

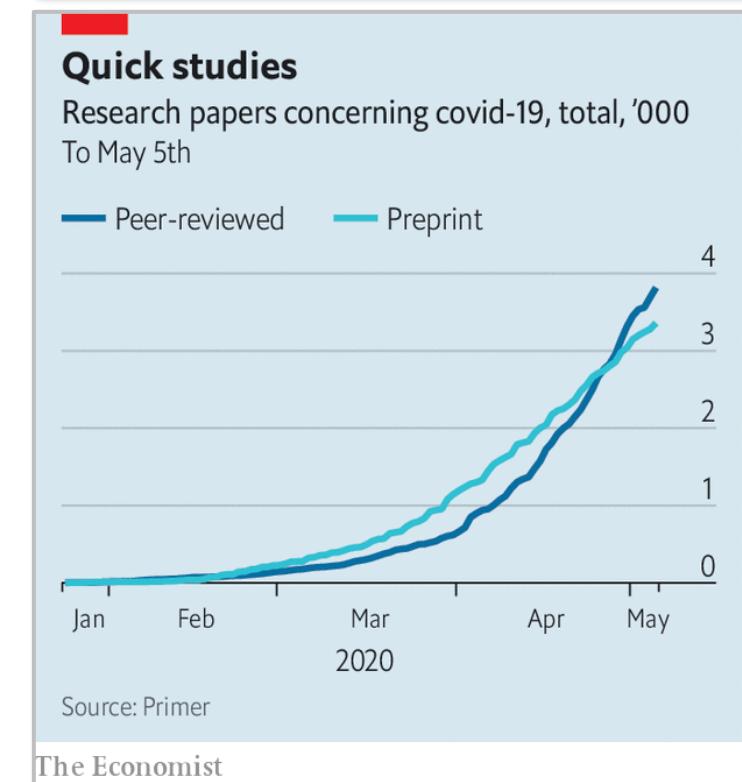
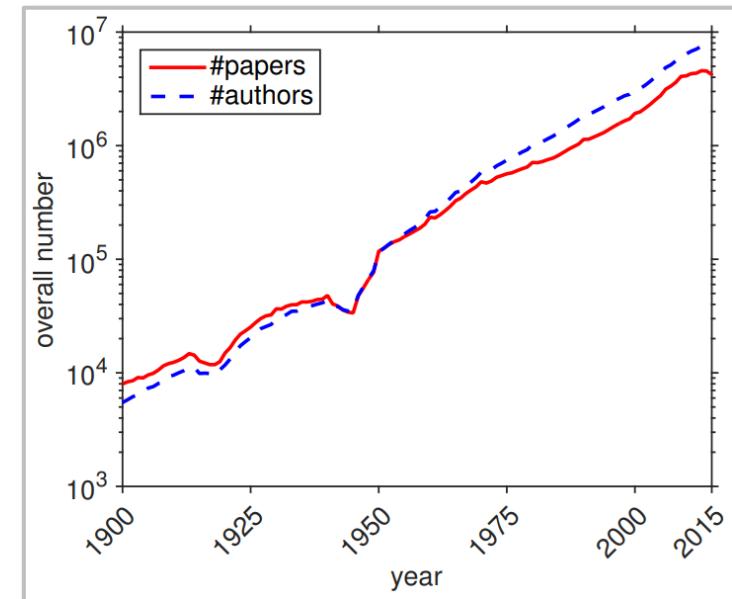
# Graph-Enhanced Text Mining for Biomedicine

Yu Zhang

November 7, 2024

# Explosion of Biomedical 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?



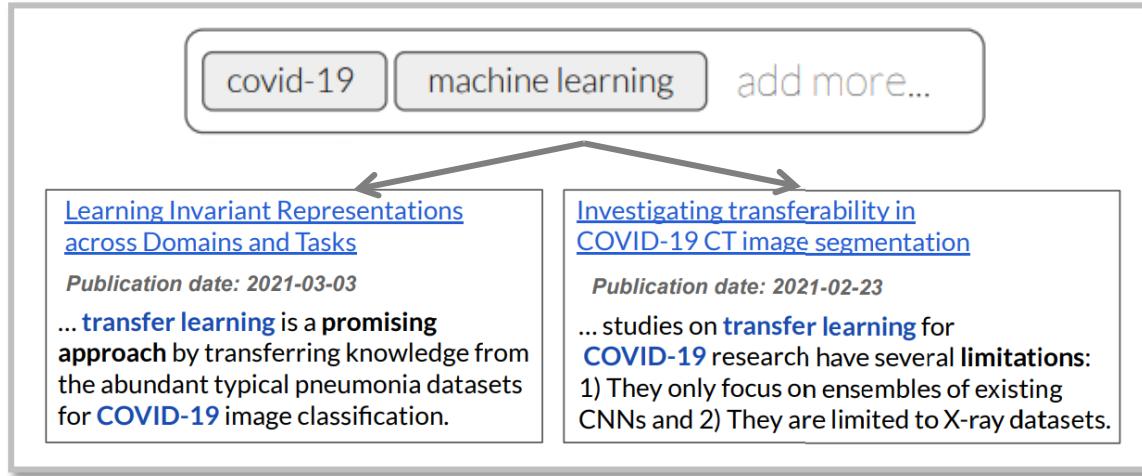
[1] "A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations." KDD 2017.

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

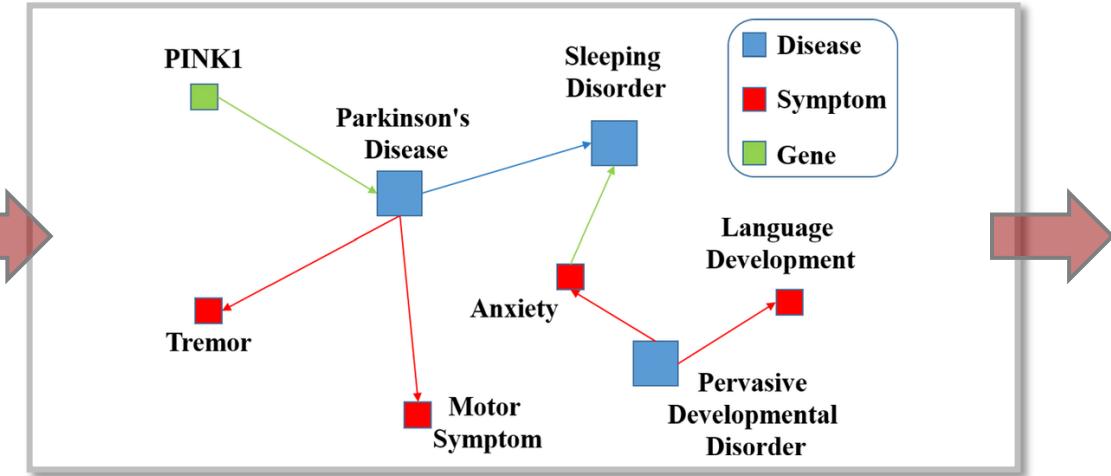
[3] <https://www.economist.com/science-and-technology/2020/05/07/scientific-research-on-the-coronavirus-is-being-released-in-a-torrent>

# How can text mining help biomedical research?

## Retrieving and Analyzing Relevant Literature



## Uncovering Knowledge Structures/Graphs



- 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 biomedical research?

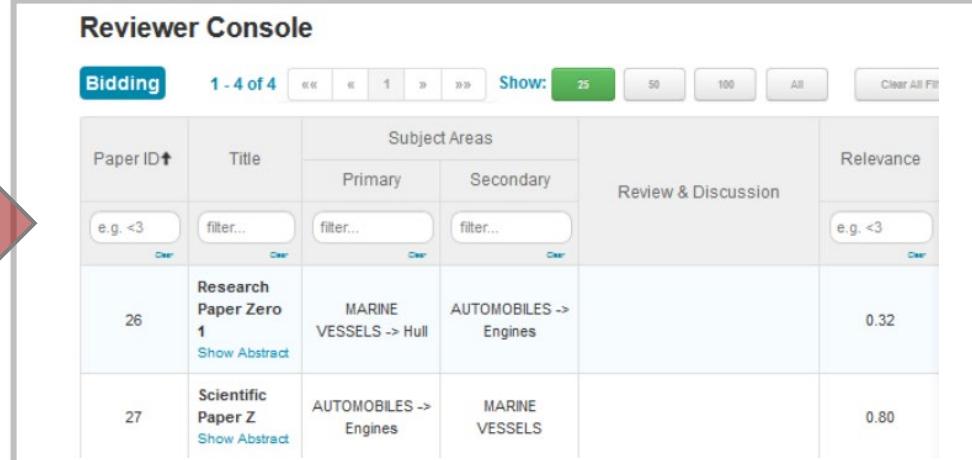
## Generating Hypotheses and Suggesting Directions



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

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

## Reviewing Research Outcomes

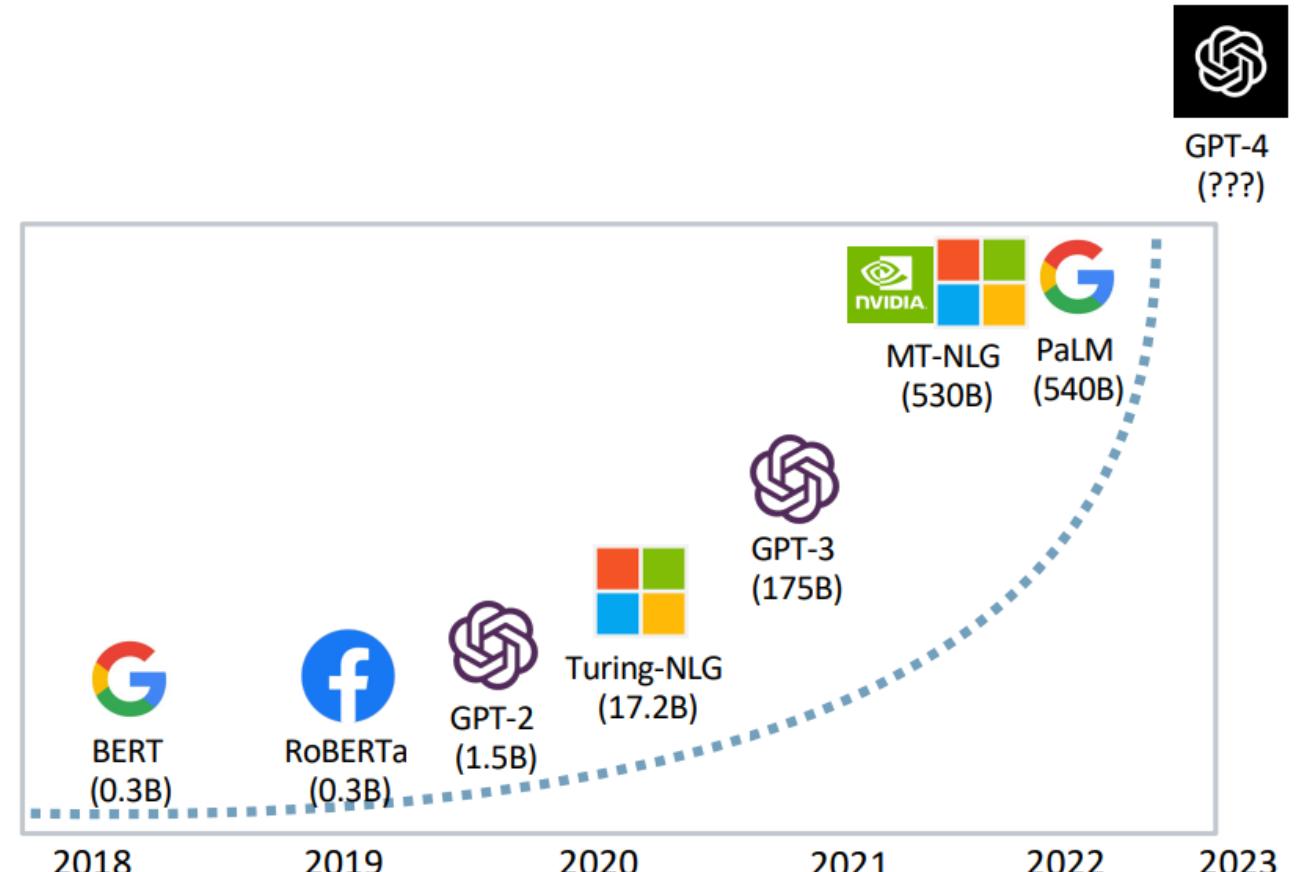


| Reviewer Console |  | Bidding 1 - 4 of 4     |                        |                     |           | Show: 25 | 50 | 100 | All | Clear All Fil |
|------------------|--|------------------------|------------------------|---------------------|-----------|----------|----|-----|-----|---------------|
| Paper ID         | Title  | Subject Areas          |                        | Review & Discussion | Relevance |          |    |     |     |               |
|                  |  | Primary                | Secondary              |                     |           |          |    |     |     |               |
| 26               | Research Paper Zero 1<br><a href="#">Show Abstract</a> | MARINE VESSELS -> Hull | AUTOMOBILES -> Engines |                     | 0.32      |          |    |     |     |               |
| 27               | Scientific Paper Z<br><a href="#">Show Abstract</a>    | AUTOMOBILES -> Engines | MARINE VESSELS         |                     | 0.80      |          |    |     |     |               |

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

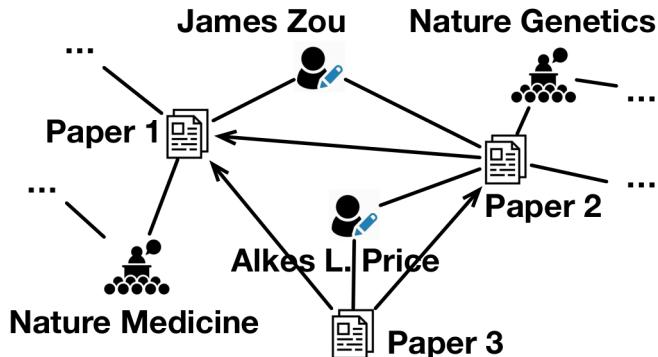
# Pre-trained Language Models (PLMs) for Biomedical Text Mining

- 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*)
  - ...

