CSCE 689 - Special Topics in NLP for Science

Lecture 7: Scientific Literature Retrieval

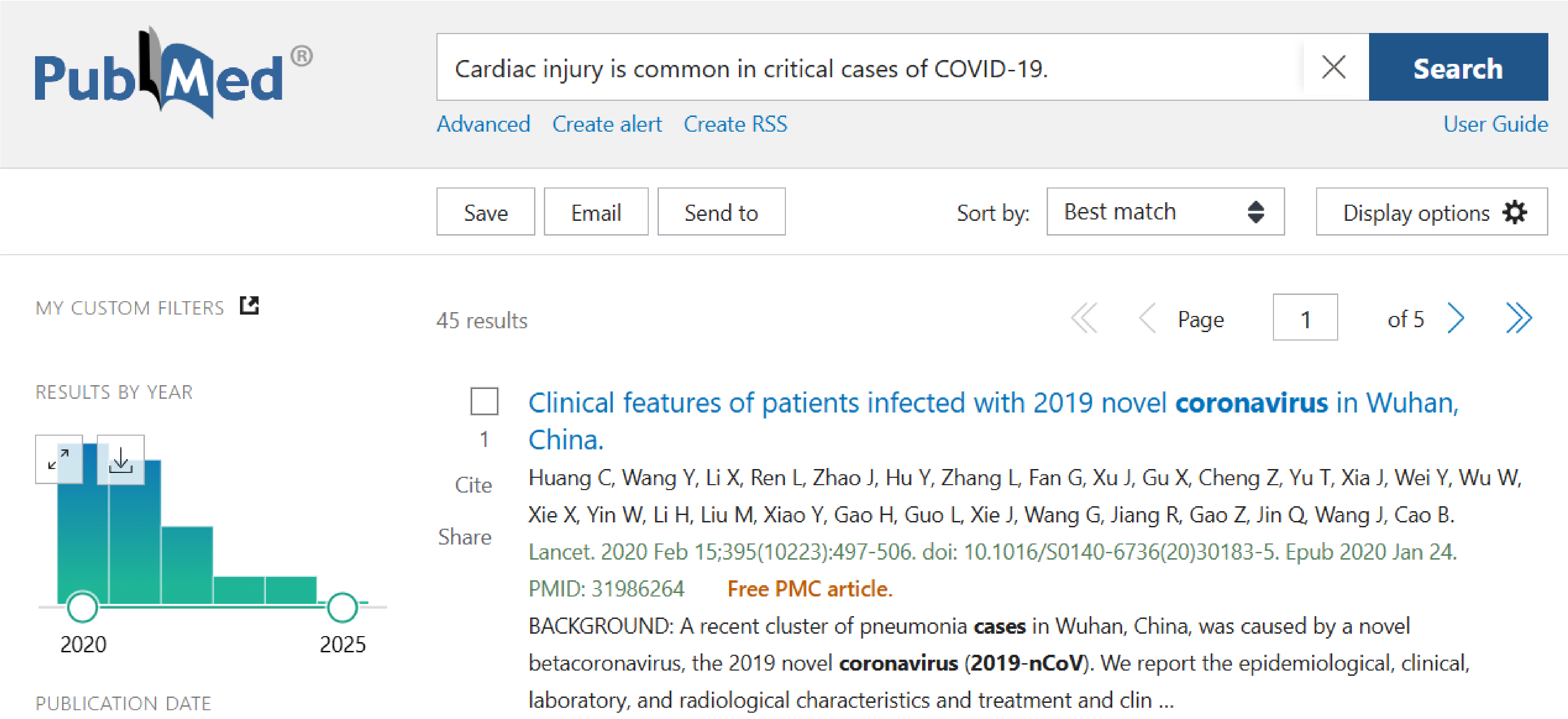
Yu Zhang

yuzhang@tamu.edu

February 6, 2025

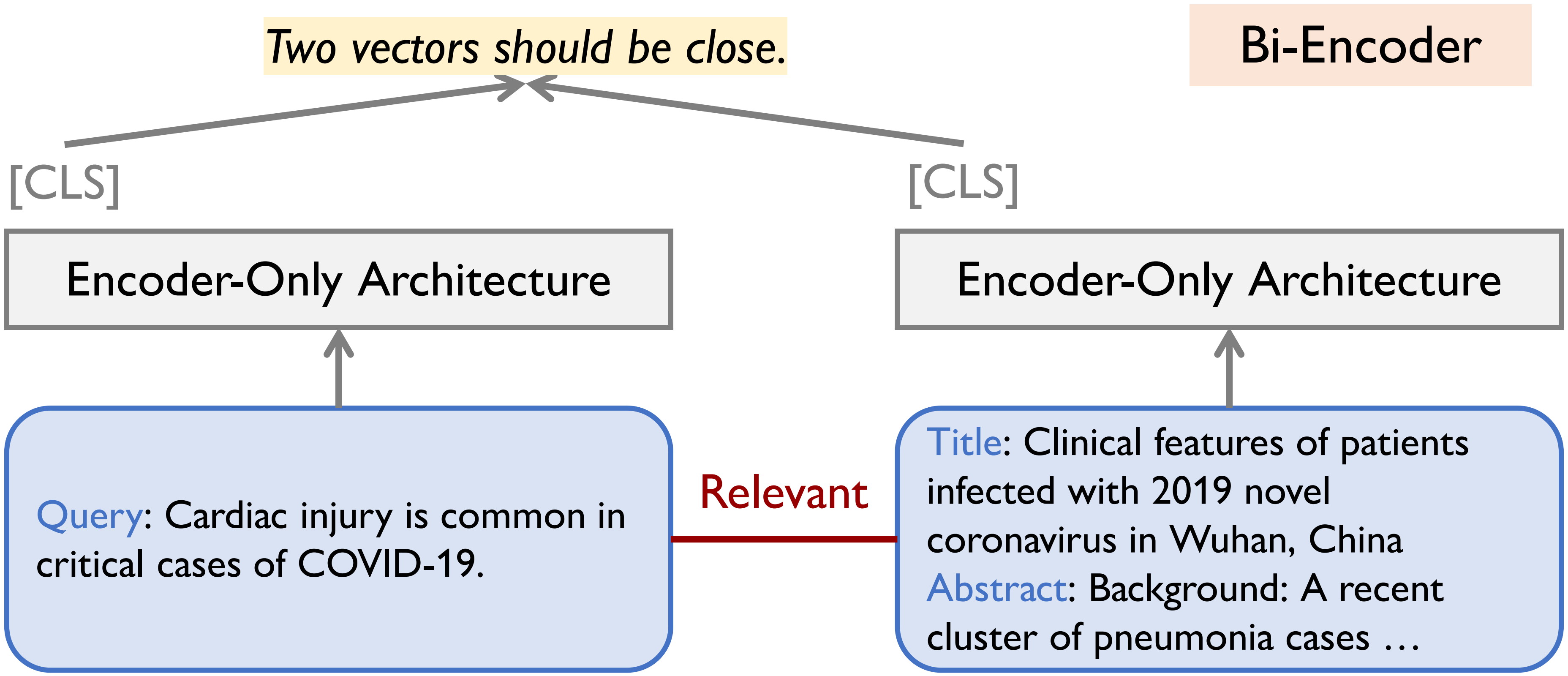
Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

Scientific Literature Retrieval



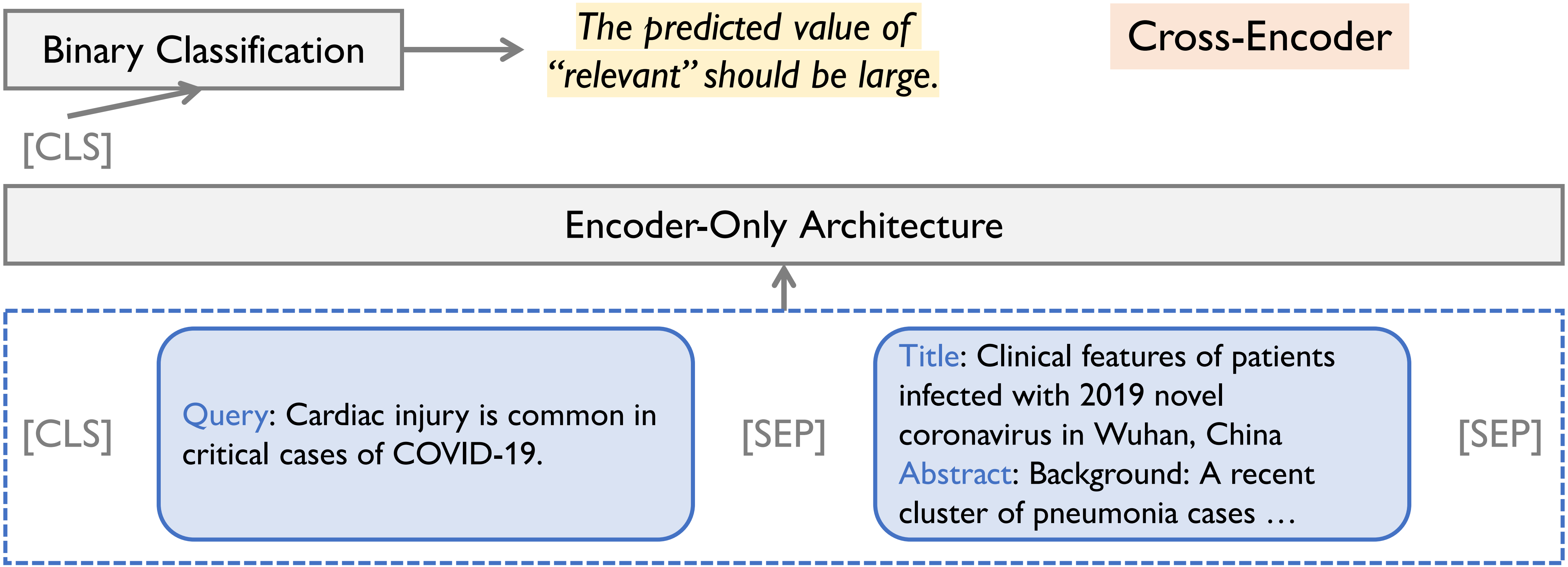
How to train an LLM to perform scientific literature retrieval?

- Step 1: Collect a large number of relevant (query, paper) pairs.
- Step 2: Train an LLM with such information (e.g., using contrastive learning).



How to train an LLM to perform scientific literature retrieval?

- Step 1: Collect a large number of relevant (query, paper) pairs.
- Step 2: Train an LLM with such information (e.g., using contrastive learning).



How to train an LLM to perform scientific literature retrieval?

- Step 1: Collect a large number of relevant (query, paper) pairs.
- How?
  - Unlike citation information that can be crawled from the academic databases or the Web, relevant (query, paper) pairs need to be derived from either user click-through data or human annotations.
  - User click-through data are proprietary.
  - Human annotations cannot be scaled up.

