

CIPS-SMP 2024 大模型时代的AI+

第十二届全国社交媒体处理大会

The Twelfth China National Conference on Social Media Processing

SMP2024 博士论坛

大模型工具学习与垂直领域应用 的一些思考

王元淳

中国人民大学

2024年10月

About Me



Yuanchun Wang (王元淳)

Ph.D. Student

📍 Beijing, China

🏛️ Renmin University of China

About Me

I am a Ph.D. student at the School of Information, Renmin University of China, under supervision of [Jing Zhang](#). At the same time, I'm an intern at THUKEG and ZHIPU.AI, under supervision of [Jifan Yu](#). Previous to that, I graduated from [Honors College](#) of Northwestern Polytechnical University, where I majored in Computer Science and Technology.

Currently, my research focuses on **Tool Intelligence of LLMs** and **AI for Education**. Besides, I love soccer🏈, photography📷, and snowboarding🏂. I also keep training for triathlon races🚴🏃. Feel free to contact me for our shared interests whether in research or life.

Selected Publications:

- **Tool Intelligence of LLMs**
 - SoAy: A Solution-based LLM API-using Methodology for Academic Information Seeking
 - R-Eval: A Unified Toolkit for Evaluating Domain Knowledge of Retrieval Augmented Large Language Models
- **AI for Education**
 - From MOOC to MAIC: Reshaping Online Teaching and Learning through LLM-driven Agents

LLM Specific Domain Application

Currently, the common requirement of LLM specific domain application is **domain QA**.

Challenge of the directly deployment of LLMs into a specific domain:

Hallucination in General-purpose LLMs:

- Lack of command of Domain Knowledge
- Difficult in Information updating

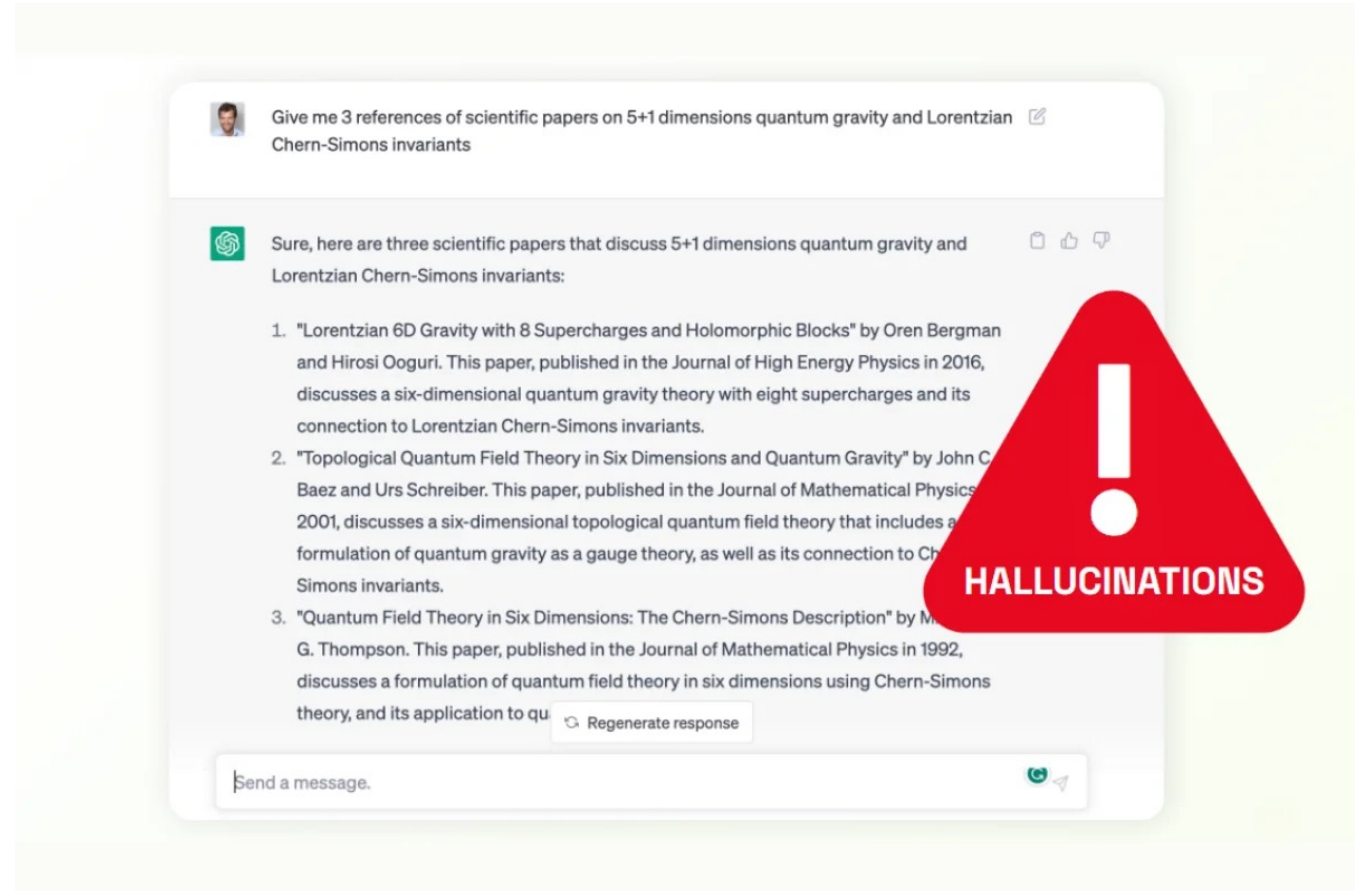


Fig.1 ChatGPT hallucinates on providing scientific reference.

LLM Specific Domain Application

Methods of apply LLM into specific domain:

- Training on collected domain dataset.
 - High cost of data collecting.
 - Still difficult in information updating
- Text2SQL
- Retrieval-augmented Generation

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

Question: How do you treat skin redness?

Complete Answer: It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: Why take simvastatin?

Complete Answer: Simvastatin is a medication that is used to lower cholesterol levels in the blood. Simvastatin is typically used to treat high cholesterol and to prevent heart disease, and to reduce the risk of heart attacks and strokes in those with the greatest risk of these happening (or re-occurring). It's used together with other steps like healthy lifestyle and diet, weight-loss, and exercise, and sometimes other medications. Simvastatin may also be used to decrease the amount of...

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: How can i split a 3mg or 4mg waefin pill so i can get a 2.5mg pill?

Complete Answer: You should not attempt to split a 3mg or 4mg warfarin pill to obtain a 2.5mg dose. If you are prescribed warfarin in a 2.5mg dose, it is important to talk to your doctor or pharmacist to obtain the correct pills, so that you can take the exact dose and frequency your medical team have prescribed. Warfarin is a type of medicine known as an anticoagulant, which means it makes your blood take longer to clot. The exact amount required to achieve the desired effect can vary.

→
Instruction
Prompt Tuning

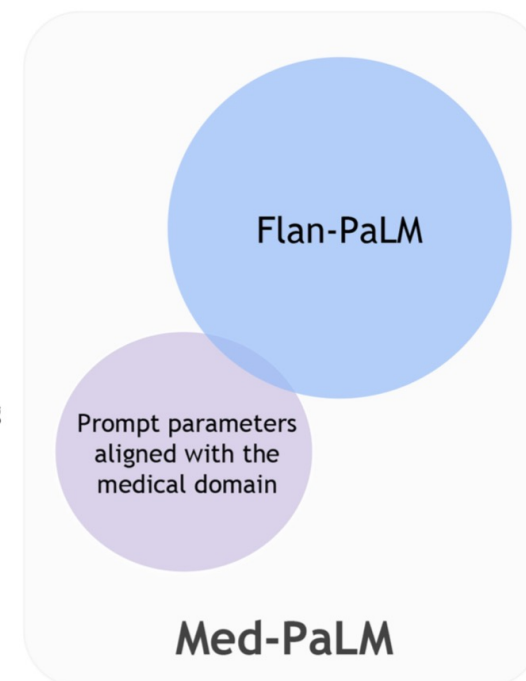


Fig.2 Flan-PaLM into Med-PaLM: Training on medical data.

LLM Specific Domain Application

Methods of apply LLM into specific domain:

- Training on collected domain dataset.
- Text2SQL
 - Limited on specific SQL Language.
 - Safety concerns about interaction directly with DB.
- Retrieval-augmented Generation

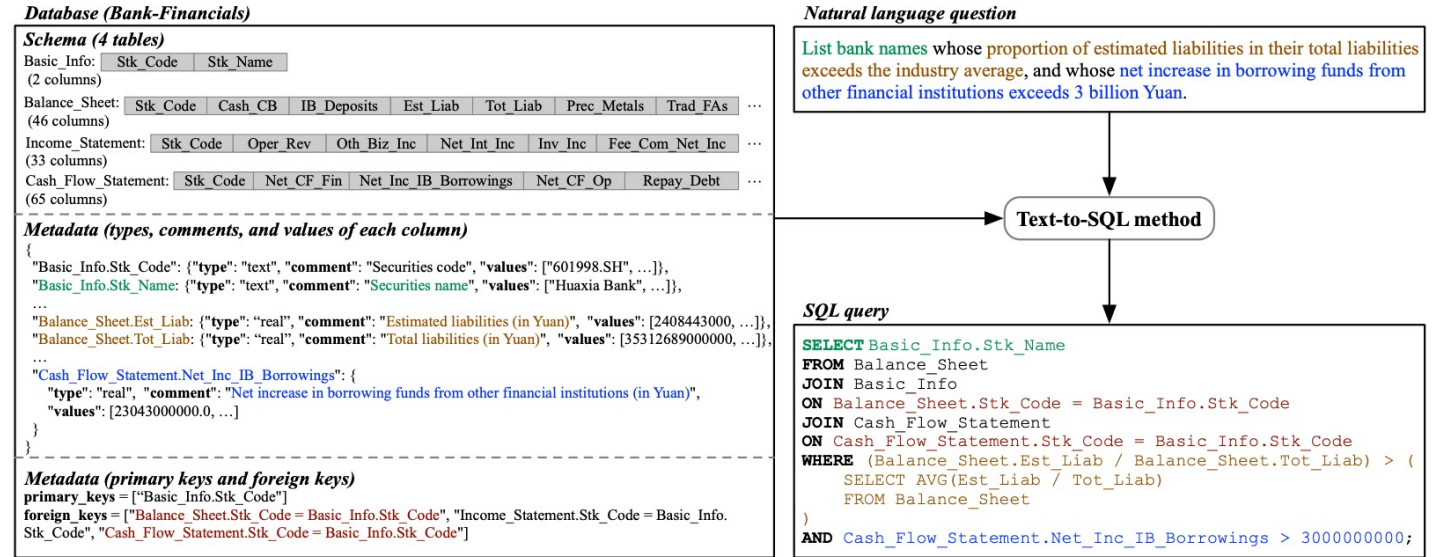


Fig.3 LLM interaction with domain KB through SQL query.

