CSCE 689 - Special Topics in NLP for Science

Lecture 23: LLMs for Research (Miscellaneous)

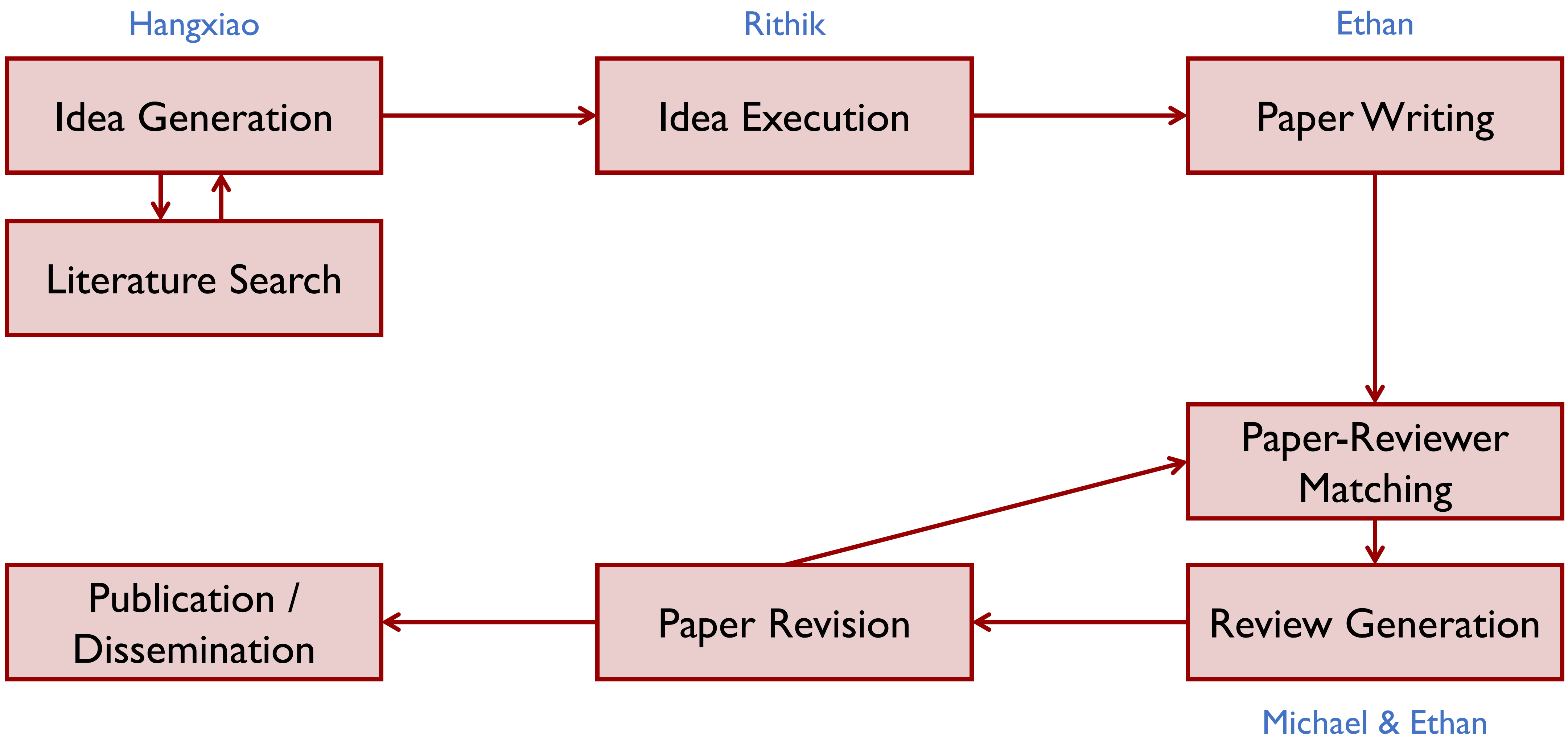
Yu Zhang

yuzhang@tamu.edu

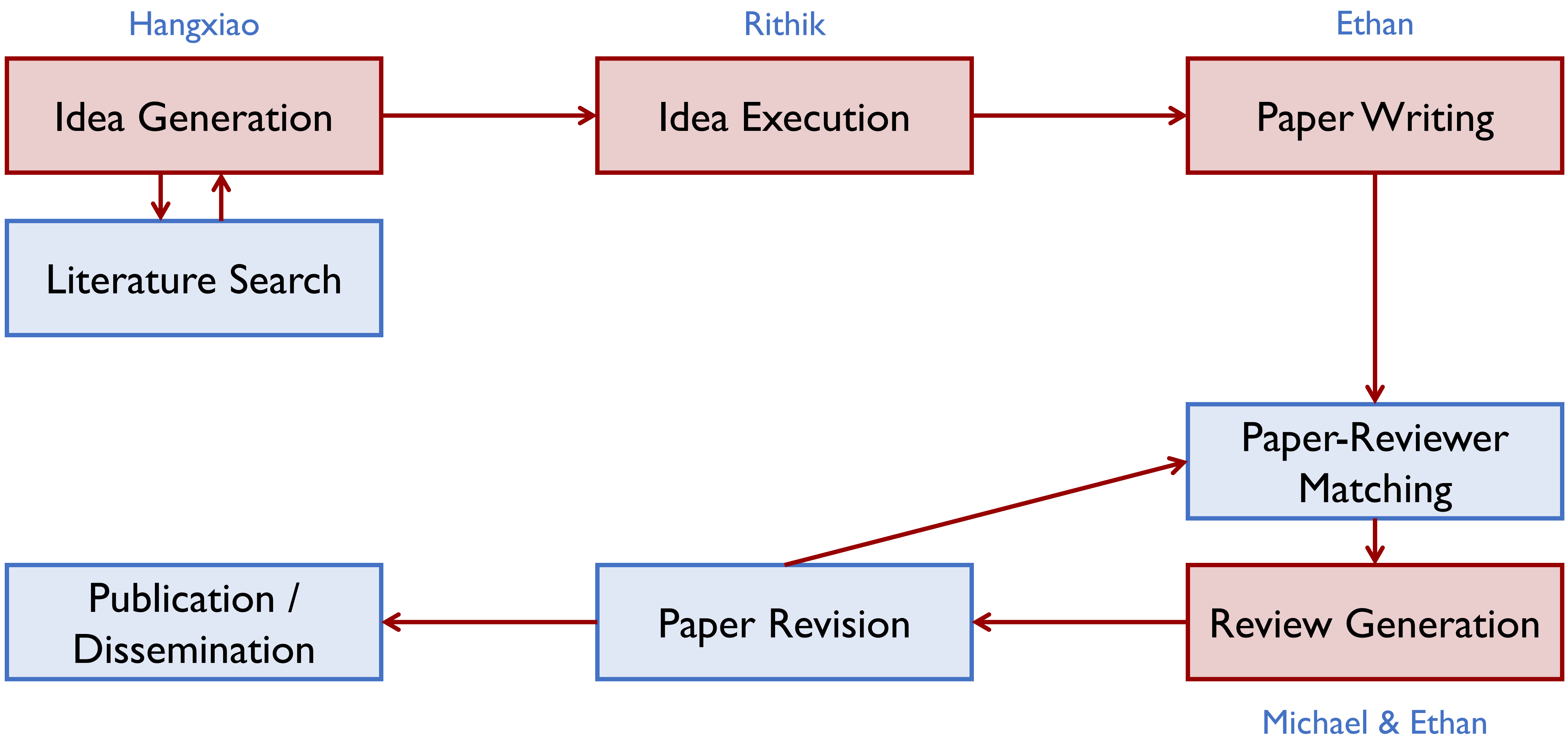
April 15, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

The Scientific Discovery Life Cycle



The Scientific Discovery Life Cycle



Agenda

- Literature Search: A Search Engine for Discovery of Scientific Challenges and Directions
- Paper-Reviewer Matching: Chain-of-Factors Paper-Reviewer Matching
- Paper Revision: ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
- Dissemination: Internal and External Impacts of Natural Language Processing Papers

Agenda

- Literature Search: A Search Engine for Discovery of Scientific Challenges and Directions
- Paper-Reviewer Matching: Chain-of-Factors Paper-Reviewer Matching
- Paper Revision: ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
- Dissemination: Internal and External Impacts of Natural Language Processing Papers

Motivation

- Scientists need to stay updated on challenges, limitations, and future directions.
- Existing tools (e.g., PubMed and Google Scholar) are not optimized for this type of discovery.

covid-19 | machine learning | add more...

Learning Invariant Representations across Domains and Tasks

Publication date: 2021-03-03

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

Investigating transferability in COVID-19 CT image segmentation

Publication date: 2021-02-23

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

Research direction: A sentence mentioning suggestions or needs for further research, hypotheses, speculations, indications or hints that an issue is worthy of exploration.

Challenge: A sentence mentioning a problem, difficulty, flaw, limitation, failure, lack of clarity, or knowledge gap.

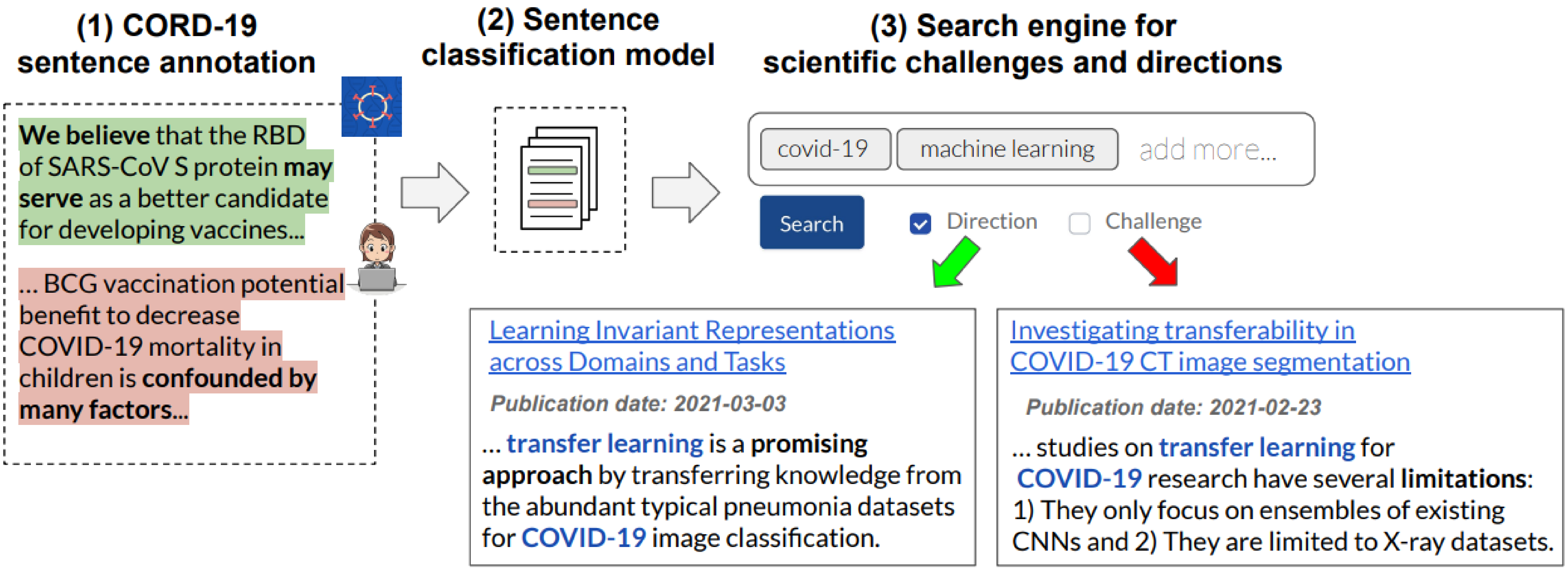
Why is detecting research directions and challenges hard?

