



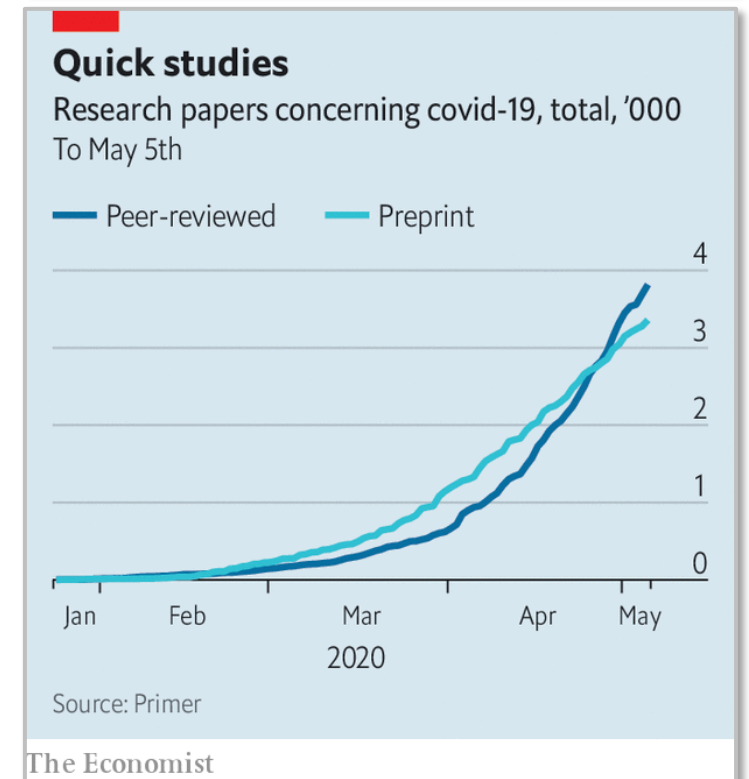
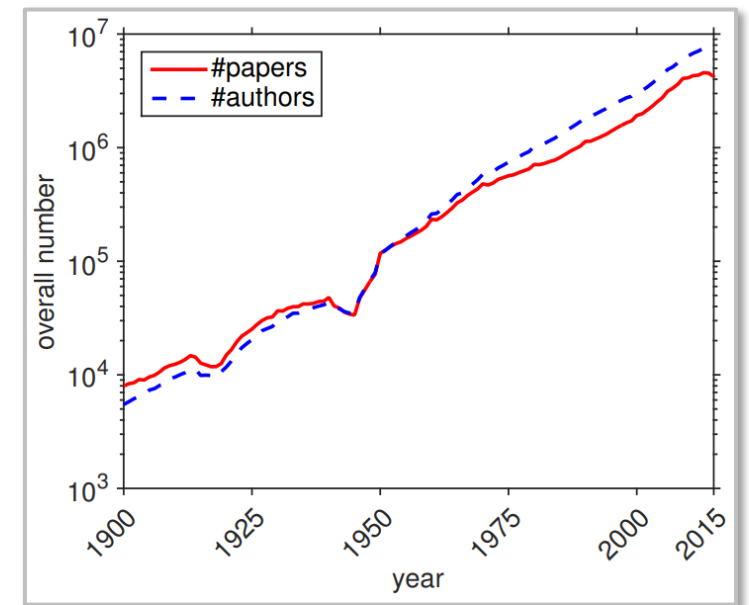
Graph-Enhanced Scientific Text Mining

Yu Zhang

University of Illinois at Urbana-Champaign

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 before the end of April 2020 [3]
- How to prevent researchers from drowning in the whole literature?



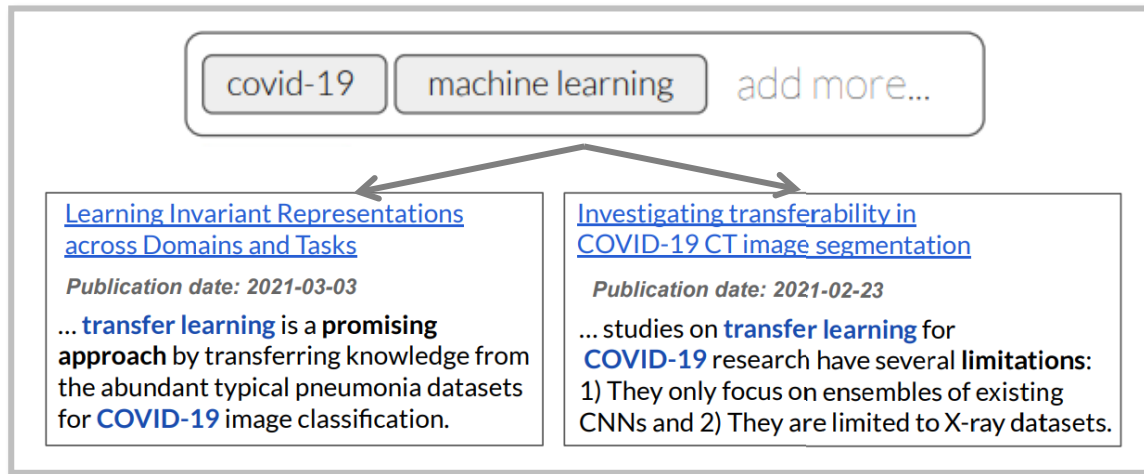
[1] "A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations." KDD 2017.

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

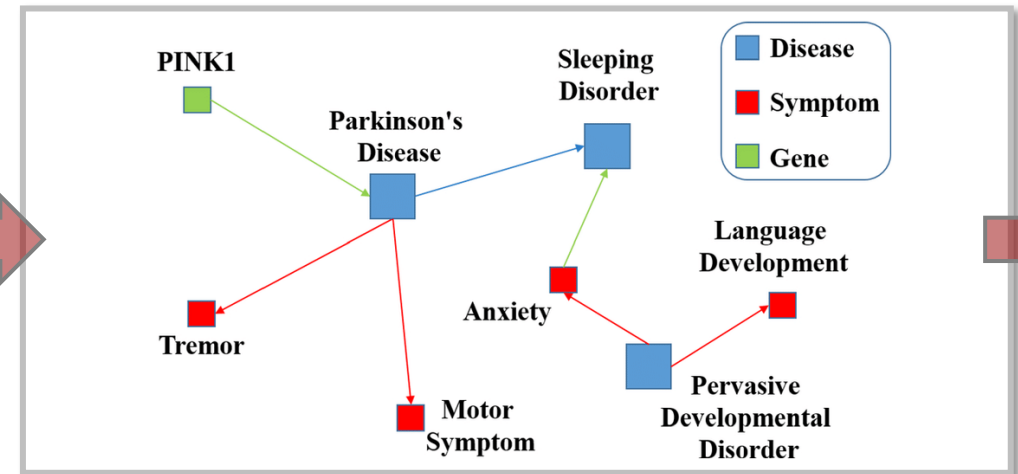
[3] <https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent>

How can text mining help scientific discovery?

Retrieving and Analyzing Relevant Literature



Uncovering Knowledge Structures



- **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*”.

- **Example tasks:**


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

How can text mining help scientific discovery?

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 « » 1 » Show: 25 50 100 All Clear All Filters

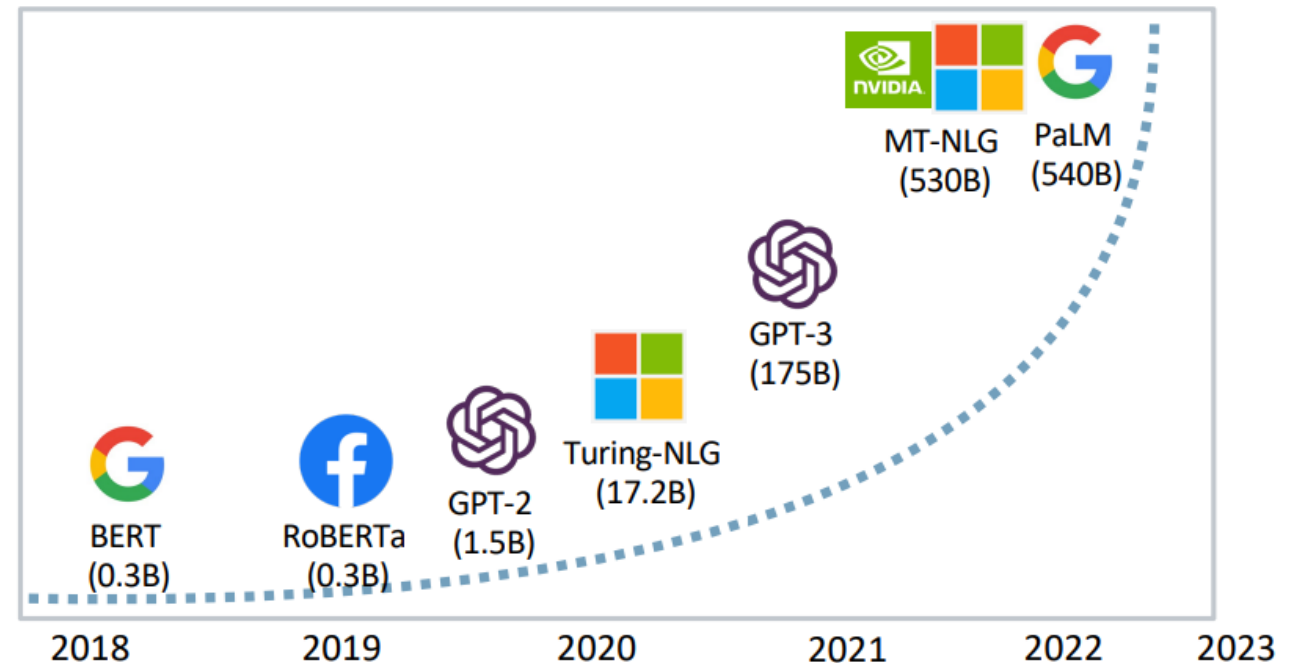
Paper ID↑	Title	Subject Areas		Review & Discussion	Relevance
		Primary	Secondary		
<input type="text" value="e.g. <3"/>	<input type="text" value="filter..."/>	<input type="text" value="filter..."/>	<input type="text" value="filter..."/>		<input type="text" value="e.g. <3"/>
26	Research Paper Zero 1 Show 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.

Pre-trained Language Models (PLMs) for Text Mining

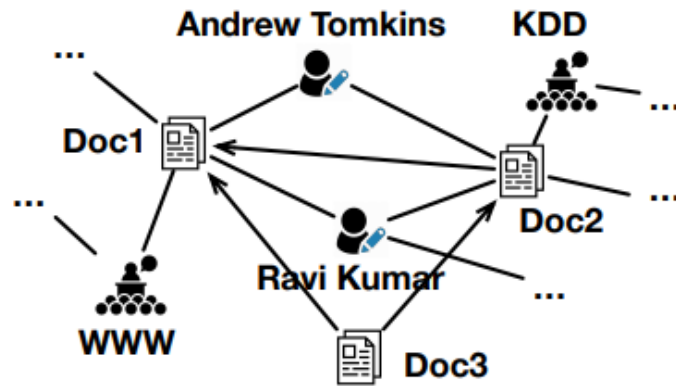
- A **unified** model to perform different text mining tasks **with a few or zero examples**
 - I went to the zoo to see giraffes, lions, and **{zebras, spoon}**. (*Lexical semantics*)
 - I was engaged and on the edge of my seat the whole time. The movie was **{good, bad}**. (*Text classification*)
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



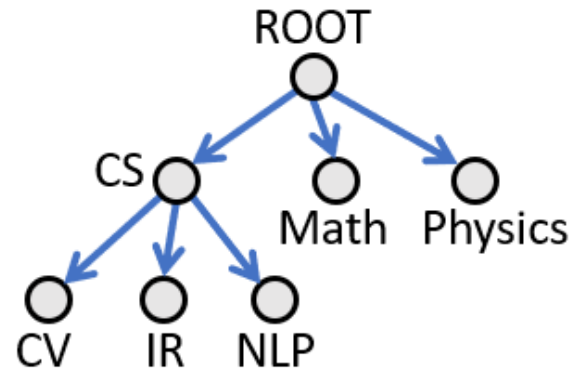
GPT-4
(???)

Are PLMs aware of **graph information**?

Graph Information Associated with Scientific Text



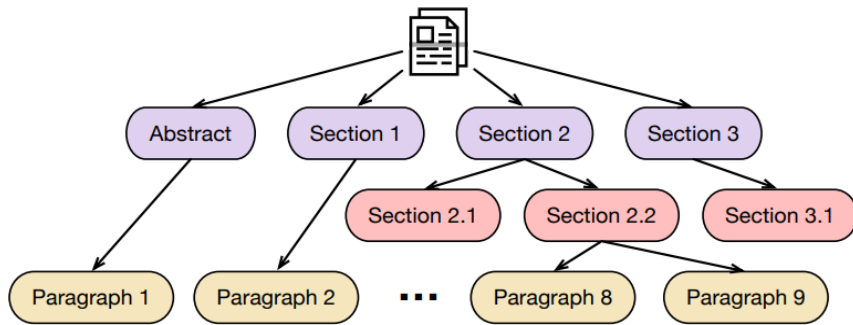
Metadata/Network



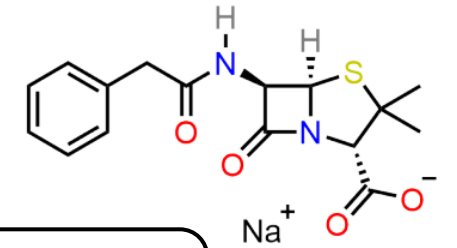
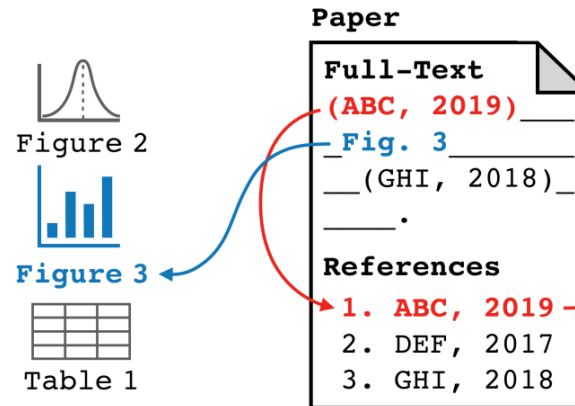
Hierarchy/Taxonomy



Knowledge Graph



In-Document Structure



Penicillin G Sodium is the sodium salt form of benzylpenicillin.

