Graph-Enhanced Scientific Text Mining

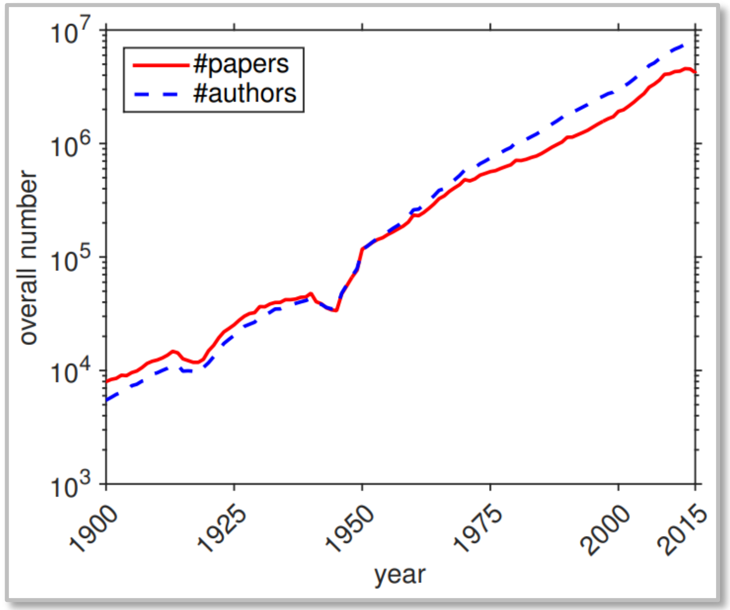
Yu Zhang

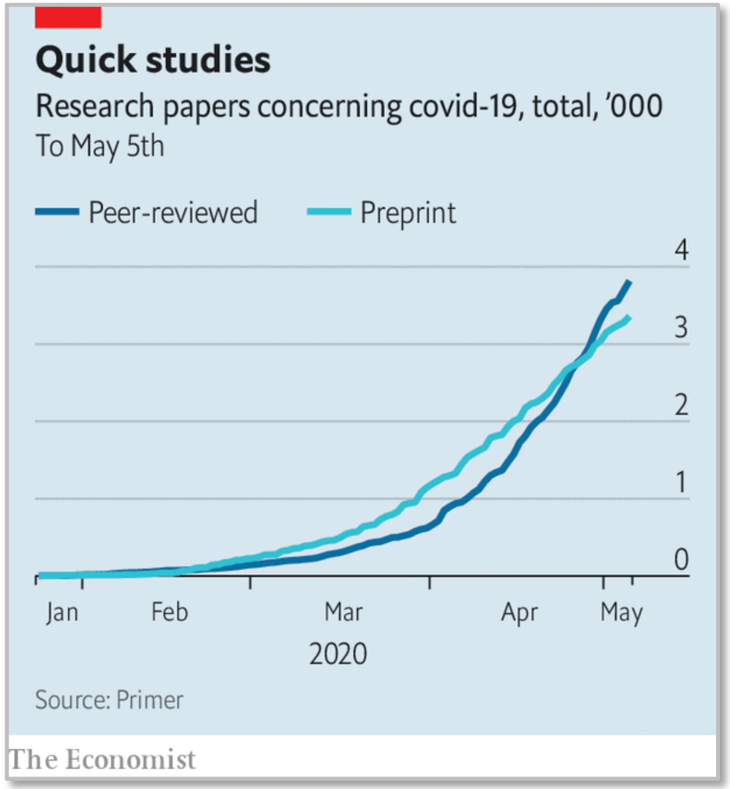
University of Illinois at Urbana-Champaign

May 15, 2024

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?

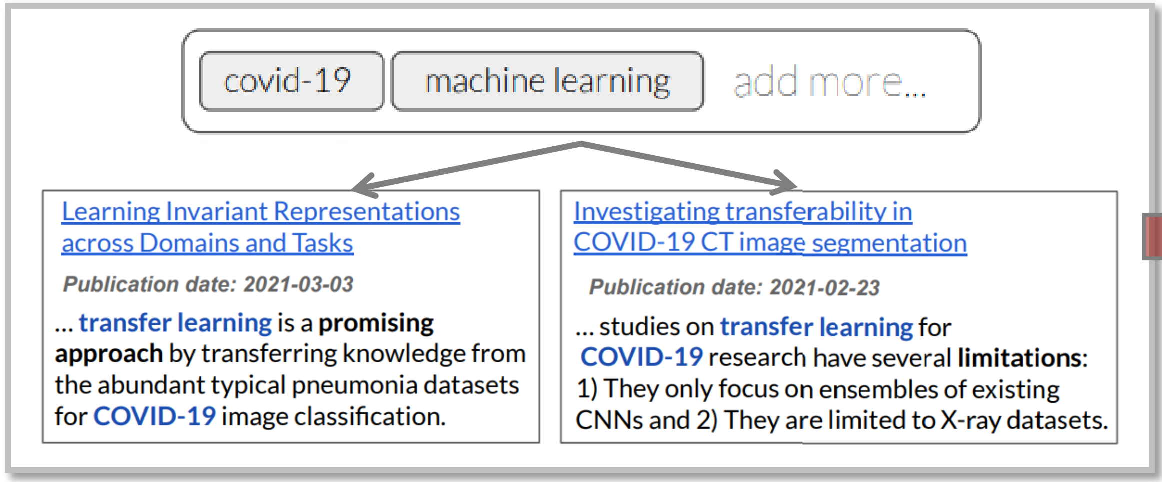




[1] "A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations." KDD 2017.
[2] "Microsoft Academic Graph: When Experts are Not Enough." Quantitative Science Studies 2020.
[3] https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent

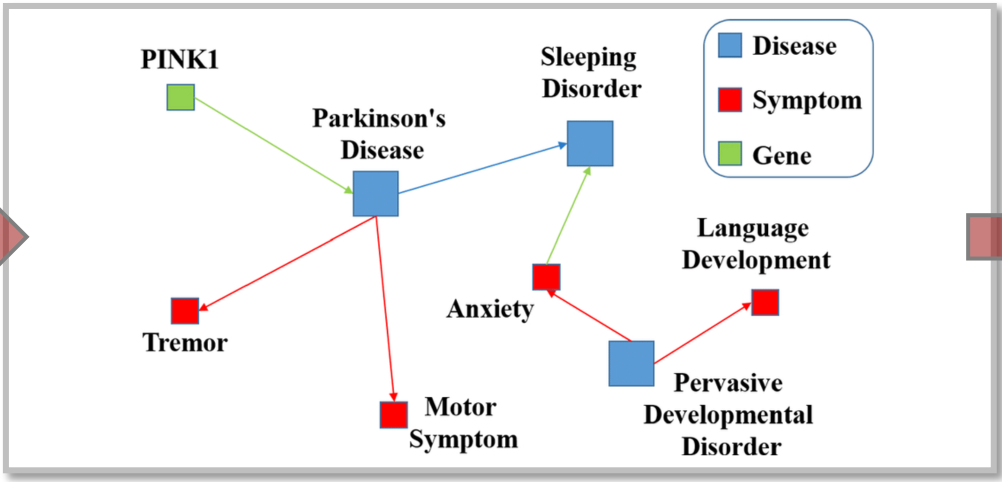
How can text mining help scientific discovery?

Retrieving and Analyzing Relevant Literature



- 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*".

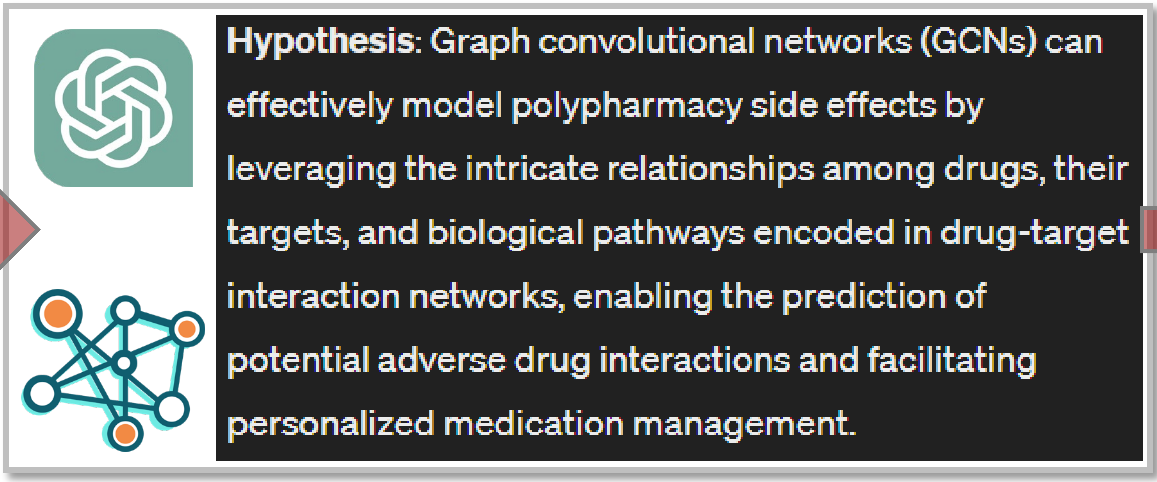
Uncovering Knowledge Structures



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

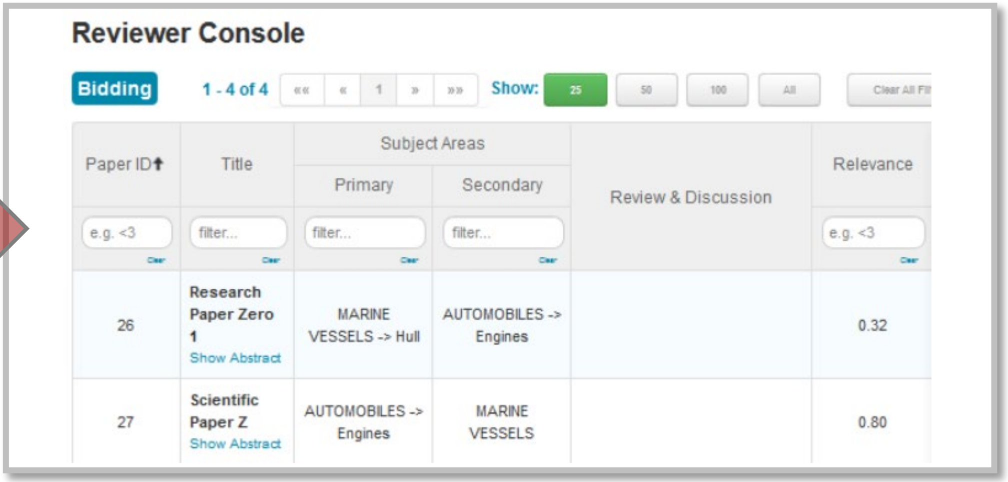
How can text mining help scientific discovery?

Generating Hypotheses and Suggesting Directions



- 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.

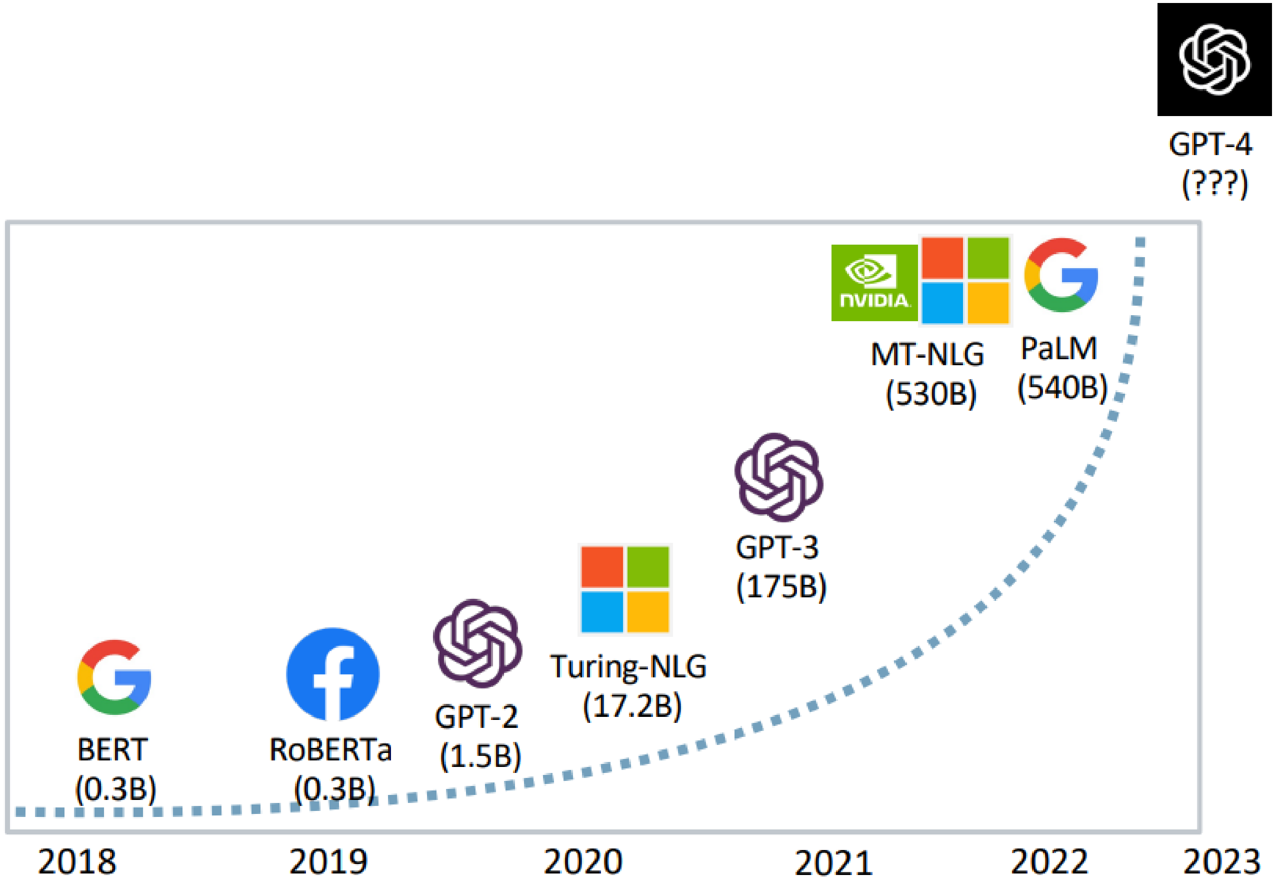
Reviewing Research Outcomes



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

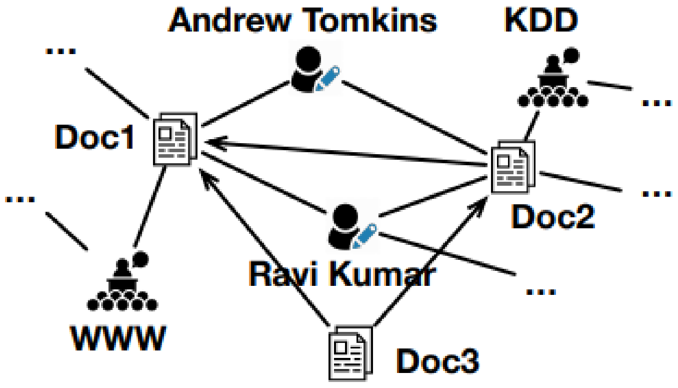
Pre-trained Language Models (PLMs) for Text Mining

- 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)*
  - …

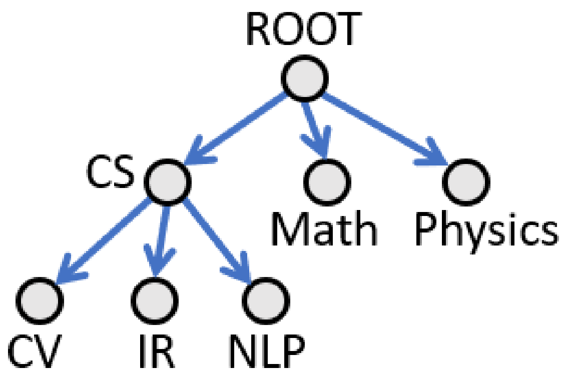


Are PLMs aware of graph information?