- Example 1 (Misleading Keywords): "*The 15-30 mg/L albumin concentration is a critical value that could indicate kidney problems when it is repeatedly exceeded*"
  - Mention a diagnostic measure that is an indicator of a problem, rather than an actual problem

- Example 2 (Context and Domain Knowledge): "*BV-2 cells expressed Mac1 (CD11b) and Mac2 but were negative for the oligodendrocyte marker GalC …*"
  - Require more context and deep domain knowledge to understand whether this outcome is problematic or not

- We need annotation!

A Search Engine for Scientific Challenges and Directions

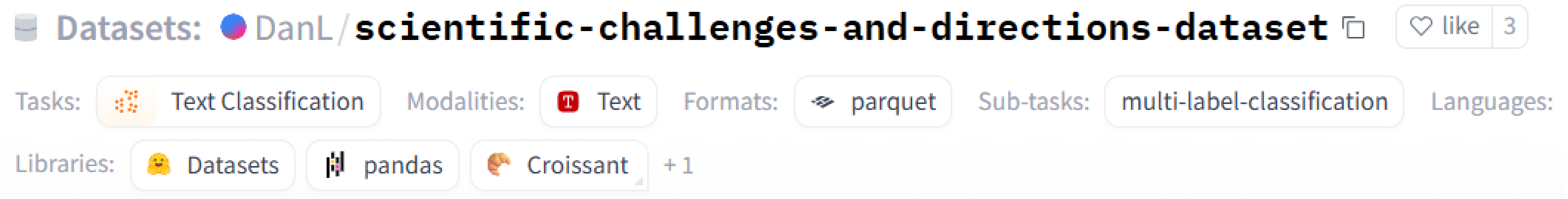


A Search Engine for Discovery of Scientific Challenges and Directions. AAAI 2022.

Annotation and Model Training

- Annotation: 2,894 sentences and their surrounding contexts (previous and next sentences) from 1,786 papers

https://huggingface.co/datasets/DanL/scientific-challenges-and-directions-dataset



- Model Training: Fine-tune a LM (e.g., PubMedBERT) on two binary classification tasks (i.e., challenge or not & direction or not)

Labels	Train	Dev	Test	All
Not Challenge, Not Direction	602	146	745	1493
Not Challenge, Direction	106	25	122	253
Challenge, Not Direction	288	73	382	743
Challenge, Direction	155	40	210	405

A Search Engine for Discovery of Scientific Challenges and Directions. AAAI 2022.

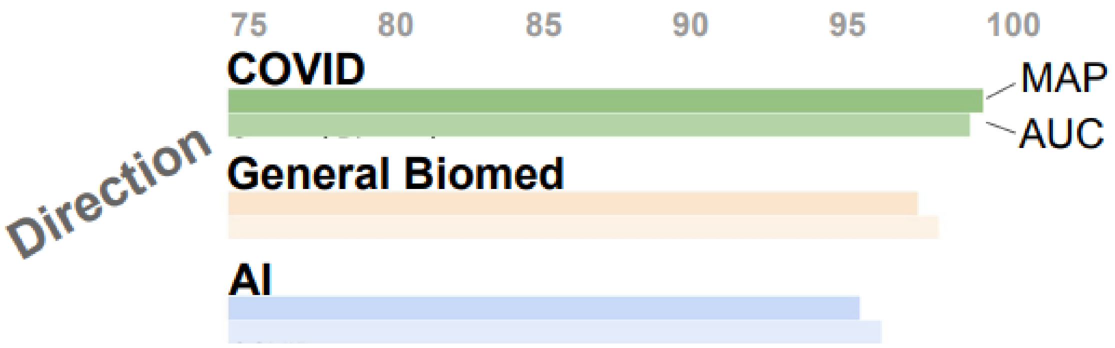
Context Slice + Combine

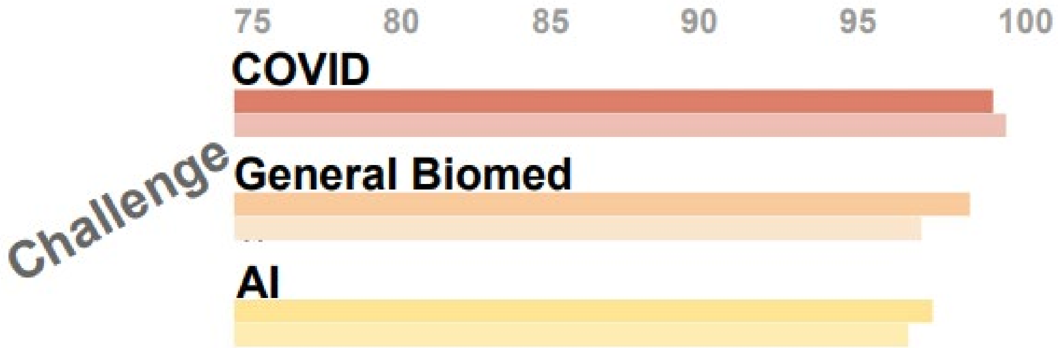
- We need context information to judge some cases.
  - [CLS] previous sentence [SEP] center sentence [SEP] next sentence [SEP]

- Train 2 models – One take the current sentence only; the other take the augmented sequence
- 2x2 predictions
  - Training on the center sentence; inference on the center sentence
  - Training on the center sentence; inference on the augmented sequence
  - Training on the augmented sequence; inference on the center sentence
  - Training on the augmented sequence; inference on the augmented sequence
- Average the output probability vector of these predictions

Performance of Challenge and Direction Sentence Classification

	Challenge			Direction		
Model	P	R	F1	P	R	F1
Keyword	0.535	0.760	0.628	0.455	0.792	0.578
Sentiment	0.405	0.966	0.571	0.239	0.837	0.371
NLI-Zeroshot	0.659	0.693	0.675	0.401	0.825	0.540
RoBERTa-large	0.723 (0.042)	0.824 (0.046)	0.769 (0.004)	0.697 (0.065)	0.825 (0.06)	0.754 (0.004)
SciBERT	0.729 (0.023)	0.799 (0.03)	0.761 (0.007)	0.719 (0.044)	0.783 (0.043)	0.749 (0.01)
PubMedBERT	0.738 (0.018)	0.804 (0.017)	**0.770** (0.006)	0.755 (0.017)	0.778 (0.015)	**0.766** (0.006)
+context	0.716 (0.048)	0.809 (0.047)	0.758 (0.007)	0.701 (0.038)	0.771 (0.026)	0.733 (0.01)
PubMedBERT-HAN	0.671 (0.02)	0.863 (0.03)	0.759 (0.01)	0.674 (0.04)	0.804 (0.04)	0.734 (0.001)
Slice-Combine	0.742 (0.011)	0.829 (0.012)	**0.783** (0.004)	0.732 (0.02)	0.82 (0.03)	**0.773** (0.005)

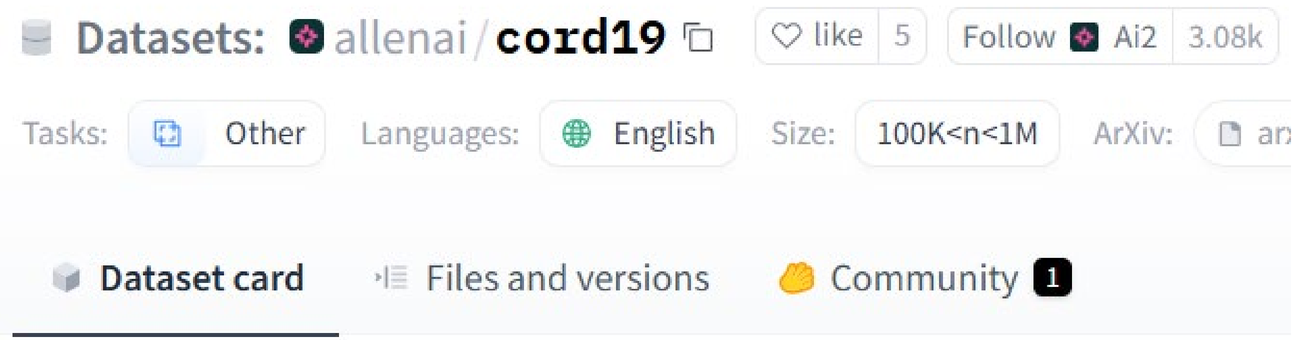




Building a Search Engine

- Step 1: Classify sentences in the CORD-19 dataset (papers related to COVID-19)

https://huggingface.co/datasets/allenai/cord19

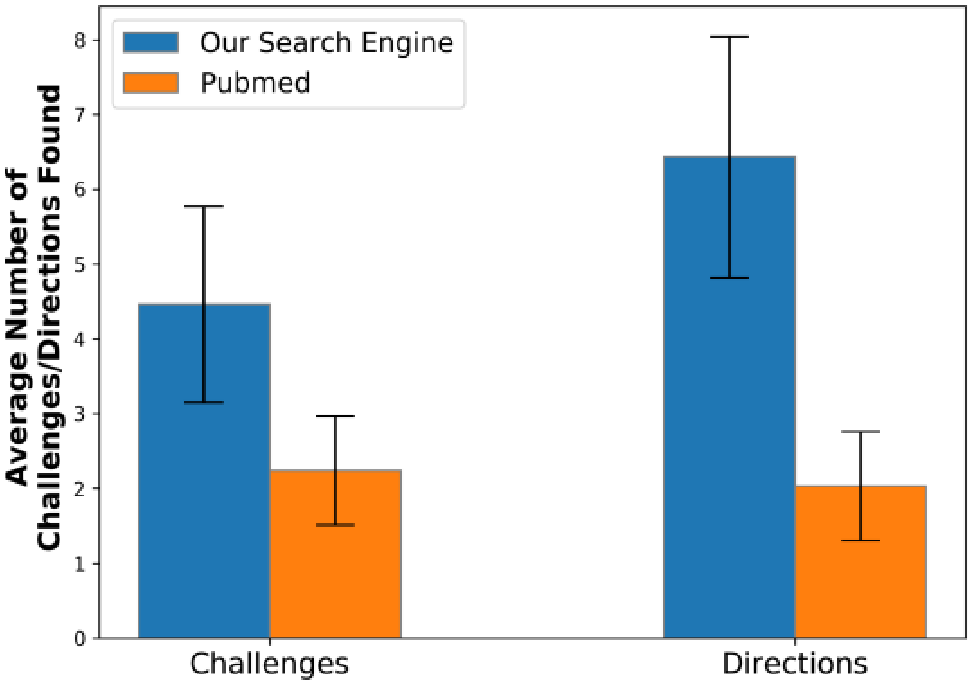


- Step 2: Extract entities from sentences predicted as challenges or research directions and link them to knowledge base entries
- Step 3: Index these sentences using linked entities

- Support entity-based faceted search (e.g., “*AI + pneumonia*”)

User Studies

- 10 participants
- Given 20 queries, find as many challenges and directions as possible in 3 minutes with the help of a search engine.