Text-Paired Molecule

PLMs may not be graph-aware!

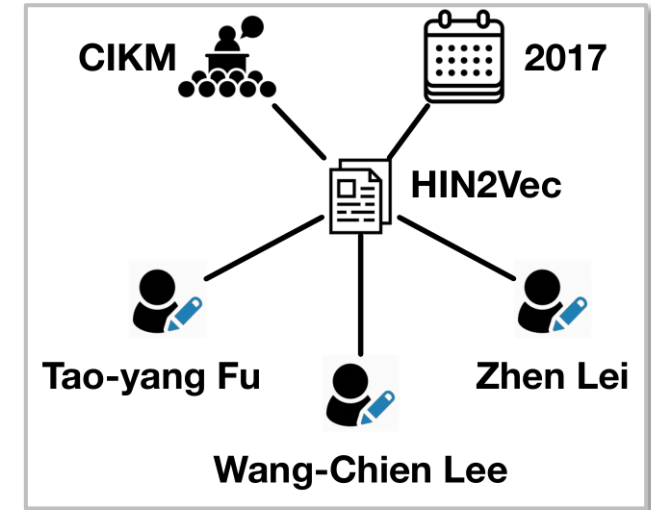
You
What is the most cited paper in WWW 2017? Please list its title and authors.

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

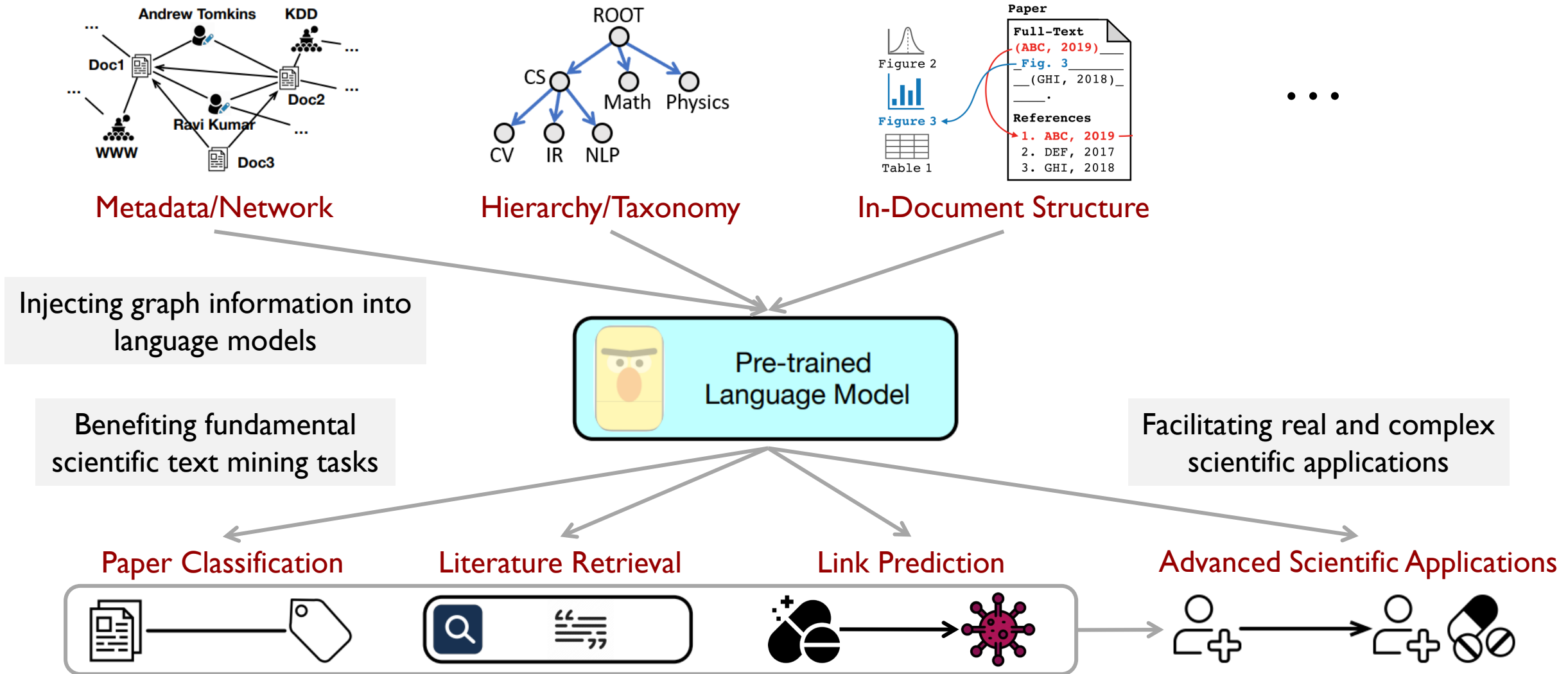
Please verify the latest citation counts and updates from reliable sources or academic databases for the most current information.



HIN2Vec was published in **CIKM 2017**.

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

Today's Talk: Overview



Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

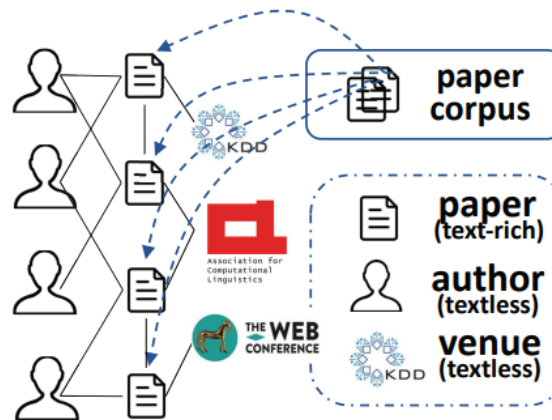
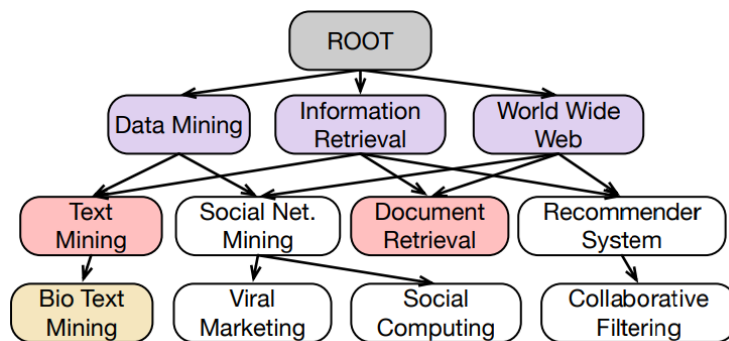
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

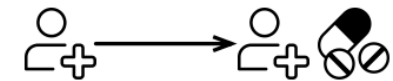
Jin, Zhang, Meng, & Han
ICLR 2023
Jin, Zhang, Zhu, & Han
KDD 2023

Part III: Advanced Scientific Applications

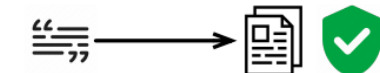
Zhang et al., EMNLP 2023
Zhang et al., arXiv 2023



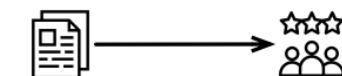
Patient-to-Patient Matching



Claim Verification



Peer Review Assignment



Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

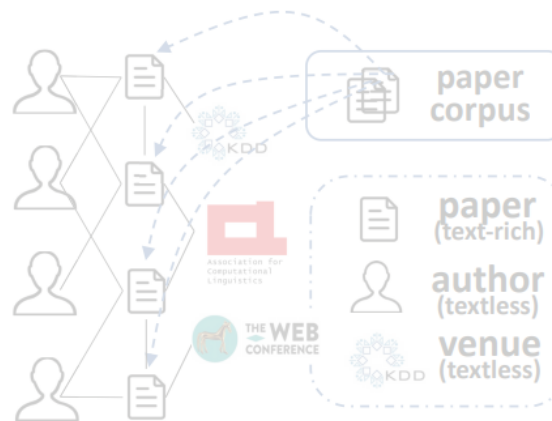
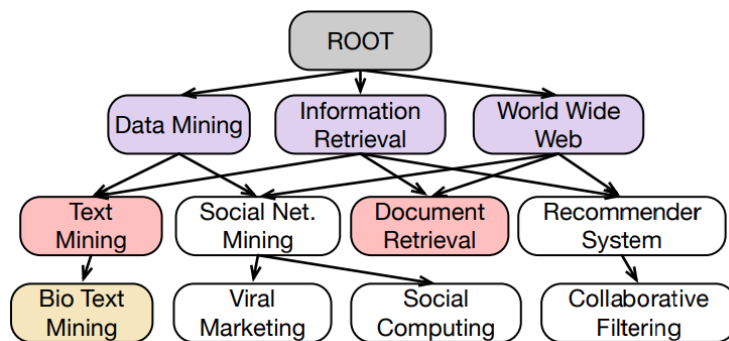
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

Jin, Zhang, Meng, & Han
ICLR 2023
Jin, Zhang, Zhu, & Han
KDD 2023

Part III: Advanced Scientific Applications

Zhang et al., EMNLP 2023
Zhang et al., arXiv 2023



Patient-to-Patient Matching



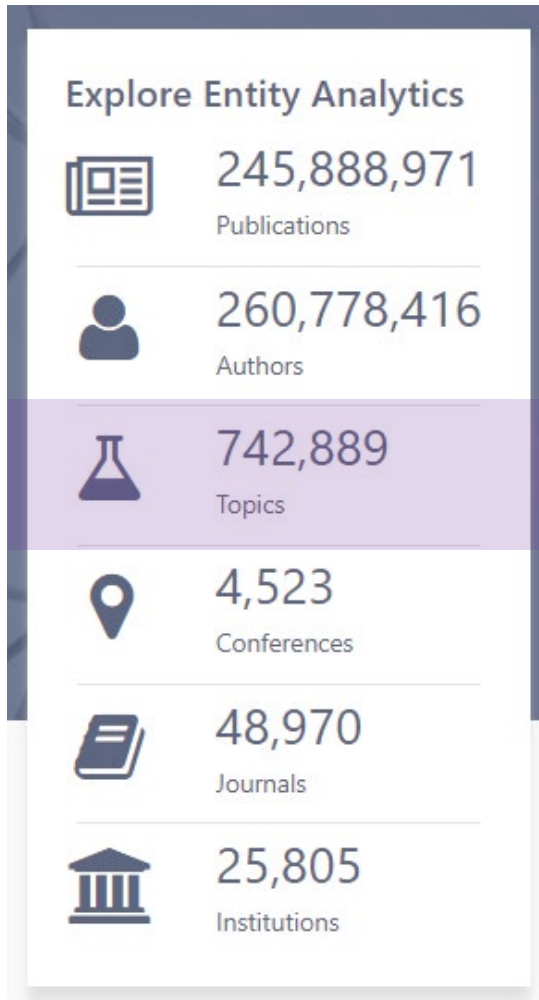
Claim Verification




Peer Review Assignment



Extremely Fine-Grained Scientific Paper Classification



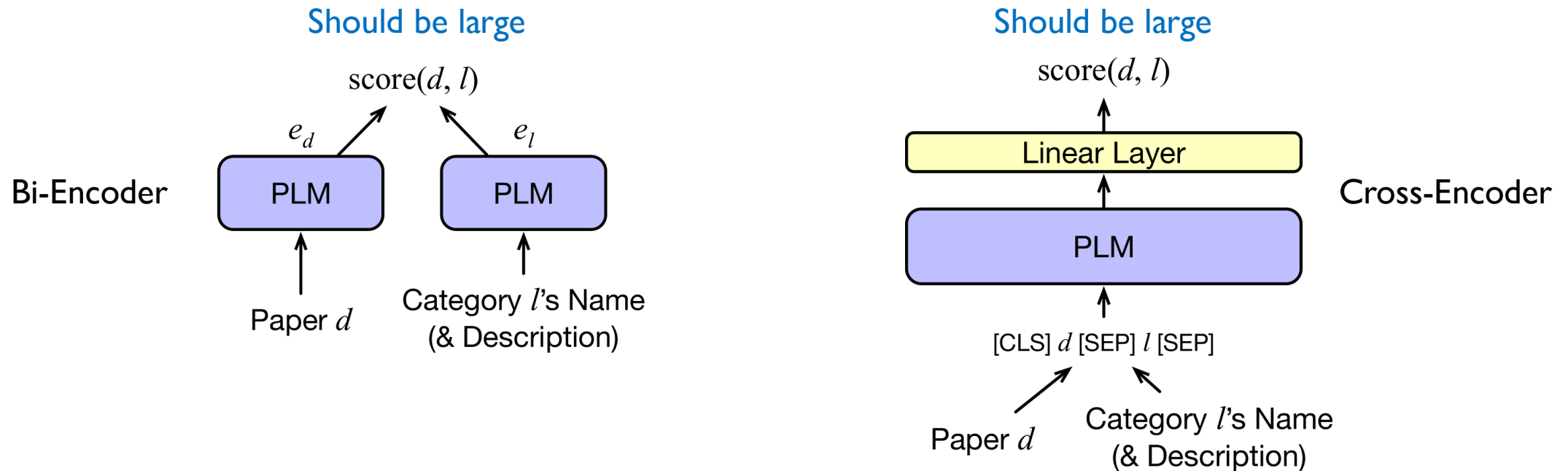
- The Microsoft Academic Graph has **740K+** categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain **30K+** categories.
- Each paper can be relevant to **more than one** category (5-15 categories for most papers).