Graph Information Associated with Scientific Text



Metadata/Network



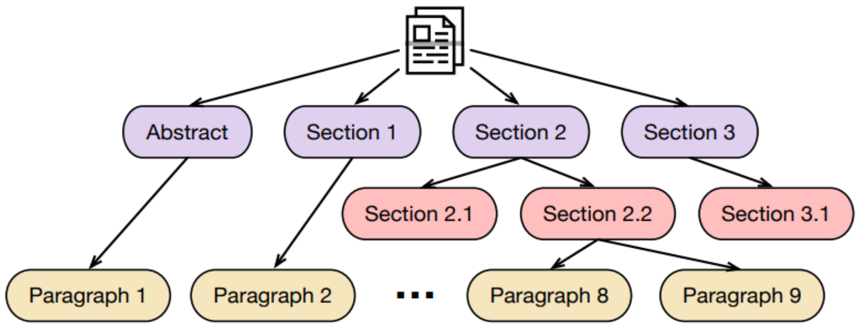
Hierarchy/Taxonomy

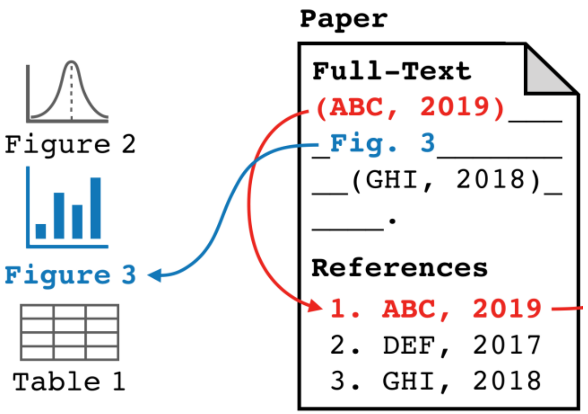
WIKIDATA

ctd

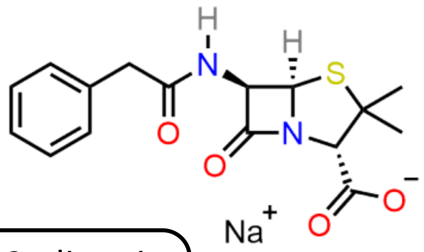
DBpedia

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.

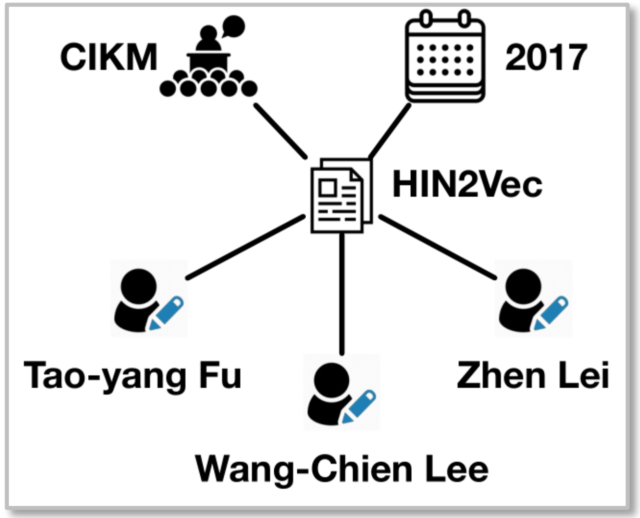
ChatGPT

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.

ChatGPT 3.5, queried on Jan 23, 2024

PLMs may not be graph-aware!

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

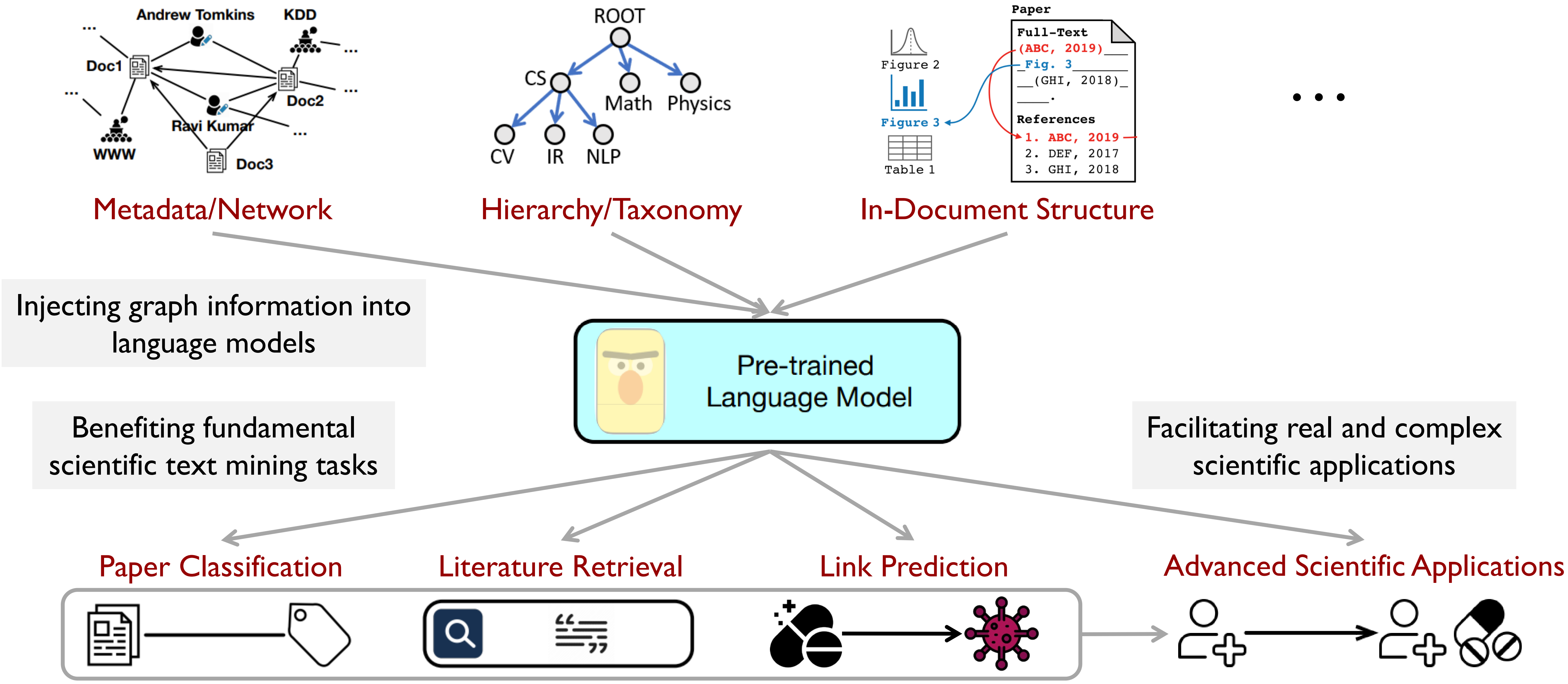
Searched 6 sites

The most cited paper from the WWW 2017 conference is "Attention Is All You Need" by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. This paper introduced the Transformer model, which has had a significant impact on the fields of natural language processing and machine learning (Nature) (Nature) (MIT Press Direct).

ChatGPT 3.5, queried on May 14, 2024

Transformer was published in NeurIPS 2017.

Today's Talk: Overview



Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

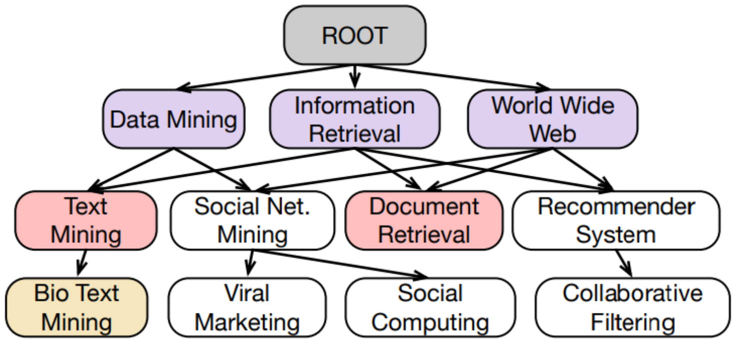
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

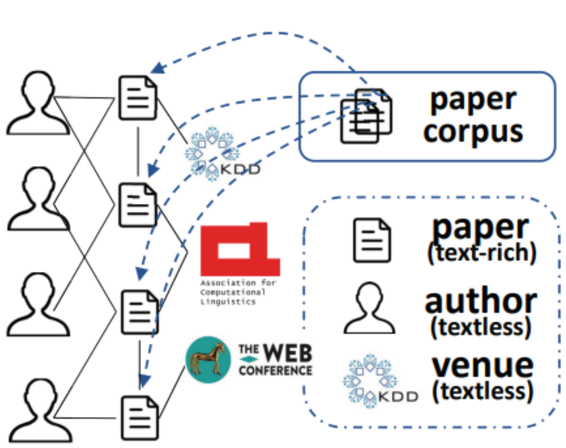
Part II: Text-Aware Link Prediction

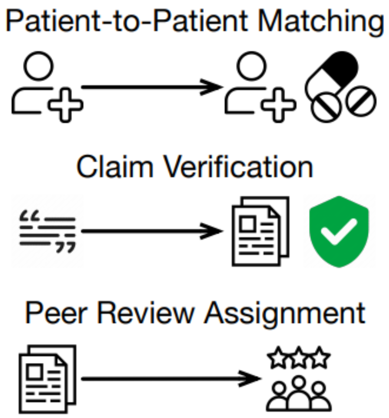
Jin et al., ACL 2023
Jin et al., KDD 2023

Part III: Advanced Scientific Applications

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







Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

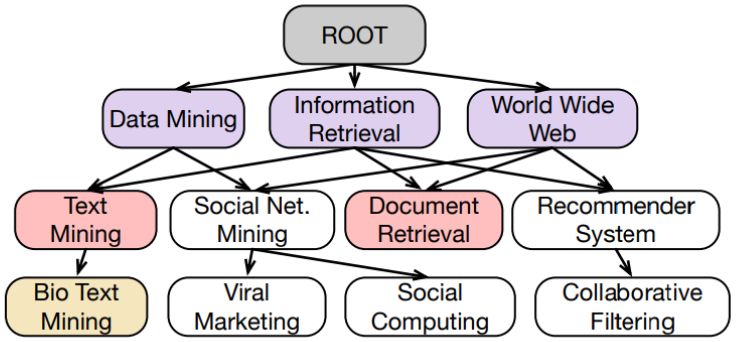
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

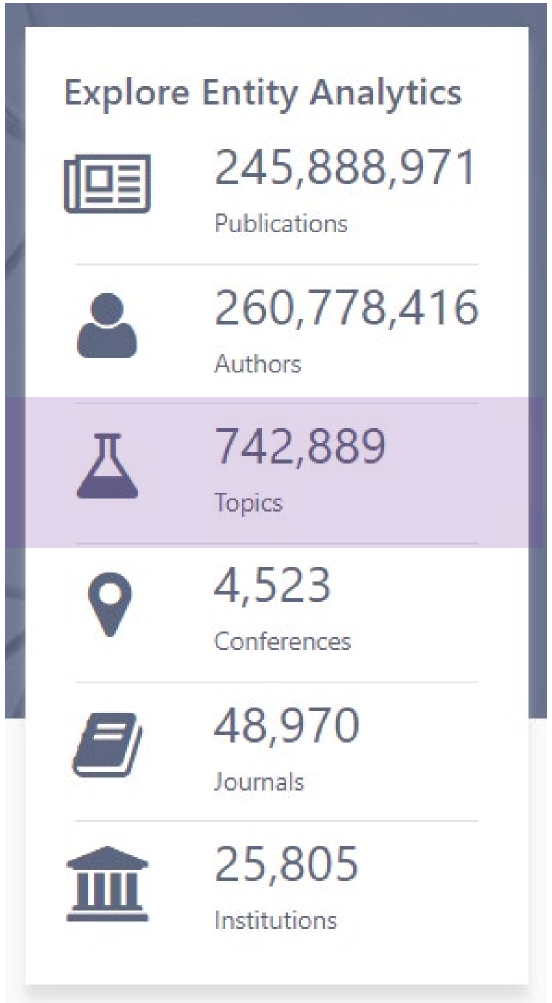
Jin et al., ACL 2023
Jin et al., KDD 2023

Part III: Advanced Scientific Applications

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



Extremely Fine-Grained Scientific Paper Classification



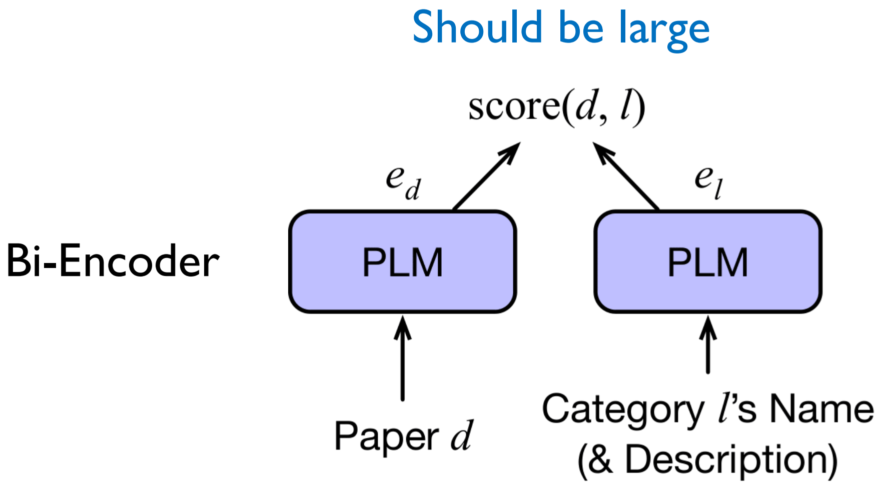
- The Microsoft Academic Graph has 740K+ categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain 30K+ categories.
- Each paper can be relevant to more than one category (5-15 categories for most papers).