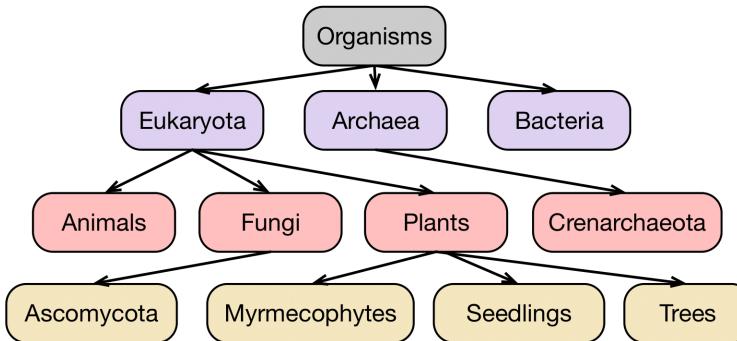


Are PLMs aware of **graph information**?

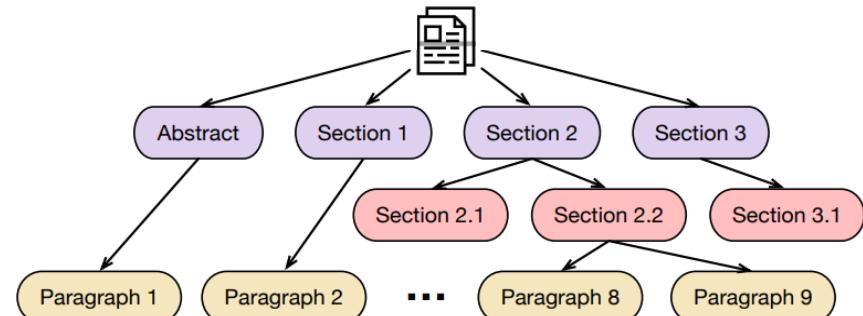
# Graph Information Associated with Biomedical Text



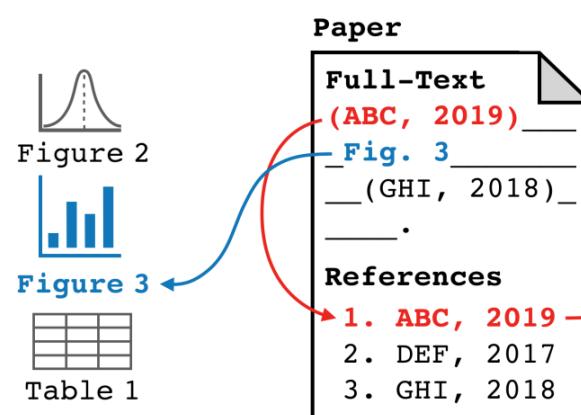
Metadata



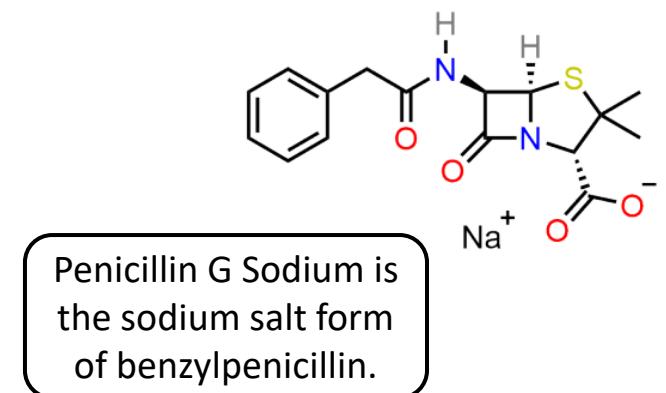
Taxonomy



In-Document Structure



Knowledge Graph



Text-Paired Molecule/Protein

# PLMs may not be graph-aware!



You

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



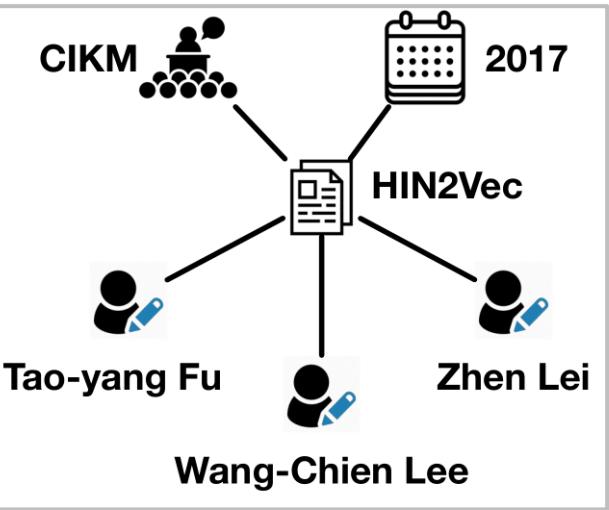
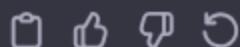
ChatGPT

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.

# PLMs may not be graph-aware!



What is the most cited paper published in *Nature Genetics* in 2018?  
Please list its title and authors.



Searched 1 site ▾

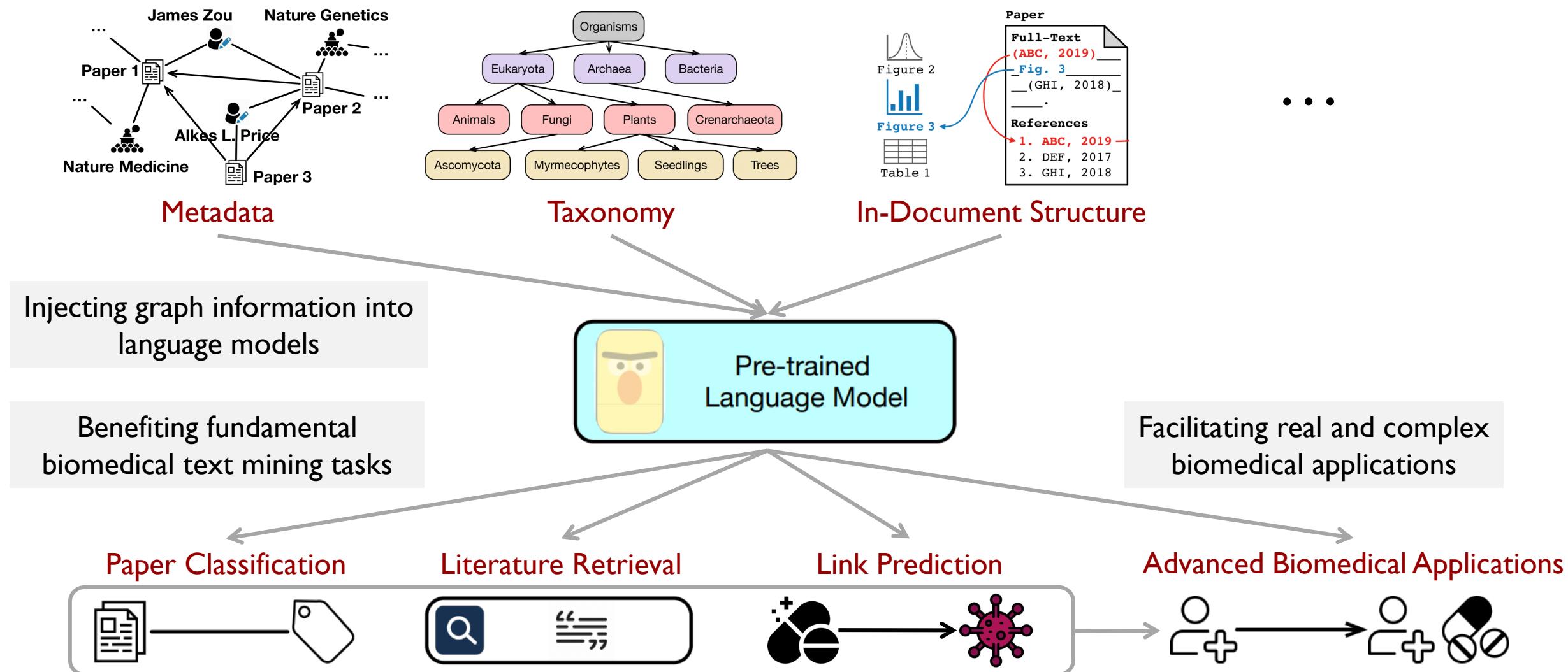
The most cited paper published in *Nature Genetics* in 2018 is titled "Analysis of protein-coding genetic variation in 60,706 humans." This landmark study, authored by Karczewski et al., from the Genome Aggregation Database (gnomAD) Consortium, analyzed extensive genetic data to identify variants linked to diseases and improve understanding of human genetic diversity. This paper has been widely influential in genetic and biomedical research due to its implications for genetic diagnostics and precision medicine PAPER DIGEST.



ChatGPT 4o, queried on November 5, 2024

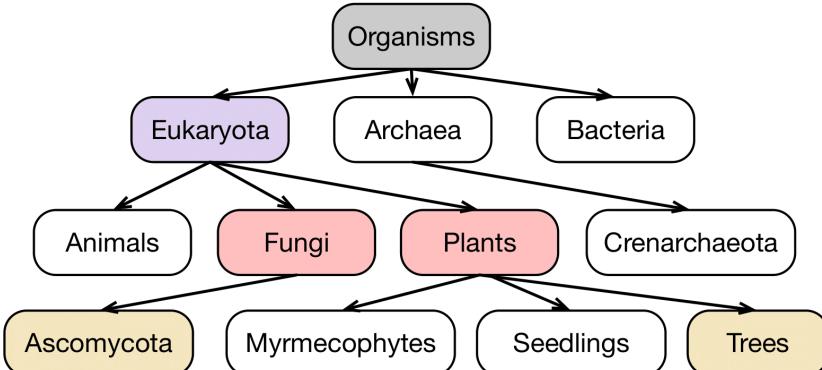
This paper was published in **Nature 2016**  
rather than **Nature Genetics 2018**.

# Today's Lecture: Overview

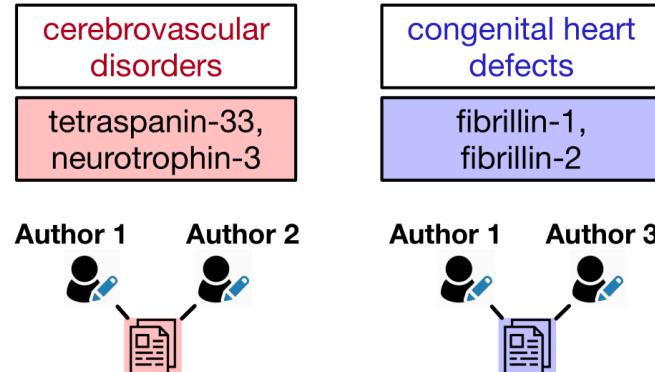


# Today's Lecture: Overview

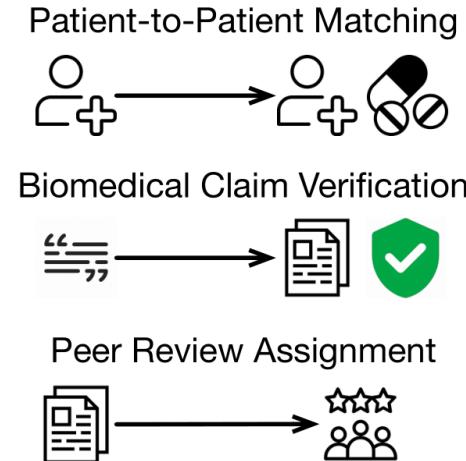
## Part I: Extremely Fine-Grained Biomedical Paper Classification



