SDM 2025 Tutorial: Integrating Textual and Graph Data: Advancing Knowledge Discovery with Semantic and Structural Insights

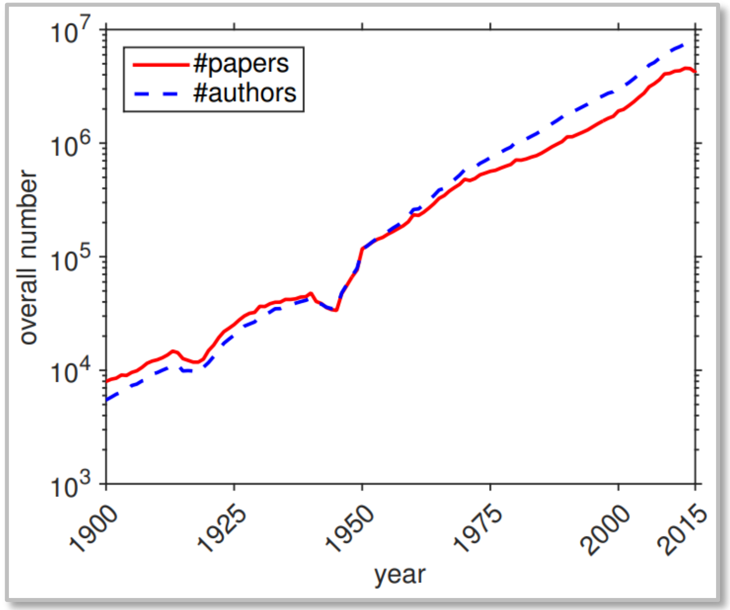
Part III: Domain-Specific Applications

Bowen Jin, Yu Zhang, Yunyi Zhang, Jiawei Han

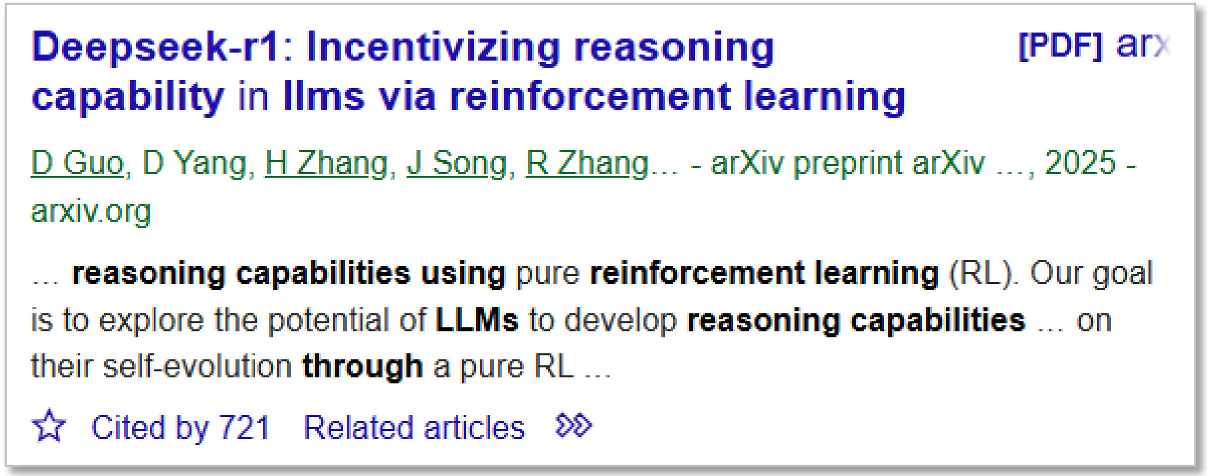
Apr 30, 2025

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 721 times on April 20, 2025.



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

D Guo, D Yang, H Zhang, J Song, R Zhang… - arXiv preprint arXiv …, 2025 - arxiv.org

… **reasoning capabilities using** pure **reinforcement learning** (RL). Our goal is to explore the potential of **LLMs** to develop **reasoning capabilities** … on their self-evolution **through** a pure RL …

---

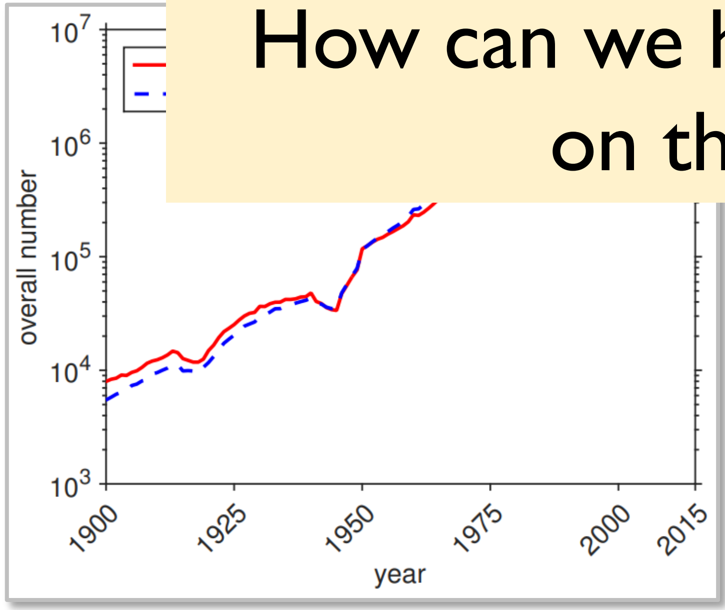
[1] *A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations*. KDD 2017.
[2] *Microsoft Academic Graph: When Experts are Not Enough*. Quantitative Science Studies 2020.
[3] https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent

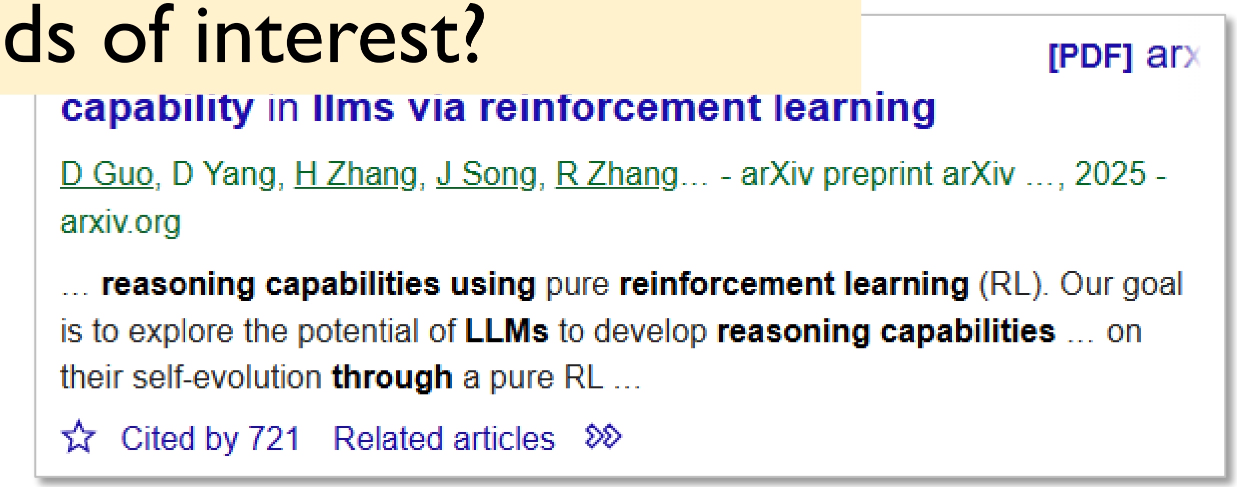
Explosion of Scientific Text Data

- 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]
  - rXiv on January
    April 20, 2025.

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




[PDF] arx

capability in llms via reinforcement learning

D Guo, D Yang, H Zhang, J Song, R Zhang… - arXiv preprint arXiv …, 2025 - arxiv.org

… **reasoning capabilities using** pure **reinforcement learning** (RL). Our goal is to explore the potential of **LLMs** to develop **reasoning capabilities** … on their self-evolution **through** a pure RL …

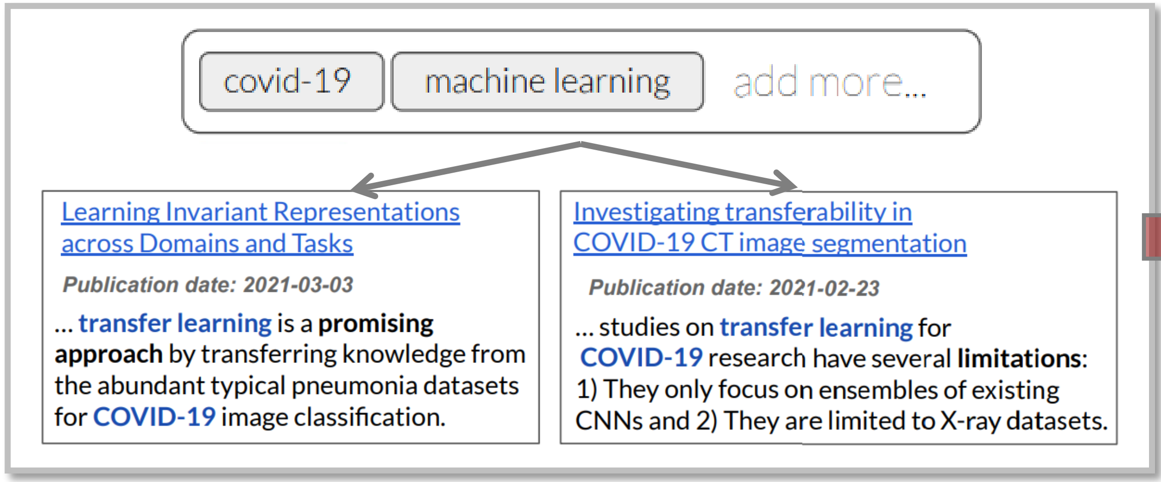
☆ Cited by 721 Related articles

---

[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 scientific text mining assist research?

Retrieving and Analyzing Relevant Literature



Learning Invariant Representations across Domains and Tasks

Publication date: 2021-03-03

... **transfer learning** is a **promising approach** by transferring knowledge from the abundant typical pneumonia datasets for **COVID-19** image classification.

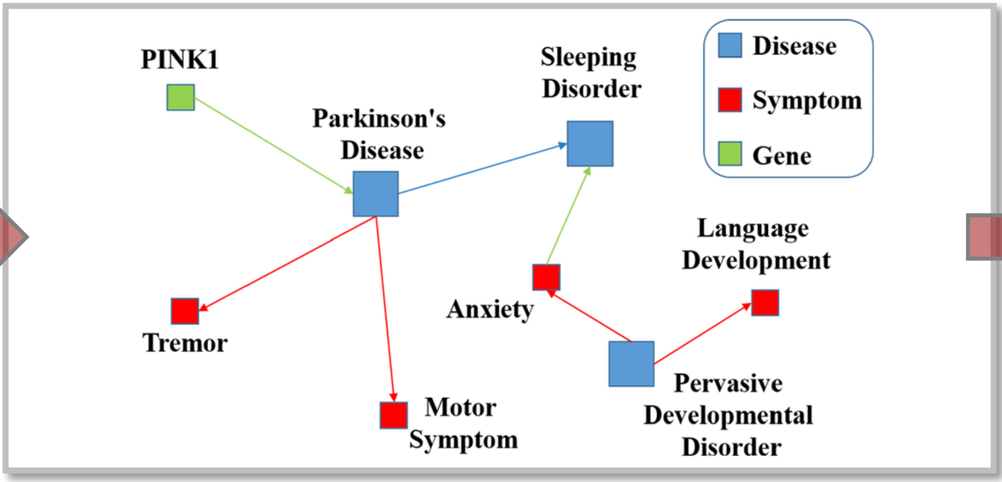
Investigating transferability in COVID-19 CT image segmentation

Publication date: 2021-02-23

... studies on **transfer learning** for **COVID-19** research have several **limitations**: 1) They only focus on ensembles of existing CNNs and 2) They are limited to X-ray datasets.

- 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*".

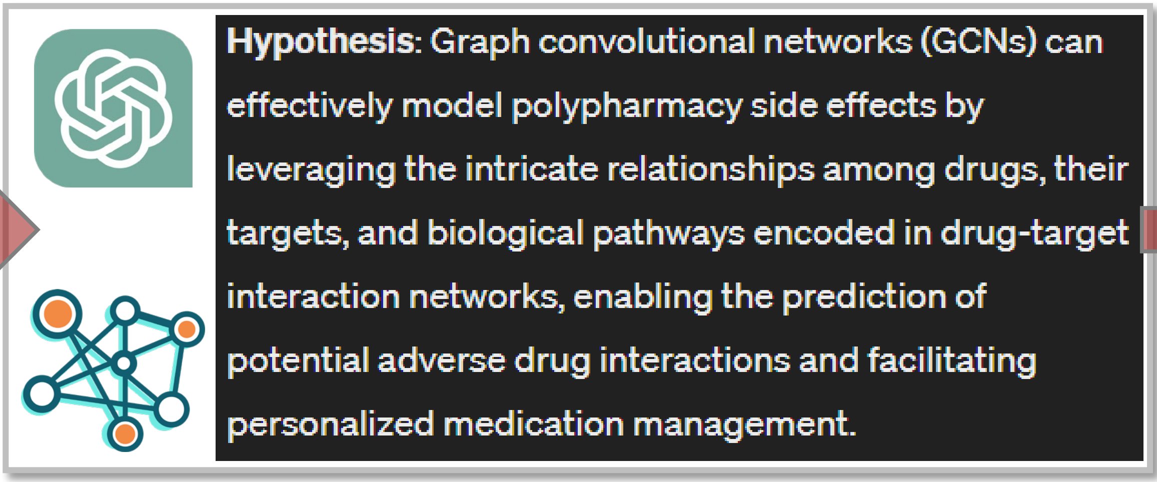
Uncovering Knowledge Structures/Graphs



- Example tasks:
  - Find protein entities relevant to "*Parkinson's disease*" from relevant literature.
  - Predict the relationship between "*Tremor*" and "*Sleeping Disorder*".

How can scientific text mining assist research?

