



Assisting Scientific Research with Structure-Aware Large Language Models

Yu Zhang

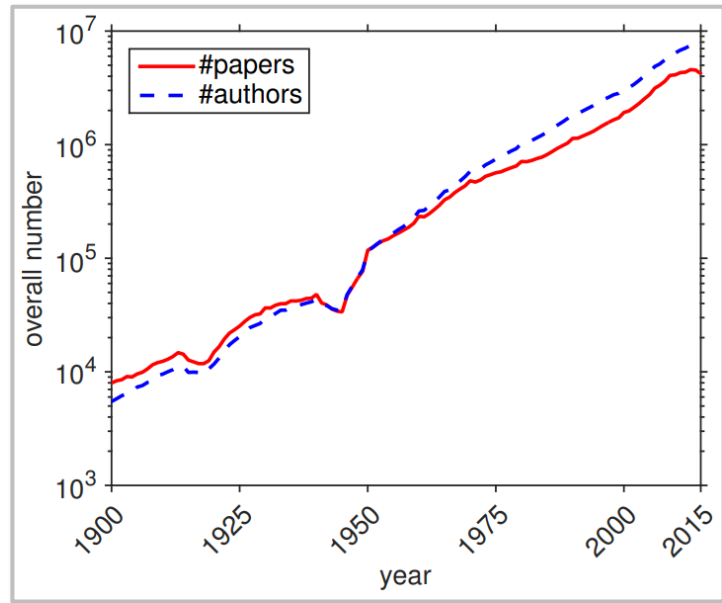
Texas A&M University

yuzhang@tamu.edu

April 11, 2025

Explosion of Scientific Text Data

- The volume of scientific publications is growing exponentially.
 - Doubling every **12** years [1]
 - Reaching **240,000,000** in 2019 [2]



- Papers on emerging topics can be released in a torrent.
 - About **4,000 peer-reviewed** papers on COVID-19 by the end of April 2020 [3]
 - DeepSeek-R1 was uploaded to arXiv on **January 22, 2025** and **cited 578 times** on **April 8, 2025**.

Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [PDF] arXiv

[D Guo, D Yang, H Zhang, J Song, R Zhang...](#) - arXiv preprint arXiv ..., 2025 - arxiv.org

... **DeepSeek-R1**, which incorporates multi-stage training and cold-start data before RL. **DeepSeekR1** ... , we open-source **DeepSeek-R1-Zero**, **DeepSeek-R1**, and six dense models (1.5B, ...

☆ Cited by 578 Related articles

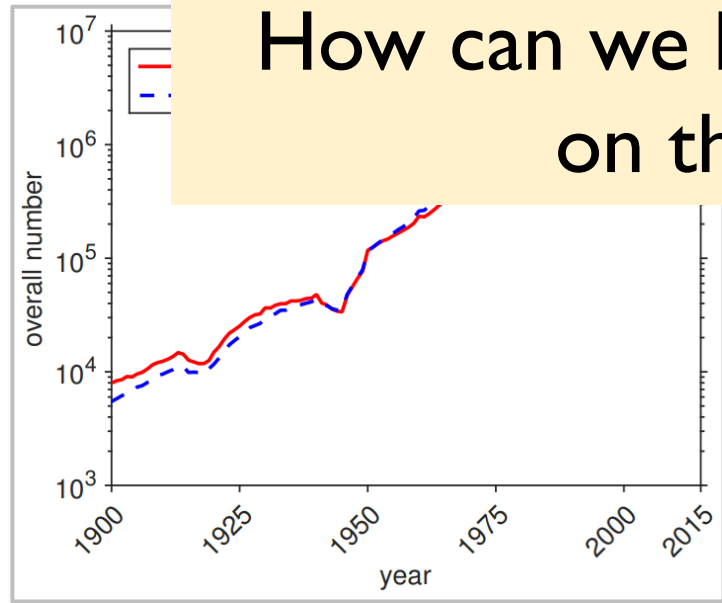
[1] A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations. KDD 2017.

[2] Microsoft Academic Graph: When Experts are Not Enough. Quantitative Science Studies 2020.

[3] <https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent>

Explosion of Scientific Text Data

- The volume of scientific publications is growing exponentially.
 - Doubling every **12** years [1]
 - Reaching **240,000,000** in 2019 [2]
- Papers on emerging topics can be released in a torrent.
 - About **4,000 peer-reviewed** papers on COVID-19 by the end of April 2020 [3]
 - DeepSeek R1 was uploaded to arXiv on **January 1, 2025** and reached over 100,000 downloads by **April 8, 2025**.



How can we help researchers **stay updated** on their fields of interest?

in llms via reinforcement learning

[D Guo, D Yang, H Zhang, J Song, R Zhang...](#) - arXiv preprint arXiv ..., 2025 - arxiv.org

... **DeepSeek-R1**, which incorporates multi-stage training and cold-start data before RL. **DeepSeekR1** ... , we open-source **DeepSeek-R1-Zero**, **DeepSeek-R1**, and six dense models (1.5B, ...

☆ Cited by 578 Related articles

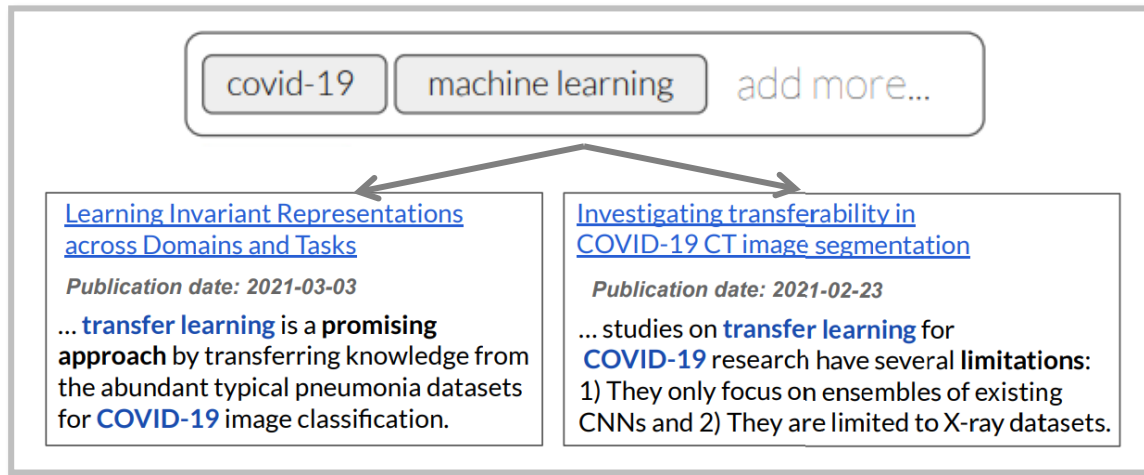
[1] A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations. KDD 2017.

[2] Microsoft Academic Graph: When Experts are Not Enough. Quantitative Science Studies 2020.

[3] <https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent>

How can NLP assist scientific research?

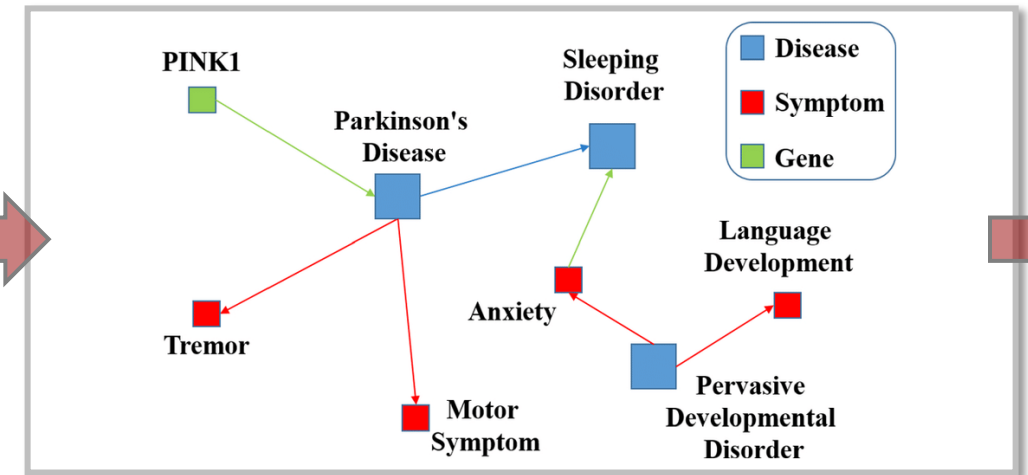
Retrieving and Analyzing Relevant Literature



- **Example tasks:**

- Predict the diseases, chemicals, and viruses relevant to each paper.
- Retrieve papers relevant to both “*Betacoronavirus*” and “*Paxlovid*”.
- Find papers refuting the claim “*CX3CR1* impairs *T* cell survival”.

Uncovering Knowledge Structures/Graphs





- **Example tasks:**

- Find protein entities relevant to “*Parkinson's disease*” from relevant literature.
- Predict the relationship between “*Tremor*” and “*Sleeping Disorder*”.

How can NLP assist scientific research?

Generating Hypotheses and Suggesting Directions



Hypothesis: Graph convolutional networks (GCNs) can effectively model polypharmacy side effects by leveraging the intricate relationships among drugs, their targets, and biological pathways encoded in drug-target interaction networks, enabling the prediction of potential adverse drug interactions and facilitating personalized medication management.

- **Example tasks:**

- Generate a new hypothesis based on the 100 most recent papers on “*Polypharmacy Side Effects*”.
- Evaluate the novelty of an idea for modeling “*Polypharmacy Side Effects*” in comparison with previous studies.

Reviewing Research Outcomes

Reviewer Console

Bidding 1 - 4 of 4 Show: 25 50 100 All Clear All Filters

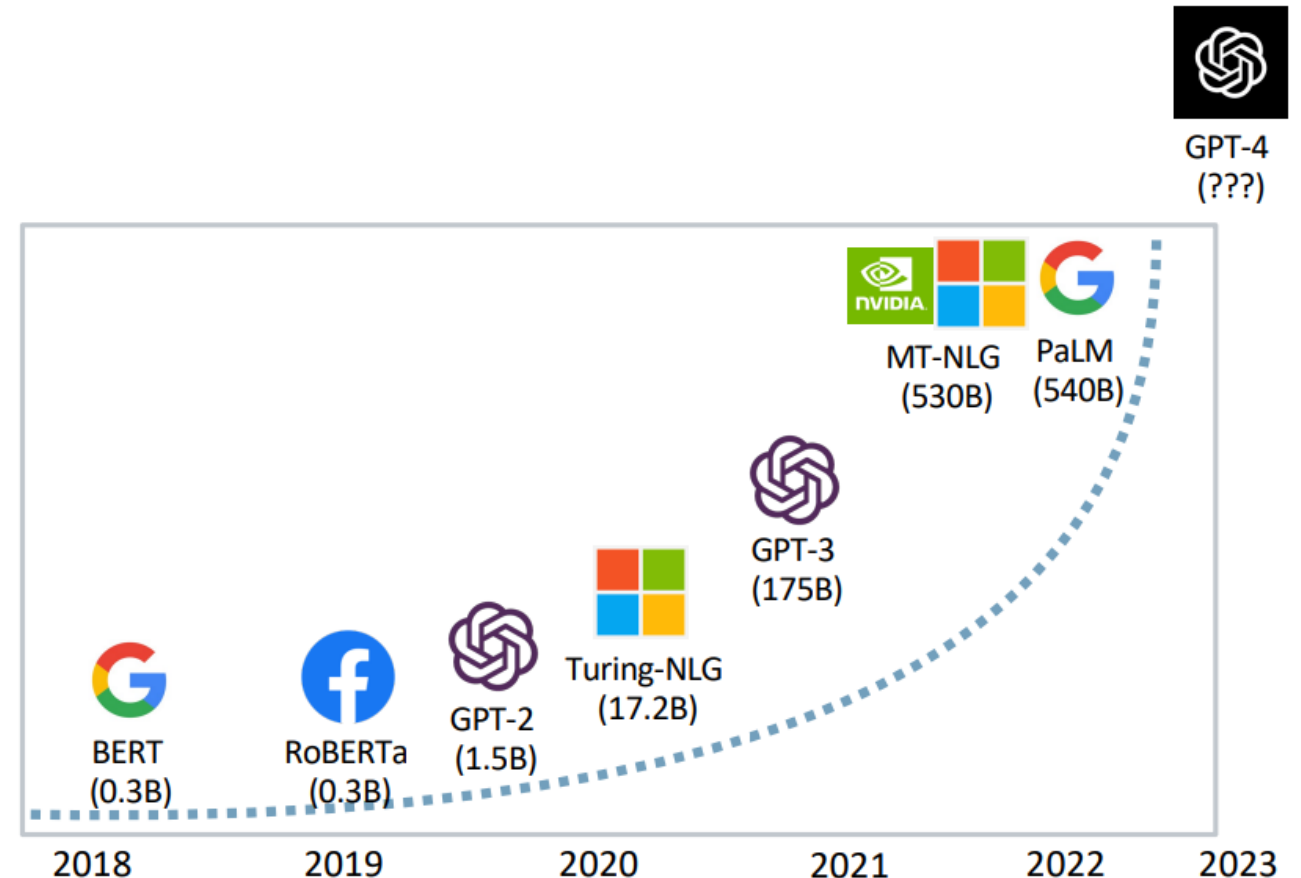
Paper ID↑	Title	Subject Areas		Review & Discussion	Relevance
		Primary	Secondary		
e.g. <3	filter...	filter...	filter...		e.g. <3
26	Research Paper Zero 1 Show Abstract	MARINE VESSELS -> Hull	AUTOMOBILES -> Engines		0.32
27	Scientific Paper Z Show Abstract	AUTOMOBILES -> Engines	MARINE VESSELS		0.80

- **Example tasks:**

- Find qualified reviewers to review a submission.
- Provide constructive feedback to a paper draft.

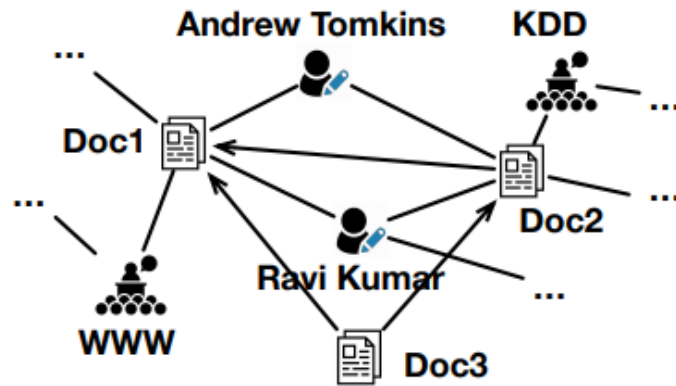
Large Language Models (LLMs) for NLP

- A **unified** model to perform different NLP tasks **with a few or zero examples**
 - I went to the zoo to see giraffes, lions, and {zebras, spoon}. (*Lexical semantics*)
 - I was engaged and on the edge of my seat the whole time. The movie was {good, bad}. (*Text classification*)
 - The word for “pretty” in Spanish is {bonita, hola}. (*Translation*)
 - $3 + 8 + 4 = \{\underline{15}, 11\}$ (*Math*)
 - ...

