Towards A Framework For Openness Score Calculation in Scholarly Research

Master Thesis
University of Neuchatel

Jennifer Swaminathan

Jennifer.Swaminathan@unine.ch

12.02.24

Supervisor:

Prof. Dr. Philippe Cudré-Mauroux

Overview

- Introduction
- Research Questions
- Methodology
- Results
- Framework
- Limitations & Future Work
- Conclusion

2

Introduction

- Open Science : Promoting transparency and accessibility
- Collaborative effort : Openness Score Team
- 'Openness' metric : Quantifying accessibility in research
- Visualization
- Framework for Openness Score

Research Questions:

RQ1 Which scholarly APIs are effective for analyzing openness in academic publications?

RQ2 How does API efficiency compare to a local server in assessing scholarly publications' openness?

RQ3 What is the efficacy of LLMs in deriving additional insights for a novel openness metric from open-access articles?

RQ1

Methodology for Comparative Analysis of Scholarly APIs

Data Collection

- Crossref (2000)- Comprehensive metadata retrieval of schorlarly articles
- Doaj (2015) - Open access journal indexing
- Unpaywall (2017) – assigns distinct status of articles
- OpenAlex (2022) – includes a wealth of information from various sources

Data Analysis

- Rate Limits - Determines data retrieval capacity
- Access Speed - Measures API responsiveness
- Metadata Content - Assesses the richness of data
- Public Dump Availability - Evaluates offline data access options

RQ1

API Comparison Results: Rate Limits, Downloads, and Records

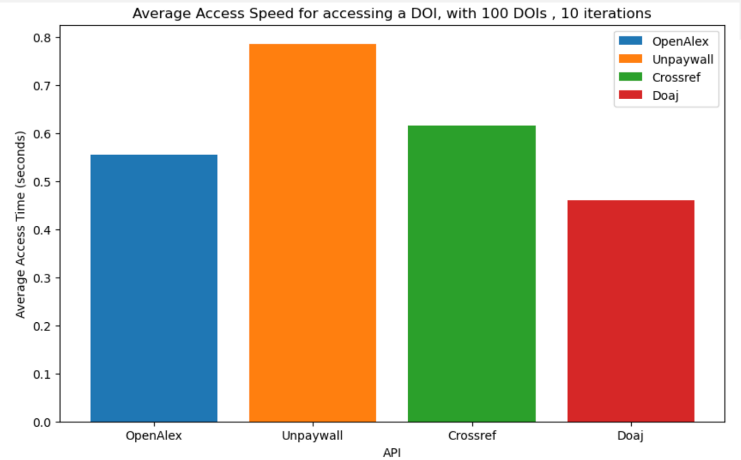
	Rate Limit	Download	Records
Crossref	Varies by authentication: Anonymous, Polite (with email), Full (paid)	185GB, snapshot	145M
Doaj	Not specified; heavy use suggests data dump	6GB on request	9M
Unpaywall	100,000 calls per day	-	48M
OpenAlex	100,000 calls per day, 10 requests per second	350GB, snapshot	253M

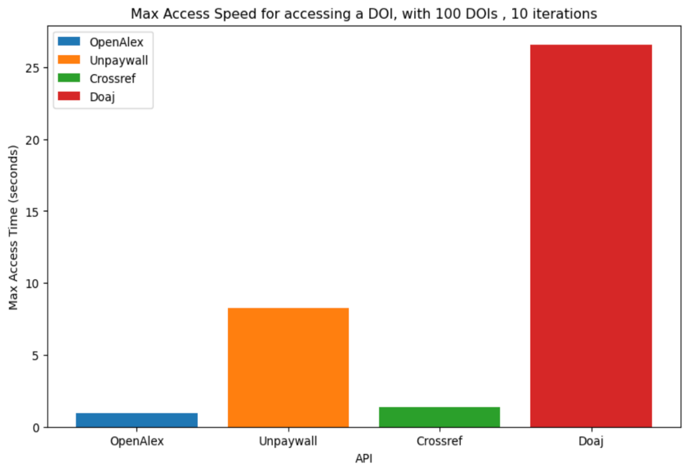
6

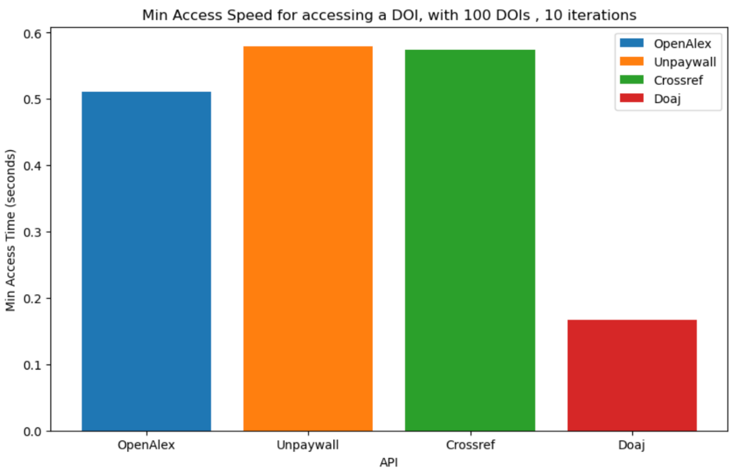
RQ1

API Metadata Access Speed Comparison

- **Experiment:**
  - 100 random articles
  - Measured retrieval speed over 10 iterations.