Agenda

- Contrastive Learning with Ground-Truth Search Logs
  - MedCPT: Bi-Encoder → Cross-Encoder
- Contrastive Learning with Data from Other Tasks
  - SciMult: Mixture-of-Experts Transformer
  - BMRetriever: Instruction Tuning
- Application
  - SciFact: Scientific Claim Verification

Agenda

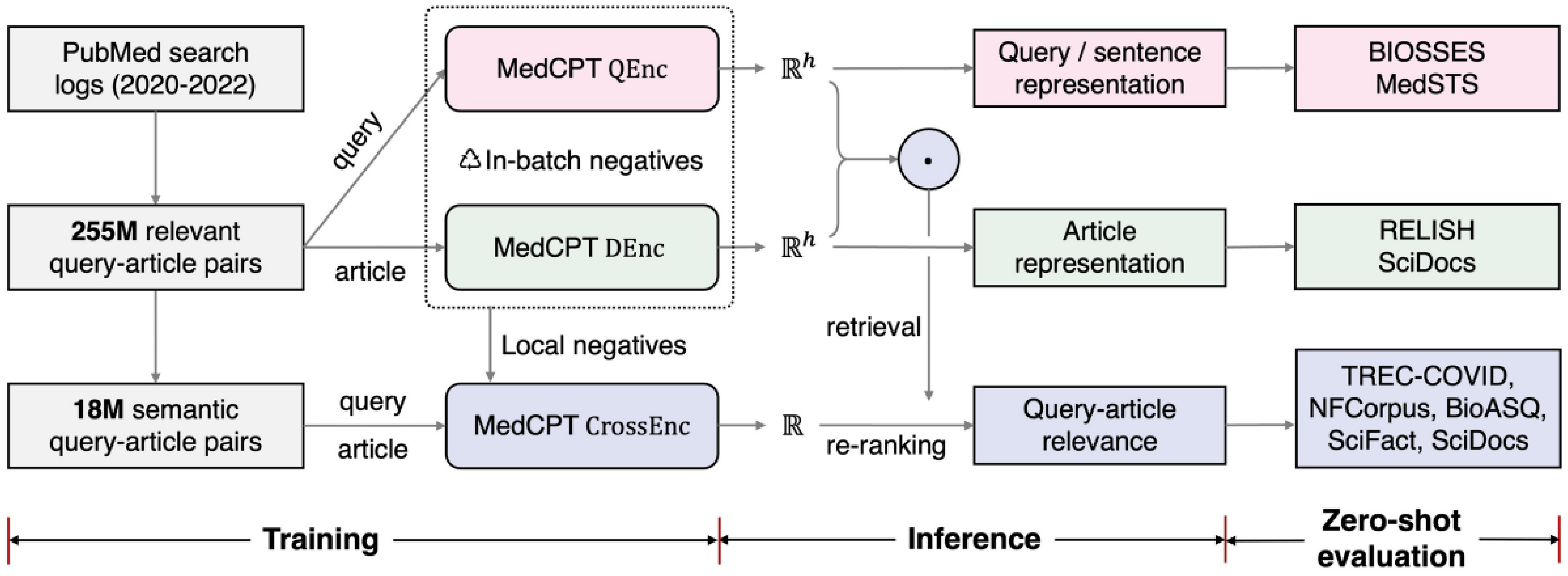
- Contrastive Learning with Ground-Truth Search Logs
  - MedCPT: Bi-Encoder → Cross-Encoder
- Contrastive Learning with Data from Other Tasks
  - SciMult: Mixture-of-Experts Transformer
  - BMRetriever: Instruction Tuning
- Application
  - SciFact: Scientific Claim Verification

PubMed Search Logs

- User click-through data from 2020 to 2022
  - A user inputted a query.
  - 20 papers were displayed on the result page.
  - The user clicked paper 1, 6, and 8.
  - Papers relevant to the query: 1, 6, 8
  - Papers irrelevant to the query: 2, 3, 4, 5, 7
  - Papers cannot be judged as relevant/irrelevant: 9, 10, …, 20
- 255M relevant (query, paper) pairs
  - Most of such queries are short keywords, and matching them to the clicked articles is a relatively simple task.
- 18M semantically relevant (query, paper) pairs
  - Remove queries either having only one word or all of the clicked articles containing exact mentions of the whole input query

The MedCPT Framework

- Bi-Encoder for retrieval (from a large candidate pool)
- Cross-Encoder for re-ranking (the retrieved papers)



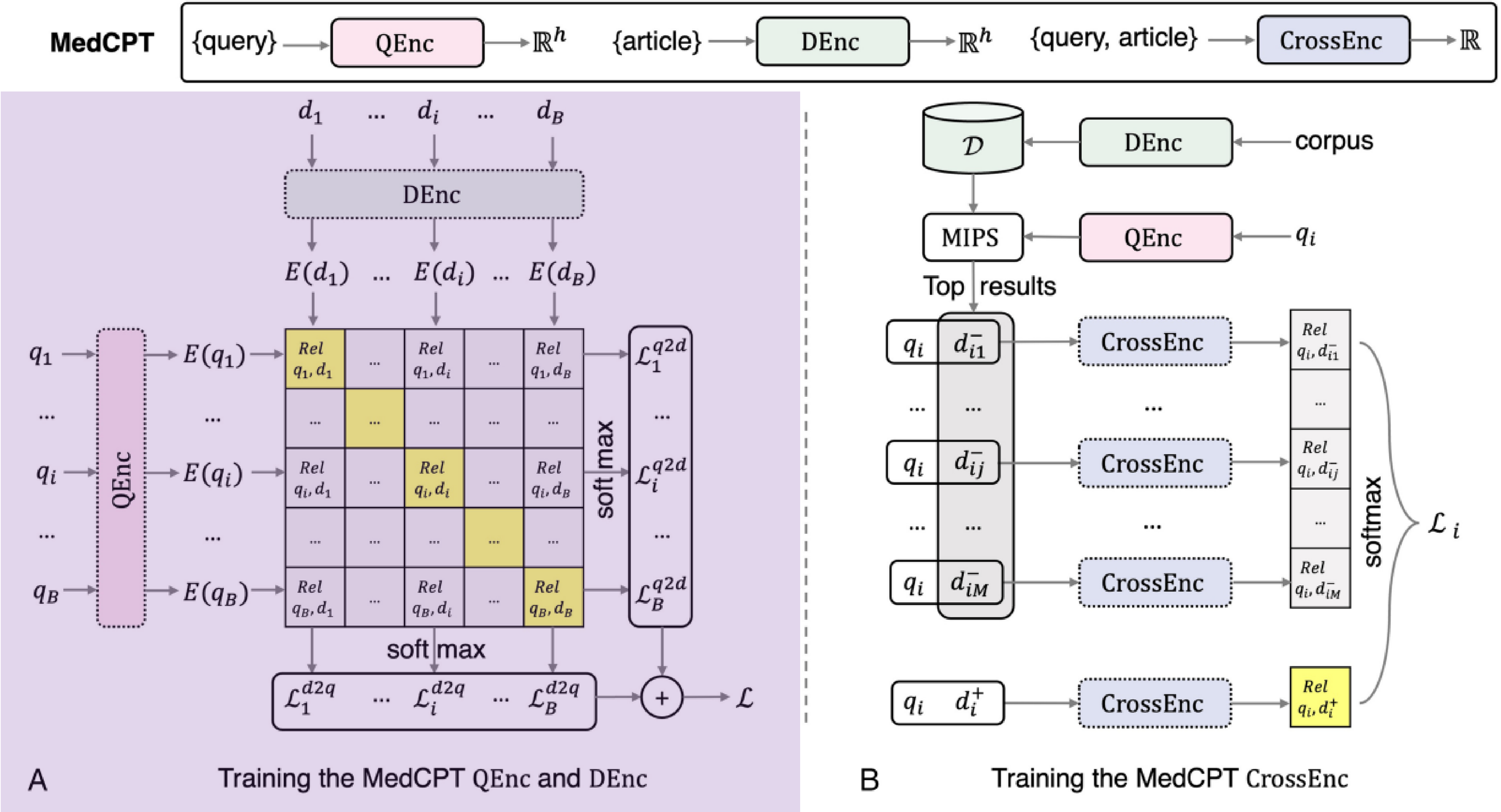
MedCPT: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot Biomedical Information Retrieval. Bioinformatics 2023.

Recap: Bi-Encoder vs. Cross-Encoder

- Bi-Encoder is much more efficient during the inference time.
- If we use Cross-Encoder, the query and the paper can serve as context of each other, so that the model can learn a better contextualized representation of each token in the input sequence.

- MedCPT: Using Bi-Encoder to remove most (e.g., 99%) of the candidates, and using Cross-Encoder to more carefully rank the remaining candidates (e.g., 1%).