 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, ...

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.



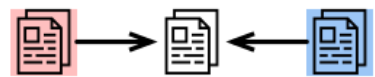
- However, human-annotated training samples are **NOT available** in many cases!
 - We are asking annotators to find ~10 relevant categories from ~100,000 candidates!

Using Graph 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



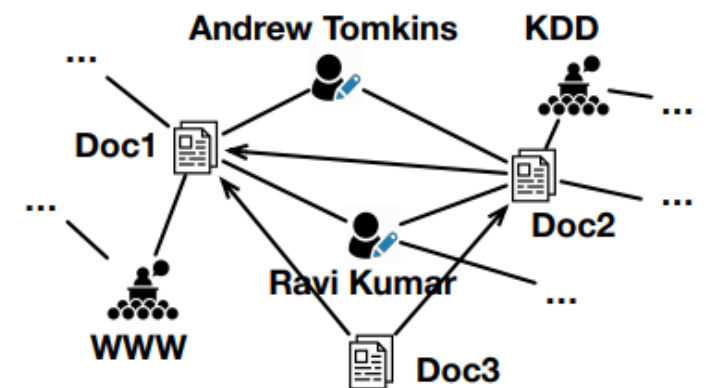
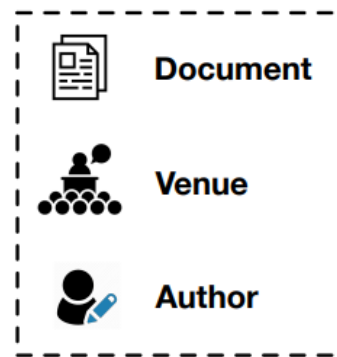
(b) meta-path: P->P<-P



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



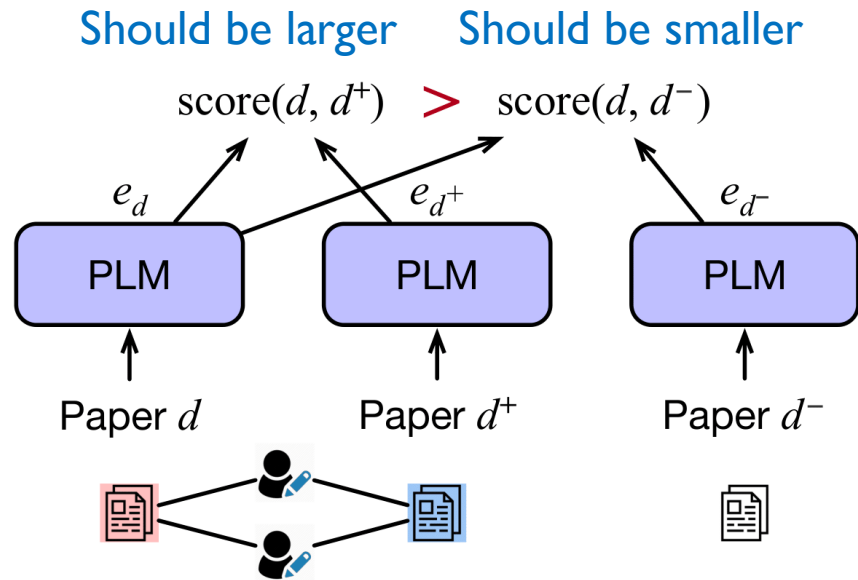
(d) meta-graph: P<- (PP) -> P



Graph-Induced Text Contrastive Learning

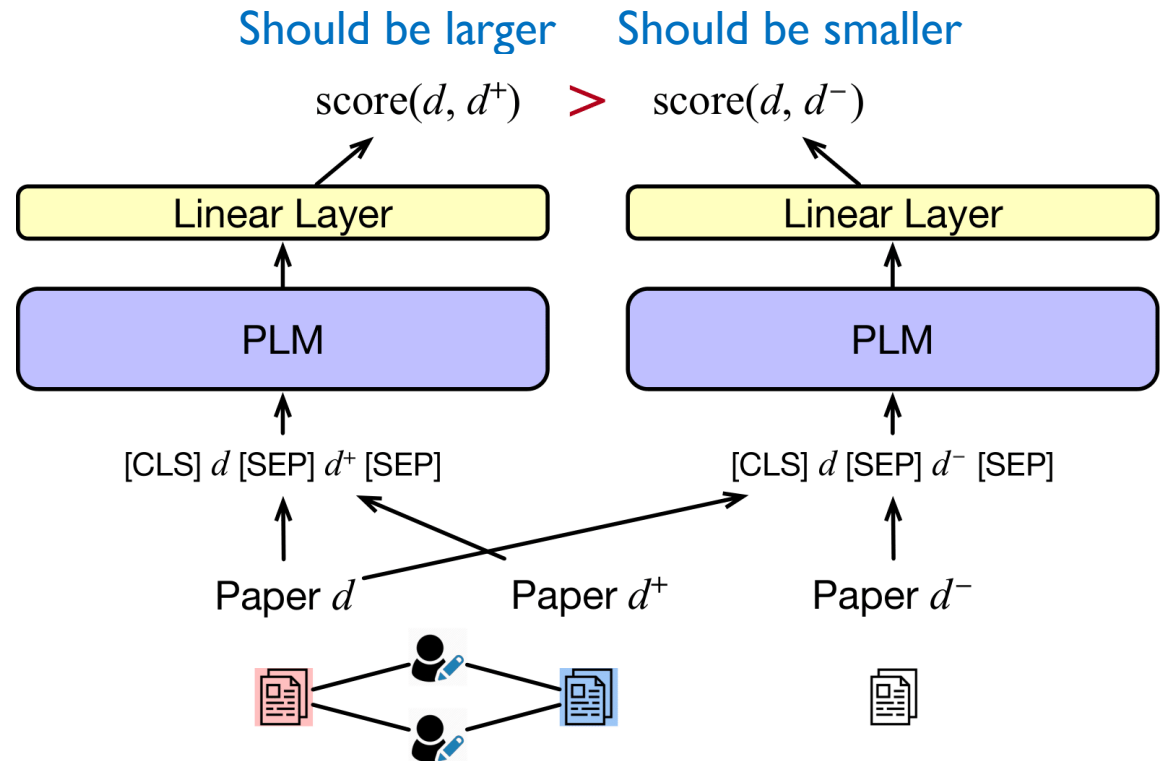
- Two papers connected via a certain meta-path/meta-graph should be more similar than two randomly selected papers.

Bi-Encoder



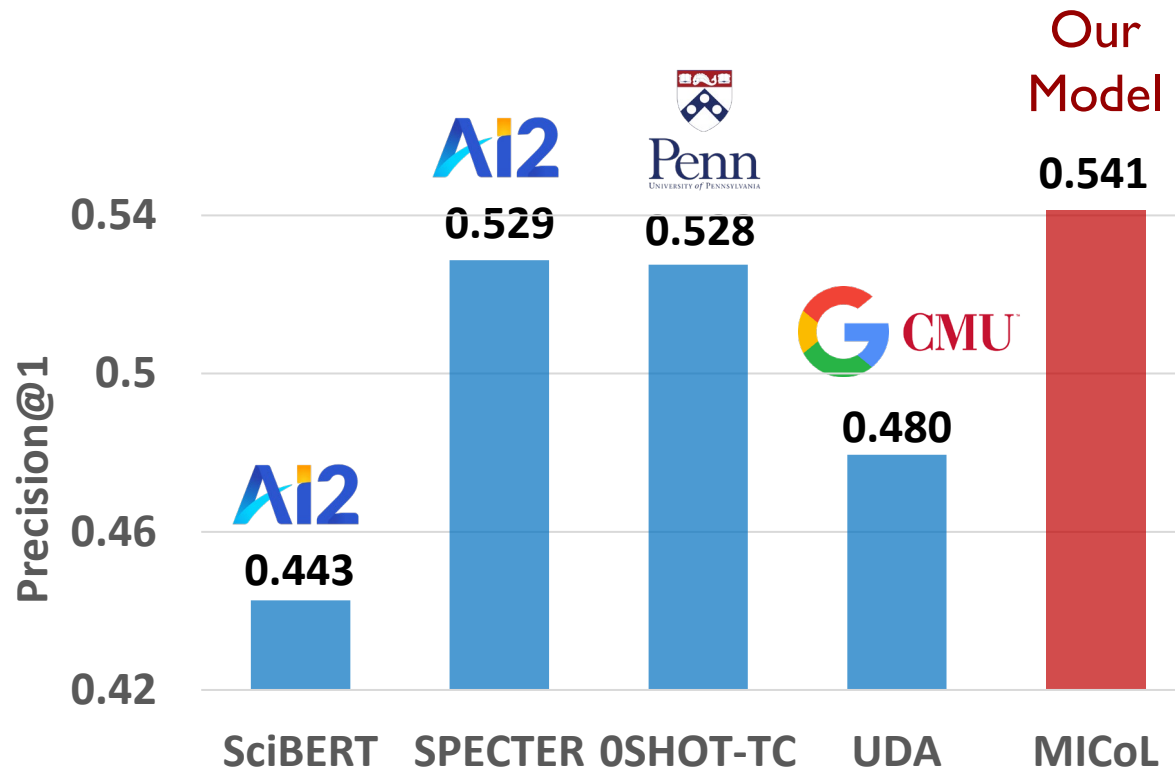
$$-\log \frac{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau)}{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau) + \sum_{i=1}^N \exp(\cos(\mathbf{e}_d, \mathbf{e}_{d_i^-})/\tau)}$$

Cross-Encoder



Comparison with Previous Approaches

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



Case Study

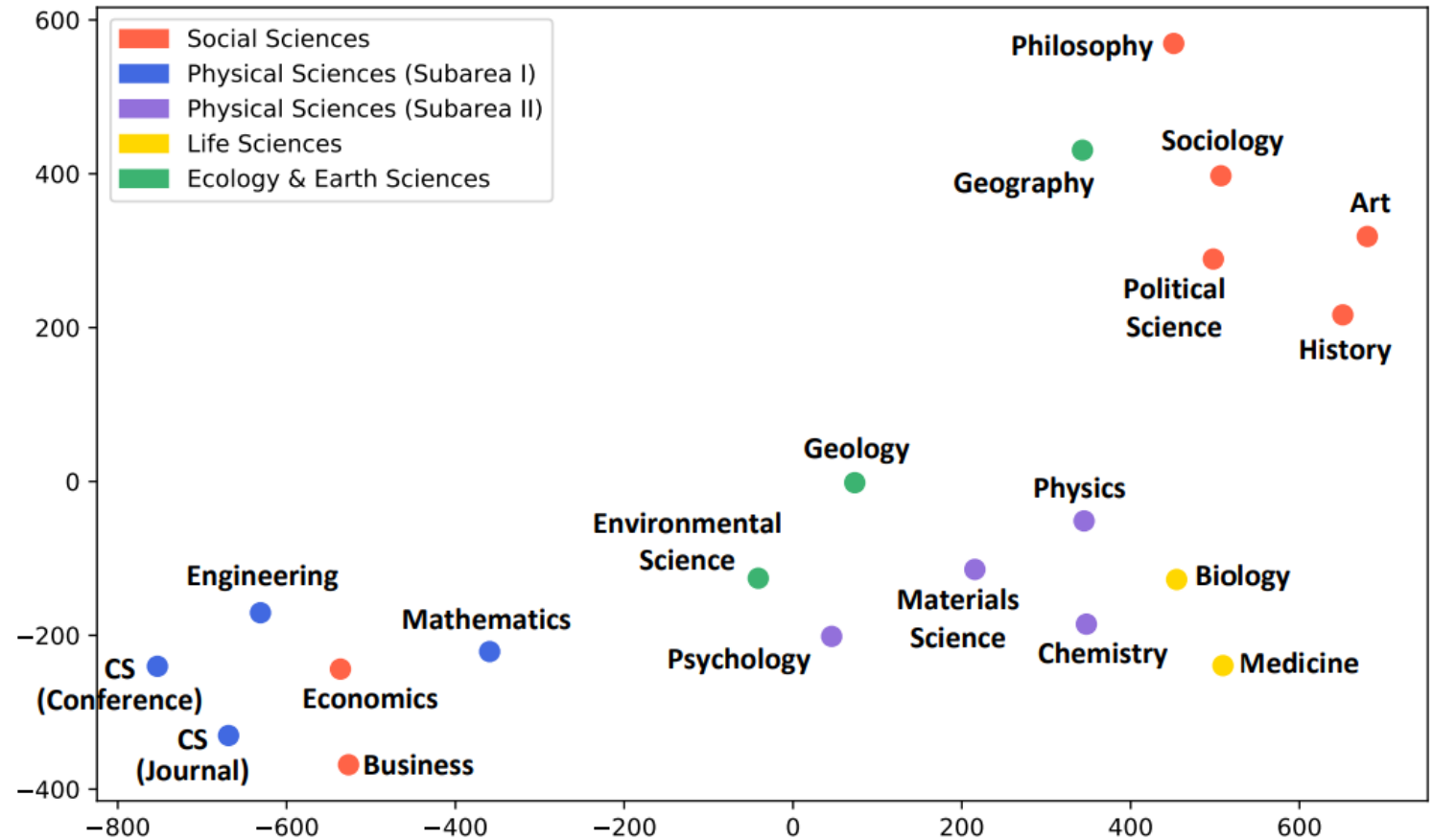
- Title: Improving Text Categorization Methods for Event Tracking
- Venue: [SIGIR](#) (2000)
- Authors: [Yiming Yang](#), Tom Ault, Thomas Pierce, Charles W. Lattimer
- Abstract: : Automated tracking of events from chronologically ordered document streams is a new challenge for statistical text classification. 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 (X)**

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

Which type of nodes is the most helpful?