LLM Specific Domain Application

Methods of apply LLM into specific domain:

- Training on collected domain dataset.
- Text2SQL
- Retrieval-augmented Generation
 - Through Retriever
 - Through External Tools (Tool Learning)

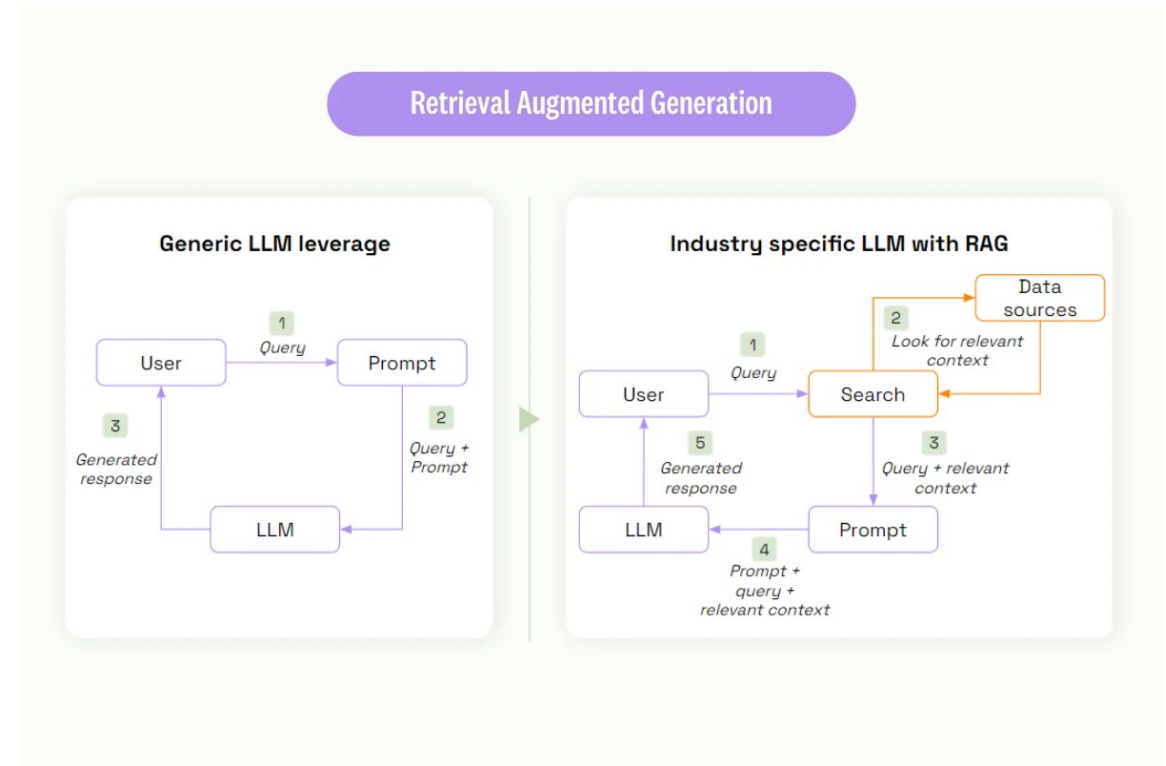


Fig.4 LLM specific domain application through RAG.

LLM Tool Learning

LLMs can use tools to interact with external environment, including **web news**, **calculator**, real-world **agent scene**, and **domain knowledge**...

- How to teach LLMs to use IR tools to get domain knowledge?
- How to evaluate the ability of the LLM domain tool using?

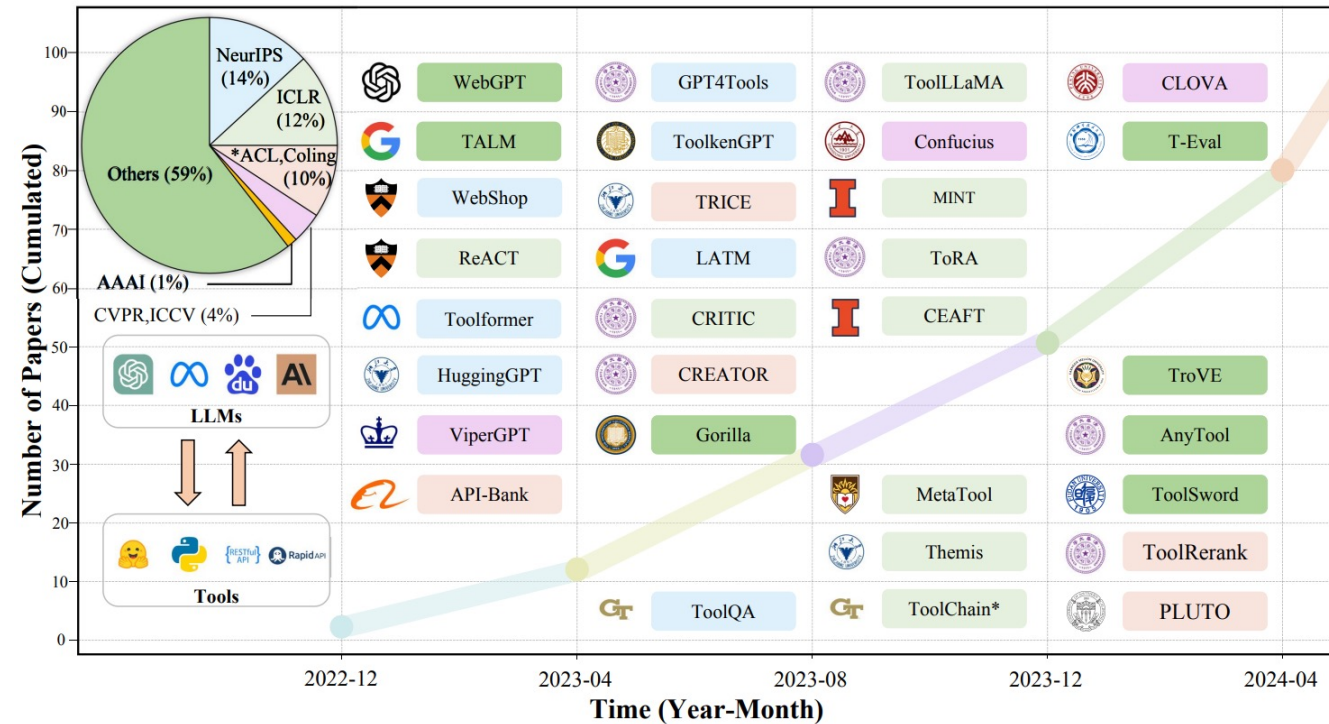


Fig.5 LLM specific domain application through Tool Learning.



×



SoAy: A Solution-based LLM API-using Methodology for Academic Information Seeking

**Yuanchun Wang^{†*}, Jifan Yu^{§*}, Zijun Yao[§], Jing Zhang[†], Yuyang Xie[§], Shangqing Tu[§]
Yiyang Fu[†], Youhe Feng[†], Jinkai Zhang[†], Jingyao Zhang[◇], Bowen Huang[◇], Yuanyao Li[◇]
Huihui Yuan[◇], Lei Hou[§], Juanzi Li[§] and Jie Tang[§]**

[†]Renmin University of China [§]Tsinghua University [◇]Zhipu AI

[Paper] <https://arxiv.org/pdf/2405.15165>

[Code] <https://github.com/RUCKBReasoning/SoAy>

[System] <https://soay.aminer.cn/>

[Model] https://huggingface.co/frederickwang99/soayllama_v2_7b

[Benchmark & Dataset] <https://huggingface.co/datasets/frederickwang99/SoAyBench>

Seeking Academic Metadata

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

Step 1: Typing **Keywords** in the searching box

The Google logo is displayed in its standard multi-colored font.A search bar containing the text "Yann Lecun". To the right of the text are a magnifying glass icon, a close button (X), and a Google Assistant icon.

Google 搜索

手气不错

Seeking Academic Metadata

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

Step 2: Find the probable item in the results list.

The image shows a Google search interface with the query 'Yann Lecun' entered in the search bar. The search results are displayed on the right side of the page. The first result is from Google Scholar, titled 'Yann LeCun', with a description: 'Yann LeCun. Chief AI Scientist at Facebook & Silver Professor at the Courant Institute, New York University. Verified email at cs.nyu.edu ...'. This result is highlighted with a red rectangular box. Below it are results from Wikipedia, lecun.com, and X (Twitter). The Wikipedia result is titled 'Yann LeCun' and describes him as a French-American computer scientist. The lecun.com result is titled 'Yann LeCun's Home Page' and lists his roles at Facebook and NYU. The X result is titled 'Yann LeCun (@ylecun) / X' and lists his titles as Professor at NYU, Chief AI Scientist at Meta, and ACM Turing Award Laureate. A small profile picture of Yann LeCun is visible next to the X result.

Google Scholar
https://scholar.google.com › citations

Yann LeCun
Yann LeCun. Chief AI Scientist at Facebook & Silver Professor at the Courant Institute, New York University. Verified email at cs.nyu.edu ...

Wikipedia
https://en.wikipedia.org › wiki › Yann_LeC...

Yann LeCun
Yann André LeCun is a French-American computer scientist working primarily in the fields of machine learning, computer vision, mobile robotics and ...

lecun.com
http://yann.lecun.com

Yann LeCun's Home Page
Yann LeCun, VP and Chief AI Scientist, Facebook Silver Professor of Computer Science, Data Science, Neural Science, and Electrical and Computer Engineering, ...
Yann's DjVu Page · MNIST handwritten digit · Fun Stuff · Publications

X
https://twitter.com › ylecun

Yann LeCun (@ylecun) / X
Professor at NYU. Chief AI Scientist at Meta. Researcher in AI, Machine Learning, Robotics, etc. ACM Turing Award Laureate.

Seeking Academic Metadata

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

Step 3: Seek the **target** information on the page

The screenshot shows a Google Scholar search for 'Yann Lecun'. The search results on the left include links to Google Scholar, Wikipedia, lecun.com, and X. The main profile page for Yann LeCun is displayed, showing his title as Chief AI Scientist at Facebook & Silver Professor at the Courant Institute, New York University. Below the profile is a table of his publications with columns for Title, Cited By, and Year. The 'Cited By' column for the publication 'Deep learning' is highlighted with a red box, showing a value of 79904. To the right of the profile is a 'Cited by' section with a bar chart showing citations from 2017 to 2024, and a table with columns for 'All' and 'Since 2019'. The table shows Citations (356553 All, 245740 Since 2019), h-index (148 All, 115 Since 2019), and i10-index (382 All, 305 Since 2019). The bar chart shows a steady increase in citations from 2017 to 2023, with a slight dip in 2024.