Bi-Encoder Contrastive Learning

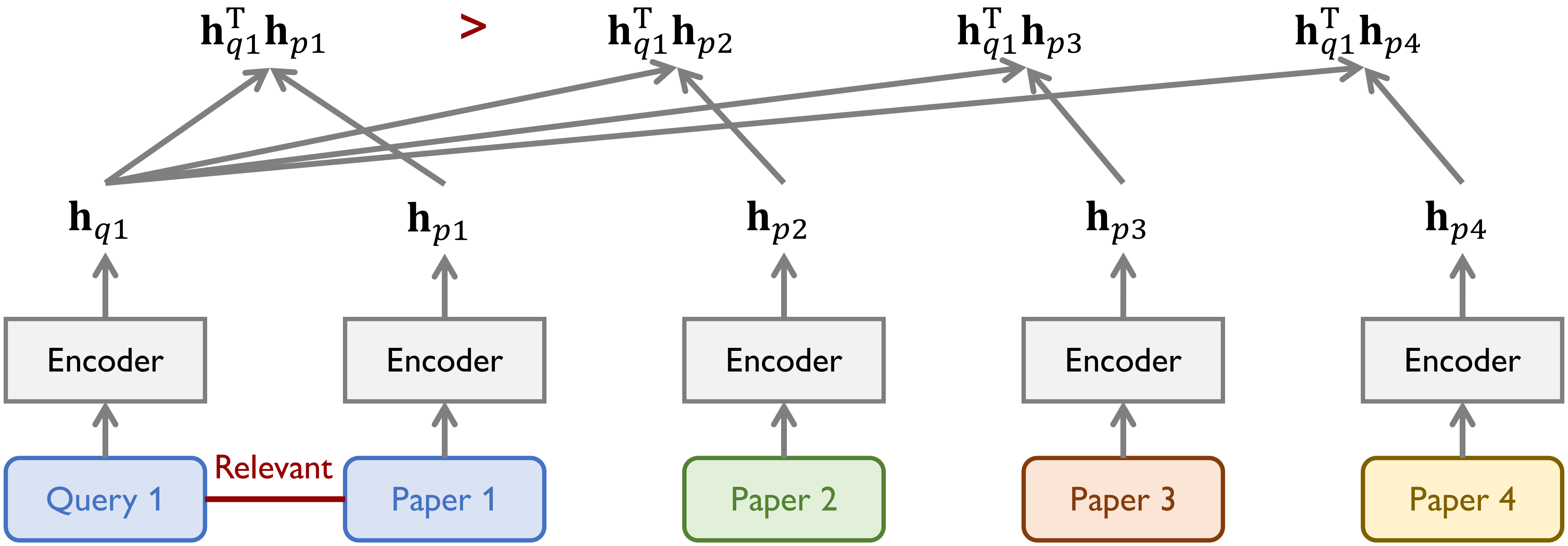


In-Batch Negative Sampling

Objective Function: maximize

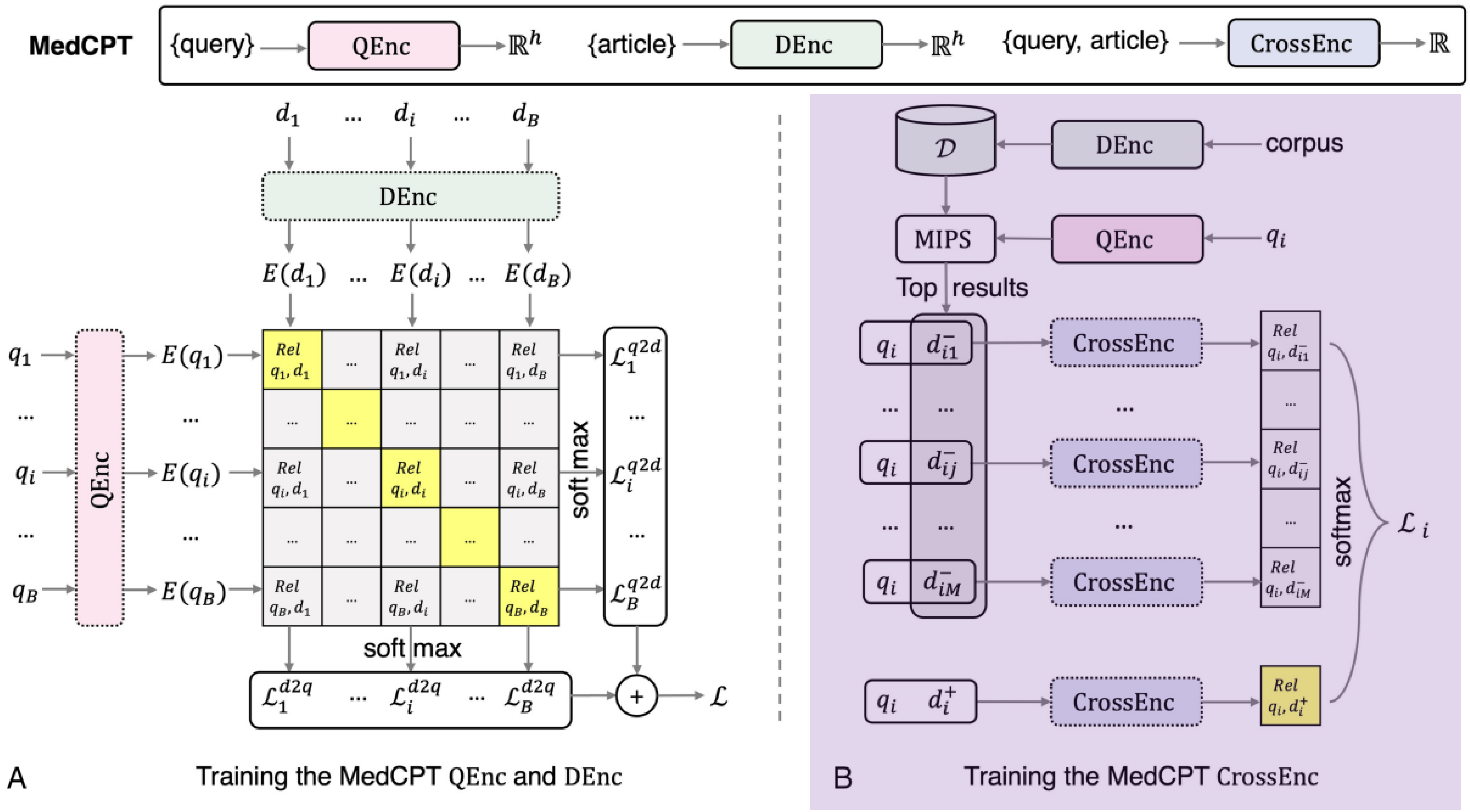
$$\frac{\exp(\mathbf{h}_{q1}^{\mathrm{T}}\mathbf{h}_{p1})}{\exp(\mathbf{h}_{q1}^{\mathrm{T}}\mathbf{h}_{p1}) + \exp(\mathbf{h}_{q1}^{\mathrm{T}}\mathbf{h}_{p2}) + \exp(\mathbf{h}_{q1}^{\mathrm{T}}\mathbf{h}_{p3}) + \exp(\mathbf{h}_{q1}^{\mathrm{T}}\mathbf{h}_{p4})}$$

- Paper 2 is relevant to Query 2, but its relevance to Query 1 is unknown.



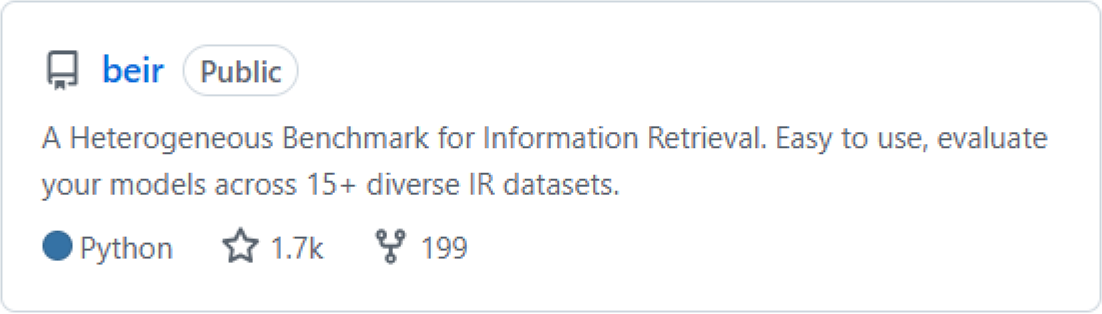
Cross-Encoder Contrastive Learning

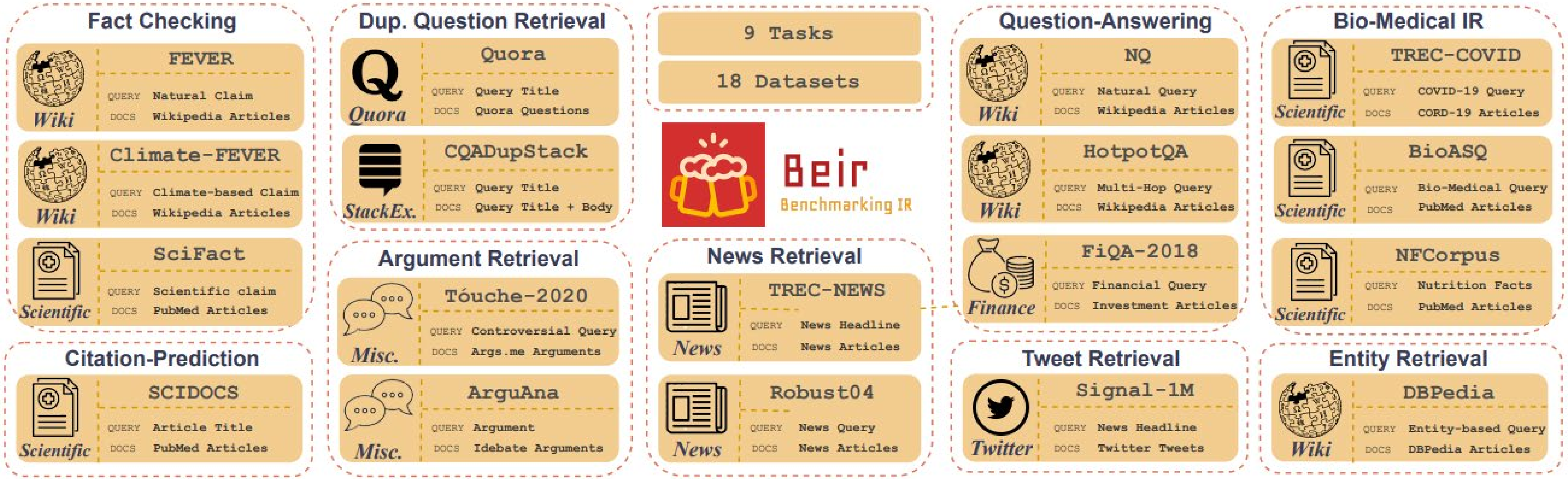
Using trained Bi-Encoder to derive hard negatives



Evaluation of MedCPT

- The BEIR benchmark
  - https://github.com/beir-cellar/beir





BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS 2021.

Performance of MedCPT: Query-Paper Relevance

Method	Size	COVID	NFC	BioASQ	SciFact	SciDocs	Avg.
Sparse retrievers							
BM25		0.656	0.325	0.465	0.665	0.158	*0.454*
BM25 + MiniLM	66M	**0.757**	0.350	0.523	0.688	*0.166*	0.497
DeepCT	110M	0.406	0.283	0.407	0.630	0.124	0.370
SPARTA	110M	0.538	0.301	0.351	0.582	0.126	0.380
docT5query	220M	0.713	0.328	0.431	0.675	0.162	0.462
Dense retrievers							
DPR	110M	0.332	0.189	0.127	0.318	0.077	0.209
ANCE	110M	0.654	0.237	0.306	0.507	0.122	0.365
TAS-B	66M	0.481	0.319	0.383	0.643	0.149	0.395
GenQ	220M	0.619	0.319	0.398	0.644	0.143	0.425
Contriever	110M	0.596	0.328		0.677	0.165	
Contriever + MiniLM	176M	*0.701*	0.344		0.692	0.171	
ColBERT	110M	0.677	0.305	*0.474*	0.671	0.145	*0.454*
Large language model retrievers							
Google GTR-Base	110M	0.539	0.308	0.271	0.600	0.149	0.373
Google GTR-Large	335M	0.557	0.329	0.320	0.639	0.158	0.401
Google GTR-XL	1.24B	0.584	0.343	0.317	0.635	0.159	0.408
Google GTR-XXL	*4.80B*	0.501	0.342	0.324	0.662	0.161	0.398
OpenAI cpt-text-S	300M	0.679	0.332		0.672		
OpenAI cpt-text-M	1.20B	0.585	*0.367*		0.704		
OpenAI cpt-text-L	6.00B	0.562	0.380		*0.744*		
OpenAI cpt-text-XL	**175B**	0.649	**0.407**		0.754		
MedCPT							
MedCPT	330M	0.709	0.355	**0.553**	**0.761**	**0.172**	**0.510**
MedCPT (retriever only)	220M	0.697	0.340	0.332	0.724	0.123	0.443
MedCPT w/o contrastive pre-training (PubMedBERT)	110M	0.059	0.015		0.010	0.004	

Performance of MedCPT: Paper-Paper and Query-Query Relevance

Table 2. Evaluation results of the MedCPT article encoder on the RELISH dataset.[a]

Method	MAP			NDCG			Avg.
	@5	@10	@15	@5	@10	@15	
Random	79.33	77.22	75.41	80.70	77.67	76.40	77.79
Sparse retrievers							
BM25	88.91	86.72	84.54	89.48	87.39	86.21	87.21
PMRA	90.30	87.57	85.75	90.95	88.40	87.45	88.40
Non-BERT embedding-based models							
fastText	85.75	82.81	81.79	86.79	83.79	83.12	84.01
BioWordVec	89.84	86.51	84.67	89.90	86.67	85.53	87.19
InferSent	85.21	82.16	80.41	86.56	83.31	82.35	83.33
WikiSentVec	87.92	85.23	83.40	88.65	85.74	84.81	85.96
BioSentVec	90.76	88.10	86.16	90.05	87.76	86.89	88.29
LDA	85.44	82.66	80.36	86.51	82.91	81.31	83.20
Doc2Vec	86.23	84.74	83.39	86.55	84.70	84.09	84.95
BERT-based models							
BioBERT	88.14	85.81	83.90	88.97	86.29	85.10	86.37
PubMedBERT	83.69	81.07	79.53	85.47	82.39	81.41	82.26
SPECTER	*92.27*	*90.00*	*88.36*	*91.47*	*89.12*	*88.42*	*89.94*
SciNCL	94.72	92.74	91.14	93.67	91.91	90.94	92.52
MedCPT DEnc	**95.58**	**93.99**	**92.39**	**94.78**	**93.12**	**92.43**	**93.72**

Evaluating relevance between papers

Table 3. Evaluation results (Pearson's correlation coefficients) of the MedCPT QEnc on the BIOSSES and MedSTS datasets.[a]

Model	BIOSSES	MedSTS
Non-BERT embedding-based models		
BioWordVec	0.694	0.747
USE	0.345	0.714
BioSentVec (PubMed)	*0.817*	0.750
BioSentVec (MIMIC-III)	0.350	*0.759*
BioSentVec (PubMed + MIMIC-III)	0.795	**0.767**
BERT-based models		
PubMedBERT	0.528	0.521
Clinical BERT	0.556	0.525
SPECTER	0.694	0.702
SciNCL	0.847	0.706
MedCPT QEnc	**0.893**	0.765

Evaluating relevance between short sentences

Take-Away Messages

- Contrastive learning with user click-through data makes small, domain-specific LMs outperform large, general LMs.
- The retrieval→re-ranking framework uses Bi-Encoder to filter out most (e.g., 99%) of the candidates, and using Cross-Encoder to more carefully rank the remaining candidates (e.g., 1%).
  - "Get the best of both worlds" by utilizing the advantages of Bi-Encoder and Cross-Encoder
- Limitation:
  - Strong reliance on proprietary data
    - Most researchers do not have access to search logs.