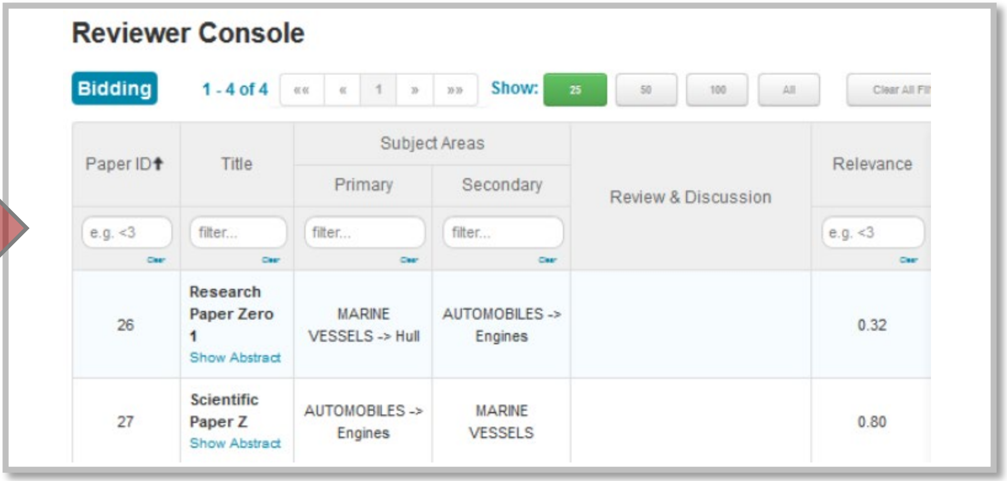
Generating Hypotheses and Suggesting Directions



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



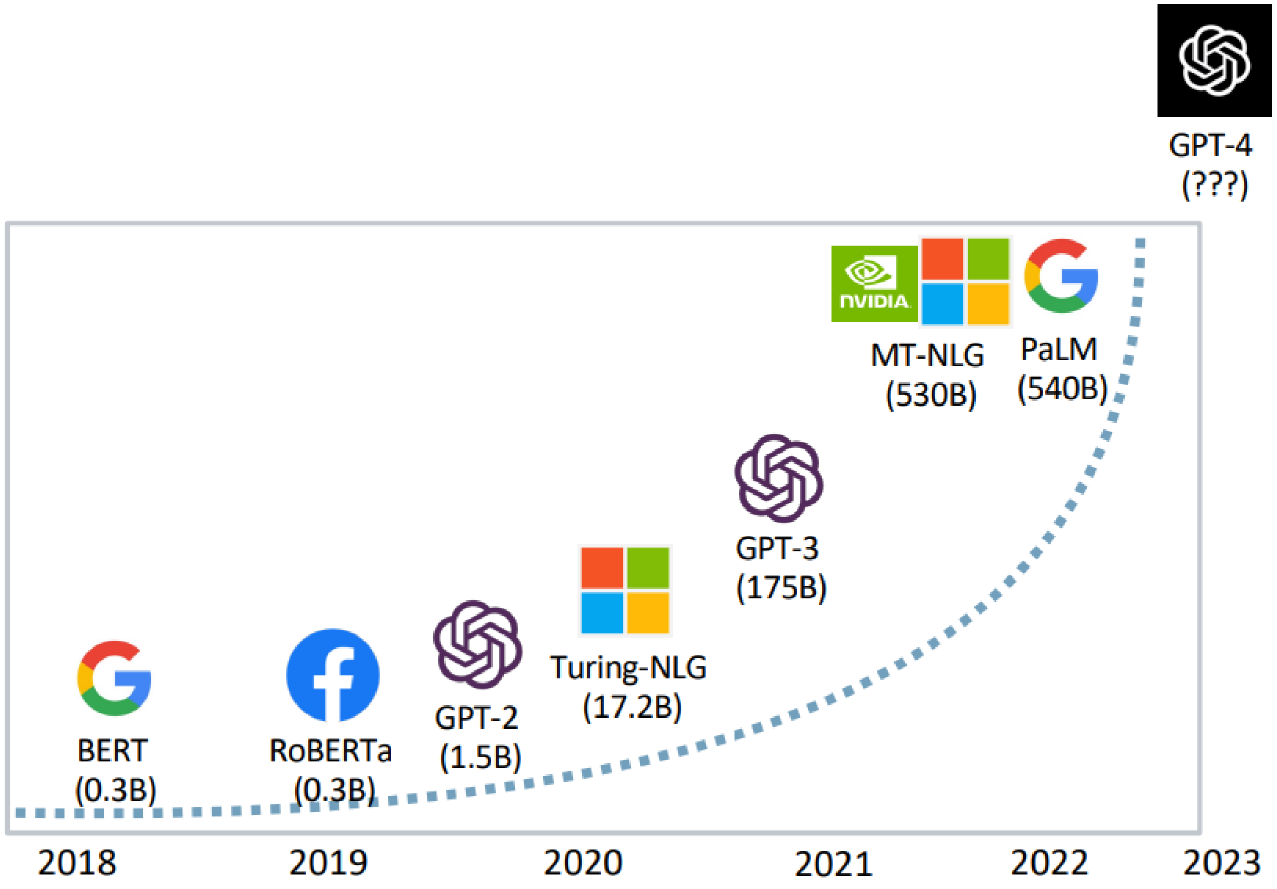
Reviewer Console

Paper ID↑	Title	Subject Areas		Review & Discussion	Relevance
		Primary	Secondary		
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.

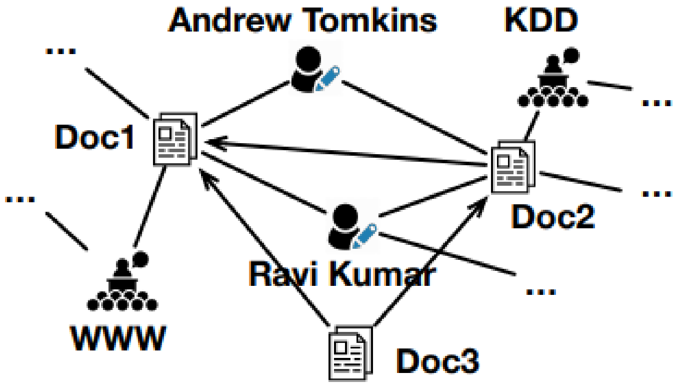
Large Language Models (LLMs) for Text Mining

- 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 = {15, 11} (*Math*)
  - …

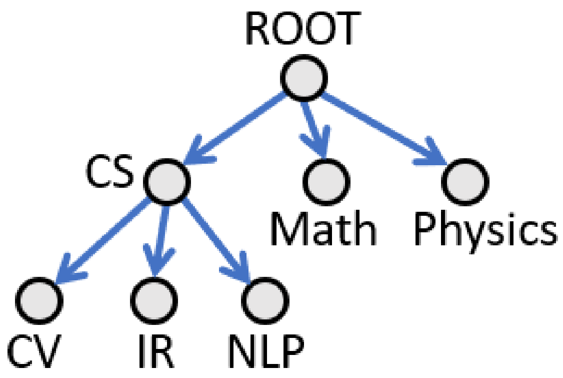


Are LLMs aware of structured information?

Structured Information Associated with Scientific Text



Metadata/Network



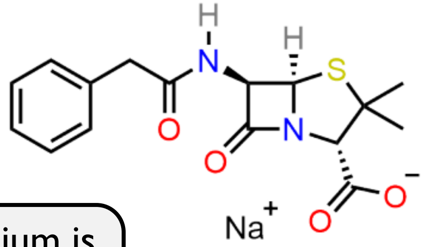
Hierarchy/Taxonomy

WIKIDATA

ctd

DBpedia

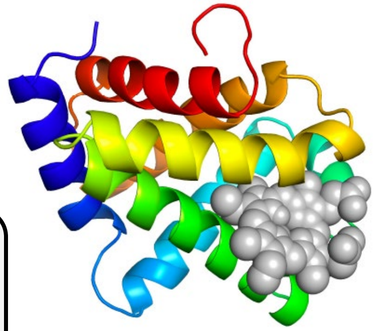
Knowledge Graph



Penicillin G Sodium is the sodium salt form of benzylpenicillin.

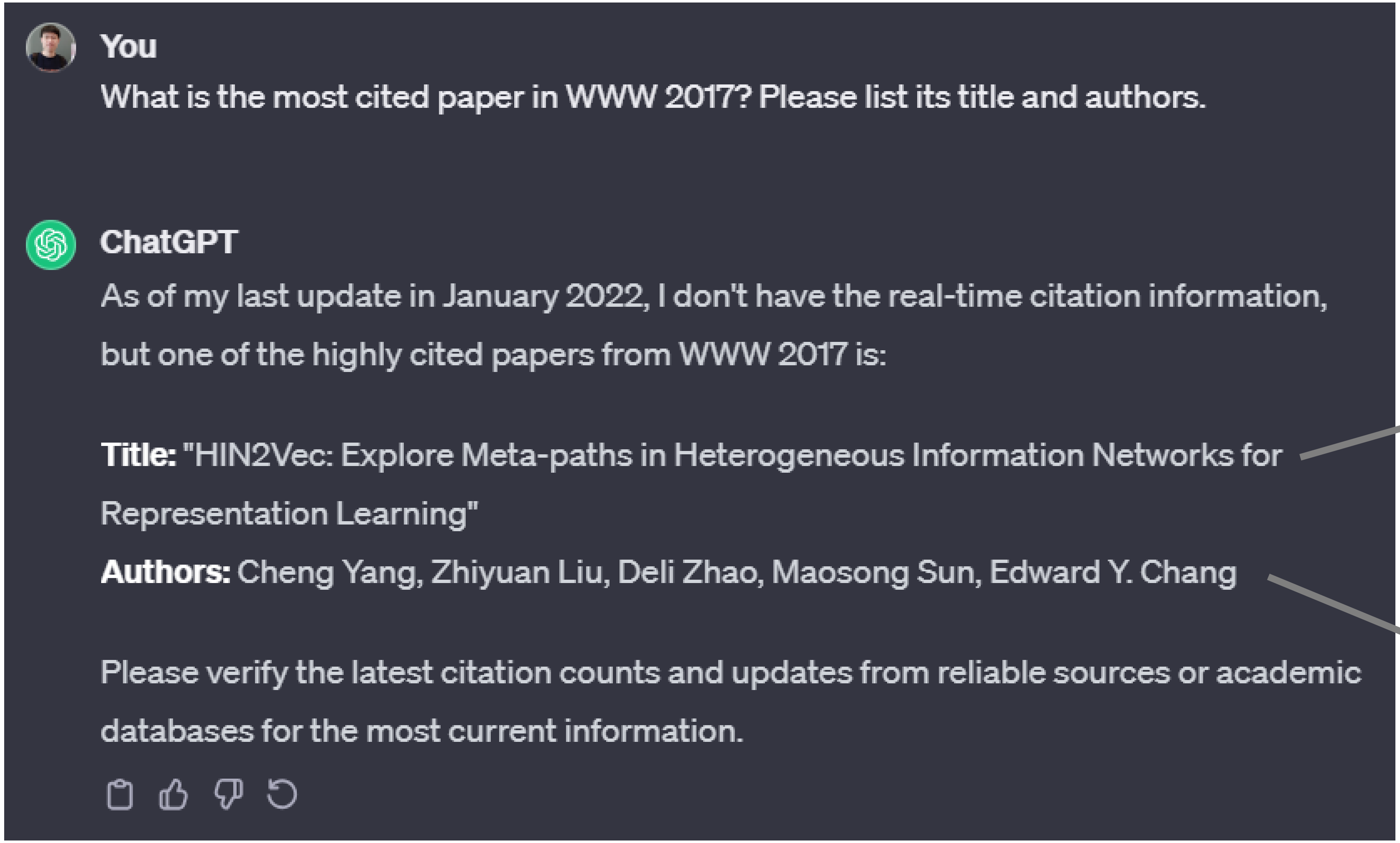
Text-Paired Molecule

1YHU: Crystal structure of Riftia pachyptila C1 …



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

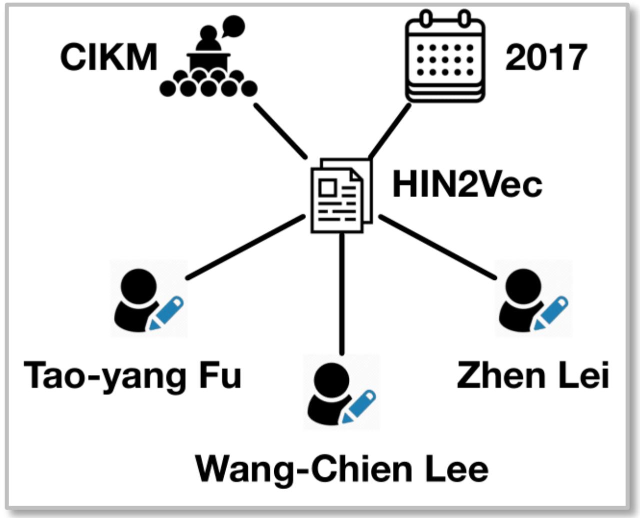
As of my last update in January 2022, I don't have the real-time citation information, but one of the highly cited papers from WWW 2017 is:

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.

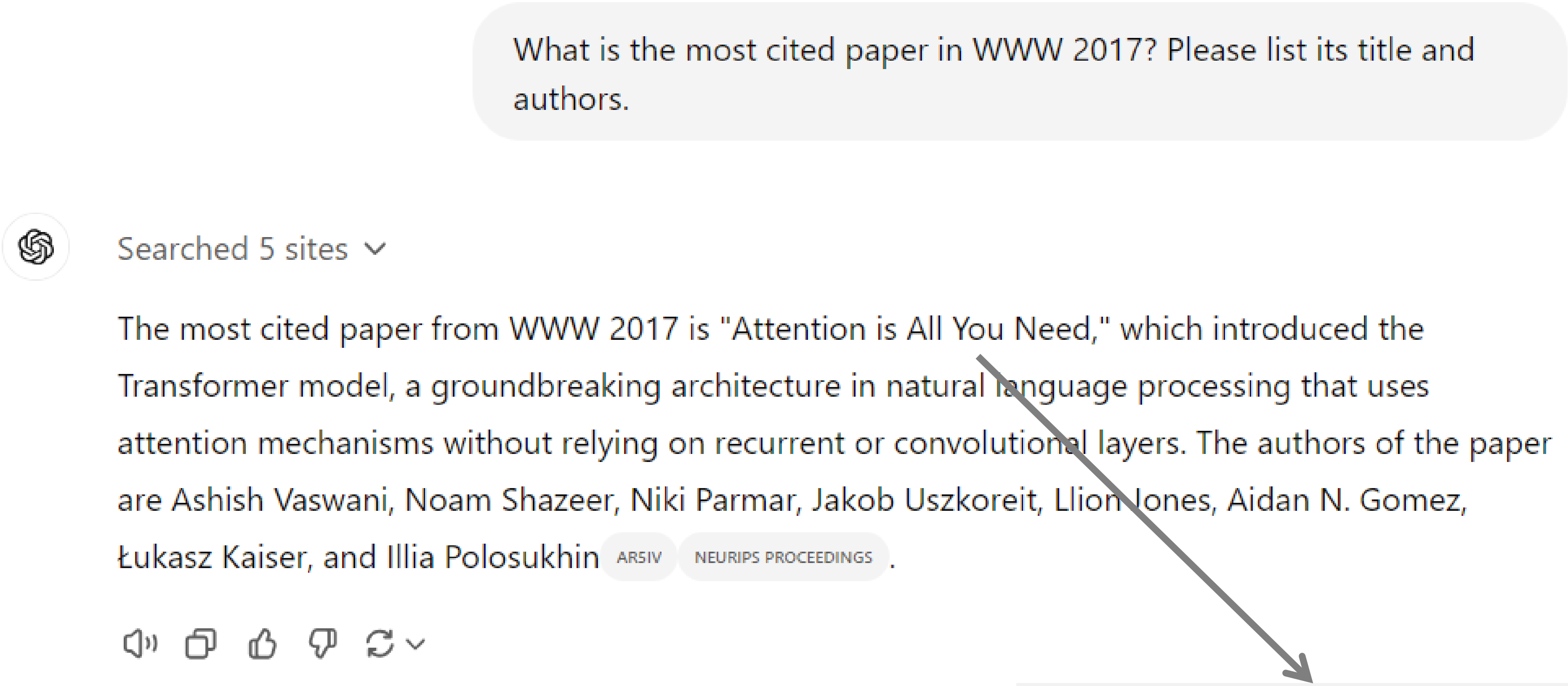
ChatGPT 3.5, queried on January 23, 2024



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.