TITLE	CITED BY	YEAR
Deep learning Y LeCun, Y Bengio, G Hinton nature 521 (7553), 436-444	79904	2015
Gradient-based learning applied to document recognition Y LeCun, L Bottou, Y Bengio, P Haffner Proceedings of the IEEE 86 (11), 2278-2324	65097	1998
Backpropagation applied to handwritten zip code recognition Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, ... Neural computation 1 (4), 541-551	16702	1989
Convolutional networks for images, speech, and time series Y LeCun, Y Bengio The handbook of brain theory and neural networks 3361 (10), 1995	8129	1995
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks P Sermanet, D Eigen, X Zhang, M Mathieu, R Fergus, Y LeCun International Conference on Learning Representations (ICLR 2014)	7710	2014
The MNIST database of handwritten digits Y LeCun, C Cortes	7592	1998
Efficient backprop Y LeCun, L Bottou, GB Orr, KR Müller Neural networks: Tricks of the trade, 9-50	7145	2002
Character-level convolutional networks for text classification X Zhang, J Zhao, Y LeCun Advances in neural information processing systems 28	6704	2015
Handwritten digit recognition with a back-propagation network Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, ...	6384	1990

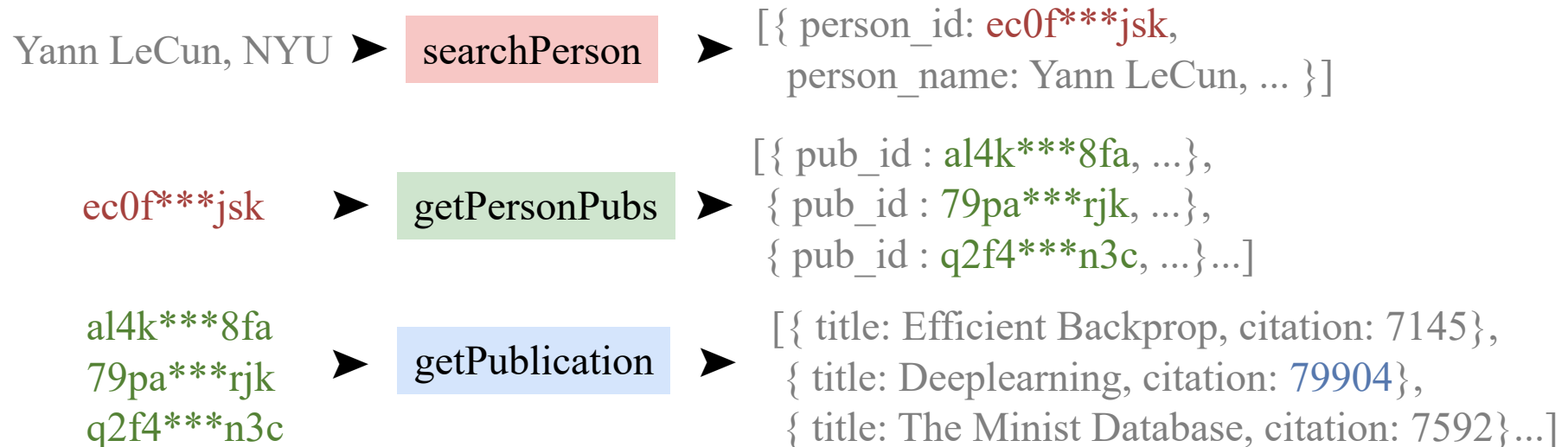
	All	Since 2019
Citations	356553	245740
h-index	148	115
i10-index	382	305

Year	Citations
2017	~12250
2018	~18750
2019	~25250
2020	~31750
2021	~38250
2022	~44750
2023	~51250
2024	~44750

Academic Information Systems API Calling

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

ID	API name	Type	Parameter(s)	Return
1	searchPerson	fuzzy	name, organization, interest	[person_id, name, num_citation, interest, num_pubs, organization]
2	searchPublication	fuzzy	publication_info	[pub_id, title, year]
3	getCoauthors	exact	person_id	[id, name, relation]
4	getPersonInterest	exact	person_id	list of interests
5	getPublication	exact	pub_id	abstract, author_list, num_citation
6	getPersonBasicInfo	exact	pub_id	person_id, name, gender, organization, position, bio, education_experience, email
7	getPersonPubs	exact	person_id	[authors_name_list, pub_id, title, num_citation, year]






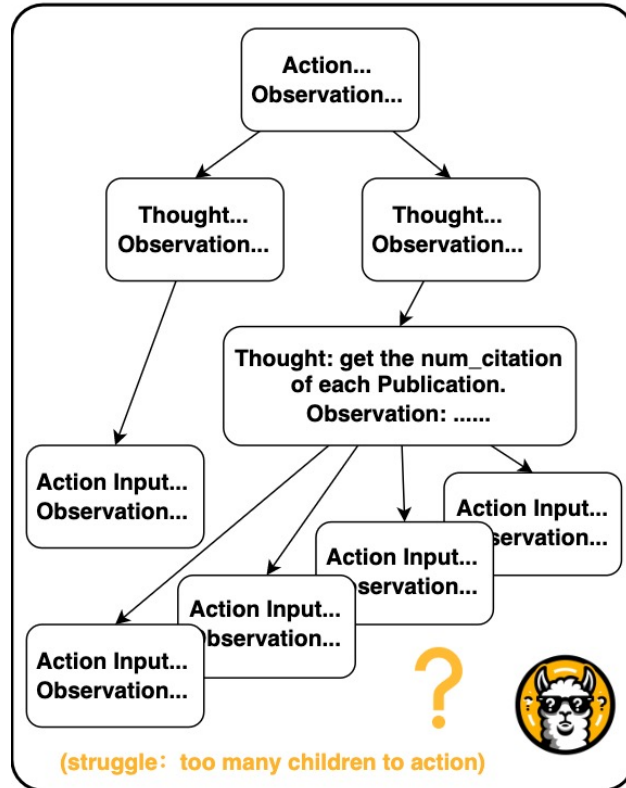
LLM API-Using

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

Retrieval & Execution:
Failed to handle
API Coupling

DFSDT Reasoning:
Could not meet
the Efficiency needs

"Functions to call":
[searchPerson, getPersonPubs, getPublication]
"parameters":
[..., person_id: not available, ...]  
Could not get person_id  failed



[solution]:
searchPerson -> getPersonPubs -> getPublication

[program]:
person_list = searchPerson(name = 'Yann LeCun')
if len(person_list) > 1:
person_list = searchPerson(name = 'Yann LeCun',
organization = 'New York University')
target_person_info = person_list[0]
target_person_id = target_person_info['person_id']
target_person_pubs = getPersonPubs(person_id =
target_person_id)
target_publication_dict = sorted(target_person_pubs,
key=lambda x: x['num_citation'], reverse=True)[0]
target_publication_id =
target_publication_dict['pub_id']
target_publication_info = getPublication(pub_id =
target_publication_id)
target_citation_count =
target_publication_info['num_citation']
final_result = target_citation_count 

[answer]:  correct
Yann LeCun's most cited publication has been cited
74731 times.

SoAy:

Pre-defined Solution
&
Solution-based Program
Generation

Fig.6 Different API-using structures facing the same academic question.

SoAy: SoAPIs Applying Framework

Query: How many times has New York University's Yann LeCun's most cited publication been cited?

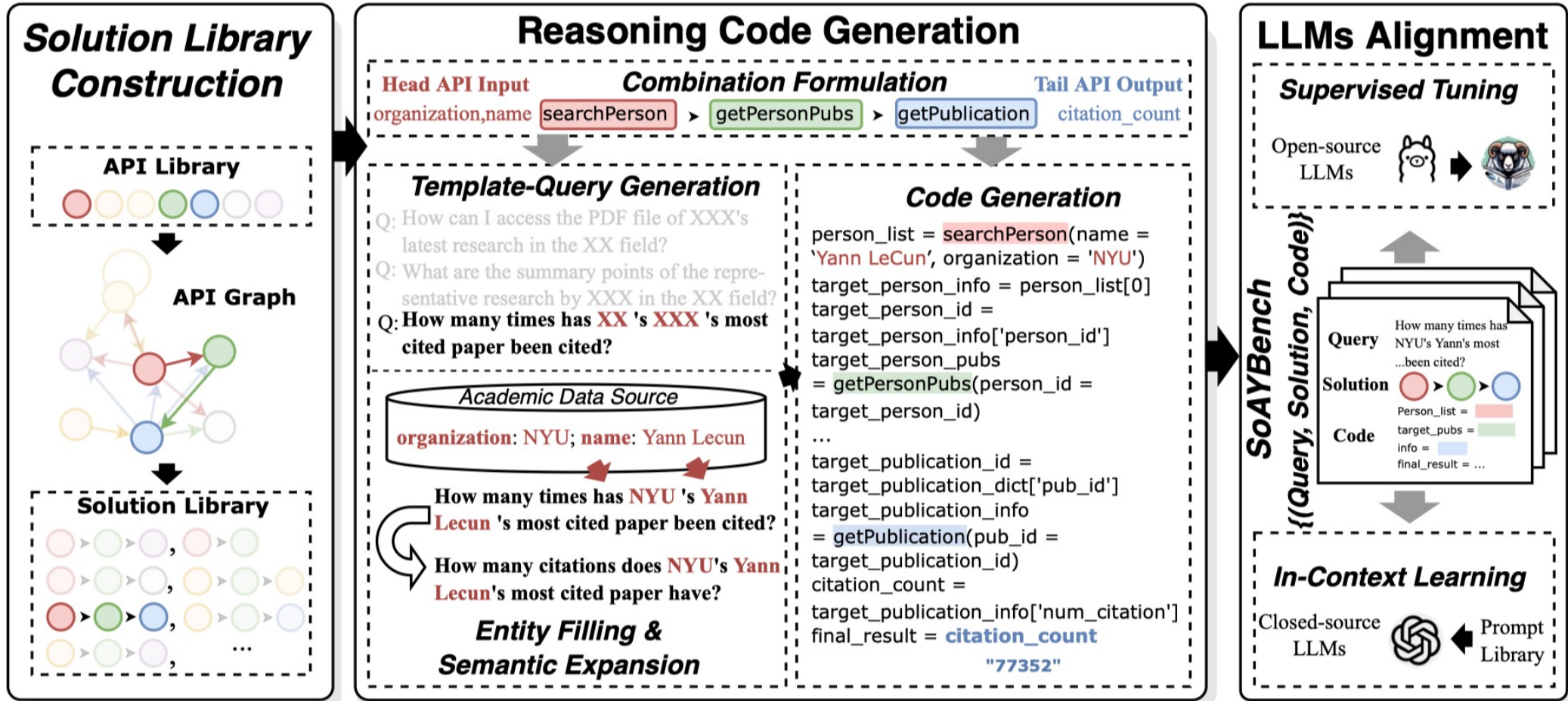


Fig.7 SoAy Framework.

SoAyBench

To assess API utilization capabilities, it is essential to publish the foundational APIs of AMiner for LLMs to invoke and provide a test set composed of academic {Query, Code, Answer} triplets for evaluation.