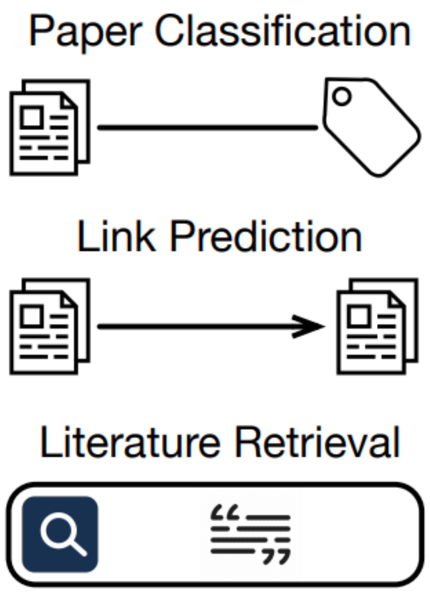
Agenda

- Contrastive Learning with Ground-Truth Search Logs
  - MedCPT: Bi-Encoder → Cross-Encoder
- Contrastive Learning with Data from Other Tasks
  - SciMult: Mixture-of-Experts Transformer
  - BMRetriever: Instruction Tuning
- Application
  - SciFact: Scientific Claim Verification

Harvesting Other Types of Relevant (Text, Text) Pairs

PubMed

Semantic Scholar



Relevant (Paper, Label) Pairs

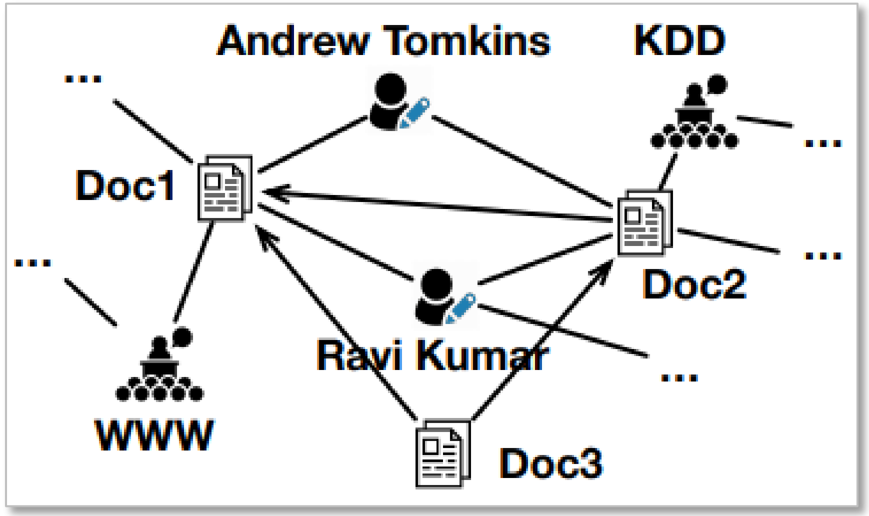
Relevant (Paper, Paper) Pairs

Relevant (Query, Paper) Pairs

- Combine all these pairs together for contrastive learning?
- Task Interference: The model is confused by different types of "relevance".

An Illustrative Example of Task Interference

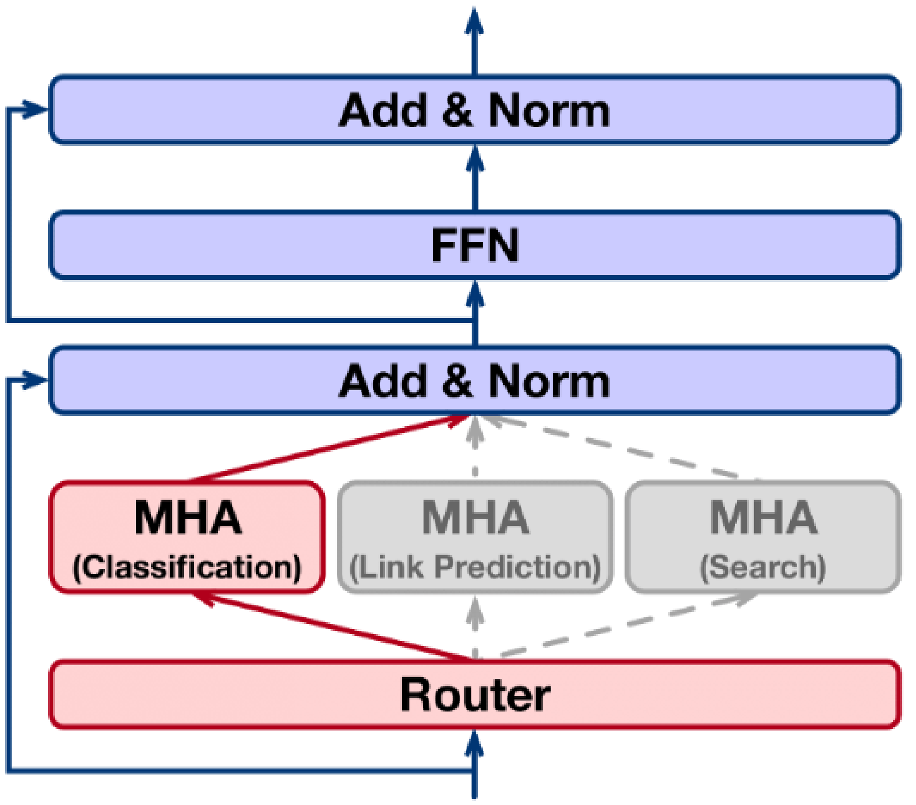
- Recall the link prediction problem
- Imagine that predicting each type of "links" is a "task"
  - Citation Prediction: Paper→Paper
  - Same Author Prediction: Paper-Author-Paper
  - Each type of "links" defines one type of "relevance".
- Directly merging the relevant (paper, paper) pairs induced by different link types?
  - The model will be confused!



(Doc3, Doc2) are relevant according to Paper→Paper but irrelevant according to Paper-Author-Paper.

Tackling Task Interference: Mixture-of-Experts Transformer

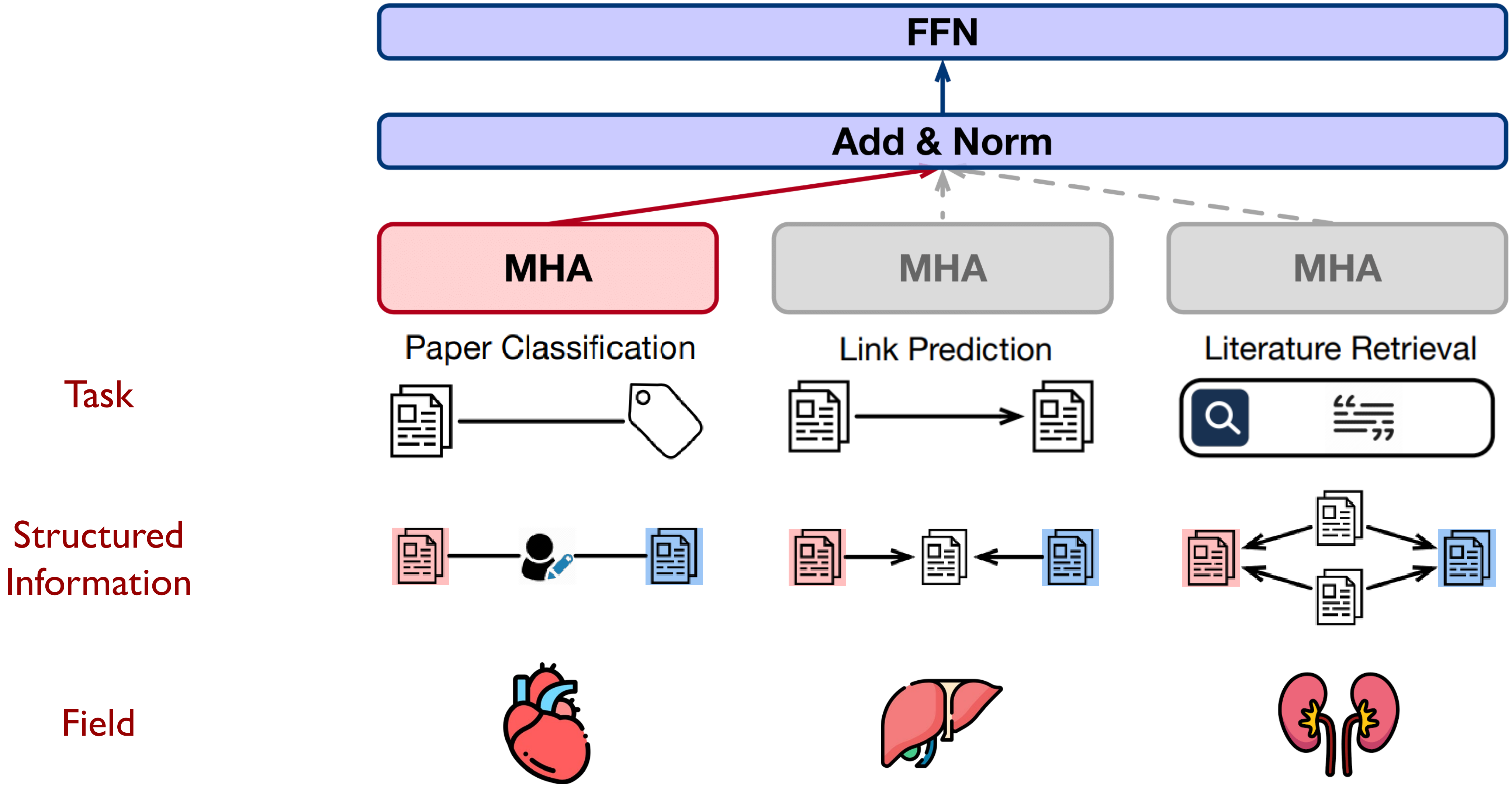
- A typical Transformer layer
  - 1 Multi-Head Attention (MHA) sublayer
  - 1 Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
  - Multiple MHA sublayers
  - 1 FFN sublayer
  - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be an "expert"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



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

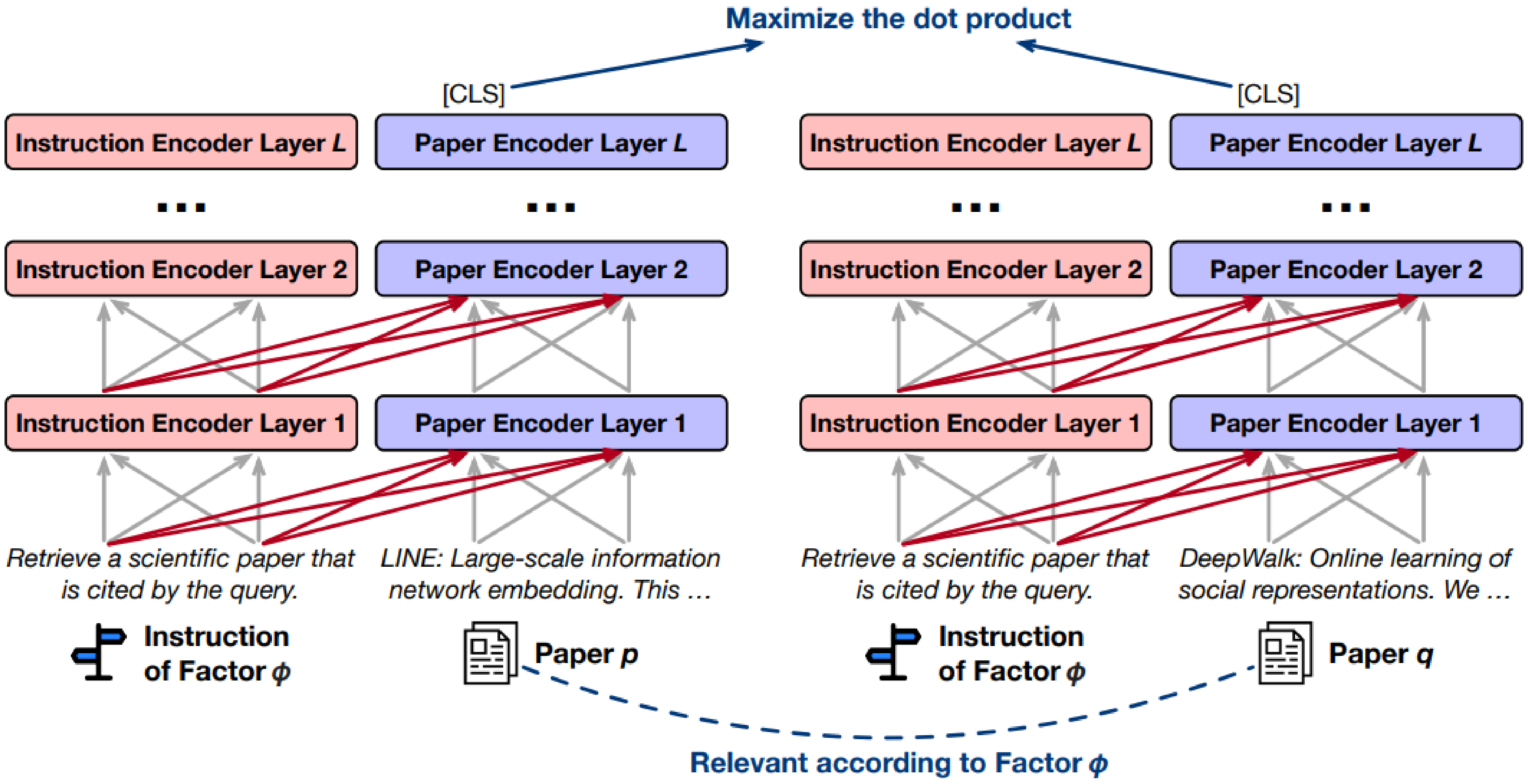
Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. EMNLP 2023 Findings.