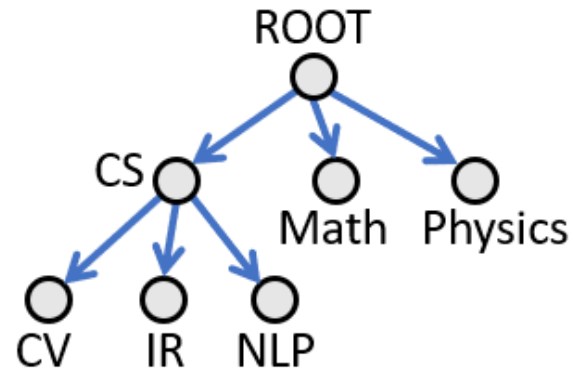


Are LLMs aware of **structured information**?

Structured Information Associated with Scientific Text



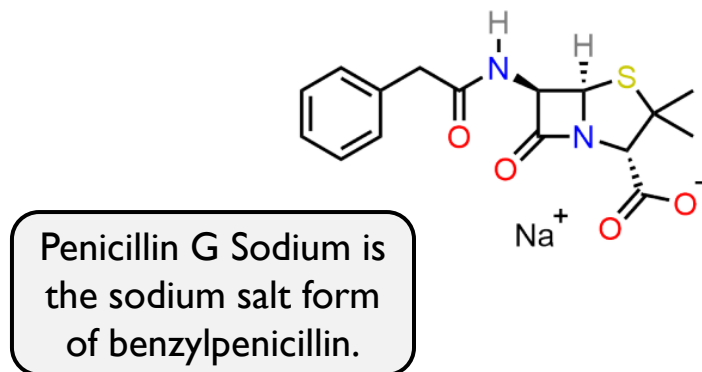
Metadata/Network



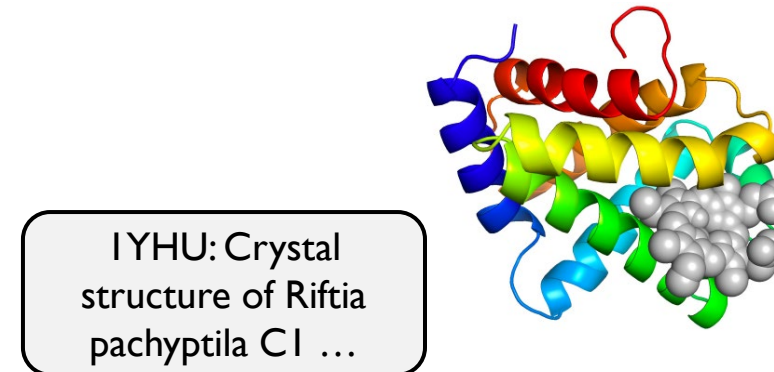
Hierarchy/Taxonomy



Knowledge Graph



Text-Paired Molecule



Text-Paired Protein

LLMs may not be structure-aware!



You

What is the most cited paper in WWW 2017? Please list its title and authors.



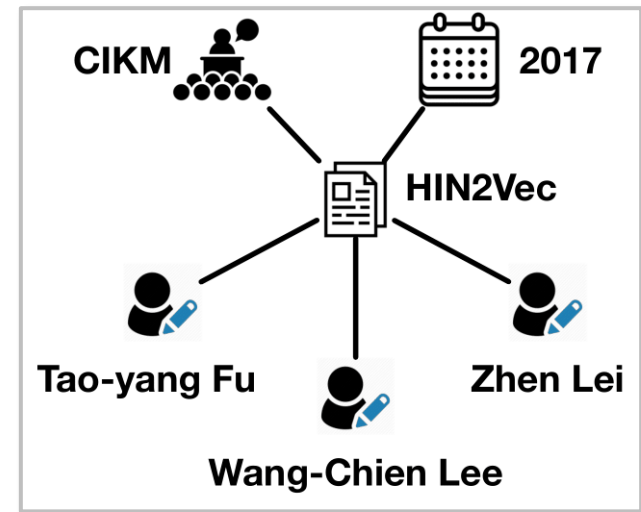
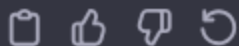
ChatGPT

As of my last update in January 2022, I don't have the real-time citation information, but one of the highly cited papers from WWW 2017 is:

Title: "HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning"

Authors: Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, Edward Y. Chang

Please verify the latest citation counts and updates from reliable sources or academic databases for the most current information.



HIN2Vec was published in **CIKM 2017**.

HIN2Vec was written by **Tao-yang Fu, Wang-Chien Lee, and Zhen Lei**.

LLMs may not be structure-aware!

What is the most cited paper in WWW 2017? Please list its title and authors.



Searched 5 sites ▾

The most cited paper from WWW 2017 is "Attention is All You Need," which introduced the Transformer model, a groundbreaking architecture in natural language processing that uses attention mechanisms without relying on recurrent or convolutional layers. The authors of the paper are Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin [ARXIV](#) [NEURIPS PROCEEDINGS](#).

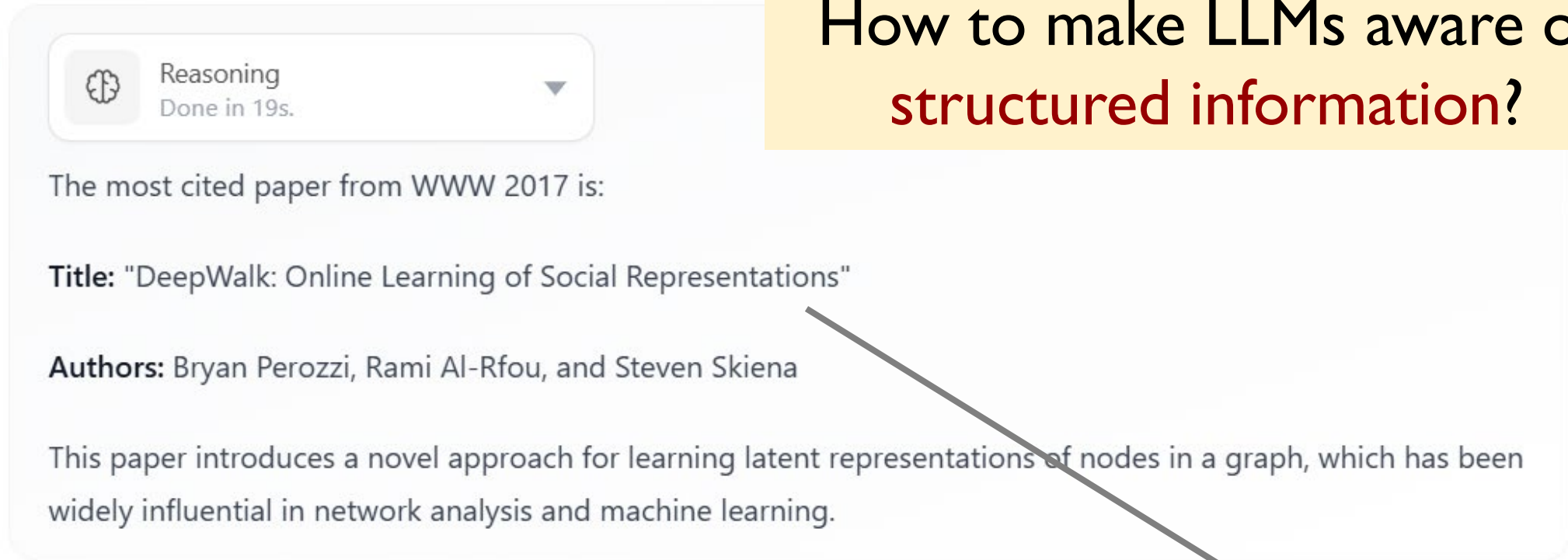


ChatGPT 4o, queried on October 12, 2024

Transformer was published
in **NeurIPS 2017**.

LLMs may not be structure-aware!

What is the most cited paper in WWW 2017? Please list its title and authors.



The screenshot shows a reasoning interface with a dropdown menu labeled 'Reasoning' and 'Done in 19s.'. Below the menu, the text reads: 'The most cited paper from WWW 2017 is:'. The response is structured as follows: 'Title: "DeepWalk: Online Learning of Social Representations"', 'Authors: Bryan Perozzi, Rami Al-Rfou, and Steven Skiena', and a paragraph: 'This paper introduces a novel approach for learning latent representations of nodes in a graph, which has been widely influential in network analysis and machine learning.' An arrow points from the title to a box at the bottom right.

Reasoning
Done in 19s.

The most cited paper from WWW 2017 is:

Title: "DeepWalk: Online Learning of Social Representations"

Authors: Bryan Perozzi, Rami Al-Rfou, and Steven Skiena

This paper introduces a novel approach for learning latent representations of nodes in a graph, which has been widely influential in network analysis and machine learning.

How to make LLMs aware of
structured information?

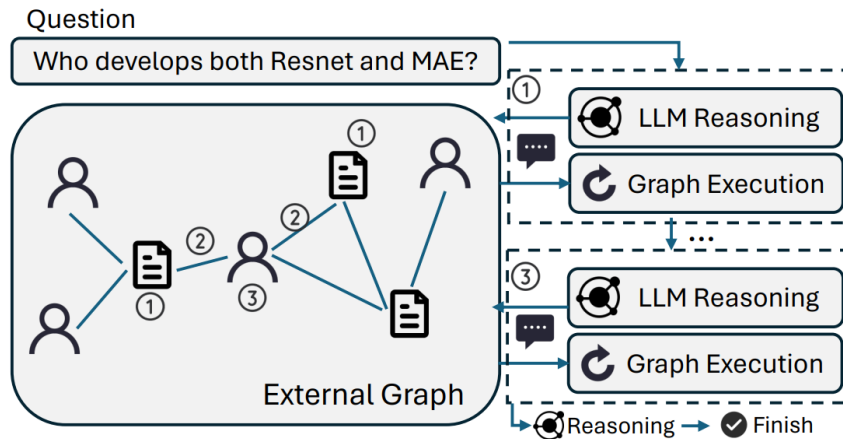
DeepSeek-R1-Distill-Llama-70B, queried on March 21, 2025

DeepWalk was published in
KDD 2014.

Today's Talk: Overview

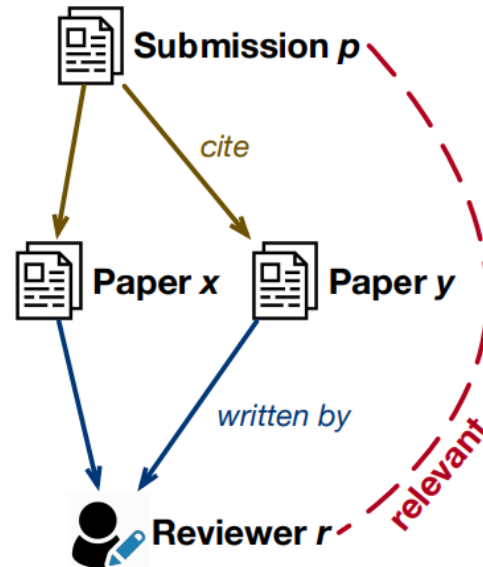
Part I: Structure-Aware LLMs for Scientific Question Answering

Jin et al., ACL 2024



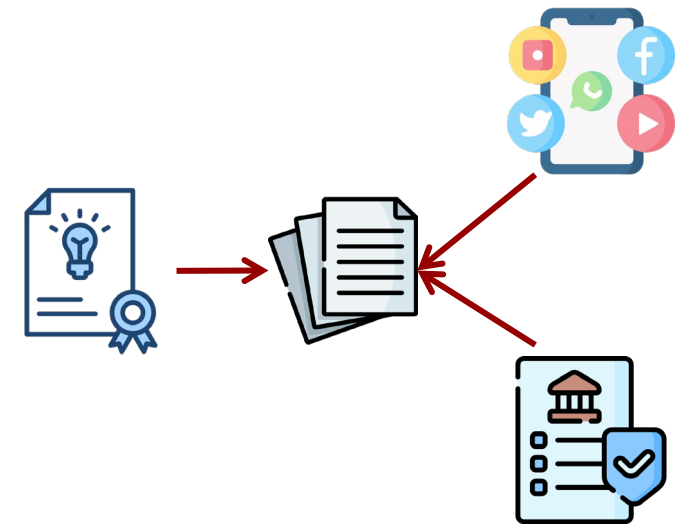
Part II: Structure-Aware LLMs for Scientific Retrieval

Zhang et al., EMNLP 2023
Zhang et al., WWW 2025



Part III: Consumption of LLM Papers by the Broader Public

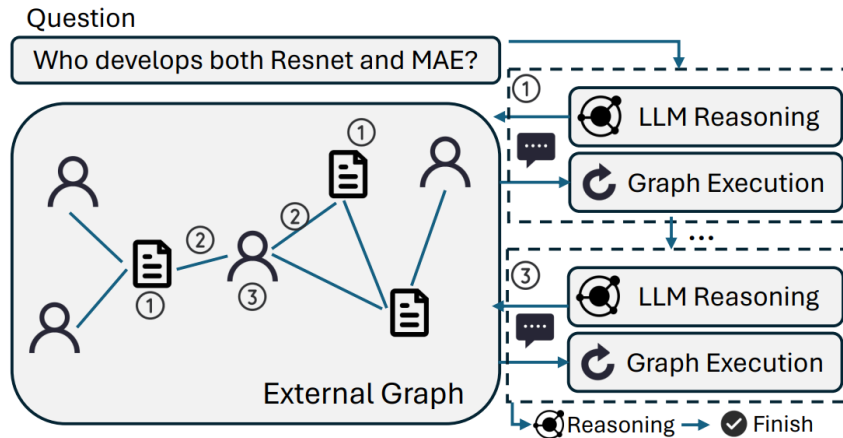
Zhang, In submission



Today's Talk: Part I

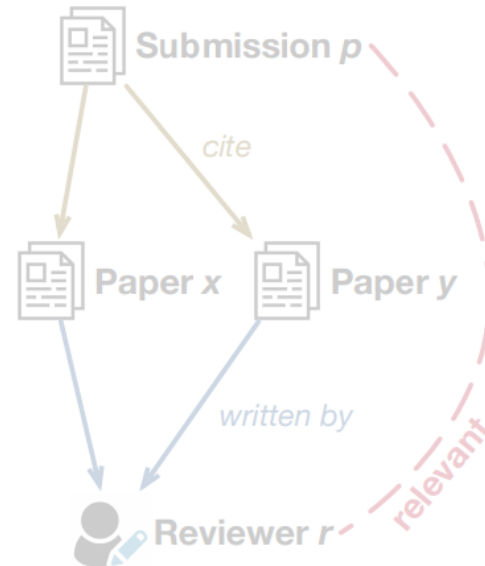
Part I: Structure-Aware LLMs for Scientific Question Answering

