both LLMs generally produce credible and logically consistent ratings and rationales, but GPT-4 appears more conservative in its ratings and provides more detail and specific recommendations in its rationales. GPT-4 also displays greater sensitivity to the presence or absence of citations compared to Mistral.

- However, both LLMs' rationales occasionally contained inaccuracies or flaws, raising concerns about the credibility of their scores.

- Moreover, the extent to which the responses are evaluated as *syntheses* and not simply as *answers*, without reliance on general knowledge, remains unclear, particularly in the case of Mistral.

References

- Evans, J., D'Souza, J., & Auer, S. (2024). Large Language Models as Evaluators for Scientific Synthesis. *arXiv preprint arXiv:2407.02977*. | Forthcoming at KONVENS 2024

SCINEXT

- Reflections on some of our research:
  - A FAIR and Free Prompt-based Research Assistant | Demo paper at NLDB 2024, preprint
  - Large Language Models for Scientific Information Extraction: An Empirical Study for Virology
    - AI-powered Virology Dashboard | In: EACL 2024 Findings – paper link
  - LLMs4OL: Large Language Models for Ontology Learning
    - Babaei Giglou, H., D'Souza, J., Auer, S. (2023). LLMs4OL: Large Language Models for Ontology Learning. In: ISWC 2023. https://doi.org/10.1007/978-3-031-47240-4_22
  - LLMs4OM: Matching Ontologies with Large Language Models
    - In: ESWC 2024 Special Track on LLMs for KE | preprint
  - Large Language Models as Evaluators for Scientific Synthesis
    - In: KONVENS 2024 short paper | preprint
  - Survey on measures of quality of crowdsourced data in the Open Research Knowledge Graph (ORKG) for six different domains
    - Quality Assessment of Research Comparisons in the Open Research Knowledge Graph: a Case Study | In: JLIS 2024

References
https://scinext-project.github.io/

Project Acknowledgements

- Open Research Knowledge Graph
  - Work was done in part under the scientific idea of the ORKG of using structured models for salient aspects of scholarly communication.
- SCINEXT
  - Large Language Model research was made possible via the junior AI research group SCINEXT
    - Federal Ministry of Education and Research (BMBF) Grant
    - Föderkennzeichen: 01IS22070
  - More info https://scinext-project.github.io/

Team Members

PhD Candidates

Julia Evans

Mahsa Shamsabadi

Hamed Babaei Giglou

Team Lead

Dr. Jennifer D’Souza

Senior Adviser

Prof. Dr. Sören Auer

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

Bundesministerium
für Bildung
und Forschung

TIB

SCINEXT: Neural-SymboliC InnovatioN EXTraction

Thank you for your attention!

Questions?

Leibniz Association