Tackling Task Interference: Mixture-of-Experts Transformer



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.



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

Evaluations of SciMult

Task	Pre-training	In-domain Evaluation	Cross-domain Evaluation
Classification	**MAPLE** (Zhang et al., 2023b) {CS-Journal, Biology-MeSH, Medicine-MeSH}	**MAPLE** (Zhang et al., 2023b) {CS-Conference, Chemistry-MeSH}, **SciDocs** (Cohan et al., 2020) {MAG Fields, MeSH Diseases}	**MAPLE** (Zhang et al., 2023b) {Geography, Psychology}
Link Prediction	**Citation Prediction Triplets** (Cohan et al., 2020)	**SciDocs** (Cohan et al., 2020) {Co-view, Co-read, Cite, Co-cite}	**Recommendation** (Kanakia et al., 2019), **PMC-Patients** (Zhao et al., 2022)
Search	**SciRepEval-Search** (Singh et al., 2022)	**SciRepEval-Search** (Singh et al., 2022)	**TREC-COVID** (Voorhees et al., 2021), **SciFact** (Wadden et al., 2020), **NFCorpus** (Boteva et al., 2016)

- For Search, both the retrieval and the re-ranking settings are evaluated.

Search		
SciRepEval-Search (Singh et al., 2022)	2,637	reranking, 10.00 for each query on average
TREC-COVID in SciRepEval (Voorhees et al., 2021)	50	reranking, 1386.36 for each query on average
TREC-COVID in BEIR (Voorhees et al., 2021)	50	171,332
SciFact (Wadden et al., 2020)	1,109	5,183
NFCorpus (Boteva et al., 2016)	3,237	3,633

Performance of SciMult: Search

Search	SciRepEval (Singh et al., 2022)	BEIR (Thakur et al., 2021)				
	Search (2022)	TREC-COVID (2021)	TREC-COVID (2021)	SciFact (2020)	NFCorpus (2016)	
	nDCG@10	nDCG@10	nDCG@10	nDCG@10	nDCG@10	Average
BM25	73.47	55.86	57.79	65.63	30.00	56.55
SciBERT	71.39	40.98	4.17	0.88	1.90	23.86
SentBERT	71.84	51.30	20.73	9.40	6.69	31.99
SPECTER	73.42	66.45	29.91	49.74	15.83	47.07
PubMedBERT	70.77	45.28	7.56	0.30	1.09	25.00
LinkBERT	71.66	52.45	2.28	0.49	1.77	25.73
BioLinkBERT	71.18	36.01	3.17	0.12	0.98	22.29
OAG-BERT	72.17	55.09	7.11	18.33	8.48	32.24
SciNCL	73.78	73.50	34.69	56.51	22.34	52.16
SPECTER 2.0	**78.22**†	79.43	58.48	67.16	22.84	61.23
SciMult-Vanilla	76.44	**86.76**	67.22	**70.76**	**31.20**	66.48
SciMult-MHAExpert	76.33	86.29	**71.18**	70.67	30.79	**67.05**
SciMult-FFNExpert	76.02	82.32	52.15	63.57	27.48	60.31
SciMult-Prefix	76.55	82.83	68.15	70.70	30.02	65.65
SciMult-Instruction	75.86	83.59	61.05	70.62	30.25	64.27

Performance of SciMult: Classification

Fine-grained classification	MAPLE (Zhang et al., 2023b)												
	CS-Conference			**Chemistry-MeSH**			**Geography**			**Psychology**			
	R@20	**R@50**	**R@100**	**R@20**	**R@50**	**R@100**	**R@20**	**R@50**	**R@100**	**R@20**	**R@50**	**R@100**	**Average**
SciBERT (Beltagy et al., 2019)	42.01	42.84	43.87	30.53	31.46	32.15	52.04	54.53	58.11	43.07	44.02	45.22	43.32
SentBERT (Reimers and Gurevych, 2019)	42.79	44.34	45.96	30.75	31.73	32.44	53.54	57.23	61.11	43.33	44.60	46.37	44.52
SPECTER (Cohan et al., 2020)	47.38	53.18	58.43	34.26	39.35	43.41	59.12	65.33	70.75	47.07	51.30	56.17	52.15
PubMedBERT (Gu et al., 2021)	41.93	42.56	43.24	30.46	31.46	31.83	52.19	54.82	56.88	43.93	46.28	49.27	43.74
LinkBERT (Yasunaga et al., 2022)	42.15	43.16	44.22	30.52	31.56	32.37	50.58	50.94	51.63	42.62	42.90	43.23	42.16
BioLinkBERT (Yasunaga et al., 2022)	42.00	42.81	43.57	30.37	31.15	31.48	50.36	50.54	50.86	42.39	42.55	42.79	41.74
OAG-BERT (Liu et al., 2022)	42.59	43.79	44.93	30.58	31.97	32.62	51.44	52.25	53.16	42.63	42.95	43.30	42.68
SciNCL (Ostendorff et al., 2022)	47.92	53.57	58.29	34.99	40.50	44.64	59.00	65.49	71.41	48.74	54.21	59.84	53.22
SPECTER 2.0 (Singh et al., 2022)	48.63	55.09	60.68	36.17	43.06	48.26	62.87	70.30	76.37	50.60	58.27	65.66	56.33
SciMult-Vanilla	53.40	64.70	74.09	**39.78**	**51.31**	**59.75**	62.08	70.65	77.79	50.42	56.58	63.17	60.31
SciMult-MHAExpert	**54.02**	**65.49**	**75.07**	39.41	50.92	59.59	**65.94**	**75.01**	**81.93**	**51.77**	**59.55**	**67.86**	**62.21**
SciMult-FFNExpert	53.73	63.79	72.46	38.01	48.76	57.43	61.90	70.69	78.81	50.09	56.94	64.28	59.74
SciMult-Prefix	53.68	63.62	72.07	37.97	48.95	57.56	62.86	71.65	79.71	50.10	57.25	64.53	60.00
SciMult-Instruction	53.78	63.99	72.72	38.81	50.12	58.96	63.26	71.74	79.52	50.86	58.47	66.46	60.72

Performance of SciMult: Link Prediction

Link Prediction (Reranking)	SciDocs (Cohan et al., 2020)								Kanakia et al. (2019)			
	Co-view		Co-read		Cite		Co-cite		Recommendation			
	MAP	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**	**nDCG@5**	**nDCG@10**	**nDCG**	**Average**
Citeomatic (Bhagavatula et al., 2018)	81.1†	90.2†	80.5†	90.2†	86.3†	94.1†	84.4†	92.8†	–	–	–	–
Kanakia et al. (2019)	–	–	–	–	–	–	–	–	83.88‡	87.71‡	93.59‡	–
SciBERT (Beltagy et al., 2019)	50.7†	73.1†	47.7†	71.1†	48.3†	71.7†	49.7†	72.6†	77.17	82.49	90.86	66.86
SentBERT (Reimers and Gurevych, 2019)	68.2†	83.3†	64.8†	81.3†	63.5†	81.6†	66.4†	82.8†	76.75	81.49	90.80	76.45
SPECTER (Cohan et al., 2020)	83.6†	91.5†	84.5†	92.4†	88.3†	94.9†	88.1†	94.8†	83.38	87.39	93.64	89.32
PubMedBERT (Gu et al., 2021)	59.43	78.23	55.59	75.63	51.81	73.43	58.19	77.80	77.30	82.21	91.09	70.97
LinkBERT (Yasunaga et al., 2022)	44.21	67.76	41.04	65.31	39.33	63.91	42.84	67.18	76.10	80.89	90.47	61.73
BioLinkBERT (Yasunaga et al., 2022)	56.46	76.38	50.76	72.18	47.73	70.55	52.94	74.44	77.02	81.78	90.73	68.27
OAG-BERT (Liu et al., 2022)	64.61	81.50	60.13	78.65	57.35	77.60	62.47	80.92	76.73	82.12	90.96	73.91
SciNCL (Ostendorff et al., 2022)	**85.3**†	**92.3**†	**87.5**†	**93.9**†	93.6†	97.3†	**91.6**†	96.4†	85.33	88.38	94.34	**91.45**
SPECTER 2.0 (Singh et al., 2022)	85.18†	**92.27**†	86.95†	93.53†	92.23†	96.84†	91.13†	96.28†	86.03	89.12	94.59	91.29
SciMult-Vanilla	83.99	91.68	86.66	93.67	91.37	96.26	91.50	**96.45**	**87.32**	89.32	**94.88**	91.19
SciMult-MHAExpert	83.92	91.60	86.45	93.55	92.58	96.92	91.47	96.36	86.68	**89.45**	94.77	91.25
SciMult-FFNExpert	83.23	91.26	85.61	93.20	93.77	97.42	90.39	95.94	85.75	88.45	94.29	90.85
SciMult-Prefix	83.43	91.48	85.89	93.27	**94.28**	**97.60**	90.73	96.09	86.05	88.85	94.66	91.12
SciMult-Instruction	82.13	90.88	84.14	92.36	92.63	96.91	89.27	95.43	86.49	88.81	94.51	90.32

PMC-Patients Leaderboard

- Given a patient summary, find the most relevant papers.

https://pmc-patients.github.io

Patient-to-Article Retrieval (PAR) Leaderboard

	Model	MRR (%)	P@10 (%)	nDCG@10 (%)	R@1k (%)
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

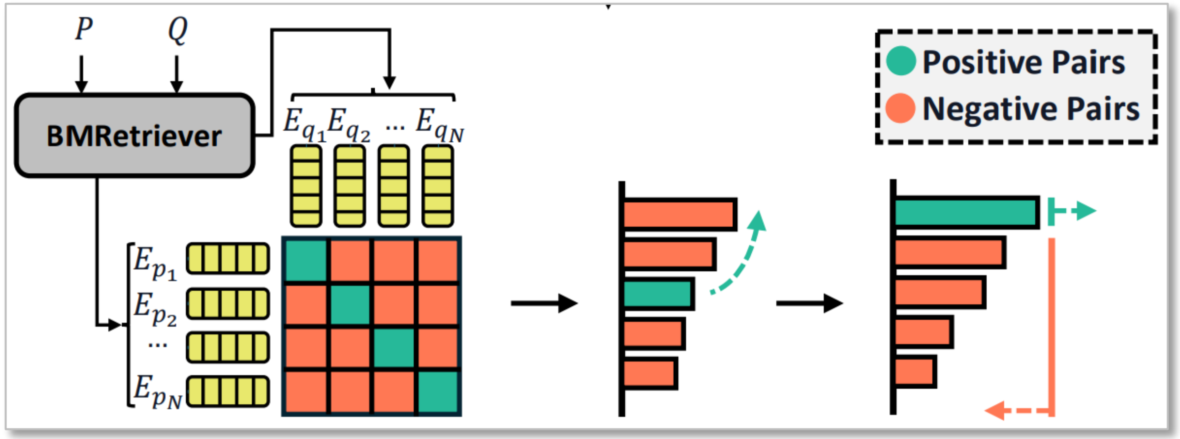
A Large-Scale Dataset of Patient Summaries for Retrieval-Based Clinical Decision Support Systems. Scientific Data 2023.