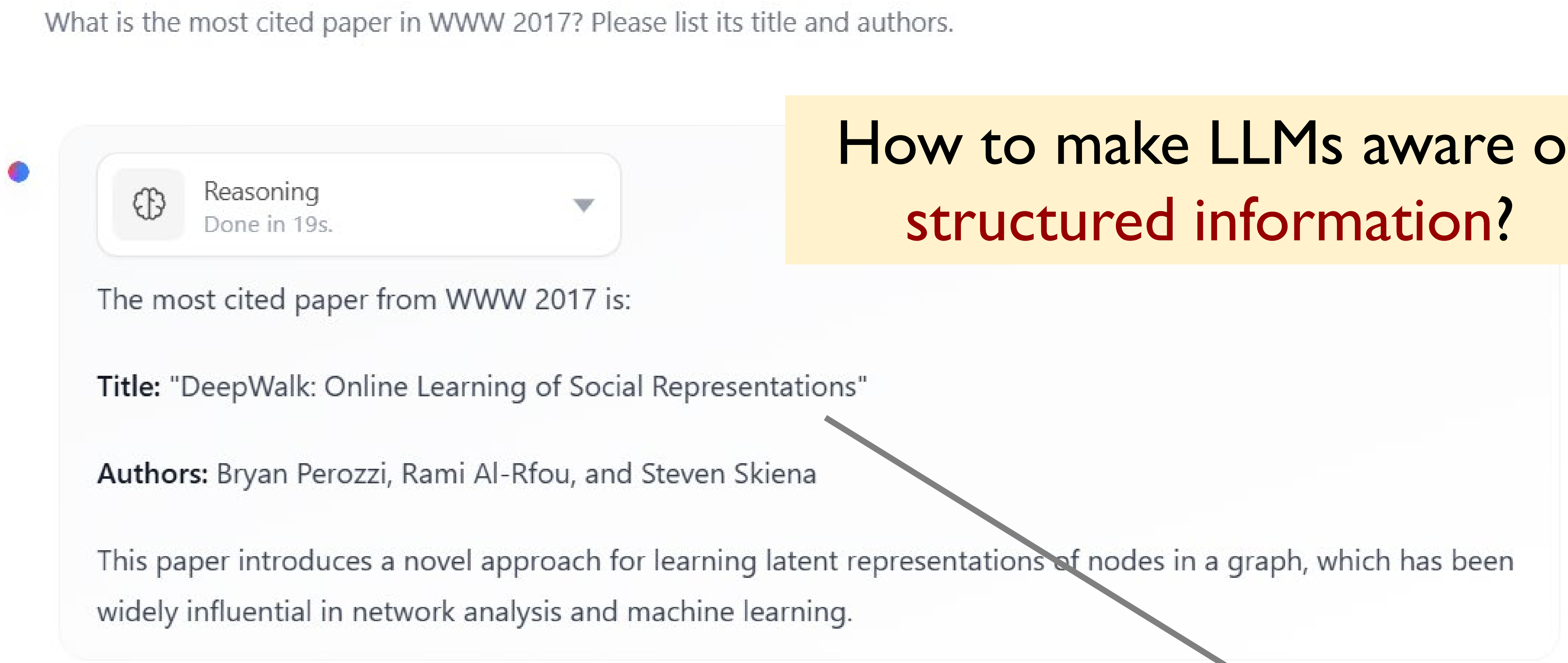
Searched 5 sites

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 ARSIV 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 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

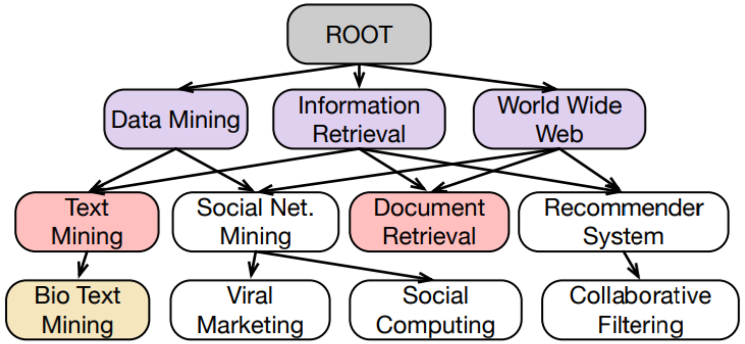
DeepWalk was published in KDD 2014.

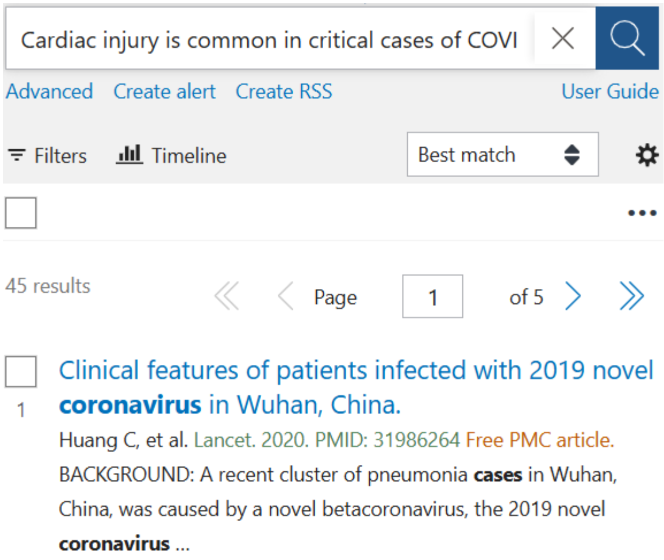
Overview of Part III

Structure-Aware LLMs for Scientific Paper Classification

Structure-Aware LLMs for Scientific Literature Search

Structure-Aware LLMs for Paper-Reviewer Matching



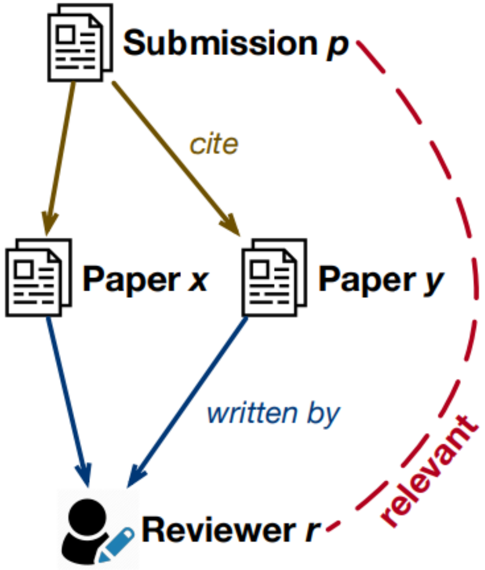


45 results

1 Clinical features of patients infected with 2019 novel **coronavirus** in Wuhan, China.

Huang C, et al. Lancet. 2020. PMID: 31986264 Free PMC article.

BACKGROUND: A recent cluster of pneumonia **cases** in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel **coronavirus** ...

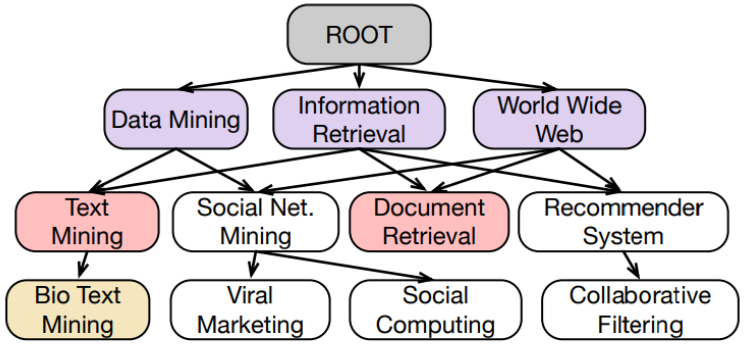


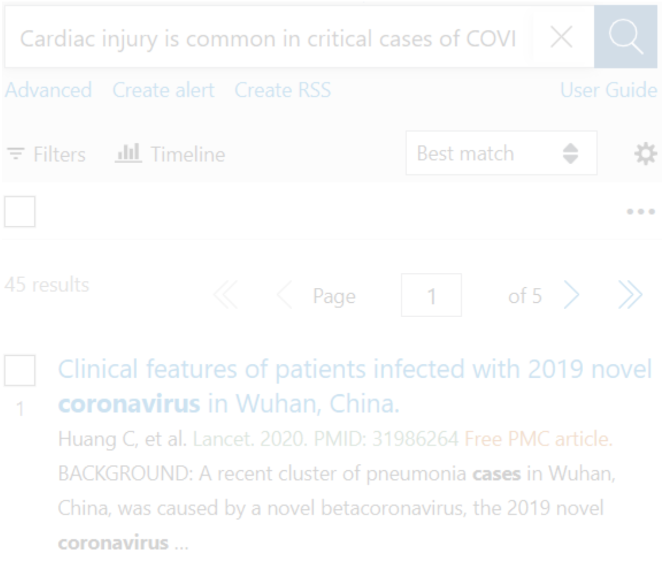
Overview of Part III

Structure-Aware LLMs for Scientific Paper Classification

Structure-Aware LLMs for Scientific Literature Search

Structure-Aware LLMs for Paper-Reviewer Matching




Cardiac injury is common in critical cases of COVI

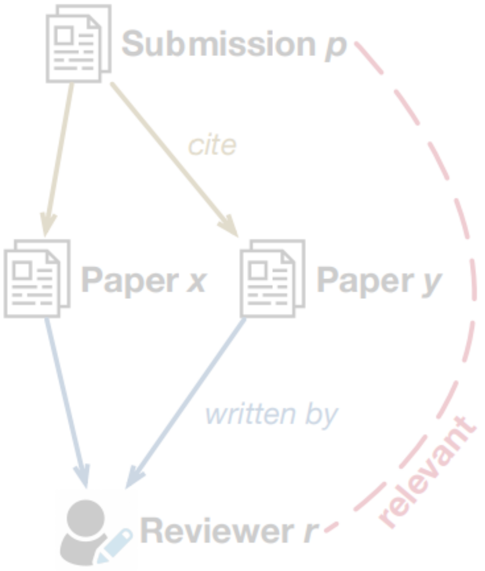
45 results

Page 1 of 5

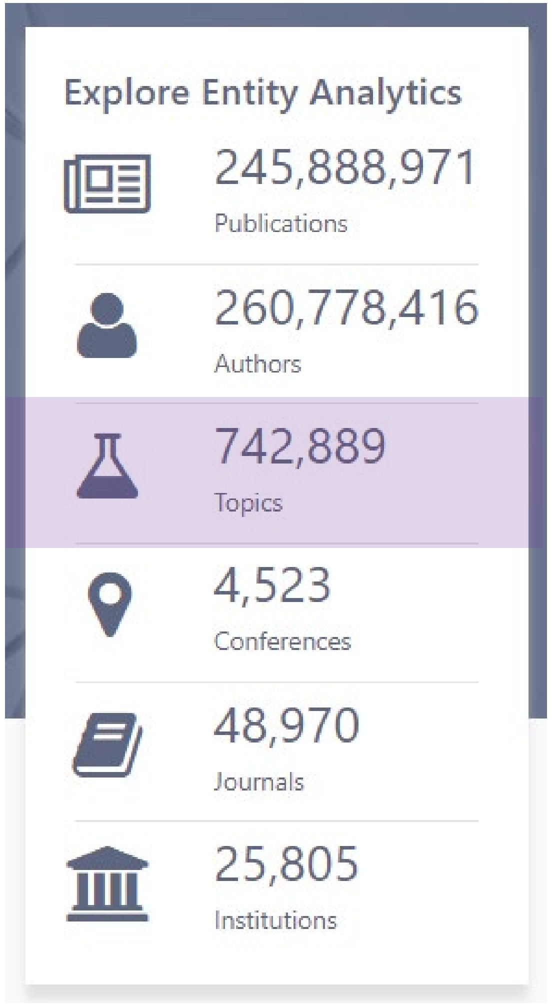
1 Clinical features of patients infected with 2019 novel **coronavirus** in Wuhan, China.

Huang C, et al. Lancet. 2020. PMID: 31986264 Free PMC article.

BACKGROUND: A recent cluster of pneumonia **cases** in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel **coronavirus** ...



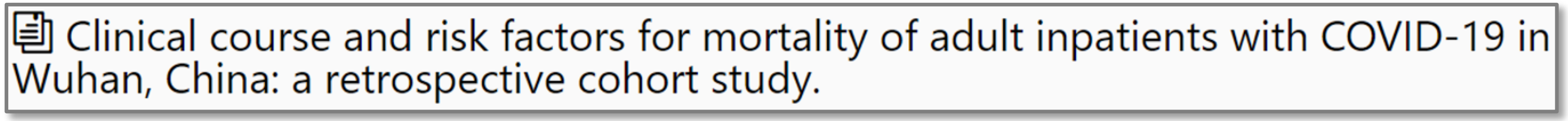
Extremely Fine-Grained Scientific Paper Classification



Explore Entity Analytics

245,888,971	Publications
260,778,416	Authors
742,889	Topics
4,523	Conferences
48,970	Journals
25,805	Institutions

- 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).



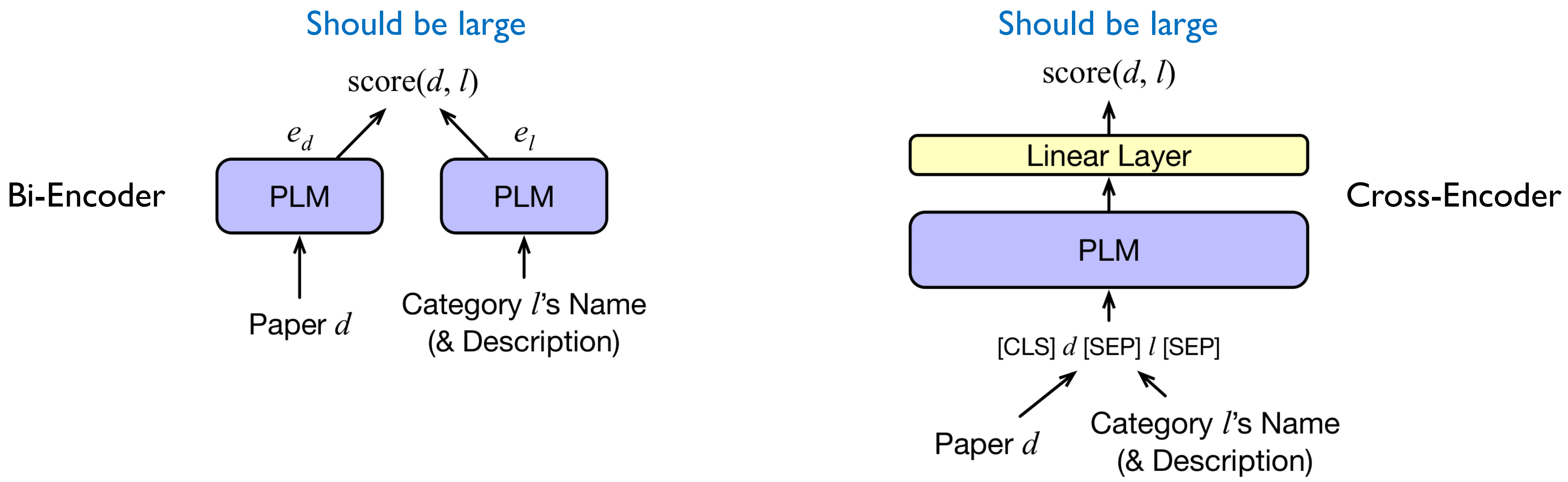
Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.