- 9 medical researchers at a large hospital
- Find problems/limitations related to COVID-19 and each of (1) hospital infections, (2) diagnosis, (3) vaccines for children, (4) probiotics and the gastrointestinal tract.
- Find directions/hypotheses related to COVID-19 and each of (1) mechanical ventilators, (2) liver, (3) artificial intelligence, (4) drug repositioning.

Metric	Chal./Dir. Search	PubMed
Search	**90%**	48%
Utility	**94%**	57%
Interface	**91%**	68%
Overall	**92%**	59%

Take-Away Messages

- Scientific research engines may focus on sentences with specific functions (e.g., directions, challenges, claims, …) in the paper rather than the overall semantics. Finding/indexing such sentences may help paper search.
  - Can GPT-4 perform this sentence classification task with a few/zero examples?
- Instead of classifying the "center" sentence only, we can classify the context-augmented sequence and jointly consider multiple predictions.

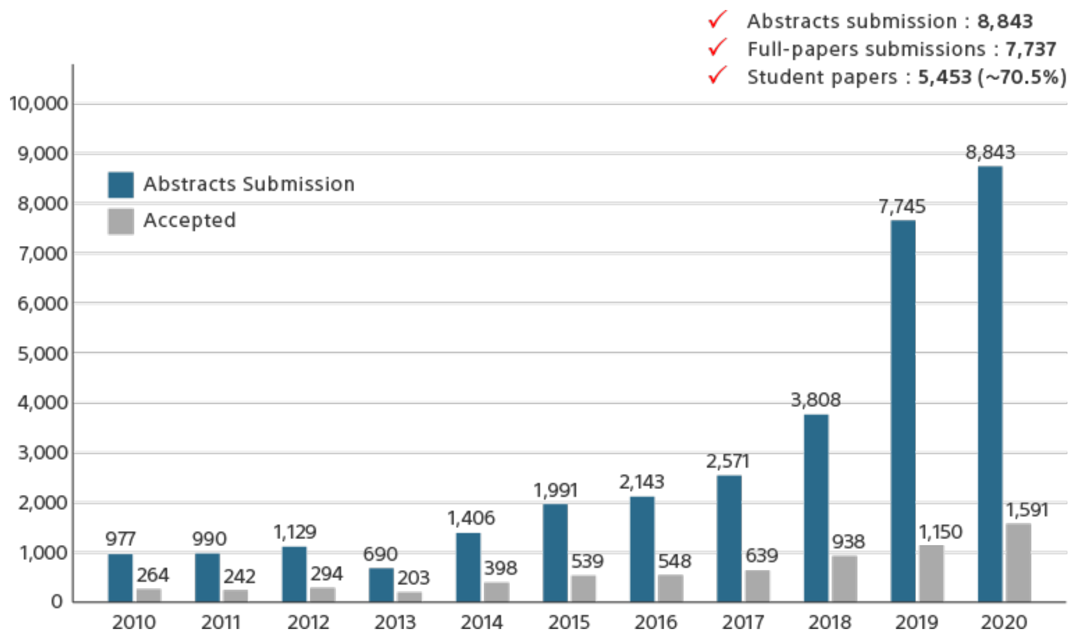
- Limitations:
  - Only support entity-based faceted search (i.e., a set of entities as the query)
  - Cannot summarize the directions and challenges from multiple papers/sentences in a generative way

Agenda

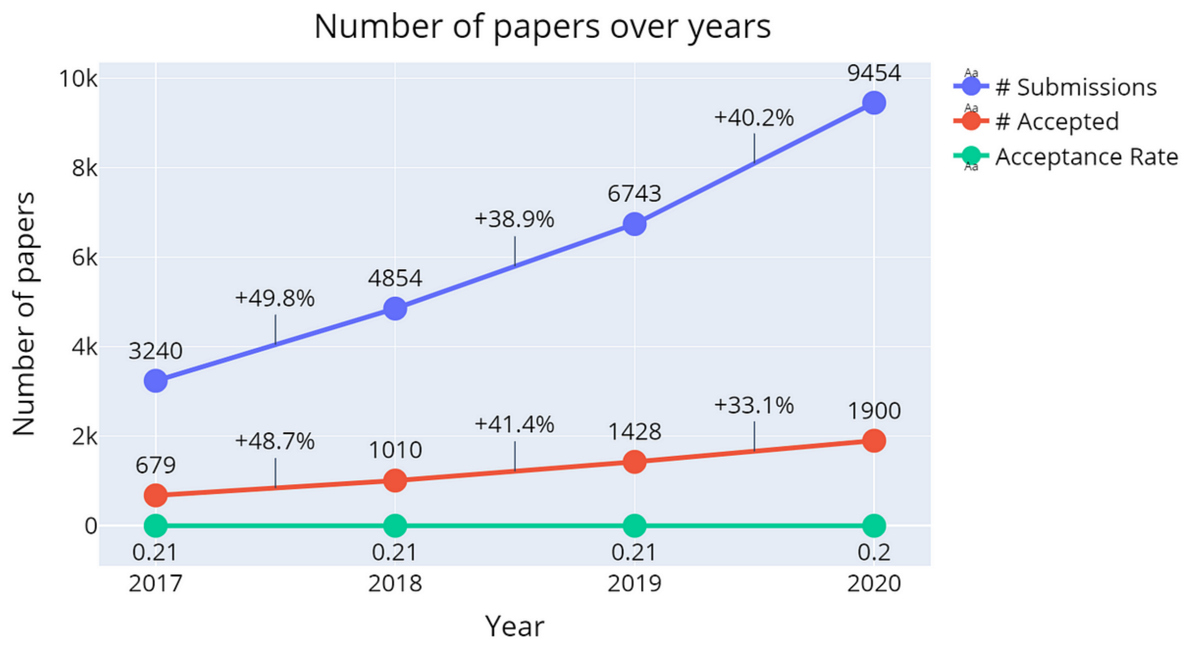
- Literature Search: A Search Engine for Discovery of Scientific Challenges and Directions
- Paper-Reviewer Matching: Chain-of-Factors Paper-Reviewer Matching
- Paper Revision: ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
- Dissemination: Internal and External Impacts of Natural Language Processing Papers

Explosion of Submissions to AI Conferences

- Given a huge volume of (e.g., 10,000) submissions, it becomes prohibitively time-consuming for chairs to manually assign papers to appropriate reviewers.



