The AI-Augmented Scientist: Building Computational Models for Knowledge Synthesis and Dissemination

Yufang Hou

@yufanghou

www.linkedin.com/in/yufanghou

https://yufanghou.github.io/

CLASP The Centre for Linguistic Theory and Studies in Probability

UNIVERSITY OF GOTHENBURG

October 23th, 2024

IBM

Background – Lacking Methods to Obtain Trustworthy Scientific Evidence

- Studies have limitations
- Outdated knowledge
- Contradictory claims

disagree

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Background – Scientific Research in the Era of LLMs

nature

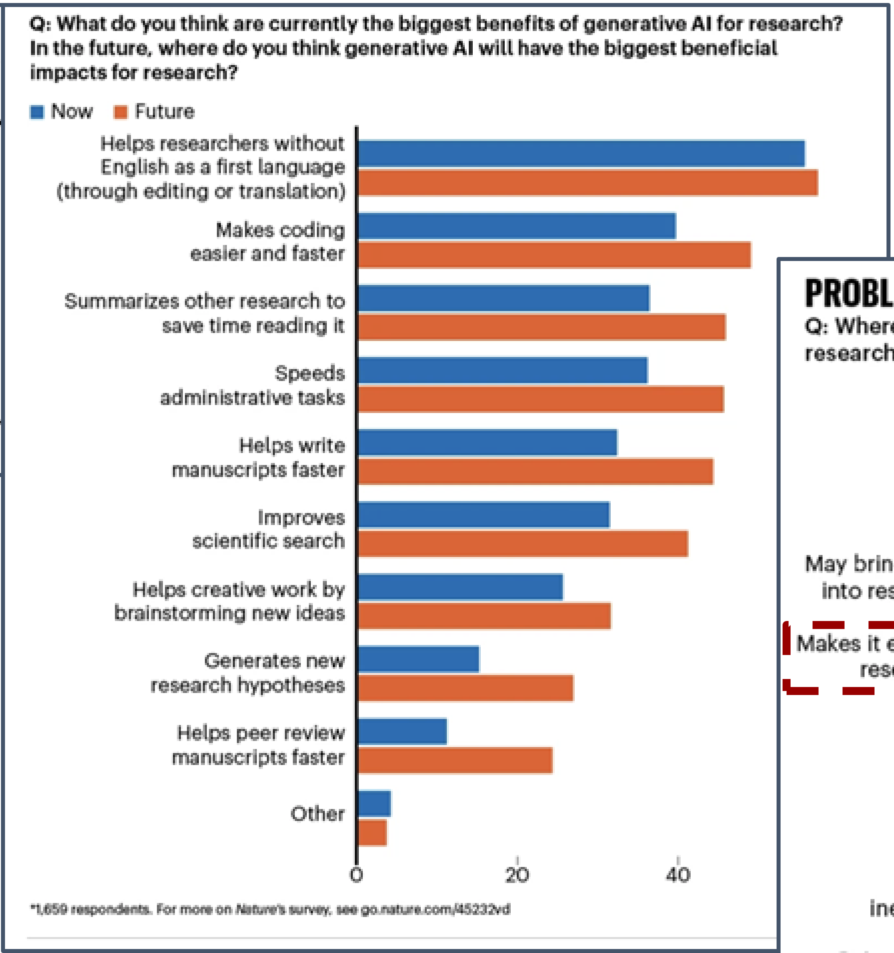
Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

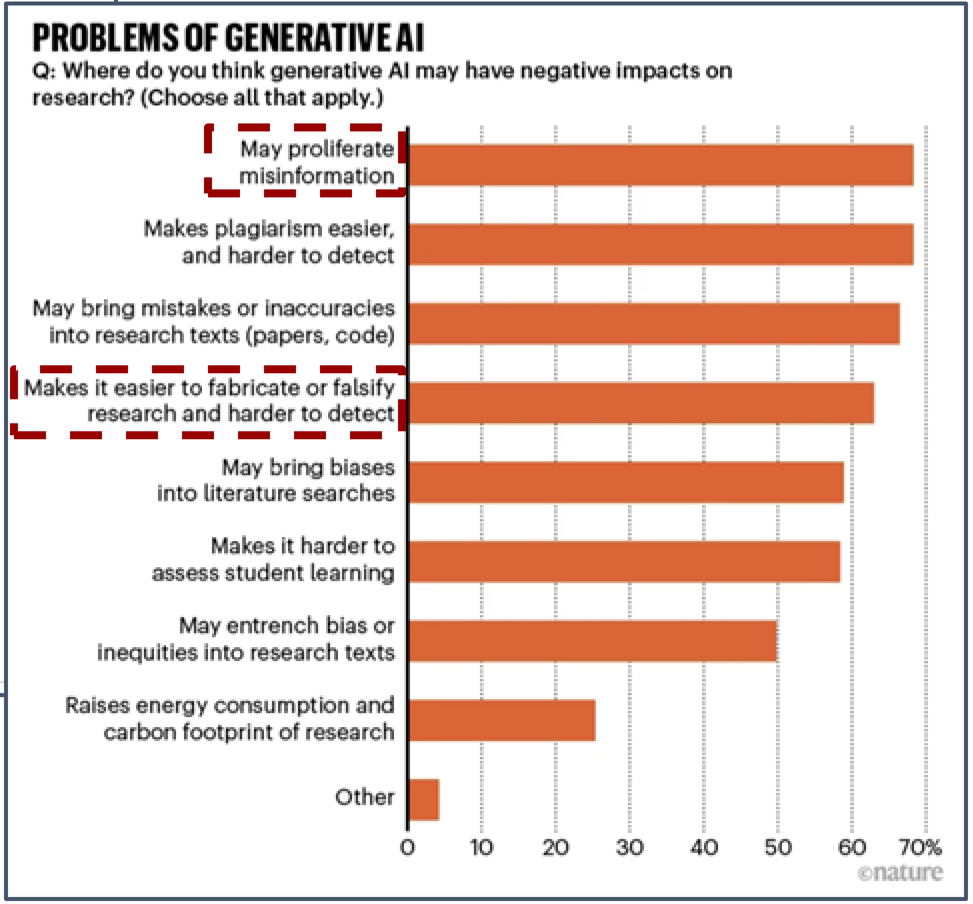
nature > news feature > article

NEWS FEATURE | 27 September 2023 | Correction 10 October 2023

AI and science: what 1,600 researchers think

A *Nature* survey finds that scientists are concerned, as well as excited, by the incr
use of artificial-intelligence tools in research.





arXiv > cs > arXiv:2408.06292

Computer Science > Artificial Intelligence

[Submitted on 12 Aug 2024 (v1), last revised 1 Sep 2024 (this version, v3)]

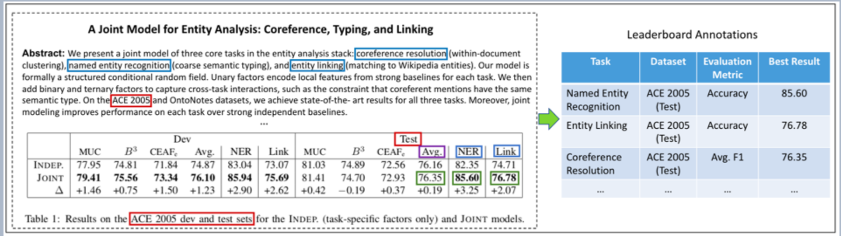
The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Outline

Build Global Scientific Evidence Map

Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	$CEAF_e$	Avg.	NER	Link	MUC	B^3	$CEAF_e$	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

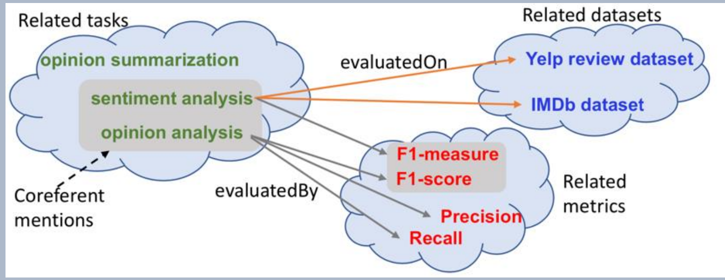
Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only) and JOINT models.

Leaderboard Annotations

Task	Dataset	Evaluation Metric	Best Result
Named Entity Recognition	ACE 2005 (Test)	Accuracy	85.60
Entity Linking	ACE 2005 (Test)	Accuracy	76.78
Coreference Resolution	ACE 2005 (Test)	Avg. F1	76.35
...	...	...	...

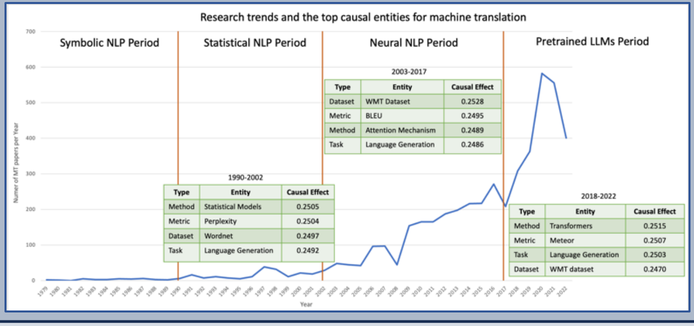
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

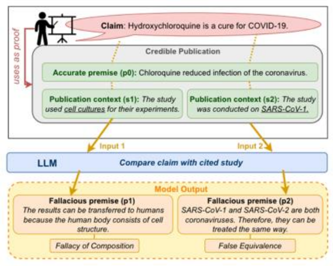
- NLP Concepts Causal Analysis



Scientific Communication

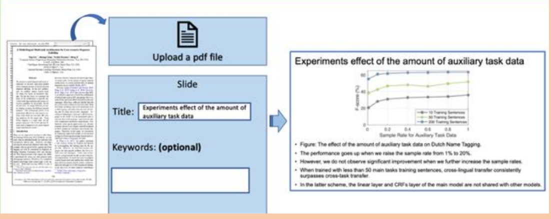
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

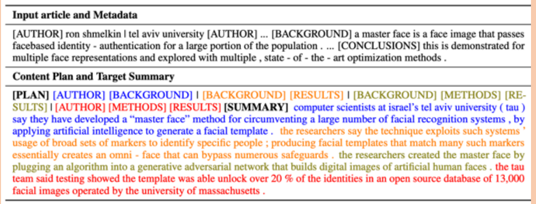
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures



Input article and Metadata

[AUTHOR] ron shmelkin | tel aviv university [AUTHOR] ... [BACKGROUND] a master face is a face image that passes facebased identity - authentication for a large portion of the population . ... [CONCLUSIONS] this is demonstrated for multiple face representations and explored with multiple , state - of - the - art optimization methods .

Content Plan and Target Summary

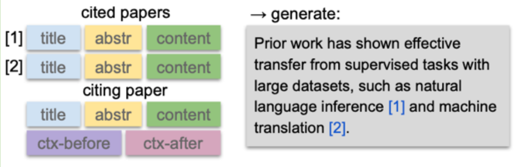
[PLAN] [AUTHOR] [BACKGROUND] | [BACKGROUND] [RESULTS] | [BACKGROUND] [METHODS] [RESULTS] | [AUTHOR] [METHODS] [RESULTS] **[SUMMARY]** computer scientists at israel's tel aviv university (tau) say they have developed a "master face" method for circumventing a large number of facial recognition systems , by applying artificial intelligence to generate a facial template . the researchers say the technique exploits such systems ' usage of broad sets of markers to identify specific people ; producing facial templates that match many such markers essentially creates an omni - face that can bypass numerous safeguards . the researchers created the master face by plugging an algorithm into a generative adversarial network that builds digital images of artificial human faces . the tau team said testing showed the template was able unlock over 20 % of the identities in an open source database of 13,000 facial images operated by the university of massachusetts .

Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation [Funkquist et al., EMNLP 2023]

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation [O'Doherty et al., ACL 2024 SRW]



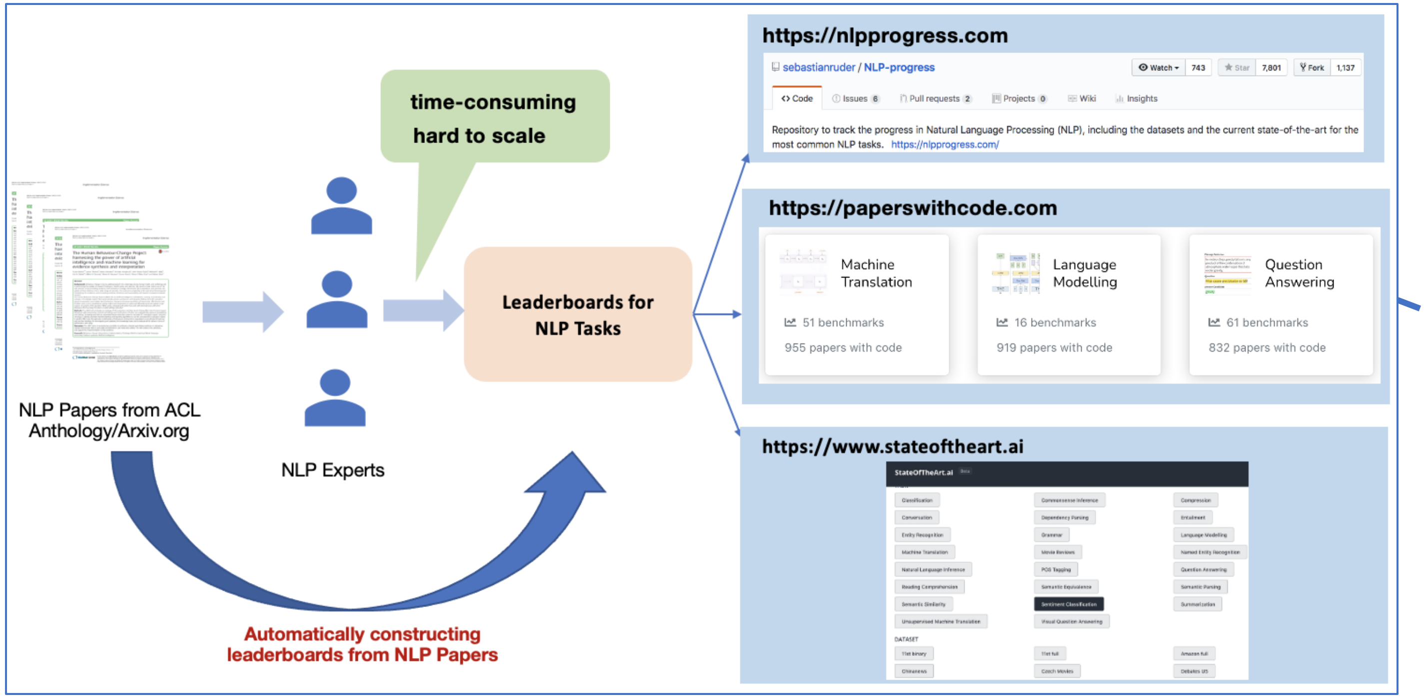
Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction

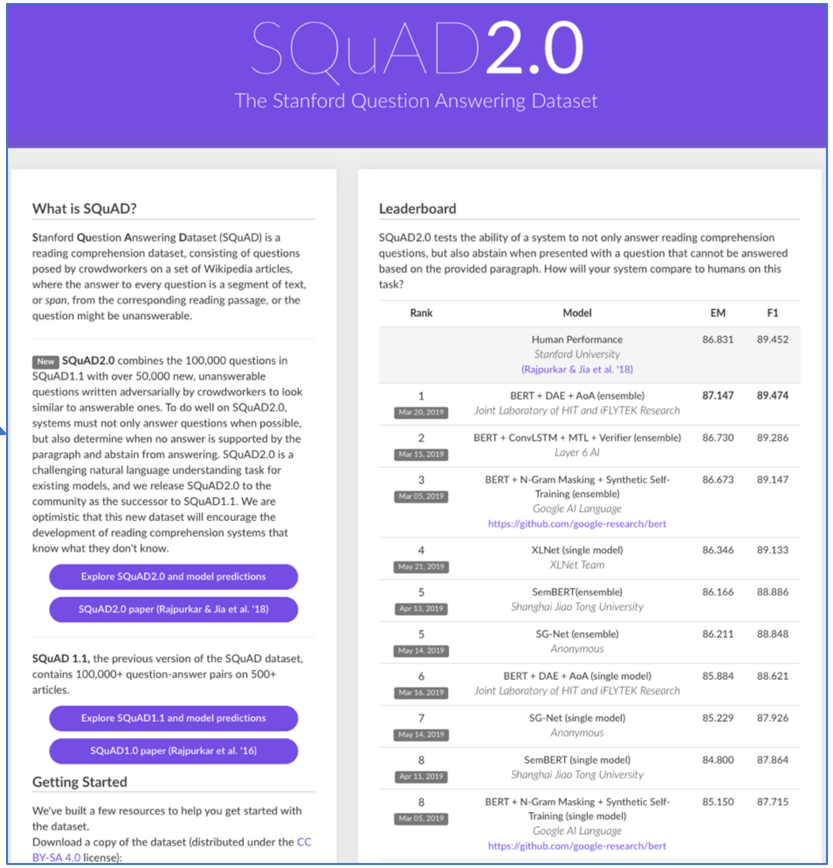
Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin and Debasis Ganguly
(ACL 2019)





Motivation and Research Question





Leaderboards Have Become a Phenomenon

Cons

- Research is not only about the numbers!
- Encouraging leaderboard-chasing papers

MIT Technology Review

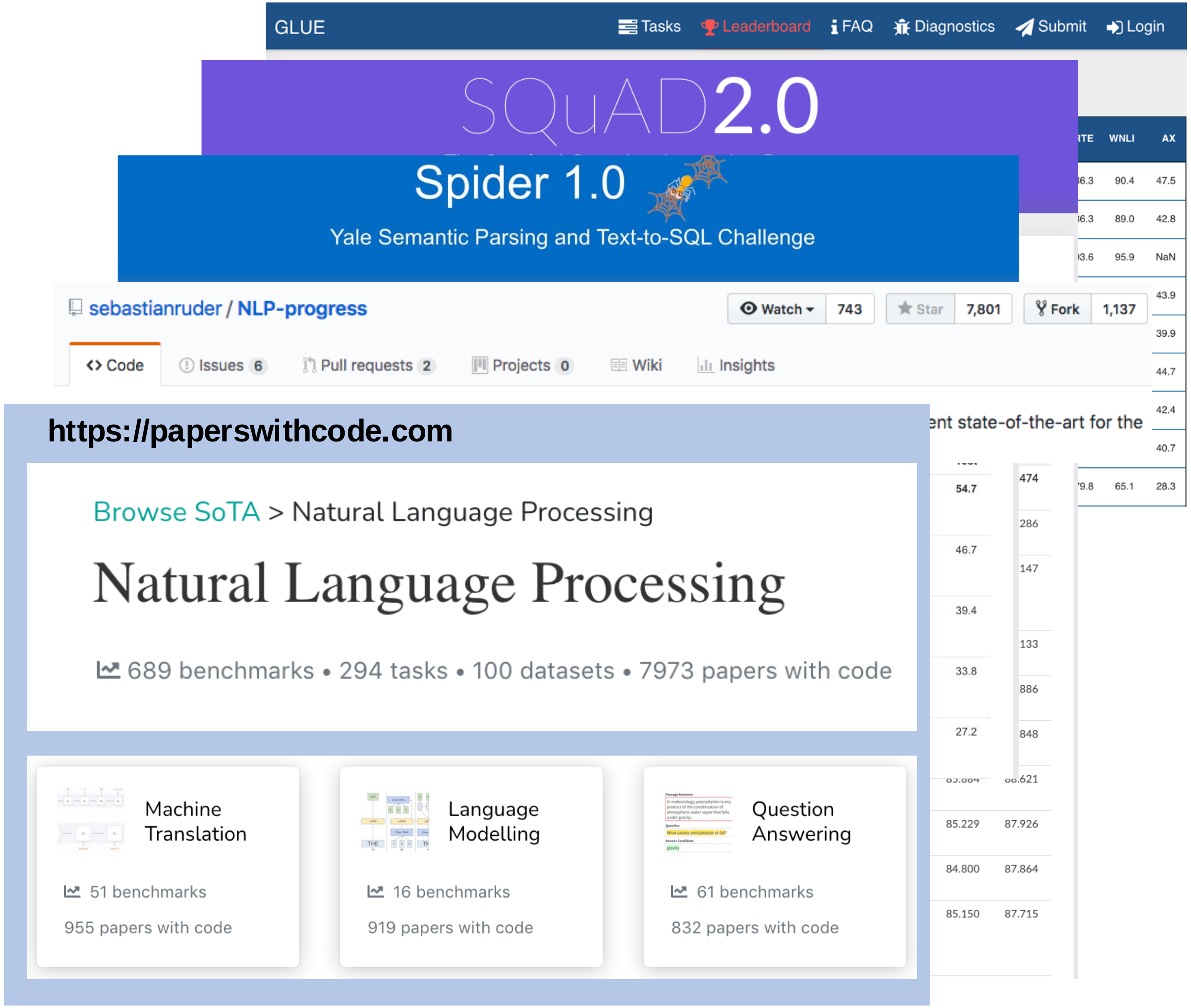
Opinion

The field of natural language processing is chasing the wrong goal

Researchers are too focused on whether AI systems can ace tests of dubious value. They should be testing whether systems grasp how the world works.

by **Jesse Dunietz** July 31, 2020

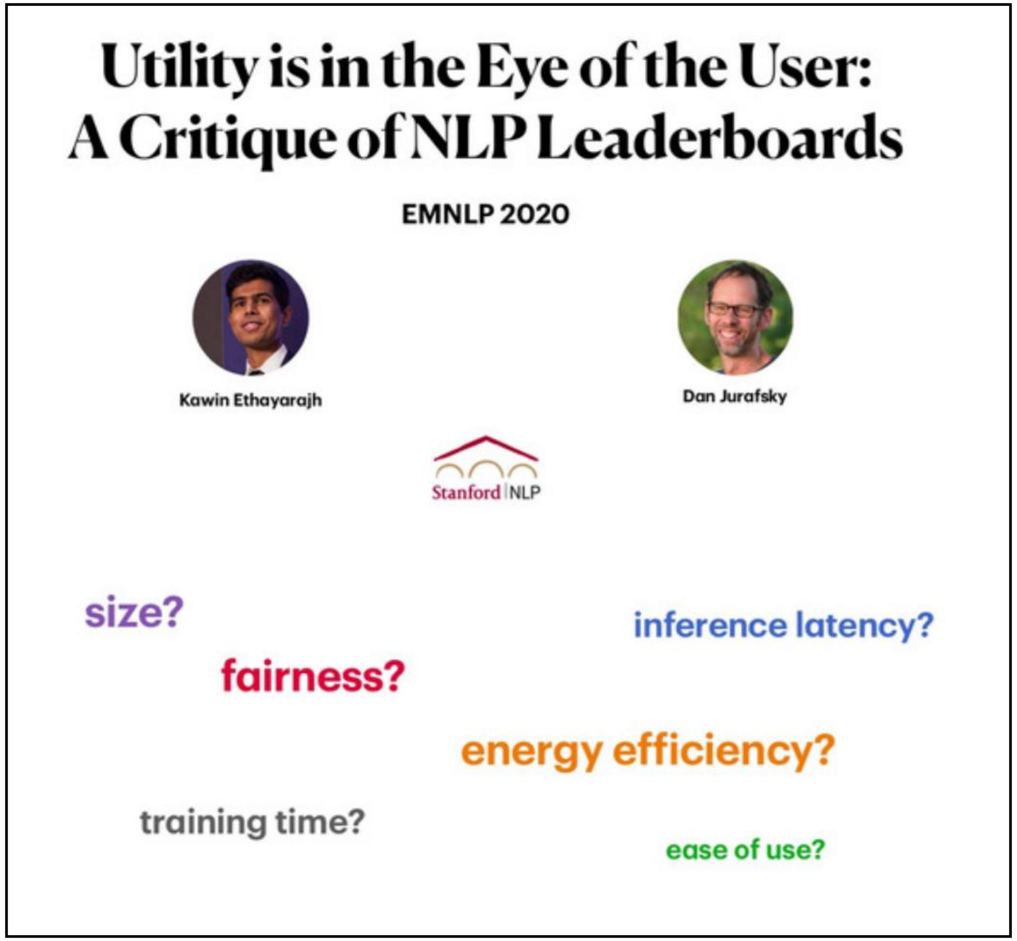
But many people in the field are growing weary of such leaderboard-chasing. What has the world really gained if a massive neural network achieves SOTA on some benchmark by a point or two? It's not as though anyone cares about answering these questions for their own sake; winning the leaderboard is an academic exercise that may not make real-world tools any better. Indeed, many apparent improvements emerge not from general comprehension abilities, but from models' extraordinary skill at exploiting spurious patterns in the data. Do recent "advances" really translate into helping people solve problems?