RQ2

Comparing API vs. Server Performance: Methodological Overview

Data Collection

- Gather data using OpenAlex snapshot

Database Setup

- Configure and optimize a local Postgres database for performance

Data Analysis

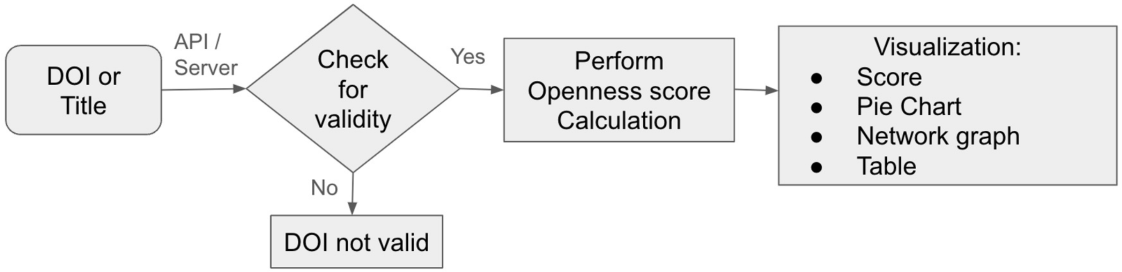
- Measure speed
- Conduct data consistency checks across API and server.

8

RQ2

Results: API vs. Server Access Speed Analysis 1/2

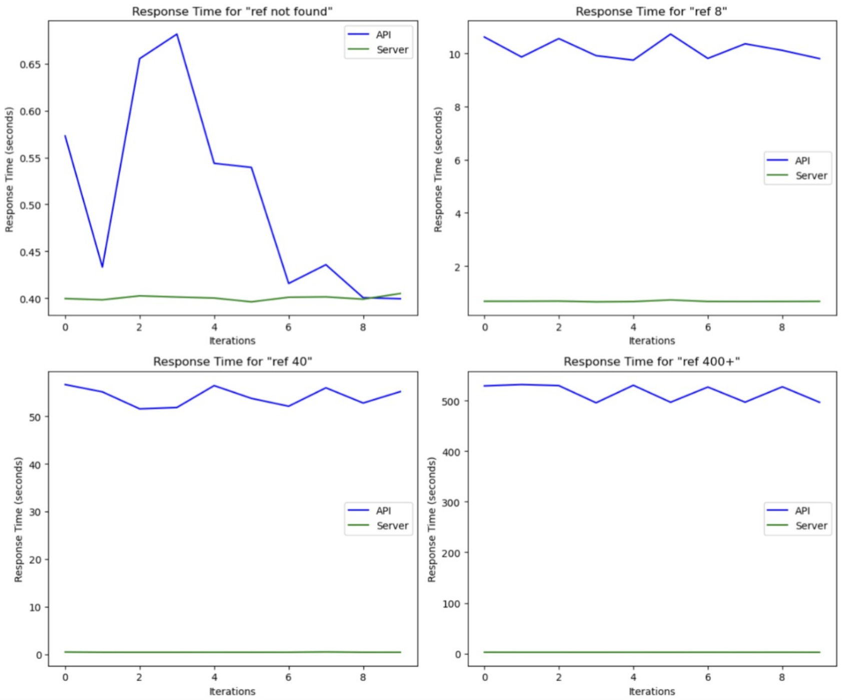
- **Experiment:**
  - 100 articles: 50 DOIs and 50 titles for 10 iterations
  - All randomly chosen to avoid caching the data
  - Gather References of the article and find the open access status