- Is the contribution of venues, authors, and references to paper classification consistent **across different fields?**
 - NO! BUT the effects of metadata tend to be similar in two similar fields.
 - The experience of using metadata in one field can be **extrapolated** to a similar field.








The MAPLE Benchmark

- Our multi-field scientific literature tagging benchmark has been **downloaded 160 times** since it was published in February 2023.



Published February 6, 2023 | Version v1

[Dataset](#) [Open](#)

The MAPLE Benchmark for Scientific Literature Tagging

Zhang, Yu¹ ; Jin, Bowen¹ ; Zhu, Qi¹ ; Meng, Yu¹ ; Han, Jiawei¹ 

[Show affiliations](#)

437  VIEWS 160  DOWNLOADS






[Show more details](#)





Published February 6, 2023 | Version v2

[Dataset](#) [Open](#)

The MAPLE Benchmark for Graph Mining

Zhang, Yu¹ ; Jin, Bowen¹ ; Zhu, Qi¹ ; Meng, Yu¹ ; Han, Jiawei¹ 

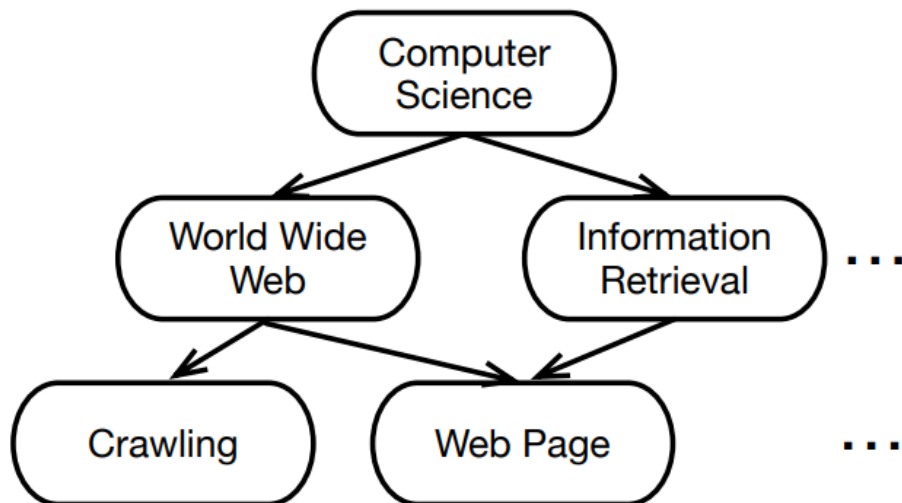
[Show affiliations](#)

209  VIEWS 29  DOWNLOADS

[Show more details](#)

How about other types of graph information?

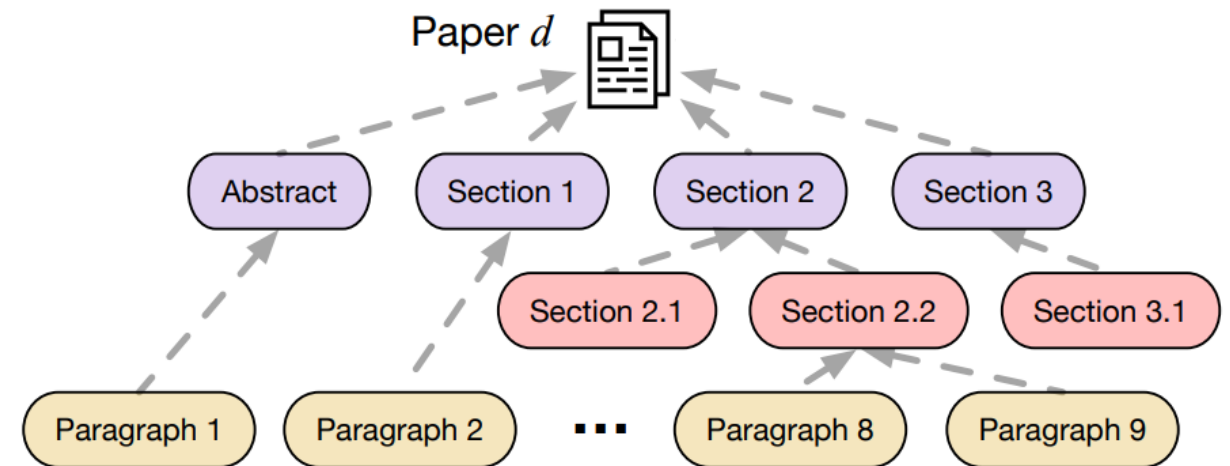
Label Hierarchy



Top-Down Pruning:

Irrelevant to **WWW** \Rightarrow Irrelevant to **Crawling**

In-Document Structure



Bottom-Up Aggregation:

Paragraphs \rightarrow **Subsections** \rightarrow **Sections** \rightarrow **Paper**

Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

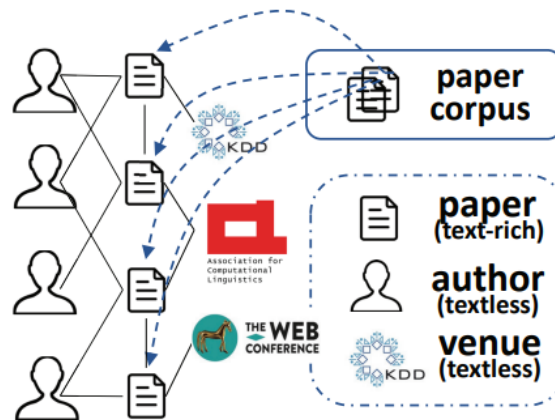
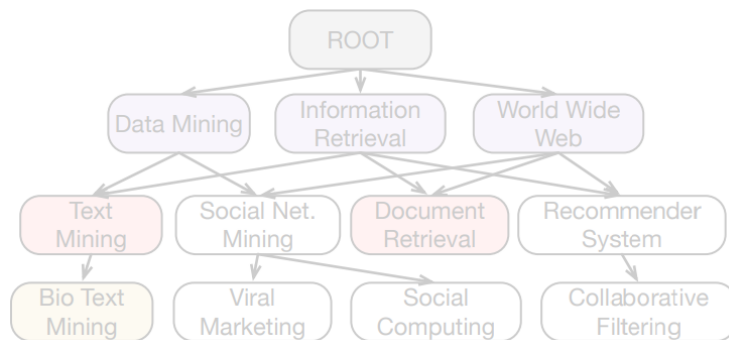
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

Jin, Zhang, Meng, & Han
ICLR 2023
Jin, Zhang, Zhu, & Han
KDD 2023

Part III: Advanced Scientific Applications

Zhang et al., EMNLP 2023
Zhang et al., arXiv 2023



Patient-to-Patient Matching



Claim Verification

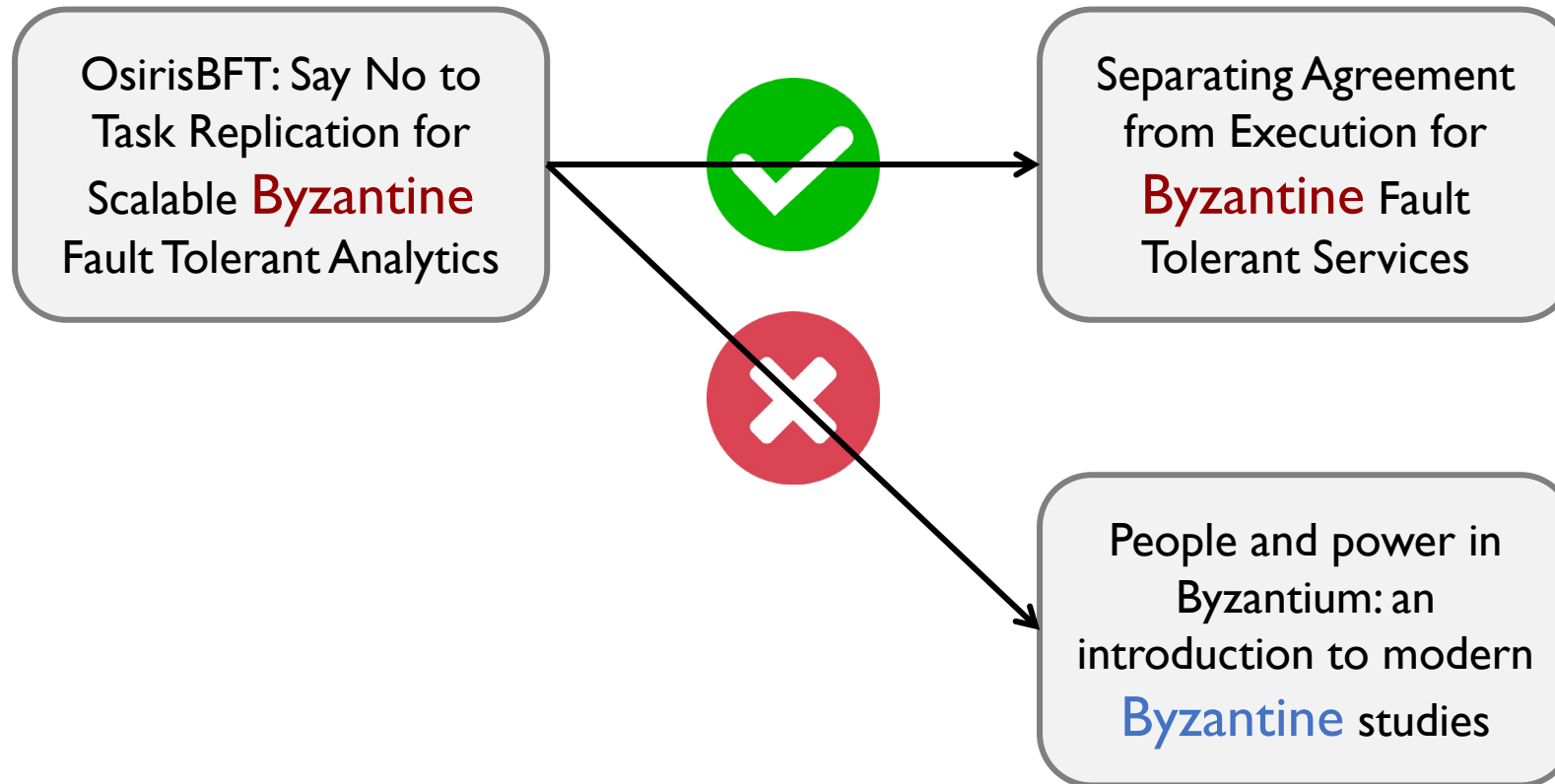


Peer Review Assignment



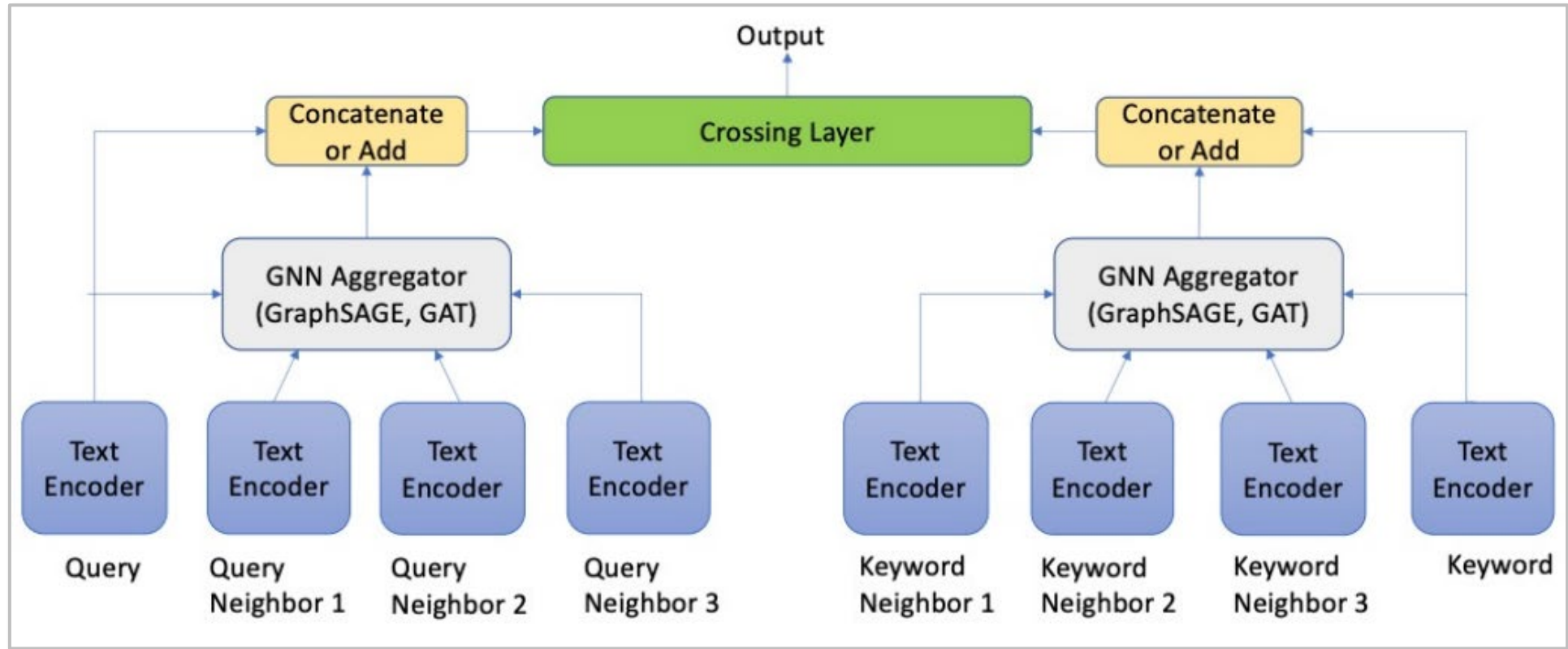
Text complements graph signals in link prediction, but ...