Jin et al., ACL 2024



Part II: Structure-Aware LLMs for Scientific Retrieval

Zhang et al., EMNLP 2023
Zhang et al., WWW 2025



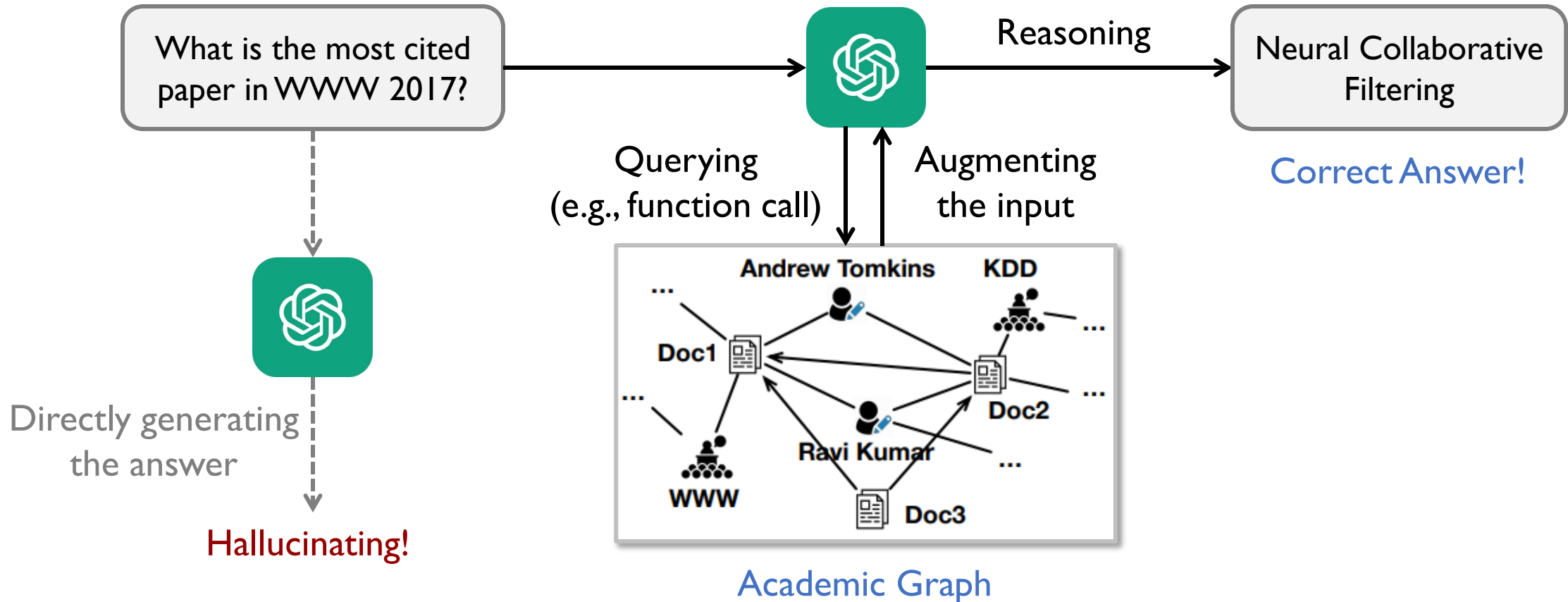
Part III: Consumption of LLM Papers by the Broader Public

Zhang, In submission



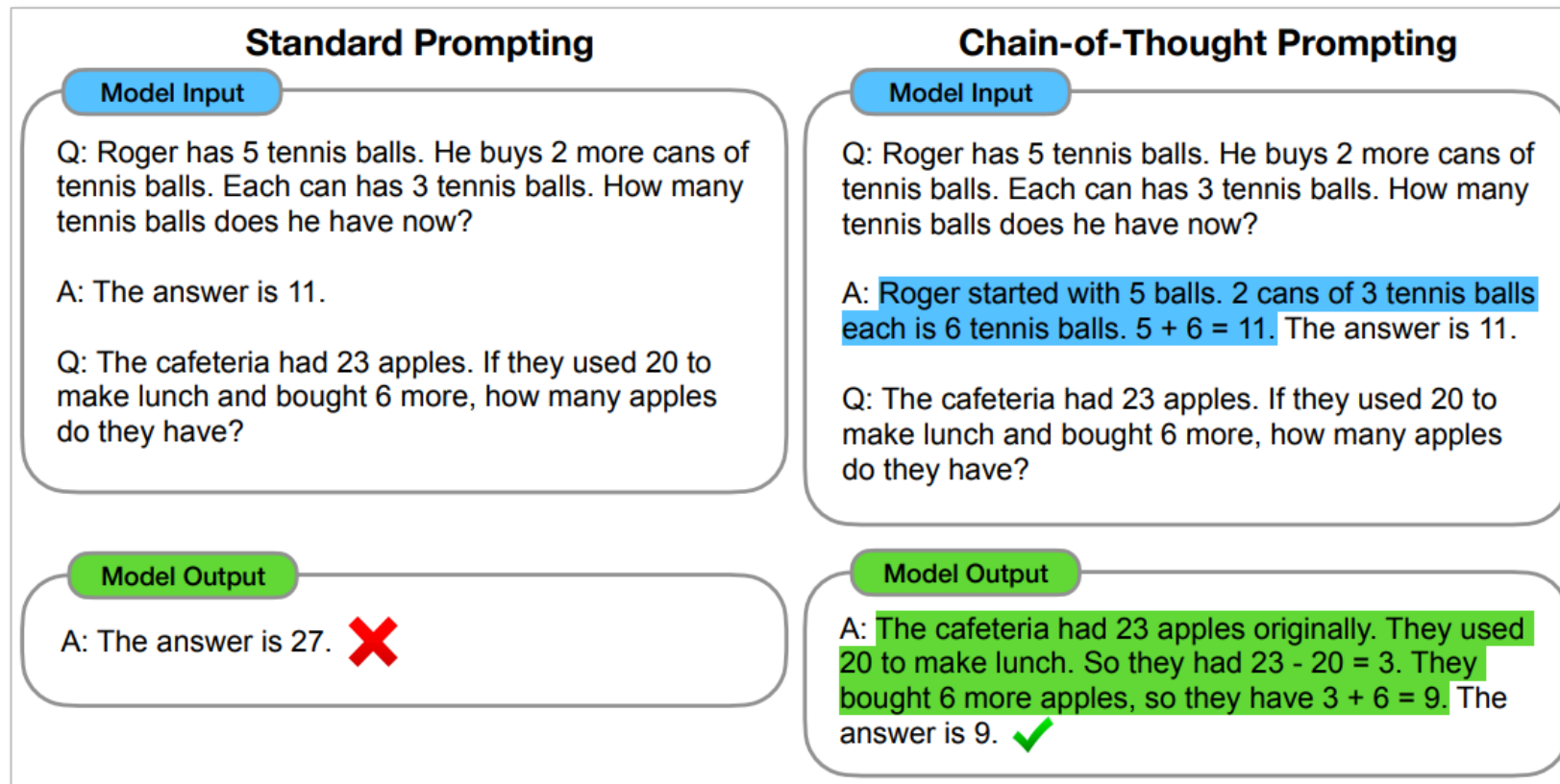
How to make LLMs aware of structured information?

- “Give a man *a fish*, and you feed him for a day; teach a man *to fish*, and you feed him for a lifetime.”
- “Teach an LLM *to know one more edge*, and you help it answer one question; teach an LLM *to explore the graph*, and you help it answer many questions.”



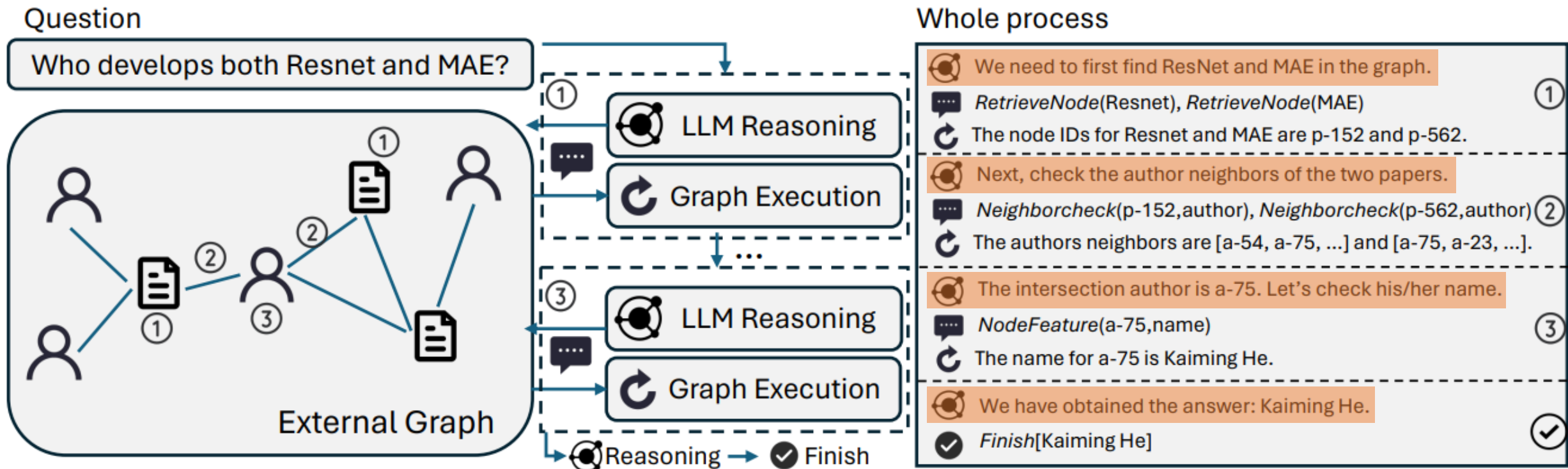
Chain-of-Thought (CoT) Prompting

- Add a series of intermediate reasoning steps in the demonstration examples(s)
- Get the model to explain its reasoning steps before making an answer



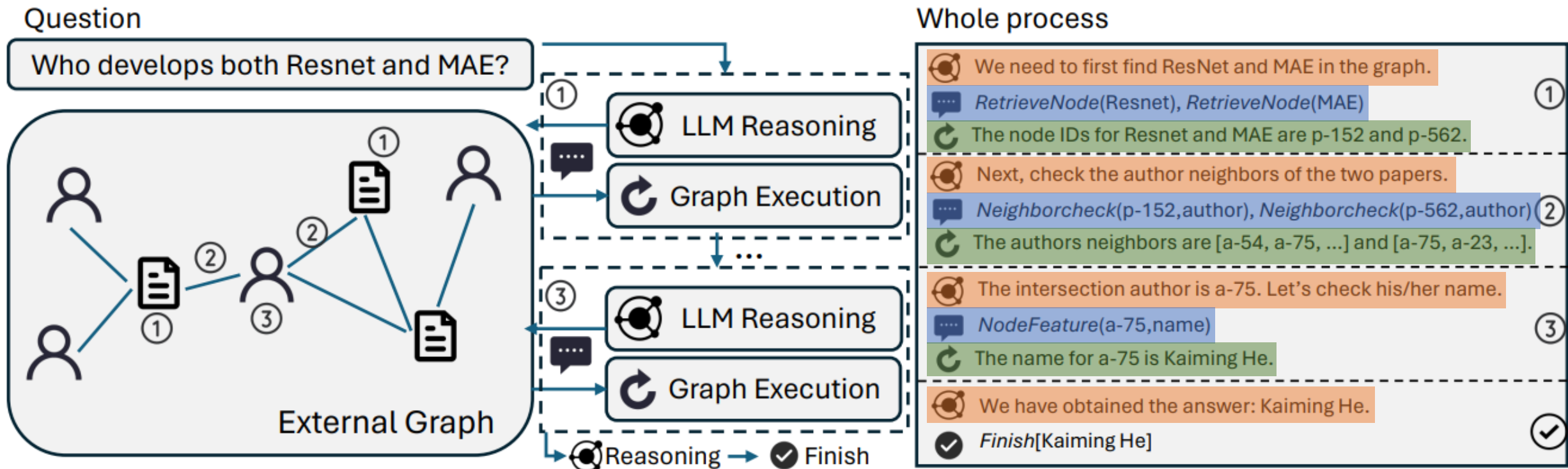
What if we need to access the graph during reasoning?

- Iterative **reasoning**, **interaction** and **execution**
- **Reasoning** – The LLM analyzes: (1) what additional external information from the graph is required; and (2) whether the question is answerable with the current contexts from graphs.






What if we need to access the graph during reasoning?

- Iterative **reasoning**, **interaction** and **execution**
- **Interaction** – The LLM writes function calls.
- **Execution** – The LLM fetches relevant information.



Comparison with Previous Approaches

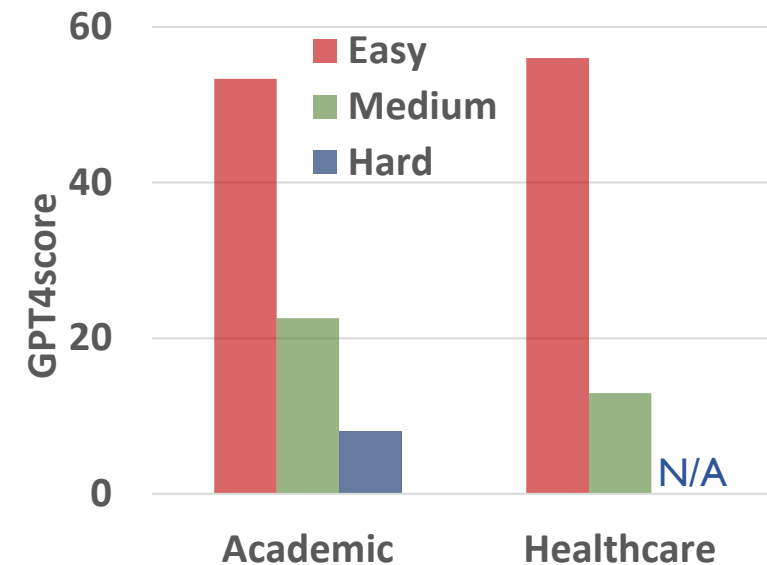
Model		Academic		Healthcare	
		R-L	GPT4score	R-L	GPT4score
Base	 LLaMA-2-13b-chat	8.13	8.03	5.25	13.70
	 Mixtral-8x7b	9.02	8.14	3.88	20.00
	 GPT-3.5-turbo	6.05	12.80	5.83	14.44
Text RAG	LLaMA-2-13b-chat	8.69	8.52	1.44	5.93
	Mixtral-8x7b	8.44	8.02	3.22	16.67
	GPT-3.5-turbo	5.83	9.91	4.57	8.52
Graph RAG	LLaMA-2-13b	22.01	22.97	2.97	4.81
	Mixtral-8x7b	27.77	31.20	8.66	15.19
	GPT-3.5-turbo	18.45	26.98	8.69	14.07
Our Model GRAPH-CoT		31.89	33.48	22.33	28.89

R-L: ROUGE-L (lexical overlap with the ground-truth answer)

GPT4score: Ask GPT-4 to judge if the prediction and the ground truth are equivalent.