- 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

If we could have some training data …

- We could use relevant (paper, category) pairs to fine-tune a pre-trained language model.
- Both Bi-Encoder and Cross-Encoder are applicable.



Should be large

$\text{score}(d, l)$

e_d

e_l

Bi-Encoder

PLM

PLM

Paper d

Category l's Name
(& Description)

Should be large

$\text{score}(d, l)$

Linear Layer

PLM

[CLS] d [SEP] l [SEP]

Paper d

Category l's Name
(& Description)

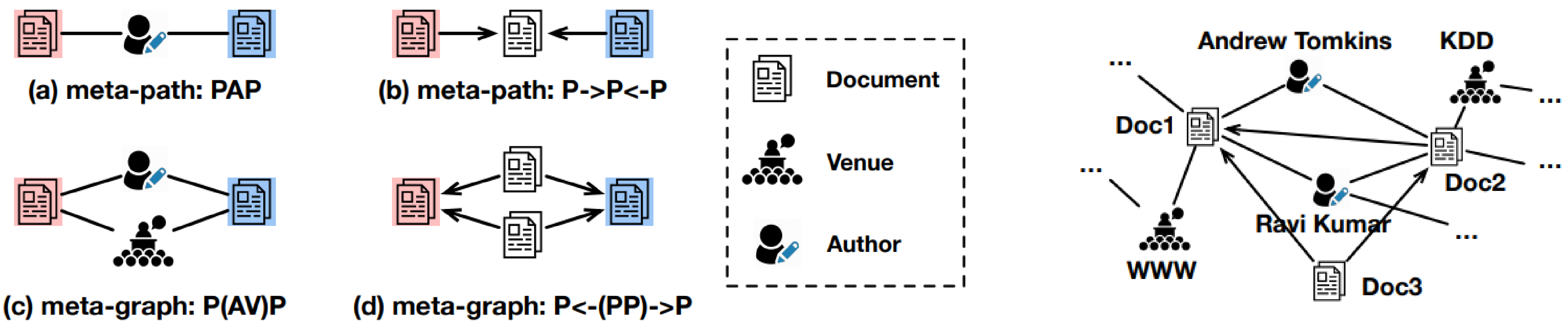
Cross-Encoder

- However, human-annotated training samples are NOT available in many cases!
  - We are asking annotators to find ~10 relevant categories from ~100,000 candidates!

Using Structured Information to Replace Annotations

- If relevant (paper, category) pairs are not available, can we automatically create relevant (paper, paper) pairs?
  - Two papers sharing the same author(s) are assumed to be similar.
  - Two papers sharing the same reference(s) are assumed to be similar.
  - …

- The notion of meta-paths and meta-graphs



(a) meta-path: PAP

(b) meta-path: P->P<-P

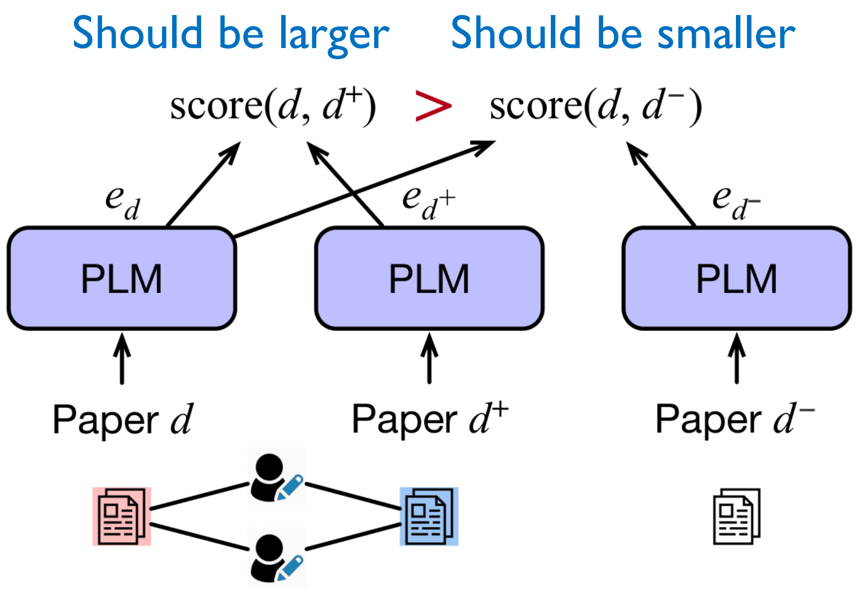
(c) meta-graph: P(AV)P

(d) meta-graph: P<-(PP)->P

Structure-Induced Contrastive Learning

- Two papers connected via a certain meta-path/meta-graph should be more similar than two randomly selected papers.

Bi-Encoder



Should be larger Should be smaller

score(d, d^+) > score(d, d^-)

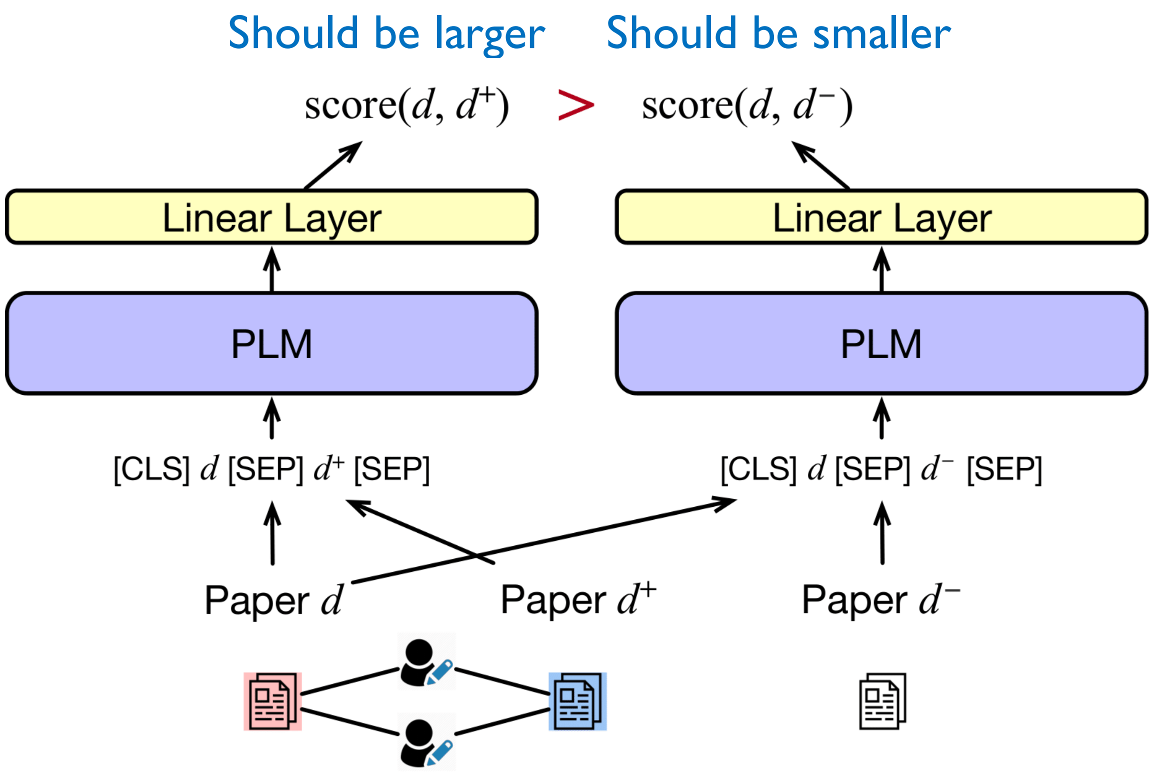
e_d e_{d^+} e_{d^-}

PLM PLM PLM

Paper d Paper d^+ Paper d^-

$$-\log \frac{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau)}{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau) + \sum_{i=1}^{N} \exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d_i^-})/\tau)}$$

Cross-Encoder



Should be larger Should be smaller

score(d, d^+) > score(d, d^-)

Linear Layer Linear Layer

PLM PLM

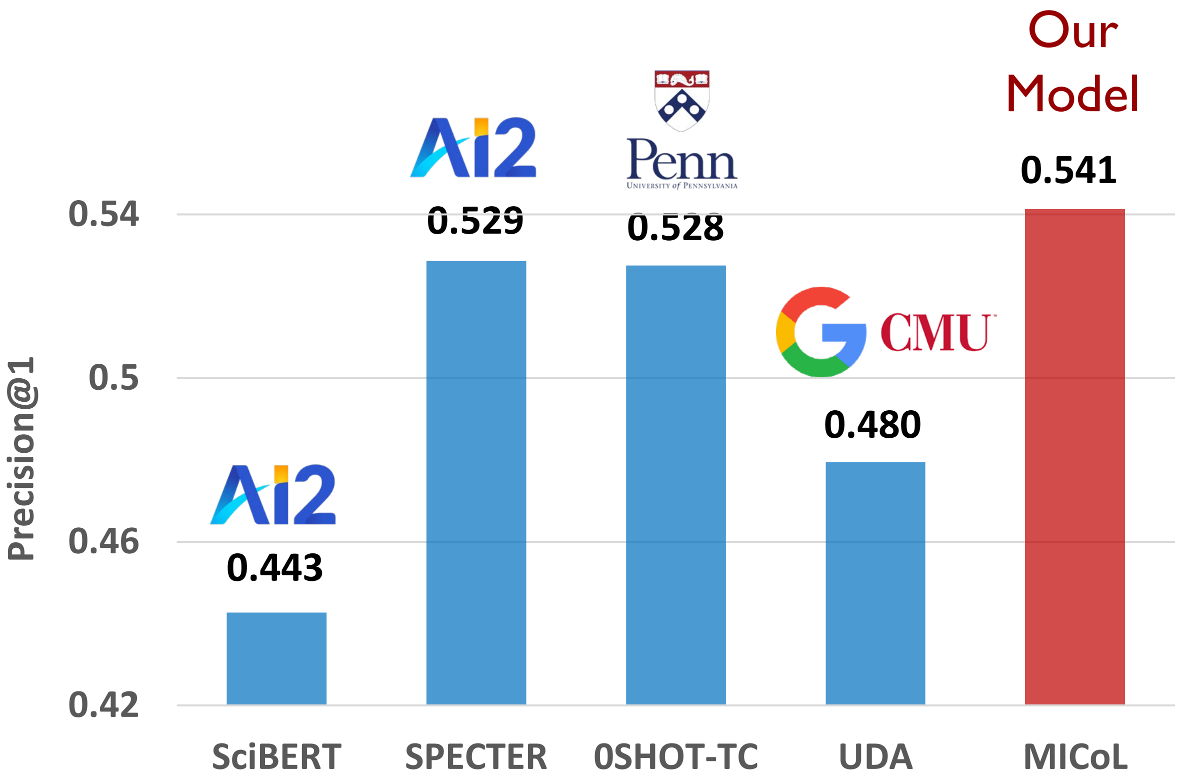
[CLS] d [SEP] d^+ [SEP] [CLS] d [SEP] d^- [SEP]

Paper d Paper d^+ Paper d^-

Zhang et al., *Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification.* WWW 2022.

Comparison with Previous Approaches

- Dataset: Microsoft Academic Graph and PubMed
- Metric: Precision@1, 3, and 5



Zhang et al., *Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification.* WWW 2022.

Case Study

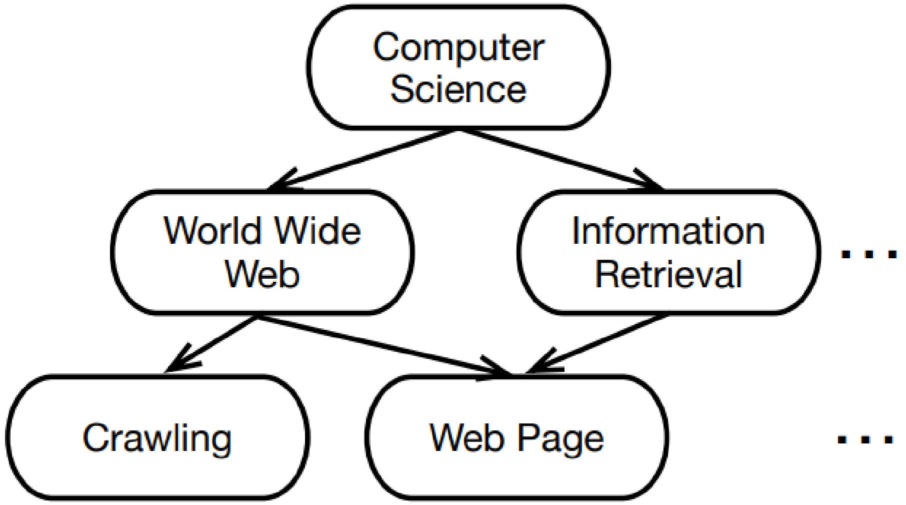
- Title: Improving Text Categorization Methods for Event Tracking
- Venue: SIGIR (2000)
- Authors: Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer
- Abstract: : Automated tracking of events from chronologically ordered document streams is a new challenge for statistical text classification. Existing learning techniques must be adapted or improved in order to effectively handle difficult situations where the number of positive training instances per event …

- Top-5 Predictions of a Text-Only Baseline: K Nearest Neighbors Algorithm (✓), Data Mining (✓), Pattern Recognition (✓), Machine Learning (✓), Nearest Neighbor Search (✗)

- Top-5 Predictions of our Metadata-Aware Method: K Nearest Neighbors Algorithm (✓), Data Mining (✓), Information Retrieval (✓), Pattern Recognition (✓), Machine Learning (✓)

How about other types of structured information?

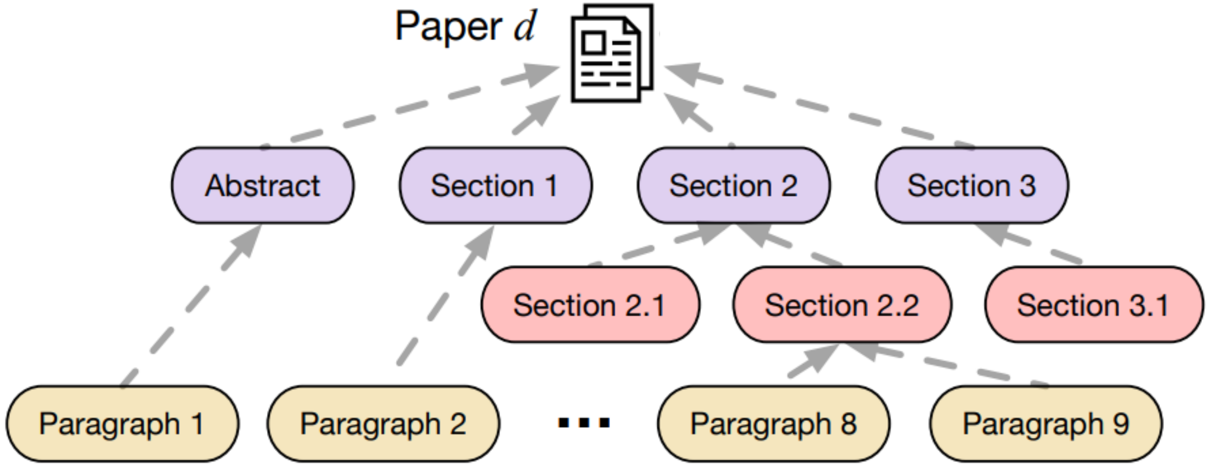
Label Hierarchy



Top-Down Pruning:

Irrelevant to WWW ⇒ Irrelevant to Crawling

In-Document Structure



Bottom-Up Aggregation:

Paragraphs → Subsections → Sections → Paper

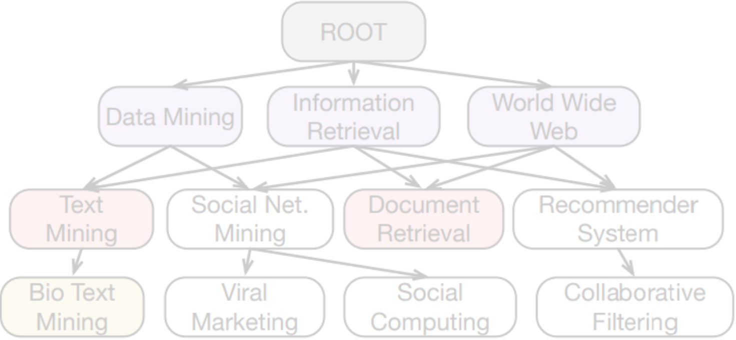
Zhang et al., *MATCH: Metadata-Aware Text Classification in A Large Hierarchy*. WWW 2021.
Zhang et al., *Weakly Supervised Multi-Label Classification of Full-Text Scientific Papers*. KDD 2023.