- We need contextualized text representations rather than bag of words!



PLM+GNN: Cascaded Architecture

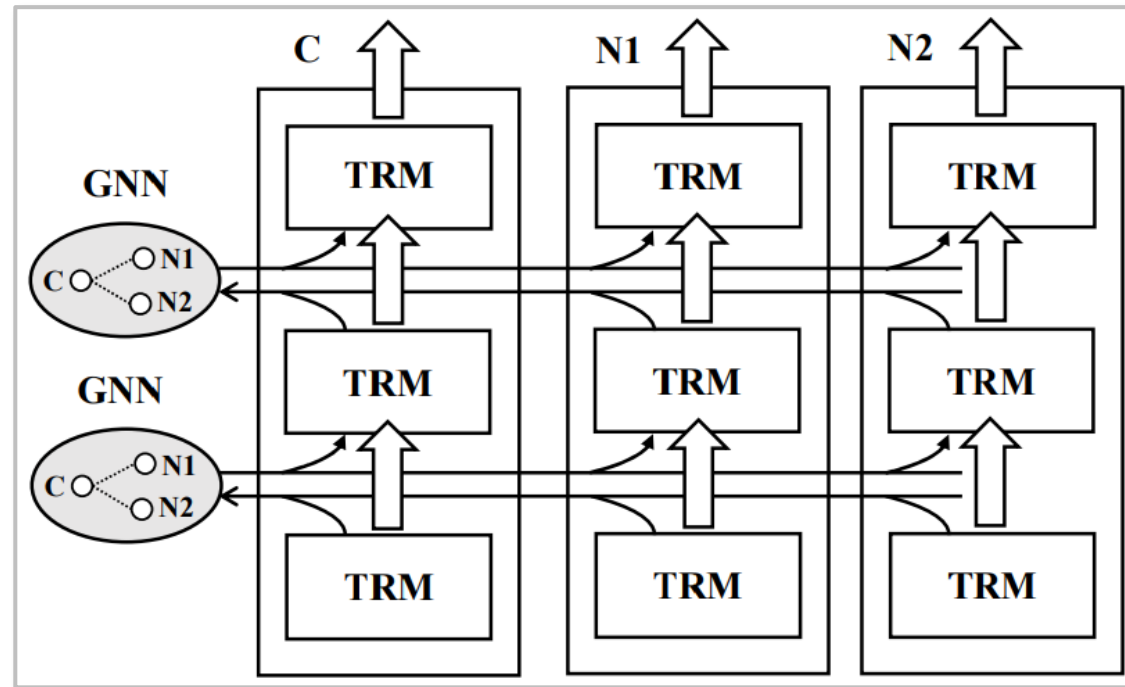
- PLM (text encoding) → GNN (graph aggregation)



- **Drawback:** Graph information is not used when encoding text.

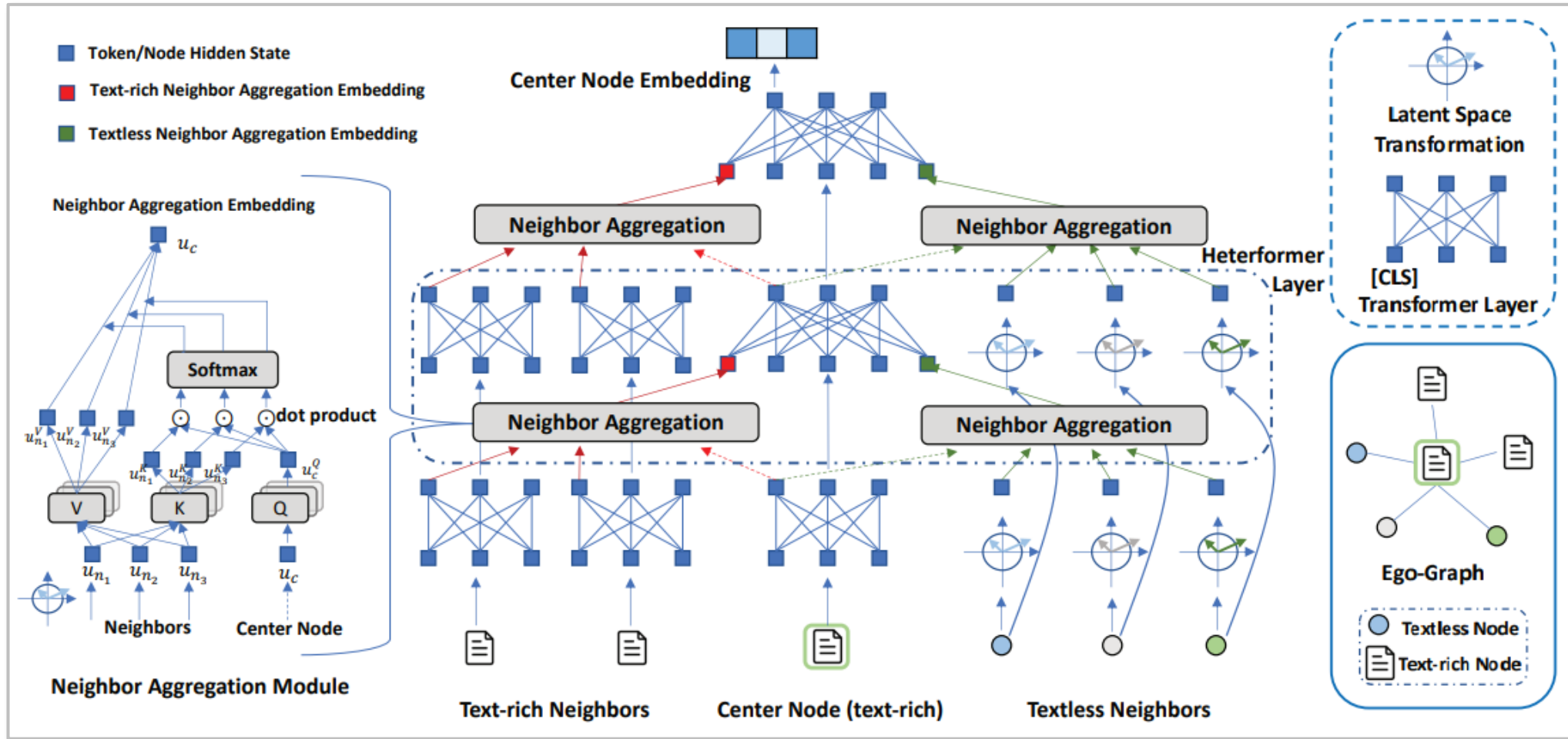
PLM+GNN: Interleaved Architecture

- Cascaded Architecture:
 - Transformer → Transformer → ... → Transformer → GNN
- Interleaved Architecture:
 - Transformer → GNN → Transformer → GNN → ... → Transformer → GNN



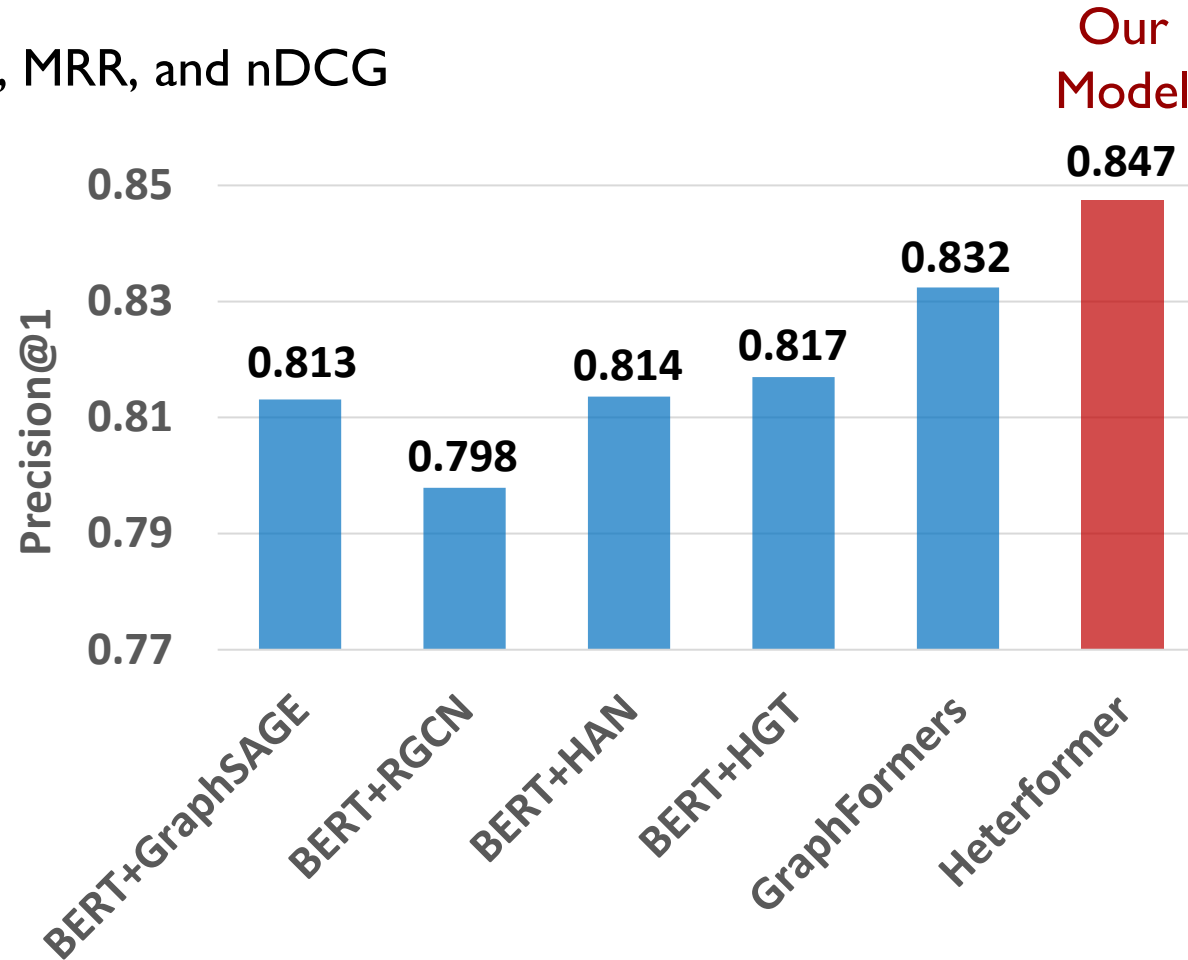
Dealing with Heterogeneity

- Some types of nodes do not have text information!