However, given the dynamic nature of academic data, with scholar and publication information **rapidly changing**, maintaining a test set with static answers proves challenging.

To address this challenge, **we clone AMiner's SoAPIs at a specific point in time** to create a static service, from which we generate a corresponding static test set.

SoAyBench now are open-sourced at : 🤗 Hugging Face:
<https://huggingface.co/datasets/frederickwang99/SoAyBench>

Question statistics in SoAyBench.

Question Type	One-hop	Two-hop	Three-hop	Total
Scholar	540	1,800	540	2,980
Publication	180	180	720	1,080
Total	720	1,980	1,230	3,960

SoAyEval

We outline five types of evaluation metrics.

- * EM: Both the retrieved solution and answer **Exactly Match** the ground truth.
- * DS: The answer is correct, but a **Different Solution** is retrieved compared to the ground truth.
- * WS: The answer is wrong due to a **Wrong Solution**.
- * WP: The solution is correct but the answer is wrong, due to a **Wrong Program** generated for the solution, which can be executed but yields the wrong answer.
- * EE: **Execution Error**, which may be caused by the generation of a nonexecutable program or network errors during the APIs request.

$$\text{ACC} = \text{EM} + \text{DS}$$

$$\text{Score} = \frac{w_1 \cdot \text{ACC}_1 + w_2 \cdot \text{ACC}_2 + w_3 \cdot \text{ACC}_3}{w_1 + w_2 + w_3}$$

Results on SoAyBench - Part I

Results on SoAyBench. DS, WS, WP and EE are different types of error, ACC denotes a accurate answer, EM means exact match, not only the answer but also the solution. Score is a weighted sum of the ACC score on different question types.

Method	Version	Question Type	Error Rate↓				EM(%)	ACC(%)	Score
			DS(%)	WS(%)	WP(%)	EE(%)			
ToolLLaMA	7B	one-hop	12.50±8.00	24.31±13.26	1.39±0.00	54.17±16.01	7.64±5.20	20.14	16.72
		two-hop	10.10±4.10	47.22±12.28	0.76±2.27	38.13±9.62	3.79±2.92	13.89	
		three-hop	11.51±6.53	38.10±14.27	1.19±3.57	43.25±13.07	5.95±4.59	17.46	
GPT-DFSdT	3.5	one-hop	55.56±21.06	15.28±7.80	4.86±0.00	21.53±10.67	2.78±0.00	58.33	43.22
		two-hop	29.55±11.47	34.34±9.23	4.29±3.64	25.76±8.65	6.06±4.11	35.61	
		three-hop	38.10±15.09	28.57±11.35	3.17±2.50	25.00±8.87	5.16±6.19	43.25	
	3.5-16k	one-hop	25.69±10.91	9.72±5.00	2.78±0.00	22.92±9.47	38.89±15.60	64.58	43.67
		two-hop	16.92±7.76	15.91±6.05	3.28±1.31	46.97±7.13	16.92±4.99	33.84	
		three-hop	18.65±7.37	15.48±5.63	2.78±0.00	38.49±10.43	24.60±8.53	43.25	
	4	one-hop	27.78±9.60	2.08±0.00	4.17±5.00	28.47±6.82	37.50±10.91	65.28	58.16
		two-hop	26.26±8.89	9.60±4.88	17.93±5.40	15.15±5.39	31.06±9.12	57.32	
		three-hop	22.22±8.65	7.54±4.46	17.06±6.96	19.05±6.45	34.13±9.87	56.35	
GPT-SoAY	3.5	one-hop	27.78±8.70	15.97±7.73	3.47±0.00	13.19±7.80	39.58±9.12	67.36	67.30
		two-hop	33.84±4.94	9.60±4.75	6.06±2.81	13.13±7.12	37.37±5.06	71.21	
		three-hop	22.22±6.43	12.70±5.91	9.52±4.42	13.10±6.72	42.46±6.00	64.68	
	3.5-16k	one-hop	28.47±11.67	15.28±6.12	1.39±0.00	17.36±7.78	37.50±9.07	65.97	66.76
		two-hop	35.86±6.01	7.32±3.41	5.30±2.18	15.91±7.16	35.61±4.65	71.46	
		three-hop	23.02±7.16	10.32±4.99	8.33±3.42	17.46±7.37	40.87±6.26	63.89	
	4	one-hop	0.00±0.00	0.00±0.00	1.39±0.00	2.78±0.00	95.83±5.70	95.83	86.57
		two-hop	15.91±4.71	1.26±0.00	9.34±1.07	2.02±1.69	71.46±3.74	87.37	
		three-hop	6.75±0.00	0.40±0.00	14.68±1.68	1.98±0.00	76.19±3.25	82.94	

Results on SoAyBench - Part II

Results on SoAyBench. DS, WS, WP and EE are different types of error, ACC denotes a accurate answer, EM means exact match, not only the answer but also the solution. Score is a weighted sum of the ACC score on different question types.

GPT-SoAY	3.5	one-hop	27.78±8.70	15.97±7.73	3.47±0.00	13.19±7.80	39.58±9.12	67.36	67.30
		two-hop	33.84±4.94	9.60±4.75	6.06±2.81	13.13±7.12	37.37±5.06	71.21	
		three-hop	22.22±6.43	12.70±5.91	9.52±4.42	13.10±6.72	42.46±6.00	64.68	
	3.5-16k	one-hop	28.47±11.67	15.28±6.12	1.39±0.00	17.36±7.78	37.50±9.07	65.97	66.76
		two-hop	35.86±6.01	7.32±3.41	5.30±2.18	15.91±7.16	35.61±4.65	71.46	
		three-hop	23.02±7.16	10.32±4.99	8.33±3.42	17.46±7.37	40.87±6.26	63.89	
	4	one-hop	0.00±0.00	0.00±0.00	1.39±0.00	2.78±0.00	95.83±5.70	95.83	86.57
		two-hop	15.91±4.71	1.26±0.00	9.34±1.07	2.02±1.69	71.46±3.74	87.37	
		three-hop	6.75±0.00	0.40±0.00	14.68±1.68	1.98±0.00	76.19±3.25	82.94	
SoAYLLaMA	Chat-7B	one-hop	0.00±0.00	0.00±0.00	0.00±0.00	0.69±0.00	99.31±2.94	99.31	85.76
		two-hop	0.00±0.00	0.00±0.00	20.20±3.84	2.53±1.97	77.27±2.70	77.27	
		three-hop	0.00±0.00	0.00±0.00	9.92±3.56	3.17±2.50	86.90±2.72	86.90	
	Code-7B	one-hop	0.69±0.00	0.00±0.00	0.69±0.00	5.56±4.37	93.06±7.50	93.75	88.95
		two-hop	0.25±0.00	3.28±0.00	7.07±2.75	4.80±3.69	84.60±5.18	84.85	
		three-hop	0.40±0.00	0.00±0.00	4.76±2.14	5.16±4.57	89.68±6.54	90.08	
	Code-13B	one-hop	0.00±0.00	0.00±0.00	1.39±0.00	0.00±0.00	98.61±4.03	98.61	92.74
		two-hop	0.00±0.00	2.27±0.00	14.14±2.14	0.51±0.00	83.08±3.32	83.08	
		three-hop	0.00±0.00	0.00±0.00	2.38±2.86	0.40±0.00	97.22±4.28	97.22	

Efficiency & Online Evaluation

To evaluate how efficient are SoAy, we compare the average response time of different methods (second).

Method	7B	13B	3.5	3.5-16k	4
ToolLLaMA	45.10	/	/	/	/
GPT-DFSDT	/	/	39.12	53.73	70.92
SoAYGPT	/	/	6.15	6.40	26.05
SoAYLLaMA-Code	1.12	1.35	/	/	/

To test whether SoAy could meet the need of real-world user requirement, we implement SoAy as an online application, gather 56 real user demands from the logs, and invite 10 annotators to conduct human evaluation.

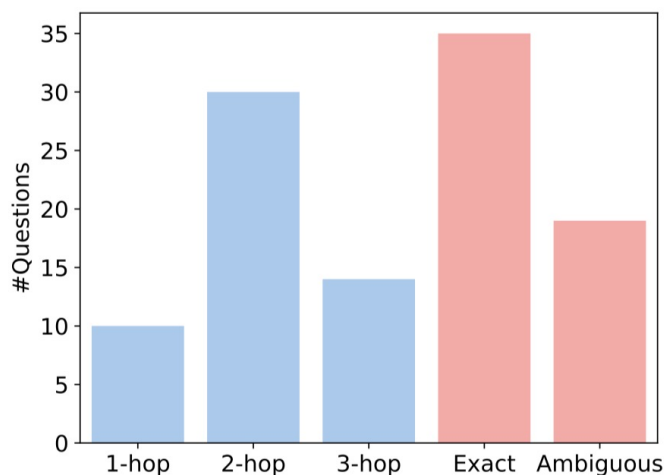


Fig.8 Online Gathered Question statistics.

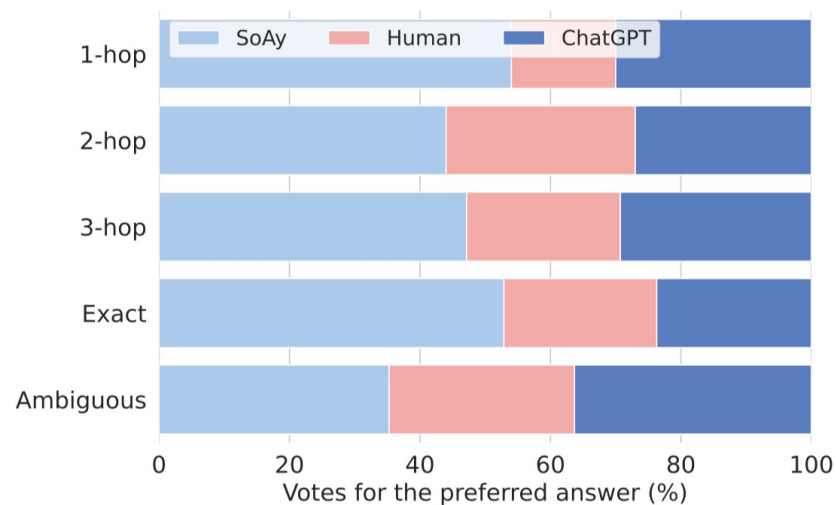


Fig.9 Results of Online Human Evaluation

Results on SoAyBench - Weak to Strong Supervision

A Small Model trained on Data generated by GPT-4 can **outperform** GPT-4, and also more **efficient**.