Performance of Graph-CoT at Different Difficulty Levels

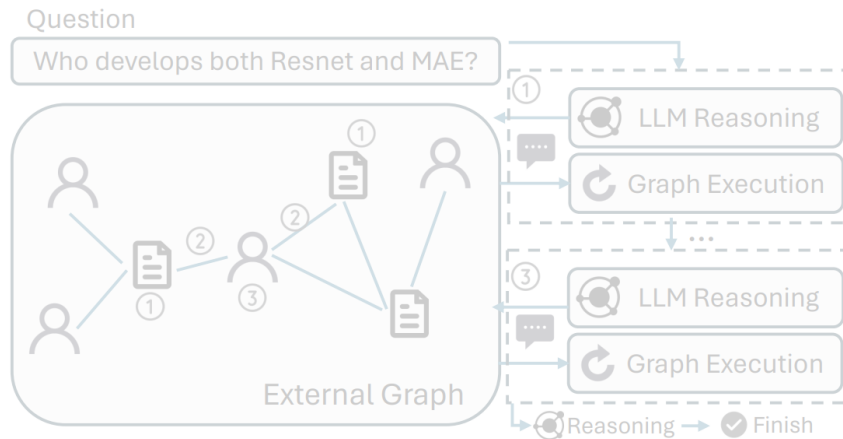
- **Easy questions:** one-step reasoning/interaction
 - “Who are the authors of {paper}?”
- **Medium questions:** multi-step reasoning/interaction
 - “Who is the closest collaborator with {author} in {year}?”
- **Hard questions:** graph information alone is not sufficient to answer the question, but the graph can be useful by providing informative context
 - “Which paper should be recommended to the reader of {paper}?”



Today's Talk: Part II

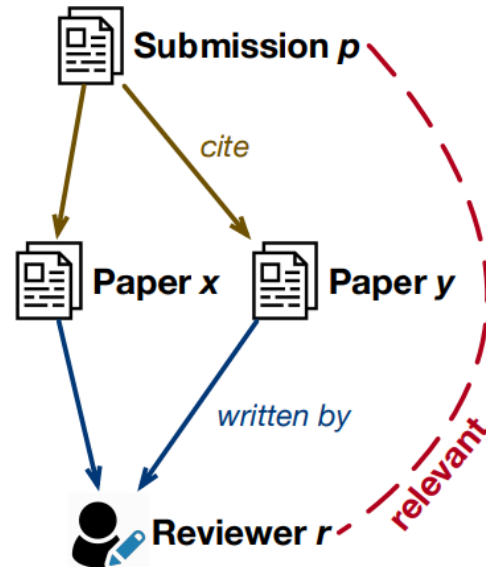
Part I: Structure-Aware LLMs for Scientific Question Answering

Jin et al., ACL 2024



Part II: Structure-Aware LLMs for Scientific Retrieval

Zhang et al., EMNLP 2023
Zhang et al., WWW 2025

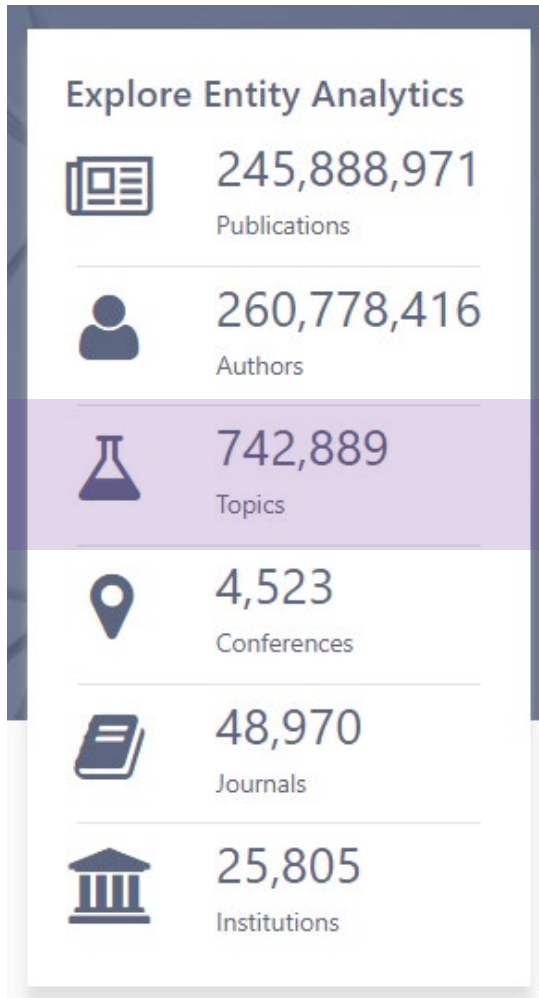


Part III: Consumption of LLM Papers by the Broader Public


Zhang, In submission



Fine-Grained Scientific Paper Classification



- The Microsoft Academic Graph has **740K+** categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain **30K+** categories.
- Each paper can be relevant to **more than one** category (5-15 categories for most papers).

 Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.

- **Relevant categories:** Betacoronavirus, Cardiovascular Diseases, Comorbidity, Coronavirus Infections, Fibrin Fibrinogen Degradation Products, Mortality, Pandemics, Patient Isolation, Pneumonia, ...

Fine-grained classification can be viewed as a retrieval task.

Query: Paper; **Candidates:** Category Names

Link Prediction

DOI: 10.48550/arXiv.2406.10833 • Corpus ID: 270560416

A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery

Yu Zhang, Xiusi Chen, +4 authors Jiawei Han • Published in [Conference on Empirical...](#) 16 June 2024 • Computer Science, Biology

TLDR This paper comprehensively survey over 260 scientific LLMs, discusses their commonalities and differences, as well as summarize pre-training datasets and evaluation tasks for each field and modality, and investigates how LLMs have been deployed to benefit scientific discovery. [Expand](#)

What papers should this survey cite?

BioBERT

Med-PaLM

DeepSeekMath

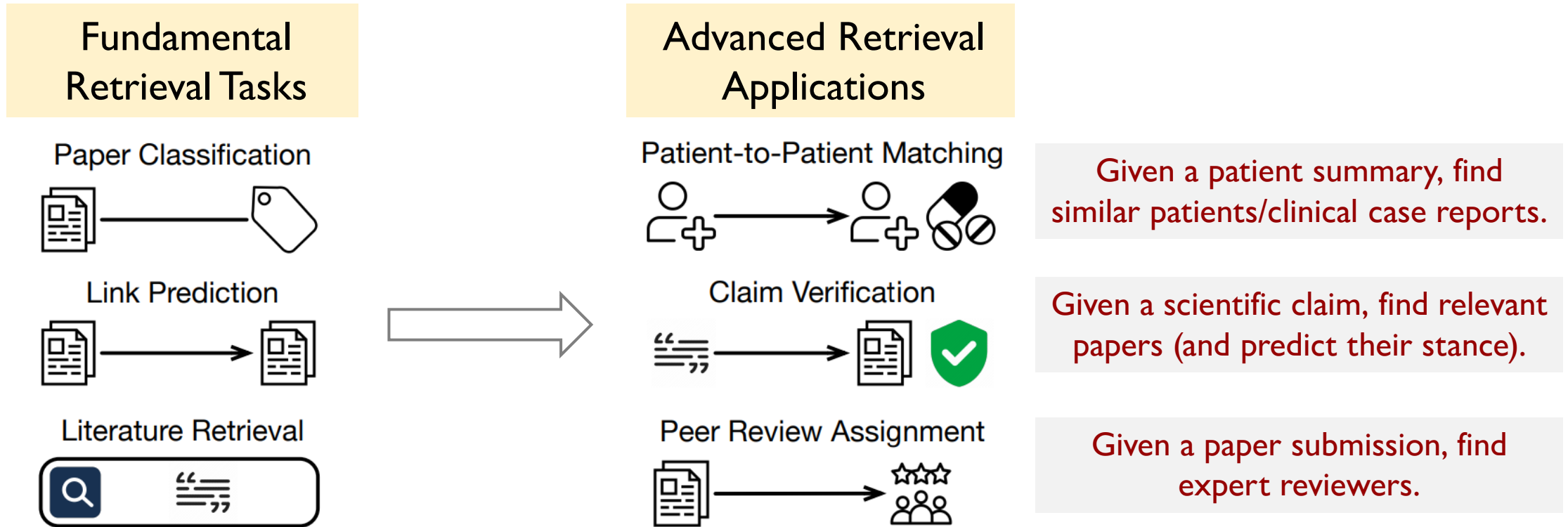
What other papers have these authors written?

What other papers have published in this venue?

Link prediction can be viewed as a retrieval task.

Query: Paper; **Candidates:** Papers

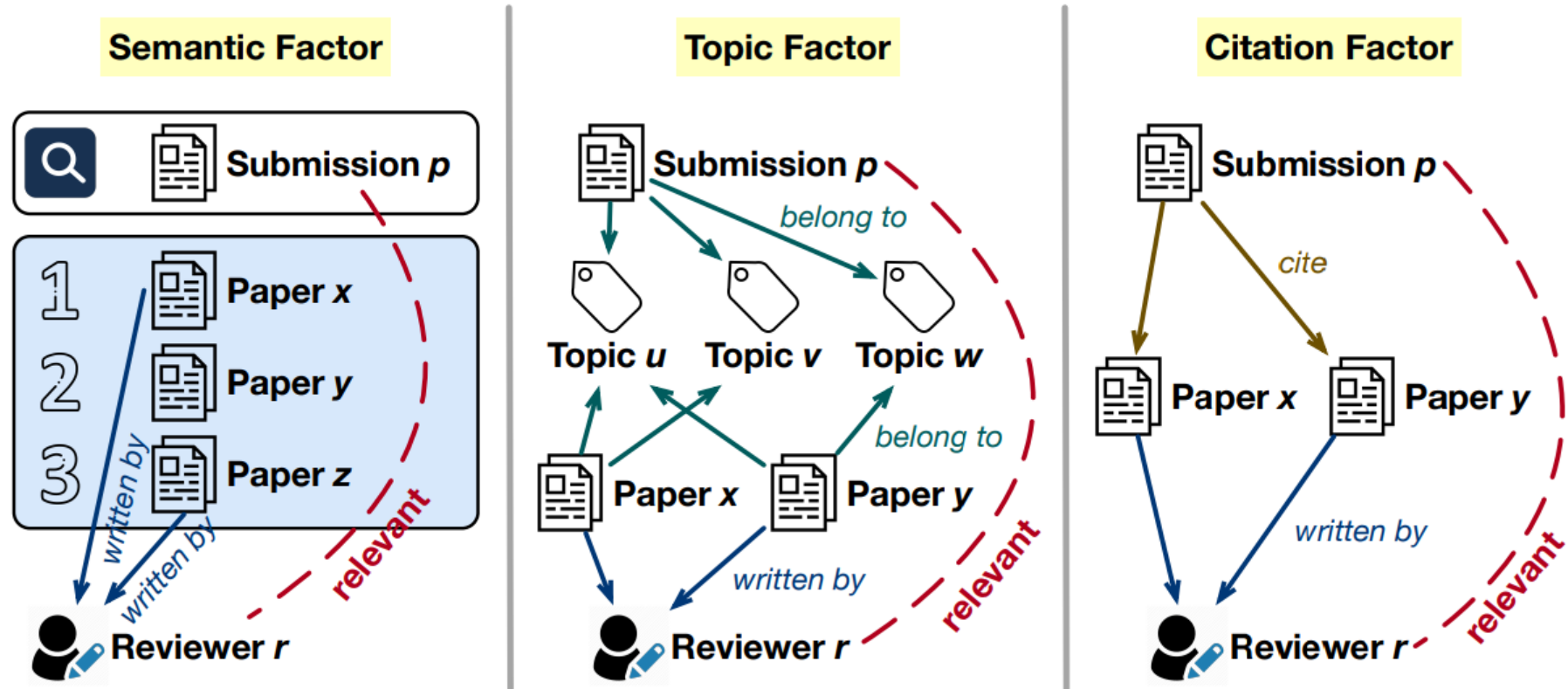
Fundamental Retrieval Tasks vs. Advanced Retrieval Applications



- Why are some tasks more complex?
 - **Multiple** factors should be considered when judging the **relevance**.