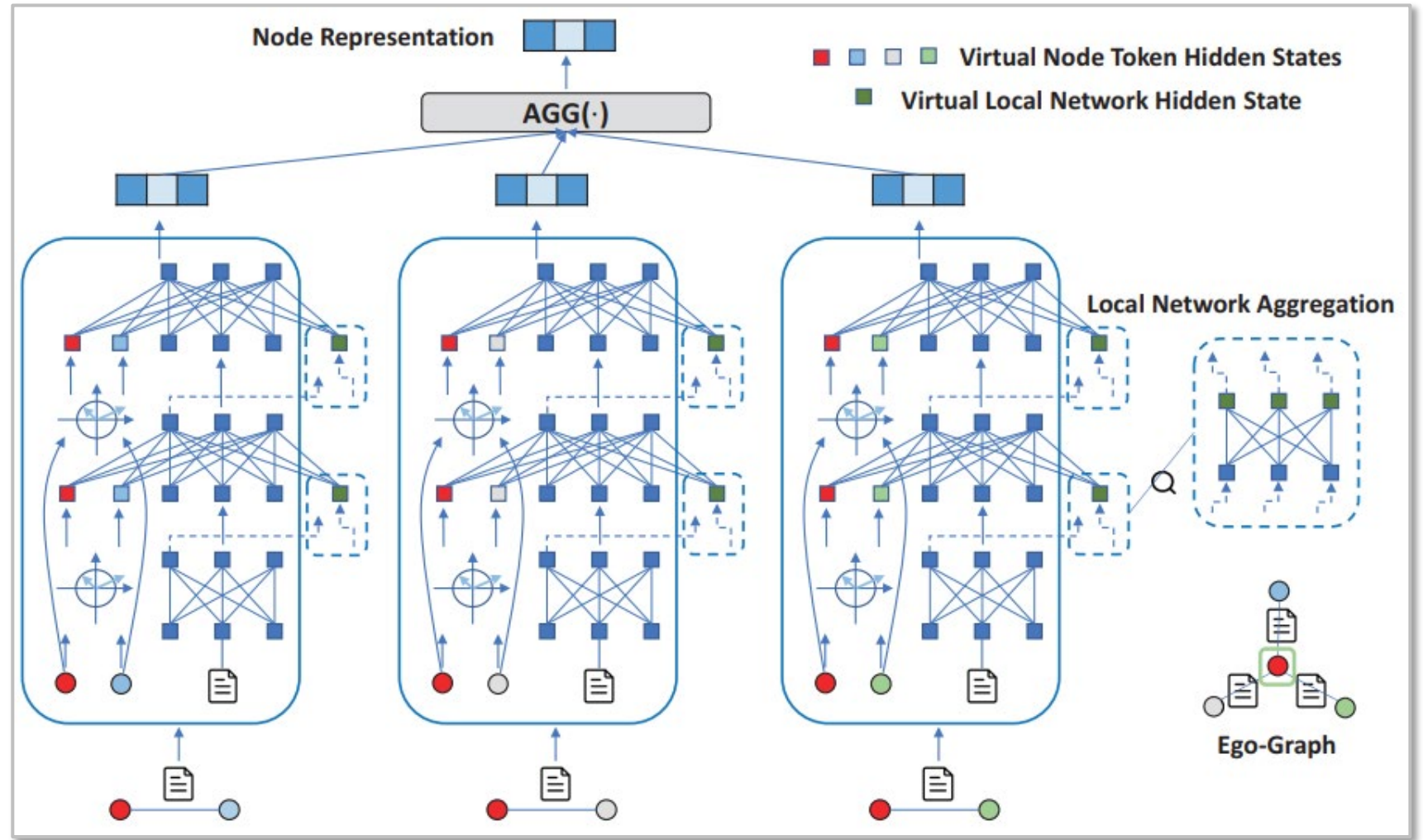
Comparison with Previous Approaches

- Dataset: DBLP
- Metric: Precision@1, MRR, and nDCG



Text Information on Edges

- One **paper** cites the other **paper** in a **sentence**.
- A **user** write a **review** for an **item**.



Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

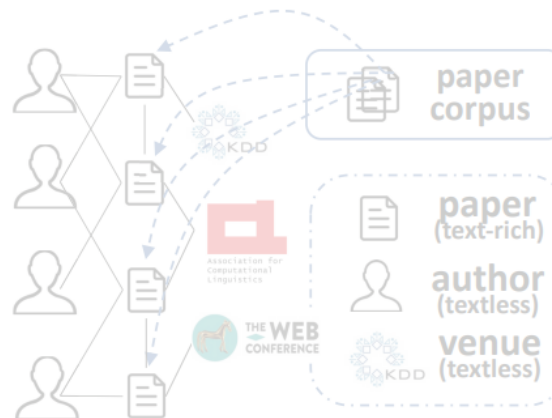
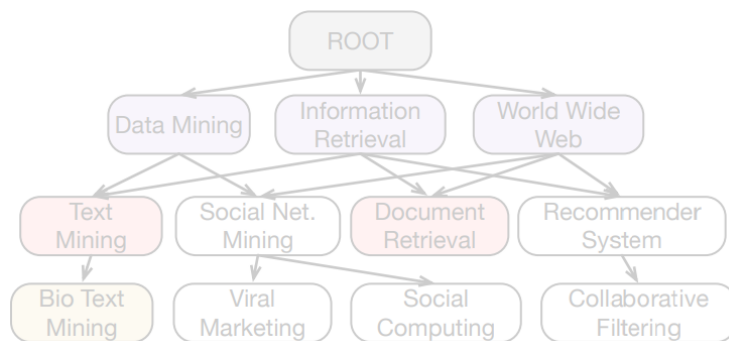
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

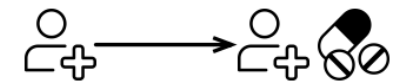
Jin, Zhang, Meng, & Han
ICLR 2023
Jin, Zhang, Zhu, & Han
KDD 2023

Part III: Advanced Scientific Applications

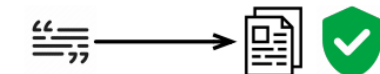
Zhang et al., EMNLP 2023
Zhang et al., arXiv 2023



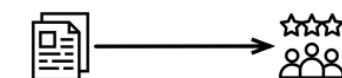
Patient-to-Patient Matching



Claim Verification



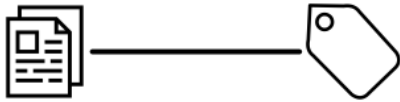
Peer Review Assignment



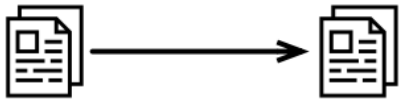
Facilitating Complex Tasks for Scientific Discovery

Fundamental Scientific Text Mining Tasks

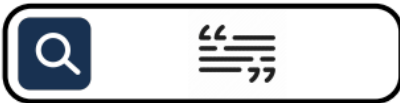
Paper Classification



Link Prediction

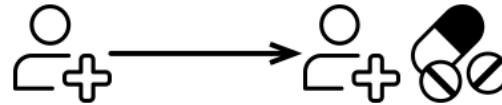


Literature Retrieval

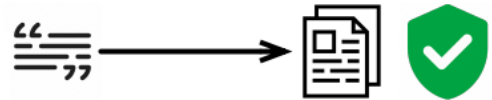


Advanced Applications for Scientific Discovery

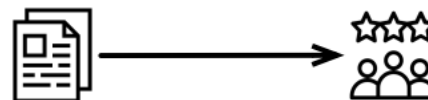
Patient-to-Patient Matching



Claim Verification



Peer Review Assignment



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

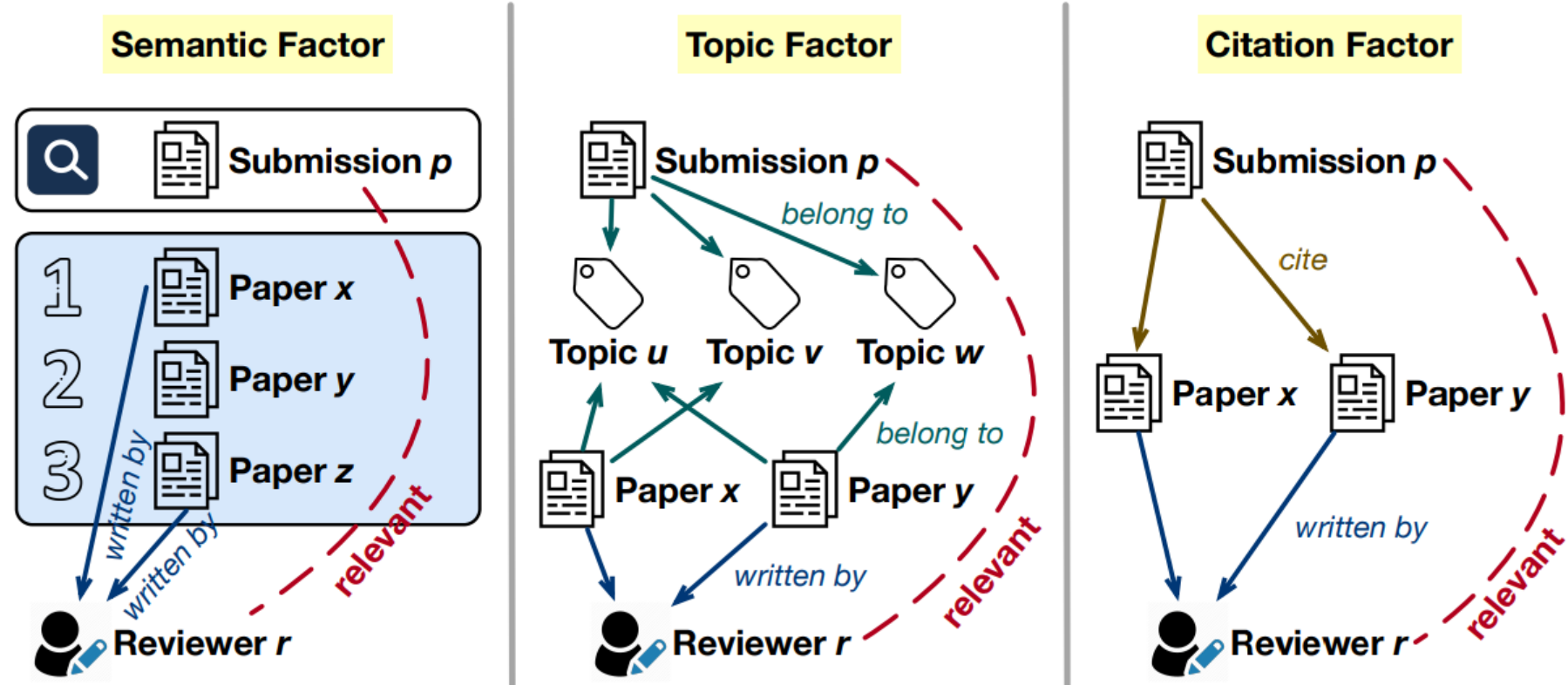
Given a scientific claim, find relevant papers (and predict their stance).

Given a paper submission, find expert reviewers.

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

Multiple Factors for Judging Relevance

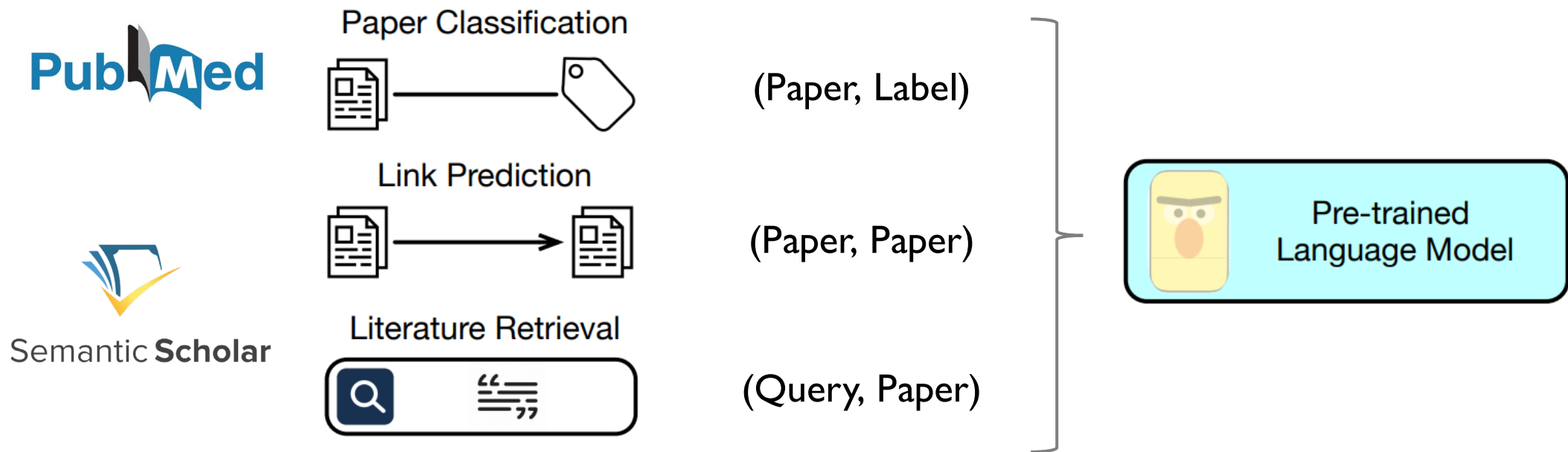
- Example: Paper-Reviewer Matching
 - Why is a pair of (Paper, Reviewer) **relevant**?