SoAYGPT	3.5	one-hop	27.78±8.70	15.97±7.73	3.47±0.00	13.19±7.80	39.58±9.12	67.36	67.30	
		two-hop	33.84±4.94	9.60±4.75	6.06±2.81	13.13±7.12	37.37±5.06	71.21		
		three-hop	22.22±6.43	12.70±5.91	9.52±4.42	13.10±6.72	42.46±6.00	64.68		
	3.5-16k	one-hop	28.47±11.67	15.28±6.12	1.39±0.00	17.36±7.78	37.50±9.07	65.97		66.76
		two-hop	35.86±6.01	7.32±3.41	5.30±2.18	15.91±7.16	35.61±4.65	71.46		
		three-hop	23.02±7.16	10.32±4.99	8.33±3.42	17.46±7.37	40.87±6.26	63.89		
	4	one-hop	0.00±0.00	0.00±0.00	1.39±0.00	2.78±0.00	95.83±5.70	95.83		86.57
		two-hop	15.91±4.71	1.26±0.00	9.34±1.07	2.02±1.69	71.46±3.74	87.37		
		three-hop	6.75±0.00	0.40±0.00	14.68±1.68	1.98±0.00	76.19±3.25	82.94		
SoAYLLaMA	Chat-7B	one-hop	0.00±0.00	0.00±0.00	0.00±0.00	0.69±0.00	99.31±2.94	99.31	85.76	
		two-hop	0.00±0.00	0.00±0.00	20.20±3.84	2.53±1.97	77.27±2.70	77.27		
		three-hop	0.00±0.00	0.00±0.00	9.92±3.56	3.17±2.50	86.90±2.72	86.90		
	Code-7B	one-hop	0.69±0.00	0.00±0.00	0.69±0.00	5.56±4.37	93.06±7.50	93.75	88.95	
		two-hop	0.25±0.00	3.28±0.00	7.07±2.75	4.80±3.69	84.60±5.18	84.85		
		three-hop	0.40±0.00	0.00±0.00	4.76±2.14	5.16±4.57	89.68±6.54	90.08		
	Code-13B	one-hop	0.00±0.00	0.00±0.00	1.39±0.00	0.00±0.00	98.61±4.03	98.61	92.74	
		two-hop	0.00±0.00	2.27±0.00	14.14±2.14	0.51±0.00	83.08±3.32	83.08		
		three-hop	0.00±0.00	0.00±0.00	2.38±2.86	0.40±0.00	97.22±4.28	97.22		

Method	7B	13B	3.5	3.5-16k	4
ToolLLaMA	45.10	/	/	/	/
GPT-DFSdT	/	/	39.12	53.73	70.92
SoAYGPT	/	/	6.15	6.40	26.05
SoAYLLaMA-Code	1.12	1.35	/	/	/

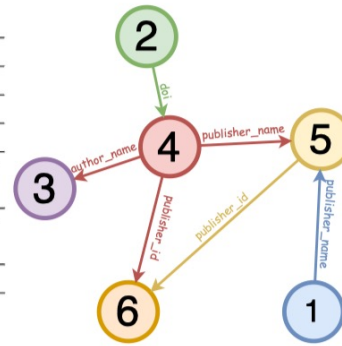
Deployment on other Academic Platforms

AMiner APIs are **NOT** the only that face the **coupling** challenges.

We also deployment SoAy on two other open-sourced scenarios: OpenLibrary and CrossRef

ID	API name	Type	Parameter(s)	Return
1	searchPublisherBySubject	fuzzy	subject	[publisher_name, doi_count]
2	searchWorksByTitle	fuzzy	work_title	[type, author, doi, publisher]
3	searchWorksByAuthor	fuzzy	author_name	[works_title, works_doi]
4	getWorksByDoi	exact	doi	[author_name, work_title, publisher_name, type, reference_count]
5	getPublisherBasicInfo	exact	publisher_name	[publisher_id, current_dois, backfile_dois, total_dois, doi_prefix]
6	getPublisherWorks	exact	publisher_id	[works_title, doi, works_author]

(a) CrossrefAPI Library



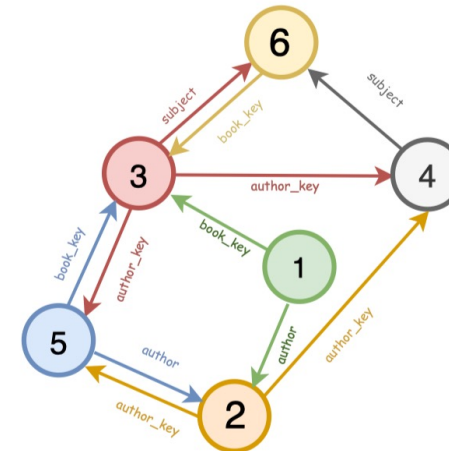
(b) CrossrefAPI Graph

Solution	Parameter(s)	Return	Question Template
searchPublisherBySubject	subject	publisher_name	Please list some publishers in the XXX field.
searchPublisherBySubject → get-PublisherBasicInfo	subject	publisher_id	Please give me some publishers' id of crossref about the field of XXX.
searchPublisherBySubject → get-PublisherBasicInfo → getPublisherWorks	subject	doi	Can you list some articles' DOI numbers in the field of XXX?
searchWorksByTitle	work_title	type	I want to know the type of work XXX.

(c) Solution Library (partly shown)

ID	API name	Type	Parameter(s)	Return
1	searchBook	fuzzy	book_info	[book_key, title, author_name, year]
2	searchAuthor	fuzzy	author_info	[author_key, name, list of alternate_names]
3	getBook	exact	book_key	description, list of author, title, first_publish, list of subjects
4	getAuthorBasicInfo	exact	author_key	name, list of alternate_names, birth_date, work_count, top_work, top_subjects
5	getAuthorWorks	exact	author_key, amount	[book_key, title, subjects]
6	searchSubject	fuzzy	subject	[book_key, title]

(a) SoAPI Library



(b) SoAPI Graph

Solution	Parameter(s)	Return	Question Template
searchSubject	subject	list of books	Please list some books on XXX topic.
searchAuthor→getAuthorWorks	author_info	list of books	Which works were written by XXX?
searchBook→getBook	book_info	book_description	Introduce some information about XXX.
searchBook→getBook→getAuthorWorks	book_info	list of books	What other books has the author of XXX written?

(c) Solution Library (partly shown)

Challenges of the SoAyBench & SoAyEval

There's still some challenges on the evaluation part of specific-domain tool using.

- The benchmark or evaluation set is limited on the Academic domain.
- The complexity of testing on the combination of the **LLMs**, Tool-using **Workflows** and the **domains**.

Method	7B	13B	3.5	3.5-16k	4
ToolLLaMA	45.10	/	/	/	/
GPT-DFSDT	/	/	39.12	53.73	70.92
SoAyGPT	/	/	6.15	6.40	26.05
SoAYLLaMA-Code	1.12	1.35	/	/	/

R-Eval: A Unified Toolkit for Evaluating Domain Knowledge of Retrieval Augmented Large Language Models

Shangqing Tu*

DCST, Tsinghua University
Beijing 100084, China
tsq22@mails.tsinghua.edu.cn

Yuyang Xie

DCST, Tsinghua University
Beijing 100084, China
xieyy21@mails.tsinghua.edu.cn

Jing Zhang

SoI, Renmin University of China
Beijing 100084, China
zhang-jing@ruc.edu.cn

Yuanchun Wang*

SoI, Renmin University of China
Beijing 100084, China
wangyuanchun@ruc.edu.cn

Yaran Shi

SIOE, Beihang University
Beijing 100084, China
syr2021@buaa.edu.cn

Lei Hou

BNRist, DCST, Tsinghua University
Beijing 100084, China
houlei@tsinghua.edu.cn

Jifan Yu

DCST, Tsinghua University
Beijing 100084, China
yujf21@mails.tsinghua.edu.cn

Xiaozhi Wang

DCST, Tsinghua University
Beijing 100084, China
wangxz20@mails.tsinghua.edu.cn

Juanzi Li

BNRist, DCST, Tsinghua University
Beijing 100084, China
lijuanzi@tsinghua.edu.cn

[Paper] Accepted by KDD'24 (ADS track), under camera ready preparation.

[Code & Toolkit] <https://github.com/THU-KEG/R-Eval>

Background - Component Selection

Given a Specific Domain, which LLM and which RAG Workflow to choose?

Shortcomings of existing evaluations:

- Insufficient exploration of combinations between LLMs and RAG workflows.
- Lack comprehensive mining of the domain knowledge.

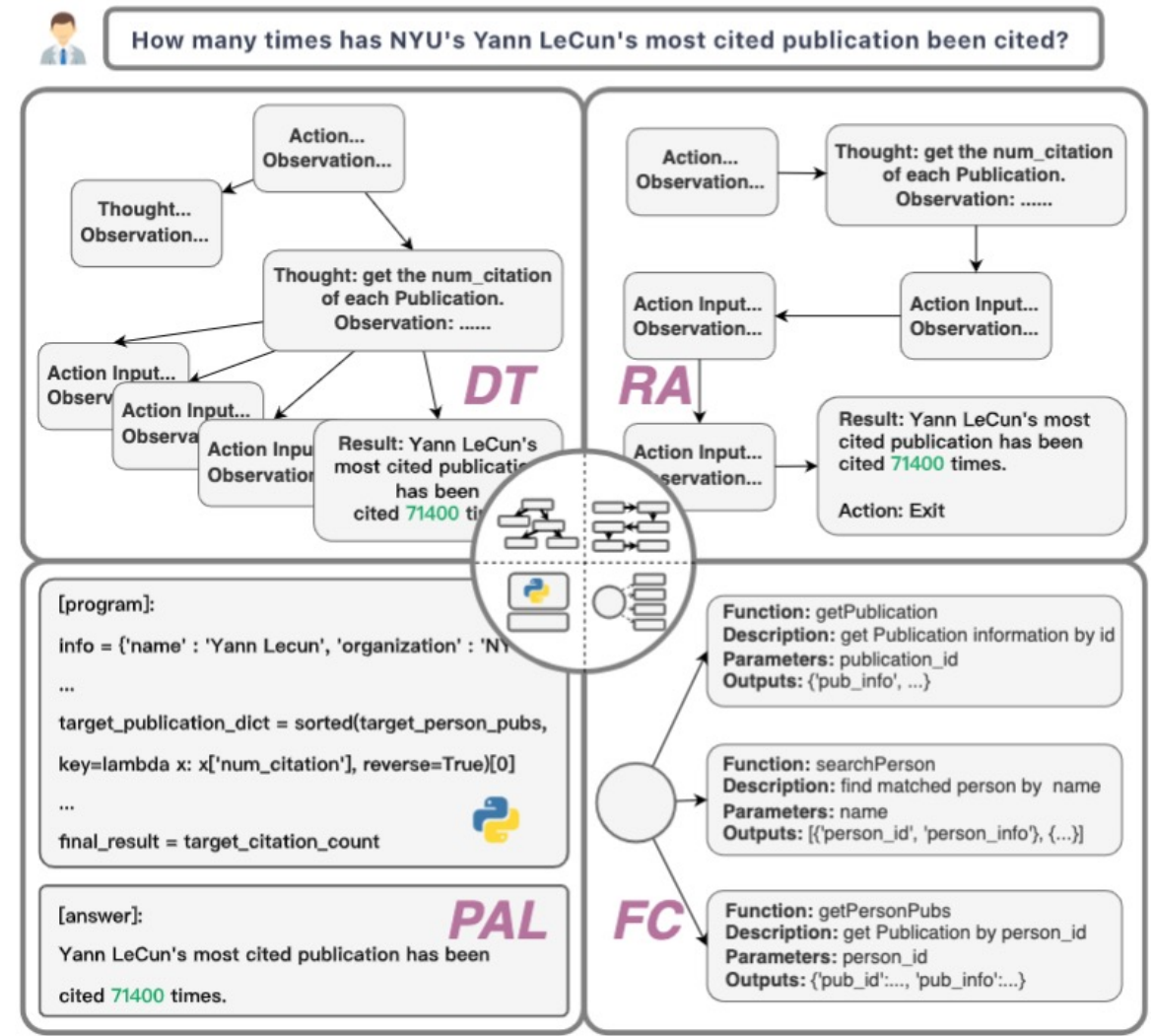


Fig.10 Four Popular RAG Workflows.