## Part II: Category-Guided Biomedical Topic Mining

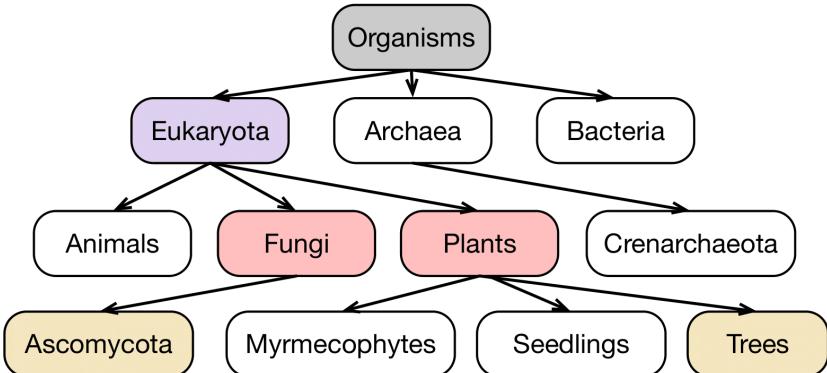


## Part III: Advanced Biomedical Applications

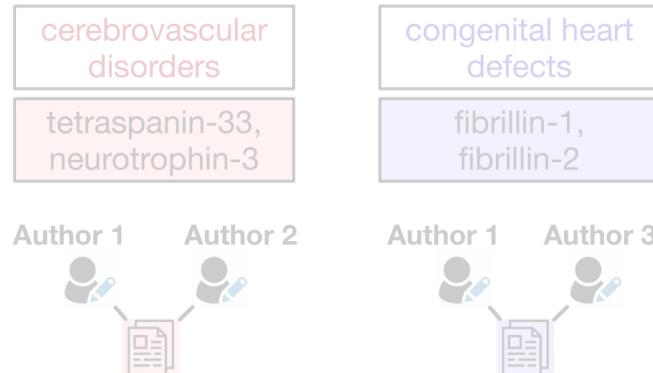


# Today's Lecture: Part I

## Part I: Extremely Fine-Grained Biomedical Paper Classification



## Part II: Category-Guided Biomedical Topic Mining



## Part III: Advanced Biomedical Applications



# Extremely Fine-Grained Biomedical Paper Classification



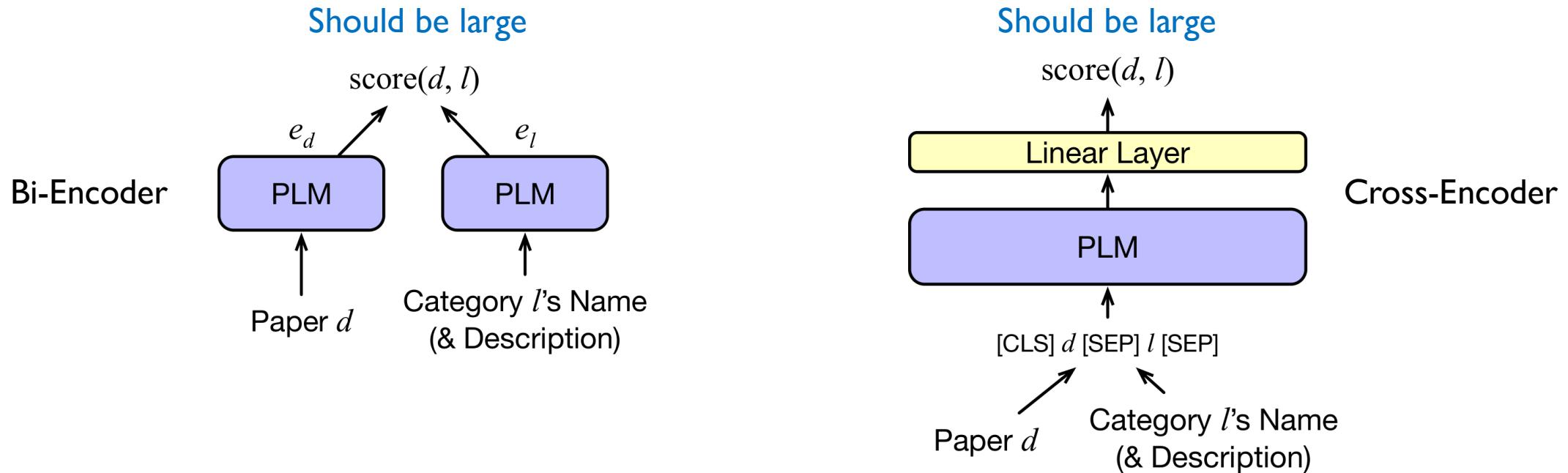
- The Microsoft Academic Graph has **740K+** categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain **30K+** categories.
- Each paper can be relevant to **more than one** category (5-15 categories for most papers).

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  $\sim 10$  relevant categories from  $\sim 100,000$  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



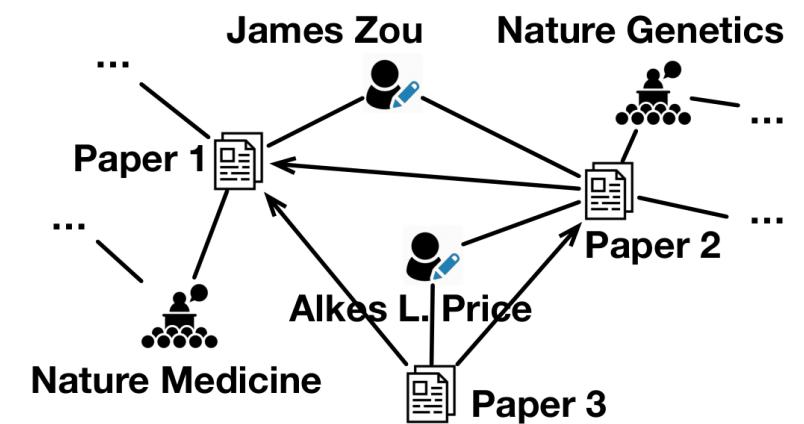
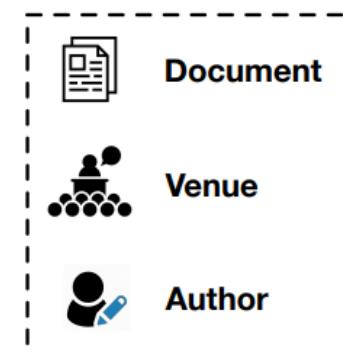
(b) meta-path: P->P<-P