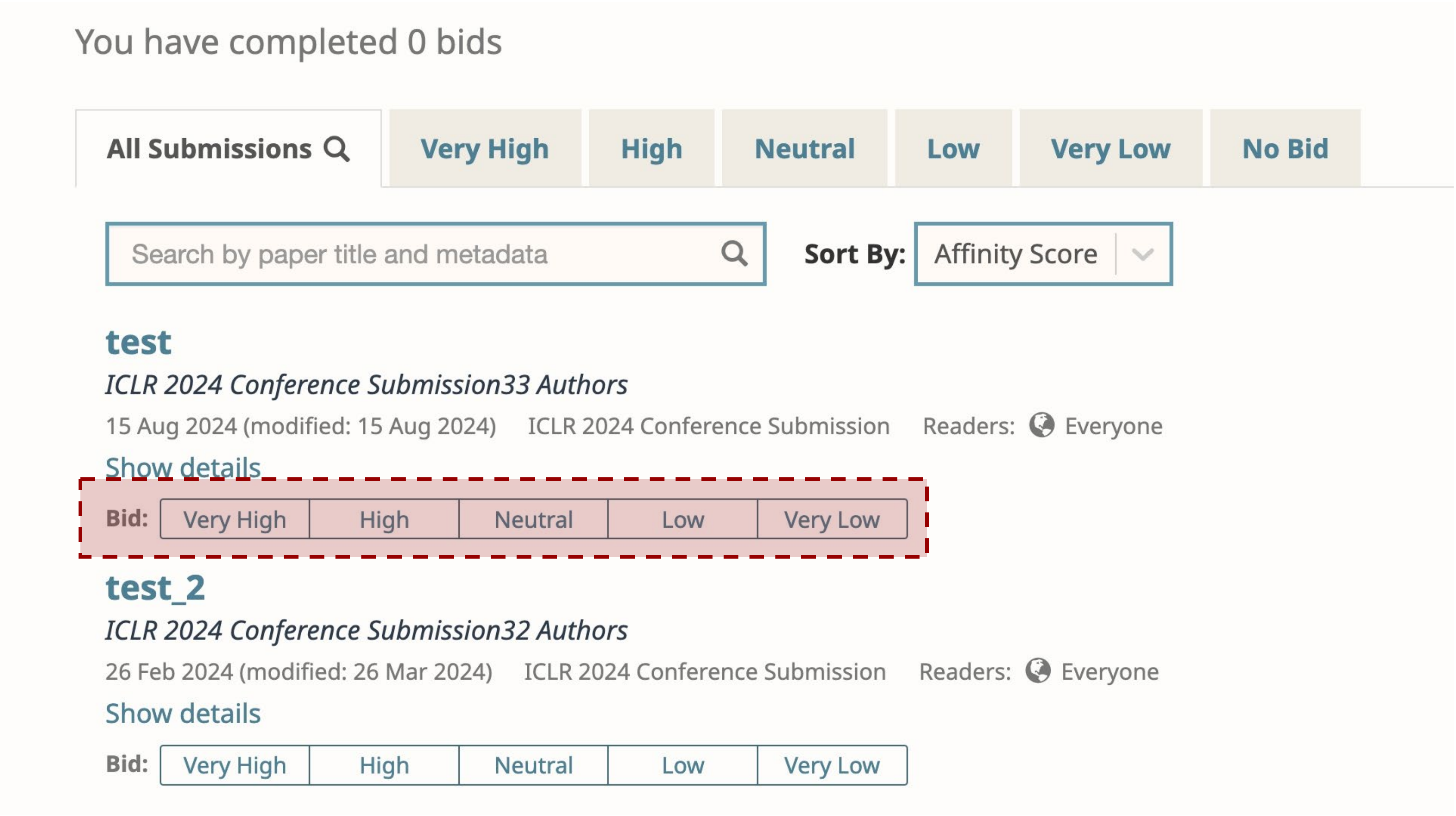
# of submissions to AAAI by year



# of submissions to NeurIPS by year

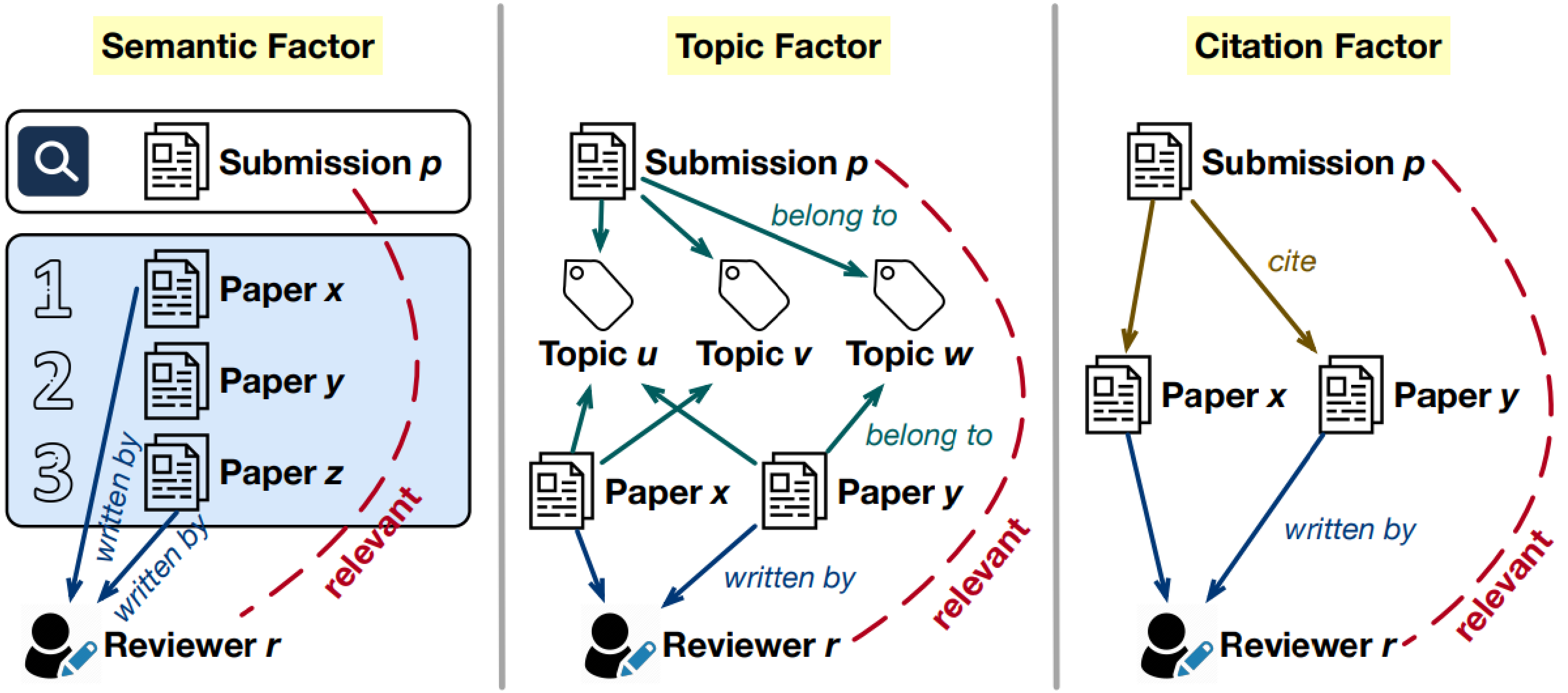
Ask Reviewers to Bid Papers?

- They can hardly scan all submissions.
- An accurate pre-ranking result should be delivered to them so that they just need to check a shortlist of papers.
- A precise scoring system that can automatically judge the expertise relevance between each paper and each reviewer is needed.



Multiple Factors for Judging Relevance

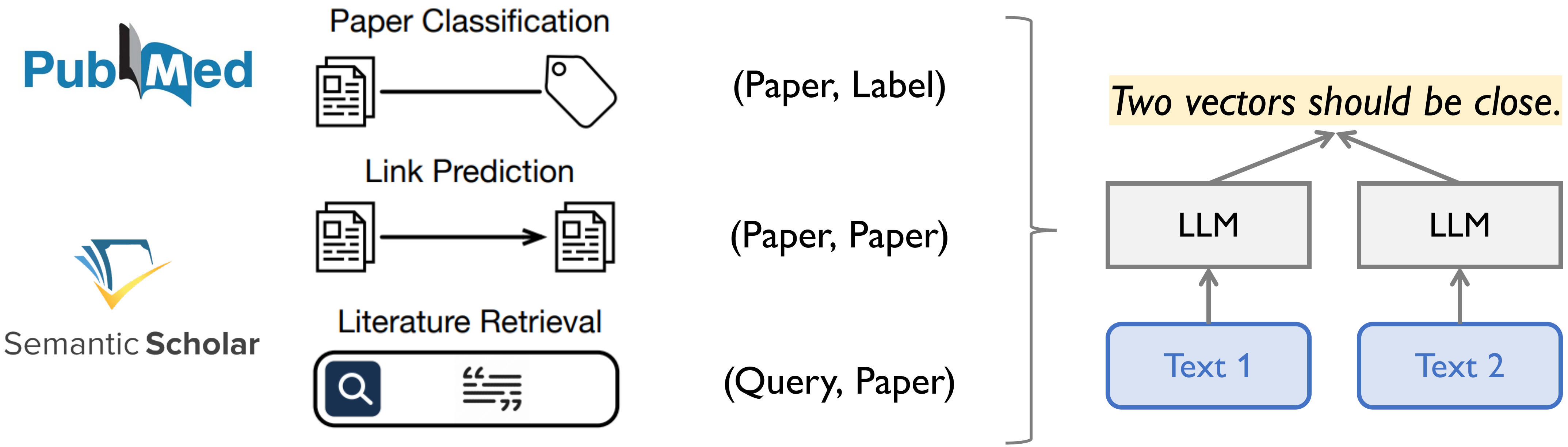
- Why is a pair of (Paper, Reviewer) relevant?



- How to make LLMs aware of these factors?

Contrastive Learning for Multiple Factors – A Naïve Way

- Directly combining pre-training data from different factors to train a model?

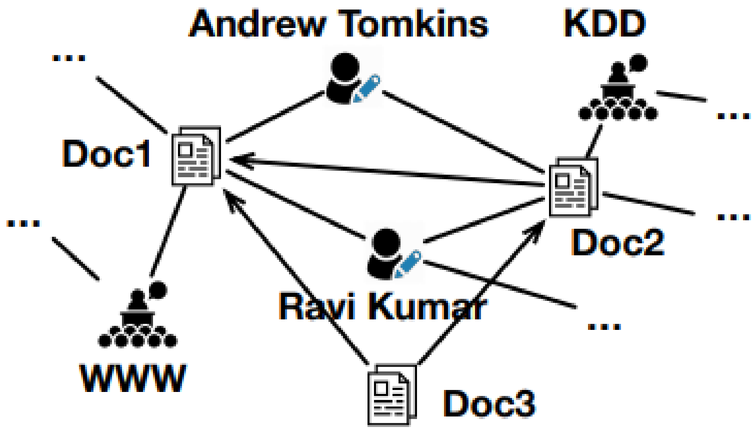


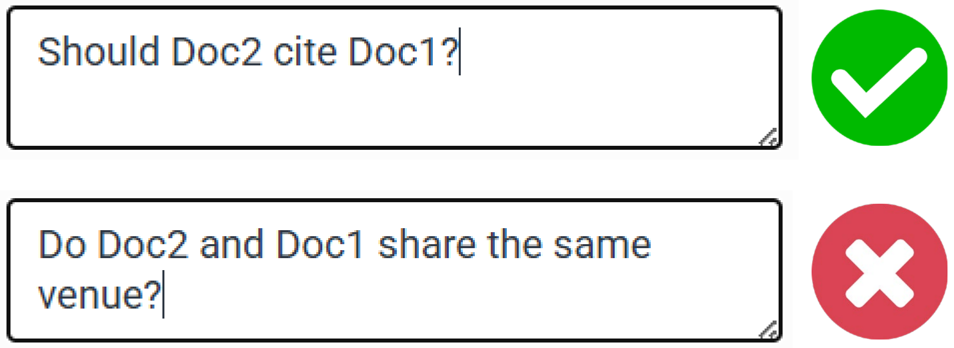
- Task Interference: The model is confused by different types of "relevance".

A Toy Example of Task Interference

- Imagine you have two “tasks”.
  - Task 1: Given Paper1 and Paper2, predict if Paper1 should cite Paper2.
  - Task 2: Given Paper1 and Paper2, predict if Paper1 and Paper2 share the same venue.

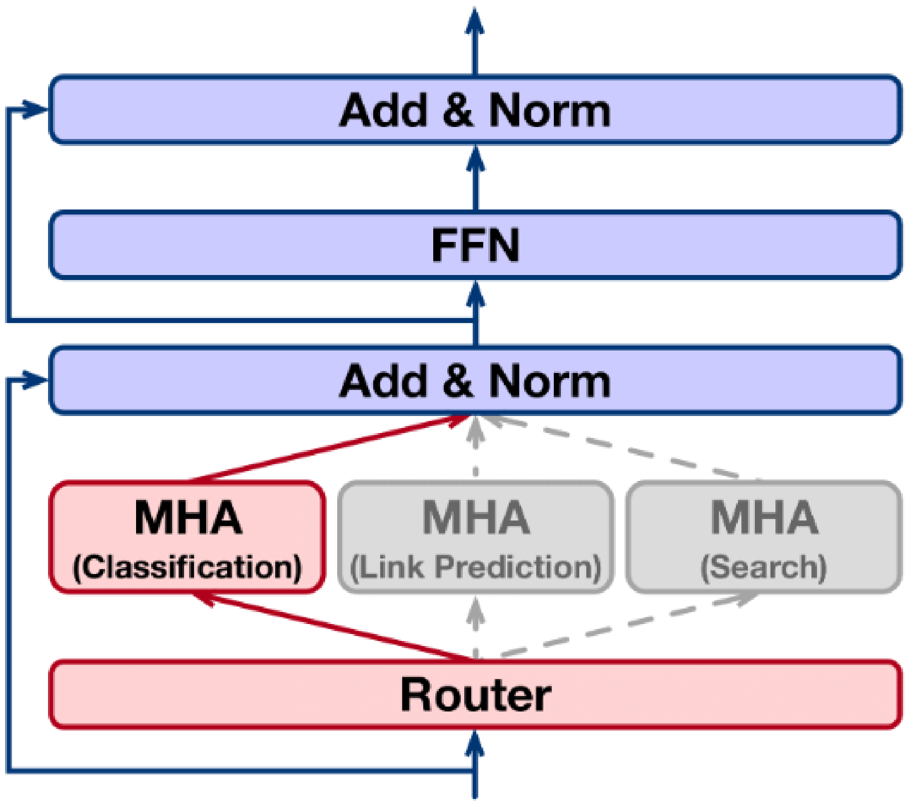
- What if we directly merge the collected relevant (paper, paper) pairs for these two tasks?
  - Is (Doc2, Doc1) relevant?
  - The model does not know which task you are referring to, so it will get confused!