Multiple Factors for Judging Relevance

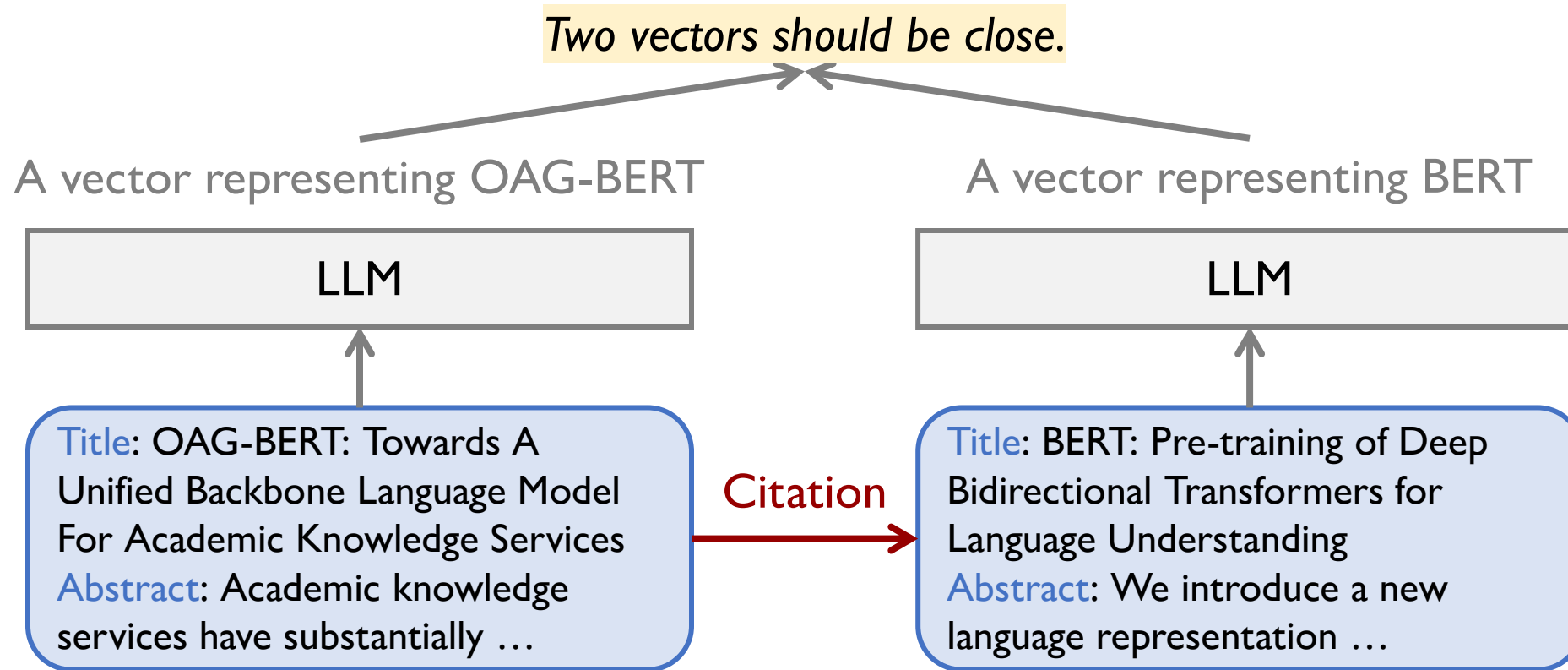
- Example: Paper-Reviewer Matching
 - Why is a pair of (Paper, Reviewer) **relevant**?



- Multiple factors exist in other tasks (e.g., Patient-to-Article Matching) as well.

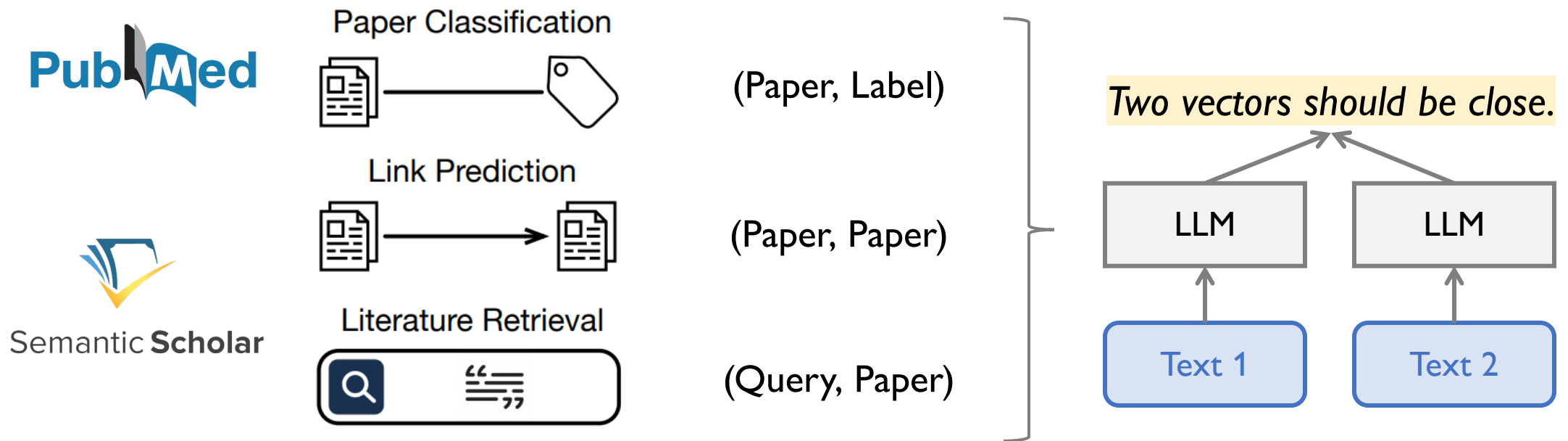
Contrastive Learning for a Fundamental Task

- E.g., Link Prediction
 - **Step 1:** Collect a large number of papers with citation information.
 - **Step 2:** Train an LLM with such citation information.



Contrastive Learning for an Advanced Task – A Naïve Way

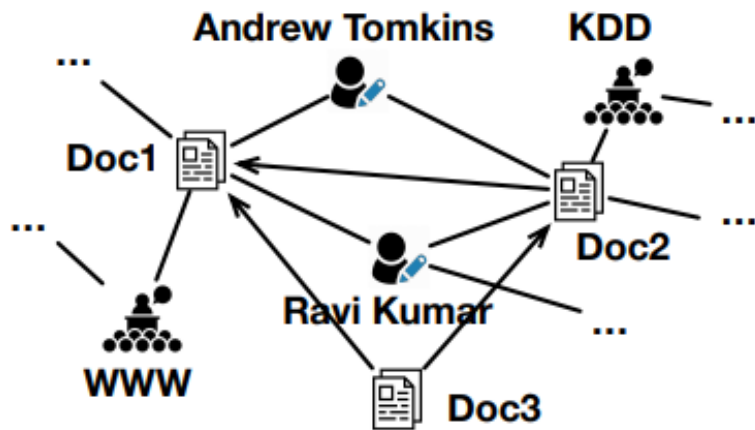
- Each factor (topic, citation, and semantic) relies on one **fundamental** retrieval task.
- Directly combining pre-training data from different tasks to train a model?



- **Task Interference:** The model is confused by different types of “relevance”.

A Toy Example of Task Interference

- Imagine you have two “tasks”.
 - **Task 1:** Given Paper1 and Paper2, predict if **Paper1 should cite Paper2**.
 - **Task 2:** Given Paper1 and Paper2, predict if **Paper1 and Paper2 share the same venue**.
- What if we directly merge the collected relevant (paper, paper) pairs for these two tasks?
 - Is (Doc2, Doc1) relevant?
 - The model does not know **which task you are referring to**, so it will get confused!



Should Doc2 cite Doc1?

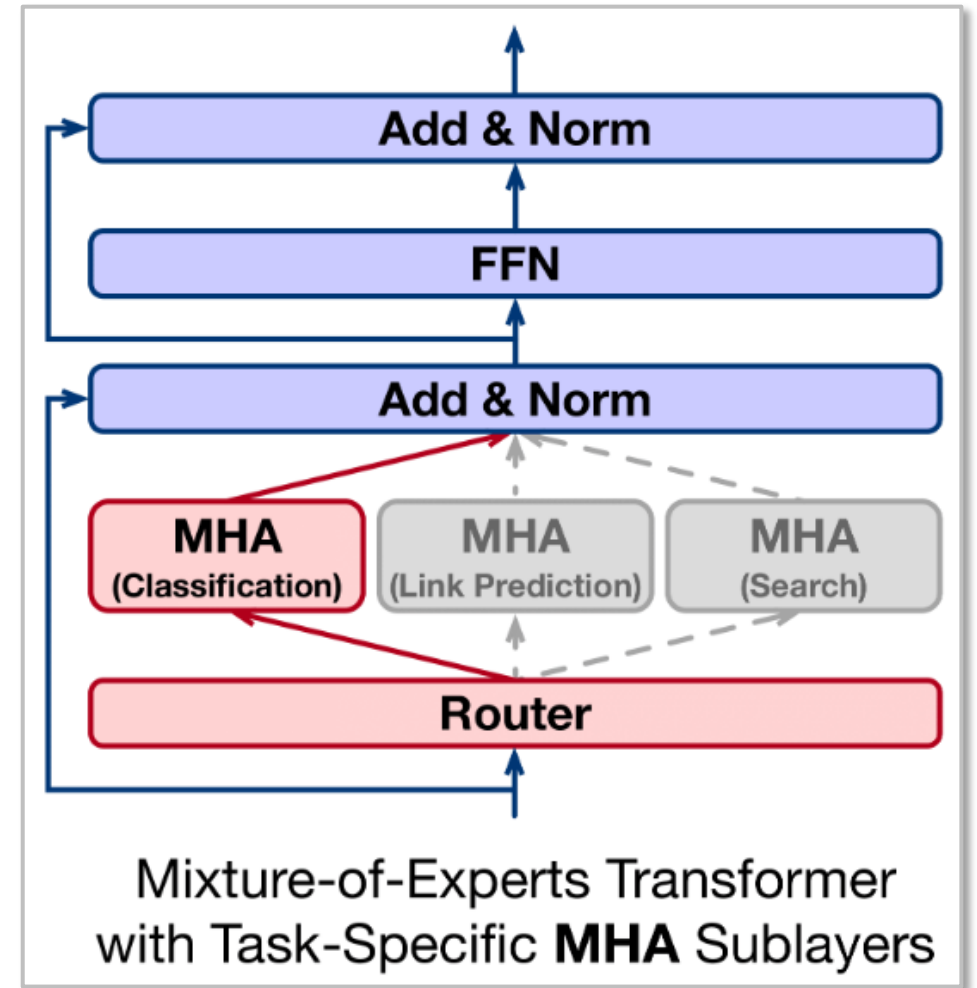


Do Doc2 and Doc1 share the same venue?



Tackling Task Interference: Mixture-of-Experts Transformer

- A typical Transformer layer
 - **1** Multi-Head Attention (MHA) sublayer
 - **1** Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
 - **Multiple** MHA sublayers
 - **1** FFN sublayer
 - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be an “expert” of one task
- The model can learn both **commonalities** and **characteristics** of different tasks.

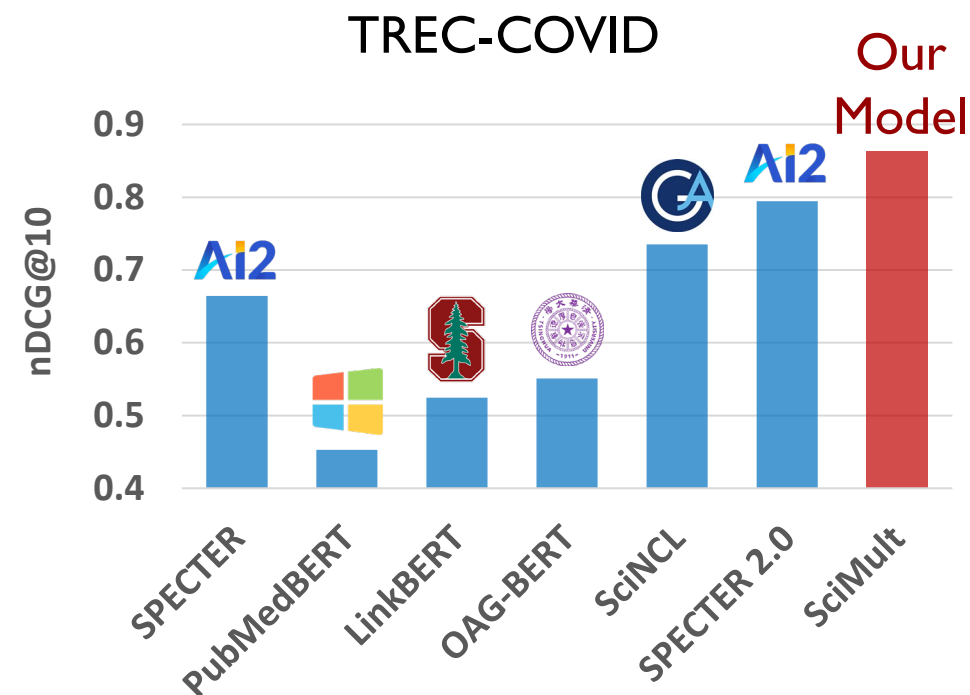


Comparison with Previous Approaches

- New **SOTA** on the PMC-Patients benchmark (**patient-to-article retrieval**)
- Outperforming previous scientific pre-trained language models in classification, link prediction, literature retrieval (**TREC-COVID**), paper recommendation, and claim verification (**SciFact**)

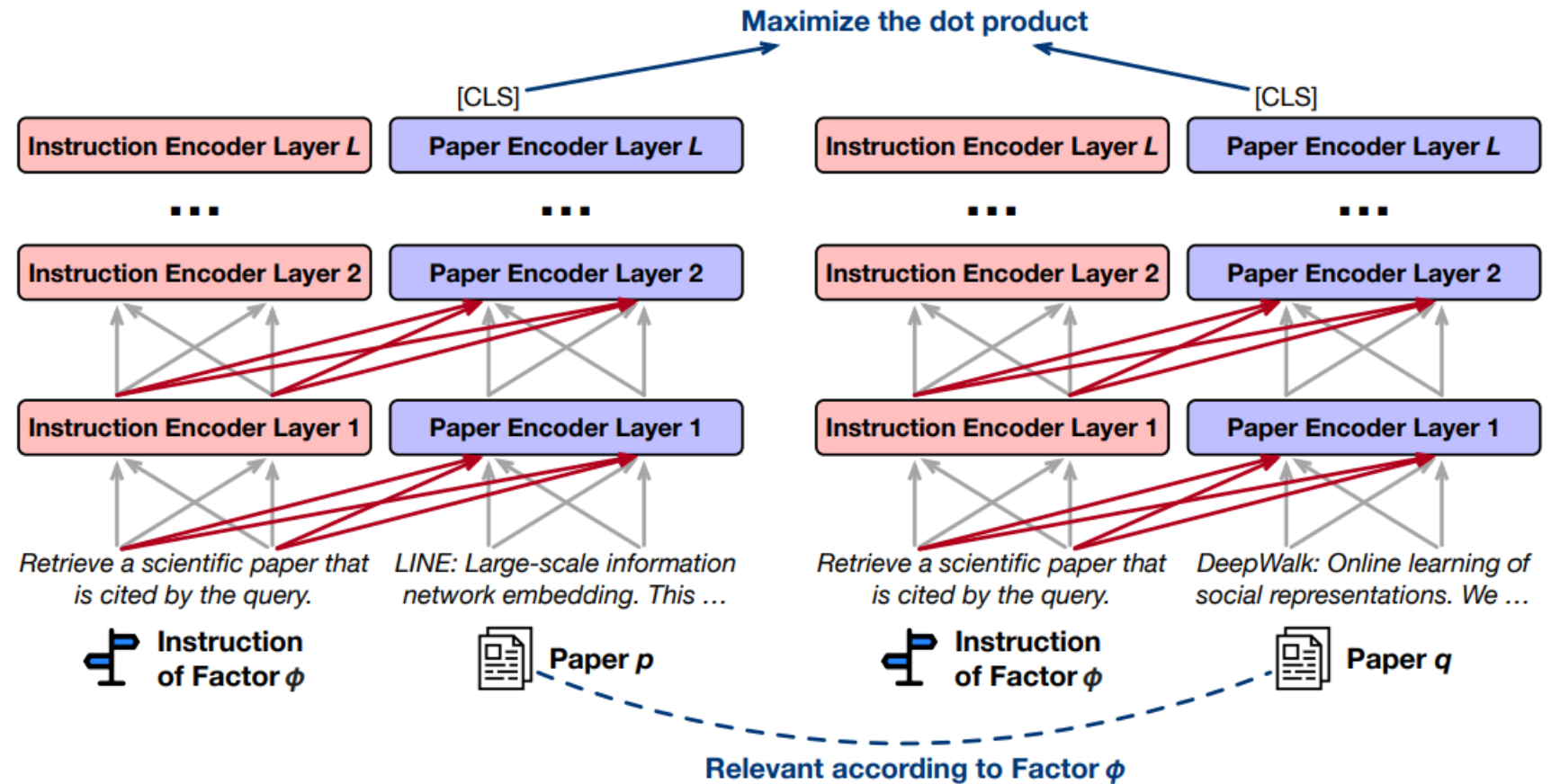
Patient-to-Article Retrieval (PAR) Leaderboard					
	Model	MRR (%)	P@10 (%)	nDCG@10 (%)	R@1k (%)
Our Model 1 June 25, 2023	DPR (SciMult-MHAExpert) <i>UIUC/Microsoft</i> (Zhang et al. 2023)	29.89	9.35	13.79	53.71
2 Apr 5, 2023	RRF <i>Tsinghua University</i> (Zhao et al. 2023)	29.86	8.86	13.36	49.45