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

Naïve Multi-task Pre-training

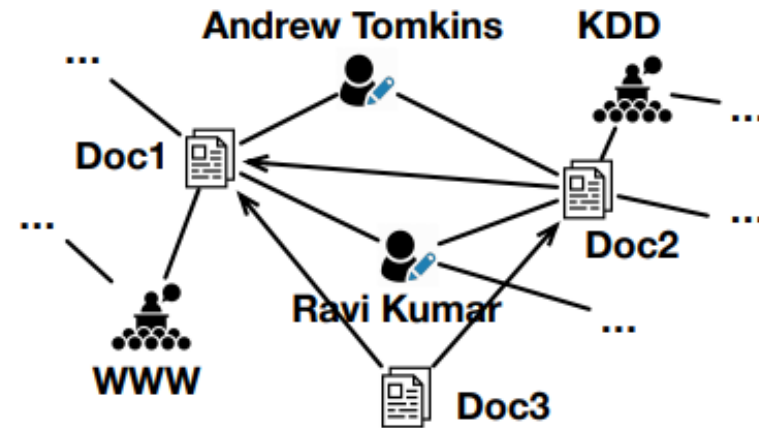
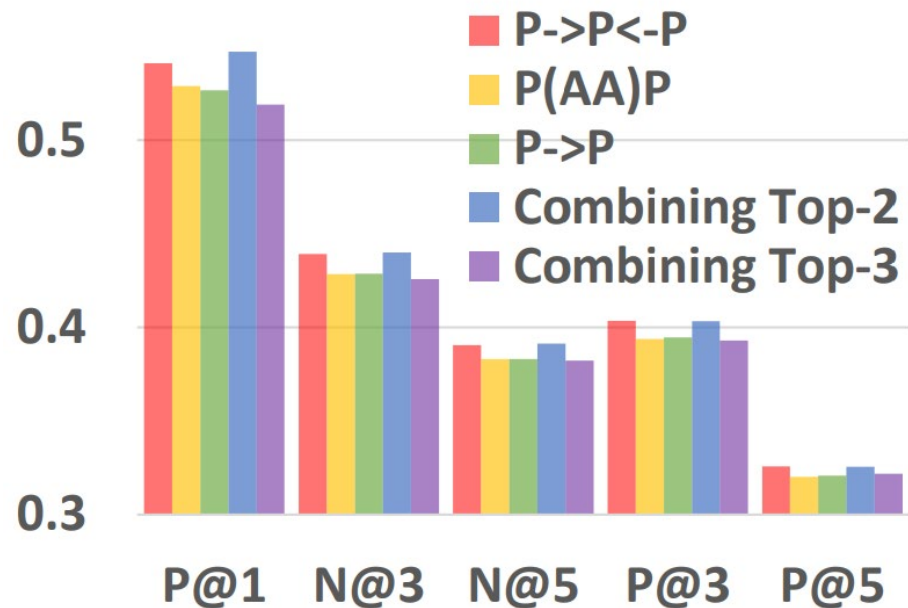
- Each factor (topic, citation, and semantic) relies on one **fundamental** text mining task.
- Directly combining pre-training data from different tasks to train a model?



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

An Illustrative Example of Task Interference

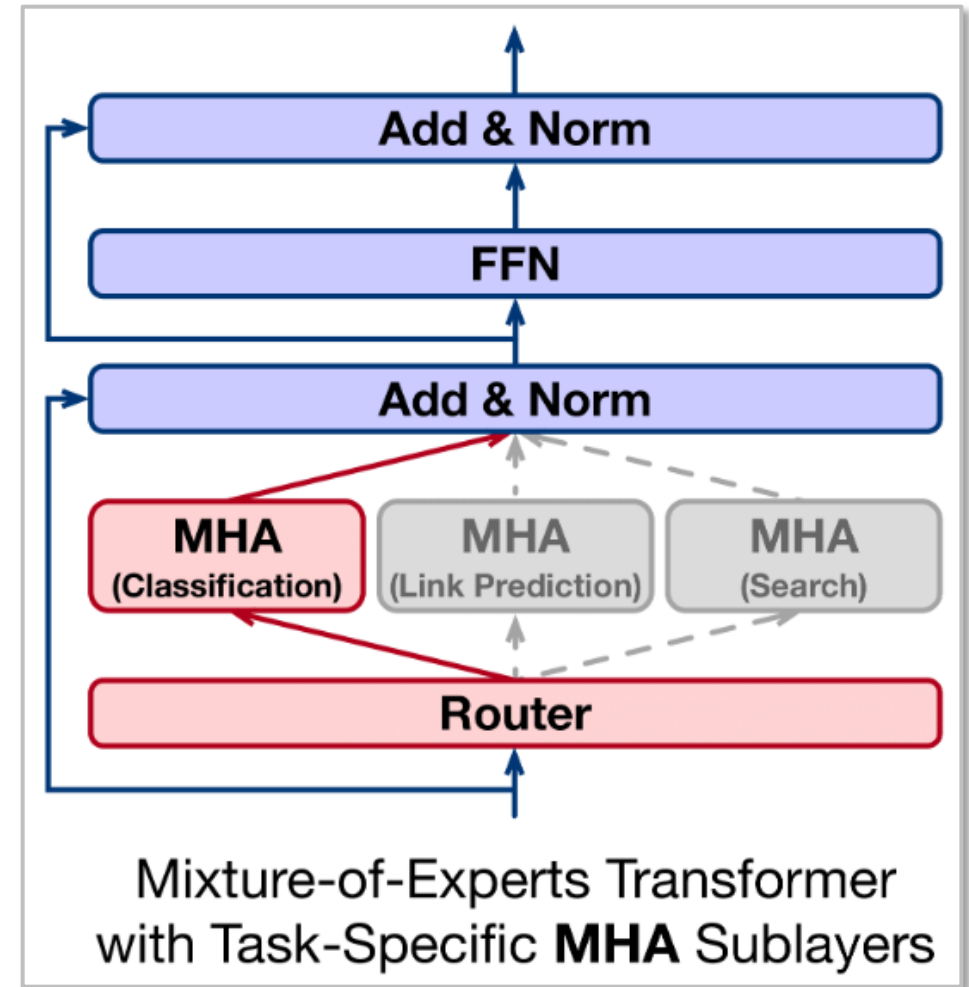
- Recall structure-induced contrastive learning
- Imagine each meta-path/meta-graph is a “task” (i.e., defines one type of “relevance”)
- Directly merging the relevant (paper, paper) pairs induced by different meta-paths for training?
 - **Cannot consistently improve the classification performance!**



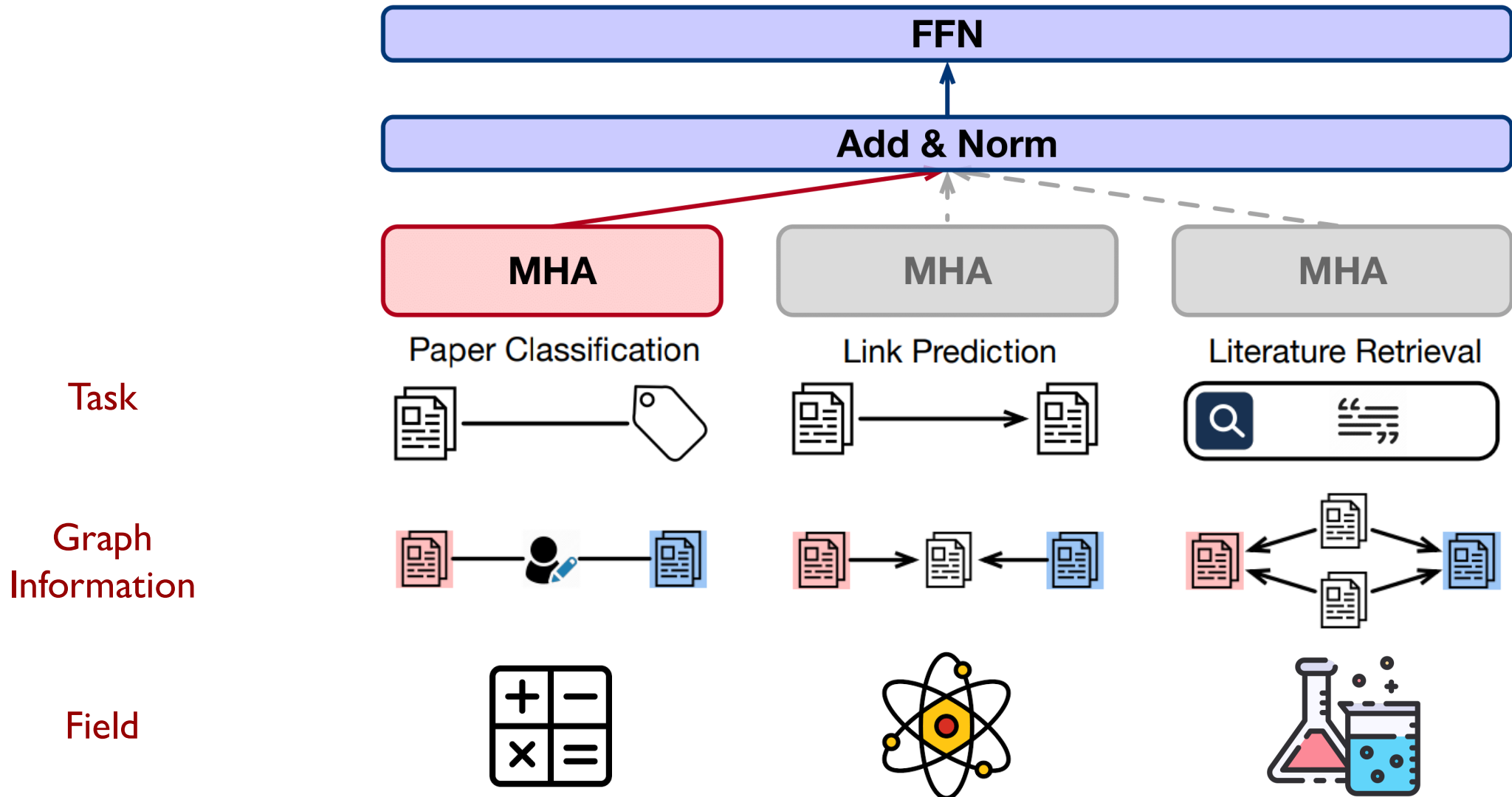
(Doc2, Doc3) are **relevant** according to $P \rightarrow P \leftarrow P$ but **irrelevant** according to $P(AA)P$.

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.



Tackling Task Interference: Mixture-of-Experts Transformer

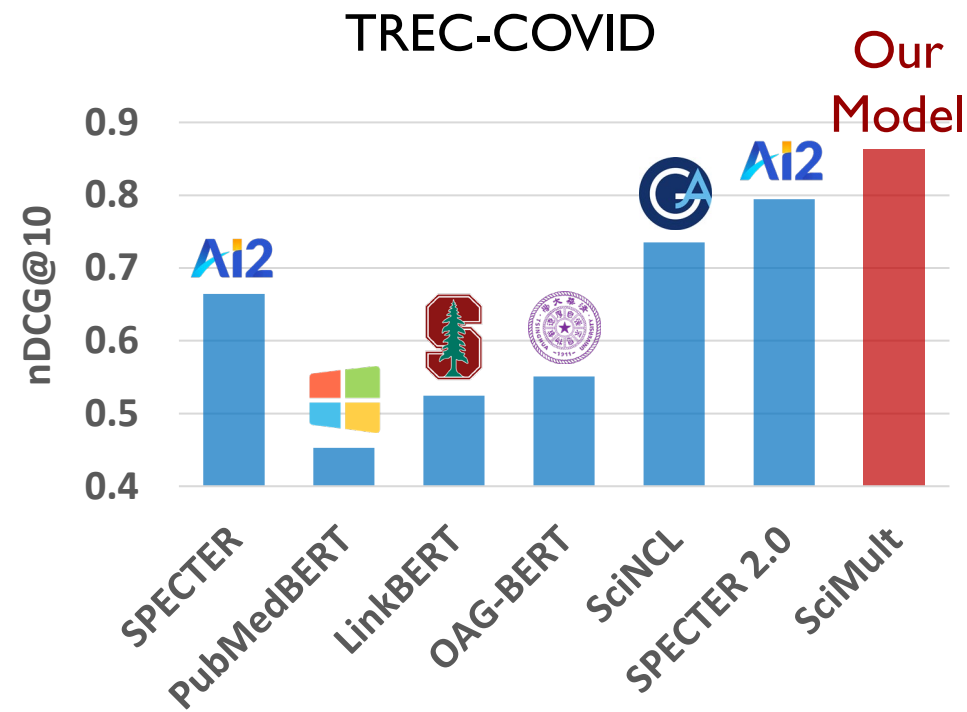


Comparison with Previous Approaches

- 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**)