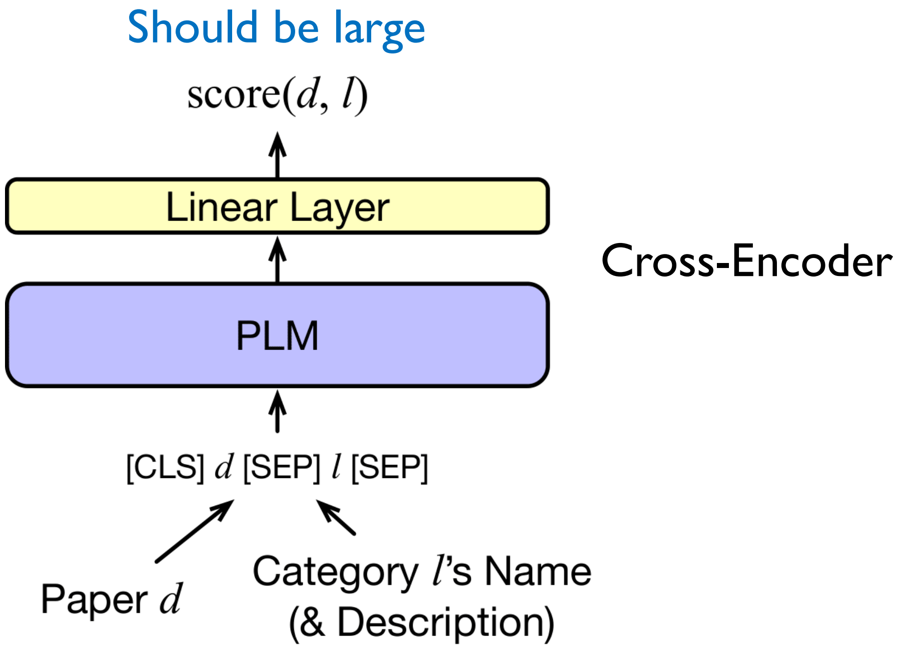
Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.

- 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.





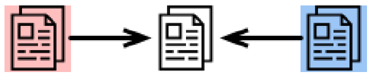
- However, human-annotated training samples are NOT available in many cases!
  - We are asking annotators to find ~10 relevant categories from ~100,000 candidates!

Using Graph Information to Replace Annotations

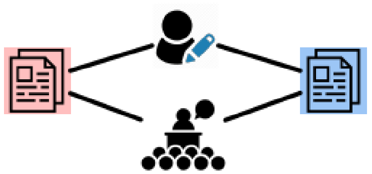
- If relevant (paper, category) pairs are not available, can we automatically create relevant (paper, paper) pairs?
  - Two papers sharing the same author(s) are assumed to be similar.
  - Two papers sharing the same reference(s) are assumed to be similar.
  - …

- The notion of meta-paths and meta-graphs

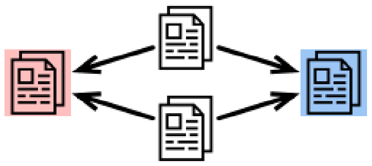
(a) meta-path: PAP



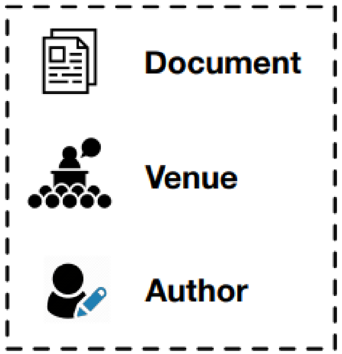
(b) meta-path: P->P<-P

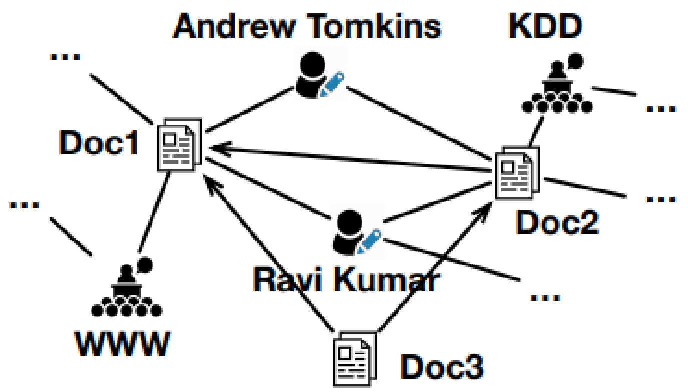


(c) meta-graph: P(AV)P



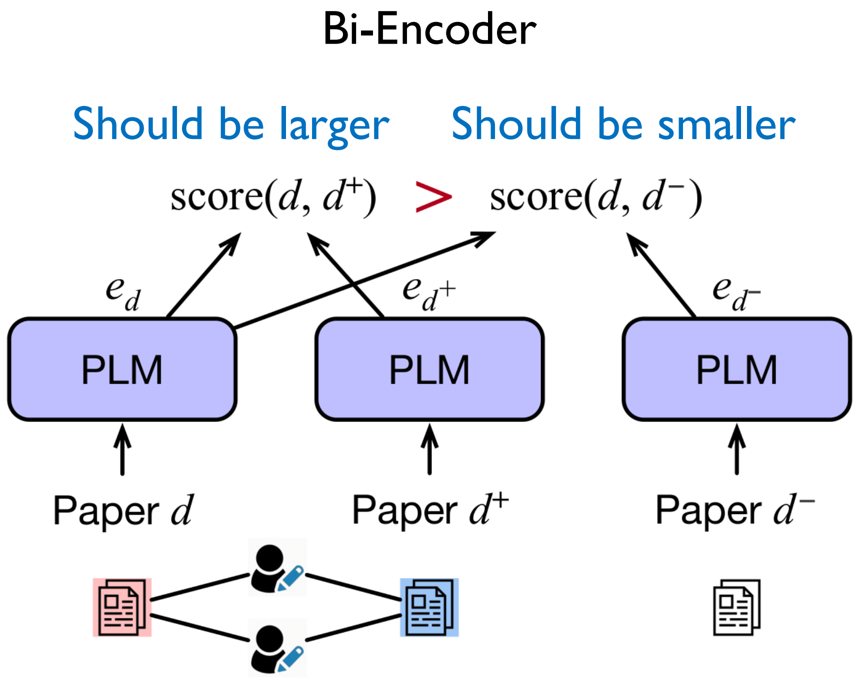
(d) meta-graph: P<-(PP)->P



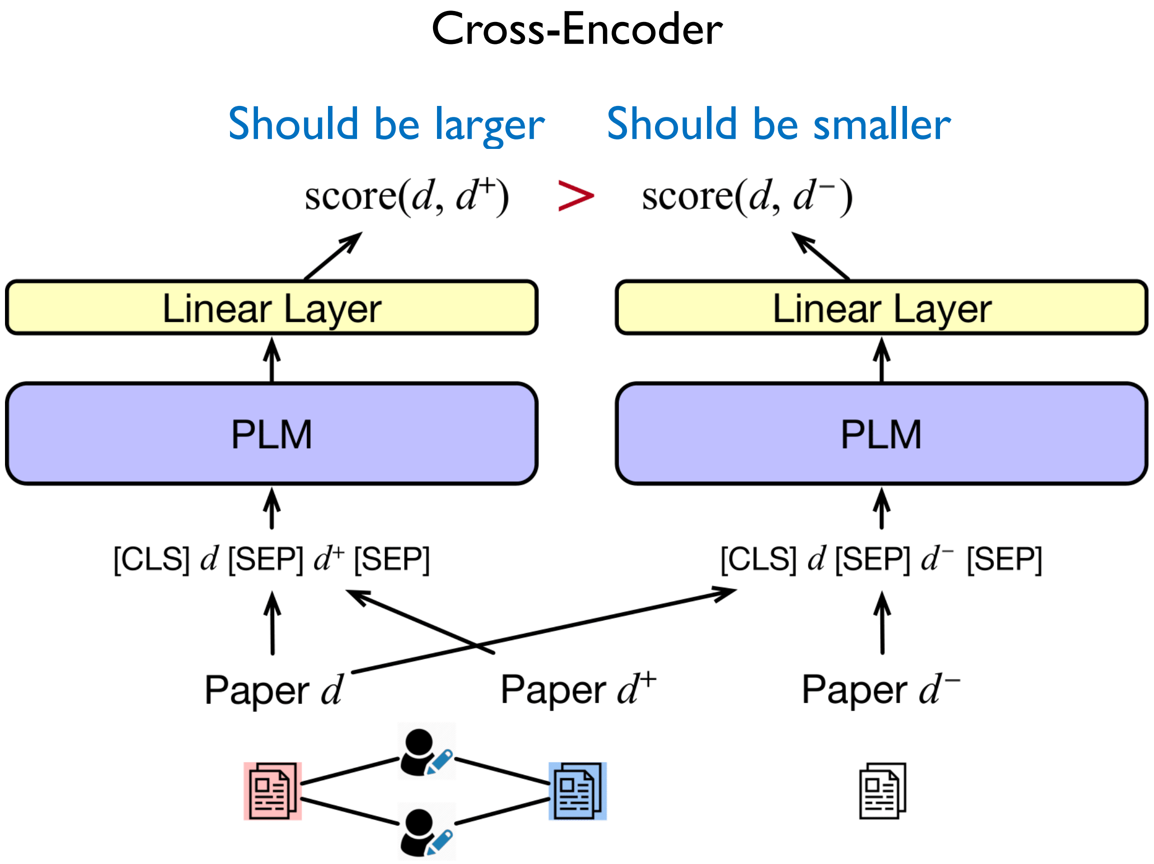


Graph-Induced Text Contrastive Learning

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



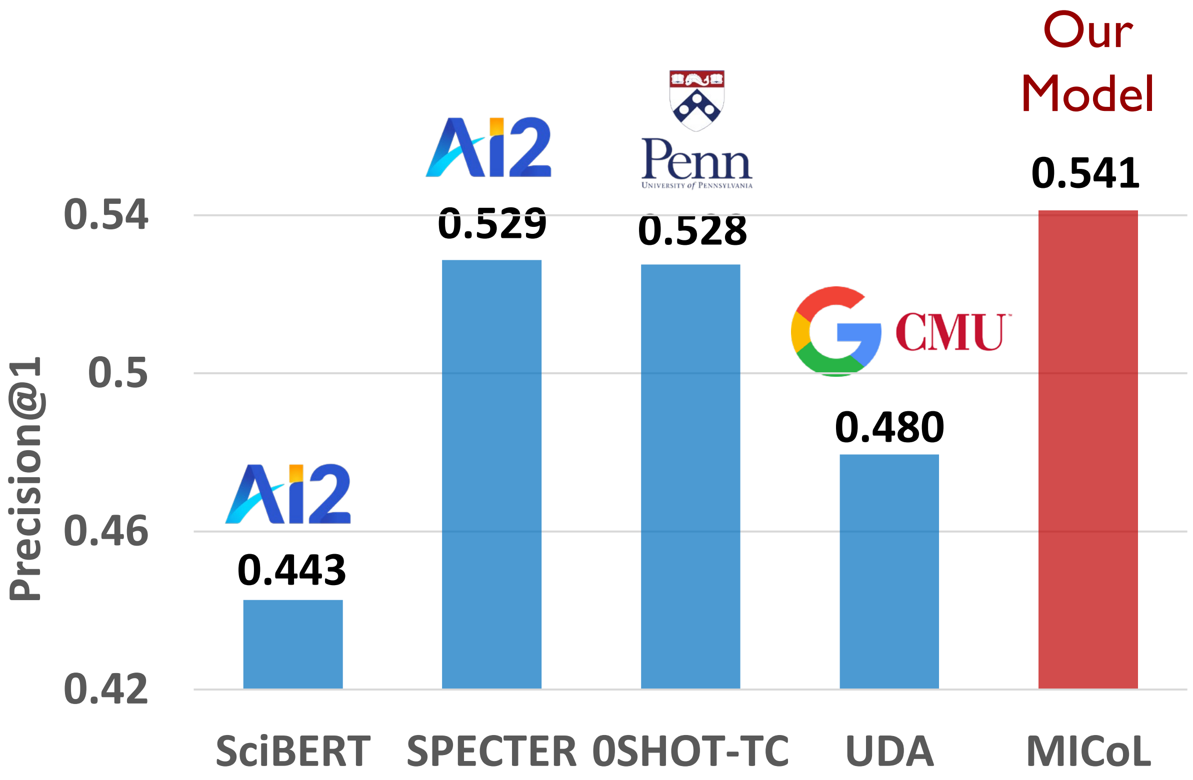
$$-\log \frac{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau)}{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau) + \sum_{i=1}^{N} \exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d_i^-})/\tau)}$$



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

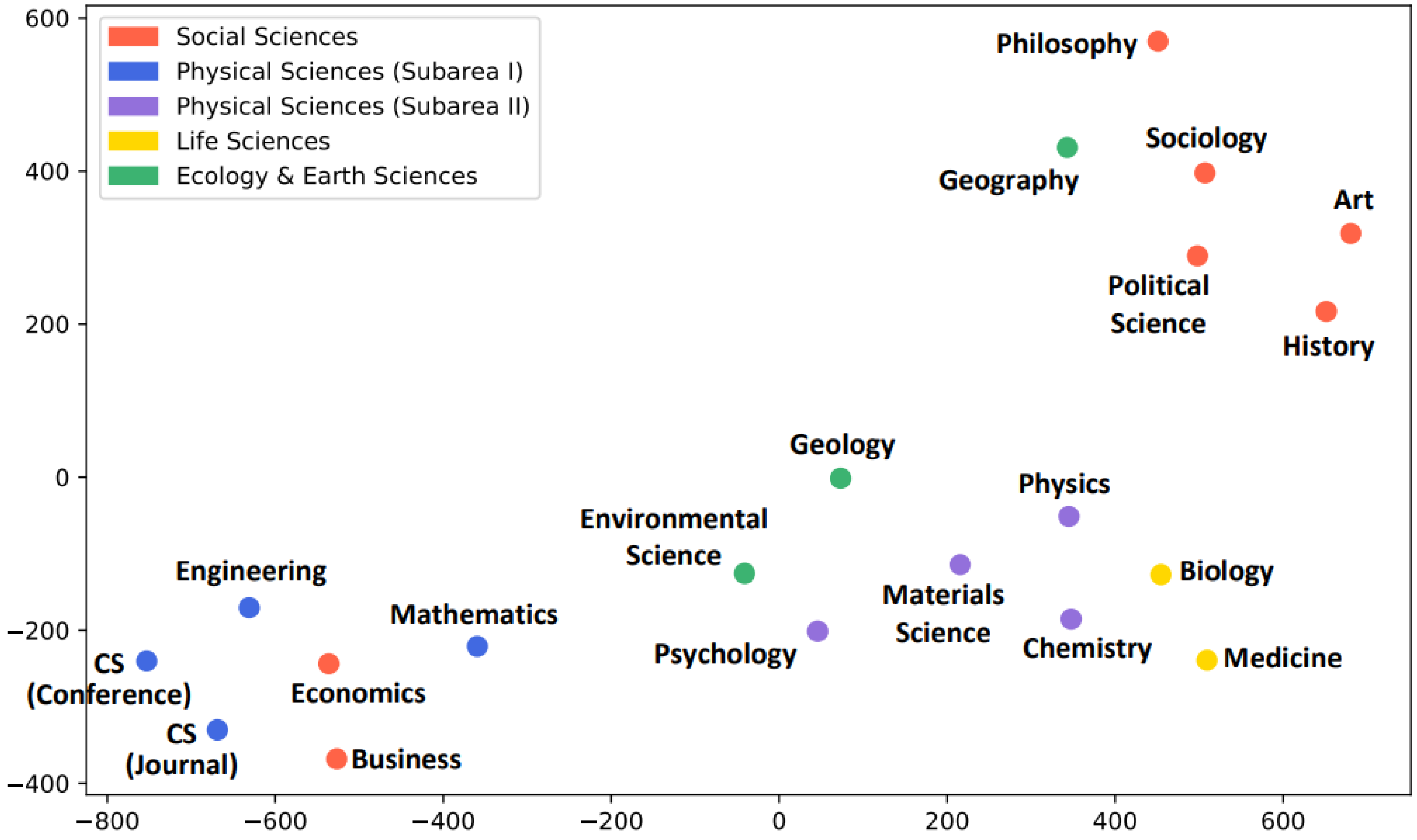
- Title: Improving Text Categorization Methods for Event Tracking
- Venue: SIGIR (2000)
- Authors: Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer
- Abstract: : Automated tracking of events from chronologically ordered document streams is a new challenge for statistical text classification.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

Which type of nodes is the most helpful?