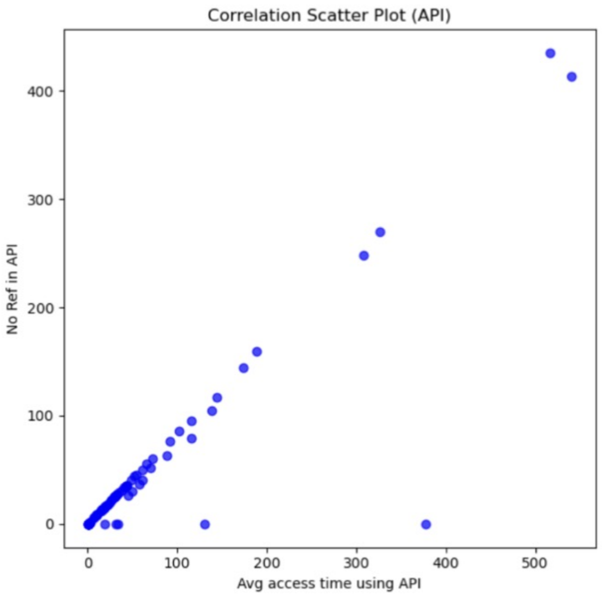


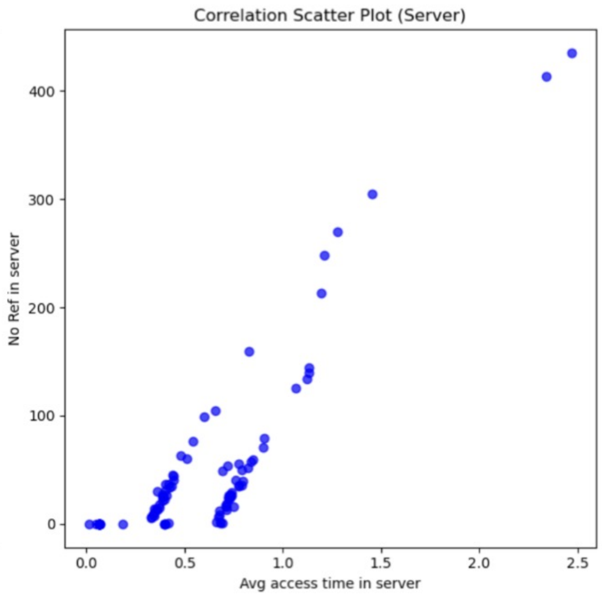
Metric	API(sec)	Server (sec)
Average	50.17	0.58
Min	0.421	0.015
Max	540.17	2.47

RQ2

Results: API vs. Server Access Speed Analysis 2/2







RQ2

Results: Server vs API Data Consistency

Server Snapshot: Data as of October 18, 2023

Endpoints	Server	API (Nov 14)	Diff	API (Nov 28)	(API28-API14)
Works	245207435	246139651	932219	246537492	397841
Author	93003987	89468168	-3535819	89565600	97432
Institutions	106956	107247	291	107252	5
Concepts	65073	65073	0	65073	0
Publishers	10250	10250	0	10250	0
Sources	247955	248650	695	248643	-7

After updating:

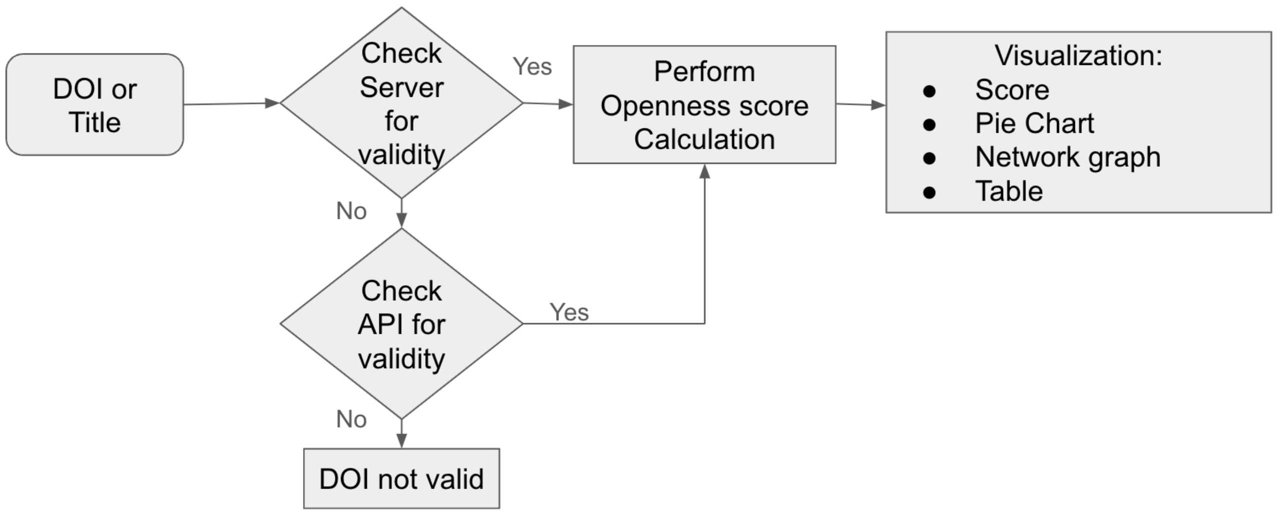
Update: Snapshot as of November 21, 2023