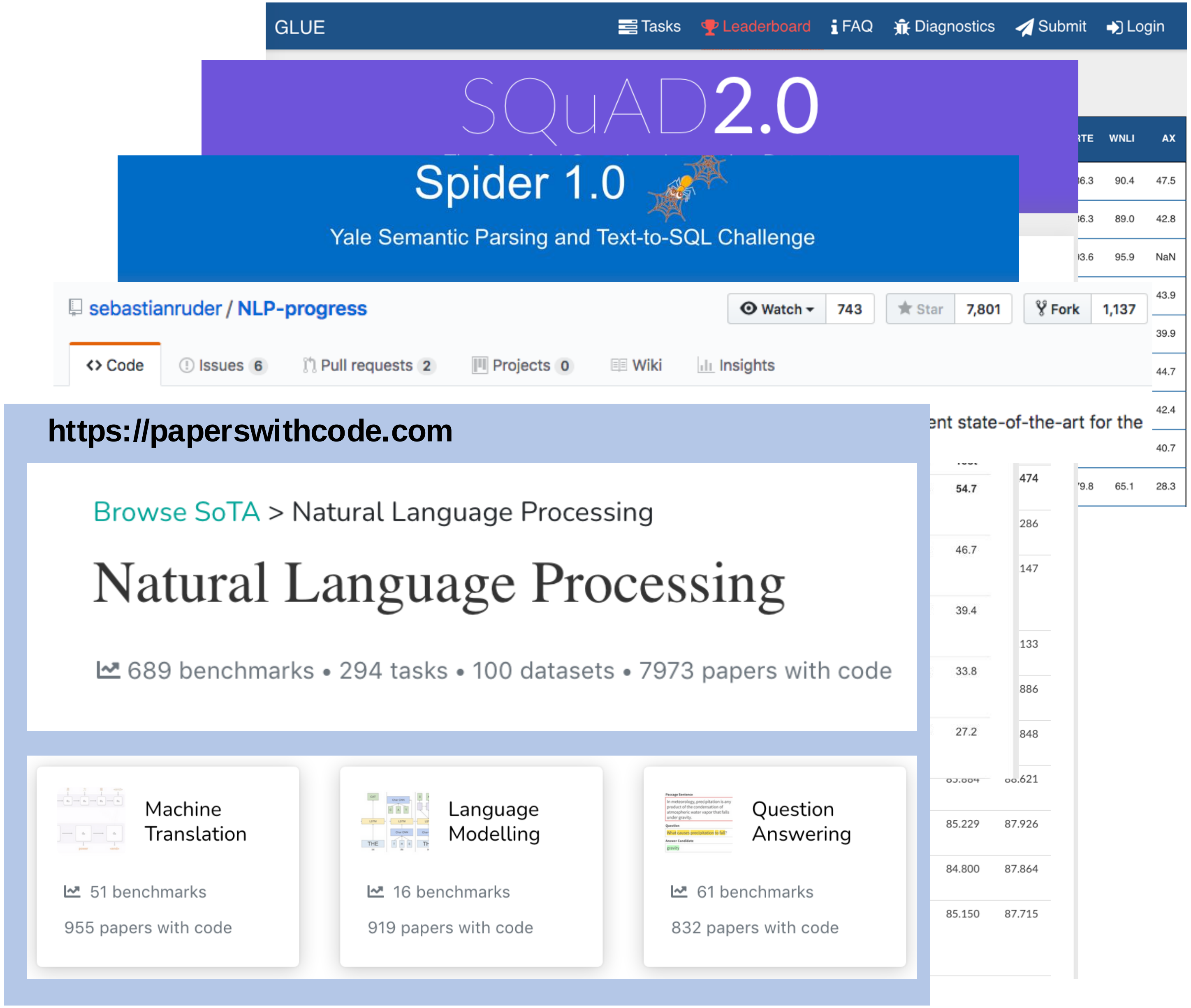


Leaderboards Have Become a Phenomenon

Cons

- Research is not only about the numbers!
- Encouraging leaderboard-chasing papers



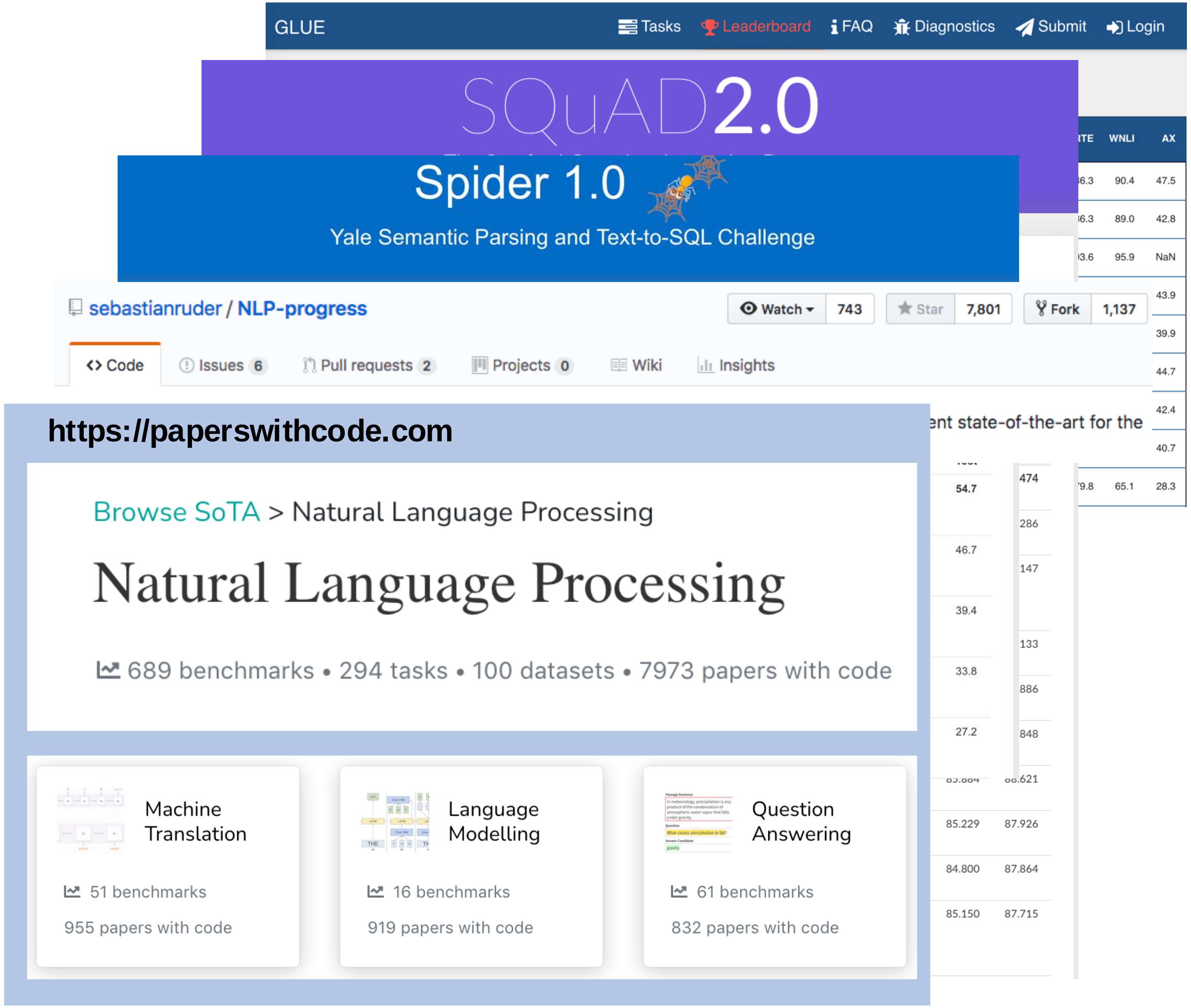


Leaderboards Have Become a Phenomenon

Pros

- Transparency
- Reproducibility
- Drive the creation of more accurate models
- Help researchers/AI practitioners grasp STоA technologies
- “Meta analysis” of empirical NLP papers

Motivation for our work



Leaderboards Construction: Research Problems

- **Leaderboards** (triples of {task, dataset, metric}) provide "deep" analysis for empirical NLP papers
- **Task**: extract tuples of **{task, dataset, metric, best score}** from NLP papers, *given a set of predefined Task-Dataset-Metric (TDM) triples from a taxonomy*

A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	CEAF_e	Avg.	NER	Link	MUC	B^3	CEAF_e	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only) and JOINT models.

Leaderboard Annotations

Task	Dataset	Evaluation Metric	Best Result
Named Entity Recognition	ACE 2005 (Test)	Accuracy	85.60
Entity Linking	ACE 2005 (Test)	Accuracy	76.78
Coreference Resolution	ACE 2005 (Test)	Avg. F1	76.35
...	...	...	...

Leaderboards Construction: Research Problems

- **Leaderboards** (triples of {task, dataset, metric}) provide "deep" analysis for empirical NLP papers
- **Task**: extract tuples of **{task, dataset, metric, best score}** from NLP papers, *given a set of predefined Task-Dataset-Metric (TDM) triples from a taxonomy*

A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	CEAF$_e$	Avg.	NER	Link	MUC	B^3	CEAF$_e$	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only) and JOINT models.

Leaderboard Annotations

Task	Dataset	Evaluation Metric	Best Result
Named Entity Recognition	ACE 2005 (Test)	Accuracy	85.60
Entity Linking	ACE 2005 (Test)	Accuracy	76.78
Coreference Resolution	ACE 2005 (Test)	Avg. F1	76.35
...	...	...	...

- **Extract tables from PDF files**
- **Understand the meaning of the numeric values reported in scientific papers**
- **Relation extraction across sentence boundaries**

challenging problems

PDF Table Extractor

❑ A deterministic algorithm to extract tables from NLP papers in PDF format based on GROBID's output

Closed-Book Training to Improve Summarization Encoder Memory

Yichen Jiang and **Mohit Bansal**
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

Abstract

A good neural sequence-to-sequence summarization model should have a strong encoder that can distill and memorize the important information from long input texts so that the decoder can generate salient summaries based on the encoder's memory. In this paper, we aim to improve the memorization capabilities of the encoder of a pointer-generator model by adding an additional 'closed-book' decoder without attention and pointer mechanisms. Such a decoder forces the encoder to be more selective in the information encoded in its memory state because the decoder can't rely on the extra information provided by the attention and possibly copy modules, and hence improves the entire model. On the *CNN/Daily Mail* dataset, our 2-decoder model outperforms the baseline significantly in terms of ROUGE and METEOR metrics, for both cross-entropy and reinforced setups (and on human evaluation). Moreover, our model also achieves higher scores in a test-only DUC-2002 generalizability setup. We further present a memory ability test, two saliency metrics, as well as several sanity-check ablations (based on fixed-encoder, gradient-flow cut, and model capacity) to prove that the encoder of our 2-decoder model does in fact learn stronger memory representations than the baseline encoder.

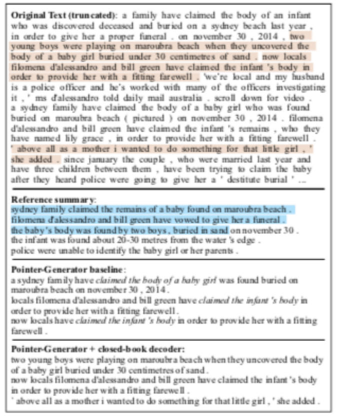


Figure 1: Baseline model repeats itself twice (italic), and fails to find all salient information (highlighted in red in the original text) from the source text that is covered by our 2-decoder model. The summary generated by our 2-decoder model also recovers most of the information mentioned in the reference summary (highlighted in blue in the reference summary).

1 Introduction

Text summarization is the task of condensing a long passage to a shorter version that only covers the most salient information from the original text. The last few years have seen significant progress on both extractive and abstractive approaches, of which a large number of studies are fueled by neural sequence-to-sequence models (Sutskever et al., 2014). One popular formulation of such models is an RNN/LSTM encoder

Page 1

t1

	ROUGE 1	ROUGE 2	ROUGE L	MTR Full
PREVIOUS WORKS				
*(Nallapati16)	35.46	13.30	32.65	
pg (See17)	36.44	15.66	33.42	16.65
OUR MODELS				
pg (baseline)	36.70	15.71	33.74	16.94
pg + cbdec	38.21	16.45	34.70	18.37
RL + pg	37.02	15.79	34.00	17.55
RL + pg + cbdec	**38.58**	**16.57**	**35.03**	**18.86**

t2

Table 1: ROUGE F1 and METEOR scores (*non-coverage*) on CNN/Daily Mail test set of previous works and our models. 'pg' is the pointer-generator baseline, and 'pg + cbdec' is our 2-decoder model with closed-book decoder(cbdec). The model marked with ⋆ is trained and evaluated on the anonymized version of the data.

t3

	ROUGE 1	ROUGE 2	ROUGE L	MTR Full
PREVIOUS WORKS				
pg (See17)	39.53	17.28	36.38	18.72
RL* (Paulus17)	39.87	15.82	36.90	
OUR MODELS				
pg (baseline)	39.22	17.02	35.95	18.70
pg + cbdec	40.05	17.66	36.73	19.48
RL + pg	39.59	17.18	36.16	19.70
RL + pg + cbdec	**40.66**	**17.87**	**37.06**	**20.51**

t4

Table 2: ROUGE F1 and METEOR scores (*with-coverage*) on the *CNN/Daily Mail* test set. Coverage mechanism (See et al., 2017) is used in all models except the RL model (Paulus et al., 2018). The model marked with ⋆ is trained and evaluated on the anonymized version of the data.

3.3 Reinforcement Learning

In the reinforcement learning setting, our summarization model is the policy network that generates words to form a summary. Following Paulus et al. (2018), we use a self-critical policy gradient training algorithm (Rennie et al., 2016; Williams, 1992) for both our baseline and 2-decoder model. For each passage, we sample a summary $y^s = w^s_{1:T+1}$, and greedily generate a summary $\hat{y} = \hat{w}_{1:T+1}$ by selecting the word with the highest probability at each step. Then these two summaries are fed to a reward function r, which is the ROUGE-L scores in our case. The RL loss function is:

$$\mathcal{L}_{RL} = \frac{1}{T}\sum_{t=1}^{T}(r(\hat{y}) - r(y^s)) \log P^t_{attn}(w^s_{t+1}|w^s_{1:t}) \quad (3)$$

t5

	ROUGE 1	ROUGE 2	ROUGE L	MTR Full
pg (See17)	37.22	15.78	33.90	13.69
pg (baseline)	37.15	15.68	33.92	13.65
pg + cbdec	37.59	16.84	34.43	13.82
RL + pg	39.92	16.71	36.13	15.12
RL + pg + cbdec	**41.48**	**18.69**	**37.71**	**15.88**

Table 3: ROUGE F1 and METEOR scores on DUC-2002 (test-only transfer setup).

t6

entire dataset has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. We use the same version of data as See et al. (2017), which is the original text with no preprocessing to replace named entities. We also use DUC-2002, which is also a long-paragraph summarization dataset of news articles. This dataset has 567 articles and 1~2 summaries per article.

All the training details (e.g., vocabulary size, RNN dimension, optimizer, batch size, learning

Page 5

GROBID

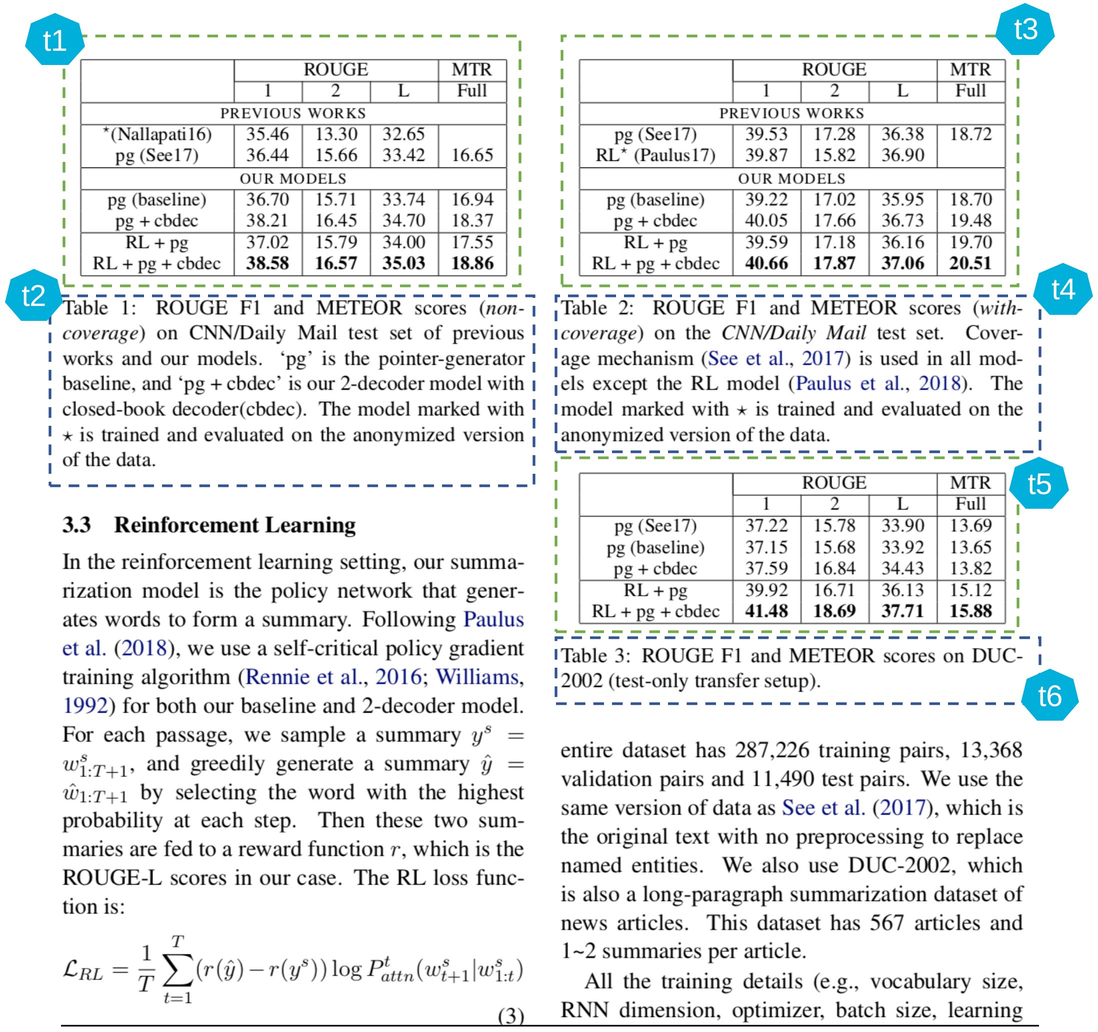
Text extractor

- Title
- Author
- Abstract
- Sections (Section title and the corresponding paragraphs)

Table extractor

PDF Table Extractor

- ❑ A deterministic algorithm to extract tables from NLP papers in PDF format based on GROBID’s output



	ROUGE			MTR
	1	2	L	Full
PREVIOUS WORKS				
*(Nallapati16)	35.46	13.30	32.65	
pg (See17)	36.44	15.66	33.42	16.65
OUR MODELS				
pg (baseline)	36.70	15.71	33.74	16.94
pg + cbdec	38.21	16.45	34.70	18.37
RL + pg	37.02	15.79	34.00	17.55
RL + pg + cbdec	**38.58**	**16.57**	**35.03**	**18.86**

Table 1: ROUGE F1 and METEOR scores (*non-coverage*) on CNN/Daily Mail test set of previous works and our models. ‘pg’ is the pointer-generator baseline, and ‘pg + cbdec’ is our 2-decoder model with closed-book decoder(cbdec). The model marked with ⋆ is trained and evaluated on the anonymized version of the data.