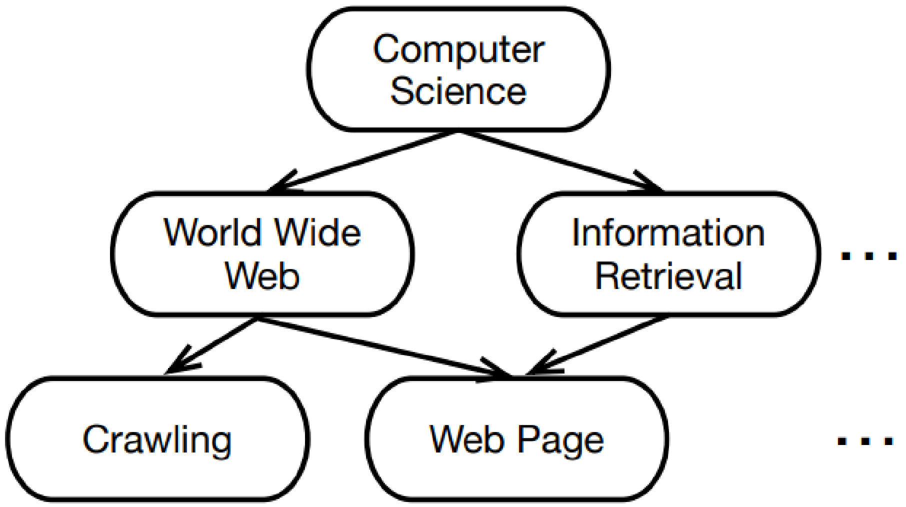
- Is the contribution of venues, authors, and references to paper classification consistent across different fields?
    - NO! BUT the effects of metadata tend to be similar in two similar fields.
    - The experience of using metadata in one field can be extrapolated to a similar field.



Zhang et al., "The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study." WWW 2023.

How about other types of graph information?

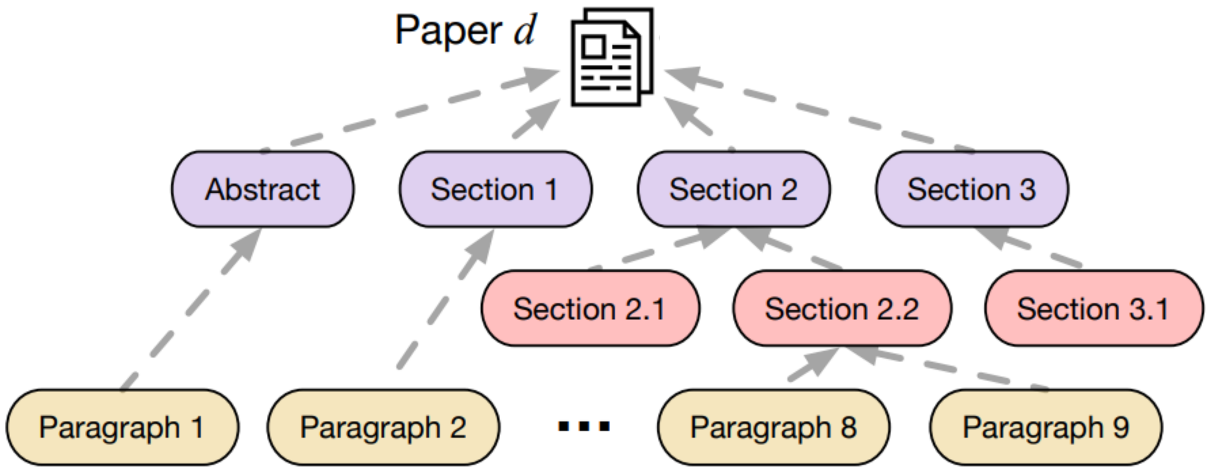
Label Hierarchy



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.

Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

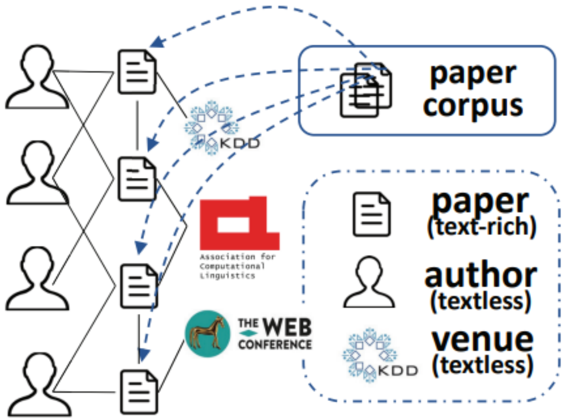
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

Jin et al., ACL 2023
Jin et al., KDD 2023

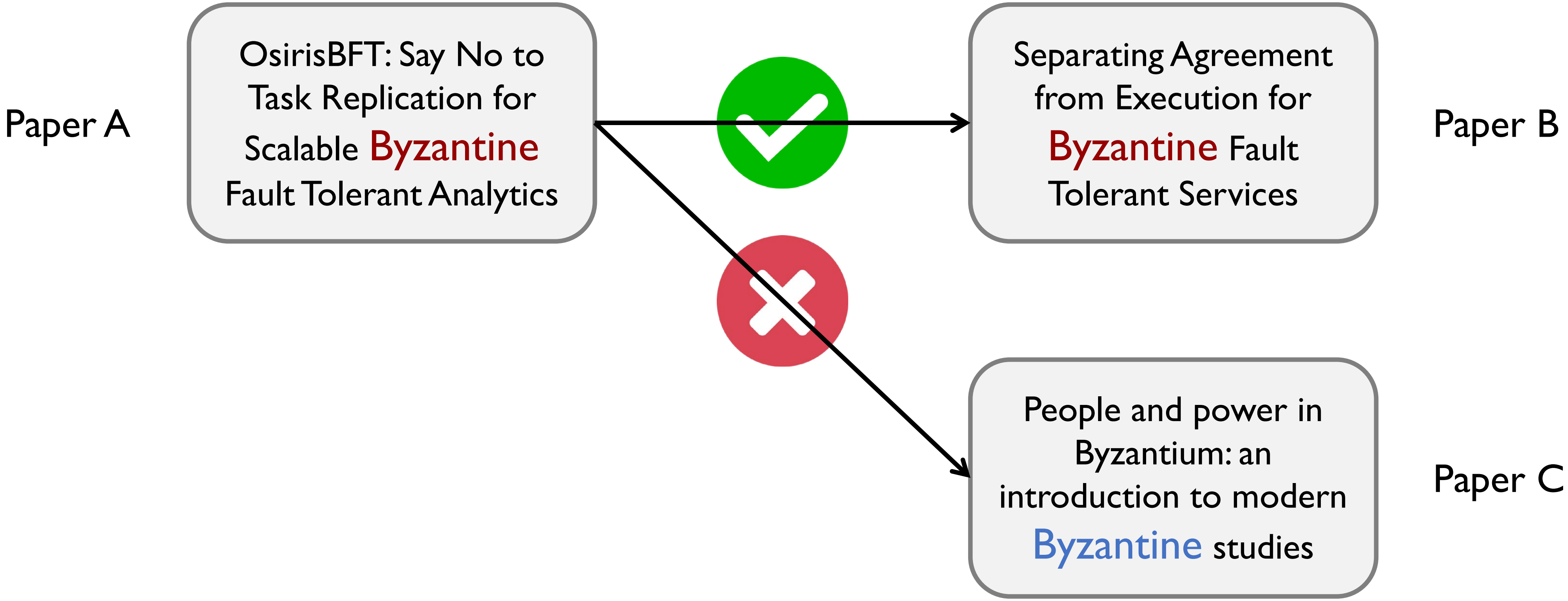
Part III: Advanced Scientific Applications

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



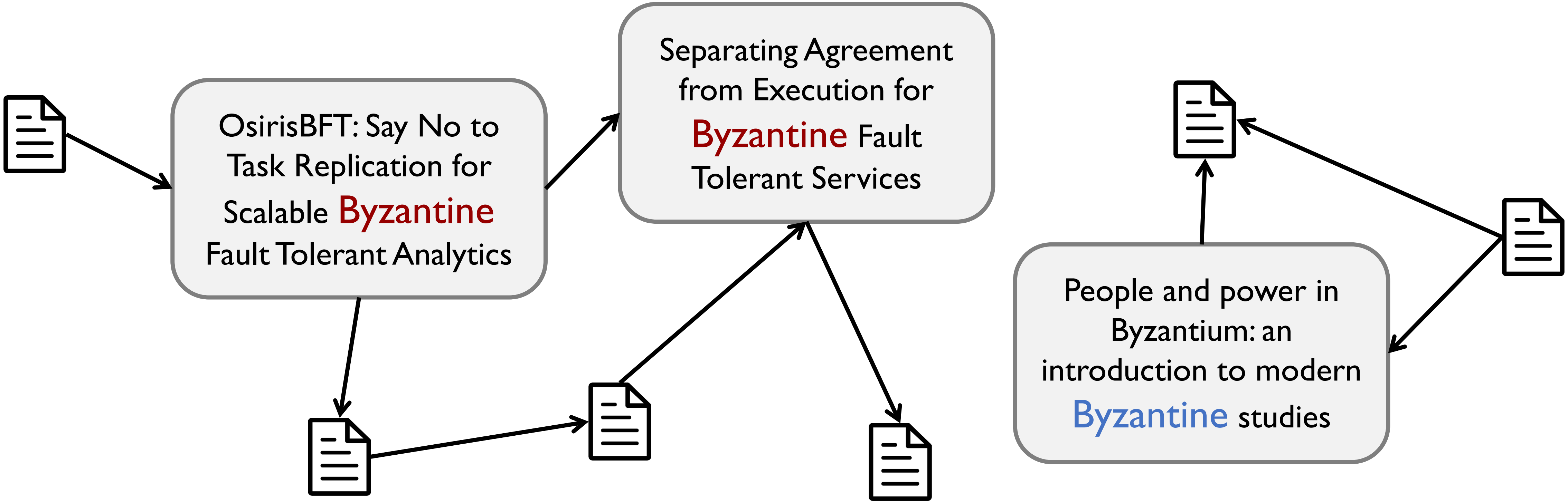
Text complements graph signals in link prediction, but …

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



Language Model Pre-training on Networks

- Given a pre-trained language model (e.g., BERT) and a network (where nodes are associated with text), we need to continue pre-training the language model to make it aware of network information.
- The network-aware pre-trained model can be used for link prediction, classification, …



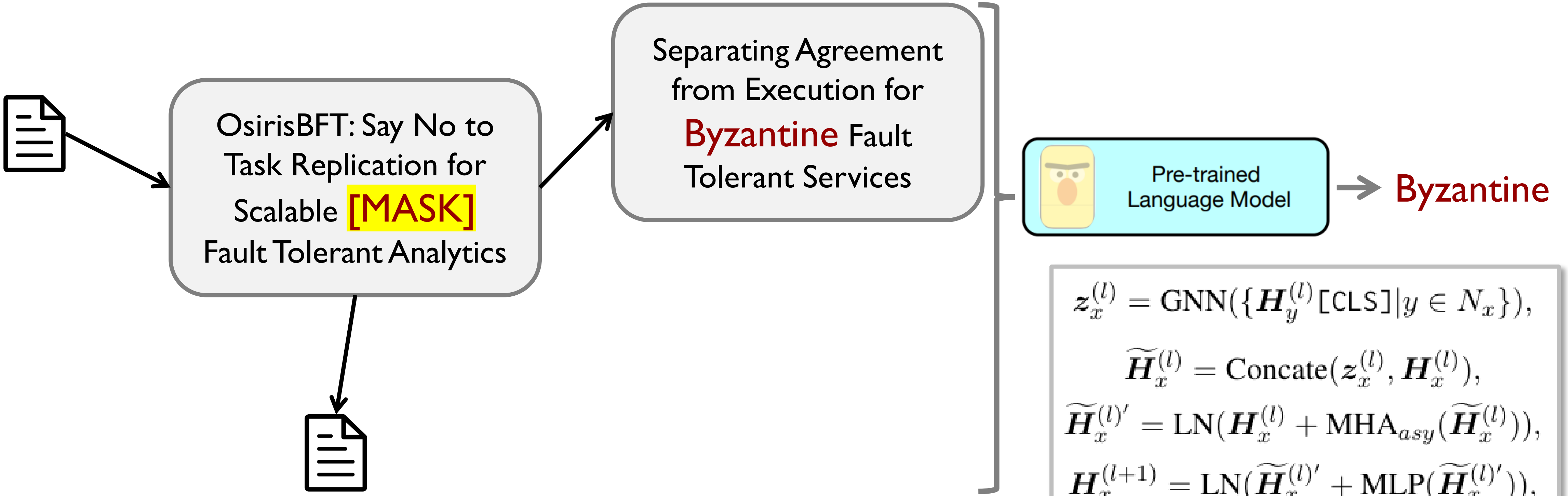
Jin et al. "Patton: Language Model Pretraining on Text-Rich Networks." ACL 2023.

Masked Language Modeling

- Masked Language Modeling in BERT pre-training:
  - Recovering the masked token given its context within the document.
  - Links between documents are not considered.