Endpoints	O_Ser	U_Ser	API Nov28	O_Ser-U_Ser	API-U_Ser
Works	245207435	245207435	246537492	0	1330057
Author	93003987	89516053	89565600	3487934	49547
Institutions	106956	107246	107252	-290	6
Concepts	65073	65073	65073	0	0
Publishers	10250	10250	10250	0	0
Sources	247955	248643	248643	-688	0

RQ2

Results: Summary

- **Tradeoff** accuracy vs speed
  - Server is faster
  - API is accurate
- Provide a **hybrid solution**



RQ3

Methodology for Assessing Scholarly Openness Using LLMs

Data Collection

- 38 Open Access PDF articles
- Questions
  - "What are the names of the datasets used in the article to perform the experiment?"
  - "Who are the authors of this article?"
  - "Are the datasets used in the experiment in the article open access?”
- Annotate the answers for the articles
- Question Answering models

Data Analysis

- Quality of responses, and score them
- F1 score, Accuracy, Recall and Precision

13

RQ3

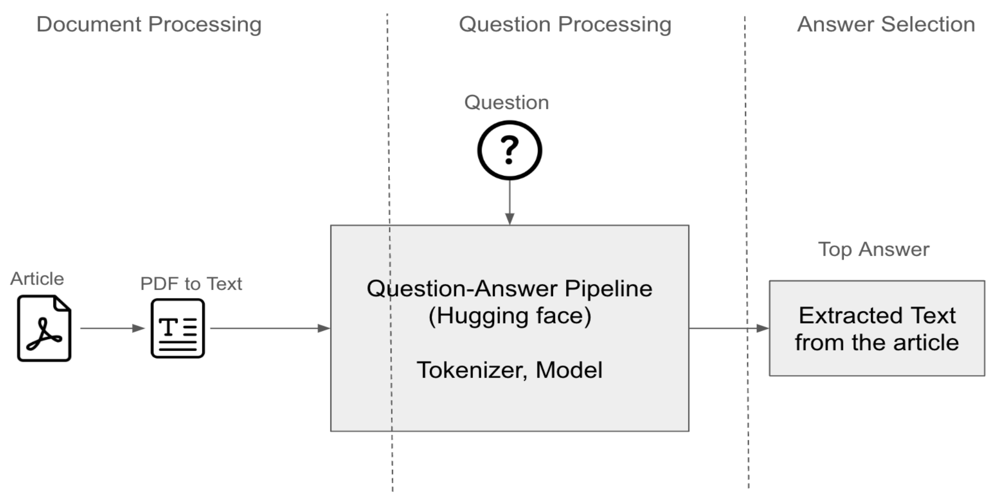
Extractive Question Answering

Components

- Document Processing
- Question Processing
- Answer Selection

Models Used

- distilbert-base-cased-distilled-squad (DistilBert)
- deepset/roberta-base-squad2(RoBERTa)



RQ3

Results on 38 articles: Extractive QA

DistilBert

- Fine-tuned on Squad dataset
- Version of BERT which is lighter(65.2M parameters)

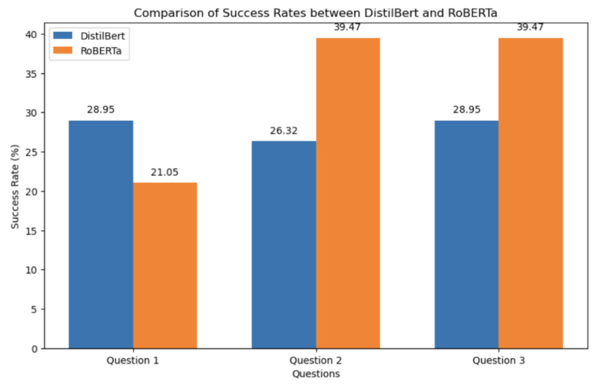
Roberta

- 124M parameters
- Fine-tuned on Squad 2.0

Tradeoff

- Roberta more accurate
- DistilBert is faster

	Correct Answers	
	DistilBert	RoBERTa
Question 1	11	8
Question 2	10	15
Question 3	10	15



Model	Average Time (s)	Max Time (s)	Min Time (s)
DistilBert	21.83	131.96	6.53
RoBERTa	42.43	262.09	12.41