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.

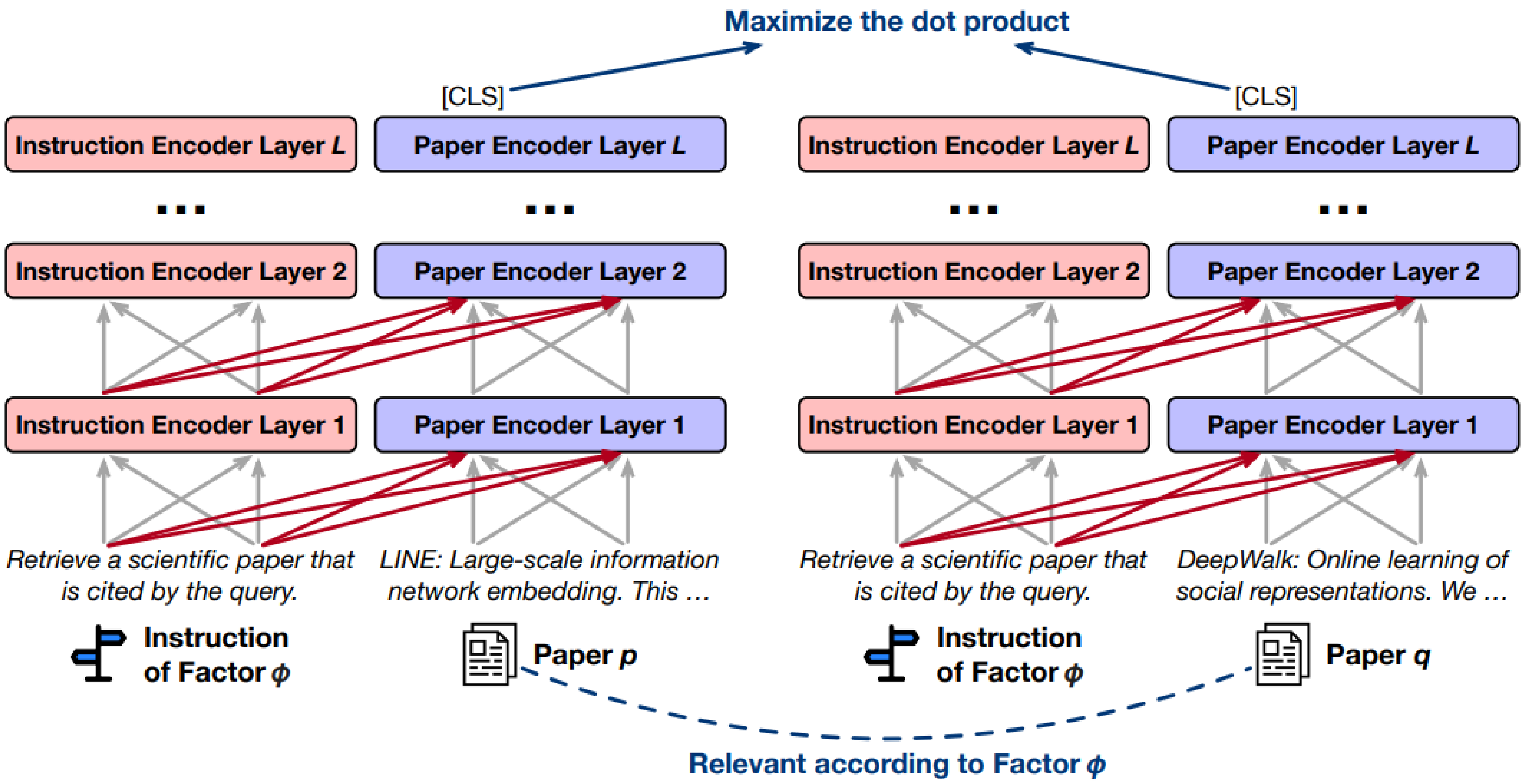


Mixture-of-Experts Transformer with Task-Specific **MHA** Sublayers

Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. EMNLP 2023 Findings.

Tackling Task Interference: Instruction Tuning

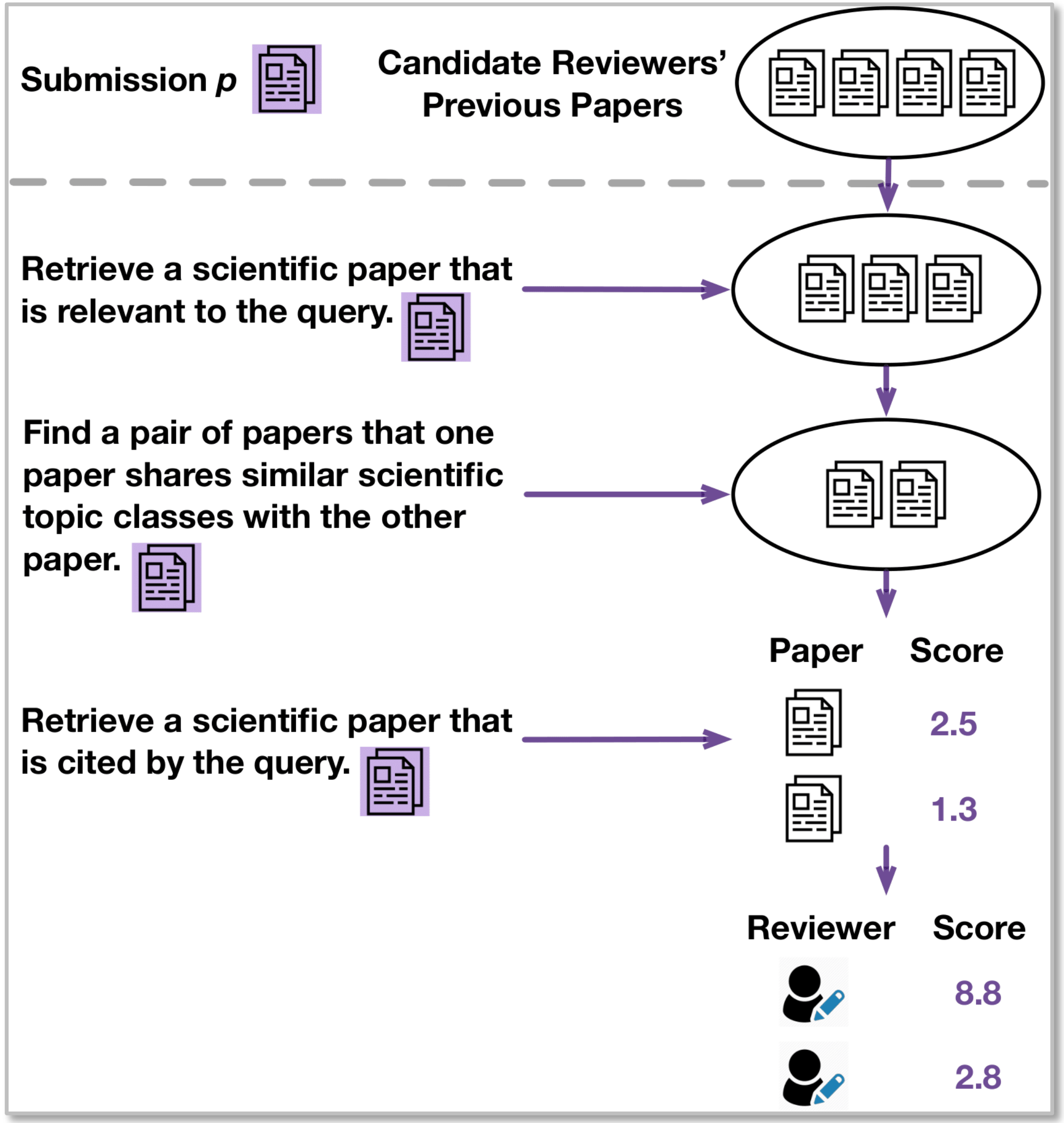
- Using a factor-specific instruction to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.



Chain-of-Factors Paper-Reviewer Matching. WWW 2025.

Chain-of-Factors Reasoning

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



Chain-of-Factors Paper-Reviewer Matching. WWW 2025.

Performance of Chain-of-Factors (CoF)

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

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

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

Performance of Chain-of-Factors (CoF)

- CoF outperforms traditional paper-reviewer matching methods

	NIPS [32]			
	Soft P@5	**Hard P@5**	**P@5 defined in [28]**	**P@5 defined in [1]**
APT200 [32]	41.18**	20.59**	–	–
TPMS [7]	49.41**	22.94**	50.59**	55.15**
RWR [28]	–	24.1**	45.3**	–
Common Topic Model [1]	–	–	–	56.6**
SciBERT [6]	47.06**	21.18**	49.61**	52.79**
SPECTER [9]	52.94**	25.29**	53.33**	58.68**
SciNCL [35]	54.12**	27.06**	54.71**	59.85**
COCO-DR [58]	54.12**	25.29**	54.51**	59.85**
SPECTER 2.0 CLF [46]	52.35**	24.71**	53.33**	58.09**
SPECTER 2.0 PRX [46]	53.53**	27.65	54.71**	59.26**
CoF	**55.68**	**28.24**	**56.41**	**61.42**