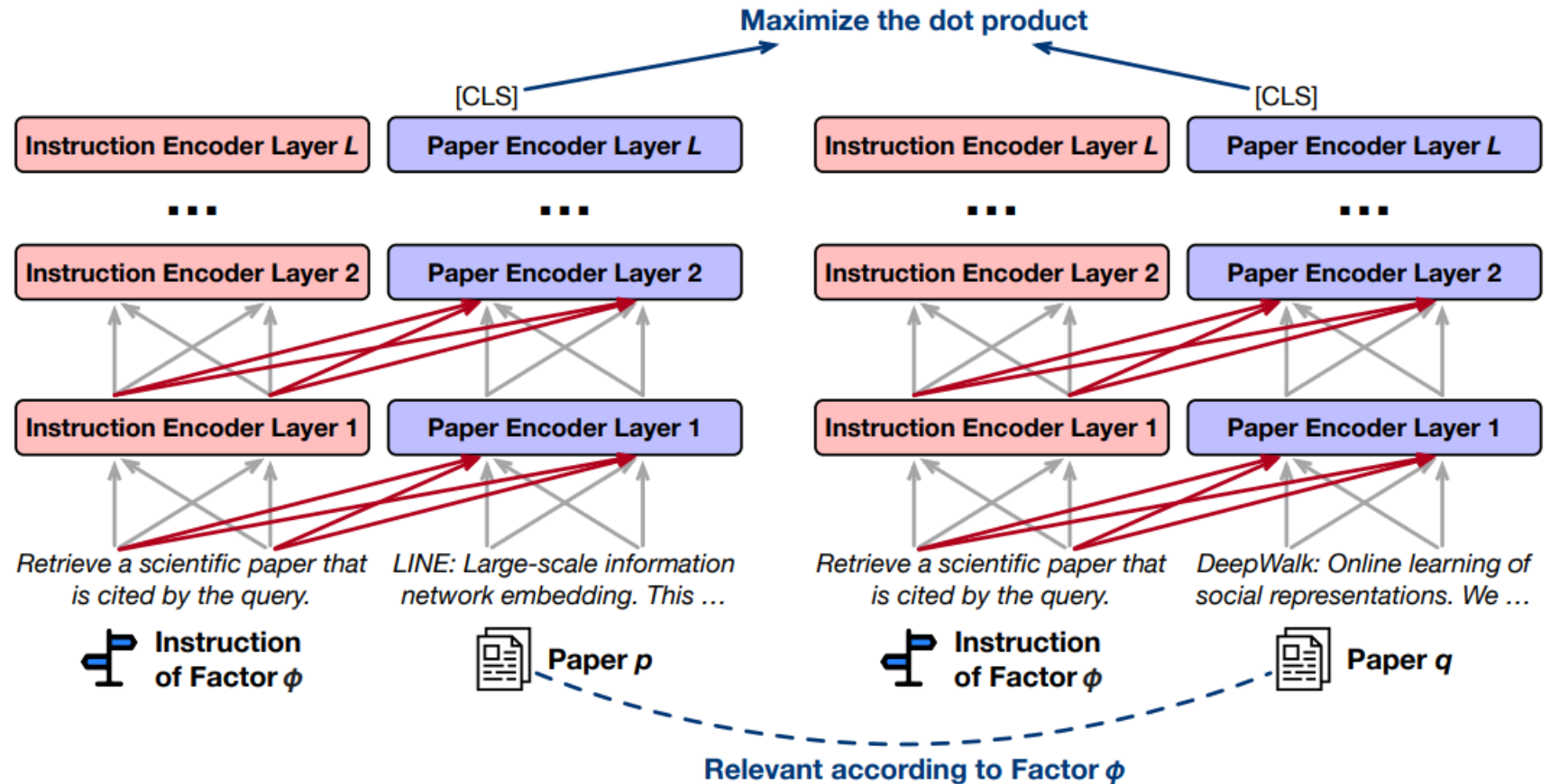
	Model	MRR (%)	P@10 (%)	nDCG@10 (%)	R@1k (%)
Our Model 1 June 25, 2023	DPR (SciMult-MHAExpert) <i>UIUC/Microsoft</i> (Zhang et al. 2023)	29.89	9.35	13.79	53.71
2 Apr 5, 2023	RRF <i>Tsinghua University</i> (Zhao et al. 2023)	29.86	8.86	13.36	49.45

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



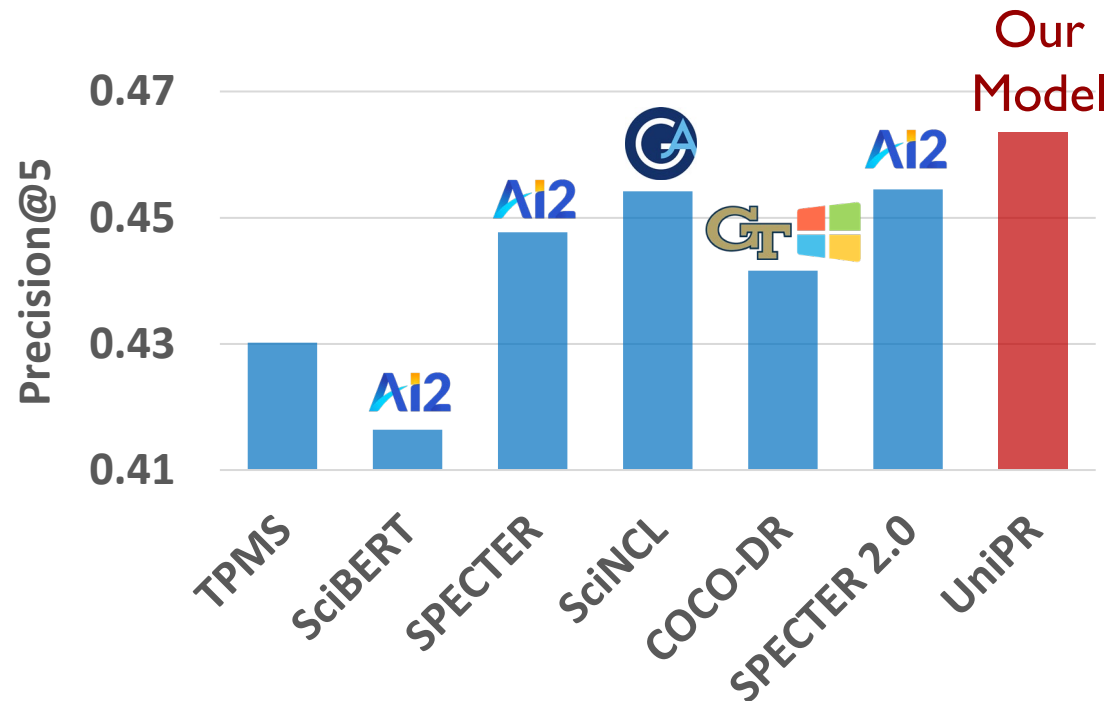
Tackling Task Interference: Instruction Tuning

- 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.




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)



Language Model on Graphs

 [Awesome-Language-Model-on-Graphs](#) Public

A curated list of papers and resources based on "Large Language Models on Graphs: A Comprehensive Survey".

☆ 384 🍷 22



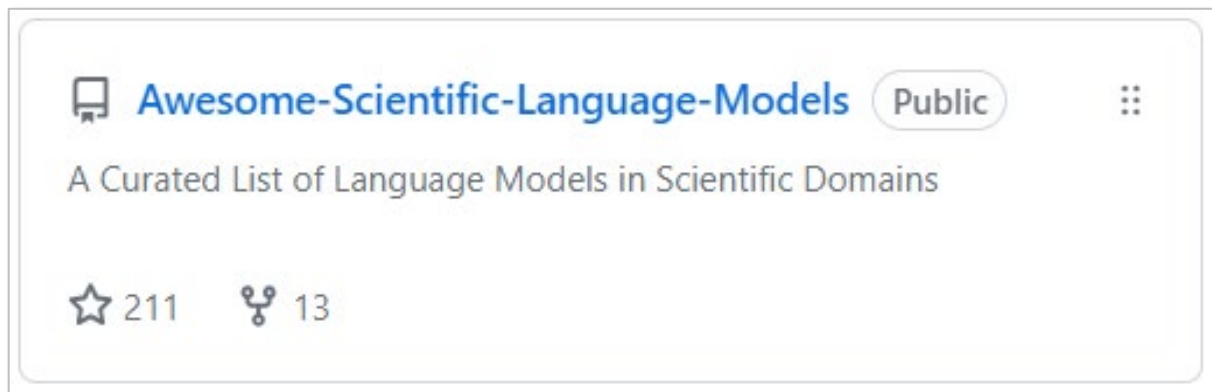
Awesome-Language-Model-on-Graphs

A curated list of papers and resources about large language models (LLMs) on graphs based on our survey paper: [Large Language Models on Graphs: A Comprehensive Survey](#).

This repo will be continuously updated. Don't forget to star ★ it and keep tuned!

Please cite the paper in [Citations](#) if you find the resource helpful for your research. Thanks!

Scientific Language Models



Awesome Scientific Language Models

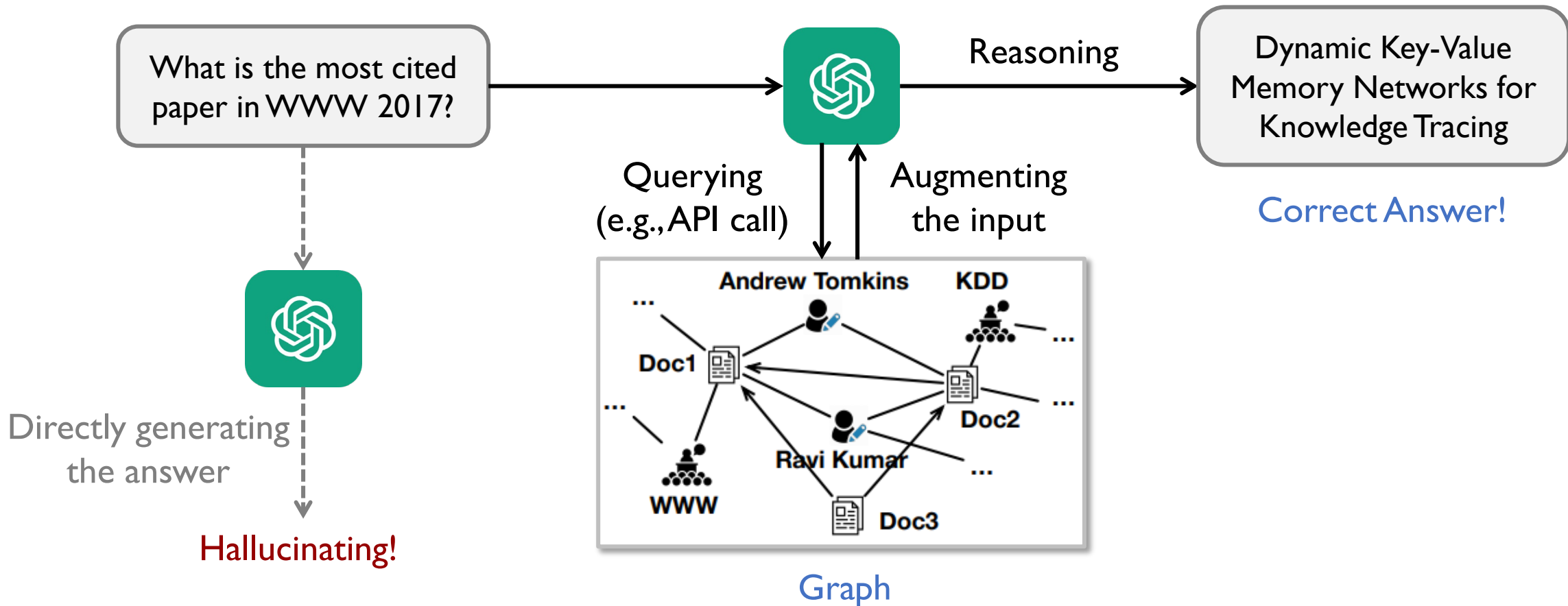
 awesome  Stars 211

PaperNumber 192 License MIT PRs Welcome

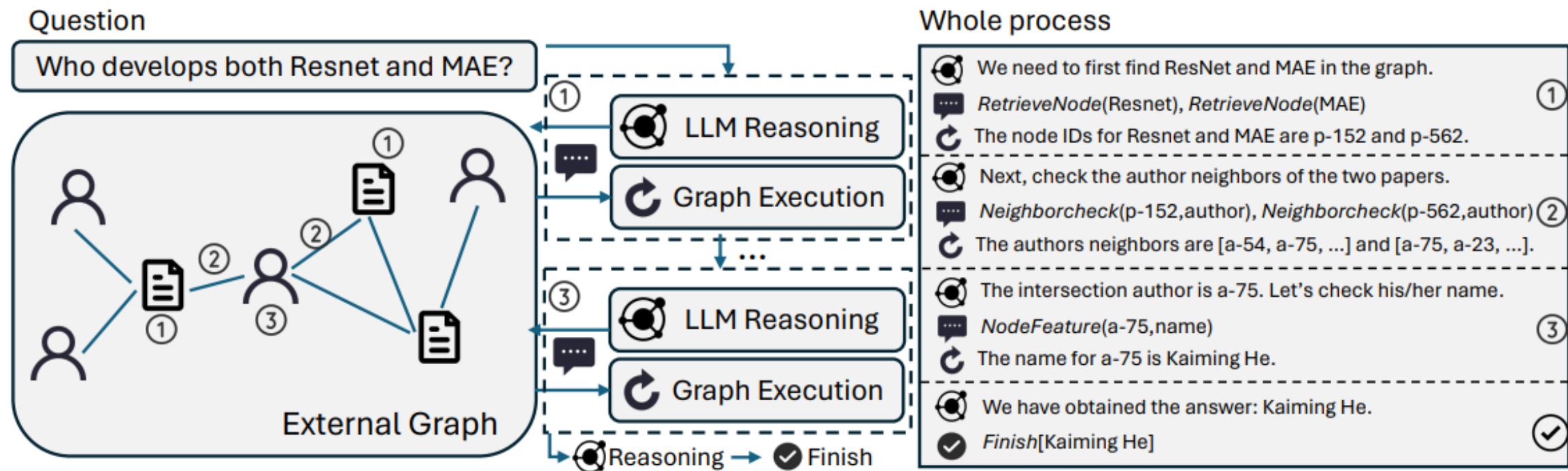
A curated list of pre-trained language models in scientific domains (e.g., mathematics, physics, chemistry, biology, medicine, materials science, and geoscience), covering different model sizes (from <100M to 70B parameters) and modalities (e.g., language, vision, molecule, protein, graph, and table). The repository will be continuously updated.

Looking Back to the Motivating Example

- Can we teach LLMs to explore graphs as **environments** / use graphs as **tools**?



Initial Trial: Graph Chain-of-Thoughts



Model	Academic		E-commerce		Literature		Healthcare		Legal	
	EM	GPT4score	EM	GPT4score	EM	GPT4score	EM	GPT4score	EM	GPT4score
Graph RAG LLaMA-2-13b	22.01	22.97	12.48	20.00	9.25	20.00	2.97	4.81	17.98	17.22
Mixtral-8x7b	27.77	31.20	32.87	37.00	20.08	33.33	8.66	15.19	23.48	25.56
GPT-3.5-turbo	18.45	26.98	17.52	28.00	14.94	24.17	8.69	14.07	18.66	22.22
Our Model	31.89	33.48	42.40	44.50	41.59	46.25	22.33	28.89	30.52	28.33



Thank you! Questions?