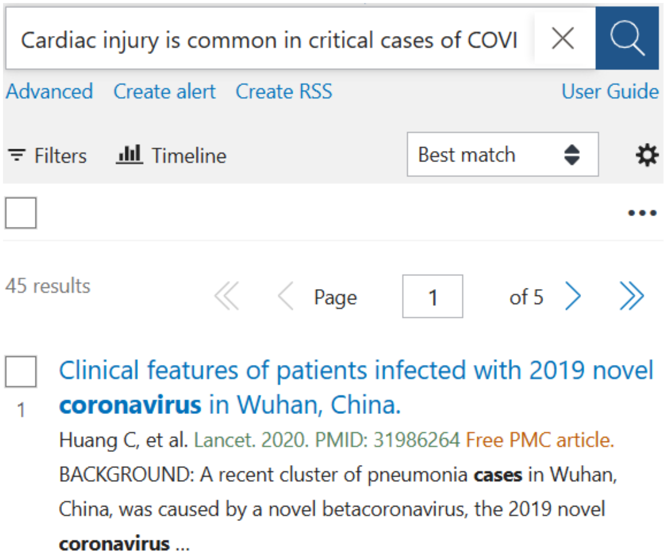
Overview of Part III

Structure-Aware LLMs for Scientific Paper Classification

Structure-Aware LLMs for Scientific Literature Search

Structure-Aware LLMs for Paper-Reviewer Matching





Cardiac injury is common in critical cases of COVI

☐

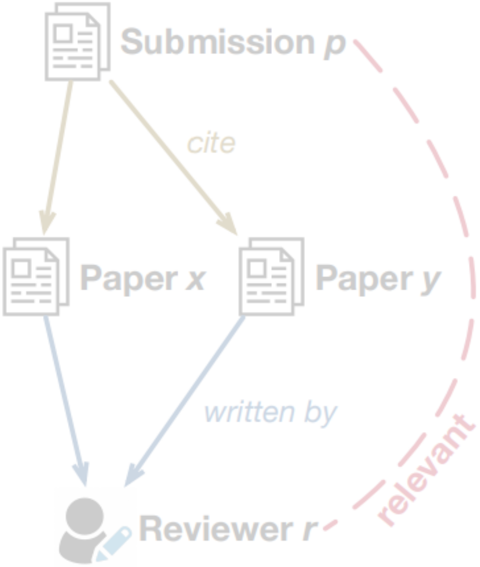
45 results

Page 1 of 5

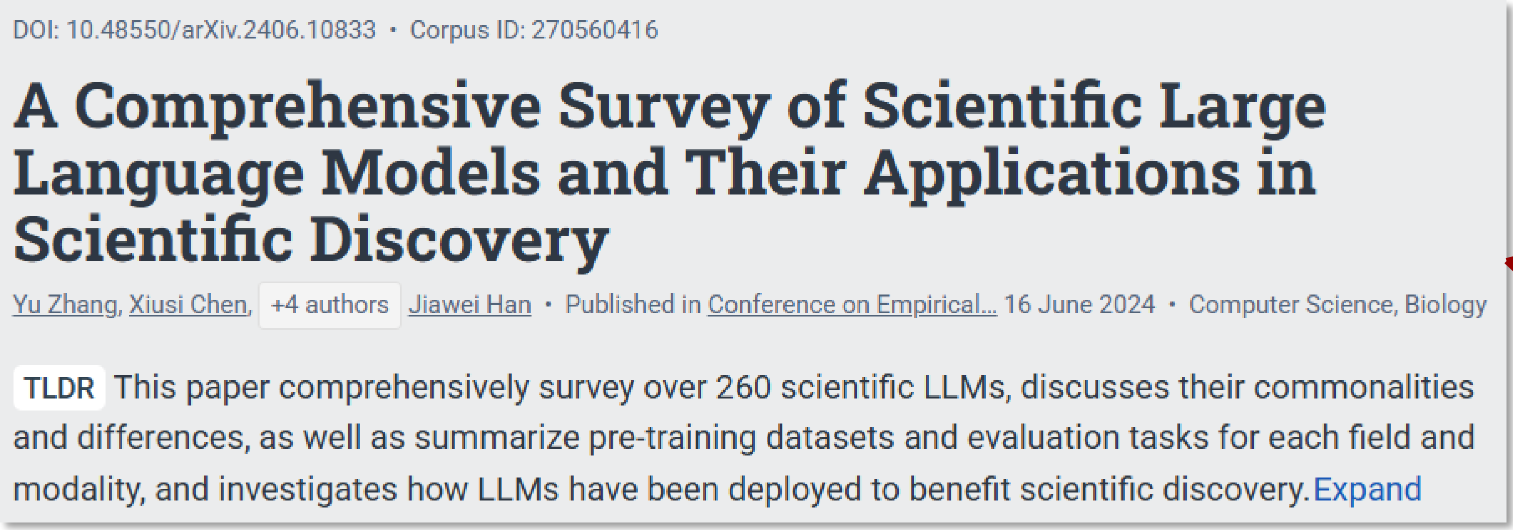
☐ 1 Clinical features of patients infected with 2019 novel **coronavirus** in Wuhan, China.

Huang C, et al. Lancet. 2020. PMID: 31986264 Free PMC article.

BACKGROUND: A recent cluster of pneumonia **cases** in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel **coronavirus** ...



Link Prediction

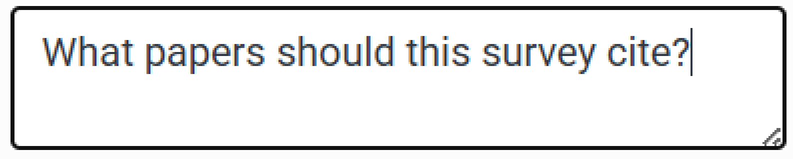


DOI: 10.48550/arXiv.2406.10833 • Corpus ID: 270560416

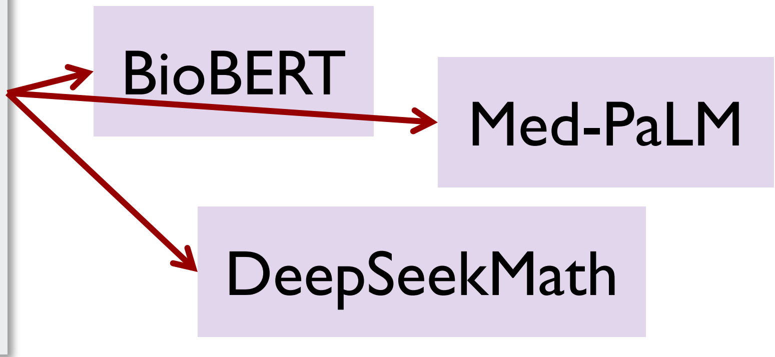
A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery

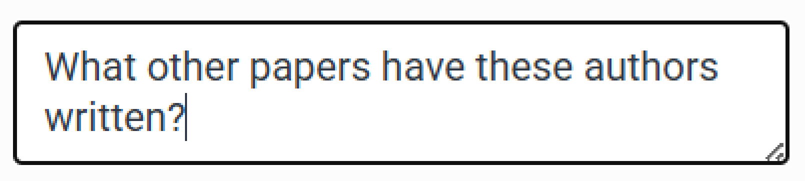
Yu Zhang, Xiusi Chen, +4 authors Jiawei Han • Published in Conference on Empirical... 16 June 2024 • Computer Science, Biology

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

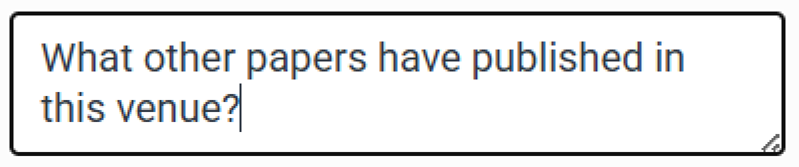


What papers should this survey cite?





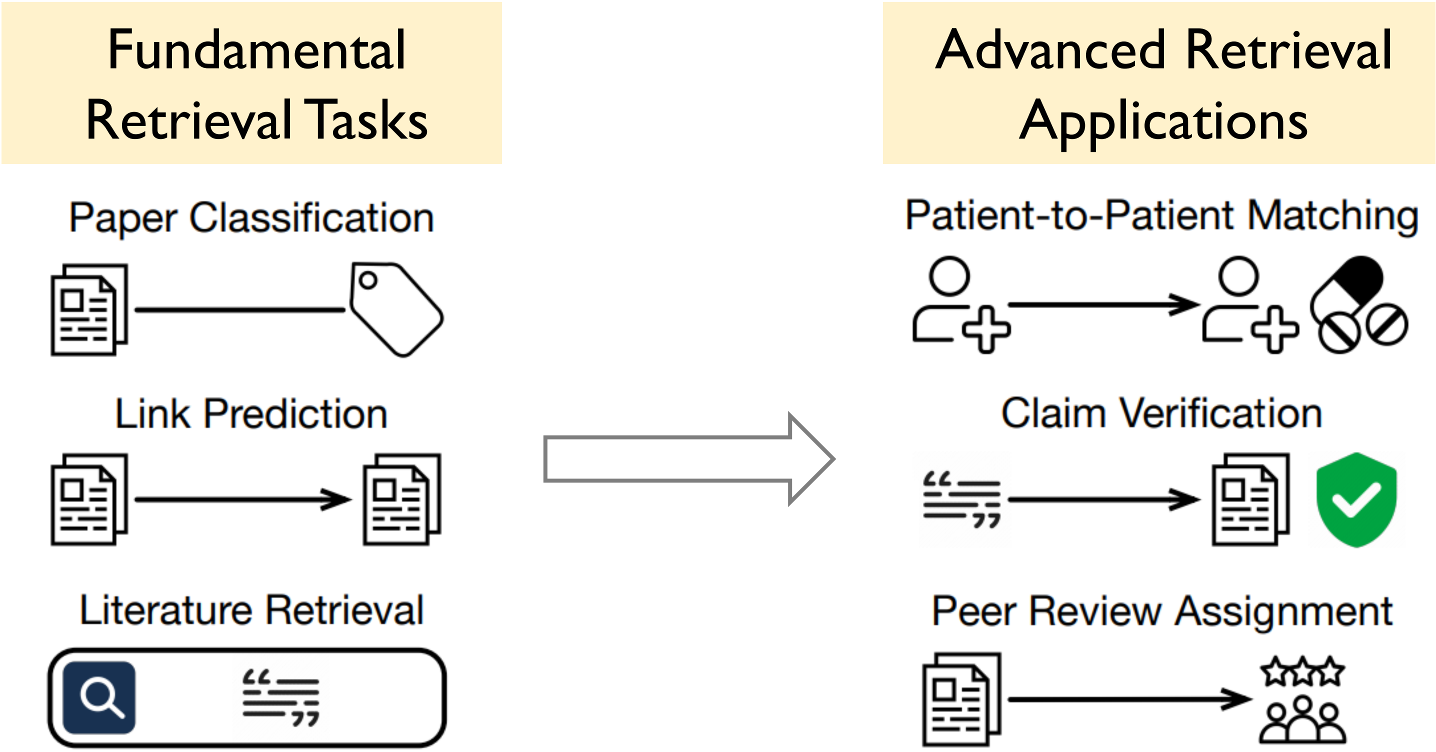
What other papers have these authors written?



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



Fundamental Retrieval Tasks	Advanced Retrieval Applications
Paper Classification	Patient-to-Patient Matching
Link Prediction	Claim Verification
Literature Retrieval	Peer Review Assignment

Given a patient summary, find similar patients/clinical case reports.

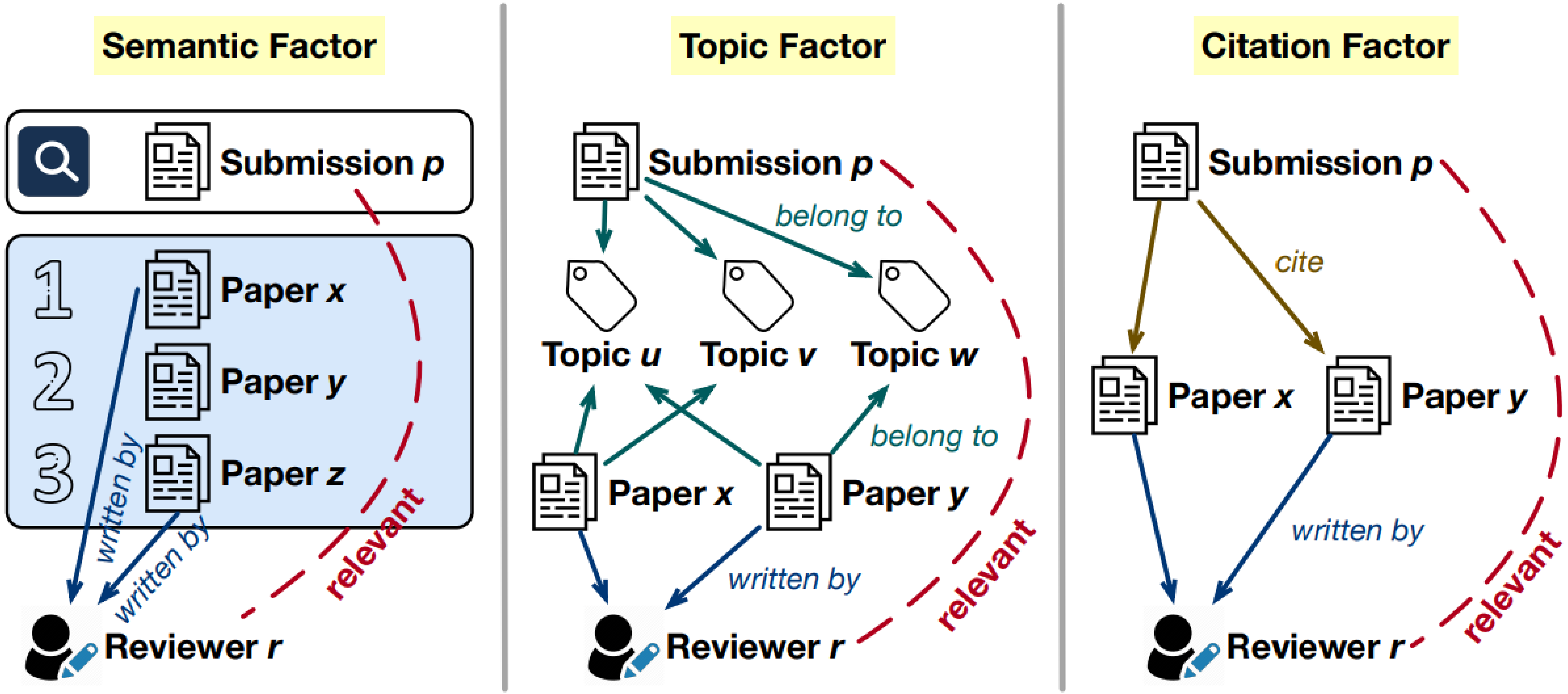
Given a scientific claim, find relevant papers (and predict their stance).