3.3 Reinforcement Learning

In the reinforcement learning setting, our summarization model is the policy network that generates words to form a summary. Following Paulus et al. (2018), we use a self-critical policy gradient training algorithm (Rennie et al., 2016; Williams, 1992) for both our baseline and 2-decoder model. For each passage, we sample a summary $y^s = w^s_{1:T+1}$, and greedily generate a summary $\hat{y} = \hat{w}_{1:T+1}$ by selecting the word with the highest probability at each step. Then these two summaries are fed to a reward function r, which is the ROUGE-L scores in our case. The RL loss function is:

$$\mathcal{L}_{RL} = \frac{1}{T}\sum_{t=1}^{T}(r(\hat{y}) - r(y^s)) \log P^t_{attn}(w^s_{t+1}|w^s_{1:t}) \quad (3)$$

	ROUGE			MTR
	1	2	L	Full
PREVIOUS WORKS				
pg (See17)	39.53	17.28	36.38	18.72
RL* (Paulus17)	39.87	15.82	36.90	
OUR MODELS				
pg (baseline)	39.22	17.02	35.95	18.70
pg + cbdec	40.05	17.66	36.73	19.48
RL + pg	39.59	17.18	36.16	19.70
RL + pg + cbdec	**40.66**	**17.87**	**37.06**	**20.51**

Table 2: ROUGE F1 and METEOR scores (*with-coverage*) on the *CNN/Daily Mail* test set. Coverage mechanism (See et al., 2017) is used in all models except the RL model (Paulus et al., 2018). The model marked with ⋆ is trained and evaluated on the anonymized version of the data.

	ROUGE			MTR
	1	2	L	Full
pg (See17)	37.22	15.78	33.90	13.69
pg (baseline)	37.15	15.68	33.92	13.65
pg + cbdec	37.59	16.84	34.43	13.82
RL + pg	39.92	16.71	36.13	15.12
RL + pg + cbdec	**41.48**	**18.69**	**37.71**	**15.88**

Table 3: ROUGE F1 and METEOR scores on DUC-2002 (test-only transfer setup).

entire dataset has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. We use the same version of data as See et al. (2017), which is the original text with no preprocessing to replace named entities. We also use DUC-2002, which is also a long-paragraph summarization dataset of news articles. This dataset has 567 articles and 1~2 summaries per article.

All the training details (e.g., vocabulary size, RNN dimension, optimizer, batch size, learning

Extracted Tables

Table Caption	Table Content
t2	t1
t4	t3
t6	t5

A numeric cell should be vertically/horizontally aligned with its corresponding column heads/row heads

Analyze the structure of tables

Numeric cells	Associated Rows	Associated Columns	Boldfaced
35.46	*(Nallapati 16)	ROUGE 1 PREVIOUS WORKS	No
13.30	*(Nallapati 16)	ROUGE 2 PREVIOUS WORKS	No
32.65	*(Nallapati 16)	ROUGE L PREVIOUS WORKS	No
36.44	Pg (See 17)	ROUGE 1 PREVIOUS WORKS	No
…	…	…	…

PDF Table Extractor

- A deterministic algorithm to extract tables from NLP papers in PDF format based on GROBID's output
- Evaluate the performance of our PDF table extractor
  - 10 papers from different venues (ACL/NAACL/EMNLP/COLING/CL/TACL)
  - 50 tables with 1,063 numeric content cells

	Macro P	Macro R	Macro F_1
Table caption	79.2	87.0	82.6
Numeric value + IsBolded + Table caption	71.1	77.7	74.0
Numeric value + Row label+ Table caption	55.5	71.4	61.4
Numeric value + Column label + Table caption	49.8	67.2	55.4
Numeric value + IsBolded + Row label + Column label + Table caption	36.6	60.9	43.0

The PDF table extractor is available at: https://github.com/IBM/science-result-extractor

Leaderboards Construction: Dataset Construction

- NLP-TDMS
  - Based on NLP-Progress Github repository which provides expert annotations of various leaderboards for a few hundred NLP papers
  - Manually clean the crawled dataset (e.g., normalize TDM annotations, such as using "*F1*" to represent "*F-score*" and "*Fscore*")
  - A leaderboard is a triple of <*task, dataset, metric*>

	Full	Exp
Papers	332	332
Extracted tables	1269	1269
"Unknown" annotations	-	90
Leaderboard annotations	848	606
Distinct leaderboards	168	77
Distinct tasks	35	18
Distinct datasets	99	44
Distinct metrics	72	30

Remove leaderboards that are associated with less than five papers

A small TDM knowledge taxonomy

Leaderboards Construction: Dataset Construction

- ARC-PDN
  - A more realistic scenario
  - Papers (in PDF format) published in ACL, EMNLP, and NAACL between 2010 to 2015 from the most recent version of the ACL Anthology Reference Corpus (ARC) (Bird et al., 2008)
  - No leaderboard annotations are available

	#Papers	#Extracted tables
ACL	1958	4537
EMNLP	1167	3488
NAACL	730	1559
Total	3855	9584

Leaderboards Construction: Problem Definition

- Prerequisite
  - An experimental NLP paper (PDF format)
  - A predefined TDM taxonomy
- Task
  - Tag the paper with relevant TDM triples from the taxonomy
  - Extract the best numeric score for each predicted TDM triple

A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	CEAF$_e$	Avg.	NER	Link	MUC	B^3	CEAF$_e$	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only) and JOINT models.

Task	Dataset	Evaluation Metric	Best Result
Named Entity Recognition	ACE 2005 (Test)	Accuracy	85.60
Entity Linking	ACE 2005 (Test)	Accuracy	76.78
Coreference Resolution	ACE 2005 (Test)	Avg. F1	76.35
...	...	...	...

Leaderboards Construction: Document and Score Context Representations

- Extract the most relevant part from a long document to predict TDM triples and the associated scores

A Joint Model for Entity Analysis: Coreference, Typing, and Linking

Abstract: We present a joint model of three core tasks in the entity analysis stack: coreference resolution (within-document clustering), named entity recognition (coarse semantic typing), and entity linking (matching to Wikipedia entities). Our model is formally a structured conditional random field. Unary factors encode local features from strong baselines for each task. We then add binary and ternary factors to capture cross-task interactions, such as the constraint that coreferent mentions have the same semantic type. On the ACE 2005 and OntoNotes datasets, we achieve state-of-the- art results for all three tasks. Moreover, joint modeling improves performance on each task over strong independent baselines.

...

	Dev						Test					
	MUC	B^3	CEAF_e	Avg.	NER	Link	MUC	B^3	CEAF_e	Avg.	NER	Link
INDEP.	77.95	74.81	71.84	74.87	83.04	73.07	81.03	74.89	72.56	76.16	82.35	74.71
JOINT	**79.41**	**75.56**	**73.34**	**76.10**	**85.94**	**75.69**	81.41	74.70	72.93	76.35	**85.60**	**76.78**
Δ	+1.46	+0.75	+1.50	+1.23	+2.90	+2.62	+0.42	−0.19	+0.37	+0.19	+3.25	+2.07

Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (task-specific factors only) and JOINT models.

...

	MUC			B^3			CEAF_e			Avg.
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1
BERKELEY	72.85	65.87	69.18	63.55	52.47	57.48	54.31	54.36	54.34	60.33
FERNANDES	–	–	70.51	–	–	57.58	–	–	53.86	60.65
BJORKELUND	74.30	67.46	70.72	62.71	54.96	58.58	59.40	52.27	**55.61**	61.63
INDEP.	72.25	69.30	70.75	60.92	55.73	58.21	55.33	54.14	54.73	61.23
JOINT	72.61	69.91	**71.24**	61.18	56.43	**58.71**	56.17	54.23	55.18	**61.71**

Table 4: CoNLL metric scores for our systems on the CoNLL 2012 blind test set, compared to Durrett and Klein

Document representation

- **Title**
- **Abstract**
- **ExpSetup**: sentences describing experimental setup
- **TableInfo**: concatenation of the table caption and column headers for all tables

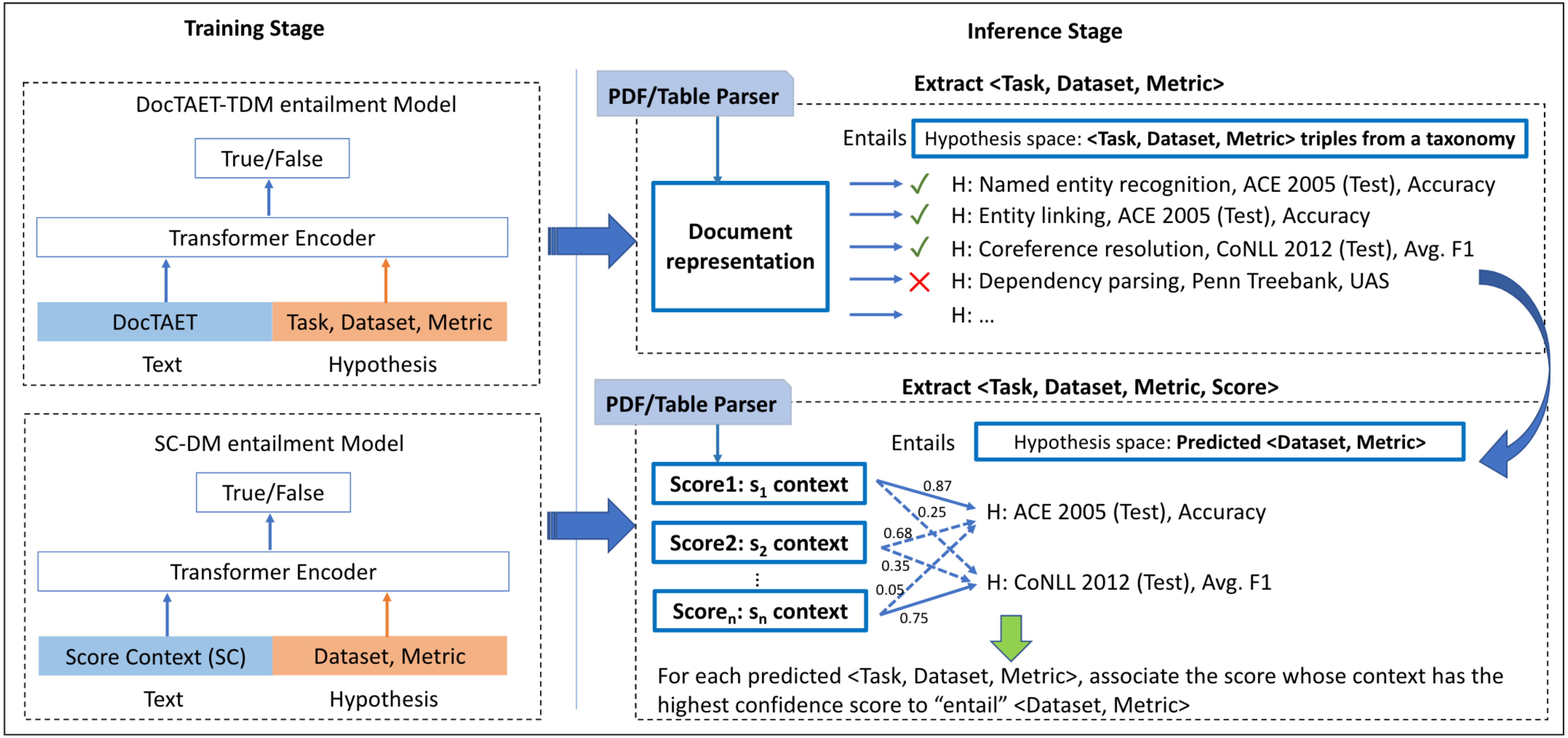
Title	A Joint Model for Entity Analysis: Coreference, Typing, and Linking
Abstract	We present a joint model of three core tasks in the entity analysis stack ...
ExpSetup	We present results on two corpora. First, we use the ACE 2005 corpus (NIST, 2005): ...
TableInfo	Table 1: Results on the ACE 2005 ... and Joint models. Dev MUC B3 CEAFe Avg. NER Link Test ... Table2: ...

Score context representation

- **Corresponding column headers**
- **Table caption**

Score	Score Context
85.60	Test NER Table 1: Results on the ACE 2005 dev and test sets for the INDEP. (Task-specific factors only) and Joint models.
61.71	Avg. F1 Table 4: CoNLL metric scores for our systems on the CoNLL 2012 blind test set. ...
...	...

Leaderboards Construction: TDMS-IE System



Leaderboards Construction: Results on NLP-TDMS

	training	test
Papers	170	162
Extracted tables	679	590
“Unknown” annotations	46	44
Leaderboard annotations	325	281
Distinct leaderboards	77	77

In realistic scenario, a predefined TDM knowledge taxonomy can’t cover all possible TDM triples

	Macro P	Macro R	Macro F_1	Micro P	Micro R	Micro F_1
(a) Task + Dataset + Metric Extraction						
SM	31.8	30.6	31.0	36.0	19.6	25.4
MLC	42.0	23.1	27.8	42.0	20.9	27.9
EL	18.1	31.8	20.5	24.3	36.3	29.1
TDMS-IE	**62.5**	**75.2**	**65.3**	**60.8**	**76.8**	**67.8**
(b) Task + Dataset + Metric Extraction (excluding papers with “Unknown” annotation)						
SM	8.1	6.4	6.9	16.8	7.8	10.6
MLC	**56.8**	30.9	37.3	56.8	23.8	33.6
EL	24.9	43.6	28.1	29.4	42.0	34.6
TDMS-IE	54.1	**65.9**	**56.6**	**60.2**	**73.1**	**66.0**
(c) Task + Dataset + Metric + Score Extraction (excluding papers with “Unknown” annotation)						
SM	1.3	1.0	1.1	3.8	1.8	2.4
MLC	6.8	6.1	6.2	6.8	2.9	4.0
TDMS-IE	9.3	11.8	9.9	10.8	13.1	11.8

Our model struggles to extract the scores for the predicted TDM pairs

Leaderboards Construction: Results on ARC-PDN

- Model trained on the *NLP-TDMS (Exp)* training set
- Evaluation: TDM triples with at least ten associated papers

Task:Dataset:Metric	P@1	P@3	P@5	P@10	#Correct Score	#Wrong Task
Dependency parsing:Penn Treebank:UAS	1.0	1.0	0.8	0.9	2	0
Summarization:DUC 2004 Task 1:ROUGE-2	0.0	0.67	0.8	0.7	0	0
Word sense disambiguation:Senseval 2:F1	0.0	0.0	0.1	0.1	0	0
Word sense disambiguation:SemEval 2007:F1	1.0	1.0	0.8	0.7	1	0
Word segmentation:Chinese Treebank 6:F1	1.0	0.67	0.4	0.2	0	2
Word Segmentation:MSRA:F1	1.0	0.67	0.6	0.7	2	3
Sentiment analysis:SST-2:Accuracy	1.0	0.67	0.6	0.3	0	3
AMR parsing:LDC2014T12:F1 on All	0.0	0.67	0.4	0.2	0	5
CCG supertagging:CCGBank:Accuracy	1.0	1.0	1.0	0.8	0	1
Machine translation:WMT 2014 EN-FR:BLEU	1.0	0.33	0.2	0.1	0	0
Macro-average	0.70	0.67	0.57	0.46	-	-

Most papers are about MT, but report results on ***WMT 2012 EN-FR*** or ***WMT 2014 EN-DE***

Extracting TDMS tuples is a challenging task

Leaderboards to Construct Leaderboards

Scientific Results Extraction

5 papers with code • 2 benchmarks • 4 datasets

Scientific results extraction is the task of extracting relevant result information (e.g., in the case of Machine learning performance results: task, dataset, metric name, metric value) from the scientific literature.

Benchmarks

Add a Result

These leaderboards are used to track progress in Scientific Results Extraction

Trend	Dataset	Best Model	Paper	Code	Compare
	NLP-TDMS (Exp, arXiv only)	AxCell			See all
	PWC Leaderboards (restricted)	AxCell			See all

Datasets

SegmentedTables · PWC Leaderboards · ArxivPapers · LinkedResults

Most implemented papers

Most implemented · Social · Latest · No code

Search for a paper, author or keyword

Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction

IBM/science-result-extractor • TensorFlow • ACL 2019

1

Paper · Code

While the fast-paced inception of novel tasks and new datasets helps foster active research in a community towards interesting directions, keeping track of the abundance of research activity in different areas on different datasets is likely to become increasingly difficult.

arXiv > cs > arXiv:2401.06233

Computer Science > Computation and Language

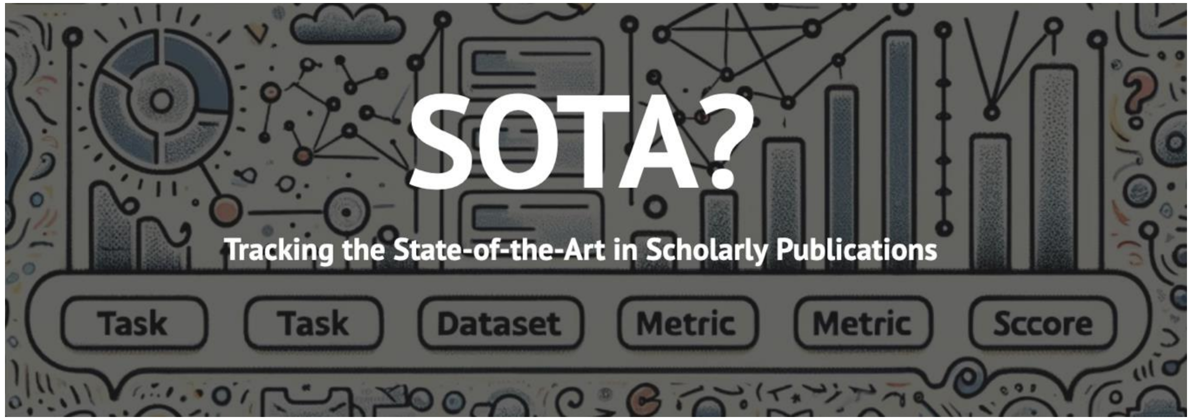
[Submitted on 11 Jan 2024 (v1), last revised 21 Feb 2024 (this version, v2)]

LEGOBench: Scientific Leaderboard Generation Benchmark

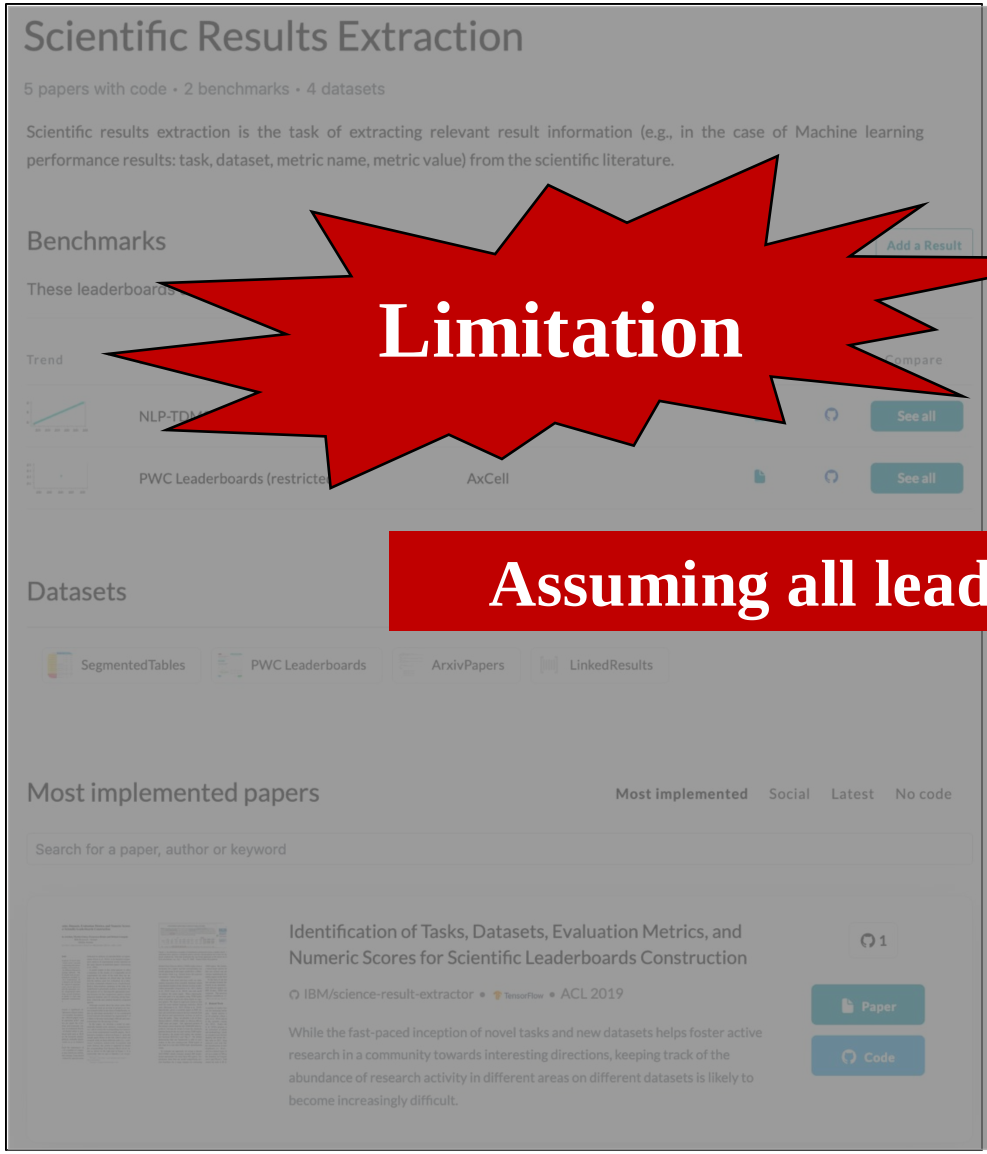
CLEF 2024 SimpleText Track:

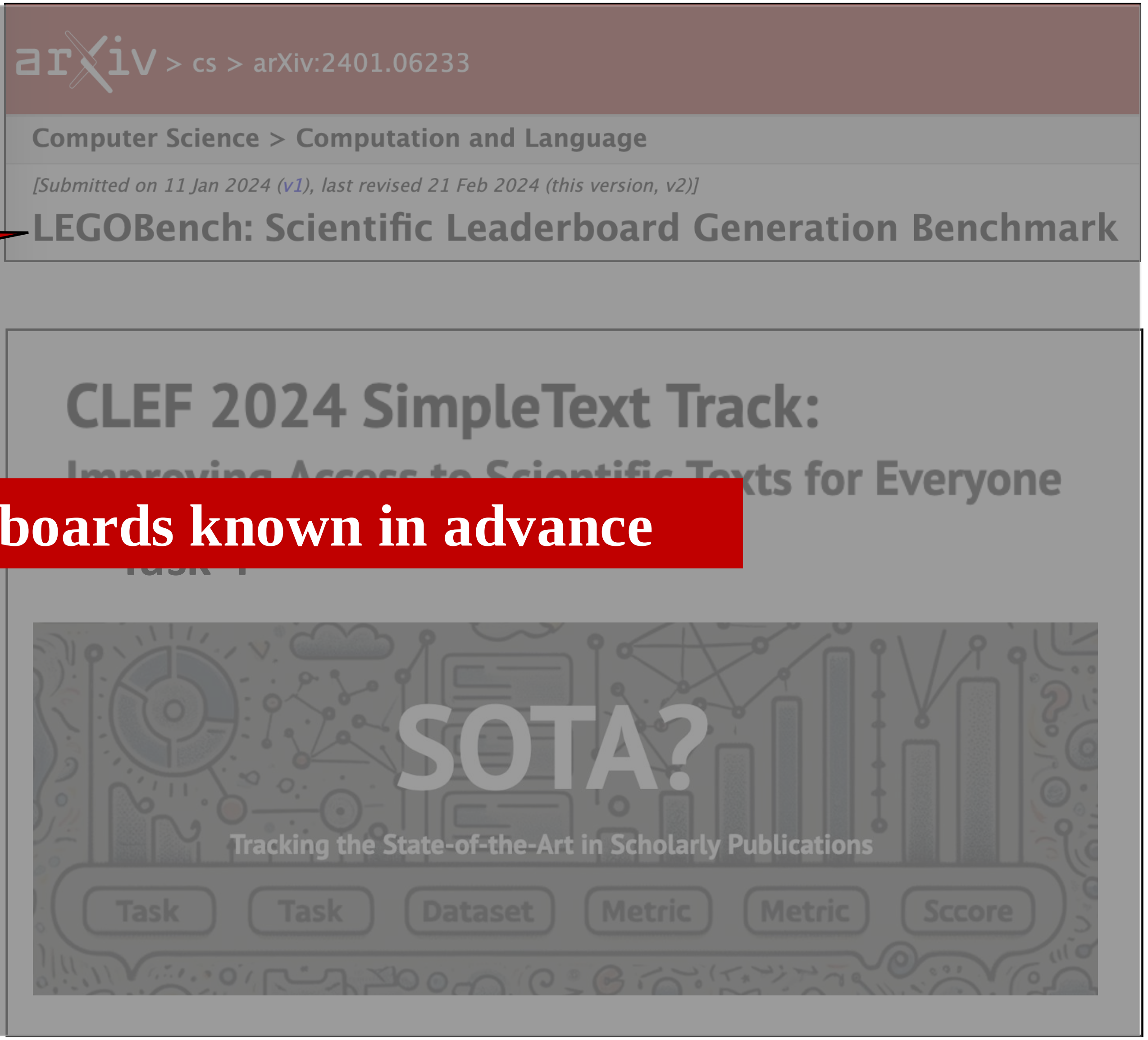
Improving Access to Scientific Texts for Everyone

- Task 4



Leaderboards to Construct Leaderboards





Limitation

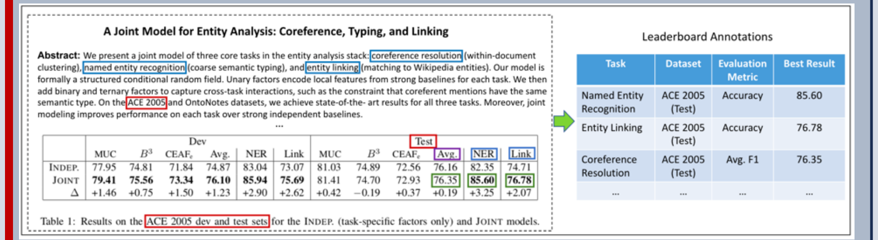
Assuming all leaderboards known in advance

Outline

Build Global Scientific Evidence Map

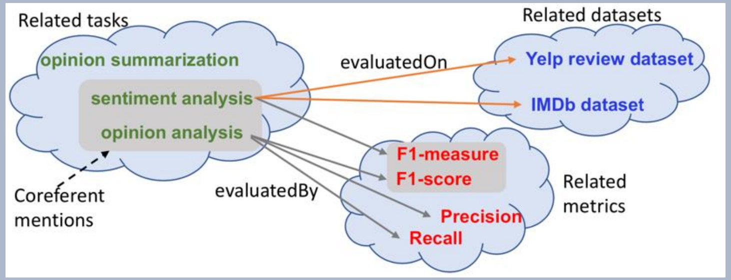
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



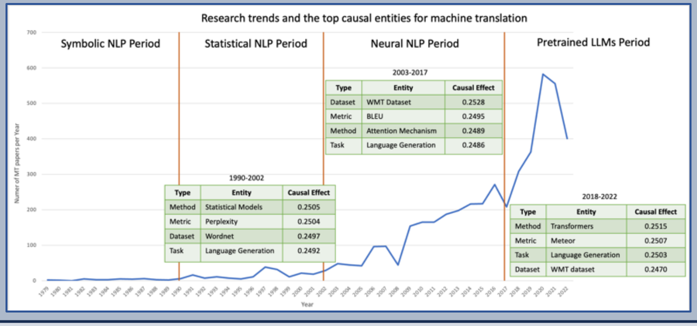
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

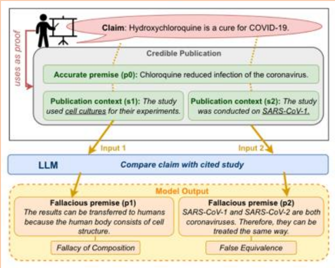
- NLP Concepts Causal Analysis



Scientific Communication

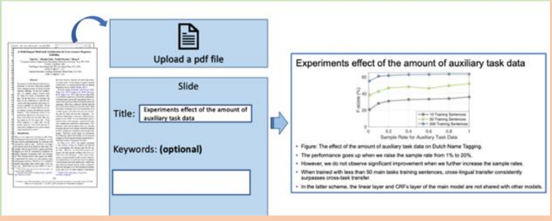
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

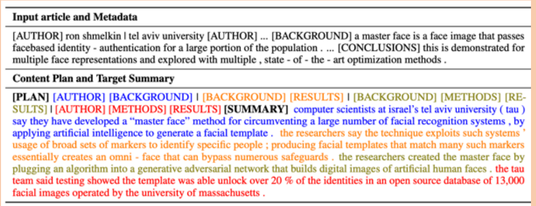
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures

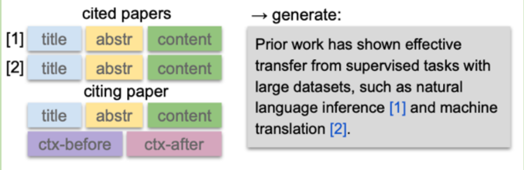


Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation **[Funkquist et al., EMNLP 2023]**

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation **[O'Doherty et al., ACL 2024 SRW]**



Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards

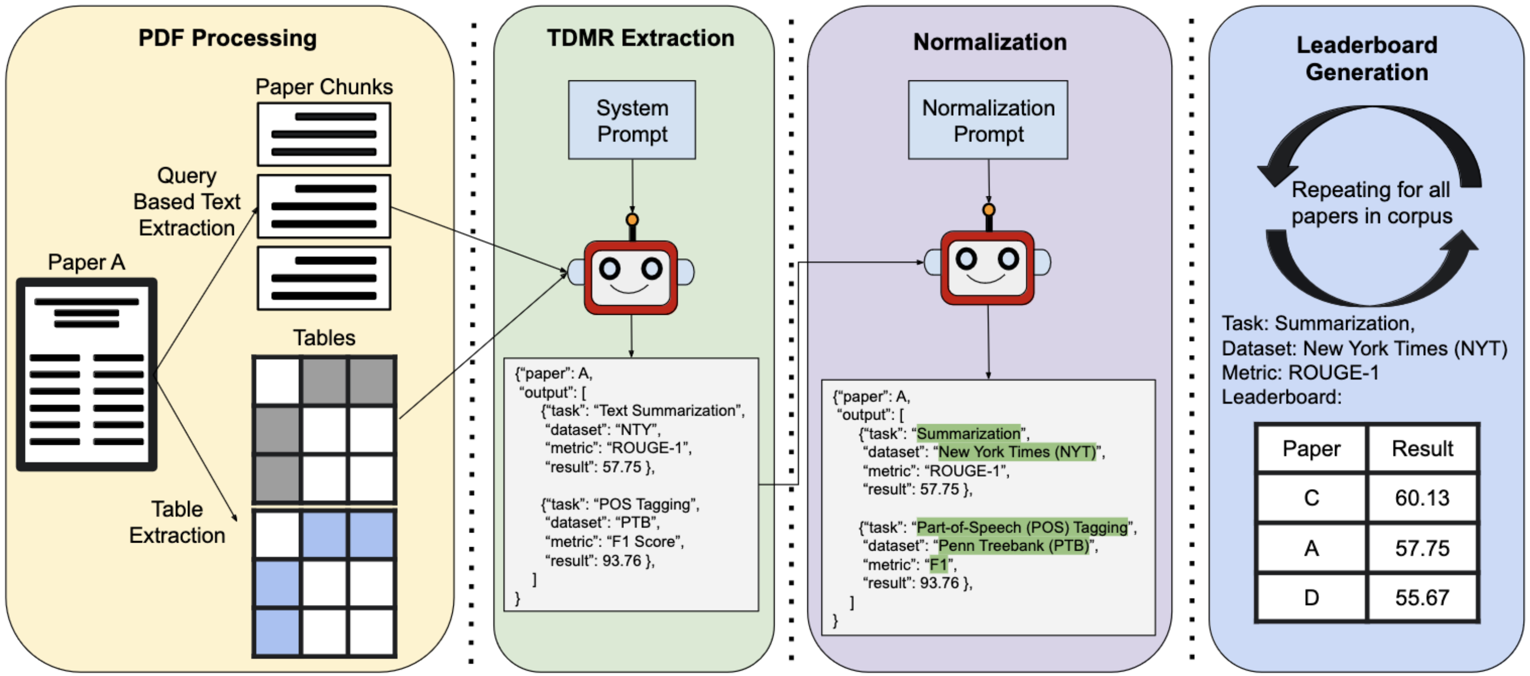
Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, Iryna Gurevych

(EMNLP 2024)

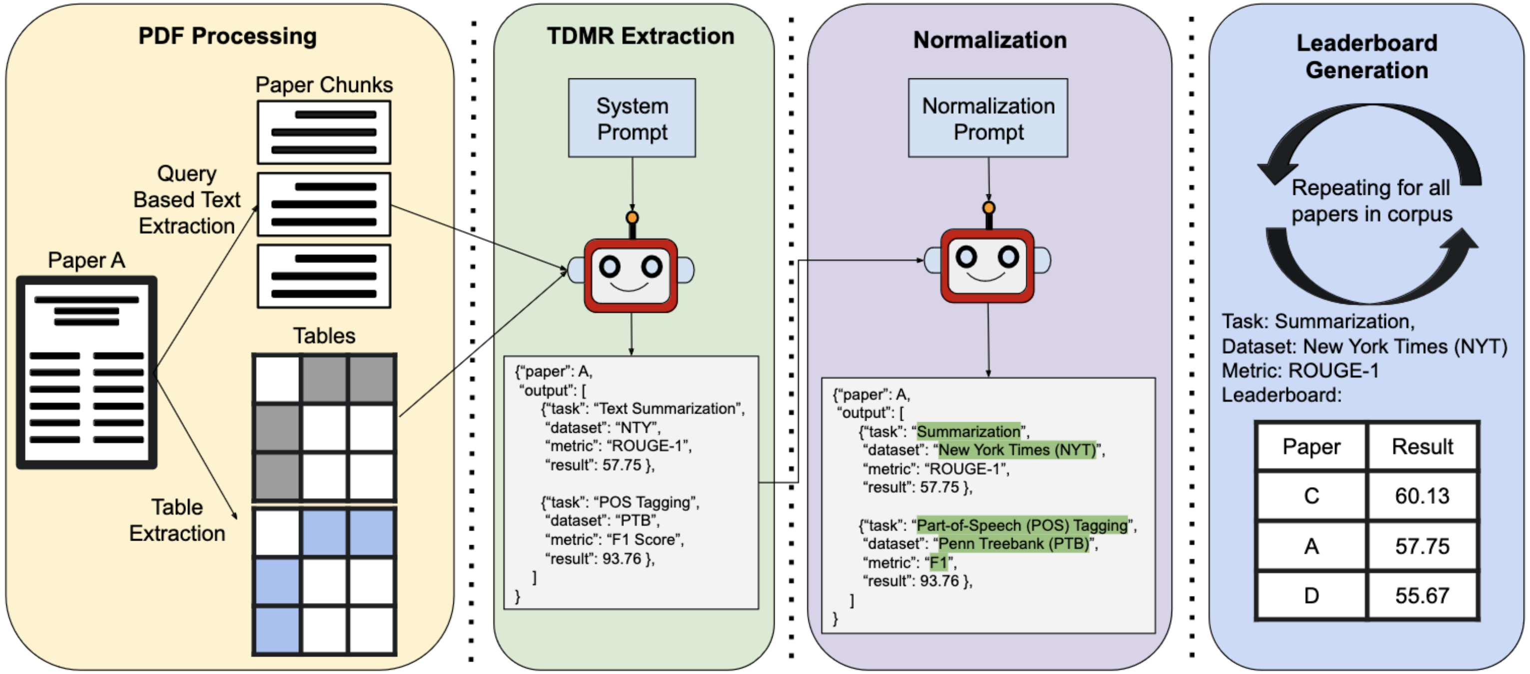





Leaderboard Construction from “Cold Start”

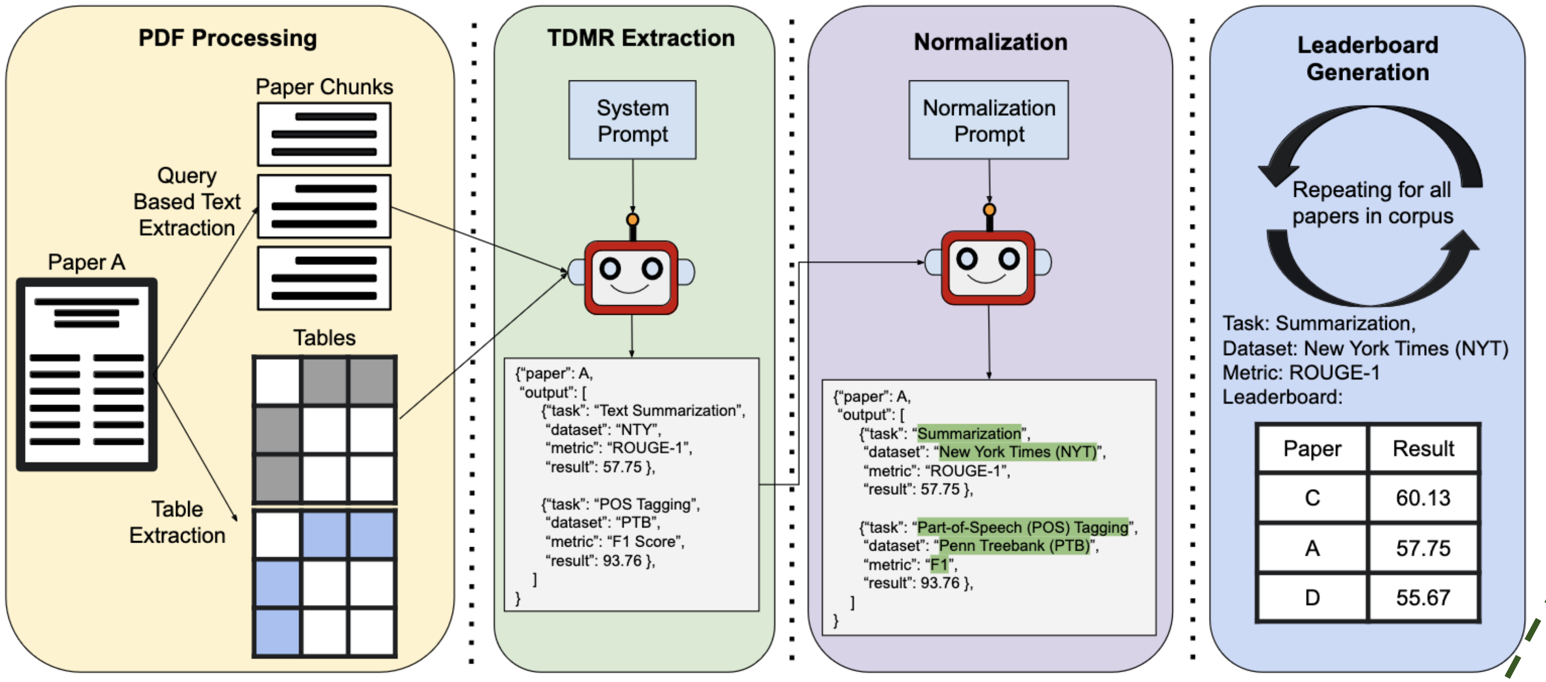


Leaderboard Construction from "Cold Start"



1. Closed-world assumption: all leaderboards are known beforehand
2. Real world scenario: keep a list of known leaderboards, add new TDM triples dynamically
3. Cold start: the list of known leaderboards is empty

Ongoing Work: Leaderboard Construction from “Cold Start”



SciLead Dataset

- Exhaustive TMDS annotations
- 43 papers
- 27 leaderboards

	Model	Leaderboard Recall	Paper Coverage	Result Coverage	Average Overlap
Fully	Llama-2	70.37	34.96	8.26	4.96
	Mixtral	88.89	46.85	16.32	11.96
	Llama-3	96.30	**79.18**	29.30	25.49
	GPT-4	**100.00**	70.37	**51.79**	**53.87**
Partially	Llama-2	74.07	30.75	4.02	1.18
	Mixtral	77.78	28.70	12.01	12.47
	Llama-3	92.59	61.90	19.01	19.52
	GPT-4	88.89	55.92	40.06	43.71
Cold Start	Llama-2	16.05	5.66	0.49	0.05
	Mixtral	49.38	20.49	8.10	3.03
	Llama-3	79.01	58.78	17.18	17.63
	GPT-4	81.48	58.74	46.13	48.15

1. Closed-world assumption: all leaderboards are known beforehand
2. Real world scenario: keep a list of known leaderboards, add new TDM triples dynamically
3. Cold start: the list of known leaderboards is empty

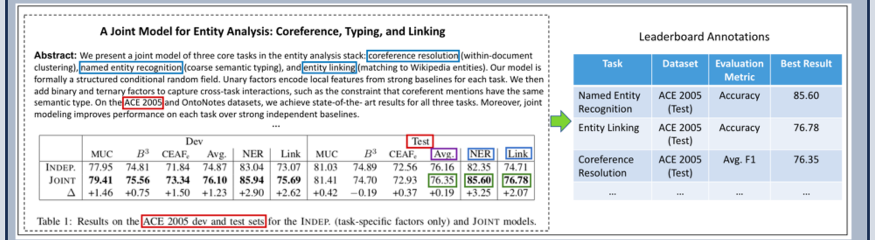
Compared to setting 2, strong LLMs perform better in setting 3 (cold start)

Outline

Build Global Scientific Evidence Map

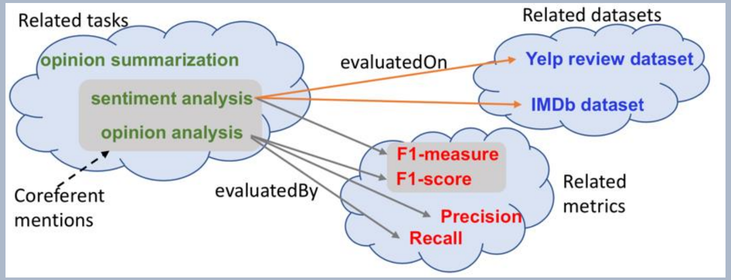
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



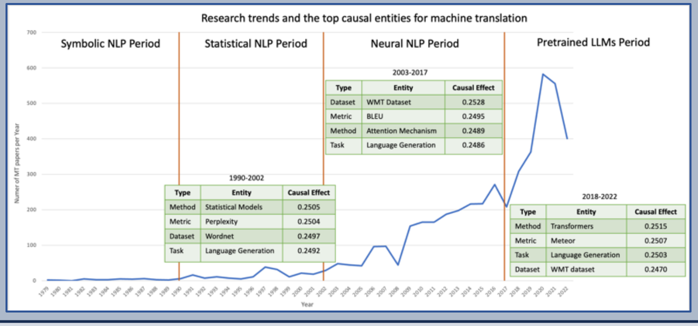
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

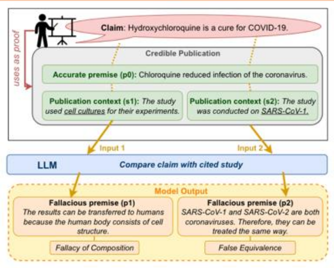
- NLP Concepts Causal Analysis



Scientific Communication

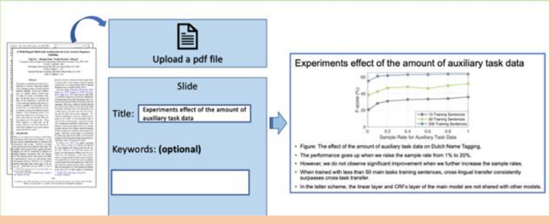
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

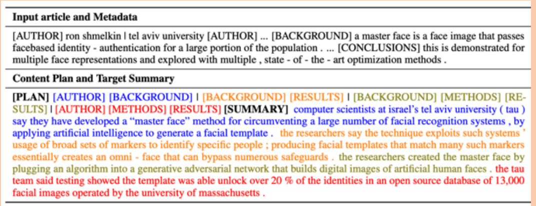
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures



Input article and Metadata

[AUTHOR] ron shmelkin | tel aviv university [AUTHOR] ... [BACKGROUND] a master face is a face image that passes facebased identity - authentication for a large portion of the population . ... [CONCLUSIONS] this is demonstrated for multiple face representations and explored with multiple , state - of - the - art optimization methods .