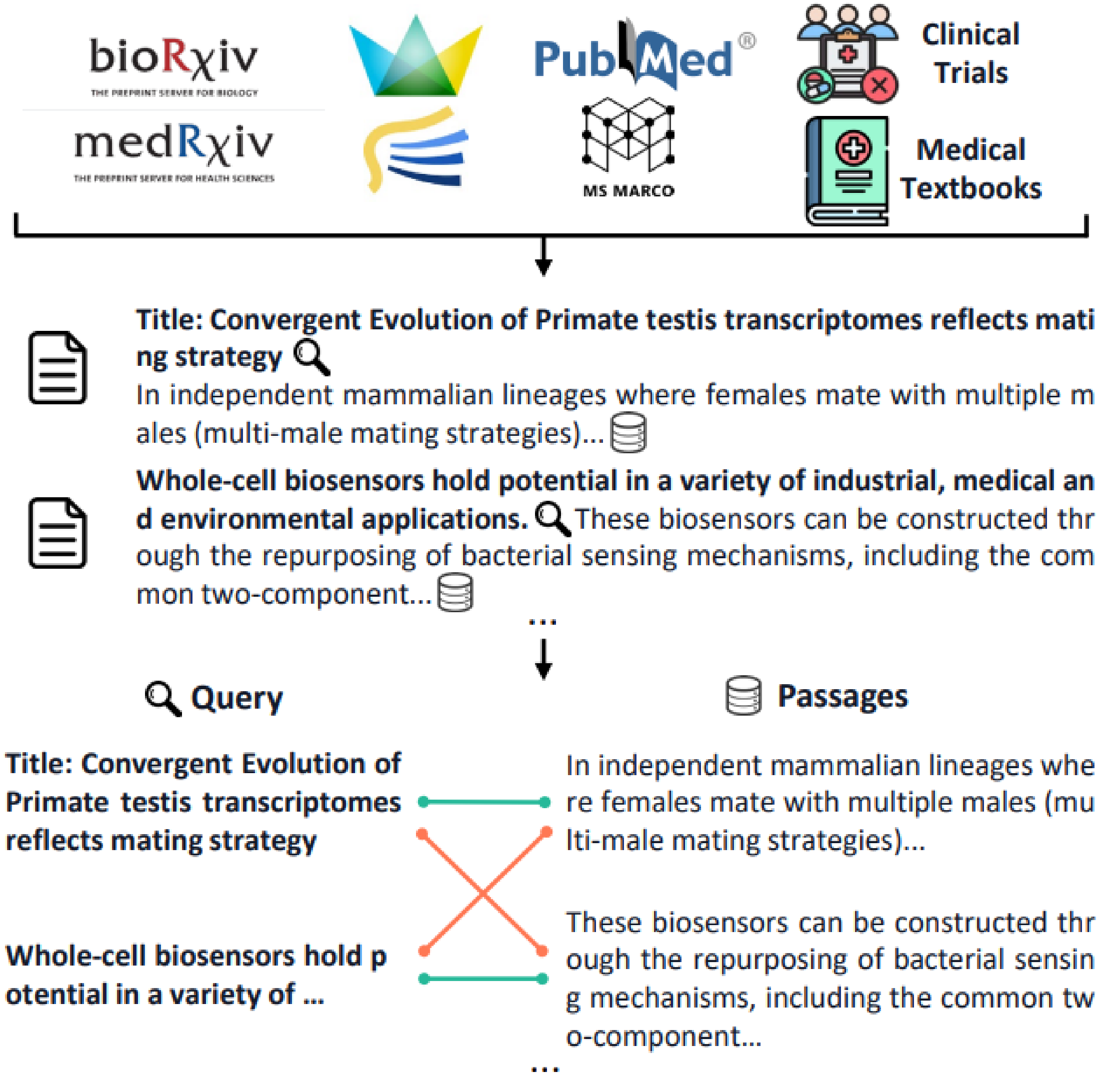
Agenda

- Contrastive Learning with Ground-Truth Search Logs
  - MedCPT: Bi-Encoder → Cross-Encoder
- Contrastive Learning with Data from Other Tasks
  - SciMult: Mixture-of-Experts Transformer
  - BMRetriever: Instruction Tuning
- Application
  - SciFact: Scientific Claim Verification

Getting Relevant (Text, Text) Pairs from One Paper

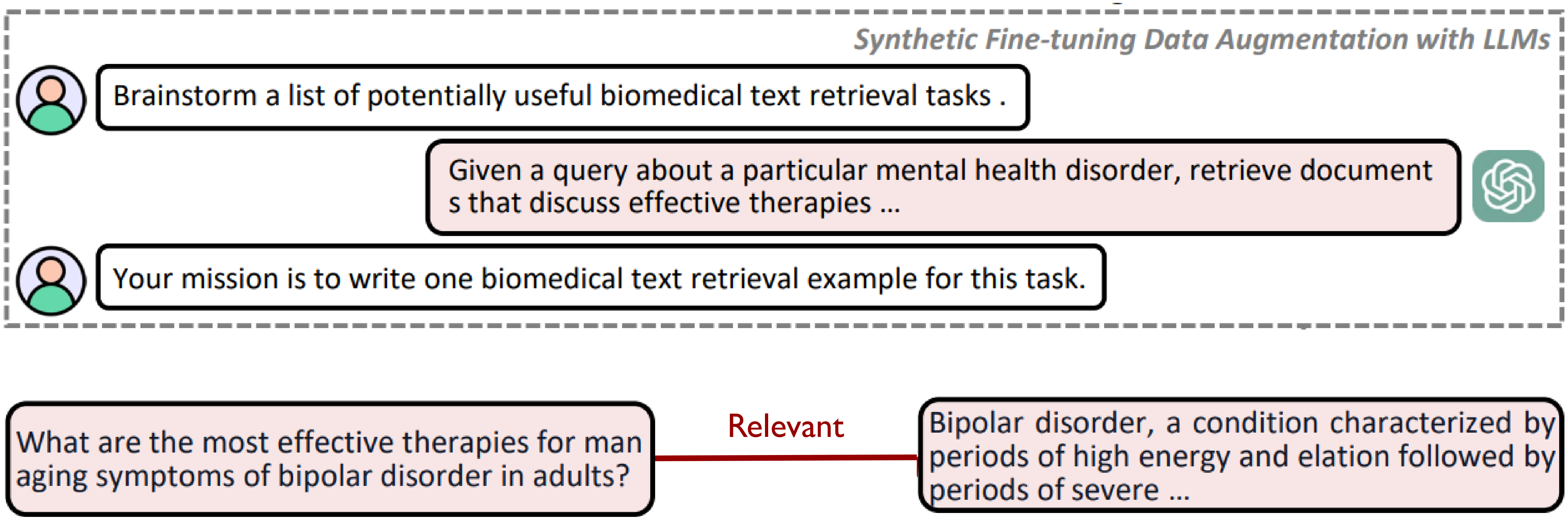
- For corpora with titles, treat the title as the query and the corresponding abstract as the passage.
- For untitled corpora, randomly sample two disjoint passages from documents, using one as the query and the other as the passage.





BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. EMNLP 2024.

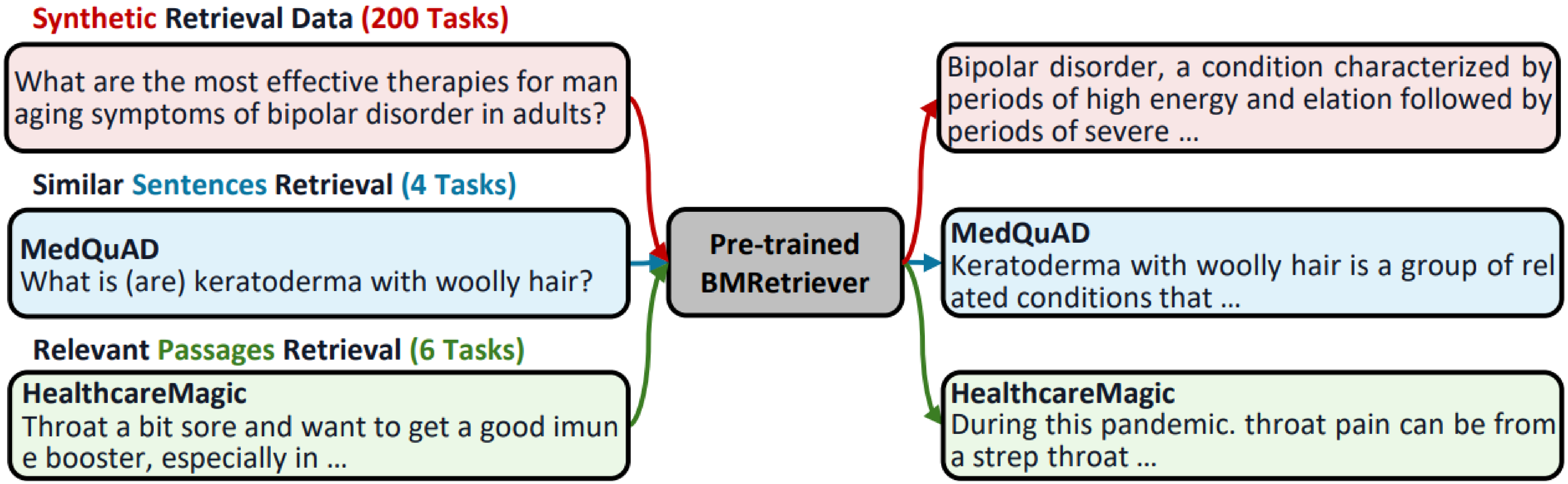
Getting Relevant (Text, Text) Pairs from LLMs



BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. EMNLP 2024.