<https://pmc-patients.github.io/>



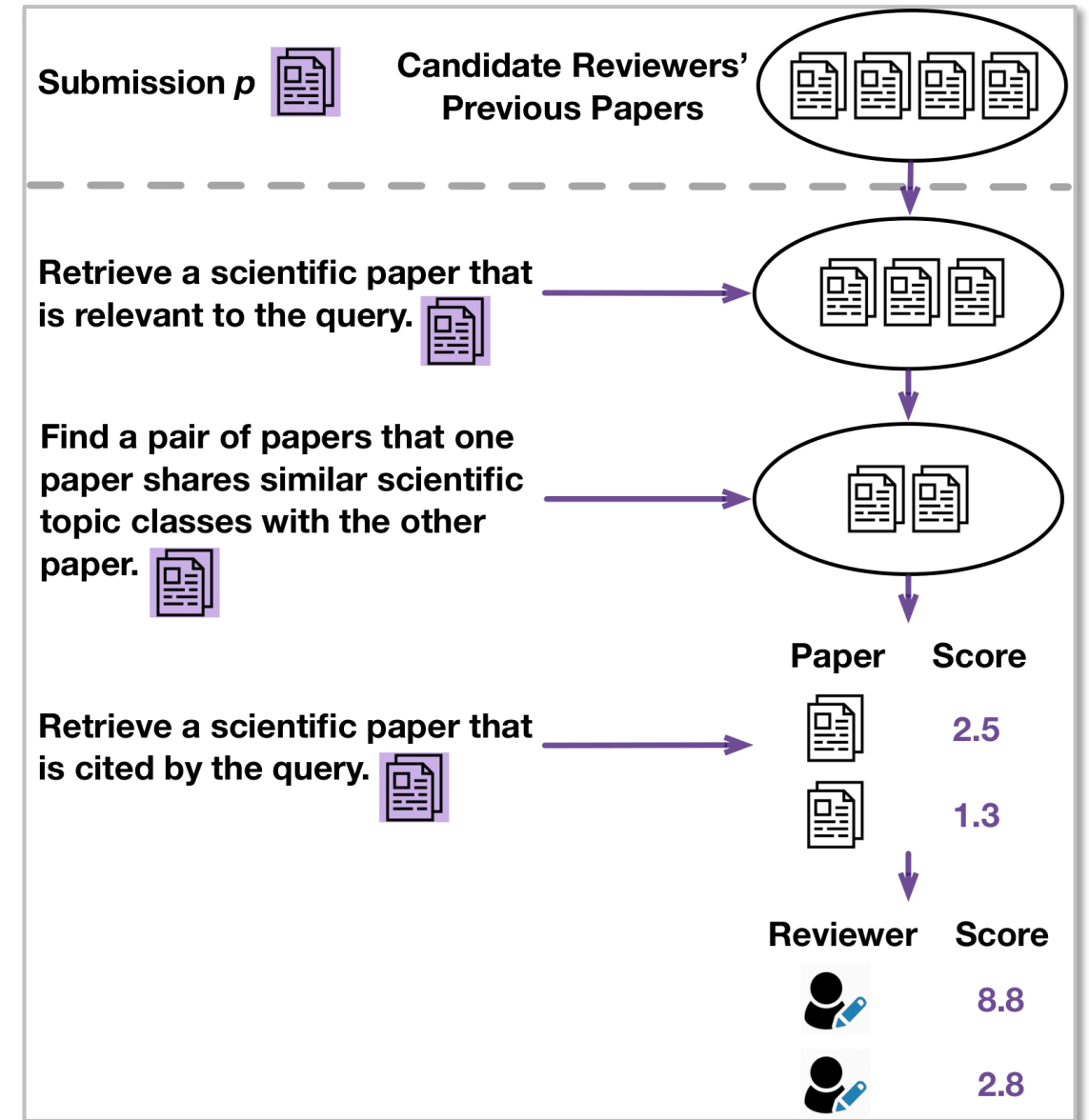
Tackling Task Interference: Instruction Tuning

- Using a **factor-specific instruction** to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.









Chain-of-Factors Reasoning

- Consider semantic, topic, and citation factors in a **step-by-step, coarse-to-fine** manner.
- Step 1:** **Semantic** relevance serves as the coarsest signal to filter totally irrelevant papers.
- Step 2:** Then, we can classify each submission and each relevant paper to a fine-grained **topic** space and check if they share common topics.
- Step 3:** After confirming that a submission and a reviewer's previous paper have common topics, the **citation** link between them will become an even stronger signal, indicating that the two papers may focus on the same task or datasets.



Comparison with Previous Approaches

- Public benchmark datasets
 - Expert C judges whether Reviewer A is qualified to review Paper B.
- Outperforming the **Toronto Paper Matching System** (TPMS, used by Microsoft CMT)

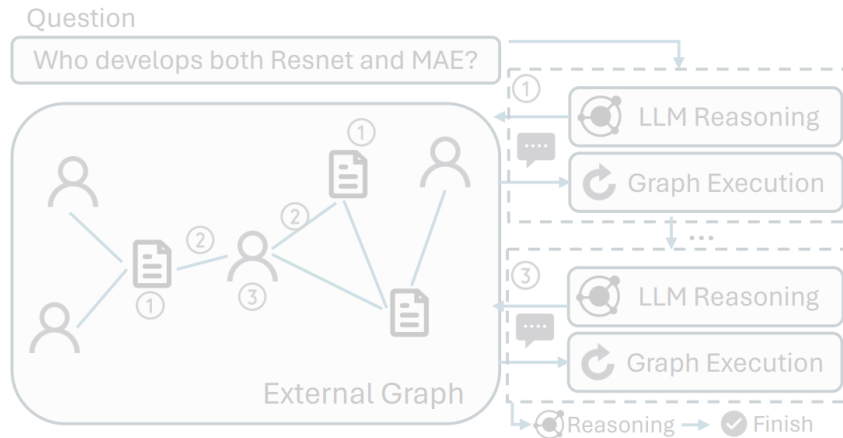
		SciRepEval [44]					SIGIR [19]					KDD				
		Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average
	TPMS [7]	62.06**	53.74**	31.40**	24.86**	43.02**	39.73**	38.36**	17.81**	17.12**	28.26**	17.01**	16.78**	6.78**	7.24**	11.95**
	SciBERT [6]	59.63**	54.39**	28.04**	24.49**	41.64**	34.79**	34.79**	14.79**	15.34**	24.93**	28.51**	27.36**	12.64**	12.70**	20.30**
	SPECTER [9]	65.23**	56.07	32.34**	25.42	44.77**	39.73**	40.00**	16.44**	16.71**	28.22**	34.94**	30.52**	15.17**	13.28	23.48**
	SciNCL [34]	66.92**	55.42**	34.02*	25.33	45.42**	40.55**	39.45**	17.81**	17.40*	28.80**	36.21**	30.86**	15.06**	12.70**	23.71**
	COCO-DR [56]	65.05**	55.14**	31.78**	24.67**	44.16**	40.00**	40.55*	16.71**	17.53	28.70**	35.06**	29.89**	13.68**	12.13**	22.69**
	SPECTER 2.0 CLF [44]	64.49**	55.23**	31.59**	24.49**	43.95**	39.45**	38.63**	16.16**	16.30**	27.64**	34.37**	30.63**	14.48**	12.64**	23.03**
	SPECTER 2.0 PRX [44]	66.36**	55.61**	34.21	25.61	45.45**	40.00**	38.90**	19.18**	16.85**	28.73**	37.13	31.03	15.86**	13.05*	24.27*
Our Model	CoF	68.47	55.89	34.52	25.33	46.05	45.57	41.69	22.47	17.76	31.87	37.63	31.09	16.13	13.08	24.48

: semantic-based method
 : topic-based method
 : citation-based method

Today's Talk: Part III

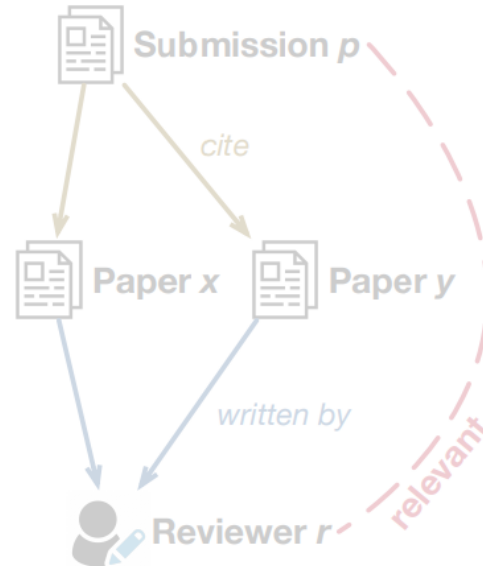
Part I: Structure-Aware LLMs for Scientific Question Answering

Jin et al., ACL 2024



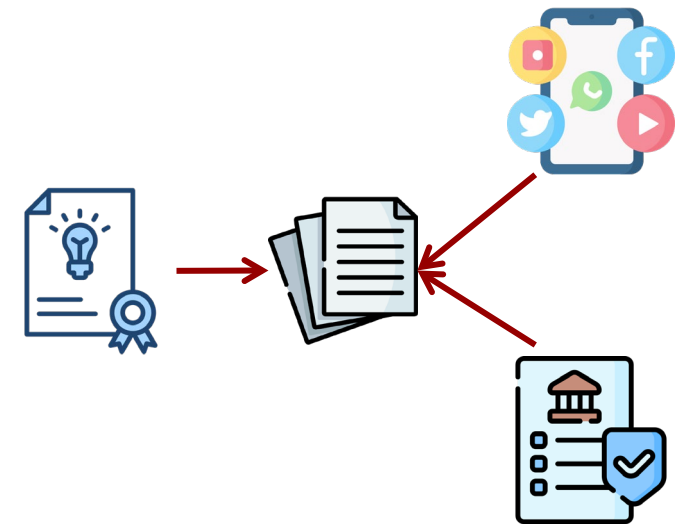
Part II: Structure-Aware LLMs for Scientific Retrieval

Zhang et al., EMNLP 2023
Zhang et al., WWW 2025



Part III: Consumption of LLM Papers by the Broader Public

Zhang, In submission



What papers should we expect at an NLP conference?

https://faculty.washington.edu/ebender/papers/ACL_2024_Presidential_Address.pdf

ACL Is Not an AI Conference

Emily M. Bender
Bangkok, Thailand
August 14, 2024

ACL 2024 Presidential Address

<https://bit.ly/EMB-ACL24>

What papers should we expect at an NLP conference?

ACL is not an AI Conference (?)

Yoav Goldberg, August 2024

In her "Presidential Address" at the ACL 2024, Emily Bender gave a talk called "ACL is not an AI Conference". For those who did not attend (or were not paying close attention), you can find the slides in the following link:

https://faculty.washington.edu/ebender/papers/ACL_2024_Presidential_Address.pdf

Somewhat surprisingly, I found myself agreeing with some core aspects of her argument. Perhaps less surprisingly, there is also a substantial part which I strongly *disagree* with. This text is a response to this address, and, beyond just responding, may also

Imagine being a CS/AI PhD student attending your first ACL, excited to present your research, only to be told by officials that ACL isn't an AI conference—you're in the wrong place. How would you feel? It's disheartening to us who've seen ACL as central to our AI/NLP journey.