Content Plan and Target Summary

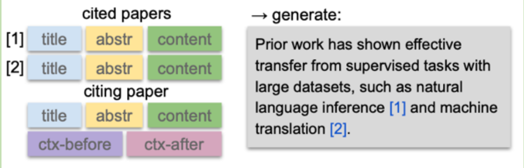
[PLAN] [AUTHOR] [BACKGROUND] | [BACKGROUND] [RESULTS] | [BACKGROUND] [METHODS] [RESULTS] | [AUTHOR] [METHODS] [RESULTS] **[SUMMARY]** computer scientists at israel's tel aviv university (tau) say they have developed a "master face" method for circumventing a large number of facial recognition systems , by applying artificial intelligence to generate a facial template . the researchers say the technique exploits such systems ' usage of broad sets of markers to identify specific people ; producing facial templates that match many such markers essentially creates an omni - face that can bypass numerous safeguards . the researchers created the master face by plugging an algorithm into a generative adversarial network that builds digital images of artificial human faces . the tau team said testing showed the template was able unlock over 20 % of the identities in an open source database of 13,000 facial images operated by the university of massachusetts .

Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation [Funkquist et al., EMNLP 2023]

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation [O'Doherty et al., ACL 2024 SRW]



A Diachronic Analysis of Paradigm Shifts in NLP Research: When, How, and Why?

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, Iryna Gurevych
(EMNLP 2023)




Diachronic Analysis of the NLP Research Areas

Develop a model to analyse NLP research areas and answer the following questions:

- What is the general trend of a research area?
- How is a research area influenced by other research concepts?
- How do researchers argue about a specific research concept? (ongoing)

We use NLP tasks to approximate research areas

- Named Entity Recognition
- Relation Extraction
- Question Answering
- Machine Translation
- Sentiment Analysis
- Coreference Resolution
- Discourse Parsing
- …

We define four types of research concepts

- Task (T)
- Dataset (D)
- Evaluation Metric (M)
- Method (M)

Diachronic Analysis of the NLP Research Areas

Step 1: Pre-processing

- Grobid Parser
- PDF Table Parser
- TDMM Entity Taggers

55K NLP Papers (1979-2022) from the ACL Anthology

Step 2: Multiple linear regression

What type of entities (X_i) influence the research trends of an NLP task?

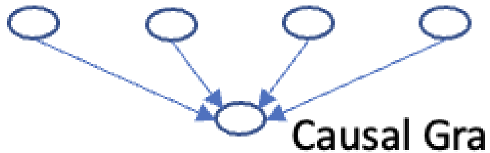
Step 3: Causal Analysis

3.1 calculate variable values for each task y

$\Delta freq(y)$ $\Delta stability(y)$ $\delta_y(x)$

Task	2018-2022			
	Δfreq	δ(Transformer)	δ(Meteor)	....
Machine Translation	....	....	....	....
....	....	....	....	....

3.2 causal structure discovery (DirectLiNGAM)

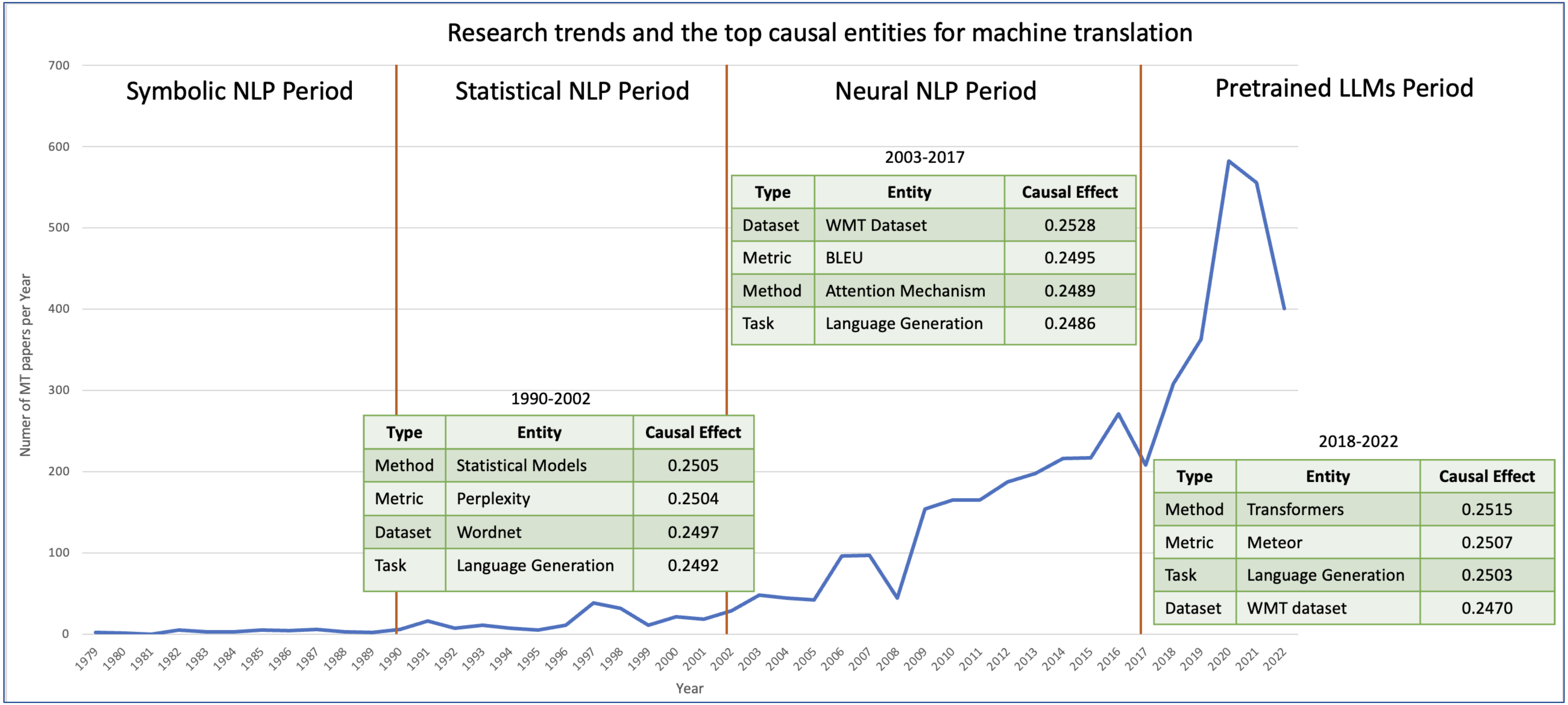


Causal Graph

3.3 causal inference analysis

Target Task	2018-2022	
	Associated Entities	Causal Effect
Machine Translation	Transformers	0.2515
	METEOR	0.2507
	Language Generation	0.2503
	WMT Dataset	0.247

Diachronic Analysis of the NLP Research Areas



1990-2002

Type	Entity	Causal Effect
Method	Statistical Models	0.2505
Metric	Perplexity	0.2504
Dataset	Wordnet	0.2497
Task	Language Generation	0.2492

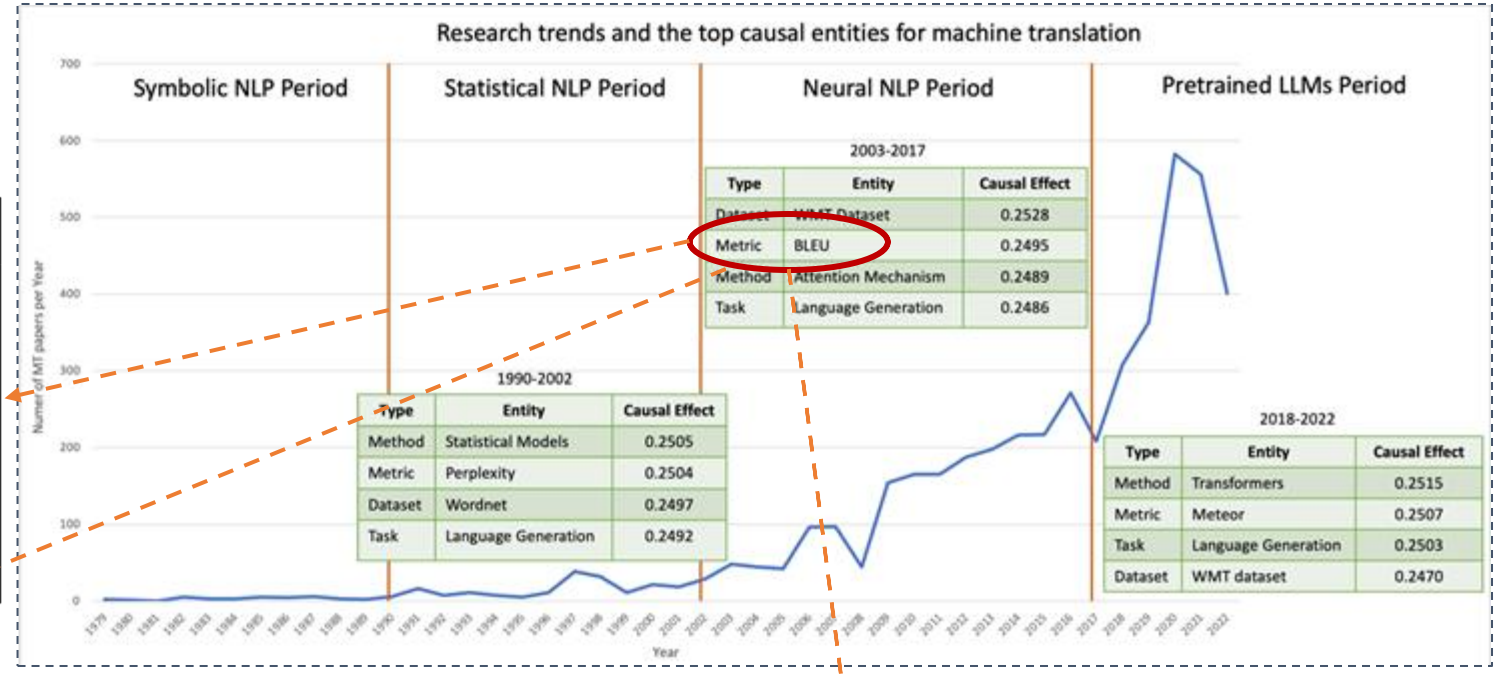
2003-2017

Type	Entity	Causal Effect
Dataset	WMT Dataset	0.2528
Metric	BLEU	0.2495
Method	Attention Mechanism	0.2489
Task	Language Generation	0.2486

2018-2022

Type	Entity	Causal Effect
Method	Transformers	0.2515
Metric	Meteor	0.2507
Task	Language Generation	0.2503
Dataset	WMT dataset	0.2470

Ongoing Work: A First Step Towards the Global Claim Veracity Summary



Topic 1: Is BLUE suitable to evaluate MT in general?

Pro (6)	Con (8)
1. BLEU will accelerate the MT R&D cycle. (2002) 2. BLEU is fast and easy to run, and it can be used as a target function in parameter optimization training procedures. (2008) 3. BLEU-4 (correctly) predicts that human subjects prefer SUMTIME-Hybrid texts to pCRU-random texts. (2009) 4. It is acceptable to use BLEU-like metrics (with caution) to estimate the linguistic quality of generated texts. (2009) 5. BLEU and NIST's strong showing in both the machine and human evaluation results indicates that they are still the best general choice for training model parameters. (2010) 6. …	1. BLEU **may** not be appropriate for comparing systems which employ different translation strategies. (2006) 2. In practice a higher Bleu score is not necessarily indicative of better translation quality. (2006) 3. BLEU **may** not be a reliable MT quality indicator. (2007) 4. Automatic metrics such as BLEU do not give a complete and reliable picture and carrying out additional evaluations is crucial. (2009) 5. Word-token BLEU is not capable of measuring the morpheme level improvements. (2010) 6. …

Topic 2: Is BLEU suitable for sentence level evaluation?

Pro (0)	Con (2)
	1. This confirms Liu and Gildea (2005)'s finding that in sentence level evaluation, long n-grams in BLEU are not beneficial. (2006) 2. BLEU is not very predictive of sentence level evaluation. (2007)

Topic 3: Can BLEU evaluate the syntactic aspect of translation quality?

Pro (1)	Con (3)
1. Higher order N-grams are used in BLEU as an indirect measure of a translation's level of grammatical wellformedness. (2005)	1. The BLEU metric **may** not be affected by the syntactic aspect of translation quality. (2002) 2. Neither the IBM models nor the BLEU metric are able to recognize long distance dependencies (such as, for example, accounting for fundamental word order differences when translating from a SOV language into a SVO language). (2003) 3. …

Topic 4: Can BLEU evaluate translation adequacy and fluency?

Pro (0)	Con (5)
	1. The standard BLEU approach tends to over-estimate its performance for translation adequacy. (2004) 2. For out-of-domain French-English, where Systran receives among the best adequacy and fluency scores, but a worse BLEU score than all but one statistical system. (2006) 3. Manual evaluation confirms our suspicion that the BLEU metric is less sensitive than QUEEN to improvements related to adequacy. (2007) 4. …

Topic 5: BLEU vs other metrics?

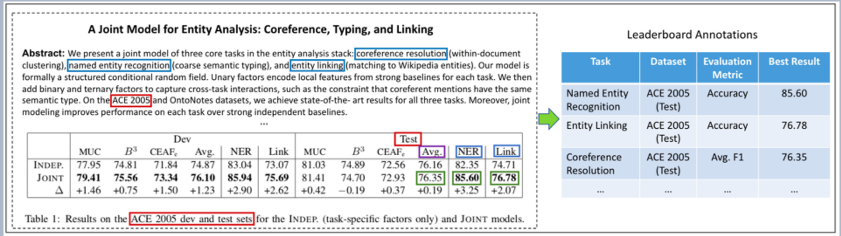
Pro (1)	Con (11)
1. Third, with the exception of BLEU:1, the performance of the BLEU, NIST, and the METEOR α=.5 models appears to be more robust across the other evaluation metrics than the standard METEOR, METEOR ranking, and edit distance-based models (WER, TER, TERp, an TERpA). (2010)	1. We have found NIST a more reliable evaluation metric than BLEU and in particular ROUGE which did not seem to offer any advantage over simple string-edit distance. (2006) 2. …

Outline

Build Global Scientific Evidence Map

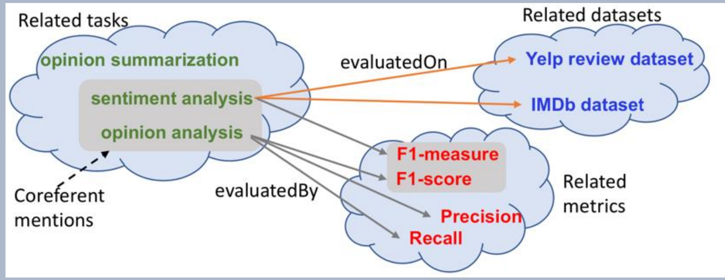
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



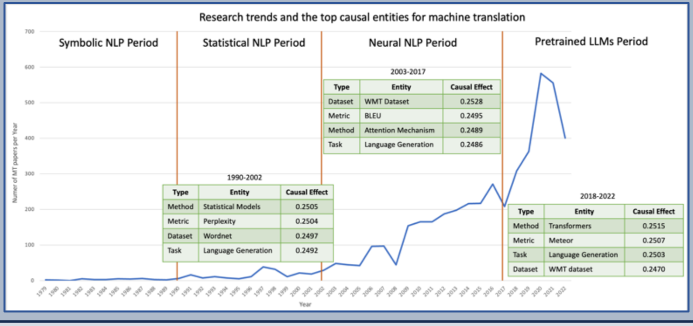
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

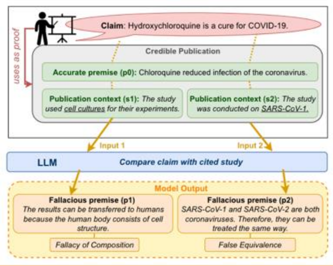
- NLP Concepts Causal Analysis



Scientific Communication

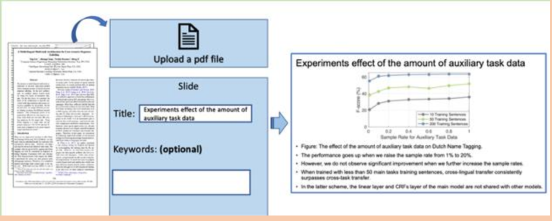
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

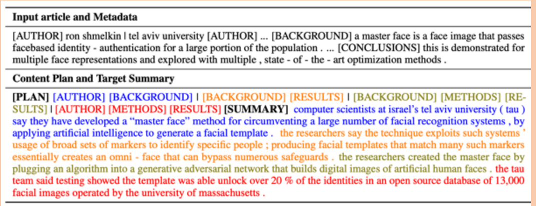
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures

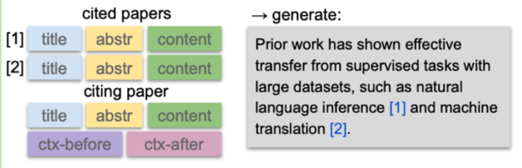


Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation [Funkquist et al., EMNLP 2023]

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation [O'Doherty et al., ACL 2024 SRW]



Missci: Reconstructing the Fallacies in Misrepresented Science

Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych

(ACL 2024)

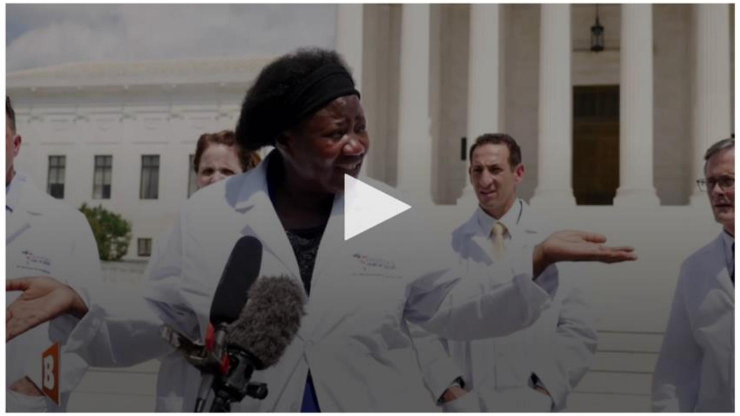
Misinformation Based on Scientific Studies

BREITBART

WATCH— TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

1,419 | EMAIL | SHARE | TWEET

Video Source: Matt Perdie



by AMY FURR | 27 Jul 2020 | 257

LISTEN TO STORY 2:19

Studies that claim hydroxychloroquine does not work when treating patients with the coronavirus are "fake science," a doctor at the "White Coat Summit" in Washington, DC, said on Monday.

Dr. Stella Immanuel of Rehoboth Medical Center in Houston, Texas, said she had 350 patients she put on hydroxychloroquine and every one of them recovered.

She continued:

> This is what I will say to all those studies — they had high doses, they were given the wrong patients — I would call them fake science. Any study that says hydroxychloroquine doesn't work is fake science. And I want them to show me how it doesn't work. How is it going to work for 350 patients for me, and they are all alive, and then somebody says it doesn't work? Guys, all them studies: fake science.

Stella Immanuel

stated on July 27, 2020 in a press conference:

> **"This virus has a cure. It is called hydroxychloroquine, zinc, and Zithromax. I know you people want to talk about a mask. Hello? You don't need (a) mask. There is a cure."**

facebook

American Medicine Today
7 October 2020 ·

Stella Immanuel M.D., notorious for stating hydroxychloroquine is a "cure" for COVID-19 has denounced White House doctors for not giving President Donald J. Trump the drug to treat his infection. Dr. Immanuel tweeted directly to President Trump and she joins us this weekend on #AmericanMedicineToday radio to discuss.

Listen Saturday on Newsradio 710 at 8AM, WBHP - The Big Talker at 10AM, Newsradio WFLA at 12PM & 6PM, News Talk 93.1 at 12pm and Sunday on News Radio 105.5 WERC at 1PM!