: semantic-based method

: topic-based method

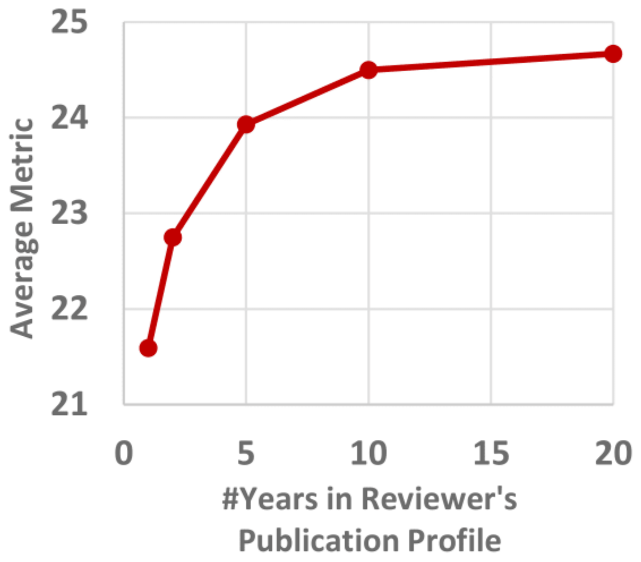
: citation-based method

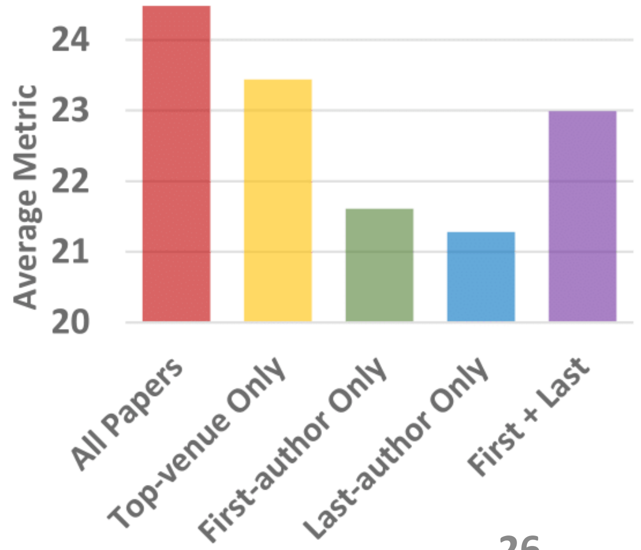
- CoF outperforms ablation versions that consider one factor only (or consider three factors simultaneously)

	NIPS	**SIGIR**	**KDD**
CoF ($\mathbb{S} \rightarrow \mathbb{T} \rightarrow \mathbb{S}+\mathbb{T}+\mathbb{C}$)	50.44	**31.87**	**24.48**
No-Instruction	49.52**	27.67**	24.07**
$\mathbb{S}$	50.29	28.07**	24.05**
$\mathbb{T}$	49.98	28.69**	24.11*
$\mathbb{C}$	50.31	28.81**	24.20*
$\mathbb{S}+\mathbb{T}+\mathbb{C}$	**50.55**	28.63**	24.26*
$\mathbb{S} \rightarrow \mathbb{T} \rightarrow \mathbb{C}$	50.11	31.79	24.36

Impact of Reviewer's Profile on the Matching Performance

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





Take-Away Messages

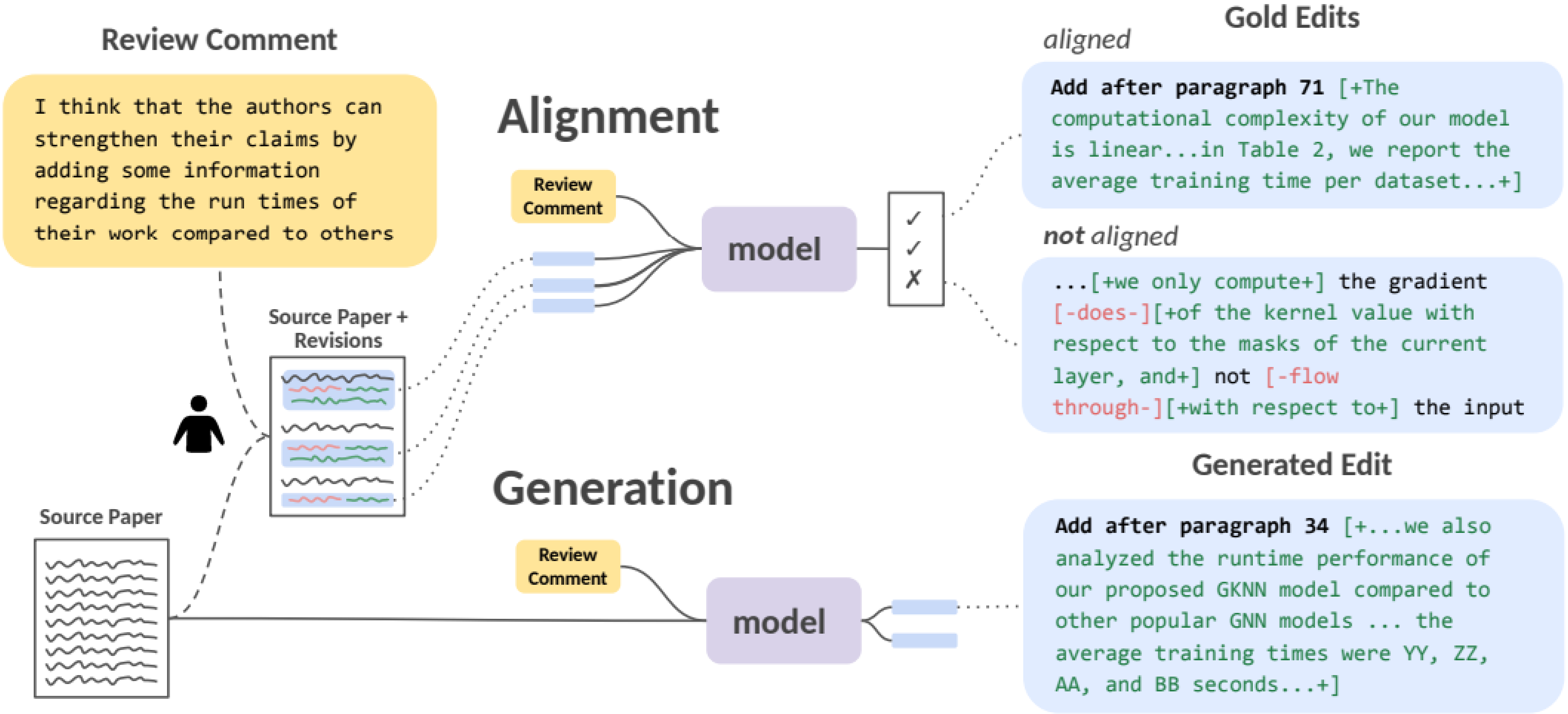
- We need to consider multiple factors (i.e., semantic, topic, and citation) for paper-reviewer matching.
- Directly combining training data from different factors for contrastive learning suffers from task interference. Instruction tuning helps the model understand the task it is performing and facilitates chain reasoning.

- Limitations:
  - Not deployed to a conference in the real world (e.g., an A/B test to compare Chain-of-Factors with TPMS or SPECTER)
  - How to perform this A/B test?
    - *Reviewer bias in single- versus double-blind peer review*. PNAS 2017.

Agenda

- Literature Search: A Search Engine for Discovery of Scientific Challenges and Directions
- Paper-Reviewer Matching: Chain-of-Factors Paper-Reviewer Matching
- Paper Revision: ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
- Dissemination: Internal and External Impacts of Natural Language Processing Papers

Two Tasks: Comment-Edit Alignment and Edit generation



ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews. ACL 2024.

Dataset Construction

- Step 1: Collect papers, reviews, and author responses from computer science conferences on OpenReview
  - Original Version: the latest PDF that was uploaded before the first review
  - Revised Version: the latest available PDF
  - Extract edits on a paragraph level
- Step 2: Identify actionable feedback & align comments to edits
  - Manually annotated by 2 annotators
  - Flexibly ways to express actionable feedback
    - Direct request: "*Apply the method to a realistic dataset*"
    - Criticism: "*The evaluation is only on a synthetic dataset*"
    - Question: "*Is the current dataset truly representative of the real-world?*"

ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews. ACL 2024.

Dataset Construction

Statistic	Manual	Synthetic
Papers	42	1678
Comments	196	3892
Aligned Edits	131	3184

- Step 3: Create synthetic data
  - Manual annotation is too costly and time-consuming!
  - Automatically identify the quoted review comments in author responses by searching for lines with a small edit distance to a contiguous span of review text
  - The corresponding response text for each comment is matched to edits with high textual overlap.

An example: consider the following author response

W2 & Q 1-3 Missing detail of the paper

Thank you for pointing out the missing detail. We added the missing detail one by one in the revision, and also place them below:

> How are the scores of dataset examples calculated in Figure 1(a)?

Match this with the review

This score is from the RULER benchmark. We added this explanation in both figure caption (Line 241) and the experiment setup (Line 206)

Match this with the revised version

ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews. ACL 2024.

Performance on Comment-Edit Alignment

Model	Micro AO-F1	Micro P	Micro R	Micro F1	Macro AO-F1	Macro P	Macro R	Macro F1
BM25	13.3 [4.7, 30.0]	12.2	10.5	11.3 [6.7, 17.1]	48.0 [36.5, 59.2]	41.7	26.4	20.9 [13.0, 29.1]
BM25-generated	14.7 [5.1, 23.6]	4.6	40.3	8.3 [5.2, 10.8]	31.0 [21.5, 41.1]	5.2	57.5	8.6 [6.6, 10.7]
Specter2 (no finetuning)	14.0 [7.6, 22.4]	8.1	14.4	10.3 [6.6, 14.7]	42.3 [31.4, 53.2]	22.2	29.0	13.0 [7.0, 18.9]
Specter2 bi-encoder	19.6 [12.8, 27.0]	17.0	29.3	21.5 [16.0, 27.4]	40.2 [31.4, 49.6]	34.6	38.1	22.6 [17.1, 28.3]
DeBERTa bi-encoder	3.1 [0.0, 8.6]	9.9	12.2	10.8 [6.0, 18.3]	42.8 [31.5, 54.2]	52.4	22.0	18.6 [11.0, 26.1]
LinkBERT cross-encoder	2.8 [0.5, 8.4]	10.1	28.4	14.4 [9.2, 20.6]	41.0 [30.3, 51.7]	14.7	40.8	12.9 [9.8, 16.2]
DeBERTa cross-encoder	8.5 [5.2, 12.5]	7.4	25.6	10.0 [6.8, 13.5]	42.6 [33.3, 52.3]	13.2	40.4	10.7 [8.1, 13.4]
GPT-4 cross-encoder 0-shot	38.7 [27.8, 51.9]	-	-	-	50.8 [40.9, 60.9]	-	-	-
GPT-4 cross-encoder 1-shot	42.1 [31.5, 54.2]	-	-	-	57.0 [47.2, 66.5]	-	-	-
GPT-4 multi-edit	36.2 [22.0, 53.4]	24.2	30.4	27.0 [18.2, 39.4]	50.6 [40.5, 60.8]	31.6	28.2	26.3 [19.4, 33.2]
Random	5.5 [3.9, 7.9]	1.5	1.7	1.6 [0.8, 2.7]	20.2 [13.7, 26.9]	10.2	17.6	4.9 [1.8, 8.1]
Human	70.6 [52.0, 83.4]	65.6	76.8	70.7 [54.9, 81.0]	75.4 [66.8, 84.0]	84.0	69.2	67.0 [58.4, 76.0]

GPT-4 significantly outperforms other baselines but still performs poorly.