Given a paper submission, find expert reviewers.

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

Multiple Factors for Judging Relevance

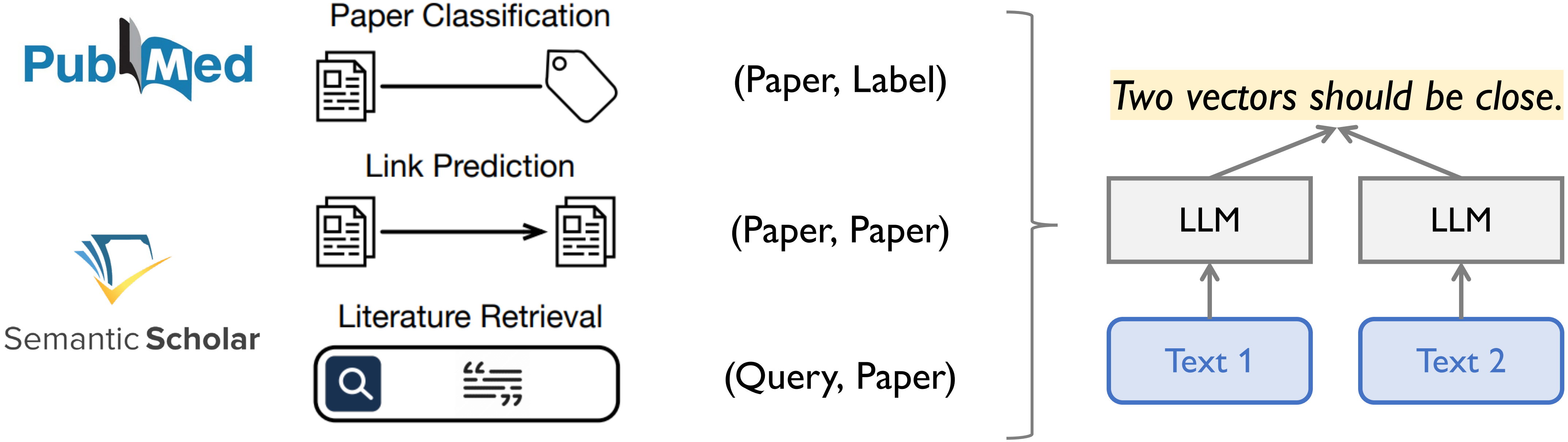
- Example: Paper-Reviewer Matching
  - Why is a pair of (Paper, Reviewer) relevant?



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

Contrastive Learning for 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?



PubMed

Semantic Scholar

Paper Classification

(Paper, Label)

Link Prediction

(Paper, Paper)

Literature Retrieval

(Query, Paper)

Two vectors should be close.

LLM

LLM

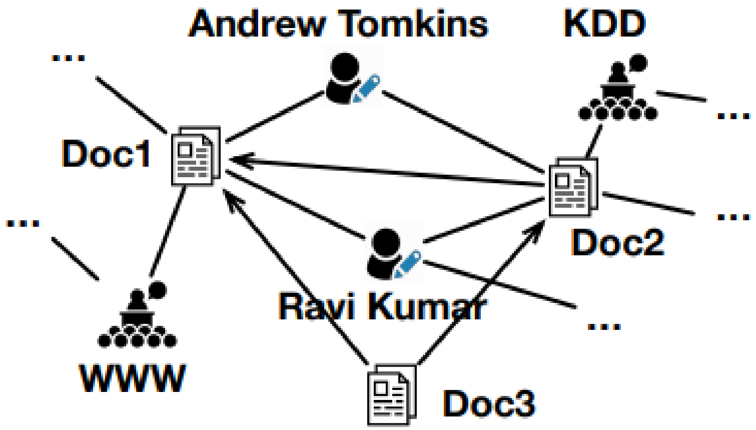
Text 1

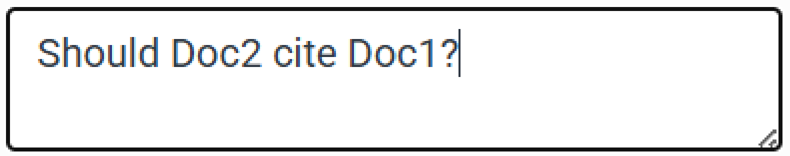
Text 2

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

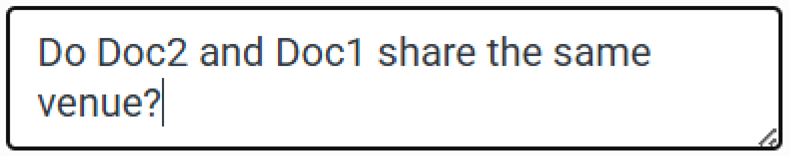
A Toy Example of Task Interference

- 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?

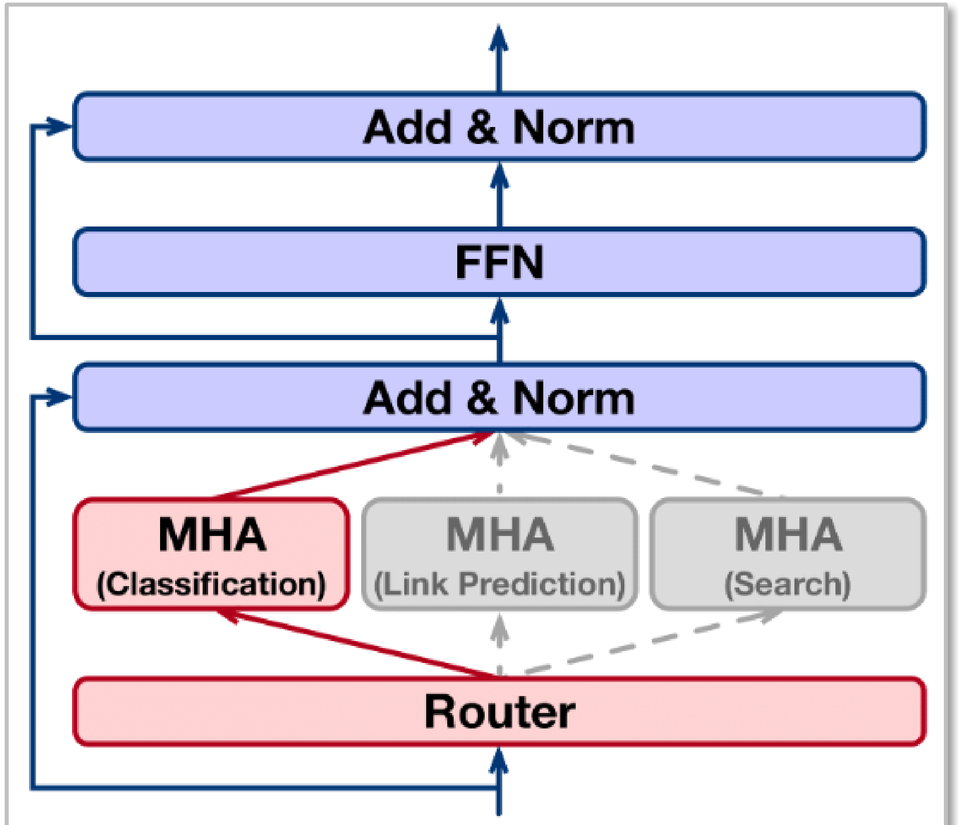
✓


Do Doc2 and Doc1 share the same venue?

✗

Tackling Task Interference: Mixture-of-Experts Transformer

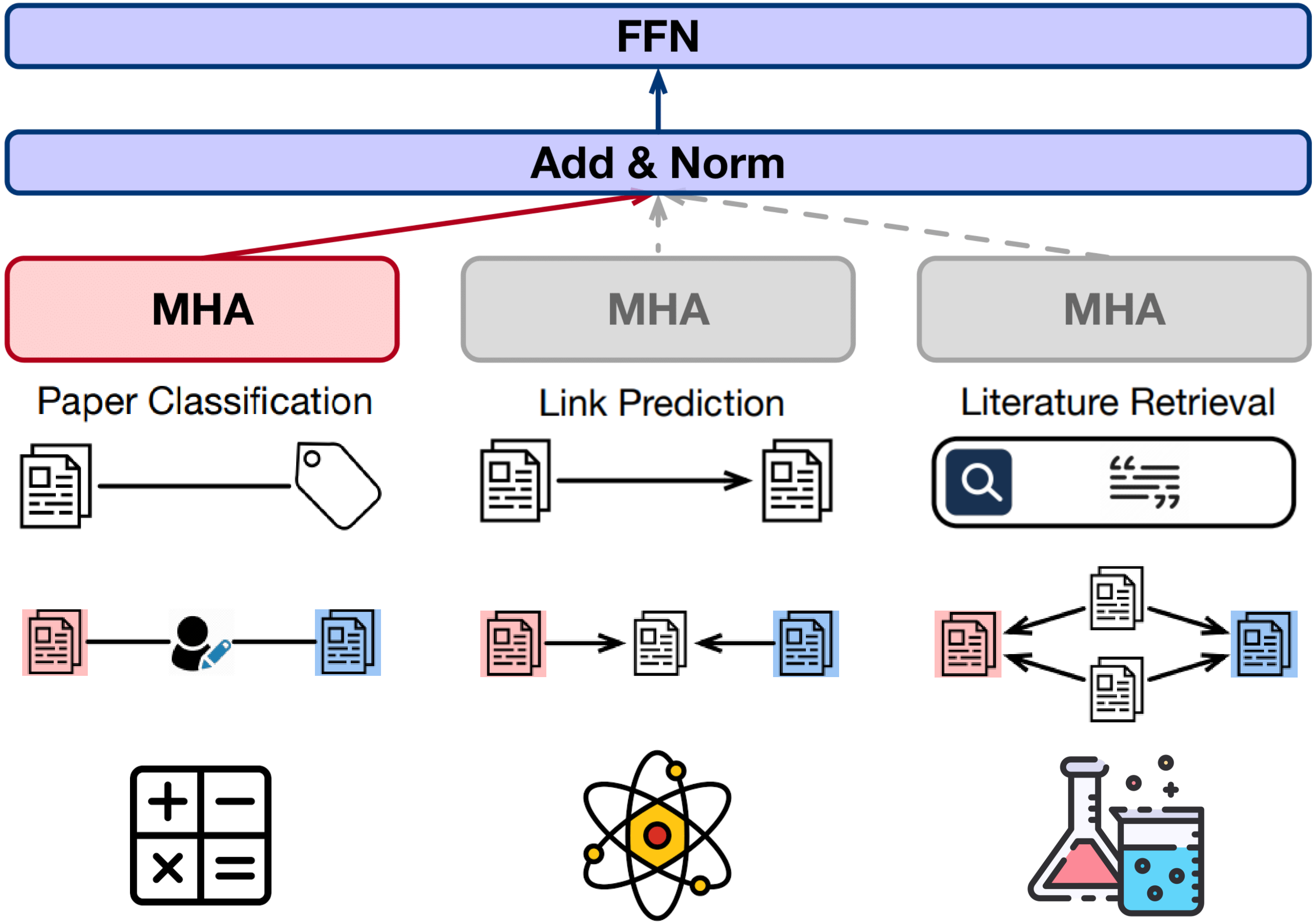
- A typical Transformer layer
  - **1** Multi-Head Attention (MHA) sublayer
  - **1** Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
  - Multiple MHA sublayers
  - **1** FFN sublayer
  - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be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



Mixture-of-Experts Transformer with Task-Specific **MHA** Sublayers

Zhang et al., *Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding*. EMNLP 2023 Findings.

Tackling Task Interference: Mixture-of-Experts Transformer



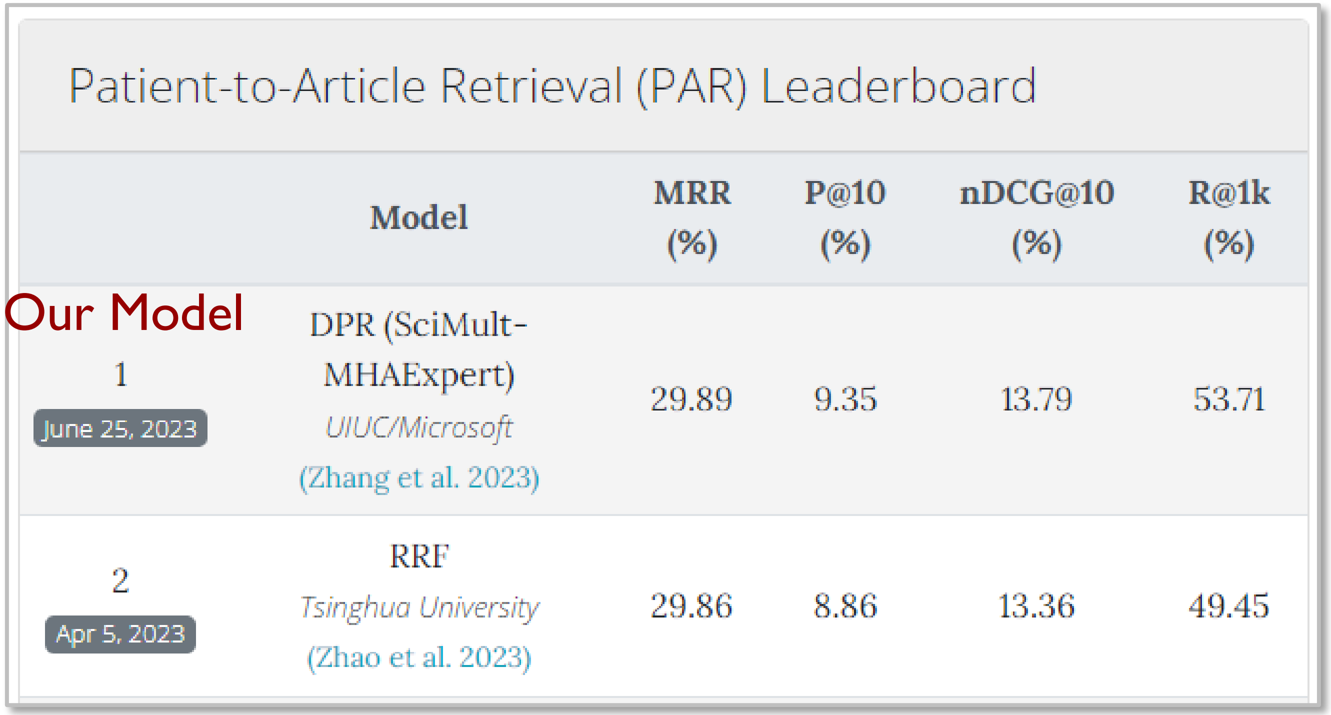
Task

Graph Information

Field

Comparison with Previous Approaches

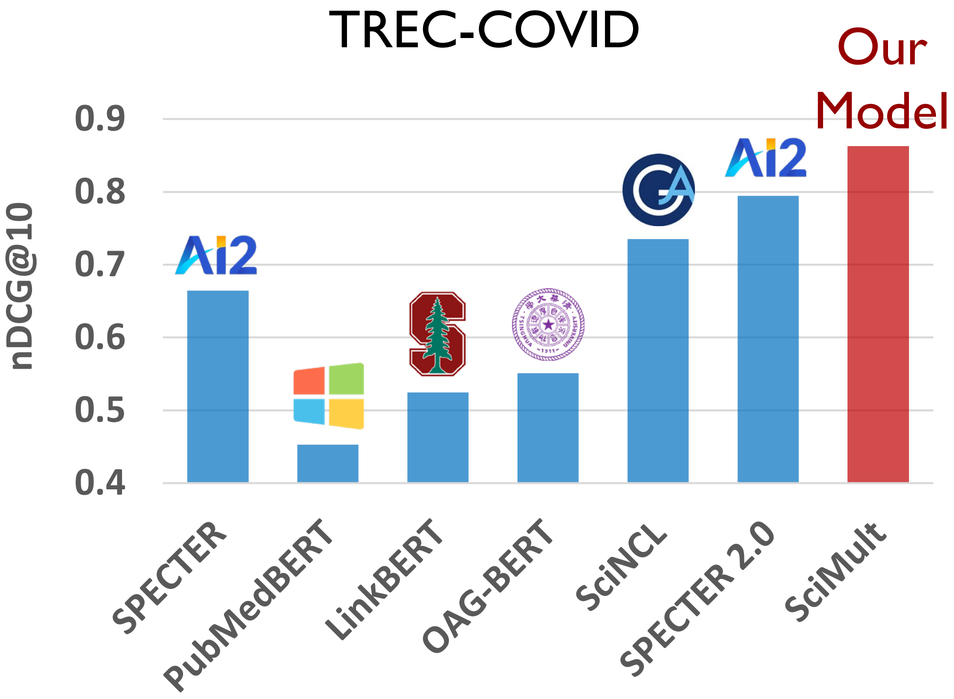
- New SOTA on the PMC-Patients benchmark (patient-to-article retrieval)
- Outperforming previous scientific pre-trained language models in classification, link prediction, literature retrieval (TREC-COVID), paper recommendation, and claim verification (SciFact)