Evaluation Framework

Inspired by [KoLA](https://iclr.cc/virtual/2024/poster/19238) and [SoAy](https://arxiv.org/pdf/2405.15165), We propose [R-Eval](#), a Python toolkit designed to streamline the evaluation of different RAG workflows in conjunction with LLMs on a specific domain's task.

- A easy-to-use evaluation of the combination between RAG Workflows and LLMs
- Customized testing data in specific domains through template-based question generation

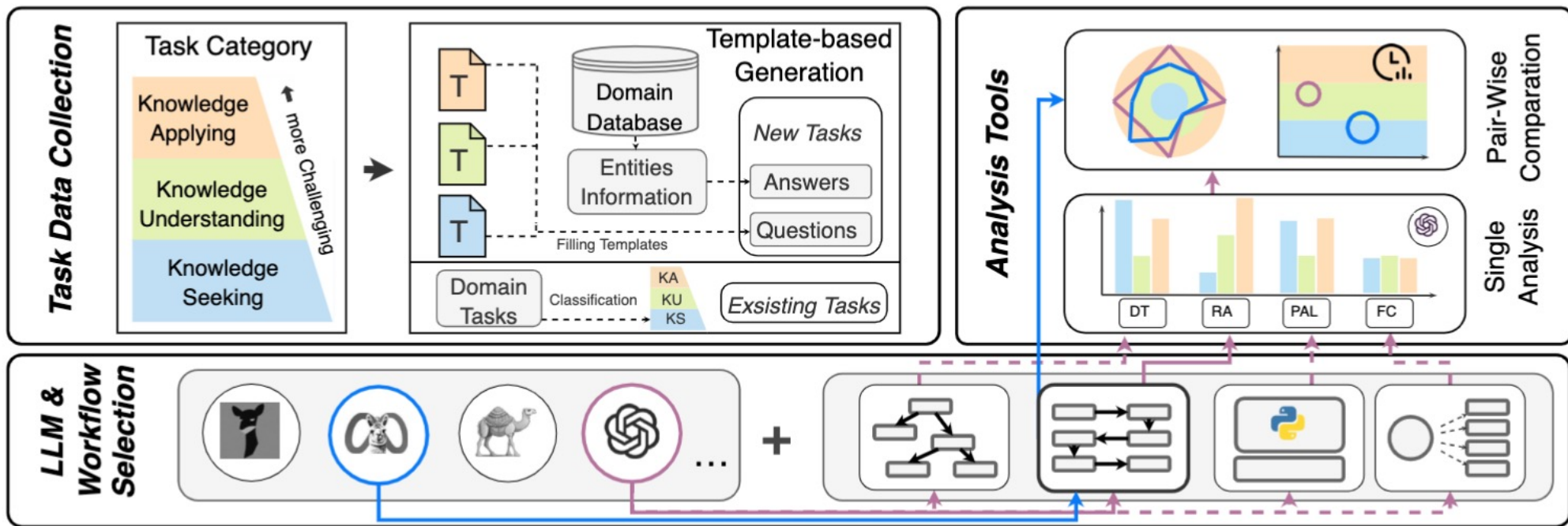


Fig.11 Framework of R-Eval.

Evaluation Result - Performance Ranking

Workflow	LLM	aminer KS		aminer KU		aminer KA		Overall Average (Level 1, 2, 3)					
		1-3	Rank	2-4	Rank	3-5	Rank	wiki	Rank	aminer	Rank	all	Rank
ReAct	gpt-4-1106	89.7	1st	46.7	3rd	57.7	1st	38.8	1st	64.7	1st	45.3	1st
PAL	gpt-3.5-turbo	80.1	3rd	50.7	2nd	54.9	2nd	19.9	6th	61.9	2nd	30.4	2nd
PAL	gpt-4-1106	59.3	4th	56.8	1st	52.7	3rd	20.3	5th	56.2	3rd	29.2	3rd
ReAct	llama2-7b-chat	45.2	5th	36.5	6th	21.5	6th	23.8	3rd	34.4	5th	26.4	4th
PAL	llama2-13b	25.3	6th	36.4	7th	20.3	7th	25.2	2nd	27.3	6th	25.7	5th
ReAct	gpt-3.5-turbo	84.6	2nd	4.0	14th	33.0	4th	19.6	7th	40.6	4th	24.9	6th
ReAct	vicuna-13b	19.9	10th	6.0	13th	7.1	16th	20.7	4th	11.0	17th	18.2	7th
PAL	tulu-7b	9.1	15th	26.8	9th	11.5	12th	18.9	8th	15.8	9th	18.1	8th
PAL	vicuna-13b	4.5	17th	40.9	4th	2.3	20th	16.7	9th	15.9	8th	16.5	9th
ReAct	llama2-13b	16.7	13th	0.7	19th	23.2	5th	15.0	10th	13.5	12th	14.6	10th
PAL	llama2-7b-chat	18.7	12th	2.8	15th	16.1	8th	12.4	11th	12.5	14th	12.4	11th
PAL	codellama-13b	4.4	18th	38.3	5th	8.1	14th	10.0	14th	16.9	7th	11.7	12th
PAL	toolllama2-7b	1.6	20th	24.4	10th	4.6	18th	12.2	12th	10.2	18th	11.7	13th
ReAct	tulu-7b	4.0	19th	27.8	8th	7.9	15th	10.3	13th	13.2	13th	11.0	14th
DFSDT	gpt-4-1106	20.6	9th	9.6	12th	11.8	11th	9.9	15th	14.0	11th	10.9	15th
FC	gpt-4-1106	24.7	7th	10.9	11th	10.2	13th	8.2	18th	15.3	10th	9.9	16th
FC	gpt-3.5-turbo	19.0	11th	1.0	17th	15.9	9th	8.8	16th	12.0	15th	9.6	17th
ReAct	toolllama2-7b	15.0	14th	2.2	16th	5.7	17th	8.3	17th	7.6	19th	8.1	18th
DFSDT	gpt-3.5-turbo	20.7	8th	0.2	20th	13.8	10th	4.8	20th	11.6	16th	6.5	19th
ReAct	codellama-13b	0.2	21th	0.8	18th	0.7	21th	7.0	19th	0.6	21th	5.4	20th
DFSDT	toolllama2-7b	7.1	16th	0.0	21th	2.3	19th	3.5	21th	3.1	20th	3.4	21th

Fig.12 Evaluation Results of R-Eval on AMiner, wiki and overall ranking.

Evaluation Result - Performance Ranking

The same
Workflow +
LLM,

Same
Domain Task

Differenct
Level

Workflow	LLM	aminer KS		aminer KU		aminer KA		Overall Average (Level 1, 2, 3)					
		1-3	Rank	2-4	Rank	3-5	Rank	wiki	Rank	aminer	Rank	all	Rank
ReAct	gpt-4-1106	89.7	1st	46.7	3rd	57.7	1st	38.8	1st	64.7	1st	45.3	1st
PAL	gpt-3.5-turbo	80.1	3rd	50.7	2nd	54.9	2nd	19.9	6th	61.9	2nd	30.4	2nd
PAL	gpt-4-1106	59.3	4th	56.8	1st	52.7	3rd	20.3	5th	56.2	3rd	29.2	3rd
ReAct	llama2-7b-chat	45.2	5th	36.5	6th	21.5	6th	23.8	3rd	34.4	5th	26.4	4th
PAL	llama2-13b	25.3	6th	36.4	7th	20.3	7th	25.2	2nd	27.3	6th	25.7	5th
ReAct	gpt-3.5-turbo	84.6	2nd	4.0	14th	33.0	4th	19.6	7th	40.6	4th	24.9	6th
ReAct	vicuna-13b	19.9	10th	6.0	13th	7.1	16th	20.7	4th	11.0	17th	18.2	7th
PAL	tulu-7b	9.1	15th	26.8	9th	11.5	12th	18.9	8th	15.8	9th	18.1	8th
PAL	vicuna-13b	4.5	17th	40.9	4th	2.3	20th	16.7	9th	15.9	8th	16.5	9th
ReAct	llama2-13b	16.7	13th	0.7	19th	23.2	5th	15.0	10th	13.5	12th	14.6	10th
PAL	llama2-7b-chat	18.7	12th	2.8	15th	16.1	8th	12.4	11th	12.5	14th	12.4	11th
PAL	codellama-13b	4.4	18th	38.3	5th	8.1	14th	10.0	14th	16.9	7th	11.7	12th
PAL	toolllama2-7b	1.6	20th	24.4	10th	4.6	18th	12.2	12th	10.2	18th	11.7	13th
ReAct	tulu-7b	4.0	19th	27.8	8th	7.9	15th	10.3	13th	13.2	13th	11.0	14th
DFSDT	gpt-4-1106	20.6	9th	9.6	12th	11.8	11th	9.9	15th	14.0	11th	10.9	15th
FC	gpt-4-1106	24.7	7th	10.9	11th	10.2	13th	8.2	18th	15.3	10th	9.9	16th
FC	gpt-3.5-turbo	19.0	11th	1.0	17th	15.9	9th	8.8	16th	12.0	15th	9.6	17th
ReAct	toolllama2-7b	15.0	14th	2.2	16th	5.7	17th	8.3	17th	7.6	19th	8.1	18th
DFSDT	gpt-3.5-turbo	20.7	8th	0.2	20th	13.8	10th	4.8	20th	11.6	16th	6.5	19th
ReAct	codellama-13b	0.2	21th	0.8	18th	0.7	21th	7.0	19th	0.6	21th	5.4	20th
DFSDT	toolllama2-7b	7.1	16th	0.0	21th	2.3	19th	3.5	21th	3.1	20th	3.4	21th

Fig.12 Evaluation Results of R-Eval on AMiner, wiki and overall ranking.