Performance on Edit Generation

	Ans.	Non-ans.	All
GPT	31%	19%	25%
Real	19%	40%	29%
Same	50%	42%	46%
Frequency	51%	49%	100%

	GPT	Real	κ	p
Compliance	2.9	2.6	0.6	10^{-4}
Promises	21%	6%	1.0	10^{-2}
Paraphrases	48%	4%	0.7	10^{-11}
Technical details	38%	53%	0.7	0.06

Factor	Comment	Edit
Compliance=1	... Isn't this percentage too much? Can't we use, e.g., 5% of all nodes for training?	[+... our split of 80% -10% -10% is a standard split+]
Compliance=2	... there is a hyperprameter in the radius decay, how it will affect the performance is crucial ...	[+... this learnable radius is not effective the in terms of an classification performance compared to that the predefined radius decay+]
Compliance=3	the experimental setup requires significantly more details on the hardware ...	[+We conducted our experiments using NVIDIA Tesla V100 GPUs ...+]*
Promises	it would be interesting to know how the proposed method would work, for instance, for node classification (e.g., Cora, Citeseer)	[+... the performance of our method on node classification tasks is beyond the scope of this paper and is left as an interesting direction for future work.+]*
Paraphrases	... it should be investigated ... with respect to more natural perturbations, e.g. noisy input, blurring, ...	[+... we also investigate their performance with respect to more natural perturbations, such as noisy input, blurring, ...+]*
Technical details	... This does put into question whether the full closed loop model is actually useful in practice	[+... we evaluated the performance of a closed-loop N-CODE model ... Here, the control parameters are a matrix of dynamic weights, $\theta(t) \in \mathbb{R}^{m\times m}$ ...+]

Take-Away Messages

- GPT-4 performs poorly on the comment-edit alignment task despite being able to generate plausible edits in the generation task.
- The kinds of edits produced by GPT-4 can be very different from the real edits authors make to their papers.
  - GPT-4 tends to paraphrase, provide a standalone response (i.e., not tightly integrated into the context of the paper), and lack specific technical details.
- Limitations:
  - Only aim to understand the differences in style and content between human edits and GPT-generated edits. Not evaluating the correctness or appropriateness of generated edits.
  - Not proposing any advanced techniques to boost the performance of comment-edit alignment

Agenda

- Literature Search: A Search Engine for Discovery of Scientific Challenges and Directions
- Paper-Reviewer Matching: Chain-of-Factors Paper-Reviewer Matching
- Paper Revision: ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
- Dissemination: Internal and External Impacts of Natural Language Processing Papers

What papers should we expect at an NLP conference?

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

ACL Is Not an AI Conference

Emily M. Bender
Bangkok, Thailand
August 14, 2024

ACL 2024 Presidential Address

https://bit.ly/EMB-ACL24

What papers should we expect at an NLP conference?

ACL is not an AI Conference (?)

Yoav Goldberg, August 2024

In her "Presidential Address" at the ACL 2024, Emily Bender gave a talk called "ACL is not an AI Conference". For those who did not attend (or were not paying close attention), you can find the slides in the following link: https://faculty.washington.edu/ebender/papers/ACL_2024_Presidential_Address.pdf

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

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

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

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

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

How does the public perceive NLP conferences?

ACL 2024
Bangkok, Thailand

EMNLP 2024

NAACL 2024

Cite

NEURAL INFORMATION PROCESSING SYSTEMS

nature

Science

"Internal"

Mention

UNITED STATES PATENT AND TRADEMARK OFFICE
uspto

CNN

X

"External"

Patent

Media (News/Social)