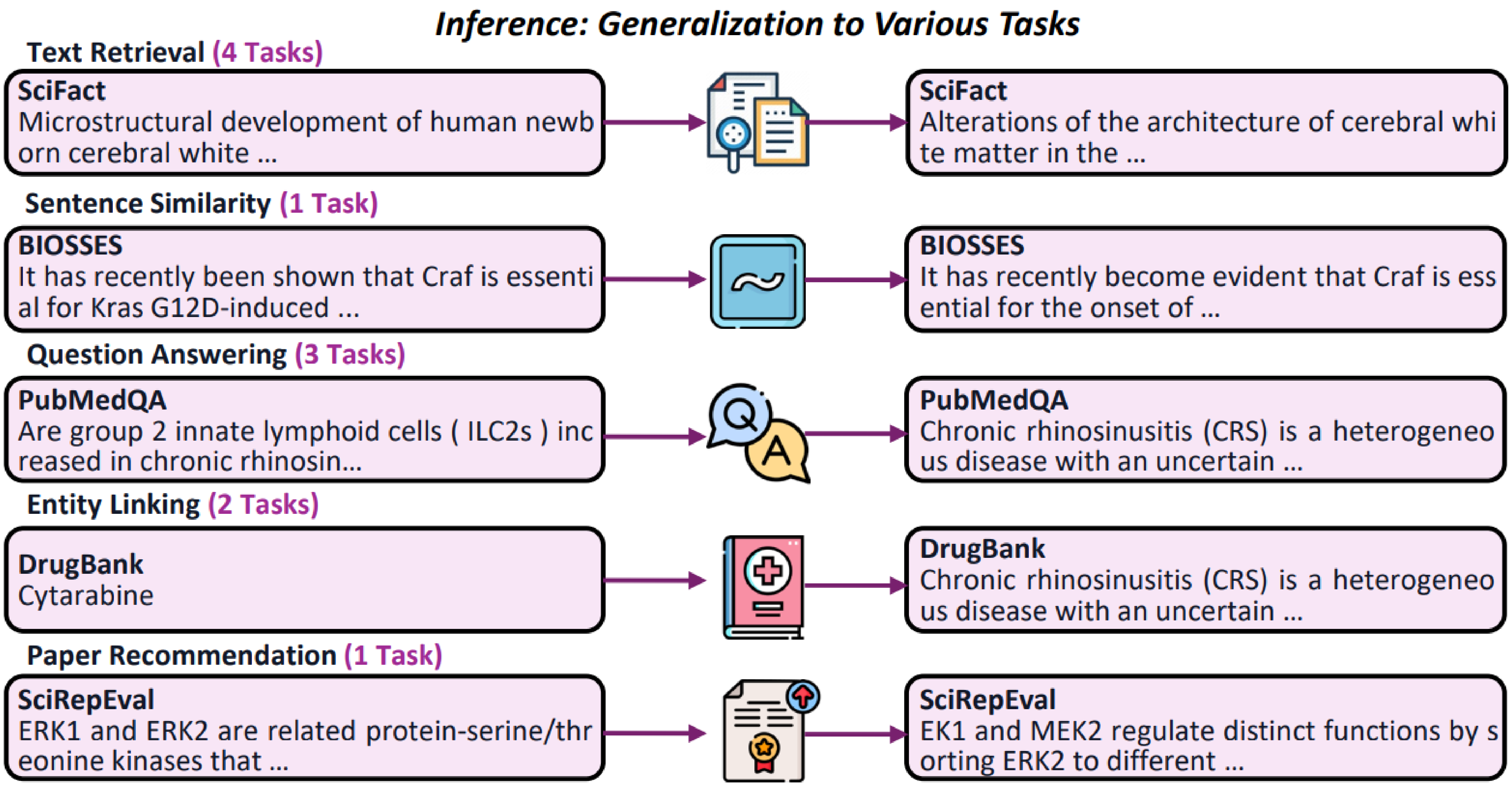
Getting Relevant (Text, Text) Pairs from Other Tasks

- Instruction Tuning



BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. EMNLP 2024.

Generalize to Unseen Retrieval Tasks



BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. EMNLP 2024.

Performance of BMRetriever: Paper Retrieval

Task / Model	Scale	# PT Pairs	# FT Pairs	Standard IR: NFCorpus	Standard IR: SciFact	Standard IR: SciDocs	Standard IR: Trec-COVID	Sent. Sim.: BIOSSES	Avg. Retr.	Avg. All
Sparse Retrieval										
BM25 (Robertson et al., 2009)	—	—	—	0.325	0.665	0.158	0.656	—	0.451	—
Base Size (< 1B)										
Contriever (Izacard et al., 2022)	110M	1B	500K	0.328	0.677	0.165	0.596	0.833	0.442	0.520
Dragon (Lin et al., 2023)	110M	—	28.5M	0.339	0.679	0.159	0.759	0.819	0.484	0.551
SPECTER 2.0 (Singh et al., 2023)	110M	3.3M	—	0.228	0.671	—	0.584	—	—	—
SciMult (Zhang et al., 2023)	110M	5.5M	—	0.308	0.707	—	0.712	—	—	—
COCO-DR (Yu et al., 2022)	110M	15M	500K	0.355	0.709	0.160	0.789	0.829	0.503	0.567
SGPT-125M (Muennighoff, 2022)	125M	unknown	500K	0.228	0.569	0.122	0.703	0.752	0.406	0.475
MedCPT (Jin et al., 2023)	220M	—	255M	0.340	0.724	0.123	0.697	0.837	0.471	0.544
GTR-L (Ni et al., 2022)	335M	2B	662K	0.329	0.639	0.158	0.557	0.849	0.421	0.506
InstructOR-L (Su et al., 2023)	335M	—	1.24M	0.341	0.643	0.186	0.581	0.844	0.438	0.519
E5-Large-v2† (Wang et al., 2022b)	335M	270M	1M	0.371	0.726	0.201	0.665	0.836	0.491	0.560
BGE-Large*‡ (Chen et al., 2024)	335M	1.2B	1.62M	0.345	0.723	0.222	0.753	0.804	**0.511**	<u>0.569</u>
BMRetriever-410M	410M	10M	1.4M	0.321	0.711	0.167	0.831	0.840	<u>0.508</u>	**0.574**
Large Size (1B - 5B)										
InstructOR-XL (Su et al., 2023)	1.5B	—	1.24M	0.360	0.646	0.174	0.713	0.842	0.473	0.547
GTR-XL (Ni et al., 2022)	1.2B	2B	662K	0.343	0.635	0.159	0.584	0.789	0.430	0.502
GTR-XXL (Ni et al., 2022)	4.8B	2B	662K	0.342	0.662	0.161	0.501	0.819	0.417	0.497
SGPT-1.3B (Muennighoff, 2022)	1.3B	unknown	500K	0.320	0.682	0.162	0.730	0.830	0.473	0.545
SGPT-2.7B (Muennighoff, 2022)	2.7B	unknown	500K	0.339	0.701	0.166	0.752	0.848	0.489	0.561
BMRetriever-1B	1B	10M	1.4M	0.344	0.760	0.180	0.840	0.858	<u>0.531</u>	<u>0.596</u>
BMRetriever-2B	2B	10M	1.4M	0.351	0.760	0.199	0.863	0.828	**0.543**	**0.600**
XL Size (> 5B)										
SGPT-5.8B (Muennighoff, 2022)	5.8B	unknown	500K	0.362	0.747	0.199	0.849	0.863	0.539	0.604
LLaRA (Li et al., 2023a)	7B	21M	500K	0.372	0.757	0.172	0.853	—	0.539	—
RepLLaMA (Ma et al., 2023)	7B	—	500K	0.378	0.756	0.181	0.847	—	0.541	—
LLM2Vec* (BehnamGhader et al., 2024)	7B	1.2M	1.5M	0.393	0.788	0.225	0.776	0.852	0.545	0.606
E5-Mistral* (Wang et al., 2024)	7B	—	1.8M	0.386	0.764	0.162	0.872	0.855	<u>0.546</u>	<u>0.608</u>
CPT-text-XL (Neelakantan et al., 2022)	175B	unknown	unknown	0.407	0.754	—	0.649	—	—	—
BMRetriever-7B	7B	10M	1.4M	0.364	0.778	0.201	0.861	0.847	**0.551**	**0.610**

Performance of BMRetriever: QA, Entity Linking & Recommendation

Task	Question Answering									Entity Linking						Paper Rec.	
Model	BioASQ			PubMedQA			iCliniq			DrugBank			MeSH			RELISH	
	R@5	R@20	nDCG@20	R@5	R@20	nDCG@20	R@5	R@20	nDCG@20	R@1	R@5	MRR@5	R@1	R@5	MRR@5	MAP	nDCG
Base Size (< 1B)																	
Dragon (2023)	36.2	**54.6**	49.1	71.8	74.0	72.0	50.6	65.2	47.4	81.0	87.6	83.3	28.2	47.0	34.8	72.6	80.6
MedCPT (2023)	34.7	54.4	45.2	66.3	71.1	60.4	26.8	42.0	24.9	75.1	88.0	80.6	27.7	54.2	37.4	83.6	89.7
E5-Large-v2† (2022b)	36.8	54.0	50.4	71.6	74.2	72.2	57.6	72.0	55.8	**81.8**	86.5	81.5	**32.8**	**55.0**	**41.3**	84.9	91.0
BMRetriever-410M	**39.9**	54.2	**53.1**	**73.8**	**74.6**	**72.4**	**60.6**	**72.8**	**56.6**	81.4	**88.2**	**83.7**	31.5	53.8	39.8	**85.2**	**91.2**
Large Size (1B - 5B)																	
InstructOR-XL (2023)	29.9	43.2	41.8	70.5	74.0	69.1	64.9	78.1	58.3	75.3	84.2	80.3	33.6	56.2	45.7	84.5	90.6
SGPT-2.7B (2022)	33.9	47.4	47.3	68.3	73.7	63.2	45.0	52.2	41.2	71.9	77.0	62.9	20.2	39.7	28.5	84.9	90.8
BMRetriever-1B	40.4	55.8	53.4	73.6	74.4	72.7	61.1	73.7	56.8	**84.7**	89.1	**86.5**	35.5	60.3	48.8	85.2	91.3
BMRetriever-2B	**42.5**	**56.5**	**55.7**	**74.0**	**74.6**	**73.1**	**70.0**	**81.2**	**65.7**	82.6	**90.2**	85.8	**45.6**	**71.3**	**59.5**	**85.4**	**91.5**
XL Size (> 5B)																	
E5-Mistral* (2024)	39.6	55.4	52.7	72.6	74.2	70.0	56.7	72.2	51.8	78.5	92.2	84.0	47.9	76.2	**61.3**	85.2	90.8
BMRetriever-7B	**43.7**	**60.2**	**57.4**	**74.2**	**74.6**	**73.8**	**68.4**	**79.7**	**63.7**	**84.7**	**92.8**	**88.0**	**49.8**	**76.5**	61.1	**86.7**	**92.2**

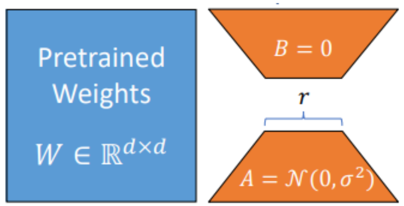
Take-Away Messages

- If you cannot access proprietary search logs but still need to train a retrieval model, get relevant (text, text) pairs from:
  - Other tasks (e.g., classification, citation prediction, question answering)
  - Different paragraphs in one document
  - LLMs
- Directly merging all these data together for contrastive learning suffers from task interference. Solutions include:
  - Mixture-of-Experts Transformers
  - Instruction Tuning

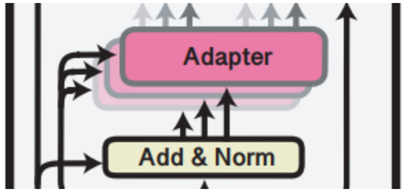
A Summary of Advanced Techniques Introduced in Recent Lectures

Parameter-Efficient Fine-Tuning (PEFT): Only tune a small number of parameters in LLMs

LoRA [ICLR 2022]

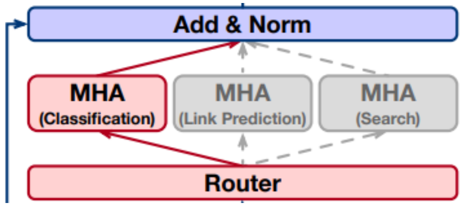


Adapter [EACL 2021]

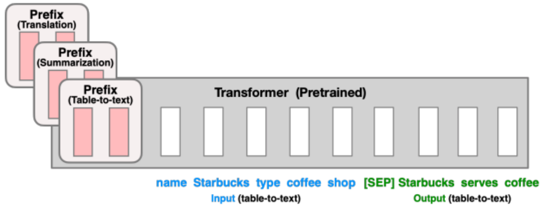


What if we tune the entire model?

Mixture-of-Experts [ICML 2022]

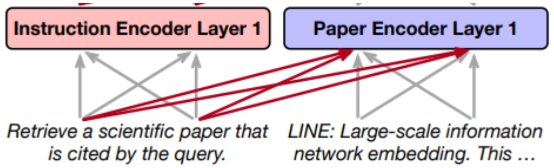


Prefix Tuning [ACL 2021]



What if we tune the entire model?