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

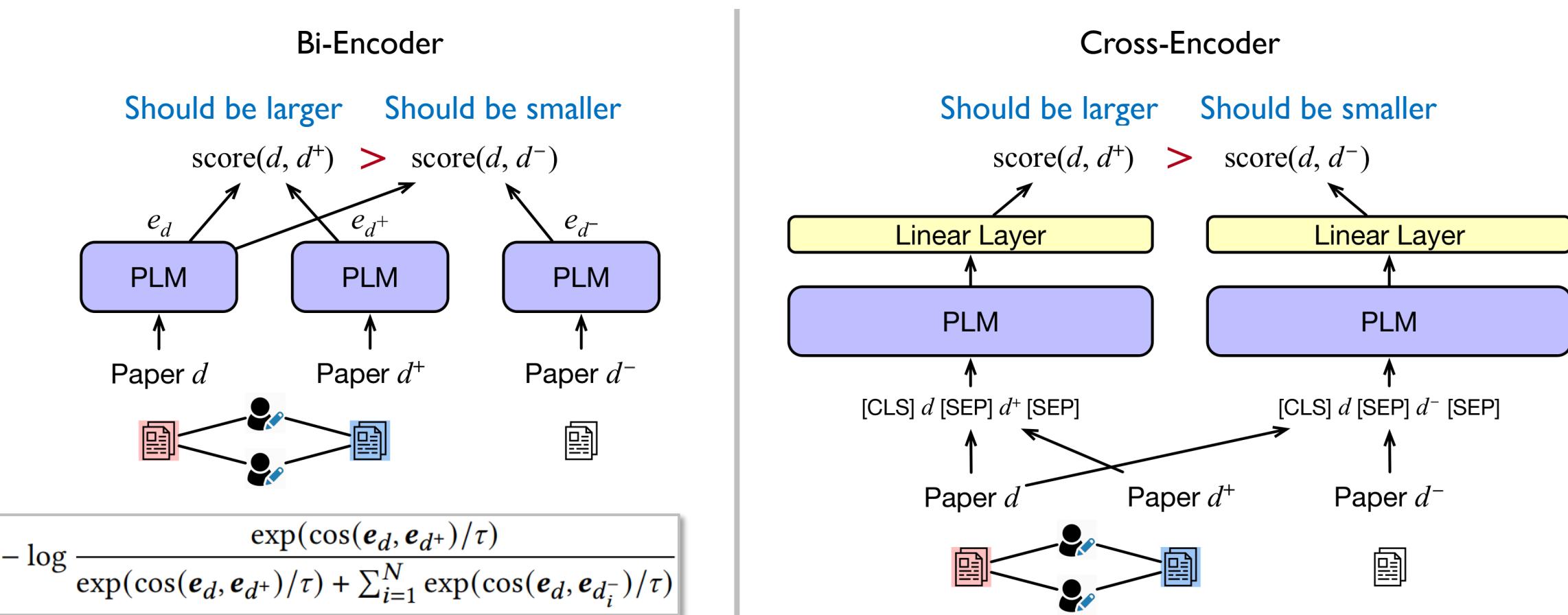


(d) meta-graph: P<-(PP)->P



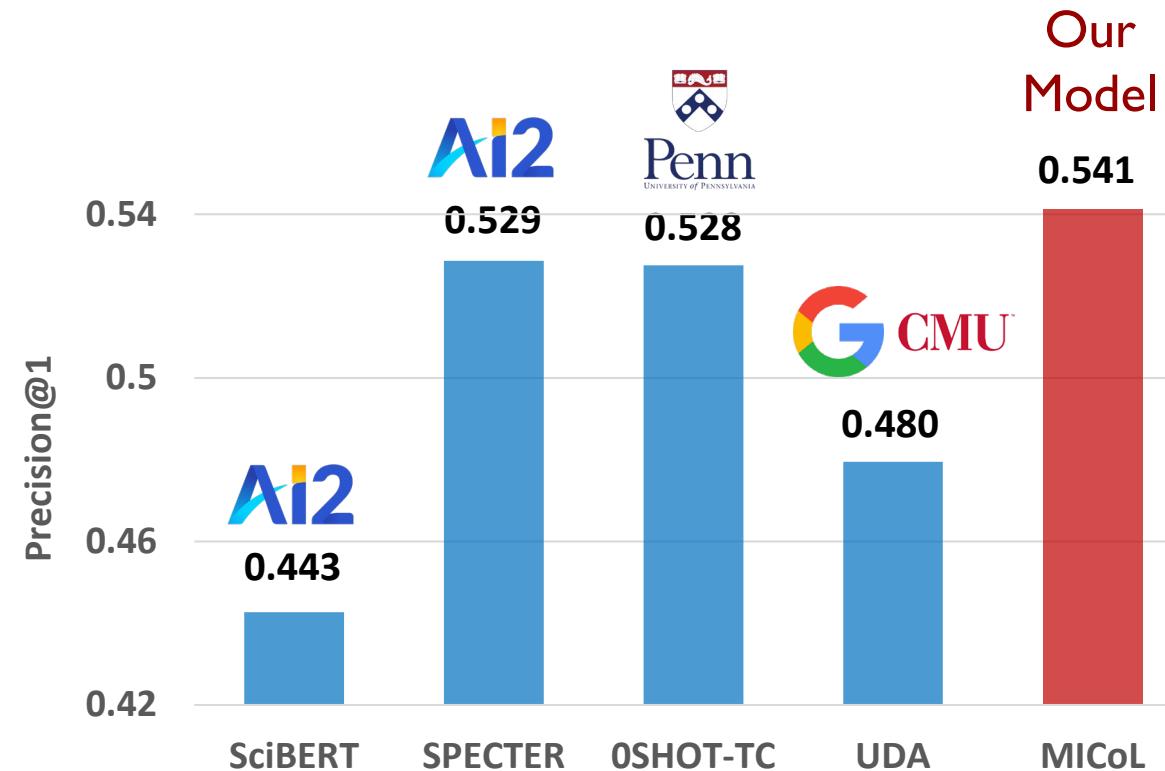
# Graph-Induced Contrastive Learning

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



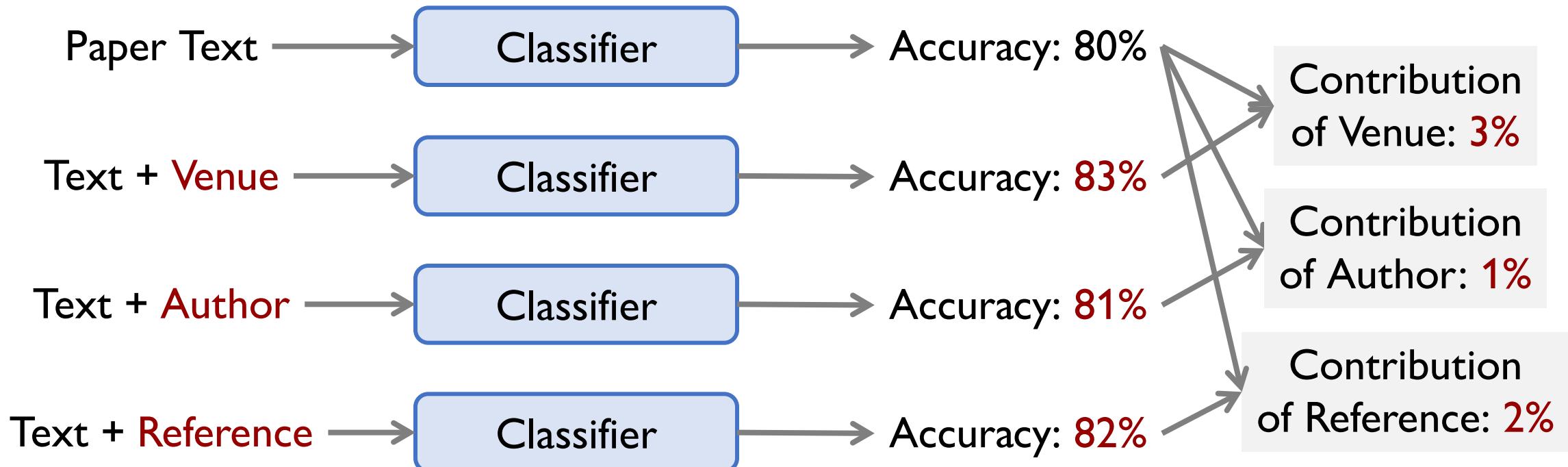
# Comparison with Previous Approaches

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



# Which type of metadata is the most helpful?

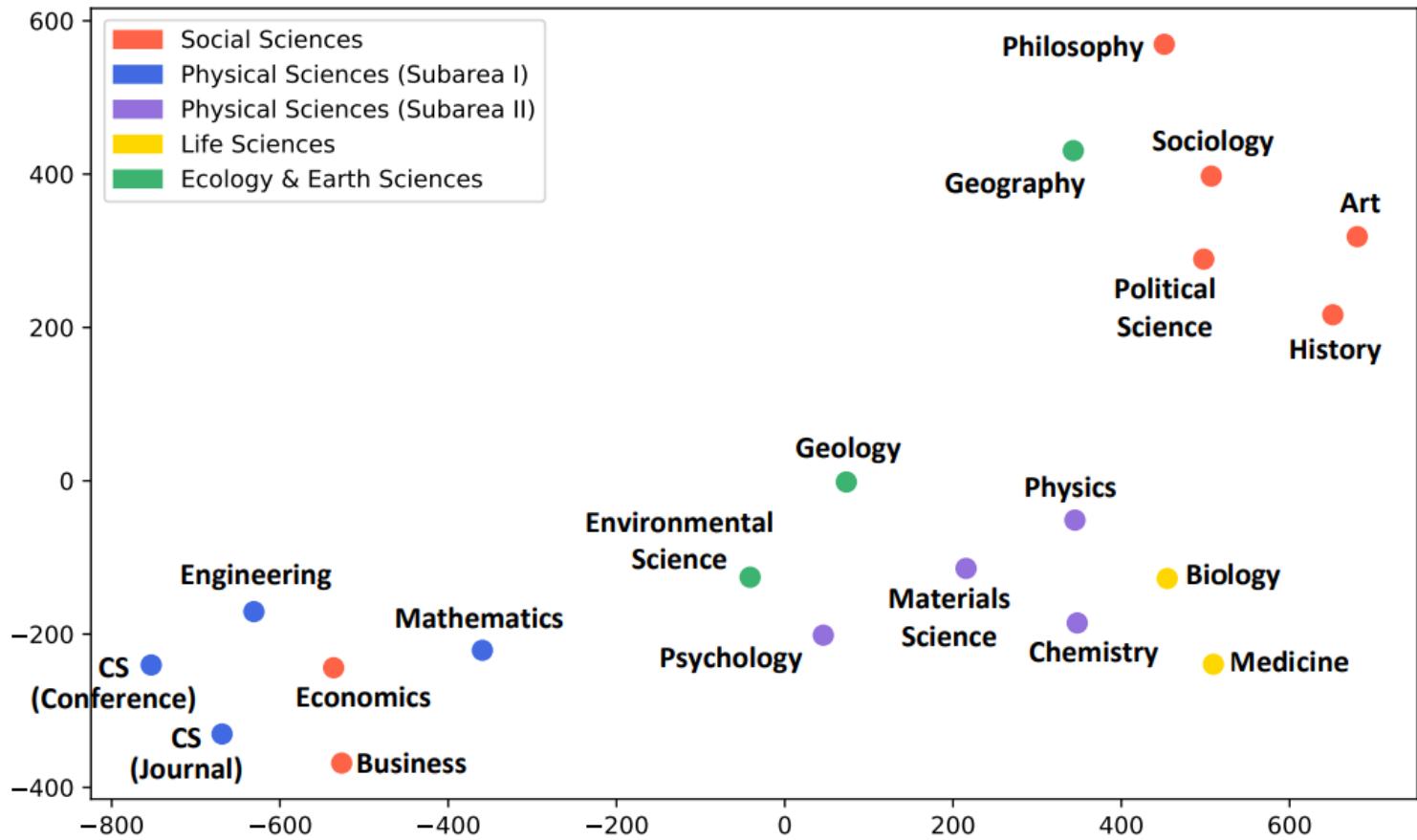
- Is the **contribution** of venues, authors, and references to paper classification consistent **across different fields**?



The effect of metadata: [3%, 1%, 2%]

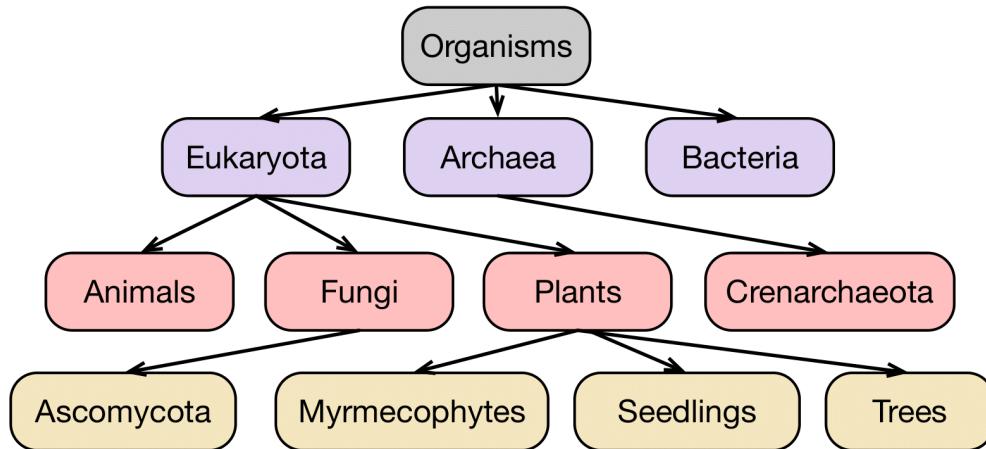
# Which type of metadata is the most helpful?

- 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.



# How about other types of graph information?

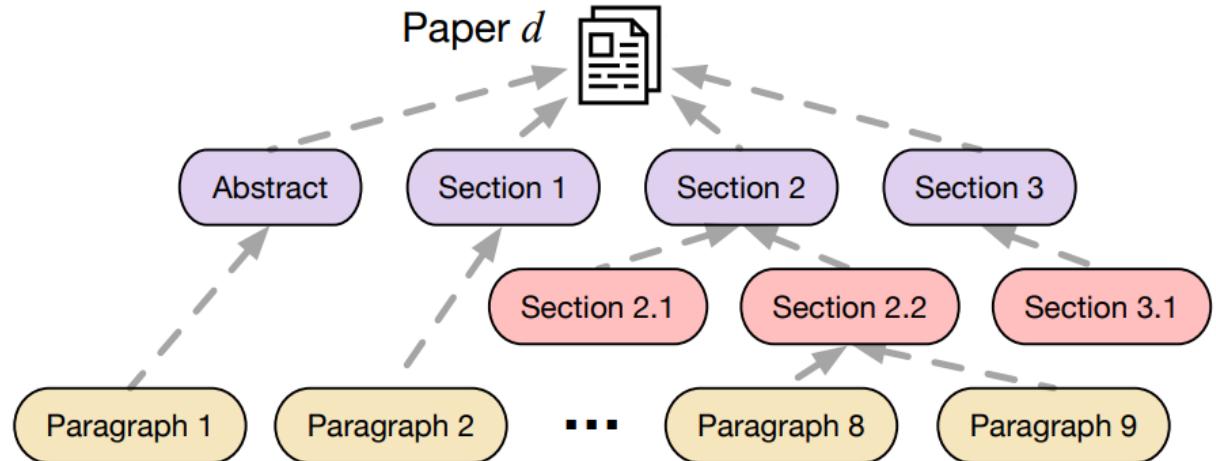
Label Hierarchy



Top-Down Pruning:

Irrelevant to **Plants**  $\Rightarrow$  Irrelevant to **Trees**

In-Document Structure

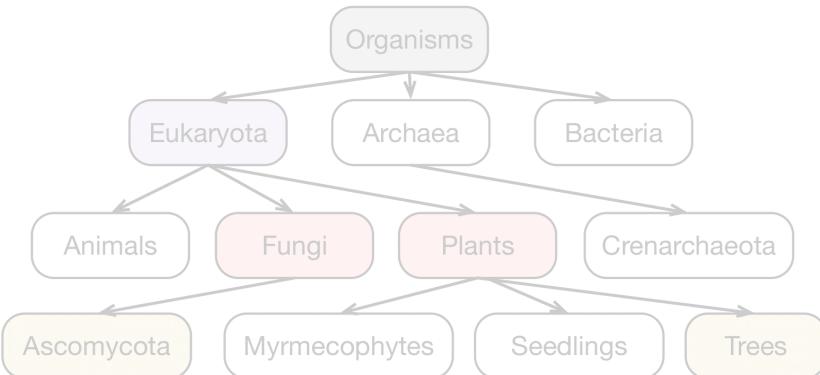


Bottom-Up Aggregation:

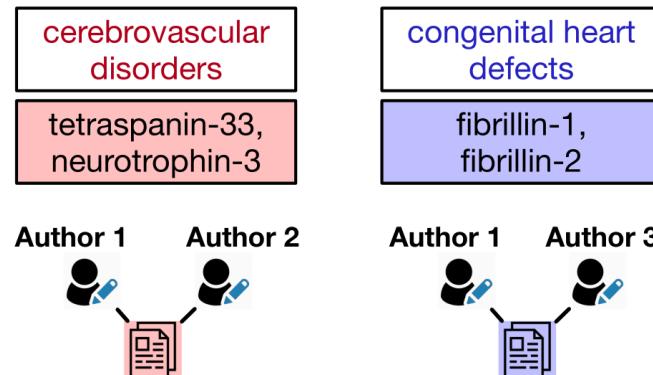
**Paragraphs**  $\rightarrow$  **Subsections**  $\rightarrow$  **Sections**  $\rightarrow$  **Paper**

# Today's Lecture: Part II

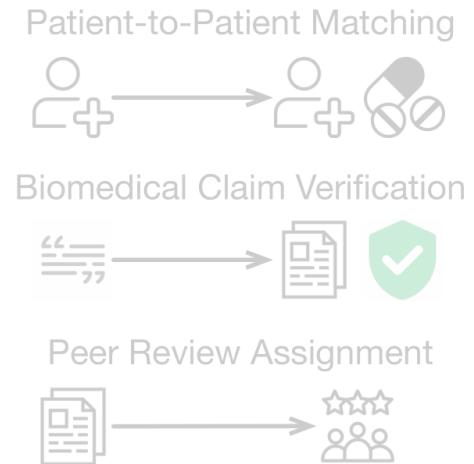
Part I: Extremely  
Fine-Grained  
Biomedical Paper  
Classification