Network-Contextualized Masked Language Modeling

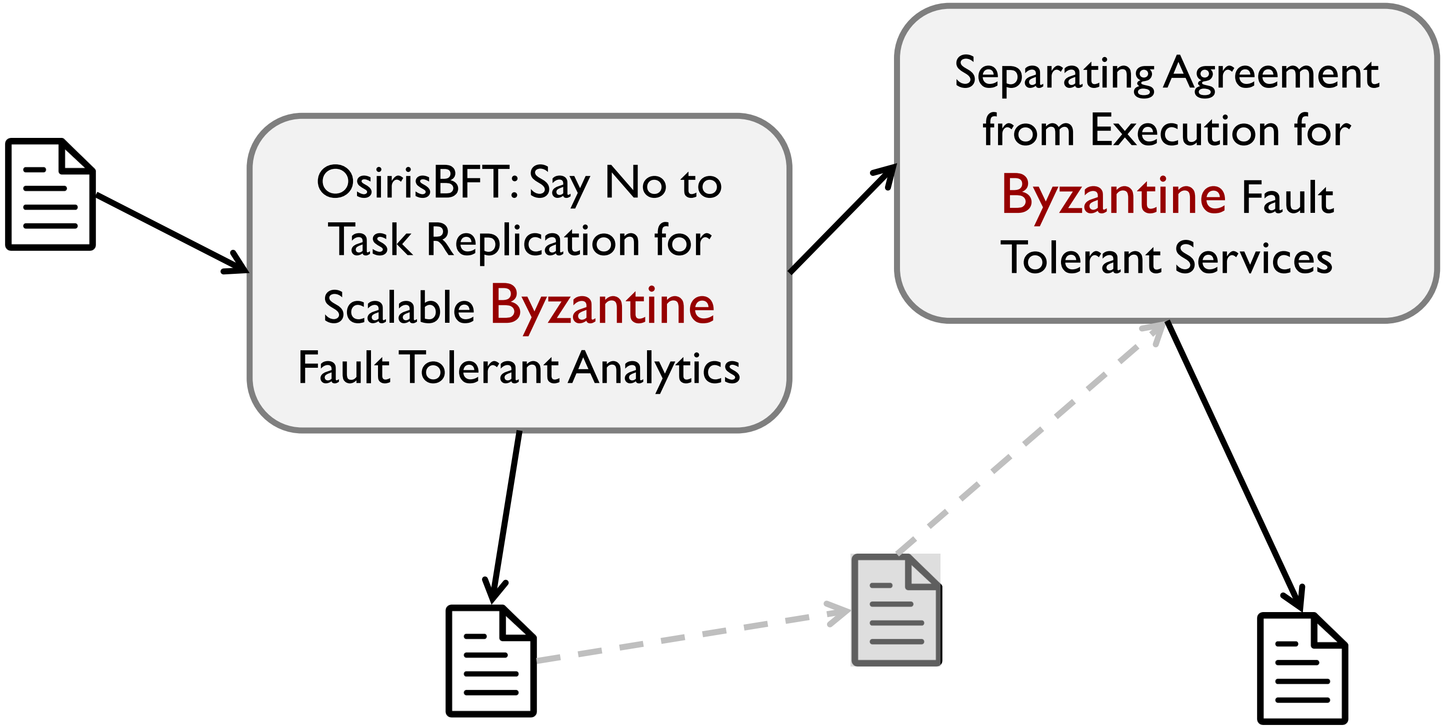
- Masked Language Modeling with network information:
  - Recovering the masked token given its context within the document AND across citation links.



$$
\begin{aligned}
\boldsymbol{z}_x^{(l)} &= \mathrm{GNN}(\{\boldsymbol{H}_y^{(l)}[\texttt{CLS}] \mid y \in N_x\}),\\
\widetilde{\boldsymbol{H}}_x^{(l)} &= \mathrm{Concate}(\boldsymbol{z}_x^{(l)}, \boldsymbol{H}_x^{(l)}),\\
\widetilde{\boldsymbol{H}}_x^{(l)\prime} &= \mathrm{LN}(\boldsymbol{H}_x^{(l)} + \mathrm{MHA}_{asy}(\widetilde{\boldsymbol{H}}_x^{(l)})),\\
\boldsymbol{H}_x^{(l+1)} &= \mathrm{LN}(\widetilde{\boldsymbol{H}}_x^{(l)\prime} + \mathrm{MLP}(\widetilde{\boldsymbol{H}}_x^{(l)\prime})),
\end{aligned}
$$

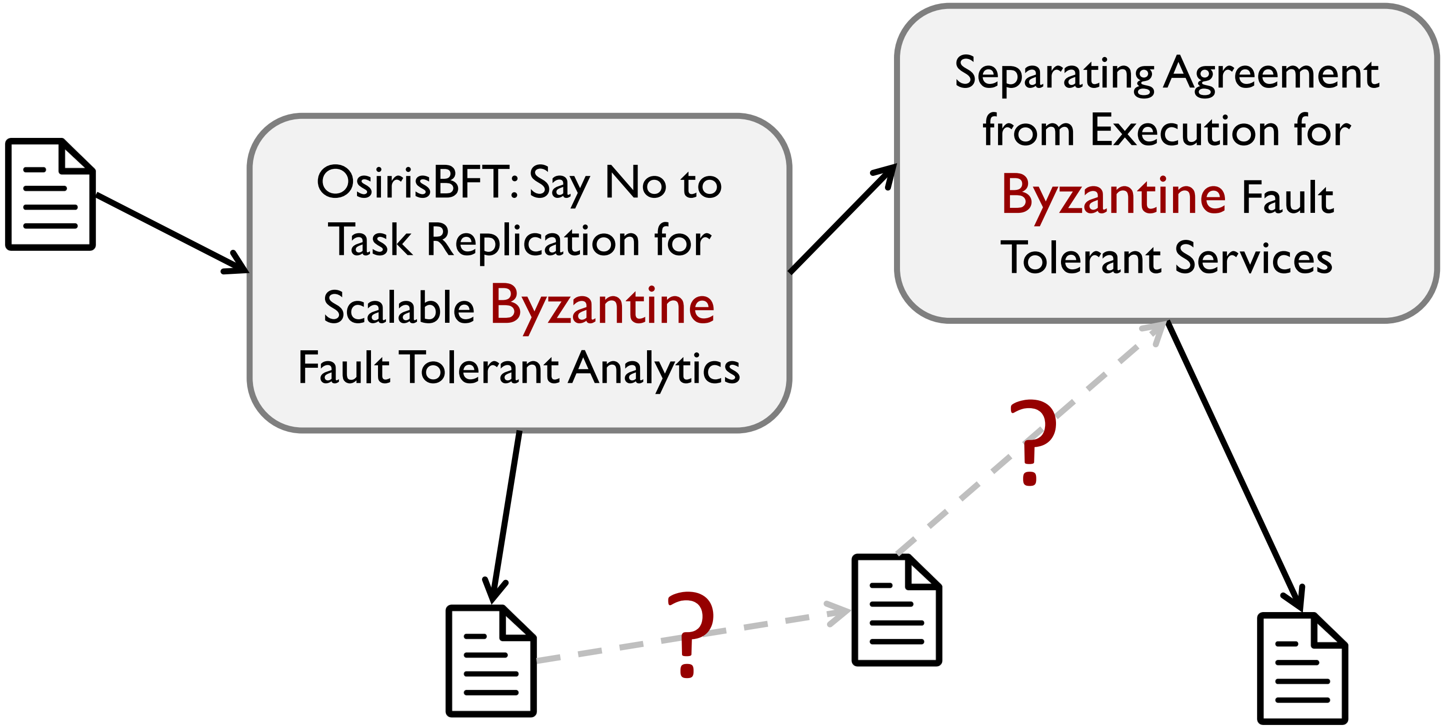
Masked Node Prediction

- If we randomly remove some nodes (and their associated edges) in the network, the model should be able to predict which removed node was in a certain position.



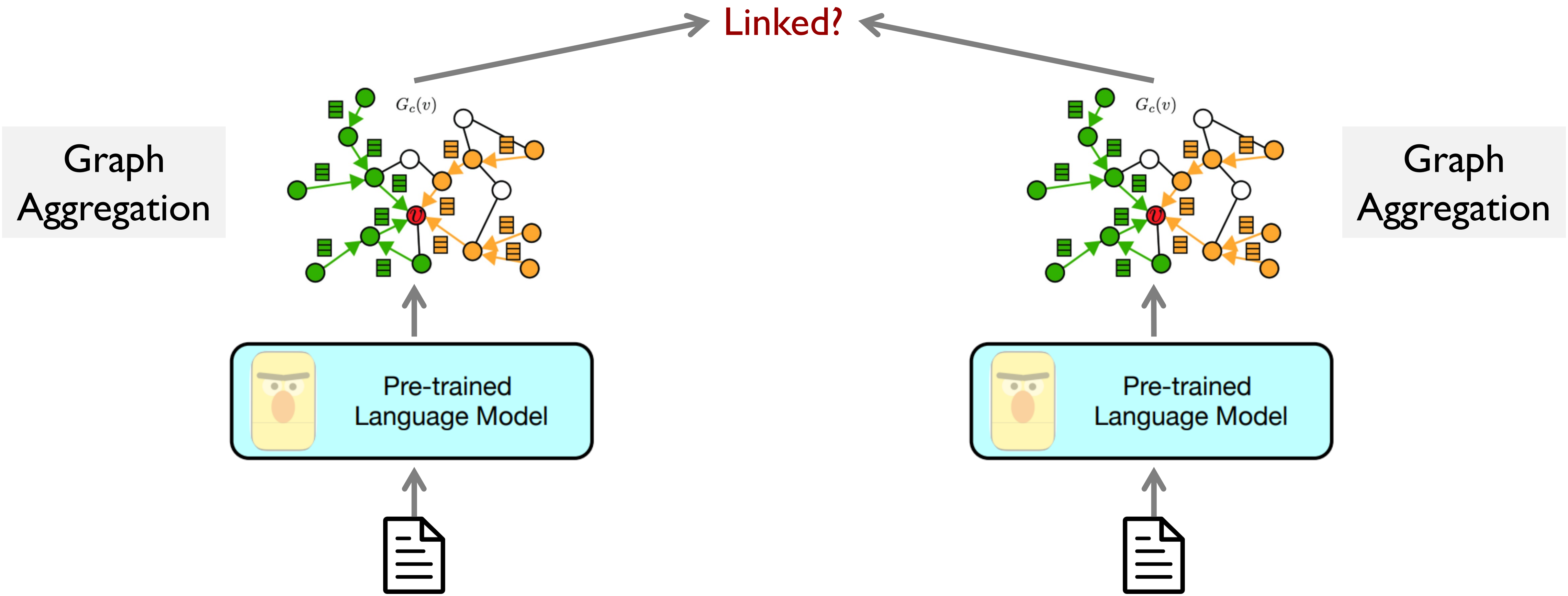
Masked Node Prediction

- If we randomly remove some nodes (and their associated edges) in the network, the model should be able to predict which removed node was in a certain position.
- Mathematically equivalent to link prediction



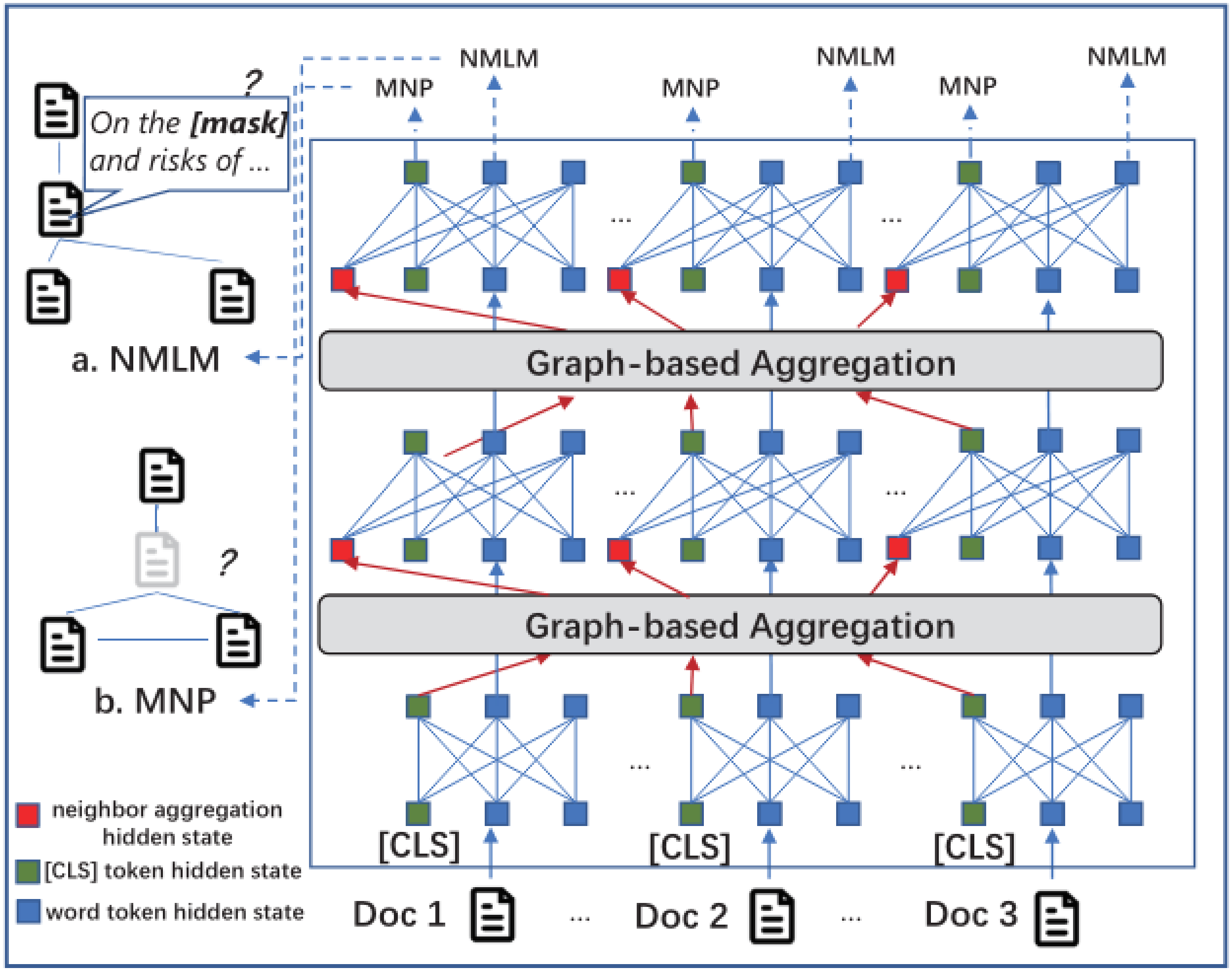
$$
\begin{aligned}
& \prod_{v_{\texttt{[MASK]}} \in M_v} p(v_{\texttt{[MASK]}} = v_i | v_k \in N_{v_{\texttt{[MASK]}}}) \\
\propto & \prod_{v_{\texttt{[MASK]}} \in M_v} p(v_k \in N_{v_{\texttt{[MASK]}}} | v_{\texttt{[MASK]}} = v_i) \\
= & \prod_{v_{\texttt{[MASK]}} \in M_v} \prod_{v_k \in N_{v_{\texttt{[MASK]}}}} p(v_k | v_{\texttt{[MASK]}} = v_i) \\
= & \prod_{v_{\texttt{[MASK]}} \in M_v} \prod_{v_k \in N_{v_{\texttt{[MASK]}}}} p(v_k \longleftrightarrow v_i)
\end{aligned}
$$

Previous PLM-GNN Cascaded Architecture



- Drawback: Graph information is not used by the PLM when encoding text.

Patton: An Interleaved Architecture



- Cascaded Architecture:
  - Transformer → Transformer → … → Transformer → Aggregation
- Interleaved Architecture:
  - Transformer → Aggregation → Transformer → Aggregation → … → Transformer → Aggregation

Jin et al. "Patton: Language Model Pretraining on Text-Rich Networks." ACL 2023.