Donald J. Trump @realDonaldTrump · Oct 2
Tonight, @FLOTUS and I tested positive for COVID-19. We will begin our quarantine and recovery process immediately. We will get through this TOGETHER!
573.5K 919K 1.8M

Stella Immanuel MD @stella_immanuel · Oct 2
Sir you exposed the effectiveness of HCQ to us. Why did you and FLOTUS not take it for prevention? You did not need to get covid19. HCQ, zinc, Vit C and D and you will not get sick. Praying for you and your team sir.
703 1.4K 8K

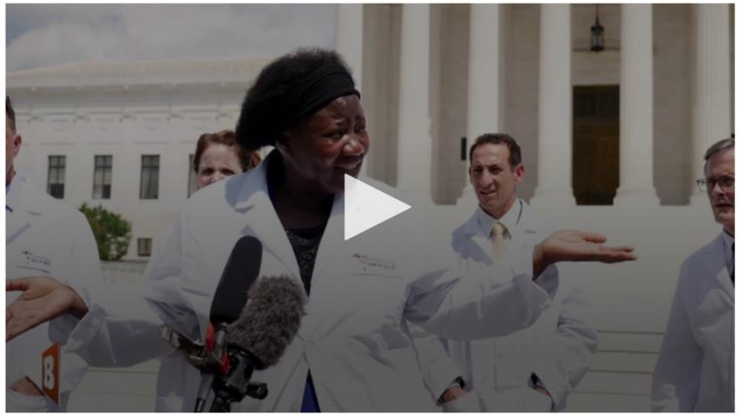
Misinformation Based on Scientific Studies

BREITBART

WATCH— TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

1,419 EMAIL SHARE TWEET

Video Source: Matt Perdie



by AMY FURR 27 Jul 2020 257

LISTEN TO STORY 2:19

Studies that claim hydroxychloroquine does not work when treating patients with the coronavirus are "fake science," a doctor at the "White Coat Summit" in Washington, DC, said on Monday.

Dr. Stella Immanuel of Rehoboth Medical Center in Houston, Texas, said she had 350 patients she put on hydroxychloroquine and every one of them recovered.

She continued:

> This is what I will say to all those studies — they had high doses, they were given the wrong patients — I would call them fake science. Any study that says hydroxychloroquine doesn't work is fake science. And I want them to show me how it doesn't work. How is it going to work for 350 patients for me, and they are all alive, and then somebody says it doesn't work? Guys, all them studies: fake science.

Robin Shipkosky · an hour ago

I watched the news conference today. Dr. Emmanuel was ON FIRE!!! Everyone should listen to her!

11 ^ | ˅ · Reply · Share ›

Wesley Alexander → Robin Shipkosky · an hour ago · edited

Doctor Stella Immanuel, my new hero. This forum of doctors have everything to lose -even their medical licenses, but they are warriors for the truth. This doctor NEEDS to meet with Trump. I was hoping Breitbart would single out her statements. She is an inspiration.

He listens to her and he will become a disciple. She is a LIGHT illuminating the darkness enveloping this country.caused by panic, fear and tyranny.

American Frontline Physicians the website to check out when you are denied to use of HQ.

6 ^ | ˅ · Reply · Share ›

Frank Galvin → Wesley Alexander · an hour ago

She needs to speak at the RNC Convention.

6 ^ | ˅ · Reply · Share ›

catwoman401k → Wesley Alexander · 12 minutes ago

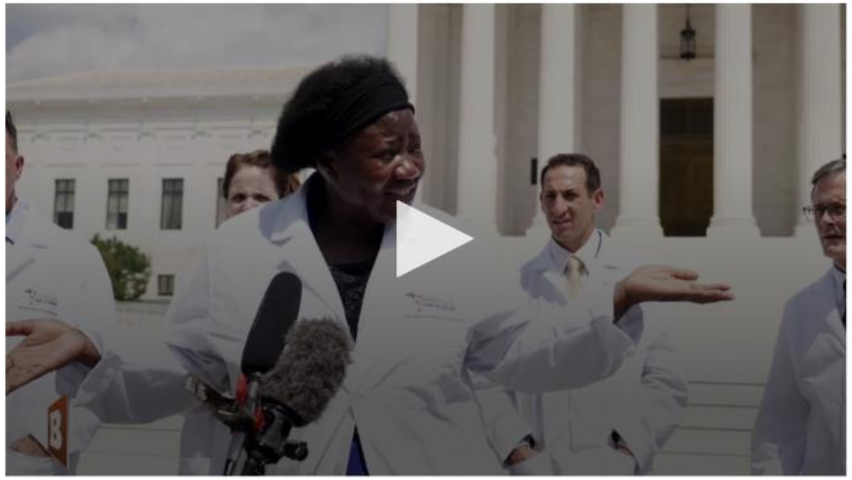
She takes no prisoners. Awesome

Misinformation Based on Scientific Studies

WATCH– TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

f 1,419 | EMAIL | SHARE | TWEET

Video Source: Matt Perdie



by AMY FURR | 27 Jul 2020 | 257

> This virus has a cure. It is called hydroxychloroquine, zinc, and Zithromax. I know you people want to talk about a mask. Hello? You don't need mask. There is a cure.

> **The study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that say it works.**

Virology Journal — BioMed Central

Research — Open Access

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent[1], Eric Bergeron[2], Suzanne Benjannet[2], Bobbie R Erickson[1], Pierre E Rollin[1], Thomas G Ksiazek[1], Nabil G Seidah[2] and Stuart T Nichol*[1]

Address: [1]Special Pathogens Brach, Division of Viral and Rickettsial Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia, 30333, USA and [2]Laboratory of Biochemical Neuroendocrinology, Clinical Research Institute of Montreal, 110 Pine Ave West, Montreal, QCH2W1R7, Canada

Email: Martin J Vincent - mvincent@cdc.gov; Eric Bergeron - bergere@ircm.qc.ca; Suzanne Benjannet - benjans@ircm.qc.ca; Bobbie R Erickson - BErickson1@cdc.gov; Pierre E Rollin - PRollin@cdc.gov; Thomas G Ksiazek - TKsiazek@cdc.gov; Nabil G Seidah - seidahn@ircm.qc.ca; Stuart T Nichol* - SNichol@cdc.gov

* Corresponding author

Published: 22 August 2005

Virology Journal 2005, **2**:69 doi:10.1186/1743-422X-2-69

Received: 12 July 2005

Accepted: 22 August 2005

This article is available from: http://www.virologyj.com/content/2/1/69

© 2005 Vincent et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Severe acute respiratory syndrome (SARS) is caused by a newly discovered coronavirus (SARS-CoV). No effective prophylactic or post-exposure therapy is currently available.

Results: We report, however, that chloroquine has strong antiviral effects on SARS-CoV infection of primate cells. These inhibitory effects are observed when the cells are treated with the drug either before or after exposure to the virus, suggesting both prophylactic and therapeutic advantage. In addition to the well-known functions of chloroquine such as elevations of endosomal pH, the drug appears to interfere with terminal glycosylation of the cellular receptor, angiotensin-converting enzyme 2. This may negatively influence the virus-receptor binding and abrogate the infection, with further ramifications by the elevation of vesicular pH, resulting in the inhibition of infection and spread of SARS CoV at clinically admissible concentrations.

Conclusion: Chloroquine is effective in preventing the spread of SARS CoV in cell culture. Favorable inhibition of virus spread was observed when the cells were either treated with chloroquine prior to or after SARS CoV infection. In addition, the indirect immunofluorescence assay described herein represents a simple and rapid method for screening SARS-CoV antiviral compounds.

Background

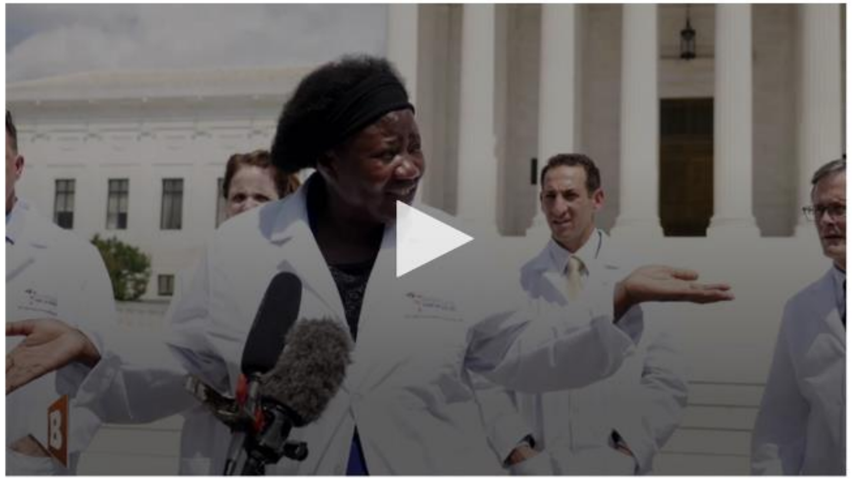
Severe acute respiratory syndrome (SARS) is an emerging disease that was first reported in Guangdong Province, China, in late 2002. The disease rapidly spread to at least 30 countries within months of its first appearance, and concerted worldwide efforts led to the identification of the etiological agent as SARS coronavirus (SARS-CoV), a novel member of the family *Coronaviridae* [1]. Complete genome sequencing of SARS-CoV [2,3] confirmed that this pathogen is not closely related to any of the

Misinformation ~~Based on~~ Misrepresents Scientific Studies

WATCH– TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

1,419 | EMAIL | SHARE | TWEET

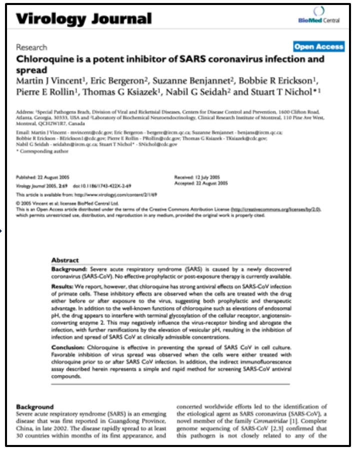
Video Source: Matt Perdie



by AMY FURR | 27 Jul 2020 | 257

> This virus has a cure. It is called hydroxychloroquine, zinc, and Zithromax. I know you people want to talk about a mask. Hello? You don't need mask. There is a cure.

> **The study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that say it works.**



Virology Journal

BioMed Central

Research

Open Access

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent[1], Eric Bergeron[2], Suzanne Benjannet[2], Bobbie R Erickson[1], Pierre E Rollin[1], Thomas G Ksiazek[1], Nabil G Seidah[2] and Stuart T Nichol*[1]

Address: [1]Special Pathogens Brach, Division of Viral and Rickettsial Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia, 30333, USA and [2]Laboratory of Biochemical Neuroendocrinology, Clinical Research Institute of Montreal, 110 Pine Ave West, Montreal, QCH2W1R7, Canada

Email: Martin J Vincent - mvincent@cdc.gov; Eric Bergeron - bergere@ircm.qc.ca; Suzanne Benjannet - benjans@ircm.qc.ca; Bobbie R Erickson - BErickson1@cdc.gov; Pierre E Rollin - PRollin@cdc.gov; Thomas G Ksiazek - TKsiazek@cdc.gov; Nabil G Seidah - seidahn@ircm.qc.ca; Stuart T Nichol* - SNichol@cdc.gov

* Corresponding author

Published: 22 August 2005

Virology Journal 2005, 2:69 doi:10.1186/1743-422X-2-69

This article is available from: http://www.virologyj.com/content/2/1/69

Received: 12 July 2005
Accepted: 22 August 2005

© 2005 Vincent et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Severe acute respiratory syndrome (SARS) is caused by a newly discovered coronavirus (SARS-CoV). No effective prophylactic or post-exposure therapy is currently available.

Results: We report, however, that chloroquine has strong antiviral effects on SARS-CoV infection of primate cells. These inhibitory effects are observed when the cells are treated with the drug either before or after exposure to the virus, suggesting both prophylactic and therapeutic advantage. In addition to the well-known functions of chloroquine such as elevations of endosomal pH, the drug appears to interfere with terminal glycosylation of the cellular receptor, angiotensin-converting enzyme 2. This may negatively influence the virus-receptor binding and abrogate the infection, with further ramifications by the elevation of vesicular pH, resulting in the inhibition of infection and spread of SARS CoV at clinically admissible concentrations.

Conclusion: Chloroquine is effective in preventing the spread of SARS CoV in cell culture. Favorable inhibition of virus spread was observed when the cells were either treated with chloroquine prior to or after SARS CoV infection. In addition, the indirect immunofluorescence assay described herein represents a simple and rapid method for screening SARS-CoV antiviral compounds.

Background

Severe acute respiratory syndrome (SARS) is an emerging disease that was first reported in Guangdong Province, China, in late 2002. The disease rapidly spread to at least 30 countries within months of its first appearance, and concerted worldwide efforts led to the identification of the etiological agent as SARS coronavirus (SARS-CoV), a novel member of the family *Coronaviridae* [1]. Complete genome sequencing of SARS-CoV [2,3] confirmed that this pathogen is not closely related to any of the

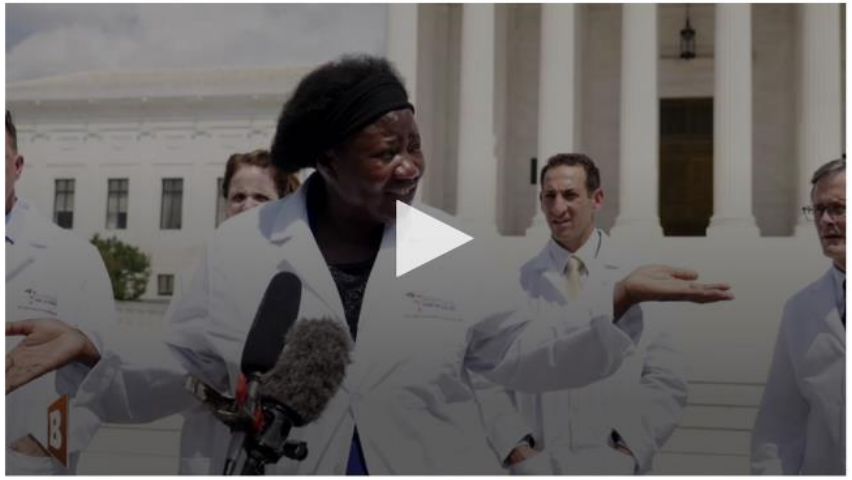
> Post-infection chloroquine treatment reduces SARS-CoV infection and spread.

Misinformation ~~Based on~~ Misrepresents Scientific Studies

WATCH— TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

1,419 EMAIL SHARE TWEET

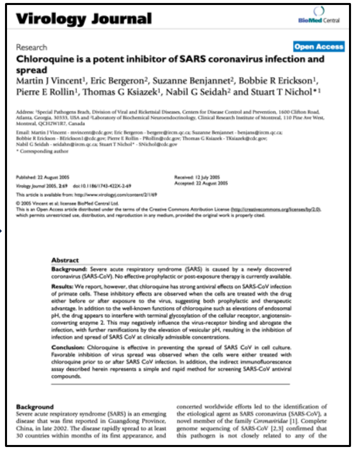
Video Source: Matt Perdie



by AMY FURR 27 Jul 2020 257

This virus has a cure. It is called hydroxychloroquine, zinc, and Zithromax. I know you people want to talk about a mask. Hello? You don't need mask. There is a cure.

The study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that say it works.



Virology Journal

Research

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent, Eric Bergeron, Suzanne Benjannet, Bobbie R Erickson, Pierre E Rollin, Thomas G Ksiazek, Nabil G Seidah and Stuart T Nichol

Post-infection chloroquine treatment reduces SARS-CoV infection and spread.

But

1. *in vitro* study (outside of a living organism)

We have provided evidence that chloroquine is effective in preventing SARS-CoV infection **in cell culture**.

2. *SARS-CoV-1 ! = SARS-CoV-2 (causes Covid-19)*

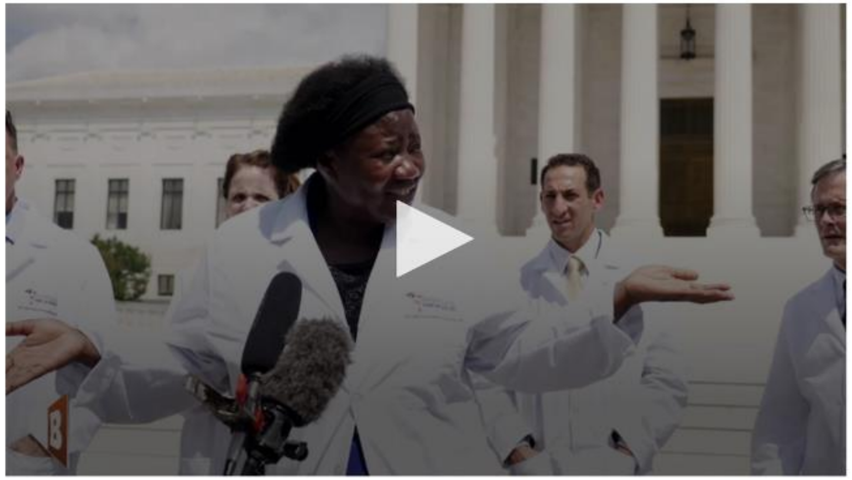
Vero E6 cells (an African green monkey kidney cell line) were infected with **SARS-CoV (Urbani strain)** at a multiplicity of infection of 0.5 for 1 h.

Misinformation ~~Based on~~ Misrepresents Scientific Studies

WATCH— TEXAS DOCTOR: STUDIES CLAIMING HYDROXYCHLOROQUINE DOES NOT WORK ARE 'FAKE SCIENCE'

1,419 EMAIL SHARE TWEET

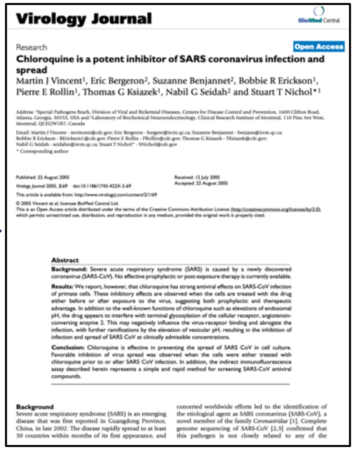
Video Source: Matt Perdie



by AMY FURR | 27 Jul 2020 | 257

This virus has a cure. It is called hydroxychloroquine, zinc, and Zithromax. I know you people want to talk about a mask. Hello? You don't need mask. There is a cure.

The study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that say it works.


Virology Journal

BioMed Central

Research

Open Access

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent[1], Eric Bergeron[2], Suzanne Benjannet[2], Bobbie R Erickson[1], Pierre E Rollin[1], Thomas G Ksiazek[1], Nabil G Seidah[2] and Stuart T Nichol*[1]

Published: 22 August 2005

Virology Journal 2005, 2:69 doi:10.1186/1743-422X-2-69

Received: 12 July 2005

Accepted: 22 August 2005

Abstract

Background: Severe acute respiratory syndrome (SARS) is caused by a newly discovered coronavirus (SARS-CoV). No effective prophylactic or post-exposure therapy is currently available.

Post-infection chloroquine treatment reduces SARS-CoV infection and spread.

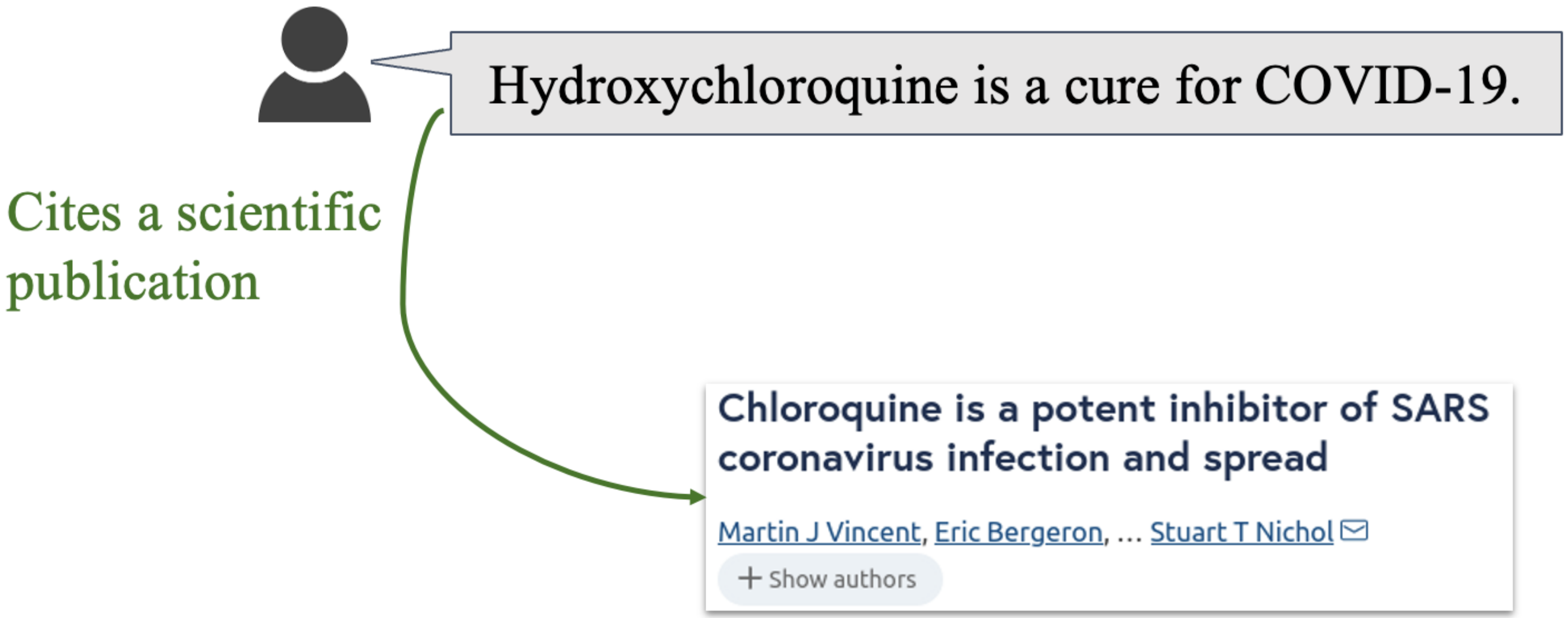
But

It is impossible to infer that a drug will work as a COVID-19 cure in a living person from an *in vitro* cell culture study on a different virus.