Policy Document

Data and Metric

NLP Papers:
ACL Anthology
ACL, EMNLP, NAACL
1979-2024

Internal Citation:
OpenAlex

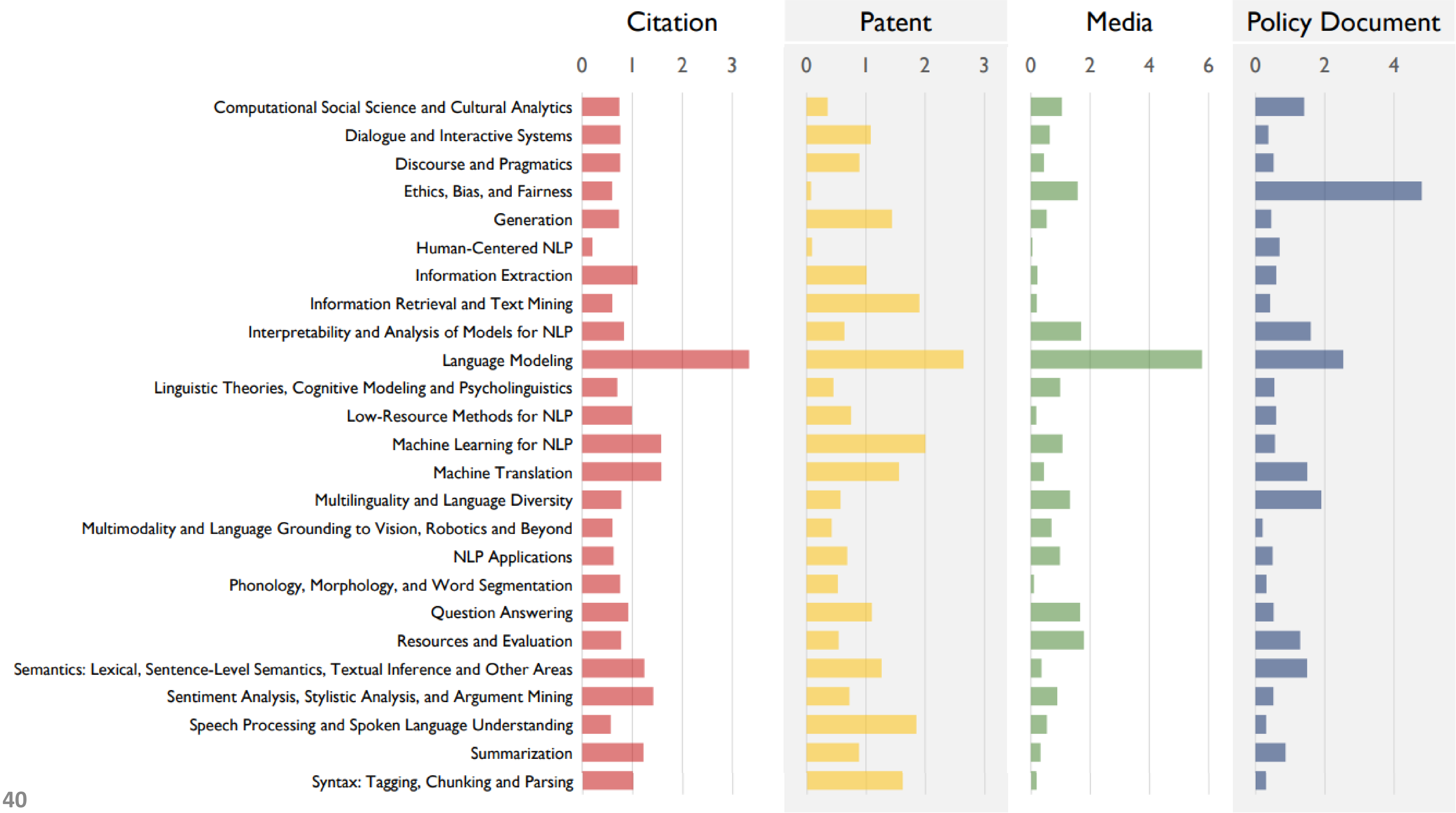
Patent-to-Paper:
Reliance on Science

Media-to-Paper:
Altmetric

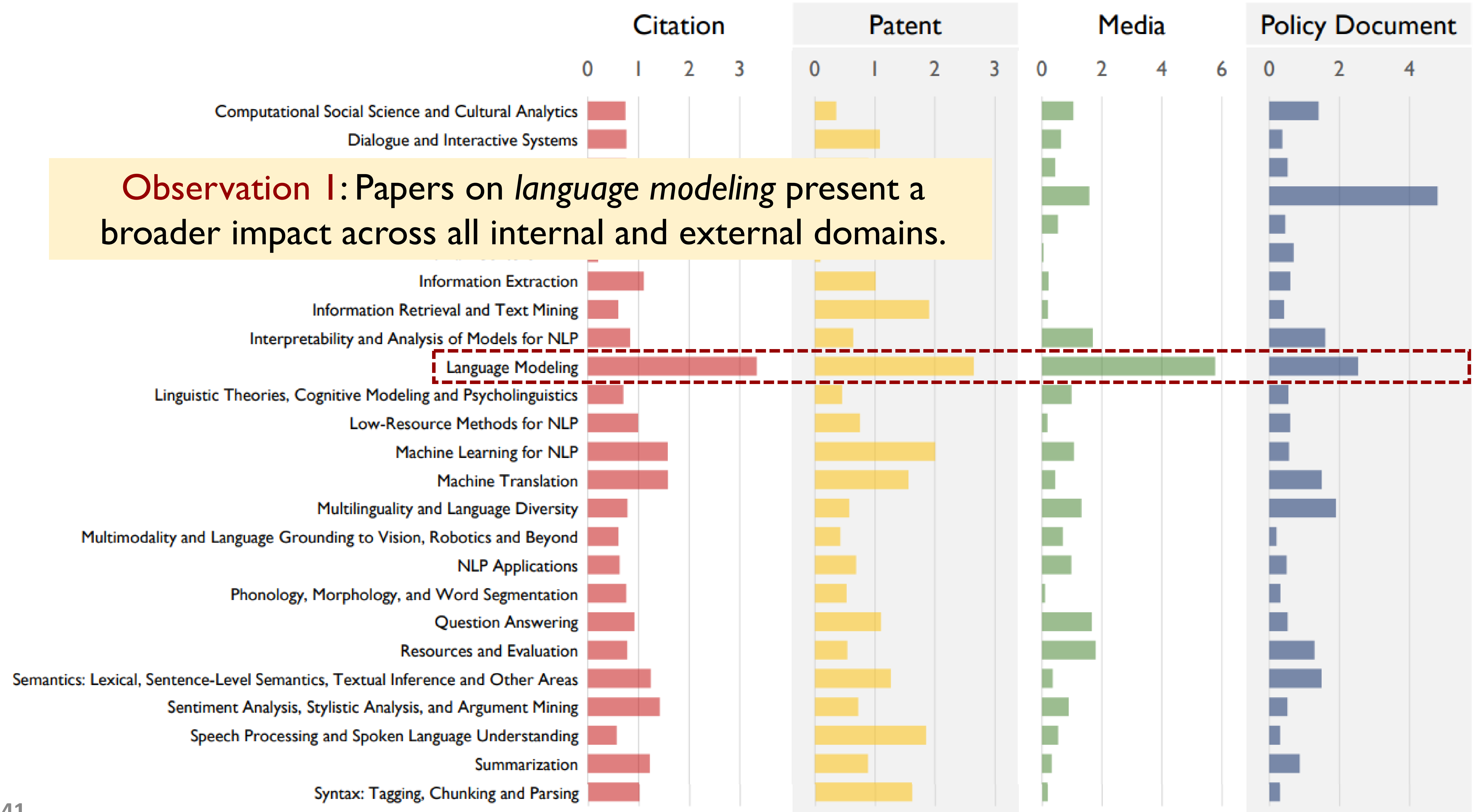
PolicyDoc-to-Paper:
Overton

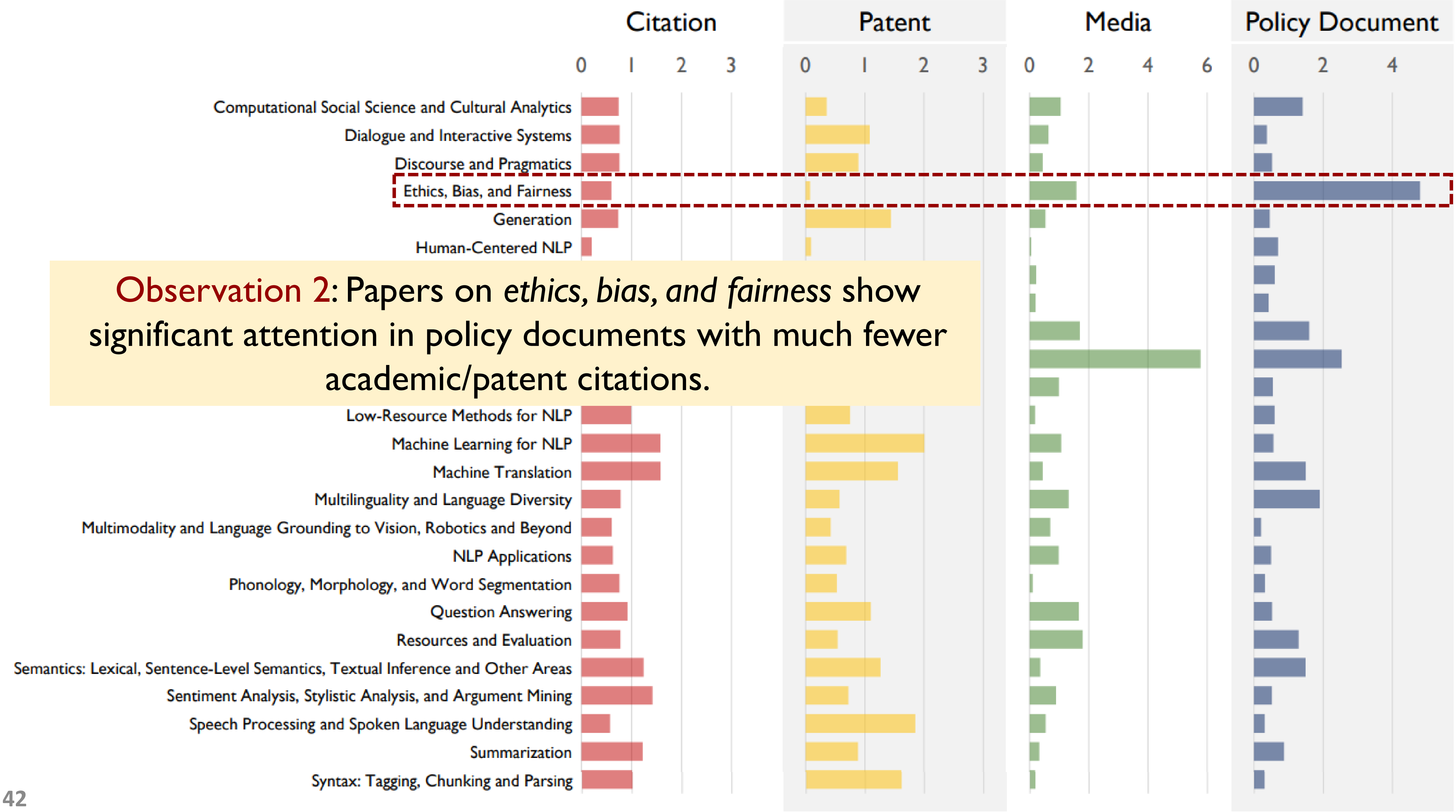
- How to quantify the impact of an NLP topic (e.g., "*Language Modeling*" and "*Ethics, Bias, and Fairness*" within a domain (e.g., "*Citation*", "*Patent*", "*Media*", and "*PolicyDocument*")?
  - Assume there are 1,000 NLP papers, collectively cited 1,000 times in media posts.
  - Among these papers,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

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

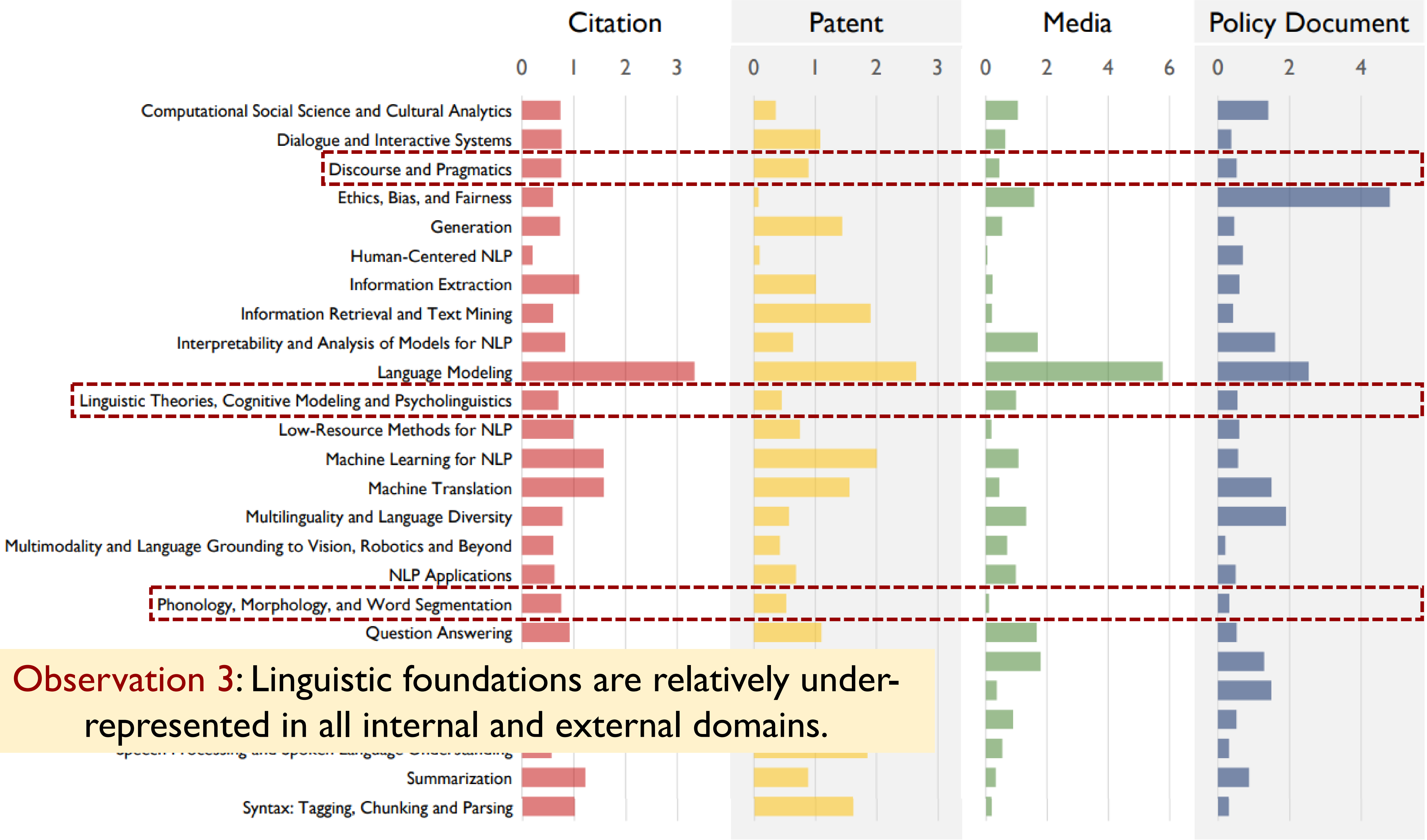

Citation
Patent
Media
Policy Document
0
1
2
3
0
1
2
3
0
2
4
6
0
2
4
Computational Social Science and Cultural Analytics
Dialogue and Interactive Systems
Discourse and Pragmatics
Ethics, Bias, and Fairness
Generation
Human-Centered NLP
Information Extraction
Information Retrieval and Text Mining
Interpretability and Analysis of Models for NLP
Language Modeling
Linguistic Theories, Cognitive Modeling and Psycholinguistics
Low-Resource Methods for NLP
Machine Learning for NLP
Machine Translation
Multilinguality and Language Diversity
Multimodality and Language Grounding to Vision, Robotics and Beyond
NLP Applications
Phonology, Morphology, and Word Segmentation
Question Answering
Resources and Evaluation
Semantics: Lexical, Sentence-Level Semantics, Textual Inference and Other Areas
Sentiment Analysis, Stylistic Analysis, and Argument Mining
Speech Processing and Spoken Language Understanding
Summarization
Syntax: Tagging, Chunking and Parsing

Observation 1: Papers on *language modeling* present a broader impact across all internal and external domains.





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



Observation 3: Linguistic foundations are relatively under-represented in all internal and external domains.

Correlation between Internal and External Impacts

	Patent	Media	Policy Document
Corr(Citation, ·)	0.654	0.725	0.247 (0.599 if excluding "*Ethics, Bias, and Fairness*")

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

Complementarity of Different External Impacts

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

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

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

Final Project Presentation (Next Tuesday & Next Thursday)

- 5 groups
- Each group has 18 minutes for presentation and 5 minutes for Q&A.
  - The number of presenters per group is not limited.

- If you would like to use the instructor's laptop, please send me the slides via email at least 30 minutes before the lecture.

- Presentation order: Last name in reverse alphabetical order
  - 1. Shuo and Hangxiao (Next Tuesday; 4/22)
  - 2. Yichen and Ethan (Next Tuesday; 4/22)
  - 3. Omnia and Michael (Next Thursday; 4/24)
  - 4. Shaohuai (Next Thursday; 4/24)
  - 5. Hasnat and Rithik (Next Thursday; 4/24)

Final Project Presentation (Next Tuesday & Next Thursday)

- Grading Criteria
  - Task background (1%)
  - Task definition (1%)
  - Related work and their limitations (1%)
  - Proposed solution (3%) – model architecture, objective function, …
  - Data (2%) – dataset statistics, collection/annotation process, …
  - Quantitative results (3%) – metric, comparisons with the baseline, ablation study
    - You should have at least one baseline and at least one ablation version
  - Qualitative results (2%) – case study, error analysis, …
  - Unfinished parts (1%) – if you have unfinished parts, explain how to finish them in ~10 days; if you have finished everything except report writing, you can skip this.
  - Conclusions and future work (1%)

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html