Patient-to-Article Retrieval (PAR) Leaderboard

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

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



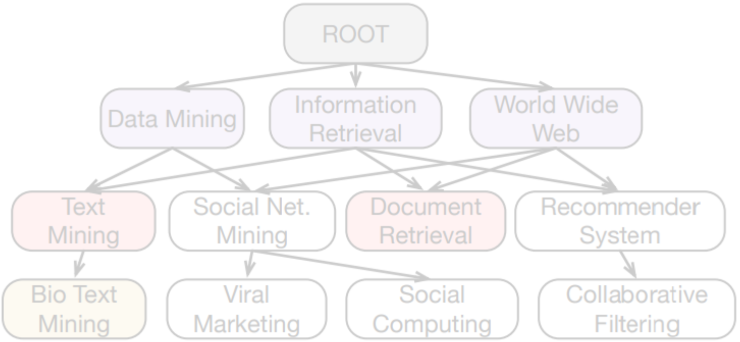
Zhang et al., *Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding*. EMNLP 2023 Findings.

Overview of Part III

Structure-Aware LLMs for Scientific Paper Classification

Structure-Aware LLMs for Scientific Literature Search

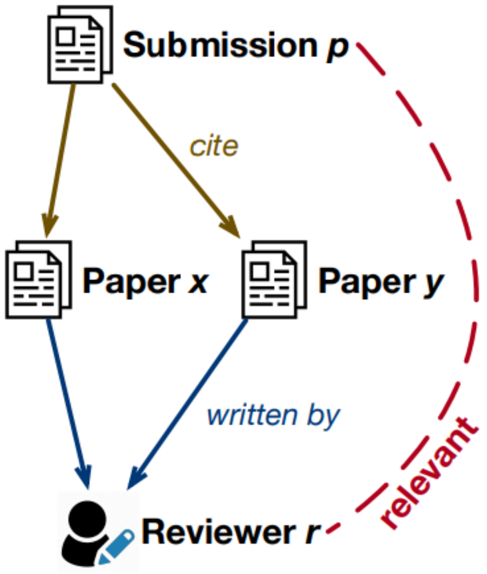
Structure-Aware LLMs for Paper-Reviewer Matching



1 Clinical features of patients infected with 2019 novel **coronavirus** in Wuhan, China.

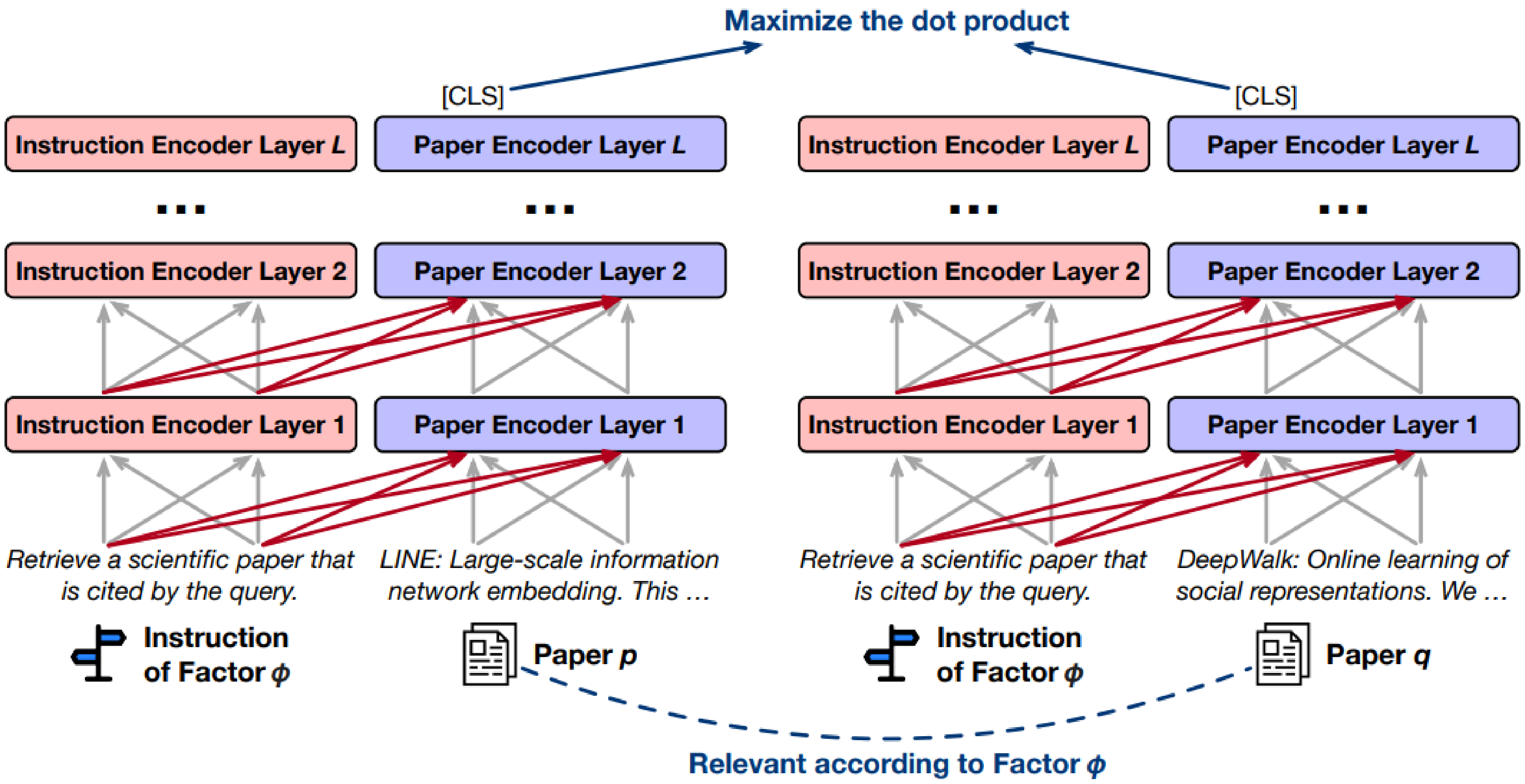
Huang C, et al. Lancet. 2020. PMID: 31986264 Free PMC article.

BACKGROUND: A recent cluster of pneumonia **cases** in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel **coronavirus** ...



Tackling Task Interference: Instruction Tuning

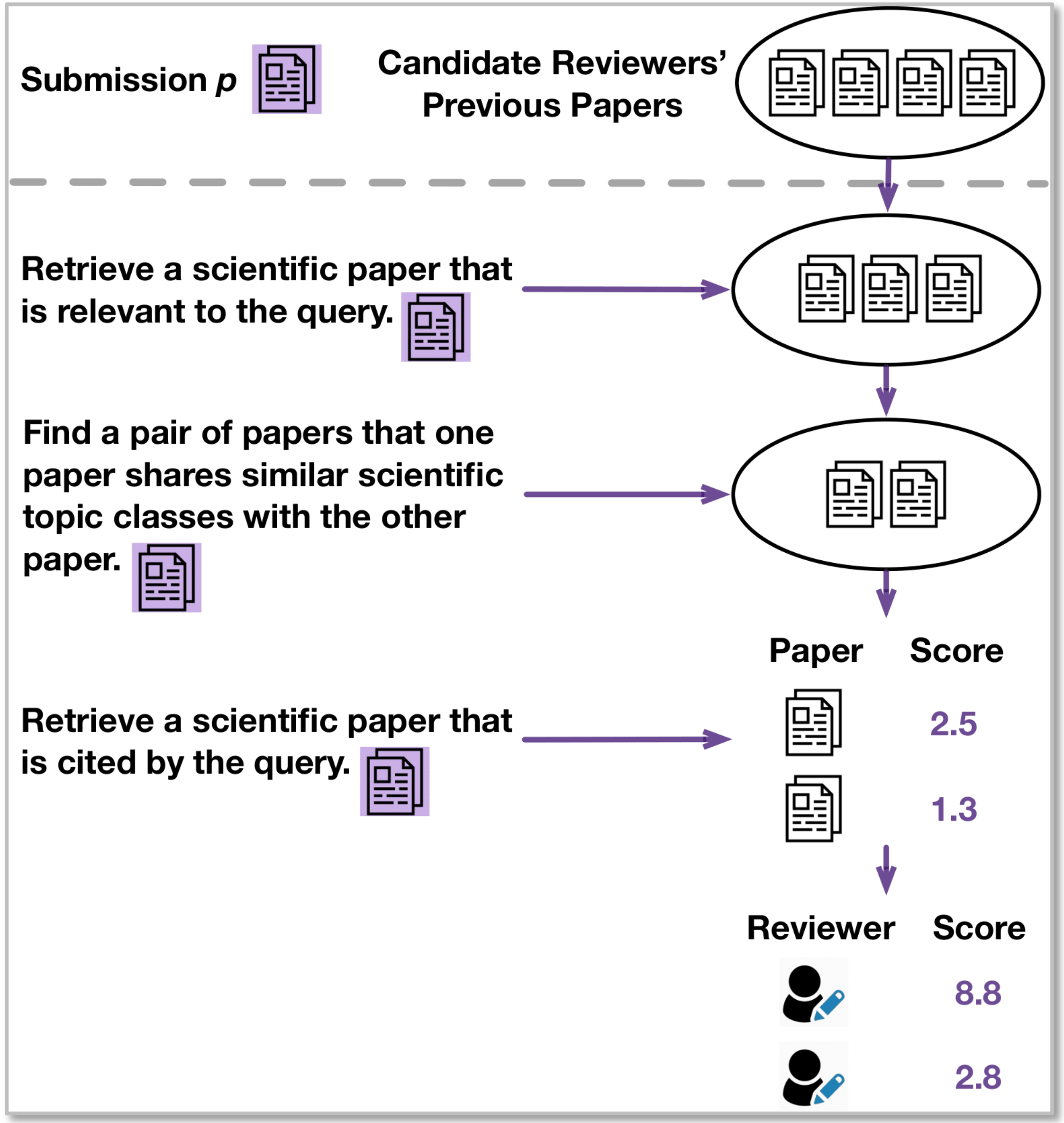
- Using a factor-specific instruction to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.


Maximize the dot product
[CLS]
[CLS]
Instruction Encoder Layer L
Paper Encoder Layer L
Instruction Encoder Layer L
Paper Encoder Layer L
...
Instruction Encoder Layer 2
Paper Encoder Layer 2
Instruction Encoder Layer 2
Paper Encoder Layer 2
Instruction Encoder Layer 1
Paper Encoder Layer 1
Instruction Encoder Layer 1
Paper Encoder Layer 1
Retrieve a scientific paper that is cited by the query.
LINE: Large-scale information network embedding. This …
Retrieve a scientific paper that is cited by the query.
DeepWalk: Online learning of social representations. We …
Instruction of Factor ϕ
Paper p
Instruction of Factor ϕ
Paper q
Relevant according to Factor ϕ

Zhang et al., *Chain-of-Factors Paper-Reviewer Matching*. WWW 2025.

Chain-of-Factors Reasoning

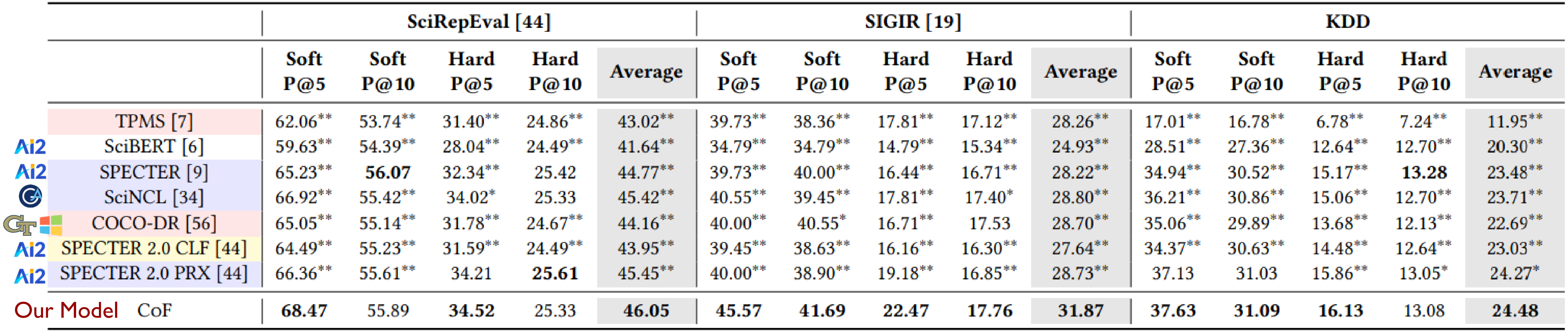
- Consider semantic, topic, and citation factors in a step-by-step, coarse-to-fine manner.
- Step 1: Semantic relevance serves as the coarsest signal to filter totally irrelevant papers.
- Step 2: Then, we can classify each submission and each relevant paper to a fine-grained topic space and check if 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.


Submission *p*
Candidate Reviewers' Previous Papers
Retrieve a scientific paper that is relevant to the query.
Find a pair of papers that one paper shares similar scientific topic classes with the other paper.
Retrieve a scientific paper that is cited by the query.
Paper
Score
2.5
1.3
Reviewer
Score
8.8
2.8

Zhang et al., *Chain-of-Factors Paper-Reviewer Matching.* WWW 2025.