RQ3

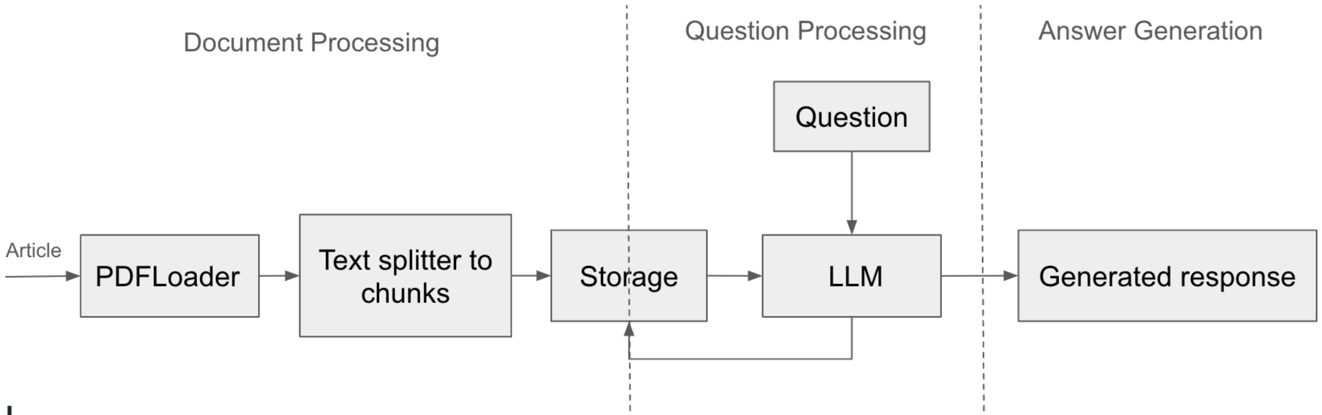
Abstractive Question Answering

Components

- Document Processing
- Question Processing
- Answer Generation

Models used:

- GPT series OpenAI
  - Text-davinci:003
  - GPT 3.5-Turbo
- Llama2- 7Billion parameter Meta AI



RQ3

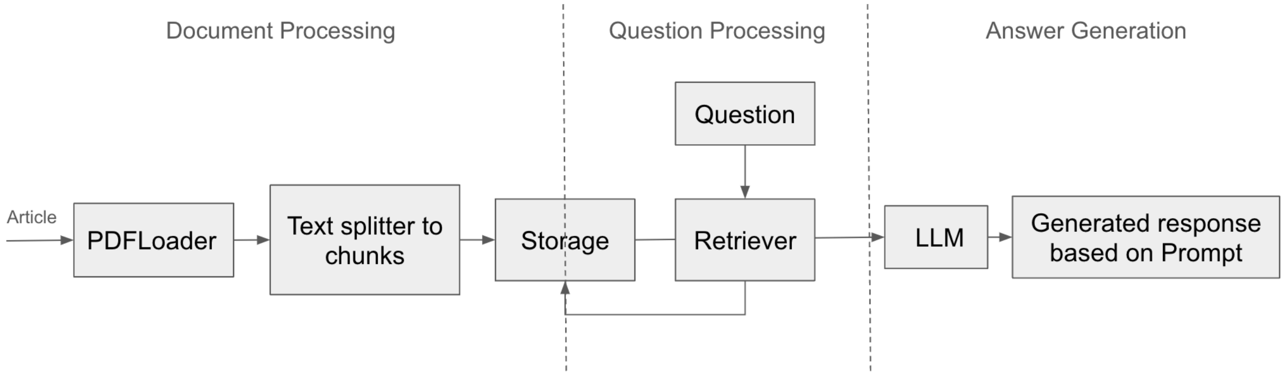
Abstractive Question Answering with Retrieval-Augmented Generation(RAG)

RAG workflow

- Relevant Information retrieval before answer generation
- Langchain modules

Evaluation

- Models evaluated on response for 'Question3'
- With & Without Prompt



Parameters:
Chunk size = 400 characters
Overlap = 50 characters
'Relevant chunks 'k' = 2
Temperature = 0.1

RQ3 Evaluating RAG Performance Across Models

Text-davinci:003

- With prompt, has more TP+TN(31 to 34)

GPT 3.5-Turbo

- Decrease TP+TN with prompts (32 to 22)

Llama2

- Prompts increases the correct response (from 22 to 30)

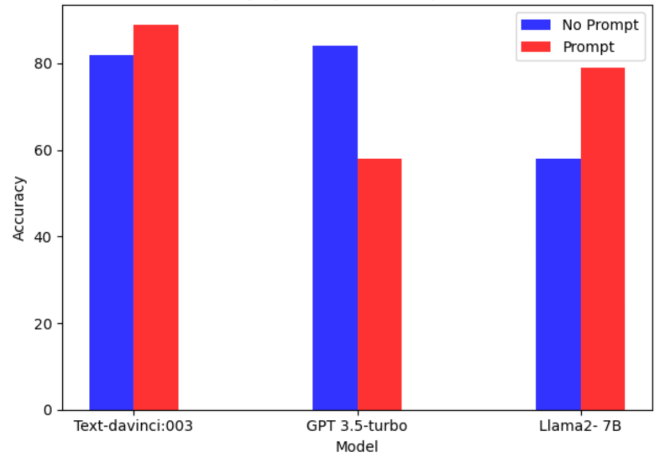
Without Prompt

Model	True Positives	True Negatives
Llama2	18	4
Text-davinci:003	16	15
GPT 3.5-Turbo	18	14