2. SARS-CoV-1 ! = SARS-CoV-2 (causes Covid-19)

Vero E6 cells (an African green monkey kidney cell line) were infected with **SARS-CoV (Urbani strain)** at a multiplicity of infection of 0.5 for 1 h.

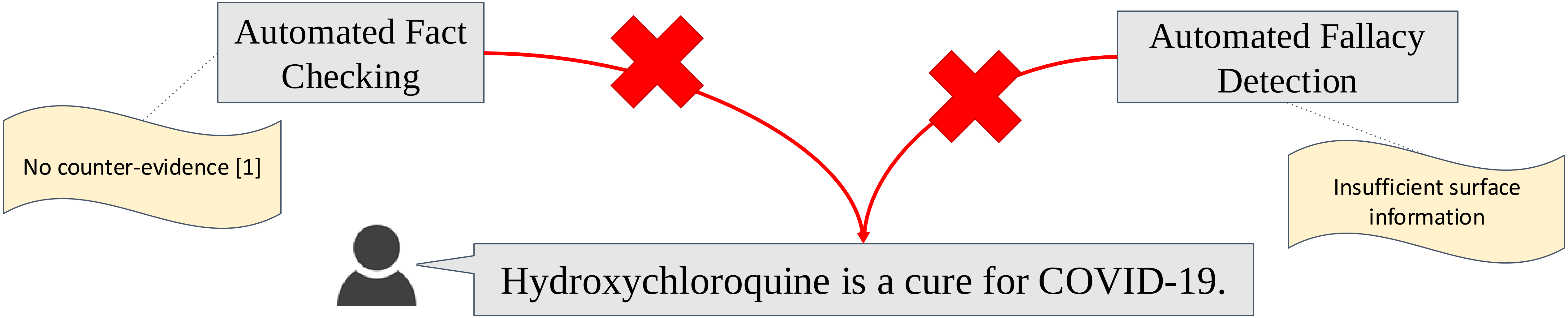
Misinformation ~~Based on~~ Misrepresents Scientific Studies



Research Question

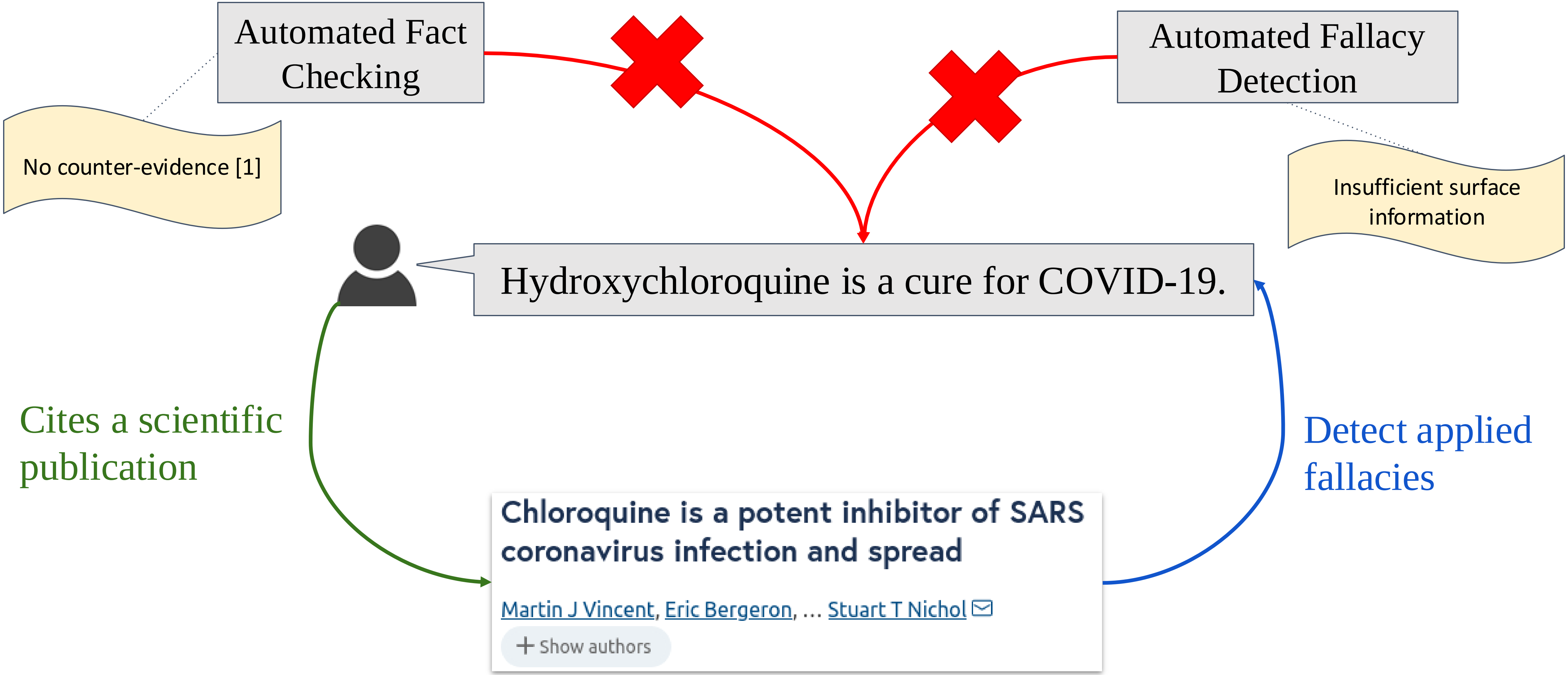
- ✓ Why is this misinformation?
- ✓ And how does it misrepresent the cited scientific study?
- ✓ Can LLMs detect such contradicts: citated scientific study $\nRightarrow$ false claim?

We Need to Assess a Claim Based on its Sources



[1] Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. Glockner et al., EMNLP

We Need to Assess a Claim Based on its Sources



[1] Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. Glockner et al., EMNLP

We Propose to Reconstruct the Fallacious Arguments

Hydroxychloroquine is a cure for COVID-19.

Support

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent, Eric Bergeron, … Stuart T Nichol

+ Show authors

We Propose to Reconstruct the Fallacious Arguments

Input

Hydroxychloroquine is a cure for COVID-19.

Support

Input

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Paraphrased content of the misrepresented publication based on articles written by human fact checkers

We Propose to Reconstruct the Fallacious Arguments

Input

Hydroxychloroquine is a cure for COVID-19.

Support

Input

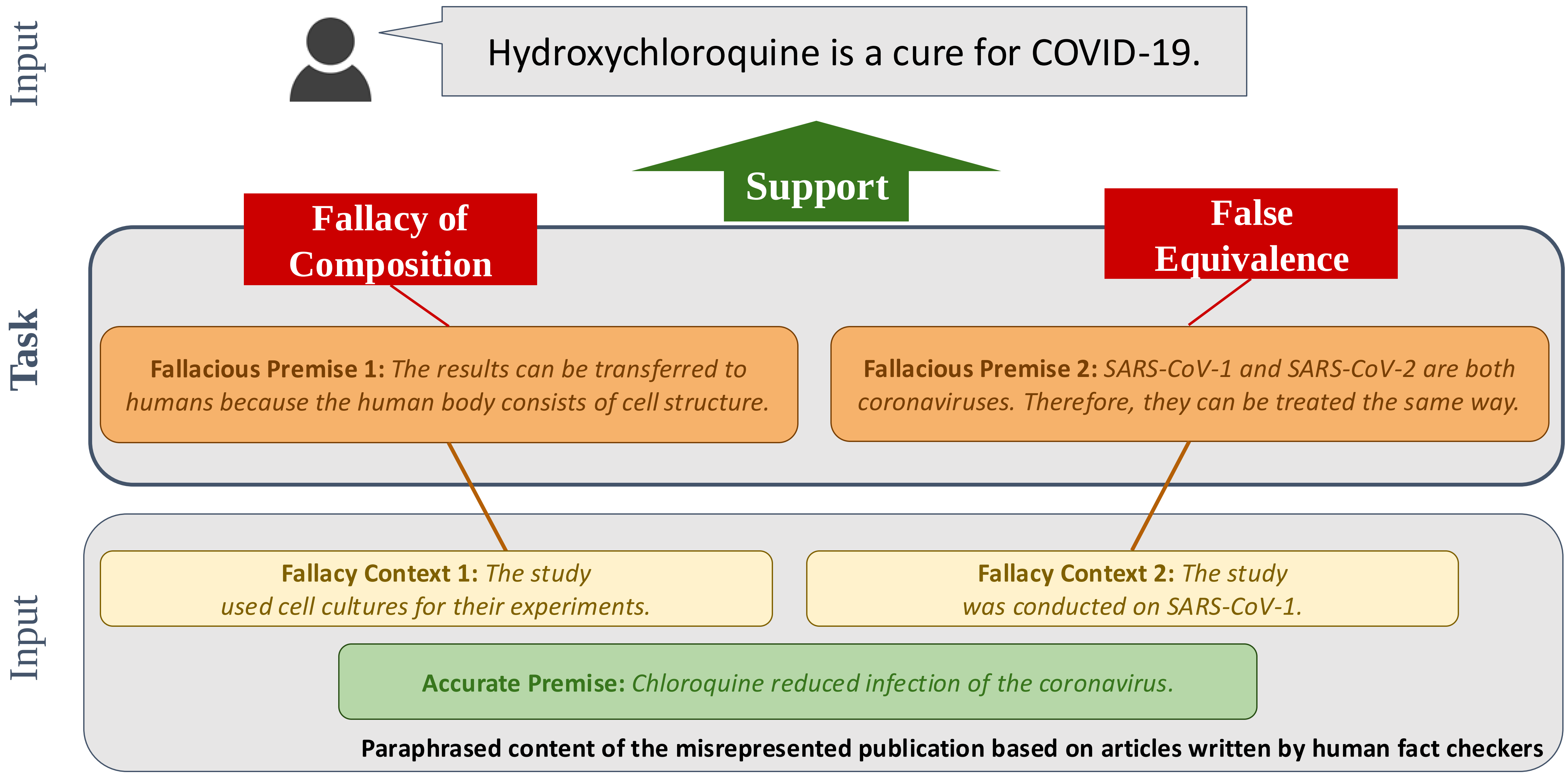
Fallacy Context 1: *The study used cell cultures for their experiments.*

Fallacy Context 2: *The study was conducted on SARS-CoV-1.*

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Paraphrased content of the misrepresented publication based on articles written by human fact checkers

We Propose to Reconstruct the Fallacious Arguments



We Create MISSCI Based on Fact-checking Articles

Collect

Rely on **expert-written fact-checking articles.**

- 8,695 linked documents in
- 527 fact-checking articles

Health Feedback

Three annotators: two M.Sc. student in biology, one M.Sc. in linguistics

We Create MISSCI Based on Fact-checking Articles

Collect

Rely on **expert-written fact-checking articles.**

- 8,695 linked documents in
- 527 fact-checking articles

Health Feedback

Select

Manually identify all cases in which a **scientific publication is misrepresented**.

- 208 links to misrepresented scientific publications
- In 150 fact-checking articles

Three annotators: two M.Sc. student in biology, one M.Sc. in linguistics

We Create MISSCI Based on Fact-checking Articles

- Annotation stage I: selecting misrepresented scientific publications from HFC articles


Health Feedback Article Reviews Claim Reviews Insights Support us

We've Moved Visit our new website at Science.Feedback.org for scientific verifications of viral claims.

Blog posts inaccurately claim that a 2005 NIH study demonstrated the effectiveness of chloroquine treatment against coronavirus infection such as COVID-19

7.9k SHARES Share Tweet

CLAIM

the NIH researched chloroquine and concluded that it was effective at stopping the SARS coronavirus in its tracks

VERDICT

INCORRECT

SOURCE: Bryan Fischer, OneNewsNow, 17 April 2020

> Several outlets claimed that Anthony Fauci, who has served as director of the U.S. National Institute of Allergy and Infectious Diseases (NIAID) since 1984 and is also a member of the White House Coronavirus Task Force, knew for 15 years that chloroquine would work on coronaviruses, and by extension COVID-19 (see examples of these articles here, here, and here). This claim is primarily based on an *in vitro* scientific study published in 2005, which examined how well chloroquine protected cells growing in petri dishes (cell culture) against SARS-CoV-1 infection[1]. SARS-CoV-1 is the virus responsible for the 2003-2005 SARS outbreak. These outlets also further assert that the study was published by the U.S. National Institutes of Health (NIH). These articles have gone viral on Facebook (see here, here, and here).

Annotation task: Is this a scientific paper that was misrepresented by a non-true claim?

We Create MISSCI Based on Fact-checking Articles

Collect

Rely on **expert-written fact-checking articles.**

- 8,695 linked documents in
- 527 fact-checking articles

Health Feedback

Select

Manually identify all cases in which a **scientific publication is misrepresented**.

- 208 links to misrepresented scientific publications
- In 150 fact-checking articles

Reconstruct

Manually reconstruct the fallacious arguments guided by the fact-checking article.

- 184 arguments
- 435 necessary fallacious reasoning steps

Three annotators: two M.Sc. student in biology, one M.Sc. in linguistics

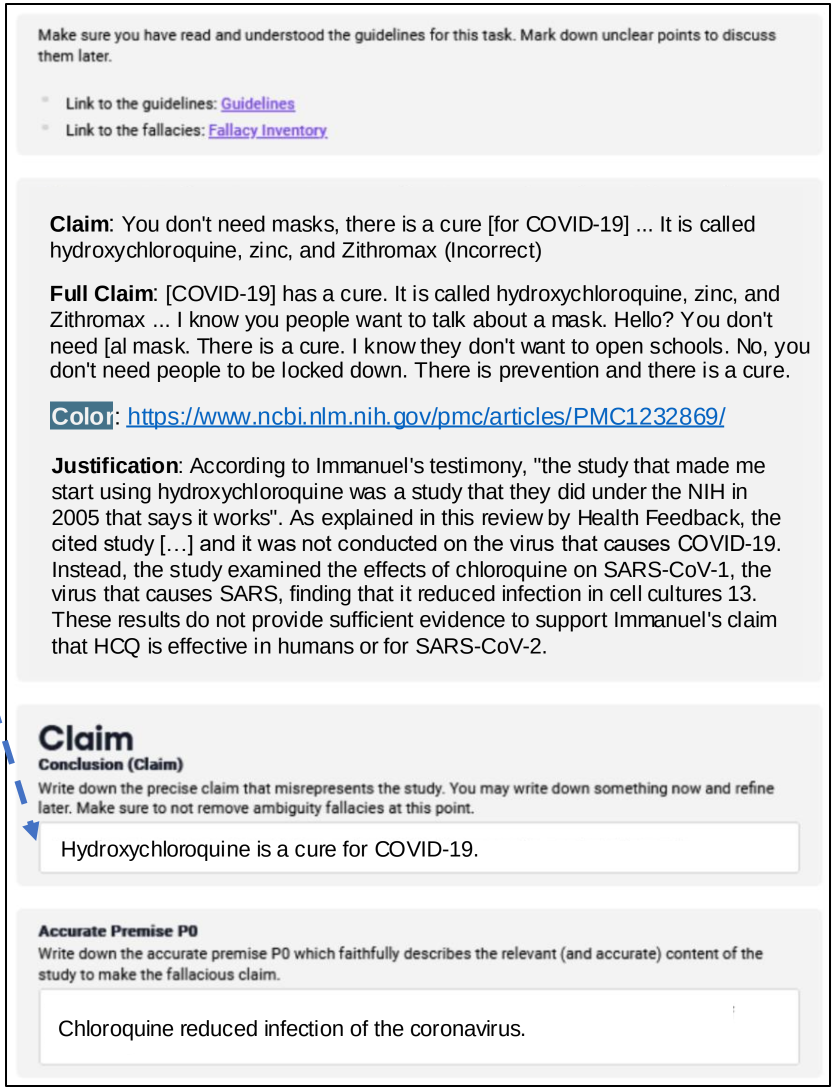
We Create MISSCI Based on Fact-checking Articles

❑ Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\bar{c}$

✓ Annotators should use the main (false) claim of the fact-checking article if possible and make minimal changes if necessary.

False claim $\bar{c}$: Hydroxychloroquine is a cure for COVID-19.



Make sure you have read and understood the guidelines for this task. Mark down unclear points to discuss them later.

- Link to the guidelines: Guidelines
- Link to the fallacies: Fallacy Inventory

Claim: You don't need masks, there is a cure [for COVID-19] ... It is called hydroxychloroquine, zinc, and Zithromax (Incorrect)

Full Claim: [COVID-19] has a cure. It is called hydroxychloroquine, zinc, and Zithromax ... I know you people want to talk about a mask. Hello? You don't need [a] mask. There is a cure. I know they don't want to open schools. No, you don't need people to be locked down. There is prevention and there is a cure.

Color: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1232869/

Justification: According to Immanuel's testimony, "the study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that says it works". As explained in this review by Health Feedback, the cited study [...] and it was not conducted on the virus that causes COVID-19. Instead, the study examined the effects of chloroquine on SARS-CoV-1, the virus that causes SARS, finding that it reduced infection in cell cultures 13. These results do not provide sufficient evidence to support Immanuel's claim that HCQ is effective in humans or for SARS-CoV-2.

Claim

Conclusion (Claim)

Write down the precise claim that misrepresents the study. You may write down something now and refine later. Make sure to not remove ambiguity fallacies at this point.

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise P0

Write down the accurate premise P0 which faithfully describes the relevant (and accurate) content of the study to make the fallacious claim.

Chloroquine reduced infection of the coronavirus.

We Create MISSCI Based on Fact-checking Articles

❑ Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\bar{c}$

Step 2: accurate premise writing → p_0

✓ The accurate premise p_0 provides a correct description of the misrepresented scientific document S

✓ p_0 offers logical support for the false claim ($p_0 \Rightarrow \bar{c}$) but it falls short due to the presence of fallacious reasoning

Accurate Premise p_0: *Chloroquine reduced infection of the coronavirus.*

Make sure you have read and understood the guidelines for this task. Mark down unclear points to discuss them later.

- Link to the guidelines: Guidelines
- Link to the fallacies: Fallacy Inventory

Claim: You don't need masks, there is a cure [for COVID-19] ... It is called hydroxychloroquine, zinc, and Zithromax (Incorrect)

Full Claim: [COVID-19] has a cure. It is called hydroxychloroquine, zinc, and Zithromax ... I know you people want to talk about a mask. Hello? You don't need [al mask. There is a cure. I know they don't want to open schools. No, you don't need people to be locked down. There is prevention and there is a cure.

Color: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1232869/

Justification: According to Immanuel's testimony, "the study that made me start using hydroxychloroquine was a study that they did under the NIH in 2005 that says it works". As explained in this review by Health Feedback, the cited study […] and it was not conducted on the virus that causes COVID-19. Instead, the study examined the effects of chloroquine on SARS-CoV-1, the virus that causes SARS, finding that it reduced infection in cell cultures 13. These results do not provide sufficient evidence to support Immanuel's claim that HCQ is effective in humans or for SARS-CoV-2.

Claim

Conclusion (Claim)

Write down the precise claim that misrepresents the study. You may write down something now and refine later. Make sure to not remove ambiguity fallacies at this point.

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise P0

Write down the accurate premise P0 which faithfully describes the relevant (and accurate) content of the study to make the fallacious claim.

Chloroquine reduced infection of the coronavirus.