## Part II: Category-Guided Biomedical Topic Mining



Part III: Advanced  
Biomedical  
Applications

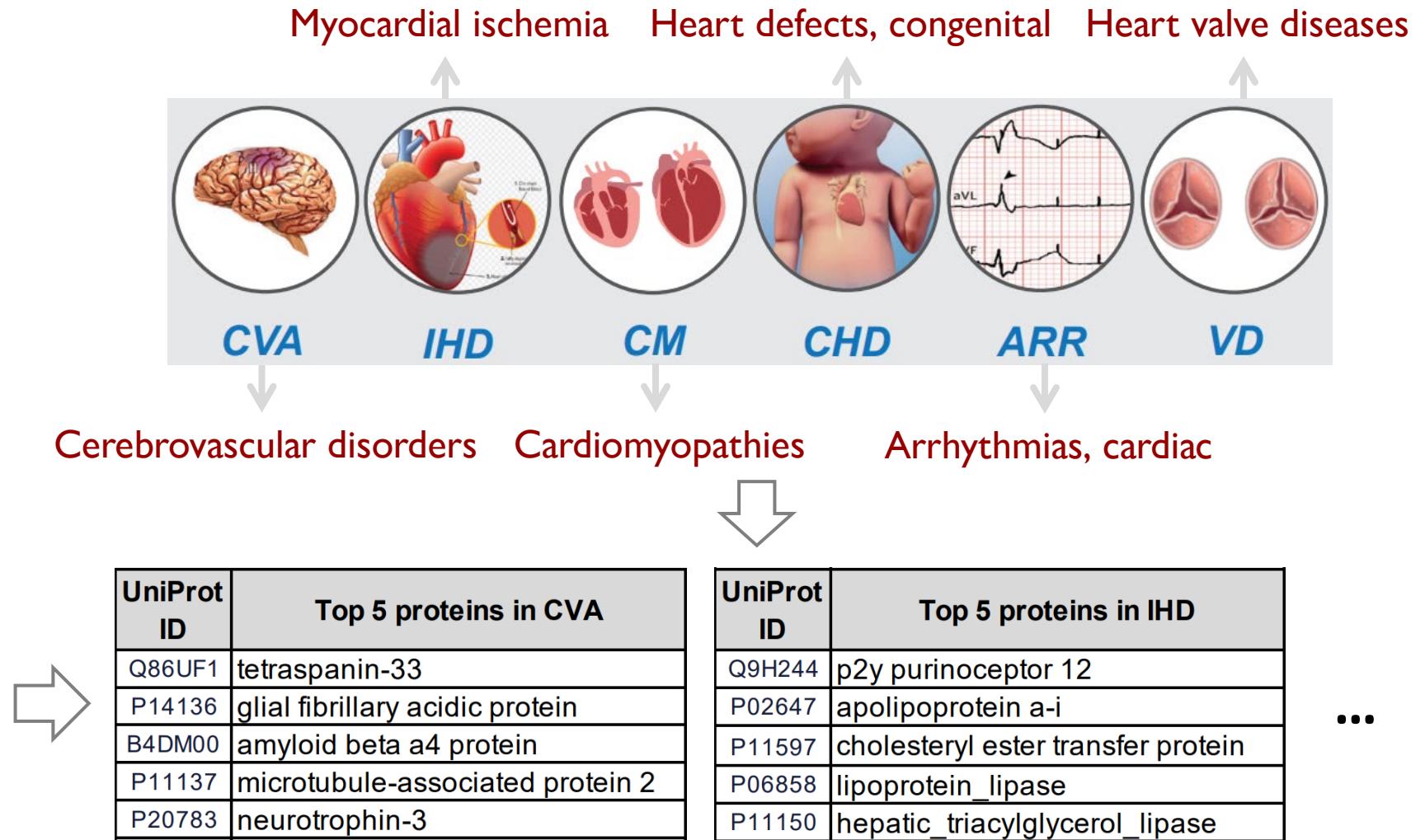


# Category-Guided Scientific Topic Mining: Motivation



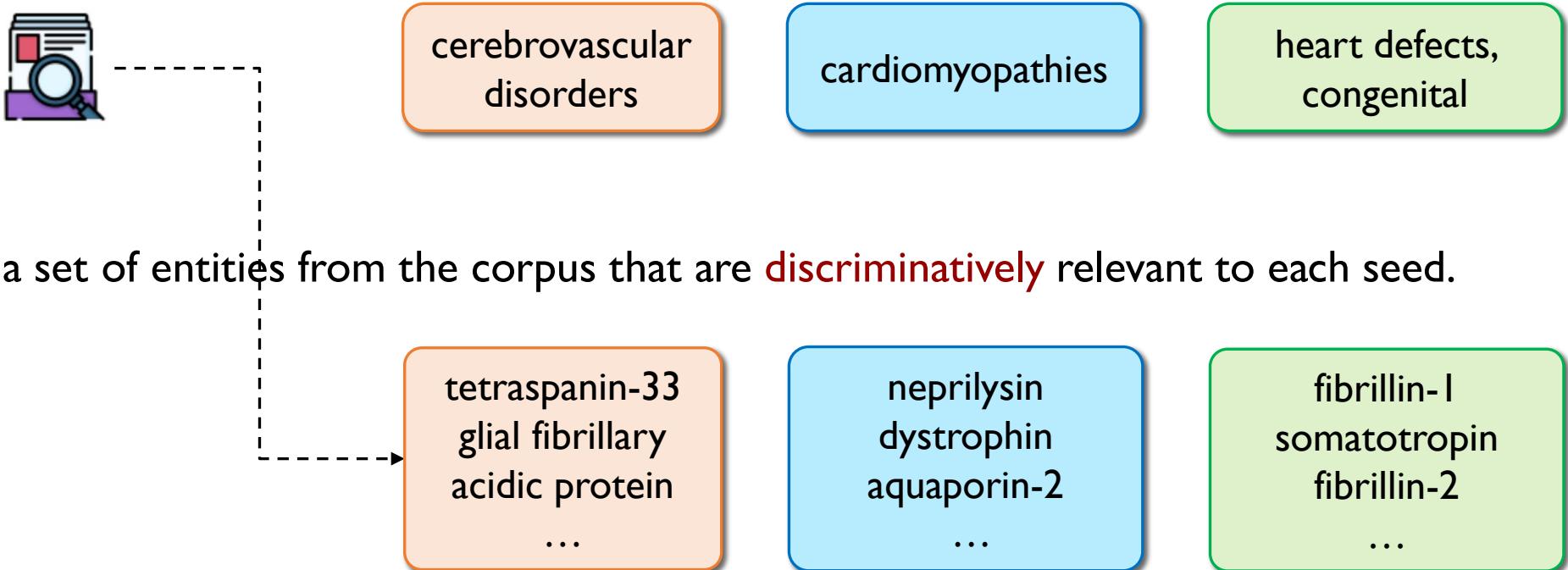
David Liem, M.D., Ph.D.,  
UC Davis

1M+ Cardiology  
Papers      Protein  
Entities



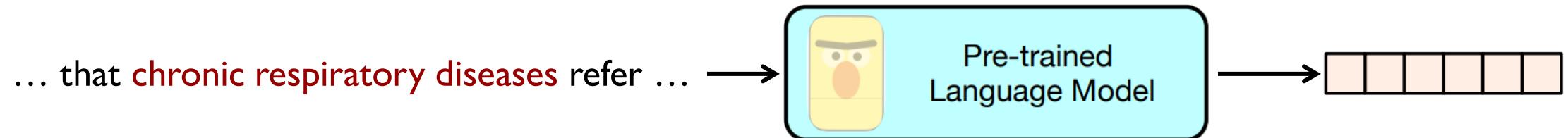
# Category-Guided Scientific Topic Mining: Task Definition

- Input: (1) A large corpus (2) The user specifies a set of category names.

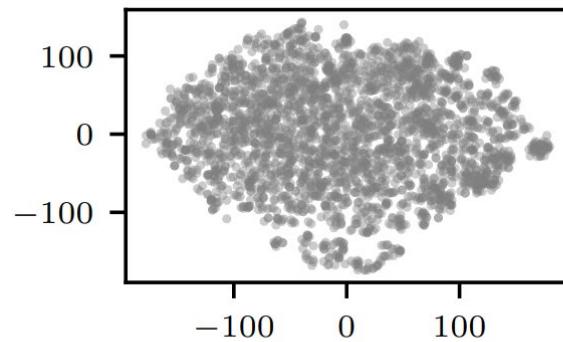


# A Naïve Approach based on a Pre-trained Language Model

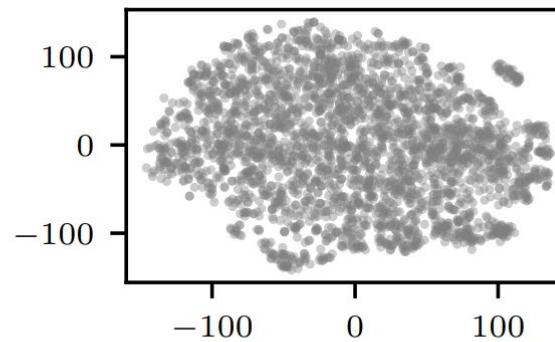
- Step 1: Using a pre-trained language model to encode each category name and each candidate entity



- Step 2: Finding the nearest entity neighbors for each category name
- **Drawback:** No clearly separated clusters in the obtained vector space



(a) New York Times.

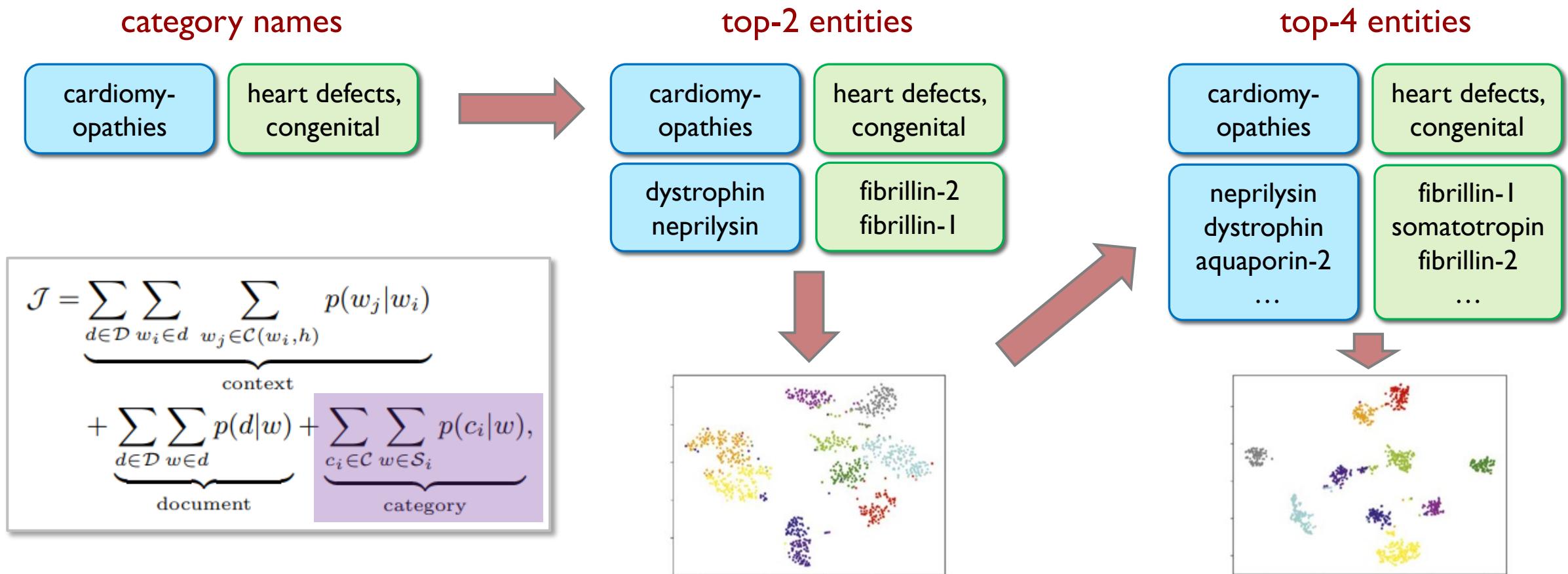


(b) Yelp Review.

**Theorem:** The Masked Language Modeling objective of BERT assumes that the learned contextualized embeddings are generated from a Gaussian Mixture Model with  $|V|$  mixture components, where  $|V|$  is the vocabulary size of BERT.

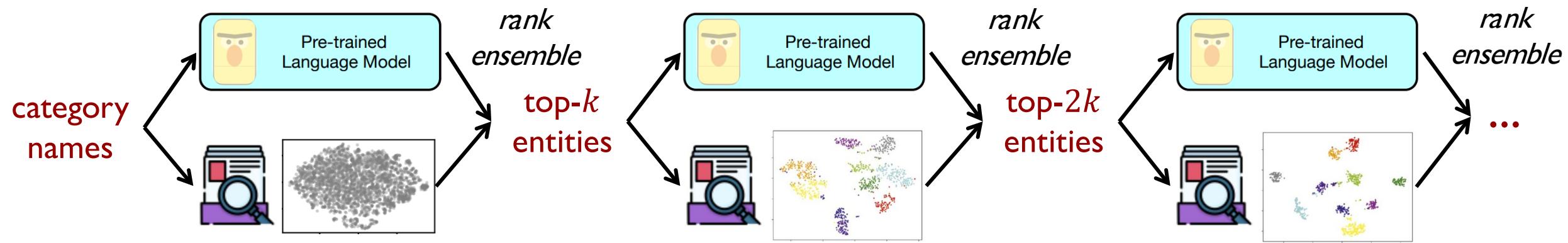
# Category-Guided Embedding Learning

- Seeking domain-specific semantic signals from the input corpus (e.g., 1M+ cardiology papers)
- Gradually expanding each topic



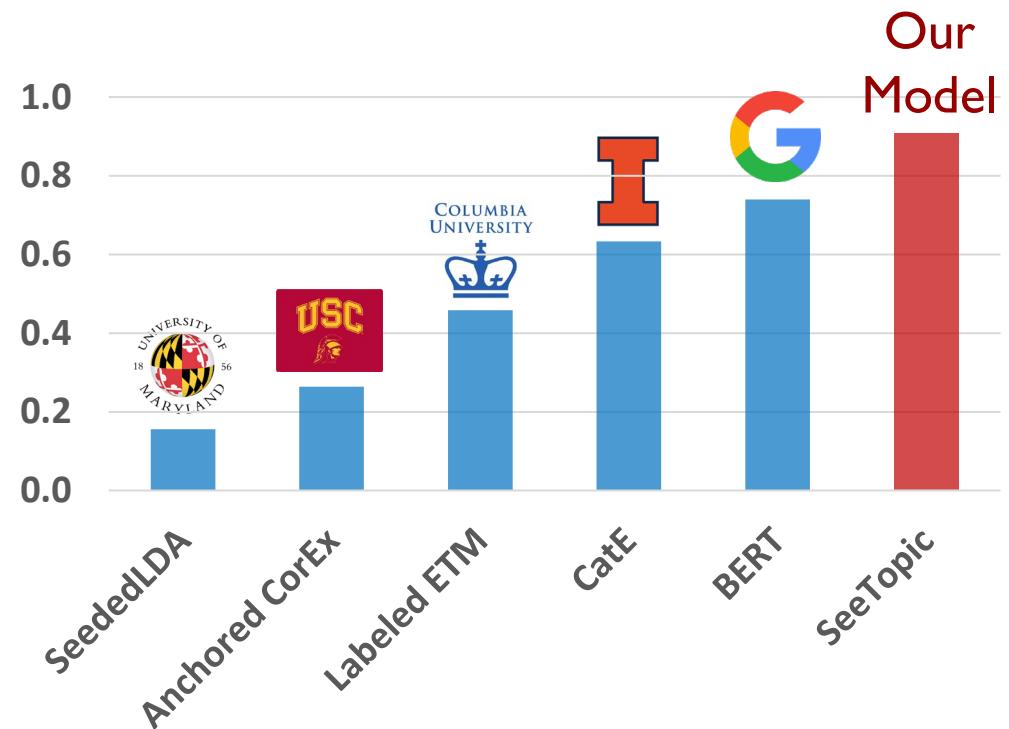
# Jointly Leveraging PLM and Category-Guided Embedding Learning

- General knowledge learned by PLM from web-scale pre-training corpora
- Domain-specific semantics learned from the input corpus



# Comparison with Previous Approaches

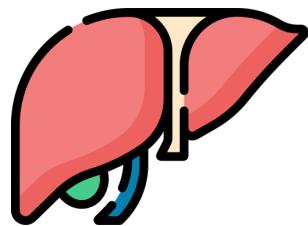
- Dataset: SciDocs (11 disease categories)
- Metric: Precision@10
  - Each method needs to find 10 entities for each disease category. Annotators judge how many of these entities are discriminatively relevant to the category.