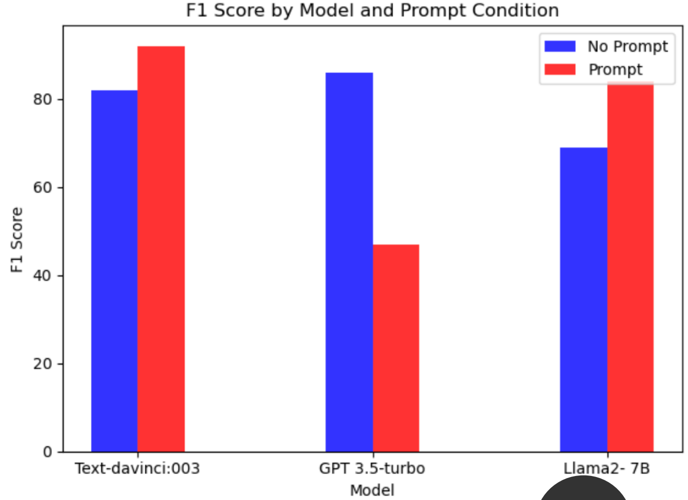
With Prompt

Model	True Positives	True Negatives
Llama2	21	9
Text-davinci:003	23	11
GPT 3.5-Turbo	7	15

Accuracy by Model and Prompt Condition



F1 Score by Model and Prompt Condition



RQ3

Comparative Performance of Language Models

Text-davinci:003

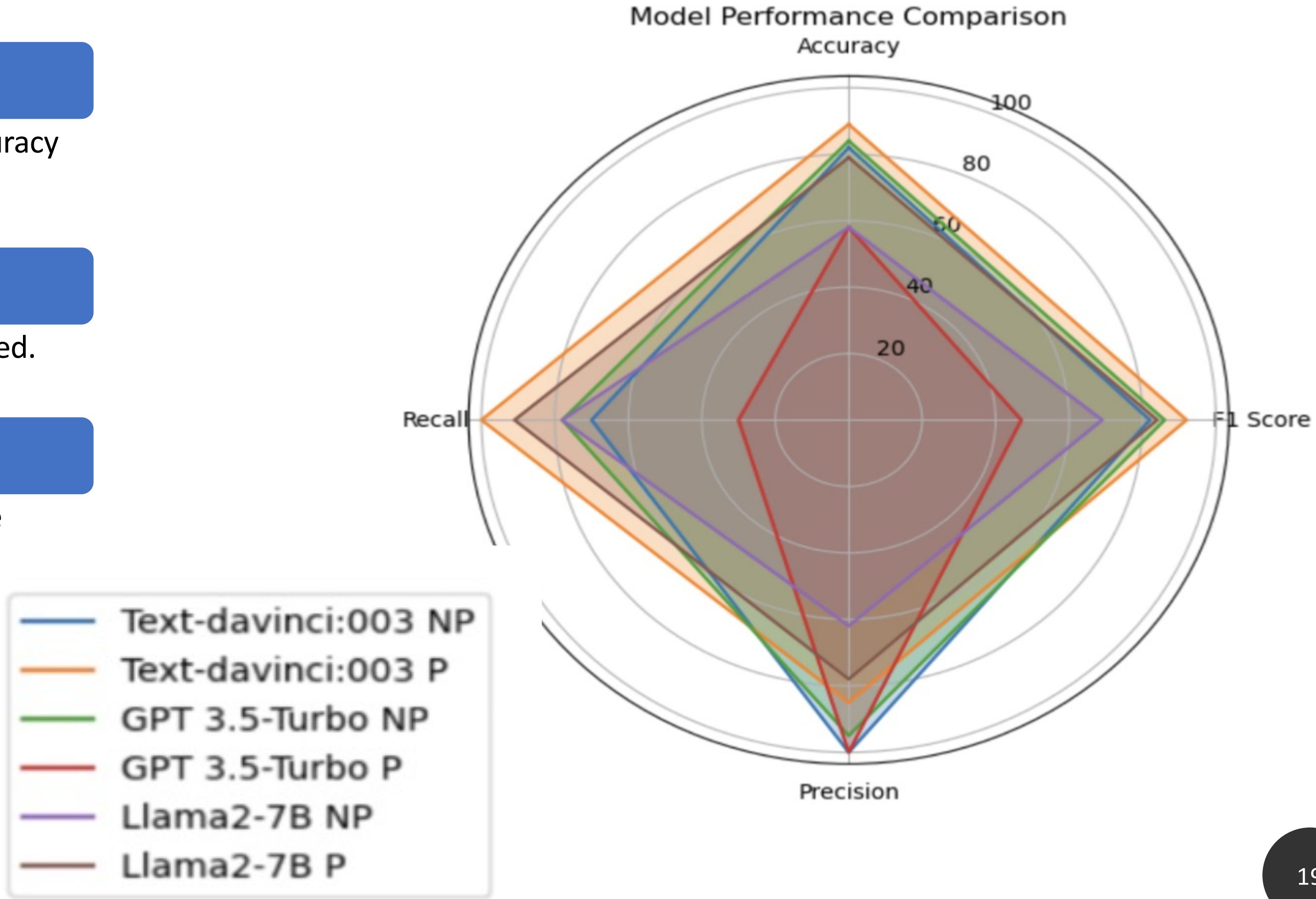
- with prompts shows superior accuracy and F1 score
- Precision is reduced with prompt.