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)



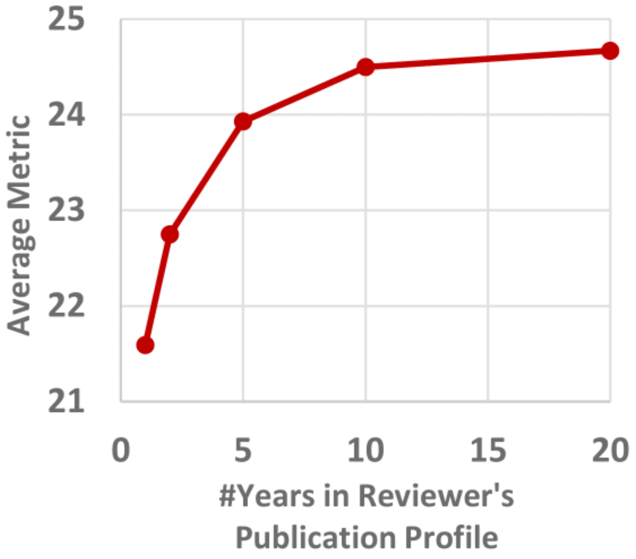
		SciRepEval [44]					SIGIR [19]					KDD				
		Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average
	TPMS [7]	62.06**	53.74**	31.40**	24.86**	43.02**	39.73**	38.36**	17.81**	17.12**	28.26**	17.01**	16.78**	6.78**	7.24**	11.95**
Ai2	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**
Ai2	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**
Ai2	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**
Ai2	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**

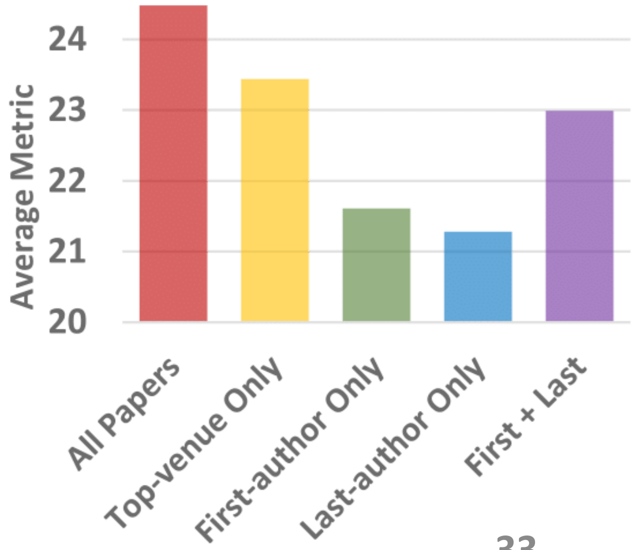
🟥 : semantic-based method 🟨 : topic-based method 🟪 : citation-based method

Zhang et al., *Chain-of-Factors Paper-Reviewer Matching.* WWW 2025.

Impact of Reviewer’s Profile on the Matching Performance

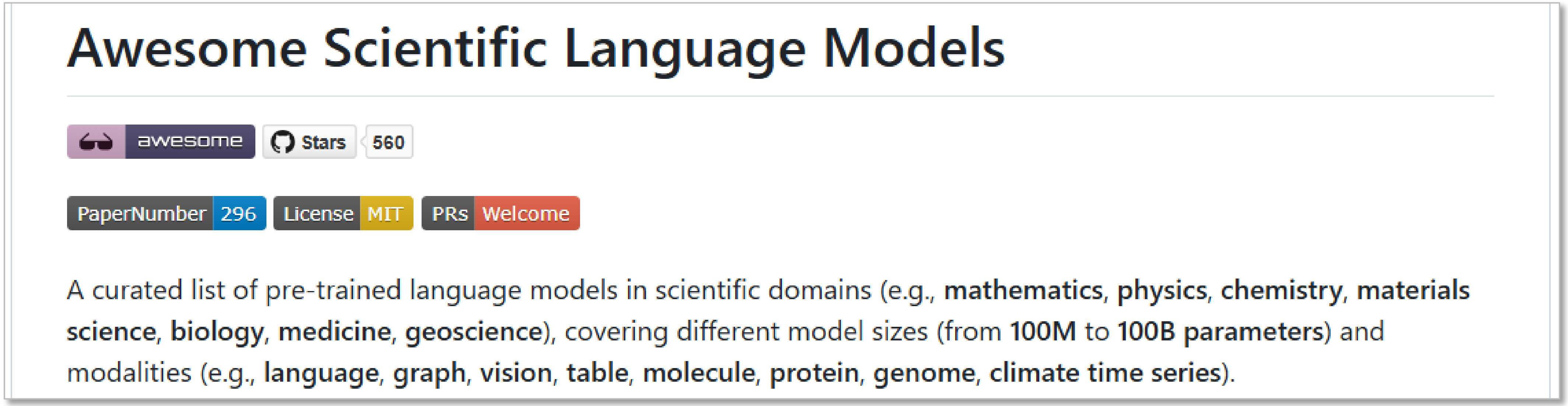
- Shall we include all papers written by a reviewer or set up some criteria?
- Timespan: What if we include papers published in the most recent Y years only (because earlier papers may have diverged from reviewers’ current interests)?
  - Earlier papers still help, but the contribution becomes subtle when $Y \geq 10$.
- Venue: What if we include papers published in top venues only?
  - Harmful!
- Rank in the author list: What if we include each reviewer’s first-author and/or last-author papers only?
  - Harmful!
- When the indication from reviewers is not available, putting the entire set of their papers into their publication profile is almost always helpful.





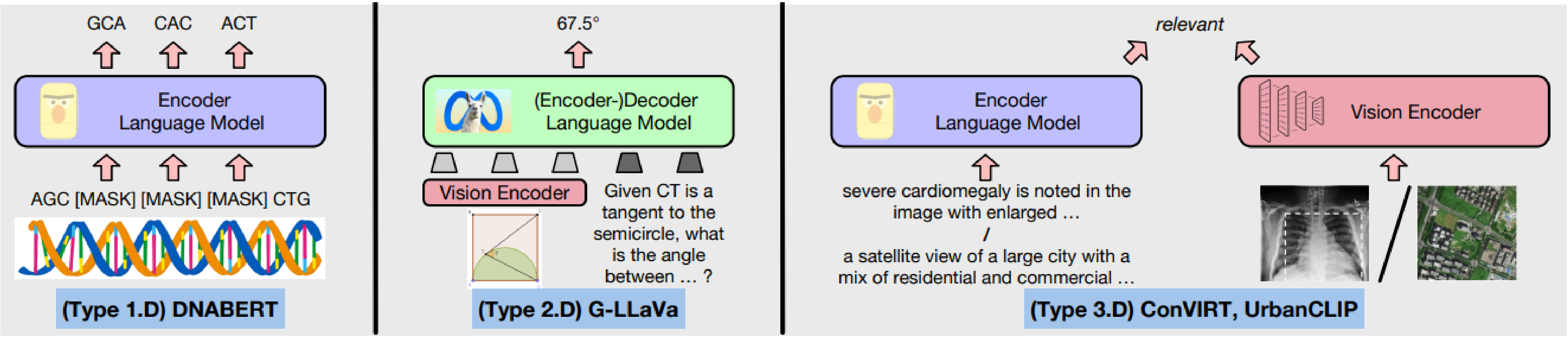
Zhang et al., *Chain-of-Factors Paper-Reviewer Matching.* WWW 2025.

Our Recent Survey Papers



Awesome Scientific Language Models

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**).

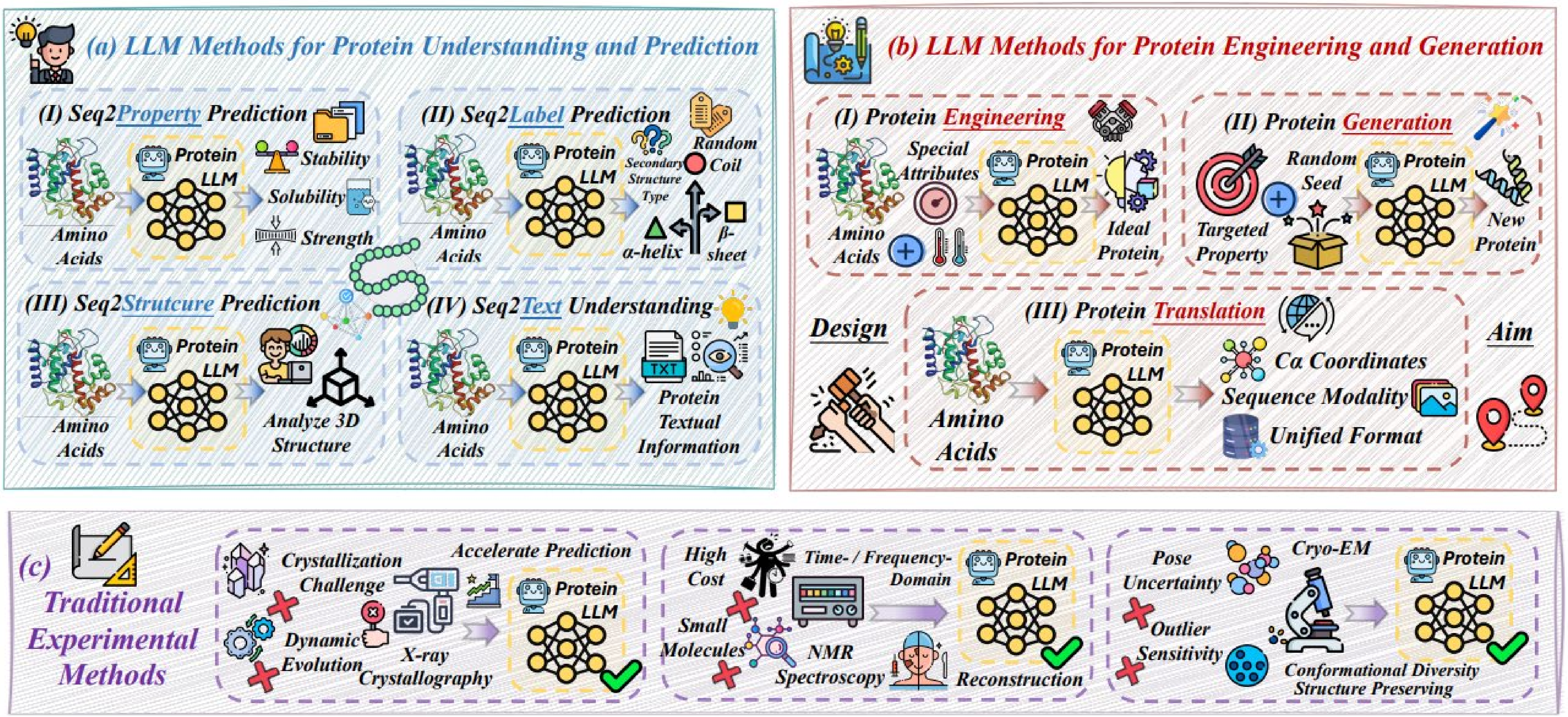


Given CT is a tangent to the semicircle, what is the angle between … ?

severe cardiomegaly is noted in the image with enlarged …
/
a satellite view of a large city with a mix of residential and commercial …

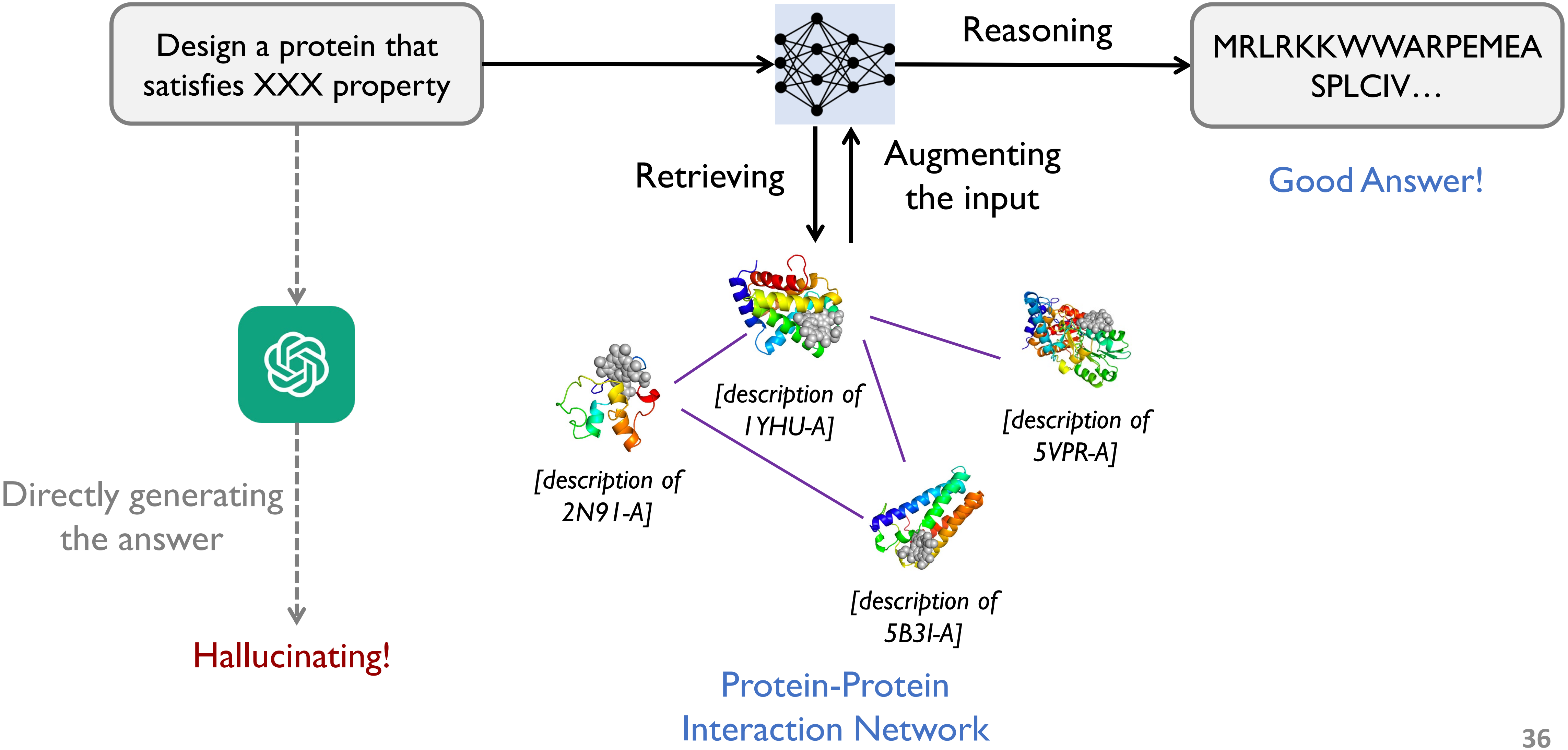
Zhang et al., *A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery*. EMNLP 2024.

Our Recent Survey Papers



Xiao et al., *Protein Large Language Models: A Comprehensive Survey*. arXiv 2025.

Ongoing Work: Expanding to Scientific Data in Other Modalities



Design a protein that satisfies XXX property

Reasoning

MRLRKKWWARPEMEA SPLCIV…

Good Answer!

Retrieving

Augmenting the input

Directly generating the answer

Hallucinating!

[description of 1YHU-A]

[description of 5VPR-A]

[description of 2N91-A]

[description of 5B3I-A]

Protein-Protein Interaction Network

Thank you! Questions?