Evaluation Result - Performance Ranking

Workflow	LLM	aminer KS		aminer KU		aminer KA		Overall Average (Level 1, 2, 3)					
		1-3	Rank	2-4	Rank	3-5	Rank	wiki	Rank	aminer	Rank	all	Rank
ReAct	gpt-4-1106	89.7	1st	46.7	3rd	57.7	1st	38.8	1st	64.7	1st	45.3	1st
PAL	gpt-3.5-turbo	80.1	3rd	50.7	2nd	54.9	2nd	19.9	6th	61.9	2nd	30.4	2nd
PAL	gpt-4-1106	59.3	4th	56.8	1st	52.7	3rd	20.3	5th	56.2	3rd	29.2	3rd
ReAct	llama2-7b-chat	45.2	5th	36.5	6th	21.5	6th	23.8	3rd	34.4	5th	26.4	4th
PAL	llama2-13b	25.3	6th	36.4	7th	20.3	7th	25.2	2nd	27.3	6th	25.7	5th
ReAct	gpt-3.5-turbo	84.6	2nd	4.0	14th	33.0	4th	19.6	7th	40.6	4th	24.9	6th
ReAct	vicuna-13b	19.9	10th	6.0	13th	7.1	16th	20.7	4th	11.0	17th	18.2	7th
PAL	tulu-7b	9.1	15th	26.8	9th	11.5	12th	18.9	8th	15.8	9th	18.1	8th
PAL	vicuna-13b	4.5	17th	40.9	4th	2.3	20th	16.7	9th	15.9	8th	16.5	9th
ReAct	llama2-13b	16.7	13th	0.7	19th	23.2	5th	15.0	10th	13.5	12th	14.6	10th
PAL	llama2-7b-chat	18.7	12th	2.8	15th	16.1	8th	12.4	11th	12.5	14th	12.4	11th
PAL	codellama-13b	4.4	18th	38.3	5th	8.1	14th	10.0	14th	16.9	7th	11.7	12th
PAL	toolllama2-7b	1.6	20th	24.4	10th	4.6	18th	12.2	12th	10.2	18th	11.7	13th
ReAct	tulu-7b	4.0	19th	27.8	8th	7.9	15th	10.3	13th	13.2	13th	11.0	14th
DFSDT	gpt-4-1106	20.6	9th	9.6	12th	11.8	11th	9.9	15th	14.0	11th	10.9	15th
FC	gpt-4-1106	24.7	7th	10.9	11th	10.2	13th	8.2	18th	15.3	10th	9.9	16th
FC	gpt-3.5-turbo	19.0	11th	1.0	17th	15.9	9th	8.8	16th	12.0	15th	9.6	17th
ReAct	toolllama2-7b	15.0	14th	2.2	16th	5.7	17th	8.3	17th	7.6	19th	8.1	18th
DFSDT	gpt-3.5-turbo	20.7	8th	0.2	20th	13.8	10th	4.8	20th	11.6	16th	6.5	19th
ReAct	codellama-13b	0.2	21th	0.8	18th	0.7	21th	7.0	19th	0.6	21th	5.4	20th
DFSDT	toolllama2-7b	7.1	16th	0.0	21th	2.3	19th	3.5	21th	3.1	20th	3.4	21th

The same
Workflow +
LLM,

Different
Domain

Fig.12 Evaluation Results of R-Eval on AMiner, wiki and overall ranking.

Evaluation Result - Performance Ranking

Workflow	LLM	aminer KS		aminer KU		aminer KA		Overall Average (Level 1, 2, 3)					
		1-3	Rank	2-4	Rank	3-5	Rank	wiki	Rank	aminer	Rank	all	Rank
ReAct	gpt-4-1106	89.7	1st	46.7	3rd	57.7	1st	38.8	1st	64.7	1st	45.3	1st
PAL	gpt-3.5-turbo	80.1	3rd	50.7	2nd	54.9	2nd	19.9	6th	61.9	2nd	30.4	2nd
PAL	gpt-4-1106	59.3	4th	56.8	1st	52.7	3rd	20.3	5th	56.2	3rd	29.2	3rd
ReAct	llama2-7b-chat	45.2	5th	36.5	6th	21.5	6th	23.8	3rd	34.4	5th	26.4	4th
PAL	llama2-13b	25.3	6th	36.4	7th	20.3	7th	25.2	2nd	27.3	6th	25.7	5th
ReAct	gpt-3.5-turbo	84.6	2nd	4.0	14th	33.0	4th	19.6	7th	40.6	4th	24.9	6th
ReAct	vicuna-13b	19.9	10th	6.0	13th	7.1	16th	20.7	4th	11.0	17th	18.2	7th
PAL	tulu-7b	9.1	15th	26.8	9th	11.5	12th	18.9	8th	15.8	9th	18.1	8th
PAL	vicuna-13b	4.5	17th	40.9	4th	2.3	20th	16.7	9th	15.9	8th	16.5	9th
ReAct	llama2-13b	16.7	13th	0.7	19th	23.2	5th	15.0	10th	13.5	12th	14.6	10th
PAL	llama2-7b-chat	18.7	12th	2.8	15th	16.1	8th	12.4	11th	12.5	14th	12.4	11th
PAL	codellama-13b	4.4	18th	38.3	5th	8.1	14th	10.0	14th	16.9	7th	11.7	12th
PAL	toollama2-7b	1.6	20th	24.4	10th	4.6	18th	12.2	12th	10.2	18th	11.7	13th
ReAct	tulu-7b	4.0	19th	27.8	8th	7.9	15th	10.3	13th	13.2	13th	11.0	14th
DFSDT	gpt-4-1106	20.6	9th	9.6	12th	11.8	11th	9.9	15th	14.0	11th	10.9	15th
FC	gpt-4-1106	24.7	7th	10.9	11th	10.2	13th	8.2	18th	15.3	10th	9.9	16th
FC	gpt-3.5-turbo	19.0	11th	1.0	17th	15.9	9th	8.8	16th	12.0	15th	9.6	17th
ReAct	toollama2-7b	15.0	14th	2.2	16th	5.7	17th	8.3	17th	7.6	19th	8.1	18th
DFSDT	gpt-3.5-turbo	20.7	8th	0.2	20th	13.8	10th	4.8	20th	11.6	16th	6.5	19th
ReAct	codellama-13b	0.2	21th	0.8	18th	0.7	21th	7.0	19th	0.6	21th	5.4	20th
DFSDT	toollama2-7b	7.1	16th	0.0	21th	2.3	19th	3.5	21th	3.1	20th	3.4	21th

The same
LLM &
Domain

Different
Workflow

Fig.12 Evaluation Results of R-Eval on AMiner, wiki and overall ranking.

Evaluation Result - Performance Ranking

Workflow	LLM	aminer KS		aminer KU		aminer KA		Overall Average (Level 1, 2, 3)					
		1-3	Rank	2-4	Rank	3-5	Rank	wiki	Rank	aminer	Rank	all	Rank
ReAct	gpt-4-1106	89.7	1st	46.7	3rd	57.7	1st	38.8	1st	64.7	1st	45.3	1st
PAL	gpt-3.5-turbo	80.1	3rd	50.7	2nd	54.9	2nd	19.9	6th	61.9	2nd	30.4	2nd
PAL	gpt-4-1106	59.3	4th	56.8	1st	52.7	3rd	20.3	5th	56.2	3rd	29.2	3rd
ReAct	llama2-7b-chat	45.2	5th	36.5	6th	21.5	6th	23.8	3rd	34.4	5th	26.4	4th
PAL	llama2-13b	25.3	6th	36.4	7th	20.3	7th	25.2	2nd	27.3	6th	25.7	5th
ReAct	gpt-3.5-turbo	84.6	2nd	4.0	14th	33.0	4th	19.6	7th	40.6	4th	24.9	6th
ReAct	vicuna-13b	19.9	10th	6.0	13th	7.1	16th	20.7	4th	11.0	17th	18.2	7th
PAL	tulu-7b	9.1	15th	26.8	9th	11.5	12th	18.9	8th	15.8	9th	18.1	8th
PAL	vicuna-13b	4.5	17th	40.9	4th	2.3	20th	16.7	9th	15.9	8th	16.5	9th
ReAct	llama2-13b	16.7	13th	0.7	19th	23.2	5th	15.0	10th	13.5	12th	14.6	10th
PAL	llama2-7b-chat	18.7	12th	2.8	15th	16.1	8th	12.4	11th	12.5	14th	12.4	11th
PAL	codellama-13b	4.4	18th	38.3	5th	8.1	14th	10.0	14th	16.9	7th	11.7	12th
PAL	toolllama2-7b	1.6	20th	24.4	10th	4.6	18th	12.2	12th	10.2	18th	11.7	13th
ReAct	tulu-7b	4.0	19th	27.8	8th	7.9	15th	10.3	13th	13.2	13th	11.0	14th
DFSDT	gpt-4-1106	20.6	9th	9.6	12th	11.8	11th	9.9	15th	14.0	11th	10.9	15th
FC	gpt-4-1106	24.7	7th	10.9	11th	10.2	13th	8.2	18th	15.3	10th	9.9	16th
FC	gpt-3.5-turbo	19.0	11th	1.0	17th	15.9	9th	8.8	16th	12.0	15th	9.6	17th
ReAct	toolllama2-7b	15.0	14th	2.2	16th	5.7	17th	8.3	17th	7.6	19th	8.1	18th
DFSDT	gpt-3.5-turbo	20.7	8th	0.2	20th	13.8	10th	4.8	20th	11.6	16th	6.5	19th
ReAct	codellama-13b	0.2	21th	0.8	18th	0.7	21th	7.0	19th	0.6	21th	5.4	20th
DFSDT	toolllama2-7b	7.1	16th	0.0	21th	2.3	19th	3.5	21th	3.1	20th	3.4	21th

The same
Workflow &
Domain

Different
LLM

Fig.12 Evaluation Results of R-Eval on AMiner, wiki and overall ranking.