GPT 3.5-Turbo

- With prompts performance reduced.
- Precision improved with prompt

Llama2-7B

- with prompts overall performance improved



RQ3

Evaluating Model Efficiency and Cost-Effectiveness

GPT 3.5-Turbo

- Fast response at lowest cost among commercial models

Text-davinci:003

- Varying speed across queries, but more expensive

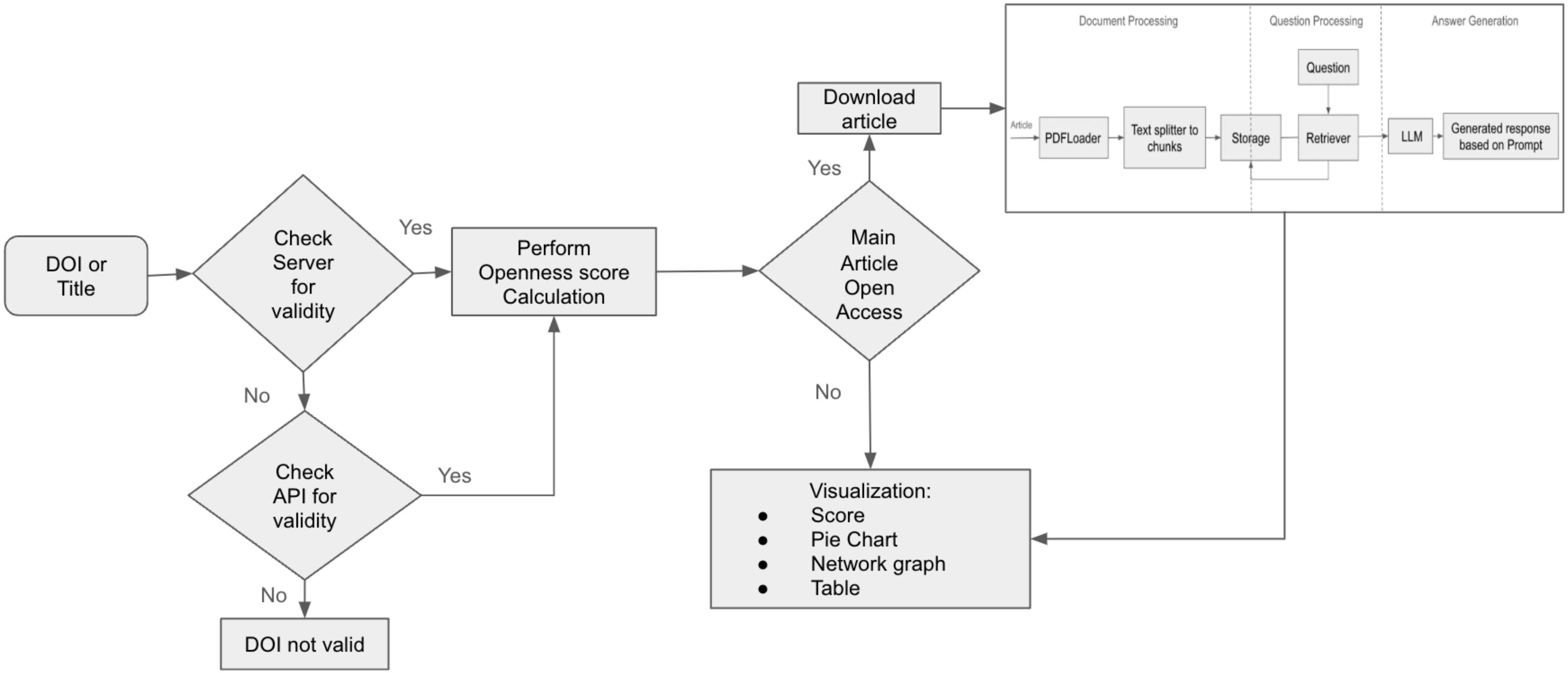
Llama2-7B

- Higher computing time, but is cost effective

Model	Max Time (s)	Min Time (s)	Avg Time (s)
GPT 3.5-Turbo	6.21	0.75	1.26
Text-davinci:003	11.61	0.52	0.96
Llama2-7B	77.53	37.03	53.54