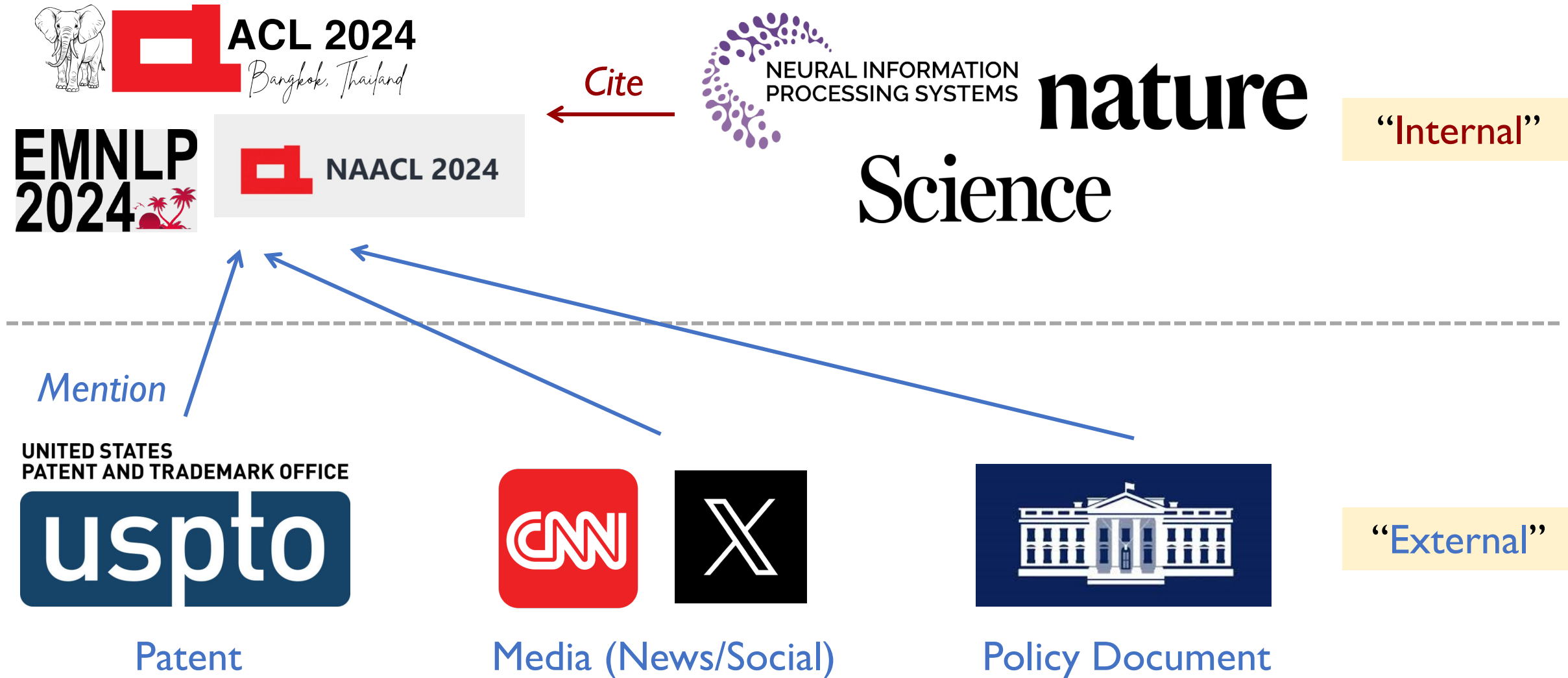
12:28 AM · Aug 15, 2024 · **44.1K** Views

I was having an identity crisis when I learned ACL isn't AI. If ACL isn't AI but NLP is, should I still submit my NLP paper to ACL? Or worse... have I not been doing NLP at all?? Turns out I'm actually a physicist! BRB, off to claim my Nobel Prize for all my physics research!



5:45 PM · Oct 8, 2024 · **15.1K** Views

How does the public perceive NLP conferences?



Data and Metric

NLP Papers:

ACL Anthology
ACL, EMNLP, NAACL
1979-2024



Internal Citation:

OpenAlex



OpenAlex

Patent-to-Paper:

Reliance on Science



Media-to-Paper:

Altmetric



Altmetric

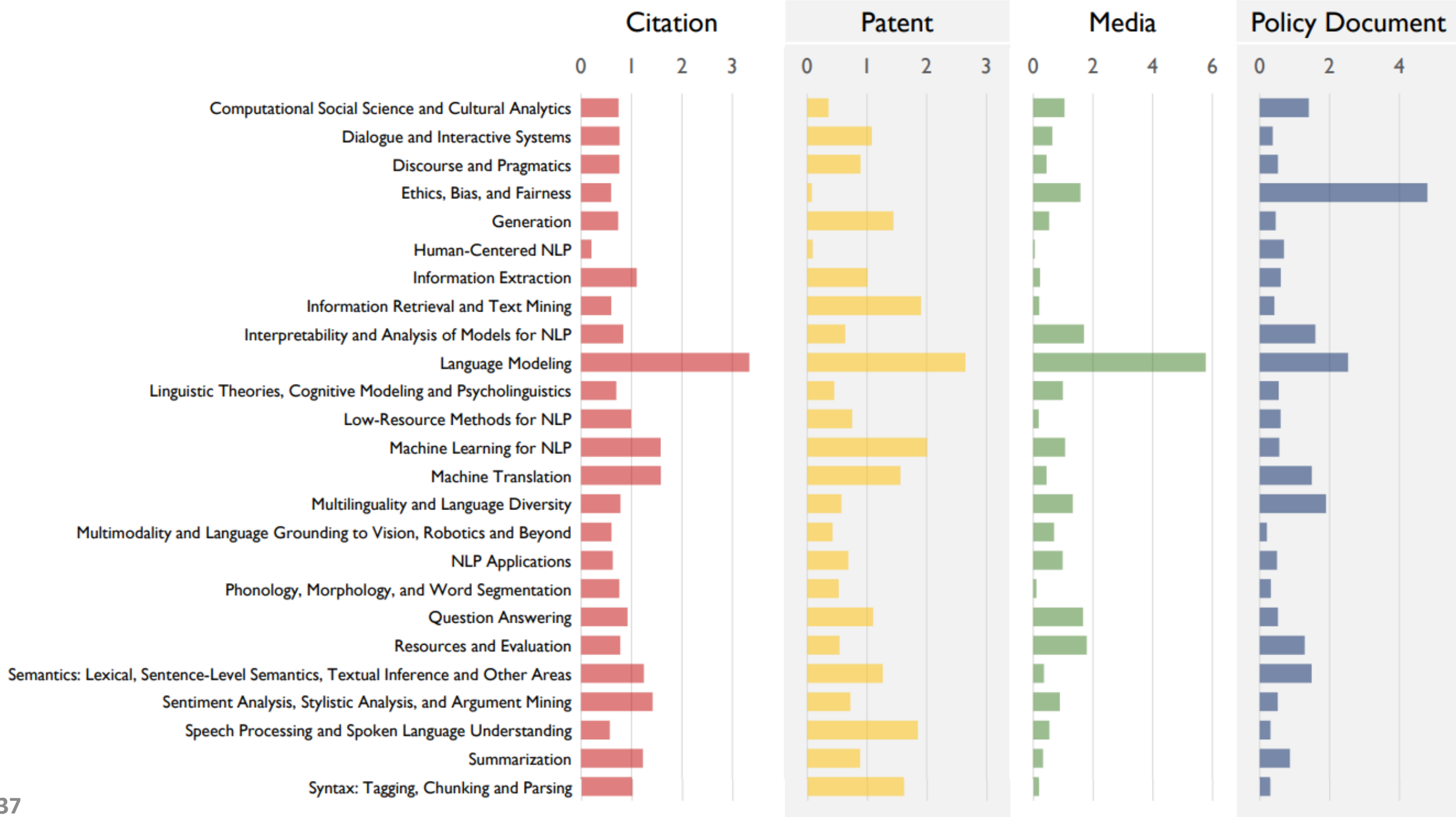
PolicyDoc-to-Paper:

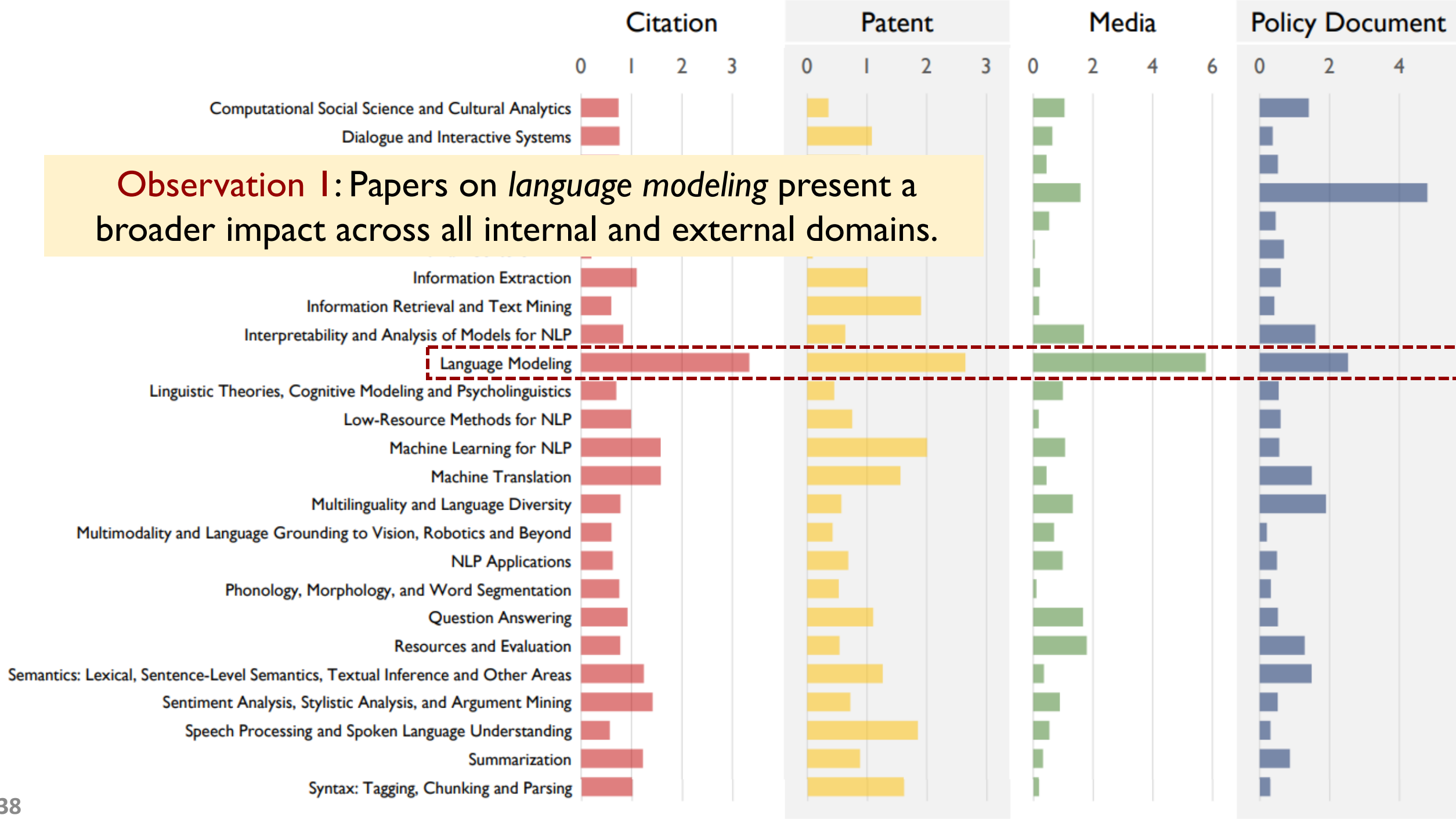
Overton

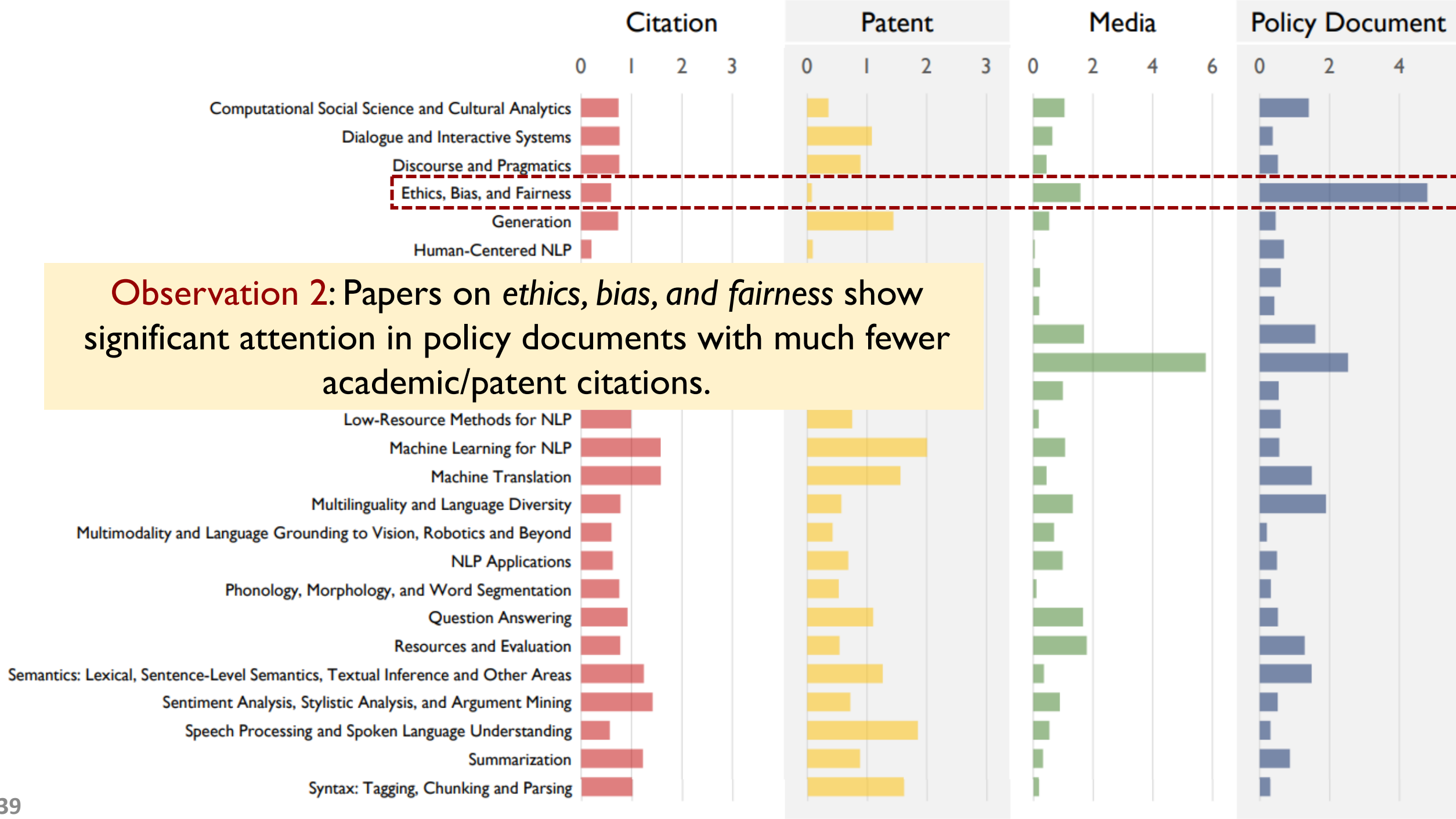


- How to quantify the impact of an NLP topic (e.g., “*Language Modeling*” and “*Ethics, Bias, and Fairness*” within a domain (e.g., “*Citation*”, “*Patent*”, “*Media*”, and “*PolicyDocument*”)?
 - Assume there are 1,000 NLP papers, collectively cited 1,000 times in media posts.
 - Among these papers, 100 are about “*Language Modeling*” and are collectively cited 200 times in media posts.