Comparison with Previous Approaches

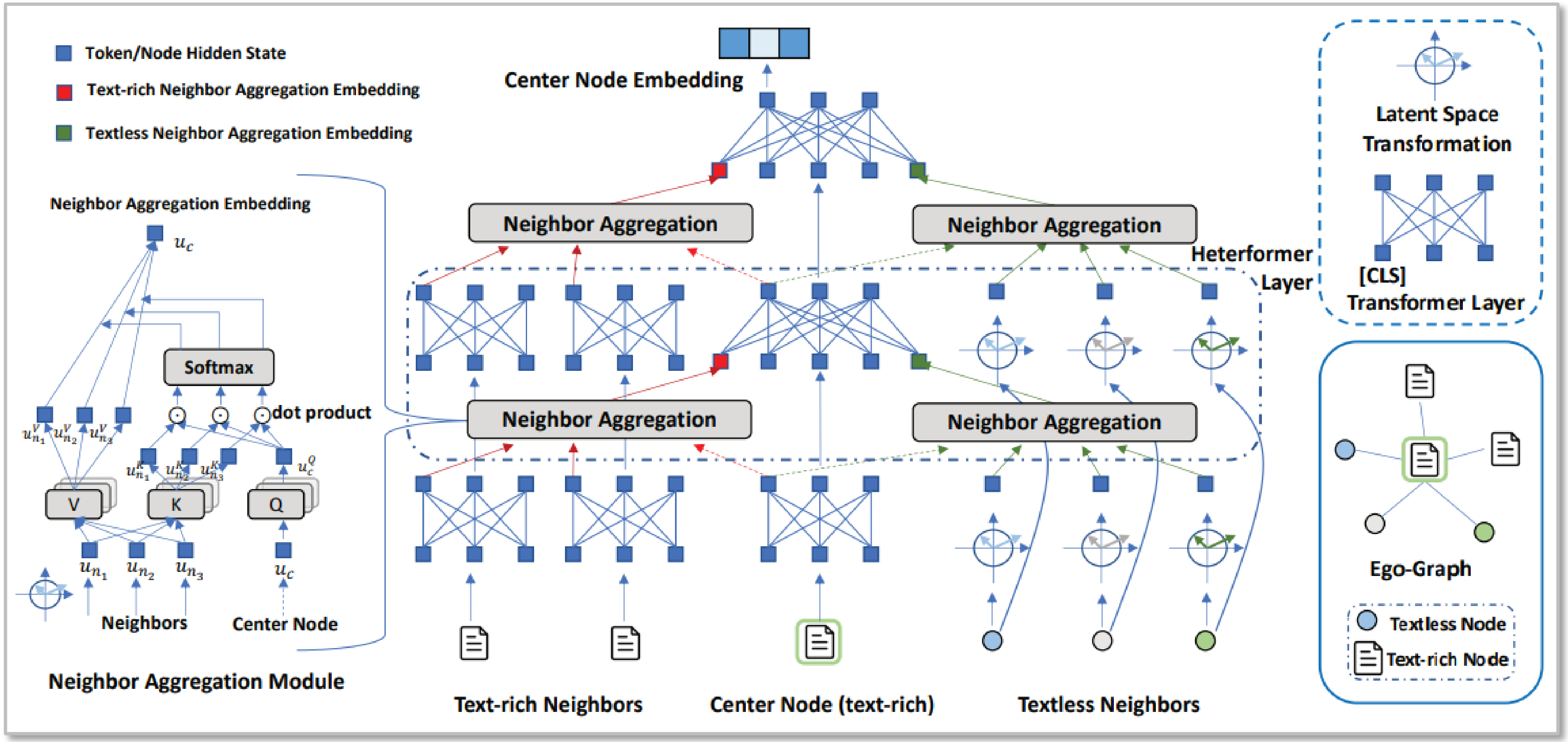
- Dataset: Microsoft Academic Graph (3 fields)

Method	Mathematics		Geology		Economics	
	PREC@1	MRR	PREC@1	MRR	PREC@1	MRR
BERT	$6.60_{0.16}$	$12.96_{0.34}$	$6.24_{0.76}$	$12.96_{1.34}$	$4.12_{0.08}$	$9.23_{0.15}$
GraphFormers	$6.91_{0.29}$	$13.42_{0.34}$	$6.52_{1.17}$	$13.34_{1.81}$	$4.16_{0.21}$	$9.28_{0.28}$
SciBERT	$14.08_{0.11}$	$23.62_{0.10}$	$7.15_{0.26}$	$14.11_{0.39}$	$5.01_{1.04}$	$10.48_{1.79}$
SPECTER	$13.44_{0.5}$	$21.73_{0.65}$	$6.85_{0.22}$	$13.37_{0.34}$	$6.33_{0.29}$	$12.41_{0.33}$
SimCSE (unsup)	$9.85_{0.10}$	$16.28_{0.12}$	$7.47_{0.55}$	$14.24_{0.89}$	$5.72_{0.26}$	$11.02_{0.34}$
SimCSE (sup)	$10.35_{0.52}$	$17.01_{0.72}$	$10.10_{0.04}$	$17.80_{0.07}$	$5.72_{0.26}$	$11.02_{0.34}$
LinkBERT	$8.05_{0.14}$	$13.91_{0.09}$	$6.40_{0.14}$	$12.99_{0.17}$	$2.97_{0.08}$	$6.79_{0.15}$
BERT.MLM	$17.55_{0.25}$	$29.22_{0.26}$	$14.13_{0.19}$	$25.36_{0.20}$	$9.02_{0.09}$	$16.72_{0.15}$
SciBERT.MLM	$22.44_{0.08}$	$34.22_{0.05}$	$16.22_{0.03}$	$27.02_{0.07}$	$9.80_{0.00}$	$17.72_{0.01}$
SimCSE.in-domain	$33.55_{0.05}$	$46.07_{0.07}$	$24.56_{0.06}$	$36.89_{0.11}$	$16.77_{0.10}$	$26.93_{0.01}$
PATTON	$70.41_{0.11}$	$80.21_{0.04}$	$44.76_{0.05}$	$57.71_{0.04}$	$57.04_{0.05}$	$68.35_{0.04}$
SciPATTON	$\mathbf{71.22}_{0.17}$	$\mathbf{80.79}_{0.10}$	$\mathbf{44.95}_{0.24}$	$\mathbf{57.84}_{0.25}$	$\mathbf{57.36}_{0.26}$	$\mathbf{68.71}_{0.31}$
w/o NMLM	$71.04_{0.13}$	$80.60_{0.07}$	$44.33_{0.23}$	$57.29_{0.22}$	$56.64_{0.25}$	$68.12_{0.16}$
w/o MNP	$63.06_{0.23}$	$74.26_{0.11}$	$33.84_{0.60}$	$47.02_{0.65}$	$44.46_{0.03}$	$57.05_{0.04}$

Jin et al. "Patton: Language Model Pretraining on Text-Rich Networks." ACL 2023.

Dealing with Network Heterogeneity

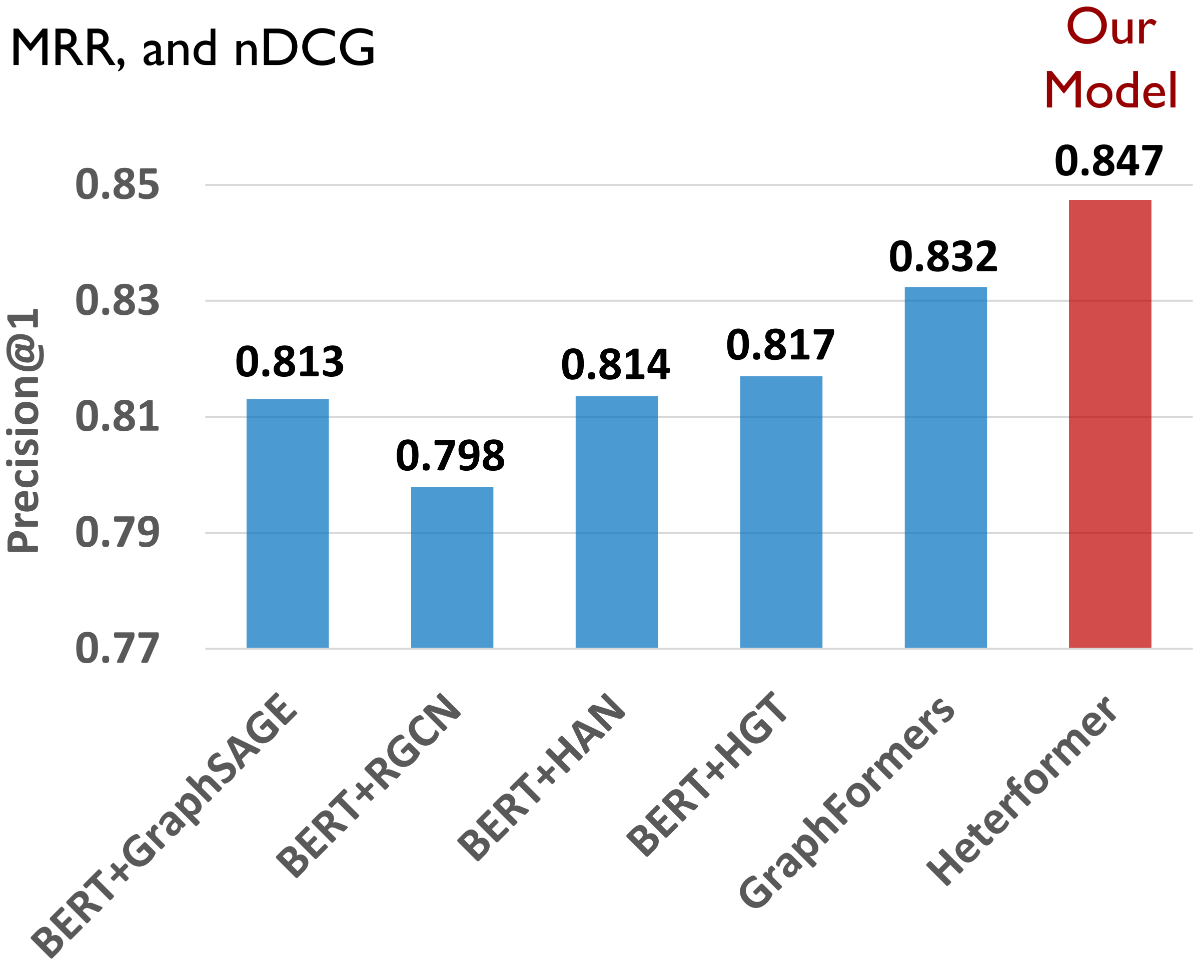
- Some types of nodes (e.g., author, year) may not have semantic-indicative text information!



Jin et al. "Heterformer: Transformer-based Deep Node Representation Learning on Heterogeneous Text-Rich Networks." KDD 2023.

Comparison with Previous Approaches

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



Jin et al. “Heterformer: Transformer-based Deep Node Representation Learning on Heterogeneous Text-Rich Networks.” KDD 2023.

Today's Talk: Overview

Part I: Extremely Fine-Grained Classification

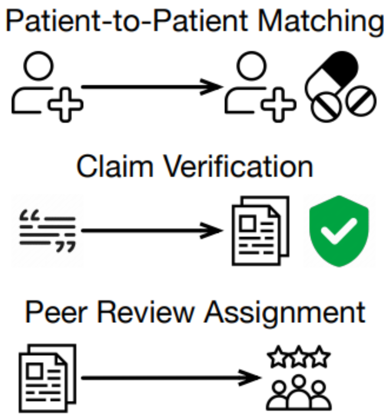
Zhang et al., WWW 2021
Zhang et al., WWW 2022
Zhang et al., WWW 2023
Zhang et al., KDD 2023

Part II: Text-Aware Link Prediction

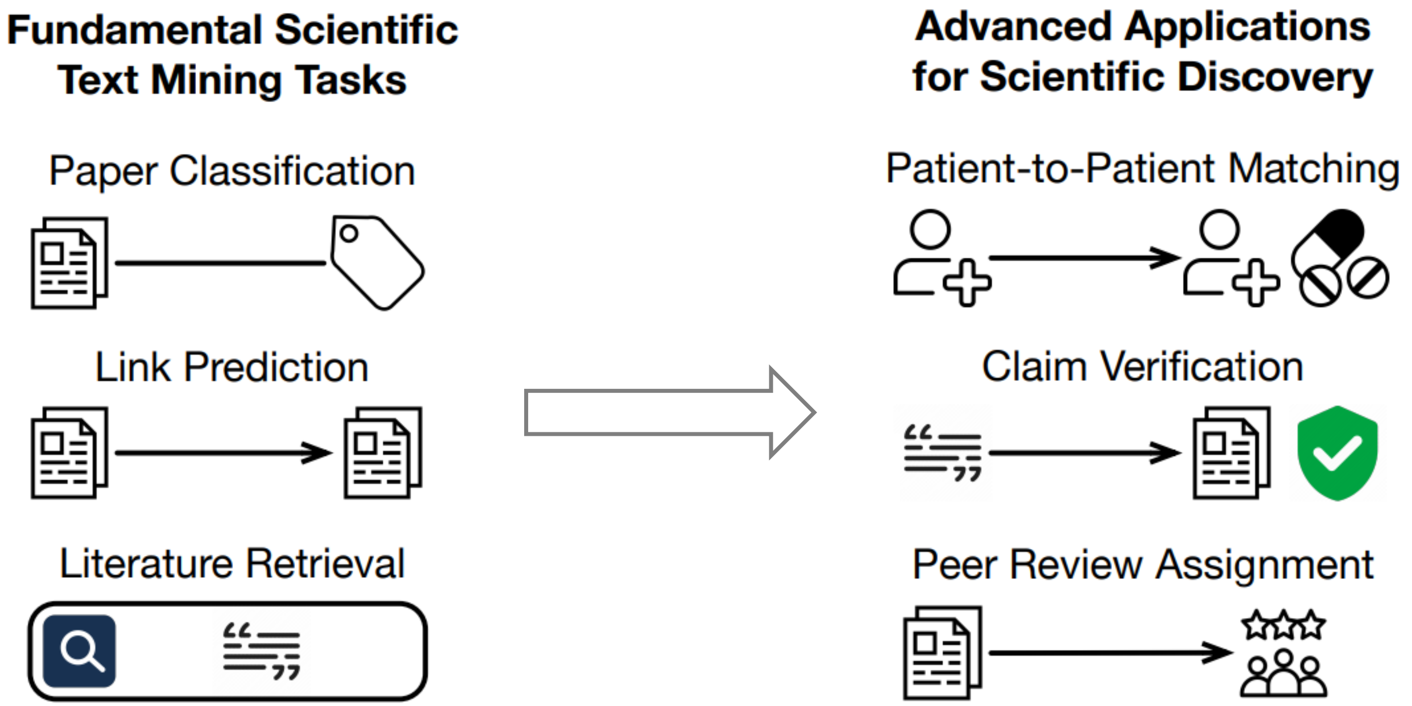
Jin et al., ACL 2023
Jin et al., KDD 2023