Visualization of the performance

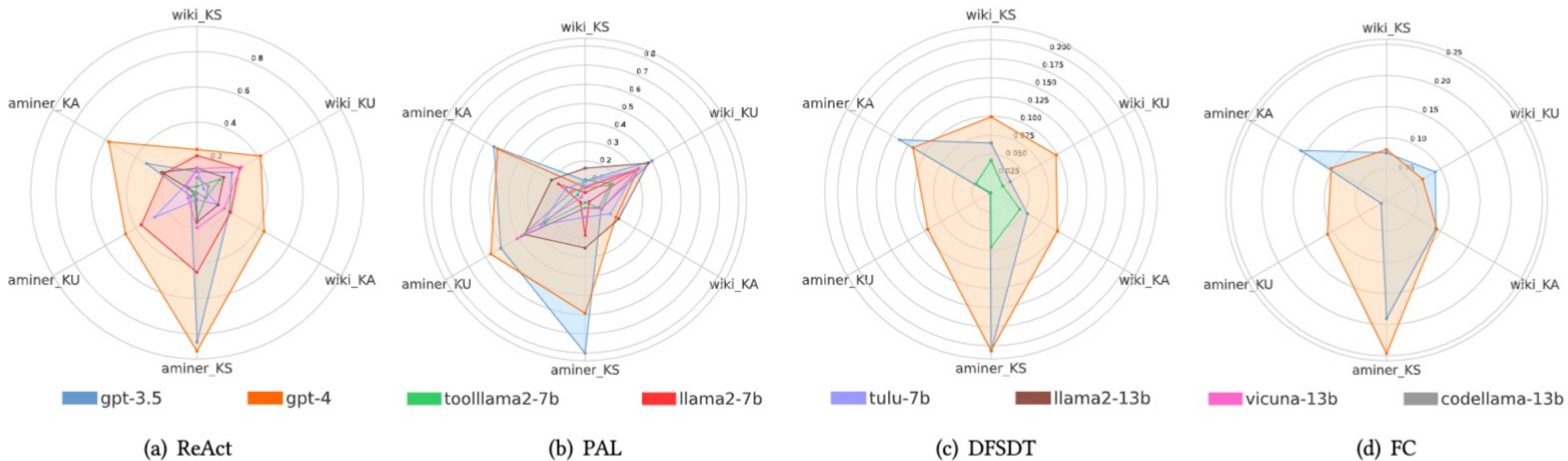


Fig.13 Radar map of single system's performance.

Error Analysis & Efficiency Evaluation

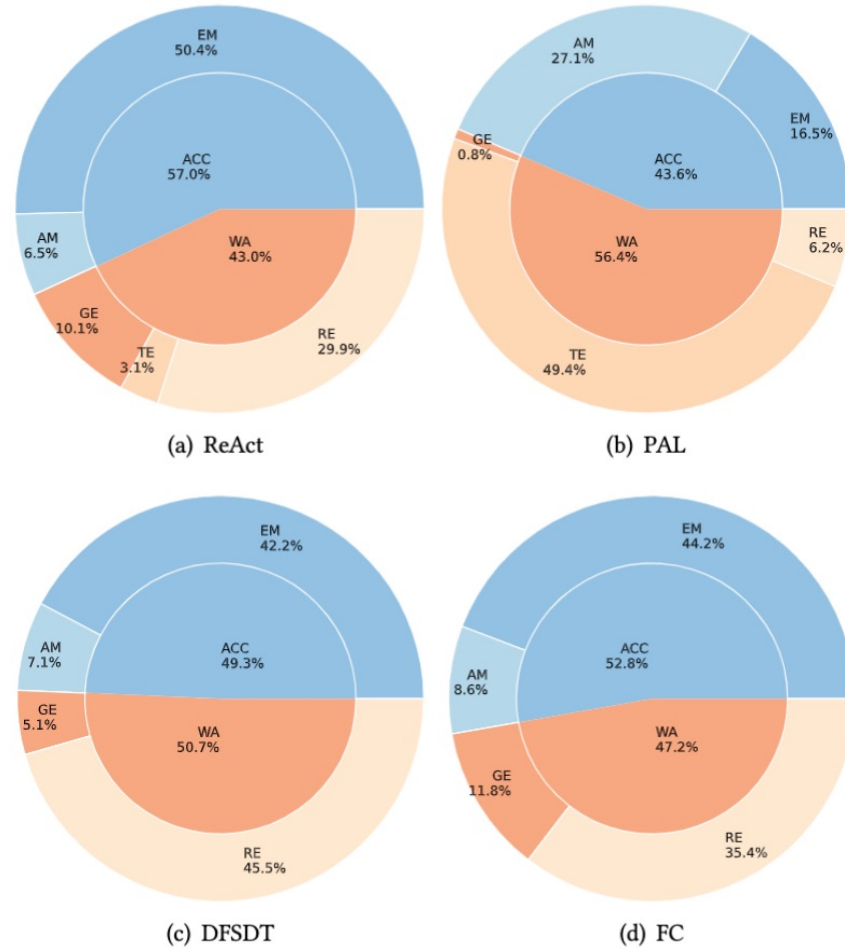


Fig.14 Error Analysis.

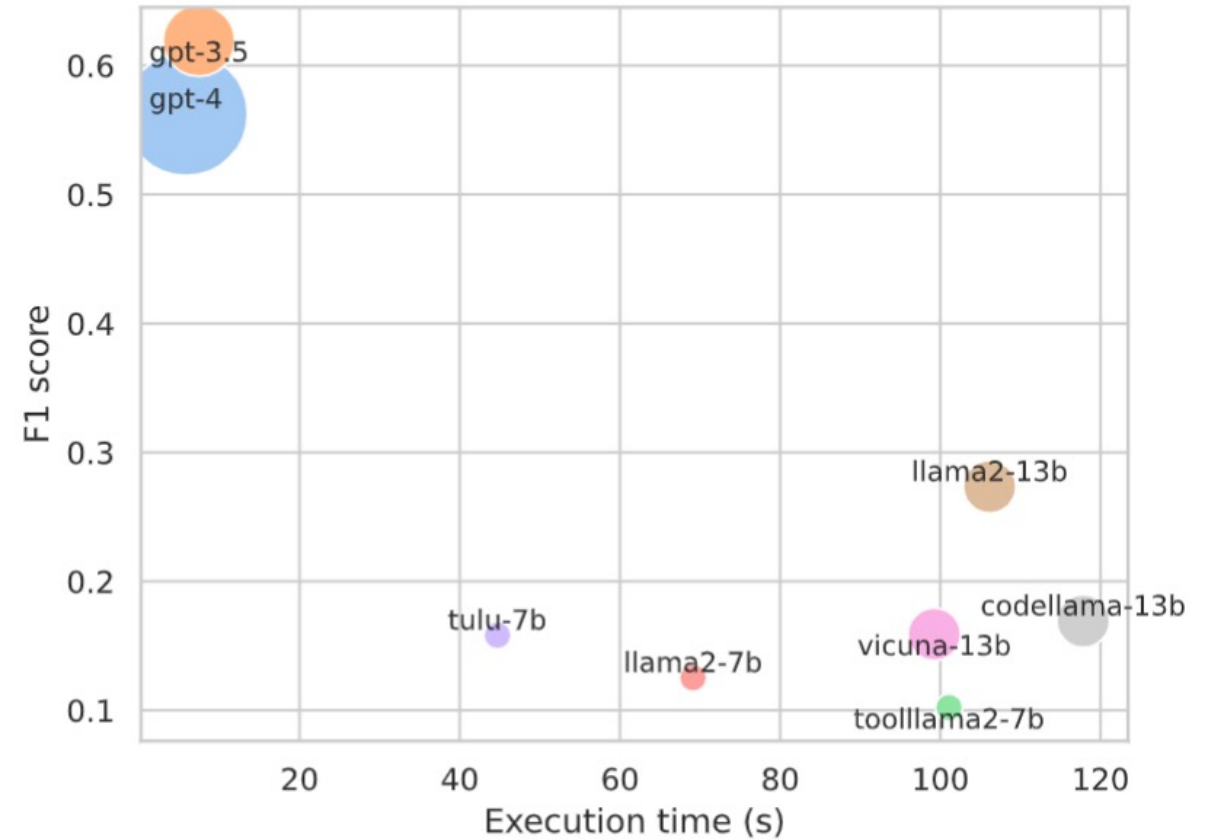


Fig.15 Efficiency Evaluation.

Challenges & Future Directions

In the area of Tool Learning and Specific Domain Application, there still remains some challenges.

- High Latency in Tool Learning,
- **Rigorous and Comprehensive Evaluation,**
- **Comprehensive and Accessible Tools,**
- Safe and Robust Tool Learning,
- **Real-Word Benchmark for Tool Learning,**
- Tool Learning with Multi-Modal

Welcome to Follow our work!

SoAy Applied System Link:

<http://soay.aminer.cn>



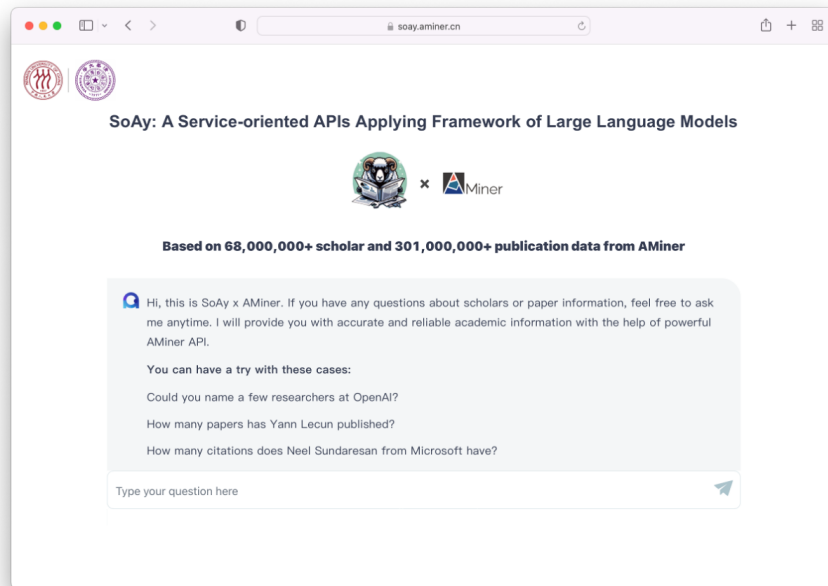
Github Link:

<https://github.com/RUCKBReasoning/SoAy>



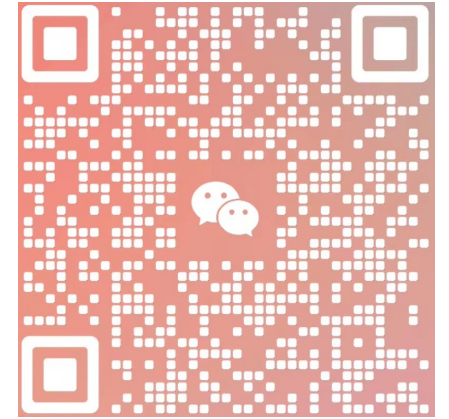
R-Eval Github Link:

<https://github.com/THU-KEG/R-Eval>



FAQ

- ▶ What would be needed for a user with their own domain-specific dataset to apply this framework on their data?
- ▶ What kind of retrieval components of dense retrieval or generative retrieval are built it?
- ▶ Can LLM based on knowledge graph retrieval also be incorporated under R-eval?
- ▶ R-eval includes the retrieval component inside? Then, how other collections can be added for RALLM?
- ▶ Is R-Eval just to collect some of the existing methods and benchmarks, and integrate them together to conduct a comprehensive evaluation?
- ▶ There are multiple LLMs missing in two rightmost figures in Figure 4.



**Yuanchun's
WeChat**

Thank you!