Model	Cost (USD)
GPT 3.5-Turbo	0.32
Text-davinci:003	5.19
Llama2	0.00

Openness Score Evaluation Framework

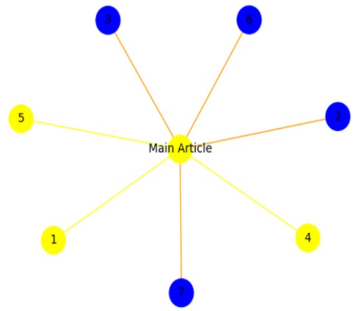


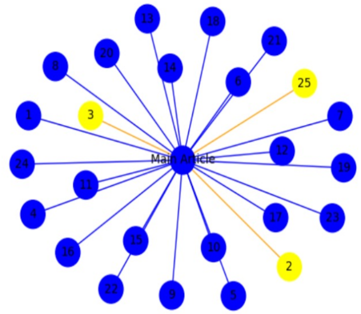
Visualizing Openness Score

Openness Score = (NDOIrOA ∗ 10/NDOIr) ∗ K

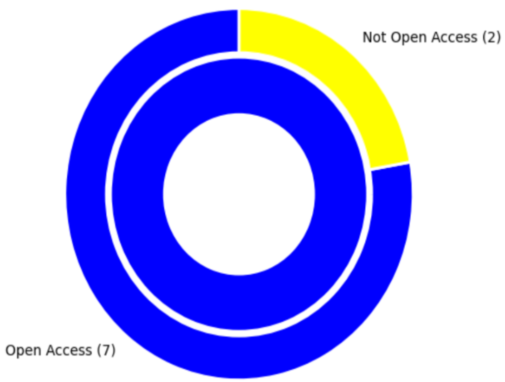
where K = 1 ,
NDOIrOA is the number of DOIs that are open-access,
NDOIr is the total number of references of the given DOI

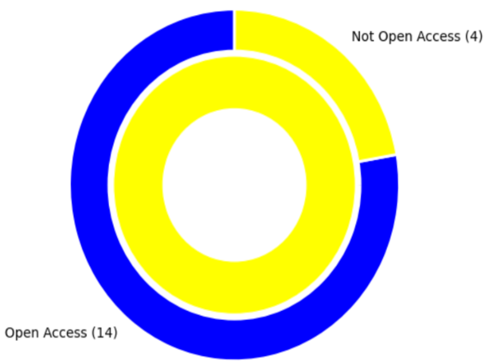
Network Graph



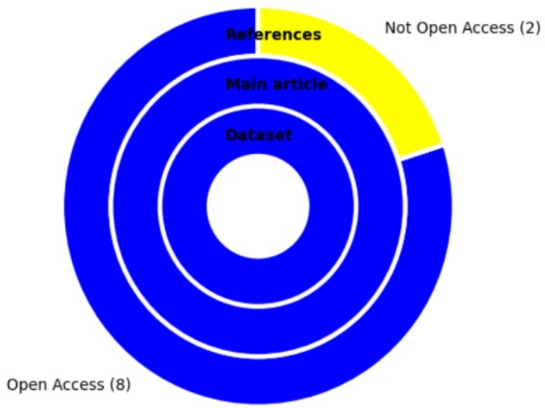


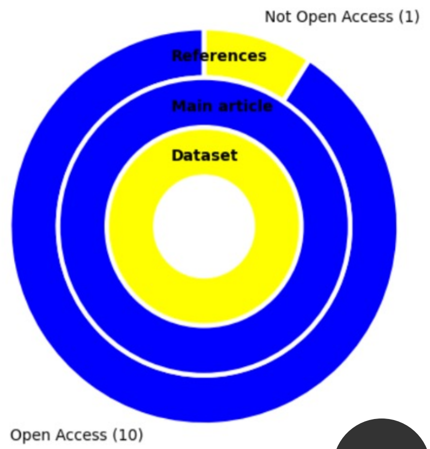
Pie Chart





Pie Chart





Limitations & Future Work

Additional APIs

Databases

Self-curated dataset of 38 articles

Human error on annotation

Scoring system

Prompts

23

Conclusion

4 Scholarly APIs

OpenAlex API vs snapshot

Leveraged LLM

Framework

24

Thank You !!!
Questions?

Towards A Framework For Openness Score Calculation in Scholarly Articles

Jennifer Swaminathan
University of Neuchatel, Switzerland
SUPERVISOR:
Prof. Dr. Philippe Cudré-Mauroux

Acknowledgements

- Prof. Dr Philippe Cudré-Mauroux
- Openness Score Team Members

26