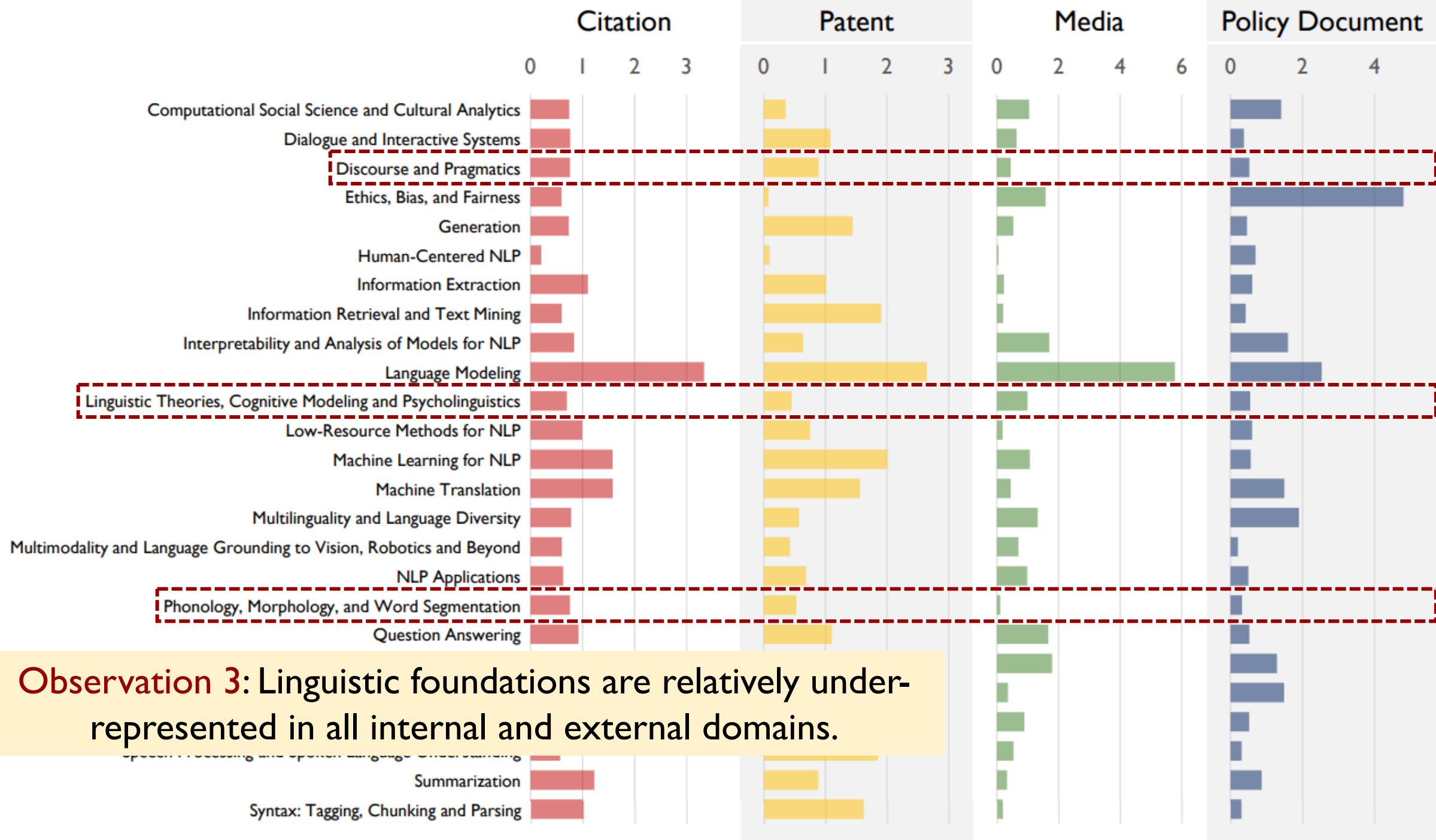
$$\text{Impact Index}(\text{“Language Modeling”} \rightarrow \text{media}) = \frac{200 \text{ total citations} / 100 \text{ papers}}{1,000 \text{ total citations} / 1,000 \text{ papers}} = 2$$







Observation 2: Papers on *ethics, bias, and fairness* show significant attention in policy documents with much fewer academic/patent citations.



Correlation between Internal and External Impacts

	Patent	Media	Policy Document
Corr(Citation, ·)	0.654	0.725	0.247 (0.599 if excluding “ <i>Ethics, Bias, and Fairness</i> ”)

Good alignment between what **the public** from external domains consume and what is regarded as impactful by **researchers themselves**.

Complementarity of Different External Impacts

- Consider the task of finding the top-1% highly cited papers.
 - Random guess? Hit Rate = 1%
 - Papers cited at least once in **patents**?
 - Papers cited at least once in **media posts**?
 - Papers cited at least once in **policy documents**?
 - Papers cited at least once in BOTH **patents** AND **media posts**?
 - ...

External Domain(s) Considered	Hit Rate
\emptyset	1.00%
{Patent}	5.46%
{Media}	9.26%
{PolicyDocument}	18.29%
{Patent, Media}	26.72%
{Patent, PolicyDocument}	34.02%
{Media, PolicyDocument}	45.71%
{Patent, Media, PolicyDocument}	71.88%

Different external domains may favor different types of NLP papers. Papers attracting attention from **multiple external domains** are more likely to be internally impactful than those attracting one domain only.

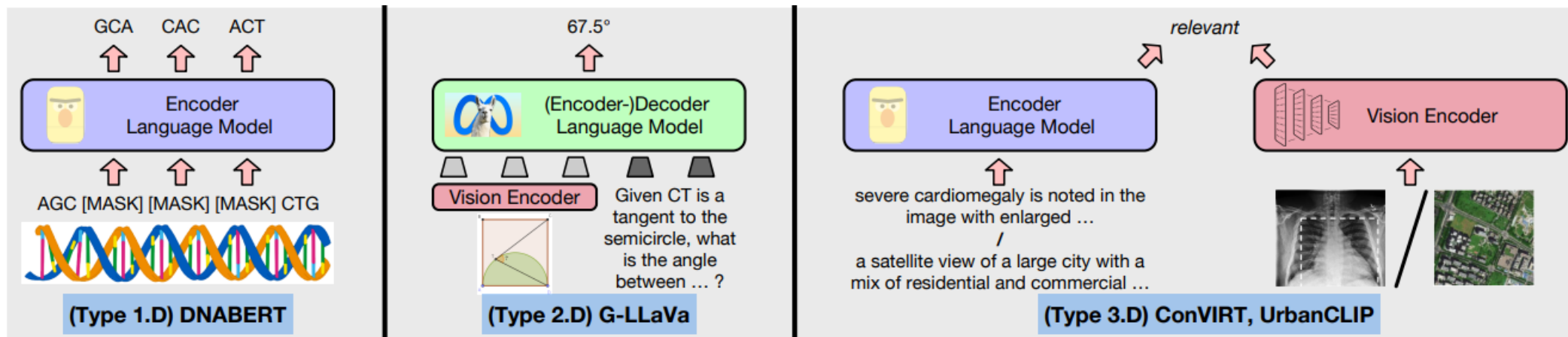
Our Recent Survey Paper

🔗 Awesome Scientific Language Models

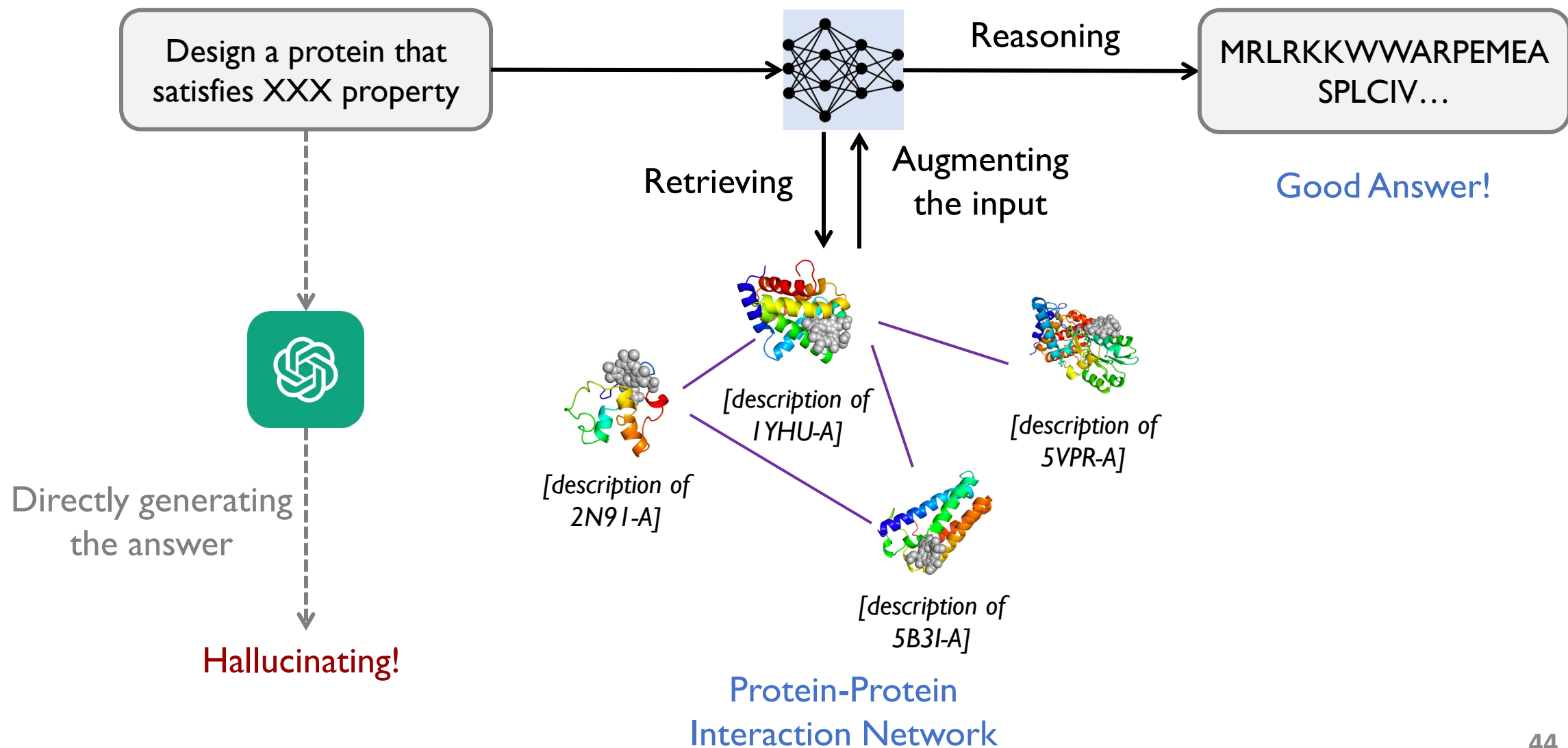
👁 awesome ⭐ Stars 551

PaperNumber 296 License MIT PRs Welcome

A curated list of pre-trained language models in scientific domains (e.g., **mathematics, physics, chemistry, materials science, biology, medicine, geoscience**), covering different model sizes (from **100M** to **100B** parameters) and modalities (e.g., **language, graph, vision, table, molecule, protein, genome, climate time series**).



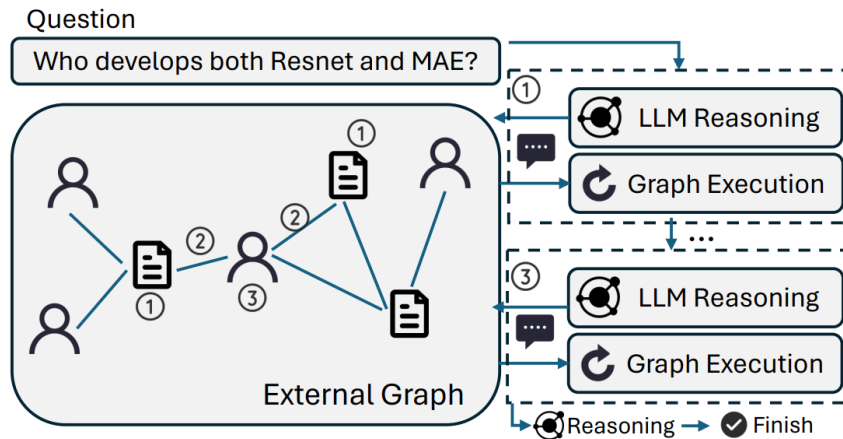
Ongoing Work: Expanding to Scientific Data in Other Modalities



Today's Talk

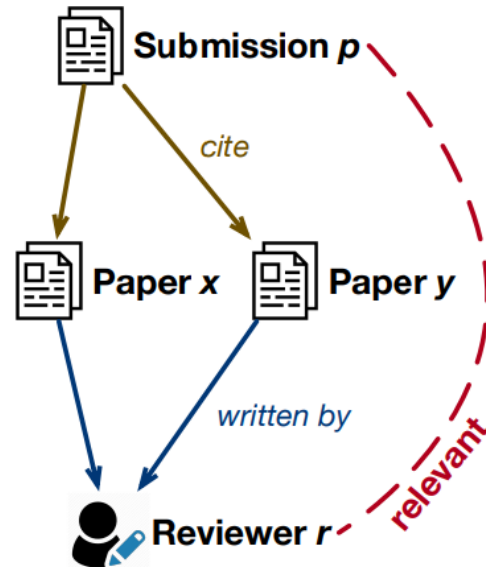
Part I: Structure-Aware LLMs for Scientific Question Answering

Jin et al., ACL 2024



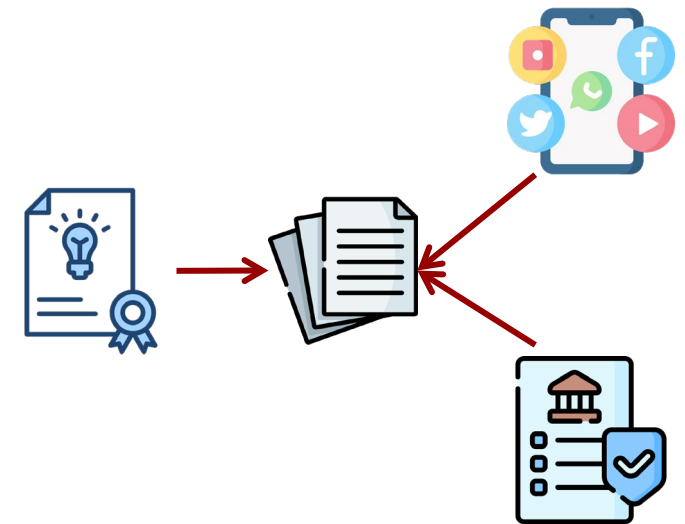
Part II: Structure-Aware LLMs for Scientific Retrieval

Zhang et al., EMNLP 2023
Zhang et al., WWW 2025



Part III: Consumption of LLM Papers by the Broader Public

Zhang, In submission





Thank you! Questions?

Yu Zhang

yuzhang@tamu.edu