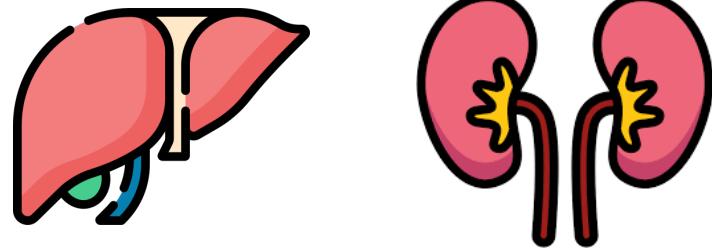
# Case Study

- Dataset: SciDocs (11 disease categories)
  - *Hepatitis A/B/C/E*
  - *Chronic Kidney Disease*
  - ...
- Baseline: Labeled ETM [Dieng, Ruiz, and Blei, TACL 2020]

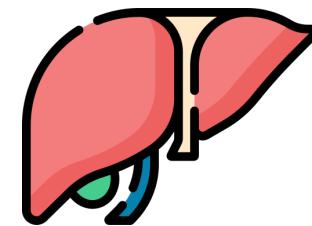
Category Name:  
*Hepatitis A/B/C/E*



Top-ranked Entity by Labeled ETM:  
*alanine aminotransferase (ALT)* ✗

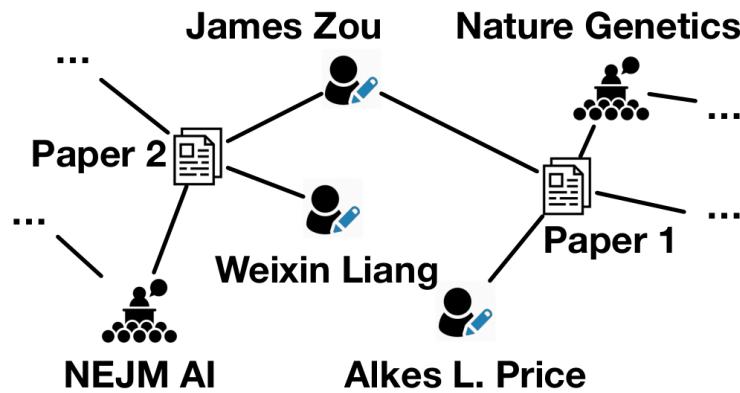


Top-ranked Entity by Our Model:  
*Hepatitis B e antigen (HBeAg)* ✓



# From Text to Graph

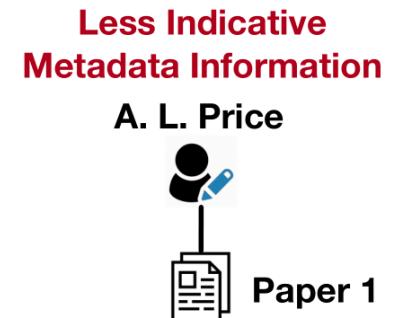
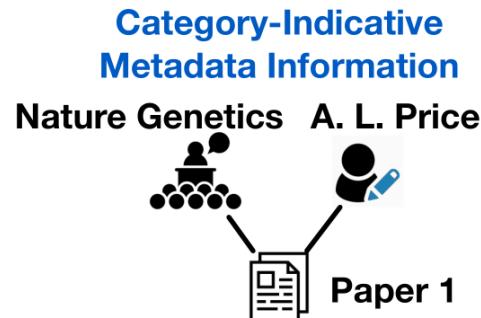
- Given some category names, how to find category-indicative graph signals?
- A graph version of category-guided topic mining!
  - Seeds: Category names
  - Candidate “entities”: Metadata or their combinations
- Benefitting text classification



---

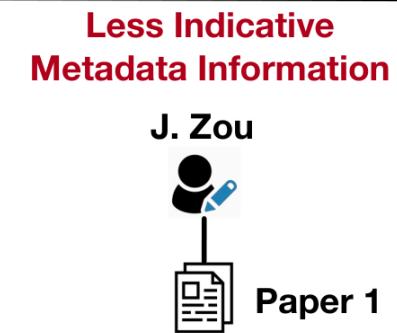
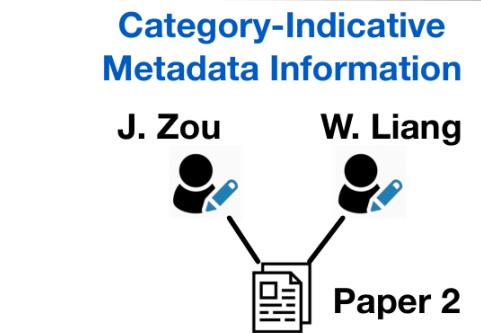
Paper 1: Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data  
Label: RNA-seq

---



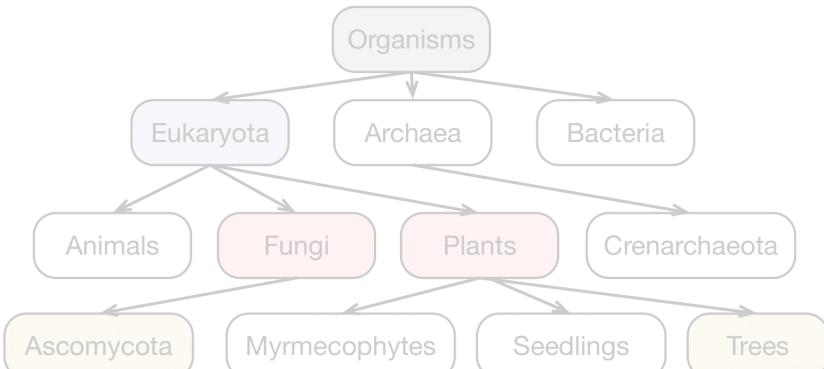
Paper 2: Can large language models provide useful feedback on research papers? A large-scale empirical analysis  
Label: Large Language Models

---

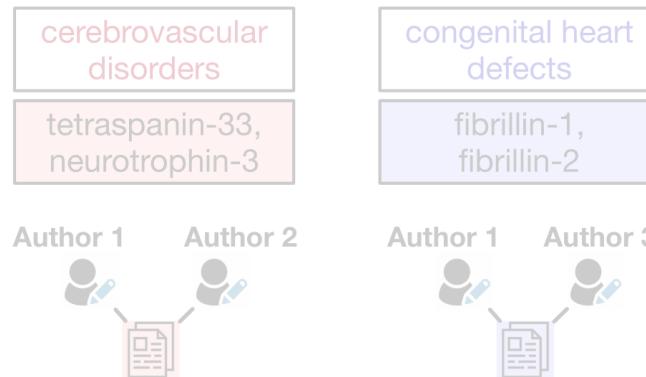


# Today's Lecture: Part III

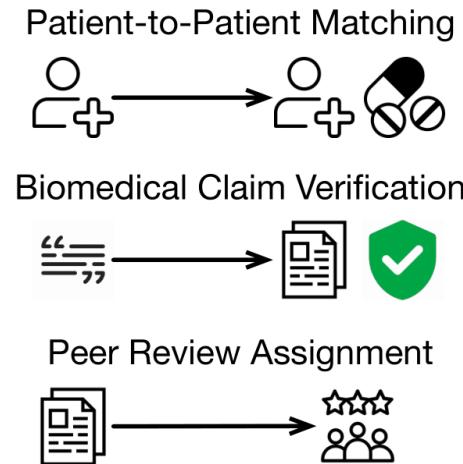
## Part I: Extremely Fine-Grained Biomedical Paper Classification



## Part II: Category-Guided Biomedical Topic Mining



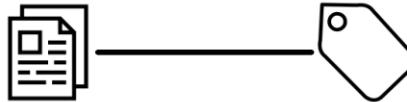
## Part III: Advanced Biomedical Applications



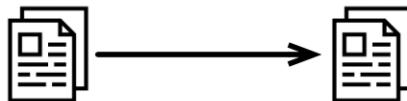
# Facilitating Complex Biomedical Applications

## Fundamental Biomedical Text Mining Tasks

Paper Classification



Link Prediction

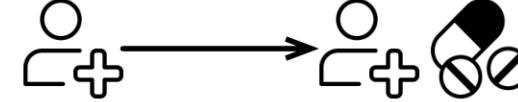


Literature Retrieval

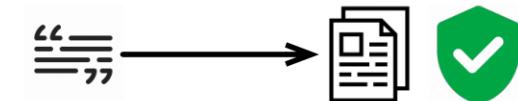


## Advanced Biomedical Applications for Scientific Discovery

Patient-to-Patient Matching



Biomedical Claim Verification



Peer Review Assignment



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

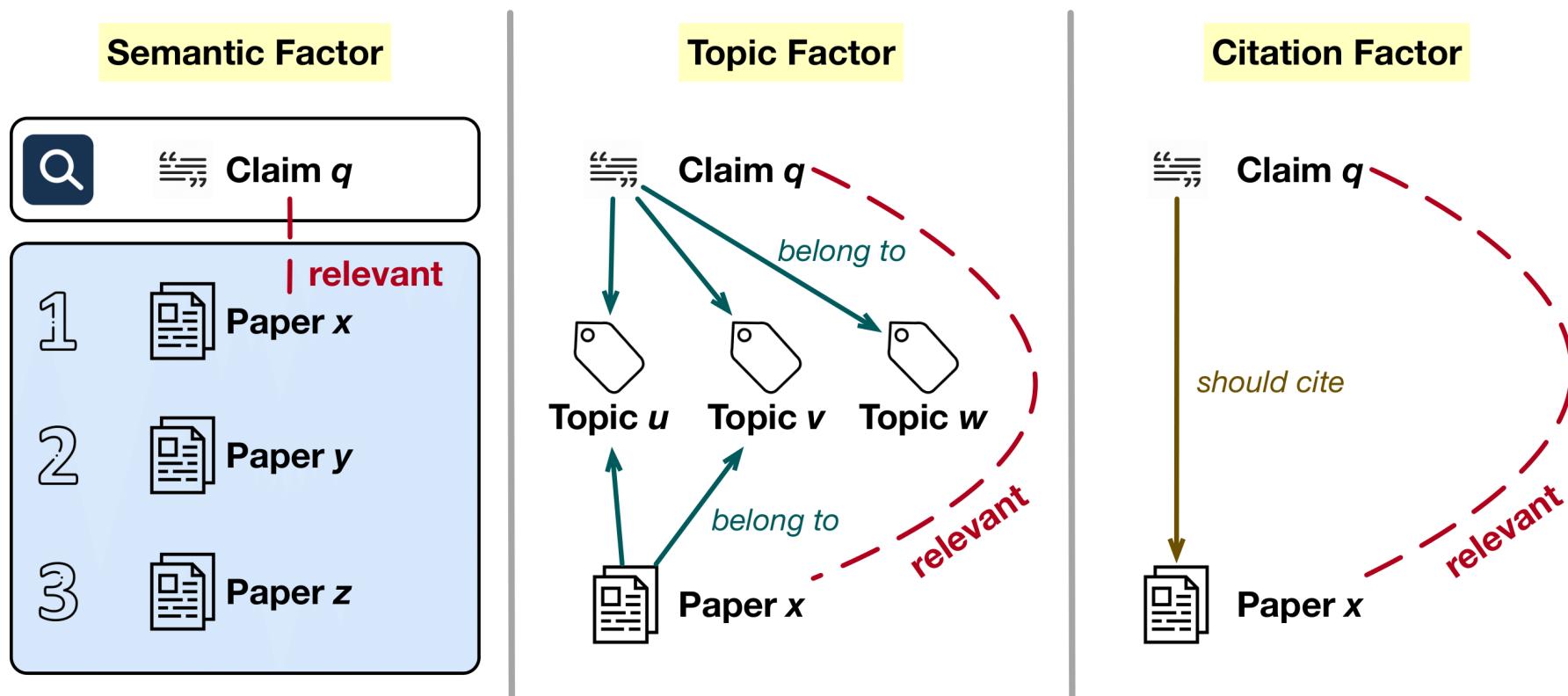
Given a biomedical claim, find relevant papers (and predict their stance).

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

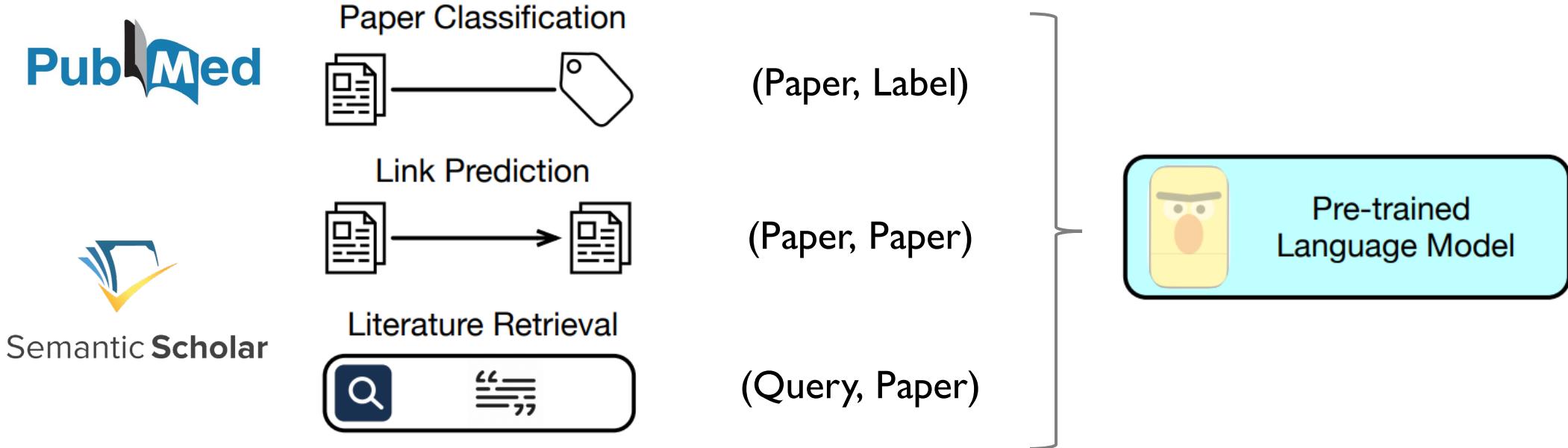
- Example: Biomedical Claim Verification
  - Why is a pair of (Claim, Paper) relevant?