Instruction Tuning [ICLR 2022]

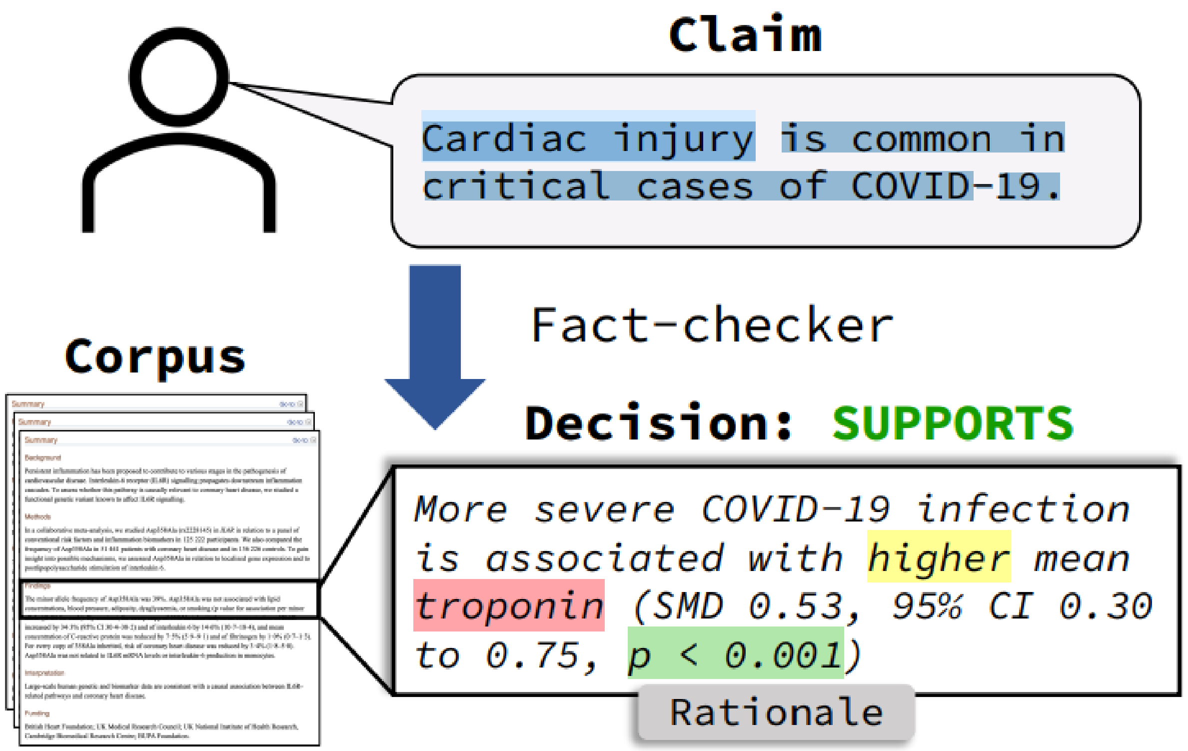


Agenda

- Contrastive Learning with Ground-Truth Search Logs
  - MedCPT: Bi-Encoder → Cross-Encoder
- Contrastive Learning with Data from Other Tasks
  - SciMult: Mixture-of-Experts Transformer
  - BMRetriever: Instruction Tuning
- Application
  - SciFact: Scientific Claim Verification

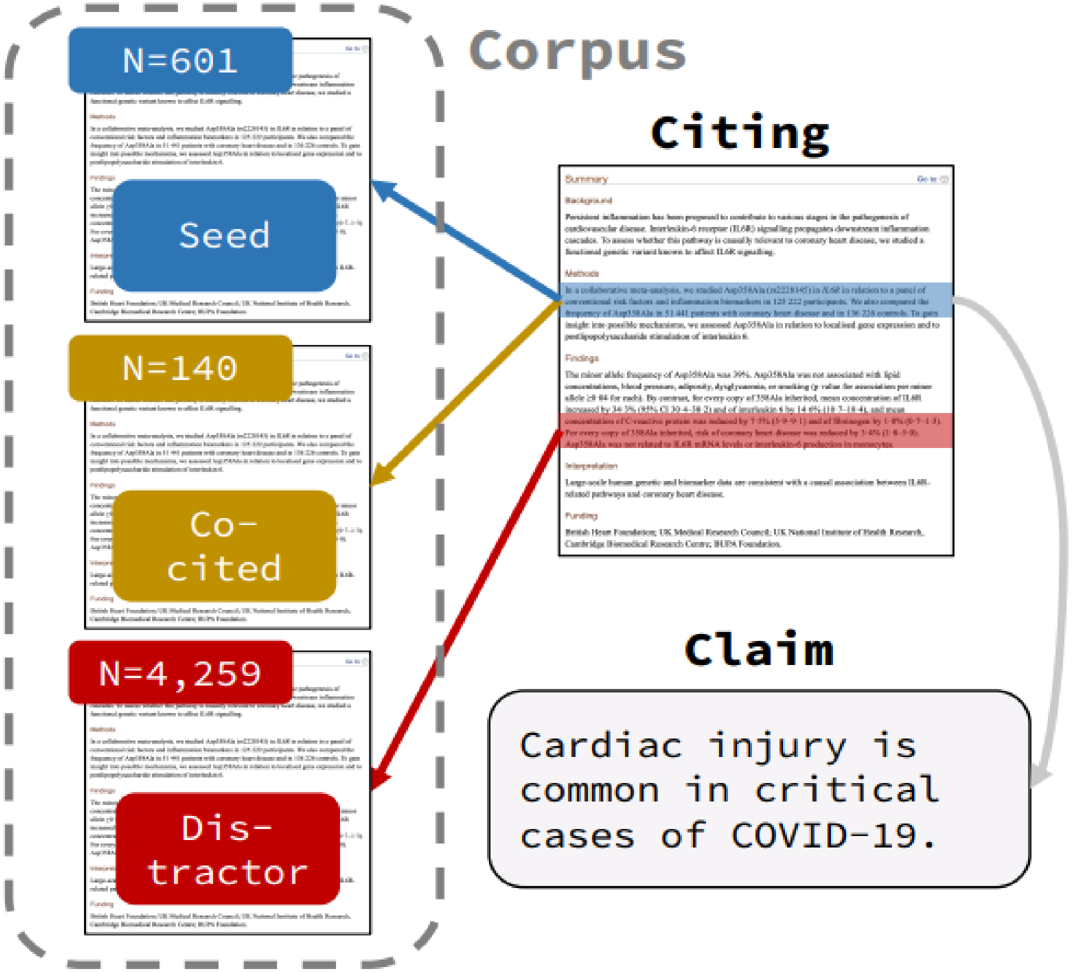
Scientific Claim Verification

- Given a scientific claim:
  - Step 1 (Relevant Paper Retrieval): Find all papers relevant to this claim.
  - Step 2 (Rationale Sentence Selection): In each relevant paper, find relevant sentences.
  - Step 3 (Stance Prediction): Based on the relevant sentences, predict if the paper supports, refutes, or is neutral towards the claim.



Fact or Fiction: Verifying Scientific Claims. EMNLP 2020.

Dataset Construction



- Data source: S2ORC
- Annotators write claims based on citation sentences.

Source citance

*"Future studies are also warranted to evaluate the potential association between WNT5A/PCP signaling in adipose tissue and atherosclerotic CVD, given the major role that IL-6 signaling plays in this condition as revealed by large Mendelian randomization studies **44, 45** ."*

Claim

IL-6 signaling plays a major role in atherosclerotic cardiovascular disease.

Framework

- Task 1 (Relevant Paper Retrieval)
  - Any retrieval model
- Task 2 (Rationale Sentence Selection)
  - For each sentence s in a relevant paper, perform binary classification (rationale sentence / not rationale sentence)
  - [CLS] claim [SEP] s [SEP]
- Task 3 (Stance Prediction)
  - Combine all rationale sentences together and perform three-class classification (support/refute/neutral)
  - [CLS] claim [SEP] rationale1 rationale2 … rationale [SEP]

Performance of Each Task

	Rational-Select.			Label-Pred.
Training data	P	R	F1	Acc.
Fever	41.5	57.9	48.4	67.6
UKP Snopes	42.5	62.3	50.5	71.3
SciFact	73.7	70.5	**72.1**	75.7
Fever + SciFact	72.4	67.2	69.7	**81.9**
Sentence encoder	P	R	F1	Acc.
SciBERT	74.5	74.3	**74.4**	69.2
BioMedRoBERTa	75.3	69.9	72.5	71.7
RoBERTa-base	76.1	66.1	70.8	62.9
RoBERTa-large	73.7	70.5	72.1	**75.7**
Model inputs	P	R	F1	Acc.
Claim-only	-	-	-	44.5
Abstract-only	60.1	60.9	60.5	53.3

End-to-End Performance

			Sentence-level						Abstract-level					
			Selection-Only			Selection+Label			Label-Only			Label+Rationale		
Retrieval	Model		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Oracle abstract	Oracle rationale	1	100.0	80.5	$89.2_{2.1}$	89.6	72.2	$79.9_{3.0}$	90.1	77.5	$83.3_{2.4}$	90.1	77.5	$83.3_{2.4}$
	Zero-shot	2	42.5	45.1	$43.8_{2.0}$	36.1	38.4	$37.2_{2.3}$	86.9	53.6	$66.3_{3.1}$	67.9	41.9	$51.8_{3.4}$
	VeriSci	3	76.1	63.8	$69.4_{2.6}$	66.5	55.7	$\mathbf{60.6}_{3.1}$	87.3	65.3	$74.7_{2.8}$	84.9	63.5	$\mathbf{72.7}_{2.9}$
Open	Oracle rationale	4	100.0	56.5	$72.2_{3.3}$	87.6	49.5	$63.2_{3.7}$	88.9	54.1	$67.2_{3.2}$	88.9	54.1	$67.2_{3.2}$
	Zero-shot	5	28.7	37.6	$32.5_{2.3}$	23.7	31.1	$26.9_{2.3}$	56.0	42.3	$48.2_{3.3}$	42.3	32.0	$36.4_{3.3}$
	VeriSci	6	45.0	47.3	$46.1_{3.0}$	38.6	40.5	$\mathbf{39.5}_{3.0}$	47.5	47.3	$47.4_{3.1}$	46.6	46.4	$\mathbf{46.5}_{3.1}$

Case Studies

Claim 1: Lopinavir / ritonavir have exhibited favorable clinical responses when used as a treatment for coronavirus.
Supports: … *Interestingly, after lopinavir/ritonavir (Kaletra, AbbVie) was administered, β-coronavirus viral loads significantly decreased and no or little coronavirus titers were observed.*
Refutes: *The focused drug repurposing of known approved drugs (such as lopinavir/ritonavir) has been reported failed for curing SARS-CoV-2 infected patients.* It is urgent to generate new chemical entities against this virus …

Claim 2: The coronavirus cannot thrive in warmer climates.
Supports: *...most outbreaks display a pattern of clustering in relatively cool and dry areas...This is because the environment can mediate human-to-human transmission of SARS-CoV-2, and unsuitable climates can cause the virus to destabilize quickly...*
Refutes: *...significant cases in the coming months are likely to occur in more humid (warmer) climates, irrespective of the climate-dependence of transmission and that summer temperatures will not substrantially limit pandemic growth*.

Take-Away Messages

- The ideas and techniques used in scientific paper retrieval can be generalized to a wide spectrum of scientific text mining tasks aiming to predict the semantic similarity between two text units, including different steps in scientific claim verification.

- Drawback
    - Not an end-to-end framework. Errors in rationale selection will propagate to stance prediction. (If you miss some rationale sentences, you loss some information in stance prediction.)
    - Can we merge these two steps? (e.g., [CLS] claim [SEP] entire paper [SEP])
    - *MultiVerS: Improving Scientific Claim Verification with Weak Supervision and Full-Document Context*. NAACL 2022 Findings.

Thank You!

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