Part III: Advanced Scientific Applications

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



Facilitating Complex Tasks for Scientific Discovery



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

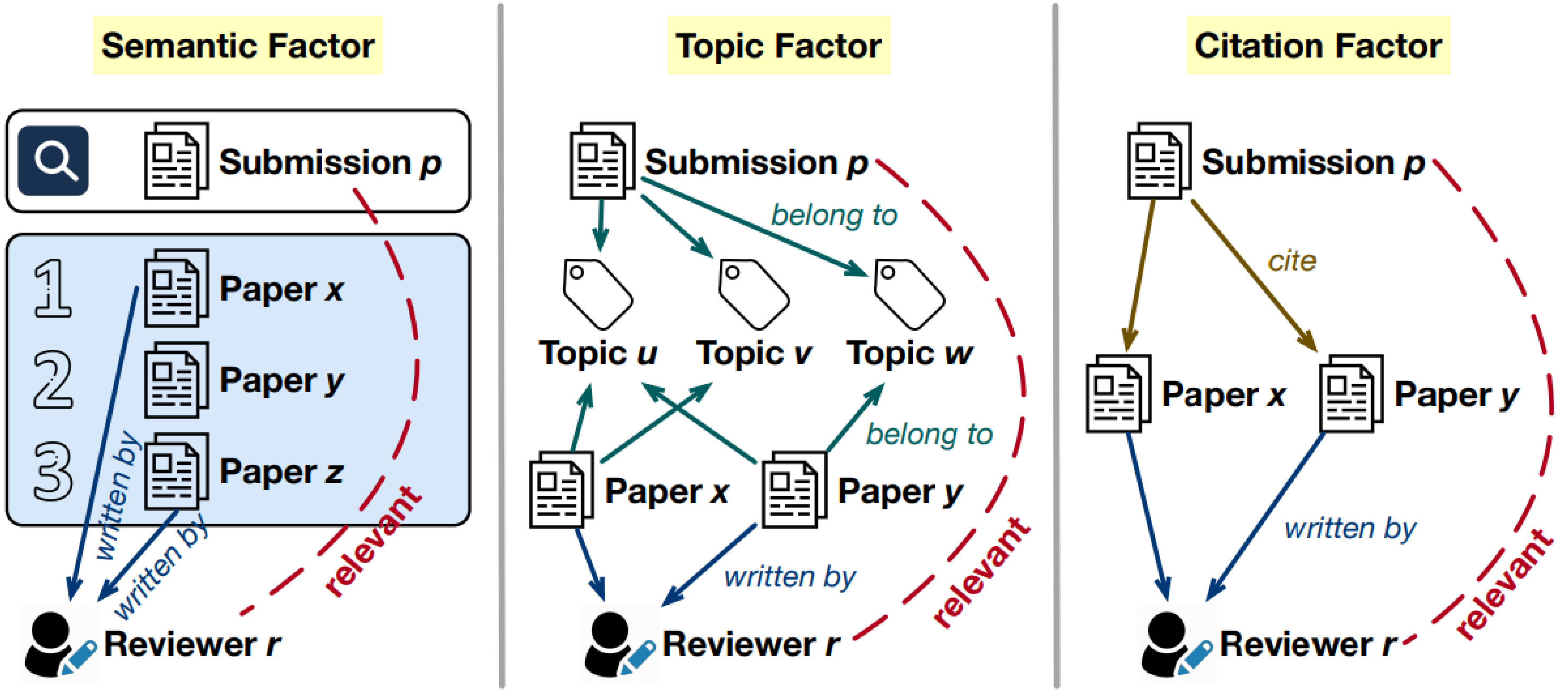
Given a scientific claim, find relevant papers (and predict their stance).

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

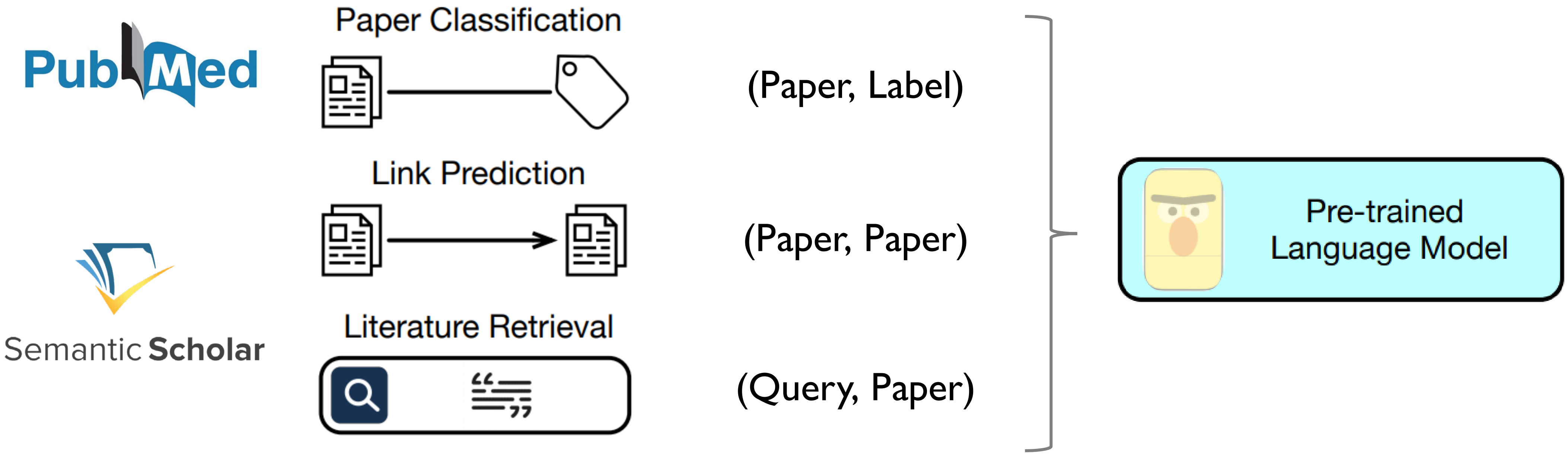
- Example: Paper-Reviewer Matching
  - Why is a pair of (Paper, Reviewer) relevant?



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

Naïve Multi-task Pre-training

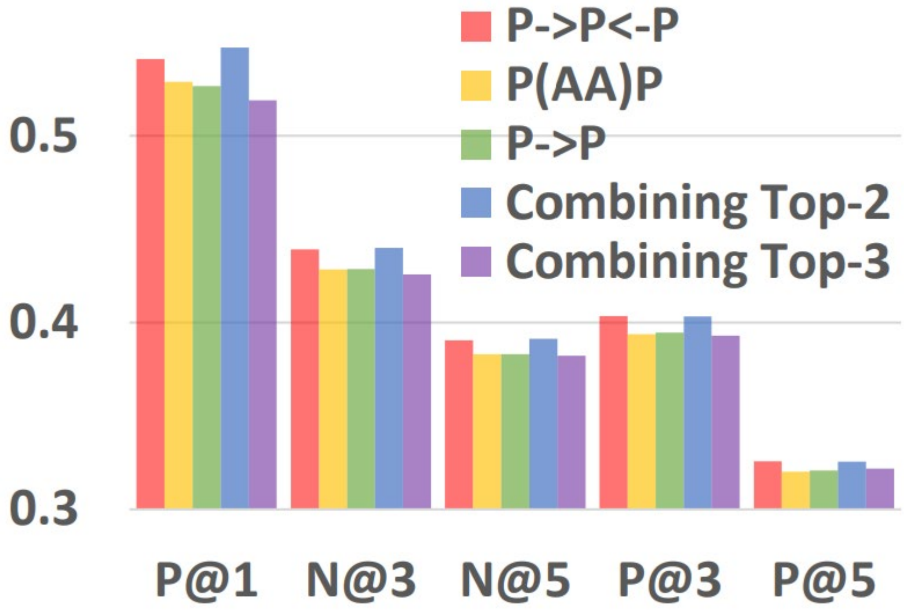
- Each factor (topic, citation, and semantic) relies on one fundamental text mining task.
- Directly combining pre-training data from different tasks to train a model?

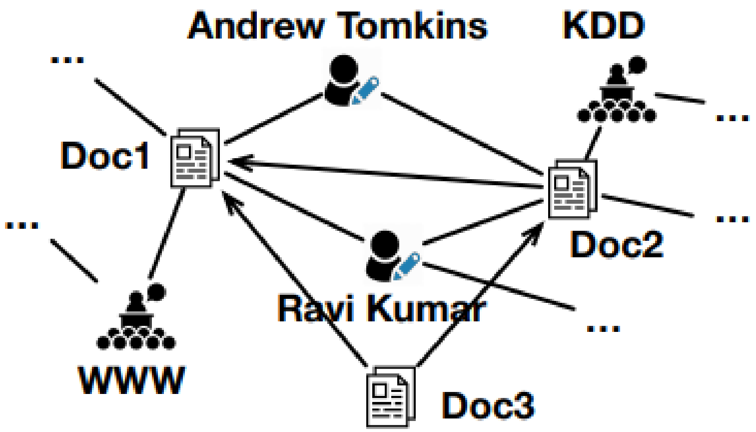


- Task Interference: The model is confused by different types of "relevance".

An Illustrative Example of Task Interference

- Recall graph-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!

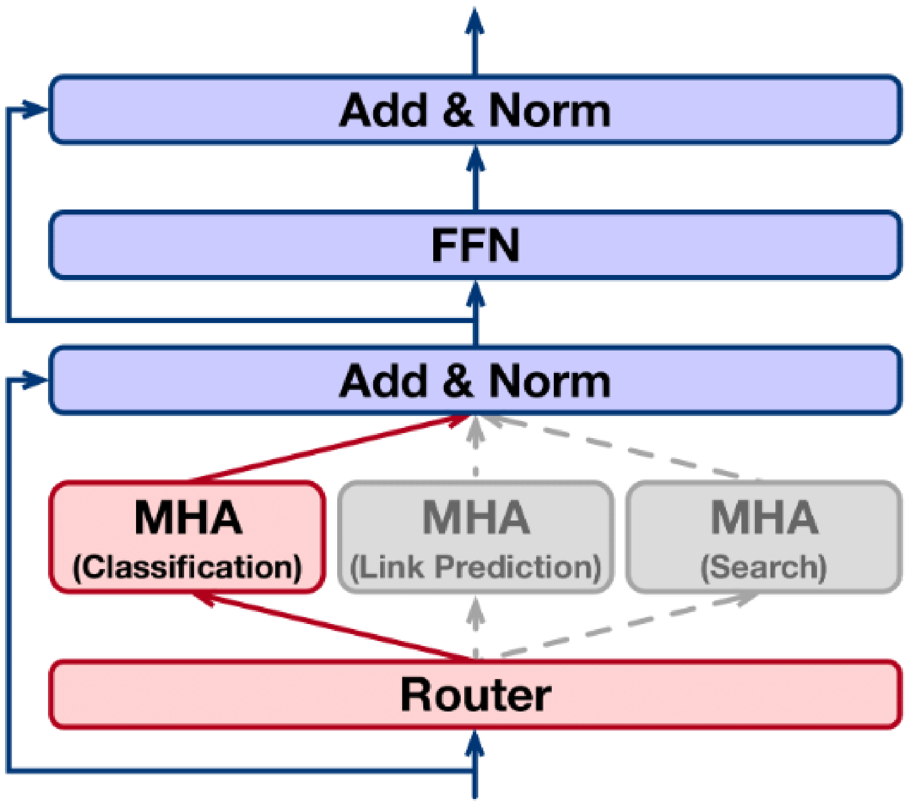




(Doc2, Doc3) are relevant according to P→P←P but irrelevant according to P(AA)P.

Tackling Task Interference: Mixture-of-Experts Transformer

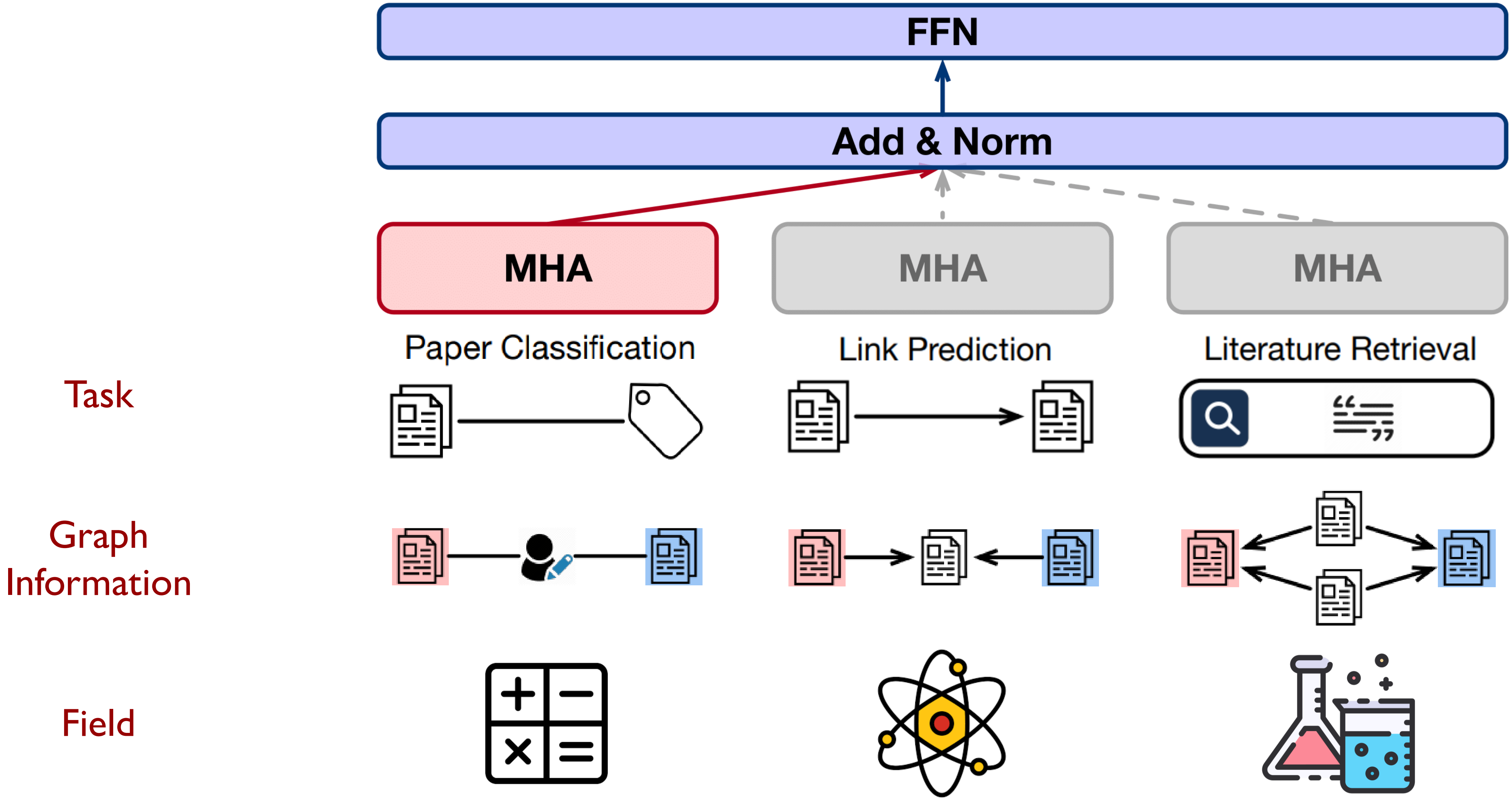
- A typical Transformer layer
  - 1 Multi-Head Attention (MHA) sublayer
  - 1 Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
  - Multiple MHA sublayers
  - 1 FFN sublayer
  - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be an "expert"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