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

# Naïve Multi-task Pre-training

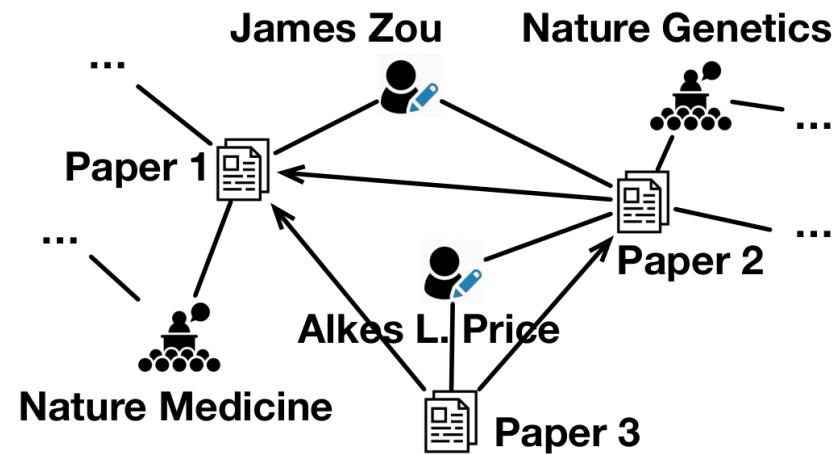
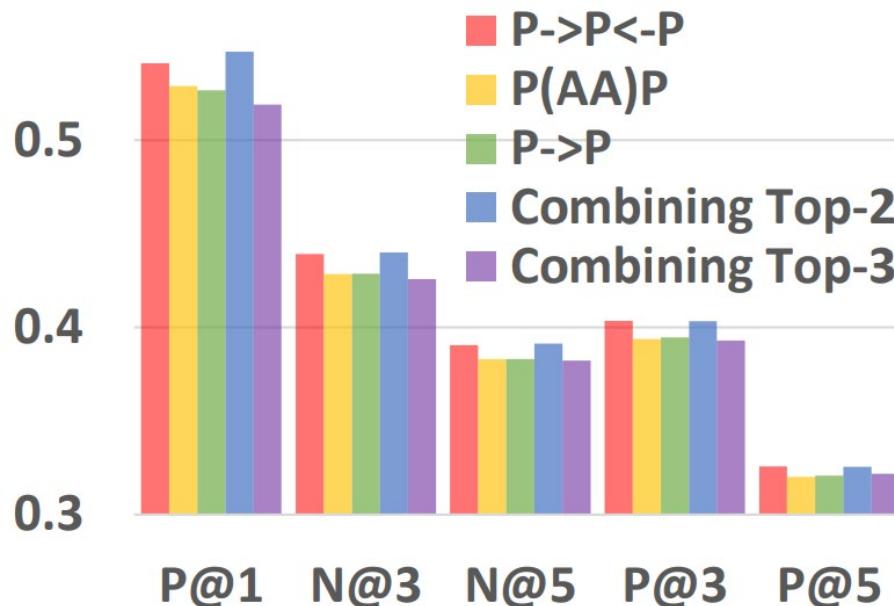
- Each factor (topic, citation, and semantic) relies on one **fundamental** text mining task.
- Directly combining pre-training data from different tasks to train a model?



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

# An Illustrative Example of Task Interference

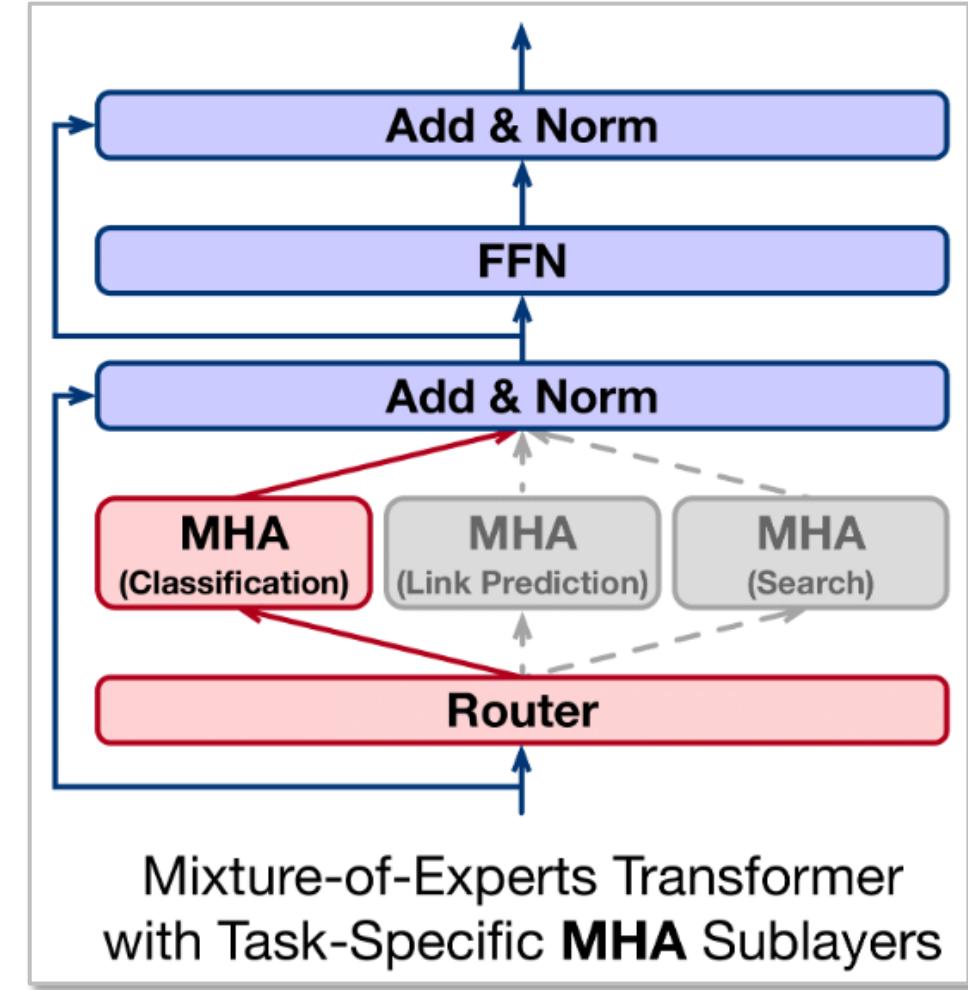
- Recall graph-induced contrastive learning
- Imagine each meta-path/meta-graph is a “task” (i.e., defines one type of “relevance”)
- Directly merging the relevant (paper, paper) pairs induced by different meta-paths for training?
  - Cannot consistently improve the classification performance!



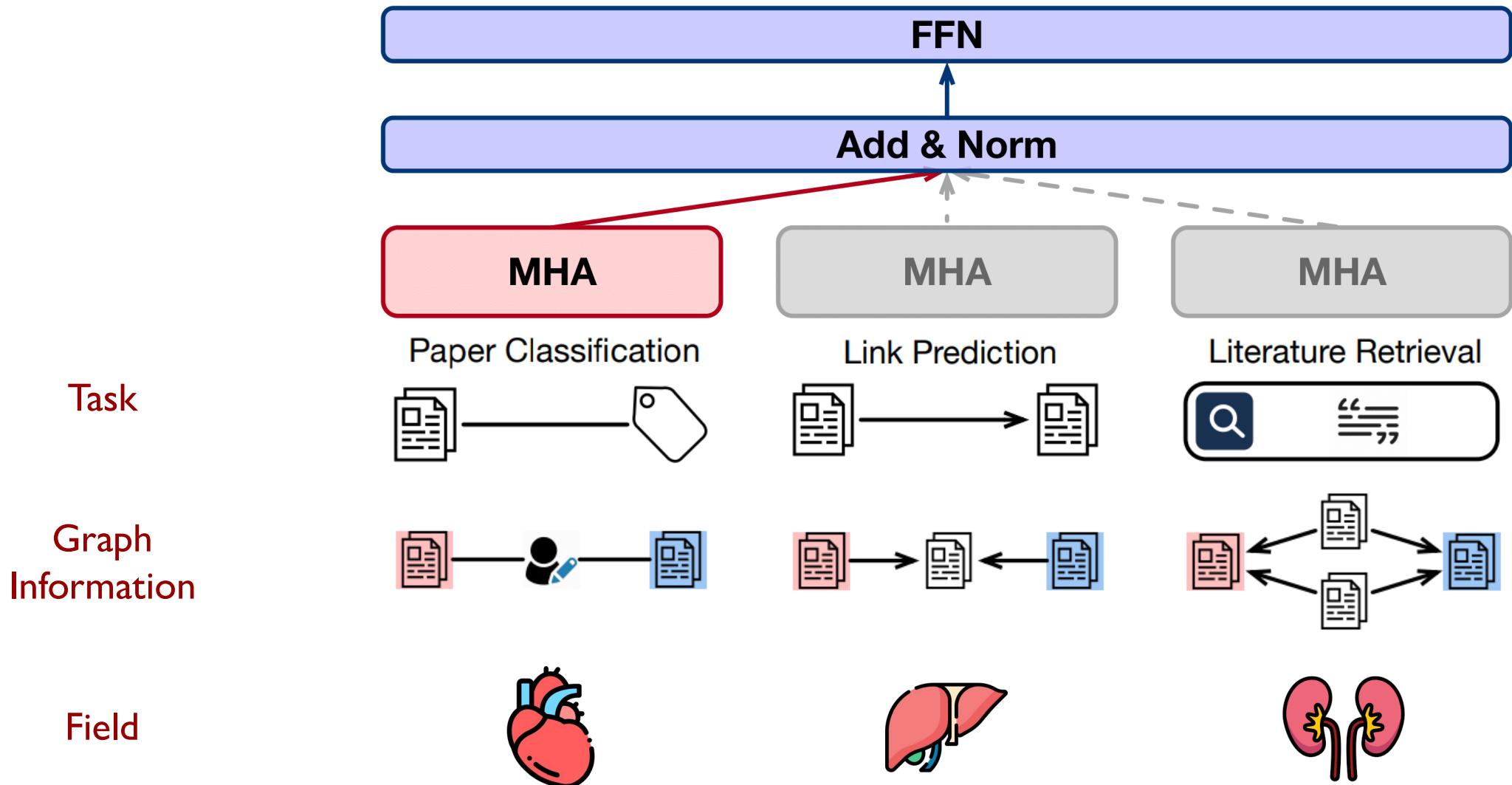
(Paper 2, Paper 3) are relevant according to  $P \rightarrow P \leftarrow P$  but irrelevant according to  $P(AA)P$ .

# 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

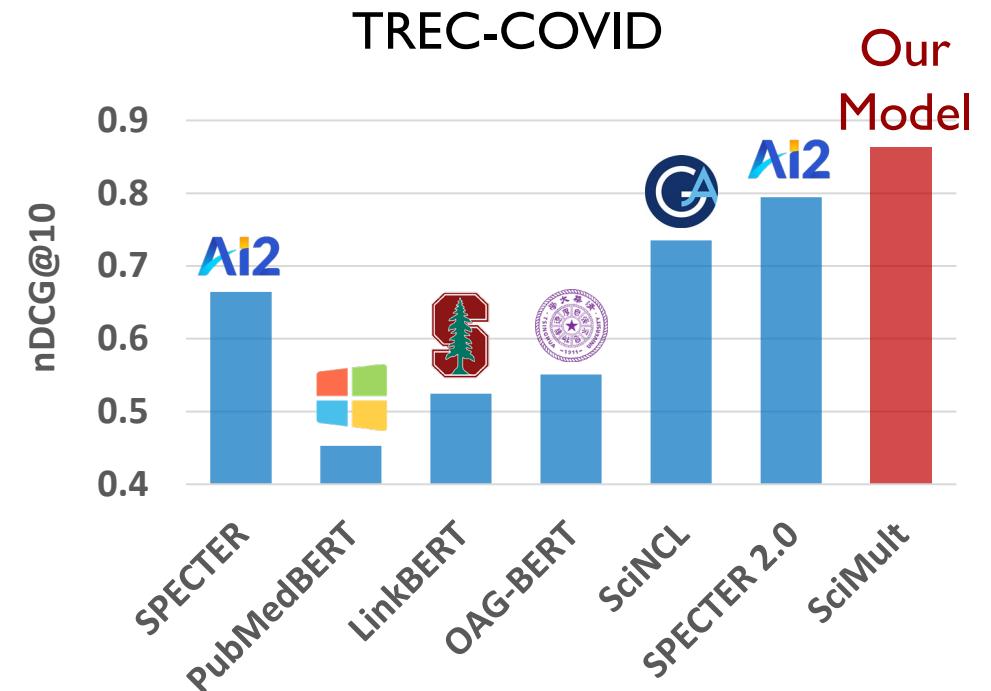


# Comparison with Previous Approaches

- 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)

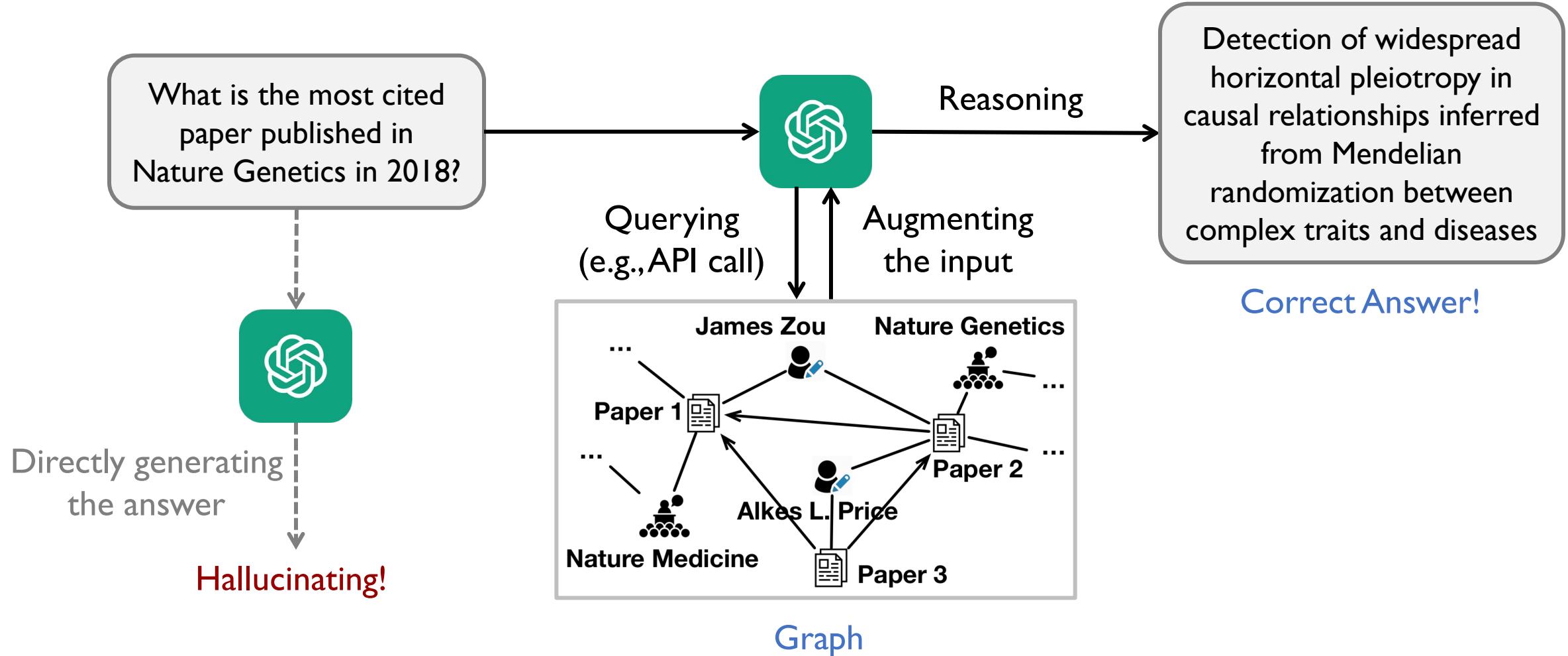
| Patient-to-Article Retrieval (PAR) Leaderboard        |   |         |          |             |          |
|---|---|---------|----------|-------------|----------|
|   | Model   | MRR (%) | P@10 (%) | nDCG@10 (%) | R@1k (%) |
| <b>Our Model</b><br>1<br><small>June 25, 2023</small> | DPR (SciMult-MHAExpert)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

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

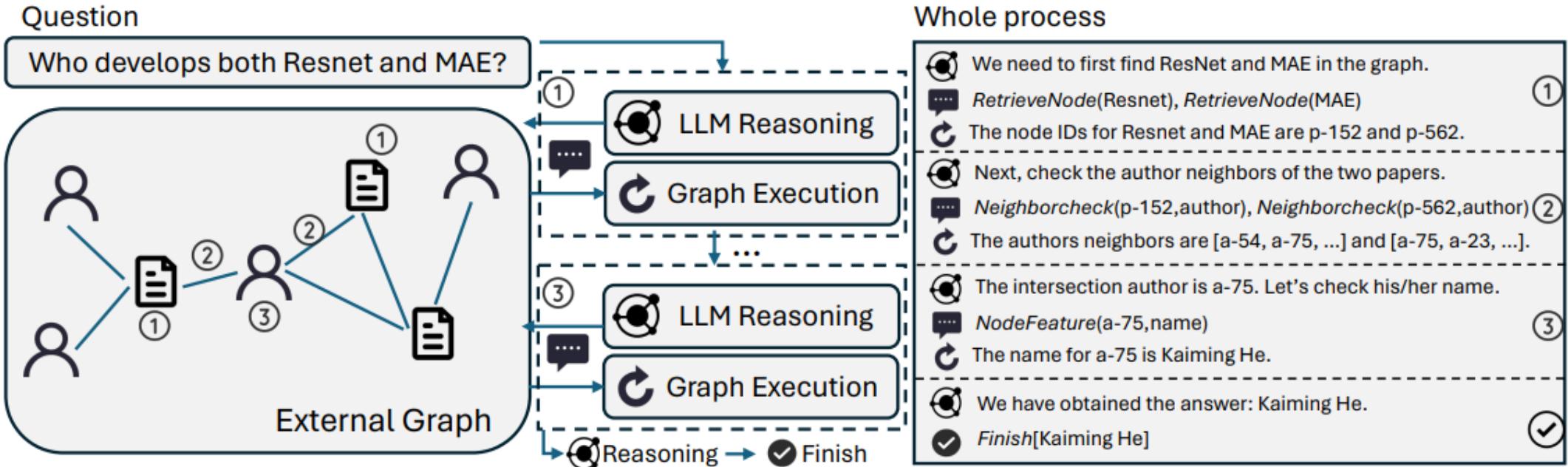


# Going Back to the Motivating Example

- Teaching LLMs to explore graphs as environments / use graphs as tools



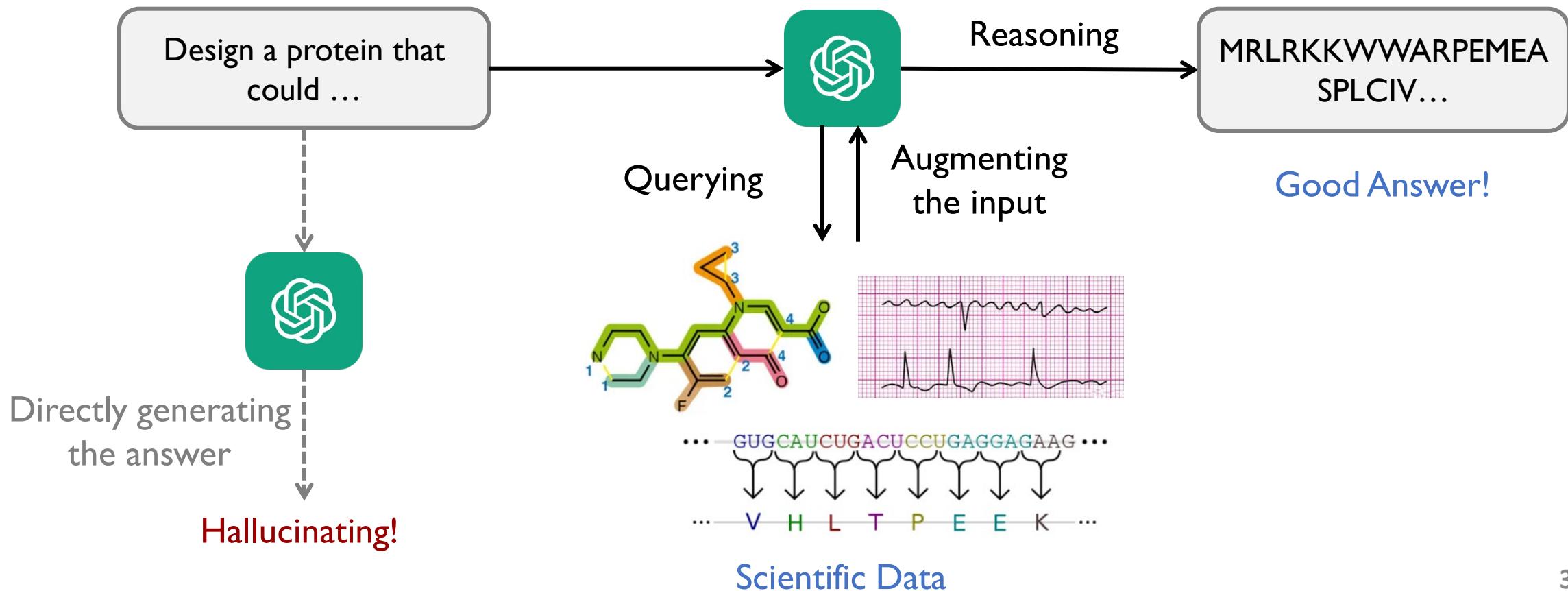
# Graph-Aware Usage of LLMs



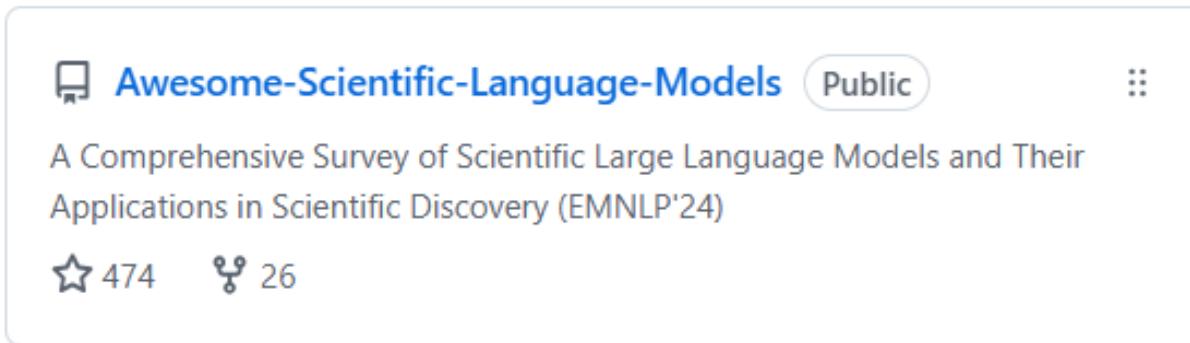
| Model            | Academic      |              | E-commerce   |              | Literature   |              | Healthcare   |              | Legal        |              |              |
|------------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  | 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         |              |
|                  | Mixtral-8x7b  | H            | 27.77        | 31.20        | 32.87        | 37.00        | 20.08        | 33.33        | 8.66         | 15.19        |              |
|                  | GPT-3.5-turbo | ∅            | 18.45        | 26.98        | 17.52        | 28.00        | 14.94        | 24.17        | 8.69         | 14.07        |              |
| <b>Our Model</b> |               | <b>31.89</b> | <b>33.48</b> | <b>42.40</b> | <b>44.50</b> | <b>41.59</b> | <b>46.25</b> | <b>22.33</b> | <b>28.89</b> | <b>30.52</b> | <b>28.33</b> |

# Building Trustworthy Biomedical Language Models

- Hallucination is particularly dangerous in high-stakes scientific domains such as chemistry and biomedicine.
  - Can we generalize RAG or CoT to **chemicals, proteins, time series, etc.**?



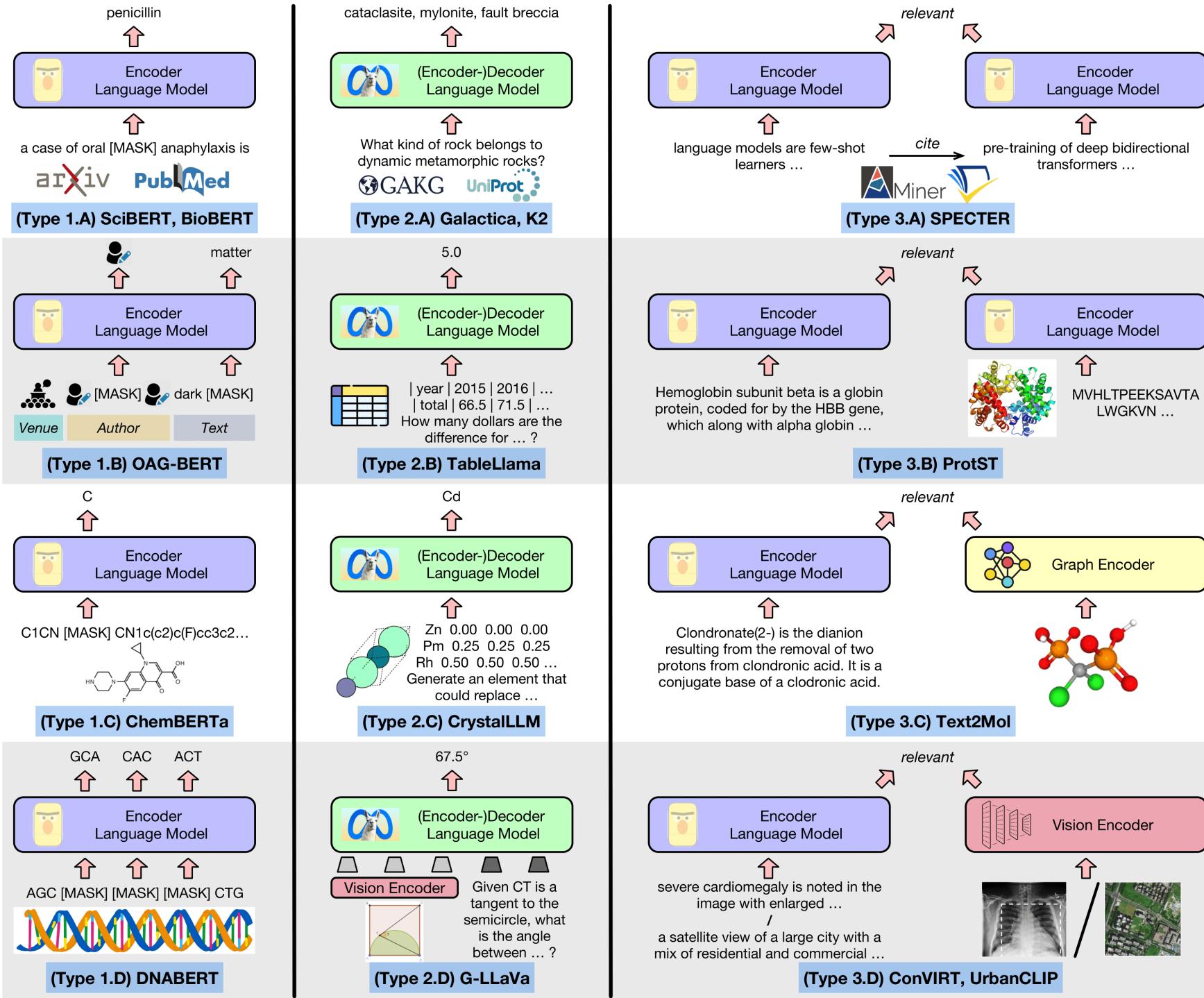
# A Survey of Scientific Large Language Models



The screenshot shows the 'README' tab of the GitHub repository. The main title is 'Awesome Scientific Language Models'. Below it, there are badges for 'awesome' and 'Stars 474'. Other buttons include 'PaperNumber 266', 'License MIT', 'PRs', and 'Welcome'. The description text is as follows:

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).

# Drawing Cross-Field Cross-Modal Connections between Scientific Large Language Models



# References

- 1) Zhang et al., “MATCH: Metadata-Aware Text Classification in A Large Hierarchy.” WWW 2021.
- 2) Zhang et al., “MotifClass: Weakly Supervised Text Classification with Higher-order Metadata Information.” WSDM 2022.
- 3) Zhang et al., “Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification.” WWW 2022.
- 4) Zhang et al., “Seed-Guided Topic Discovery with Out-of-Vocabulary Seeds.” NAACL 2022.
- 5) Zhang et al., “The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study.” WWW 2023.
- 6) Zhang et al., “Weakly Supervised Multi-Label Classification of Full-Text Scientific Papers.” KDD 2023.
- 7) Zhang et al., “Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding.” EMNLP 2023 Findings.
- 8) Jin et al., “Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs.” ACL 2024 Findings.
- 9) Zhang et al., “A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery.” EMNLP 2024.

Thank you! Questions?