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



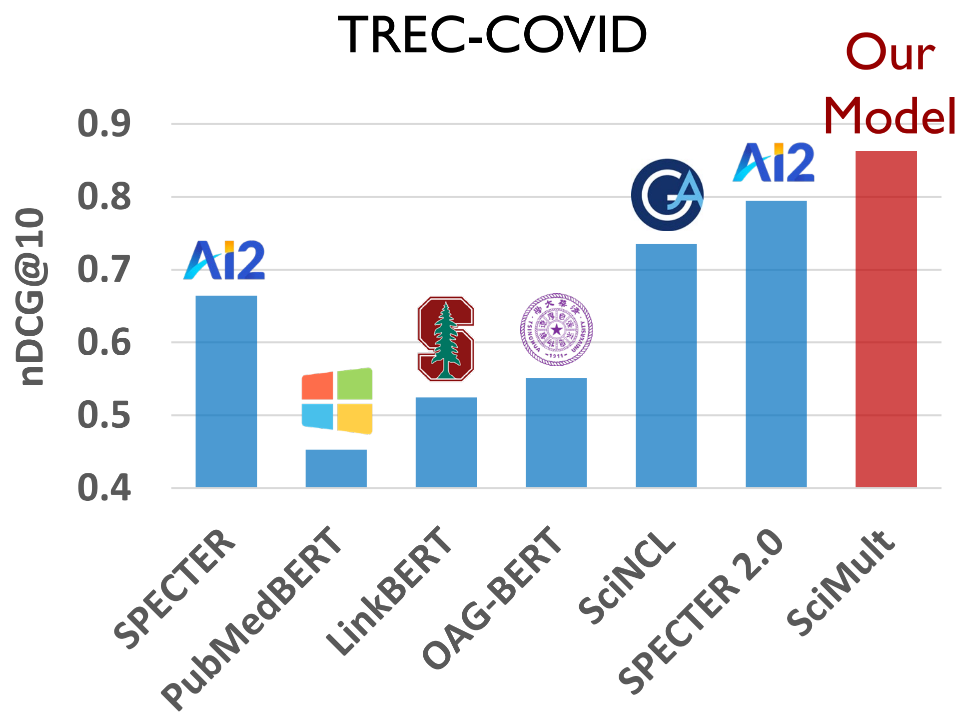
Comparison with Previous Approaches

- 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

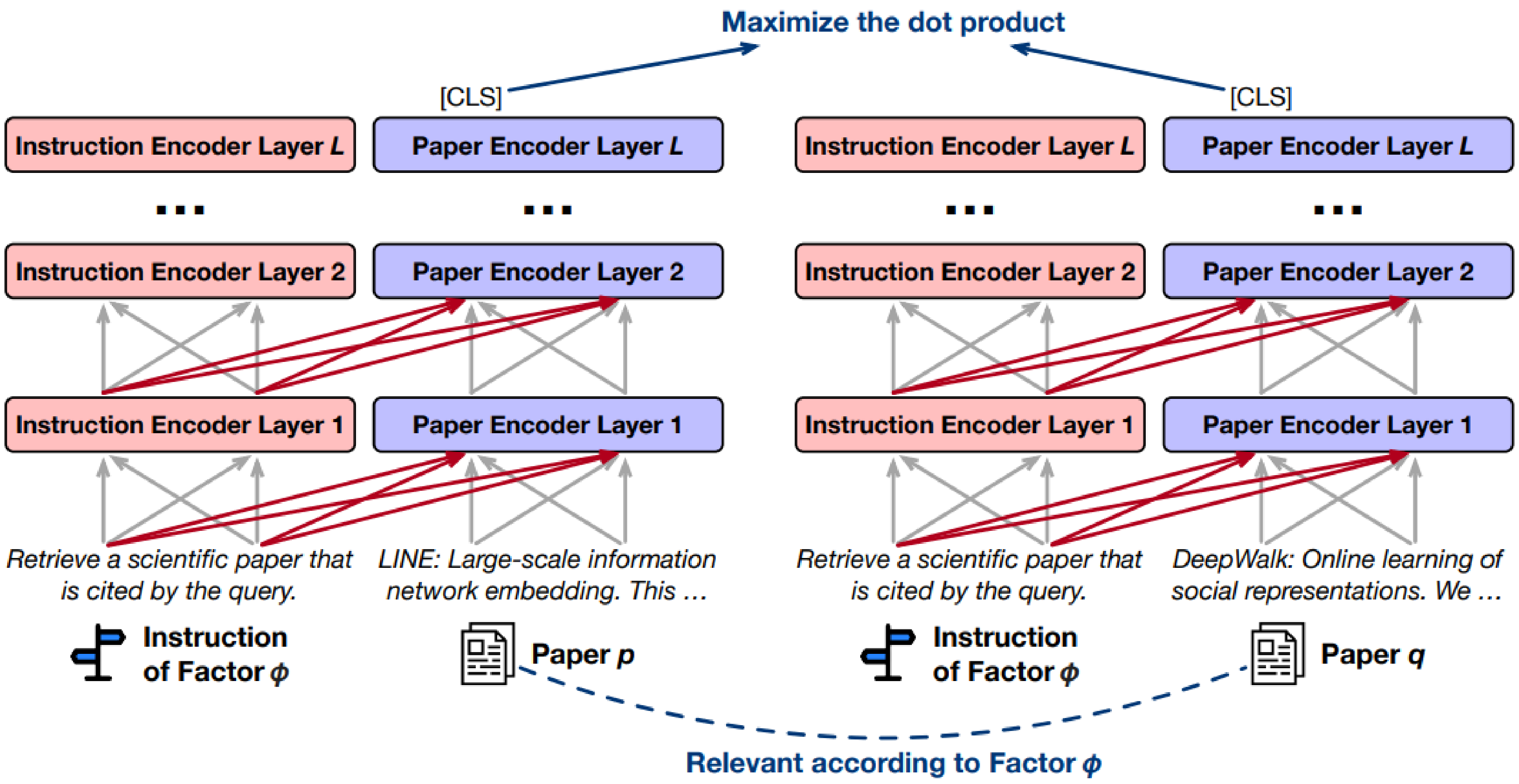
https://pmc-patients.github.io/



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

Tackling Task Interference: Instruction Tuning

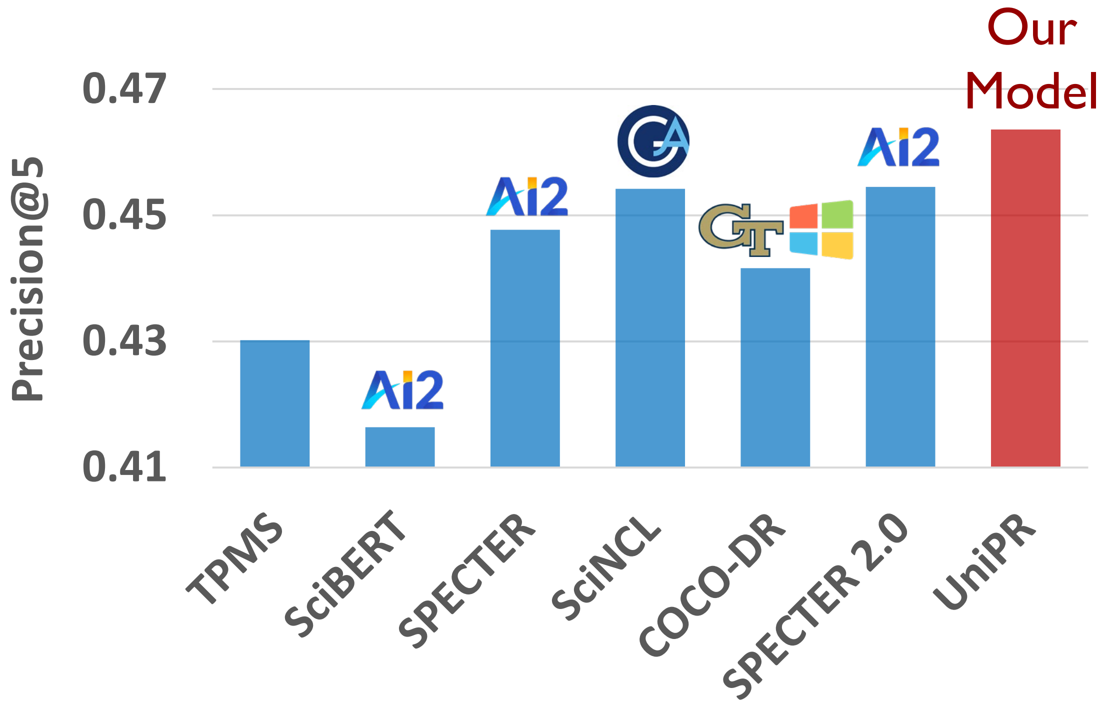
- Using a factor-specific instruction to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.



Zhang et al., ""Why Should I Review This Paper?" Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching." arXiv 2023.

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)



Zhang et al., ""Why Should I Review This Paper?" Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching." arXiv 2023.

Scientific Language Models: A Survey

Awesome-Scientific-Language-Models Public

A Curated List of Language Models in Scientific Domains

271 15

https://github.com/yuzhimanhua/Awesome-Scientific-Language-Models

Awesome Scientific Language Models

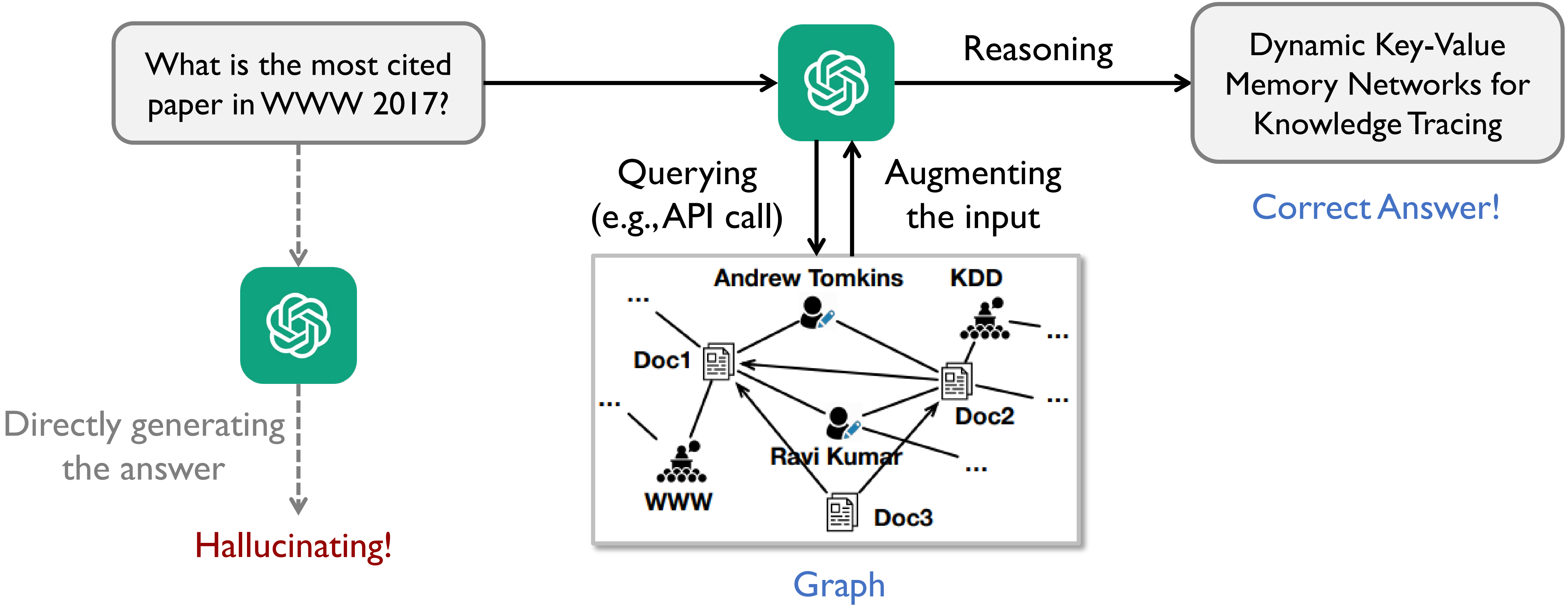
awesome Stars 271

PaperNumber 221 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**, **graph**, **molecule**, **protein**, **genome**, and **climate time series**). The repository will be continuously updated.

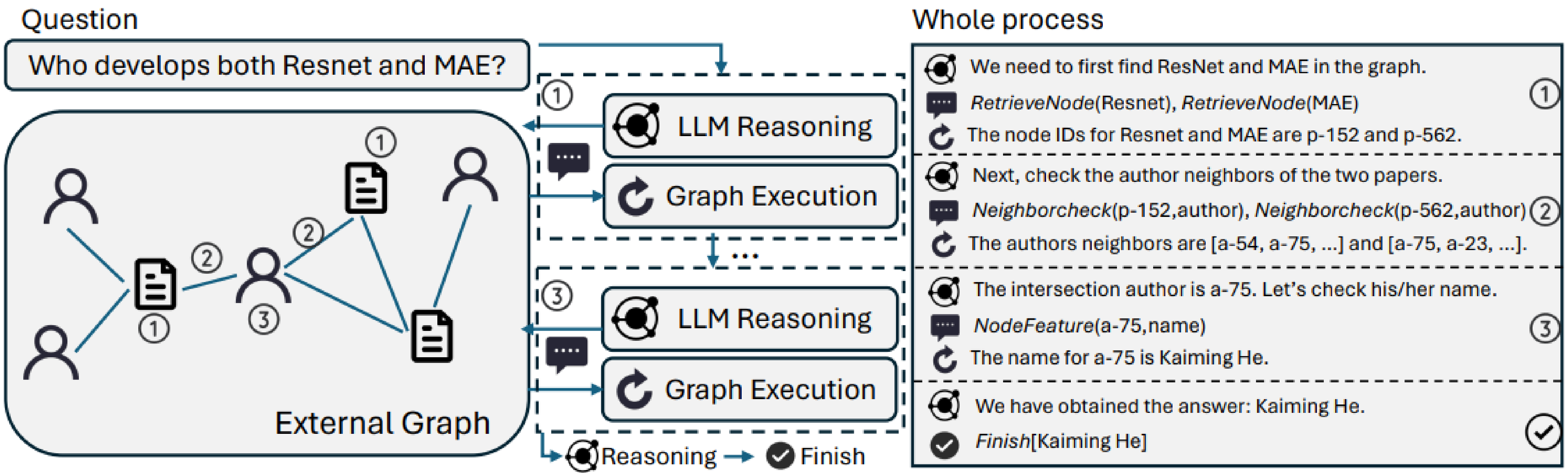
Looking Back to the Motivating Example

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



Initial Trial: Graph Chain-of-Thoughts

- *RetrieveNode*(Text): Identify related nodes in the graph with semantic search.
- *NeighborCheck*(NodeID, NeighborType): Return the neighboring information in the graph for a specific node.
- *NodeFeature*(NodeID, FeatureName): Extract the textual feature information from the graph for a specific node.



Jin et al., "Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs." arXiv 2024.

Comparison with Previous Approaches

- Easy questions: one-node / one-hop
  - "Who are the authors of {paper}?"
- Medium questions: multi-hop
  - "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}?"

	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
Graph RAG	Mixtral-8x7b	27.77	31.20	32.87	37.00	20.08	33.33	8.66	15.19	23.48	25.56
Graph RAG	GPT-3.5-turbo	18.45	26.98	17.52	28.00	14.94	24.17	8.69	14.07	18.66	22.22
	GRAPH-COT	**31.89**	**33.48**	**42.40**	**44.50**	**41.59**	**46.25**	**22.33**	**28.89**	**30.52**	**28.33**

Jin et al., "Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs." arXiv 2024.

Thank you! Questions?