We Create MISSCI Based on Fact-checking Articles

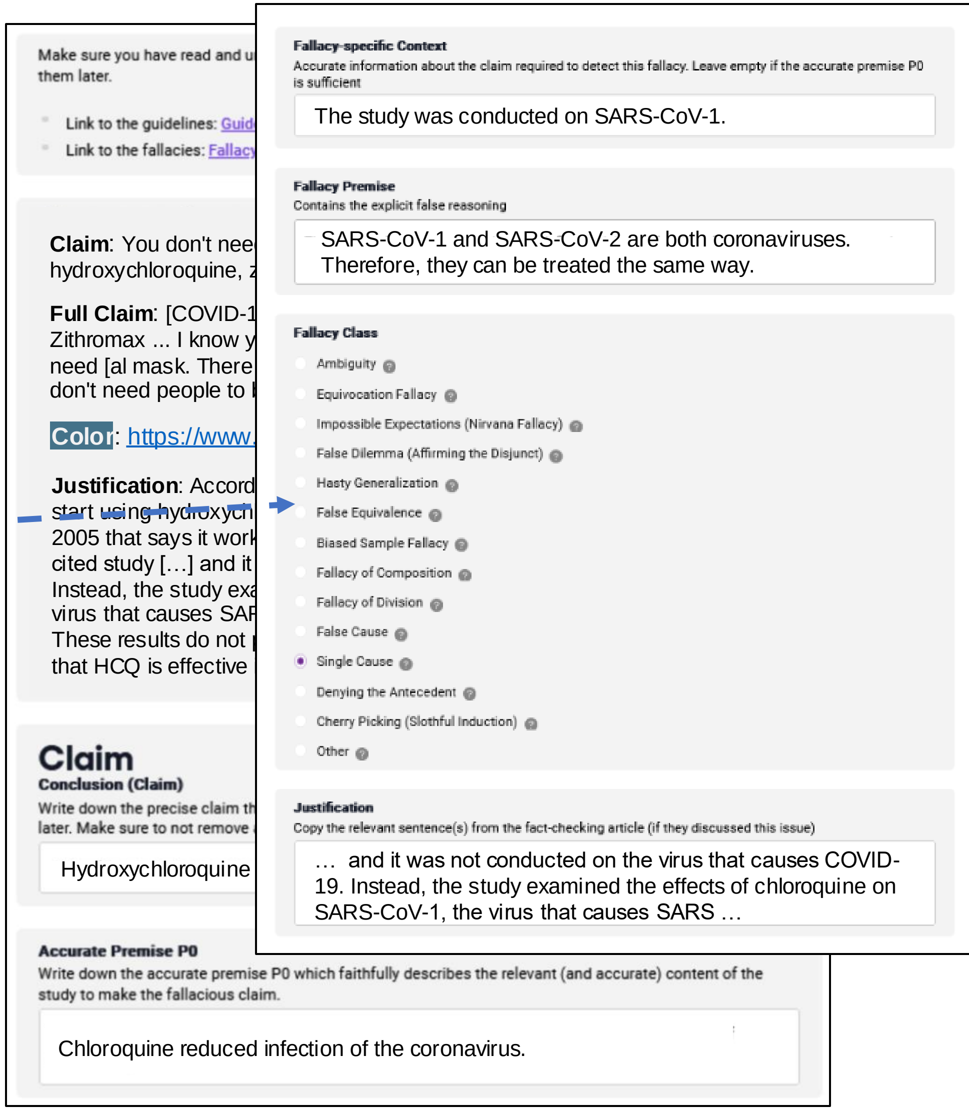
- ❑ Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\overline{c}$

Step 2: accurate premise writing → p_0

Step 3: fallacy class selection → f_i

- ✓ A taxonomy adopted from Bennett (2012) and Cook et al. (2018)
- ✓ 12 fallacy classes in a tree structure to guide the annotators to choose the more specific fallacy class if multiple apply



We Create MISSCI Based on Fact-checking Articles

- Fallacy taxonomy adopted from Bennett (2012) and Cook et al. (2018)

Definition	Logical Form
AMBIGUITY When an unclear phrase with multiple definitions is used within the argument; therefore, does not support the conclusion.	*Claim X is made. Y is concluded based on an ambiguous understanding of X.*
EQUIVOCATION (merged with **AMBIGUITY**) When the same word (here used also for phrase) is used with two different meanings. Equivocation is a subset of the ambiguity fallacy.	*Term X is used to mean Y in the premise. Term X is used to mean Z in the conclusion.*
IMPOSSIBLE EXPECTATIONS / NIRVANA FALLACY Comparing a realistic solution with an idealized one, and discounting or even dismissing the realistic solution as a result of comparing to a "perfect world" or impossible standard, ignoring the fact that improvements are often good enough reason.	*X is what we have. Y is the perfect situation. Therefore, X is not good enough.*
FALSE EQUIVALENCE Assumes that two subjects that share a single trait are equivalent.	*X and Y both share characteristic A. Therefore, X and Y are [behave] equal.*
FALSE DILEMMA Presents only two alternatives, while there may be another alternative, another way of framing the situation, or both options may be simultaneously viable.	*Either X or Y is true.*
BIASED SAMPLE FALLACY Drawing a conclusion about a population based on a sample that is biased, or chosen in order to make it appear the population on average is different than it actually is.	*Sample S, which is biased, is taken from population P. Conclusion C is drawn about population P based on S.*

Definition	Logical Form
HASTY GENERALIZATION Drawing a conclusion based on a small sample size, rather than looking at statistics that are much more in line with the typical or average situation.	*Sample S is taken from population P. Sample S is a very small part of population P. Conclusion C is drawn from sample S and applied to population P.*
FALSE CAUSE FALLACY (use as **CAUSAL SIMPLIFICATION**) Post hoc ergo propter hoc — after this therefore because of this. Automatically attributes causality to a sequence or conjunction of events.	*A is regularly associated with B; therefore, A causes B.*
SINGLE CAUSE FALLACY (use as **CAUSAL SIMPLIFICATION**) Assumes there is a single, simple cause of an outcome.	*X is a contributing factor to Y. X and Y are present. Therefore, to remove Y, remove X.*
FALLACY OF COMPOSITION Inferring that something is true of the whole from the fact that it is true of some part of the whole.	*A is part of B. A has property X. Therefore, B has property X.*
FALLACY OF DIVISION (merged with **FALLACY OF COMPOSITION**) Inferring that something is true of one or more of the parts from the fact that it is true of the whole.	*A is part of B. B has property X. Therefore, A has property X.*
FALLACY OF EXCLUSION / CHERRY PICKING / SLOTHFUL INDUCTION When only select evidence is presented in order to persuade the audience to accept a position, and evidence that would go against the position is withheld (Cherry Picking). Ignores relevant and significant evidence when inferring to a conclusion (Slothful Induction – focus on neglect).	*Evidence A and evidence B is available. Evidence A supports the claim of person 1. Evidence B supports the counterclaim of person 2. Therefore, person 1 presents only evidence A.*

We Create MISSCI Based on Fact-checking Articles

- ❑ Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\bar{c}$

Step 2: accurate premise writing → p_0

Step 3: fallacy class selection → f_i

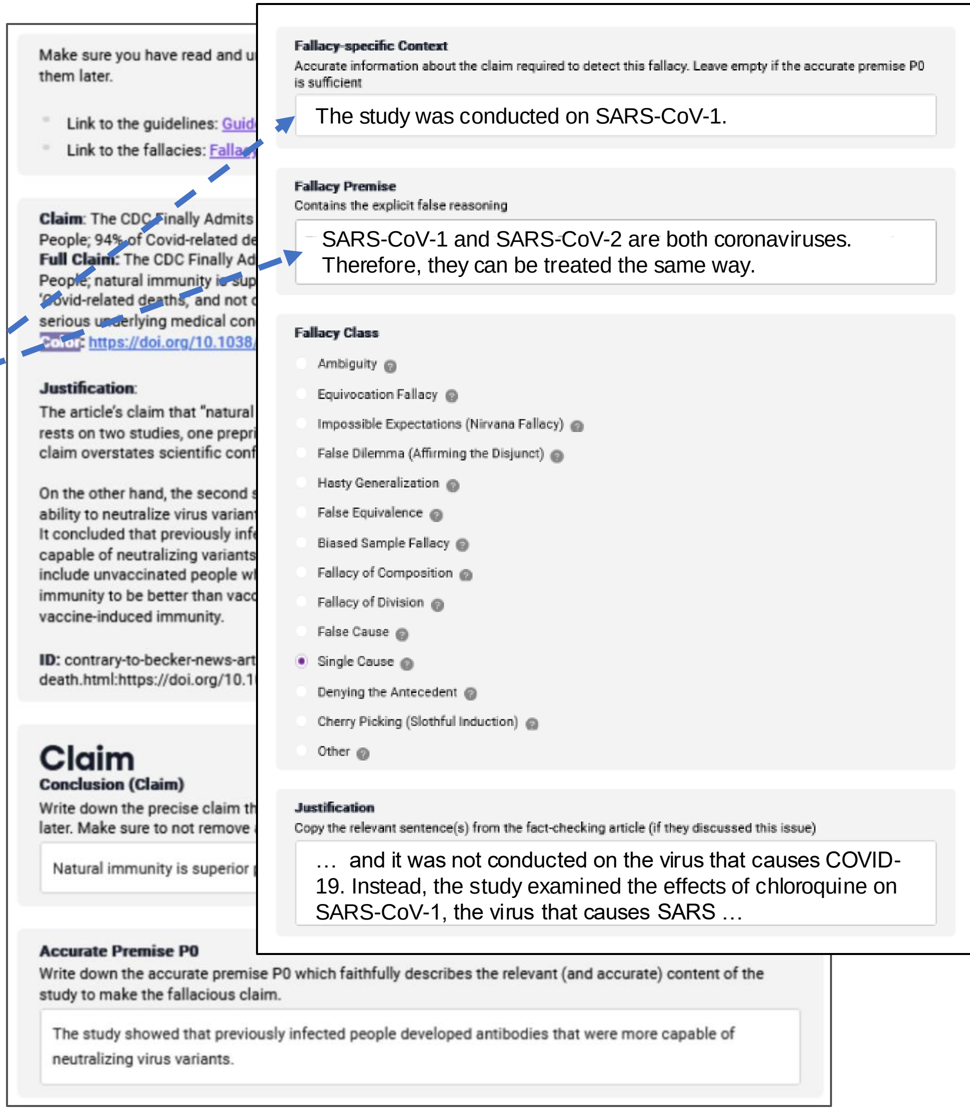
Step 4: fallacious premise + publication context writing → $\bar{p}_i$ and $\bar{s}_i$

- ✓ Identify all passages in HFC articles discussing the claim $\bar{c}$ misrepresenting the scientific publication S
- ✓ The fallacious premise must align with the selected fallacy class and make the fallacious reasoning explicit

$$S \cup \bar{p}_i \Rightarrow \bar{c}$$

Fallacy of Composition

Fallacious Premise 1: *The results can be transferred to humans because the human body consists of cell structure.*



We Create MISSCI Based on Fact-checking Articles

- ❑ Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\bar{c}$

Step 2: accurate premise writing → p_0

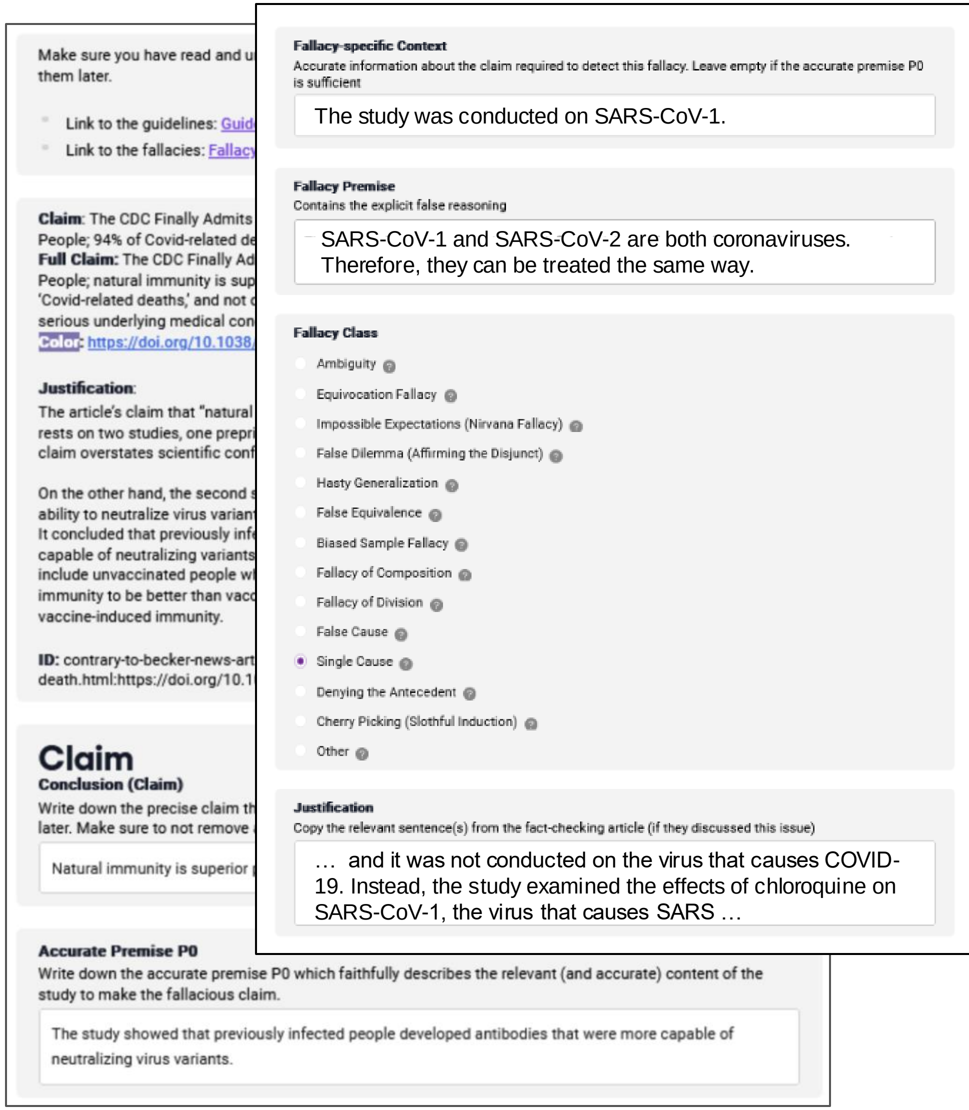
Step 3: fallacy class selection → f_i

Step 4: fallacious premise + publication context writing → $\bar{p}_i$ and $\bar{s}_i$

Step 5: argument consolidation

✓ The most experienced annotator aligned all annotated fallacious reasoning lines, select the best verbalized candidate for each $\bar{c}$ and $\bar{p}_\mathrm{i}$

✓ Each consolidated argument underwent double-checking by an author



We Create MISSCI Based on Fact-checking Articles

- Annotation stage II: fallacious argument reconstruction based on HFC articles

Step 1: false claim rewriting → $\bar{c}$

Step 2: accurate premise writing → p_0

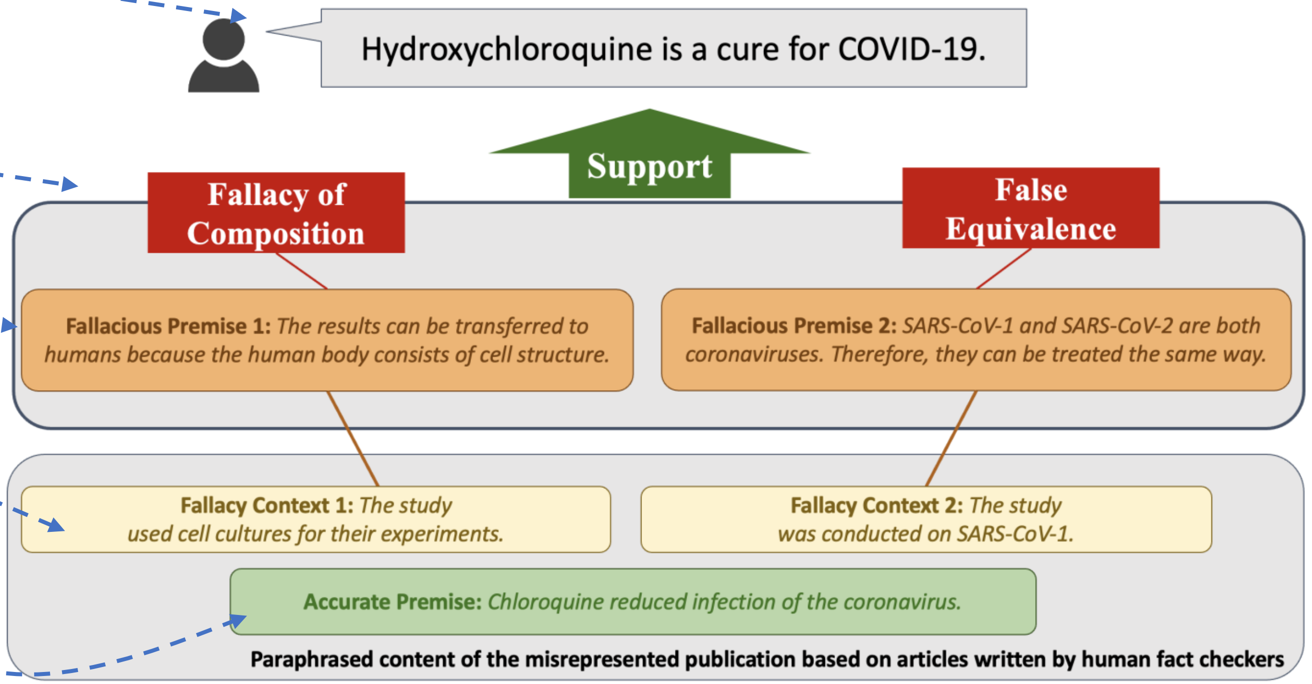
Step 3: fallacy class selection → f_i

Step 4: publication context and fallacious premise writing → $\bar{s}_i$ and $\bar{p}_i$

Step 5: argument consolidation

Fallacious argument definition

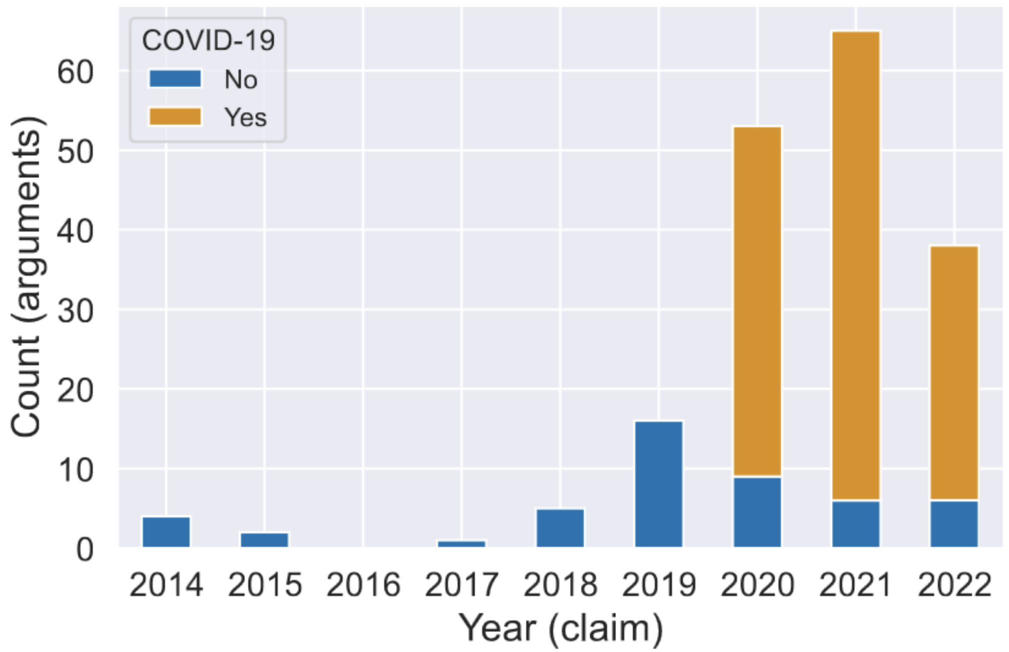
$$\left\{ \overset{s_0}{\underset{}{\overset{=}{p_0}}},\ \overset{s_1}{\underset{f_1}{\overset{\downarrow}{\underset{\downarrow}{\bar{p}_1}}}},\ \overset{s_2}{\underset{f_2}{\overset{\downarrow}{\underset{\downarrow}{\bar{p}_2}}}},\ \ldots,\ \overset{s_N}{\underset{f_N}{\overset{\downarrow}{\underset{\downarrow}{\bar{p}_N}}}} \right\} \Rightarrow \bar{c}$$



We Create MISSCI Based on Fact-checking Articles

❑ MISSCI dataset construction

	Collect	Select	Reconstruct
HFC articles	527	150	147
Links	8,695	208	184
Arguments	–	–	184
Fall. Reasoning R_i	–	–	435

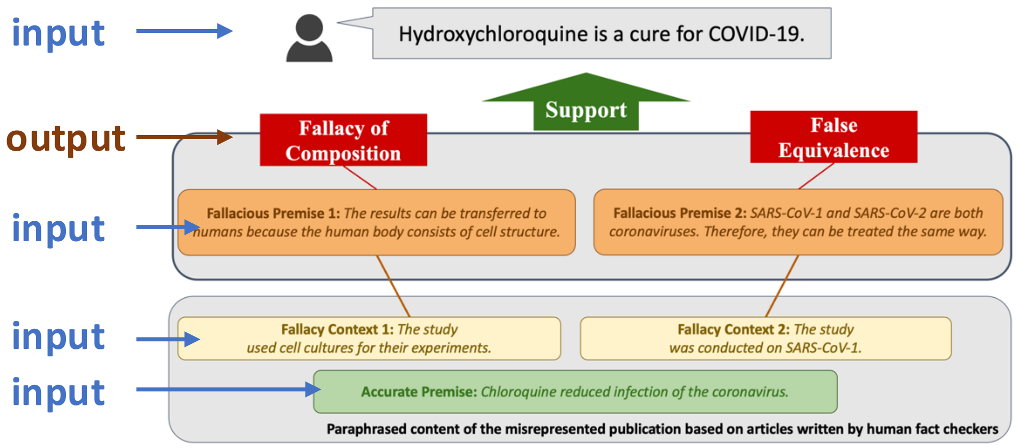


- ✓ Krippendorff's α is 0.52 for assigning fallacious class f_i
- ✓ On average, each annotator identified 72.5% of the fallacious reasoning lines in the consolidated argument

LLMs Can Predict the Fallacy Class Over Provided Premises

Simplified Task:
Predict the applied fallacy class when the fallacious premise is provided.

Explore prompts containing:
Definition, Logical Form, Example



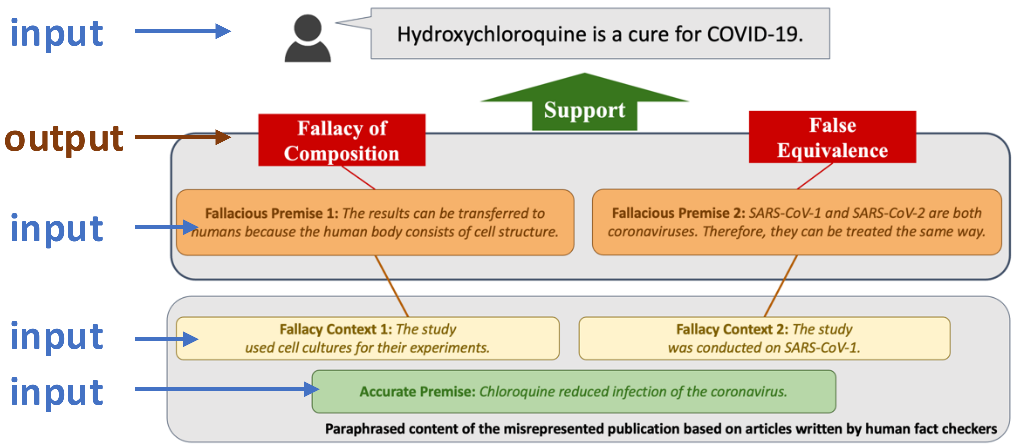
LLMs Can Predict the Fallacy Class Over Provided Premises

Simplified Task:

Predict the applied fallacy class when the fallacious premise is provided.

Explore prompts containing:

Definition, Logical Form, Example



Example: Fallacy of Composition

Definition:
Inferring that something is true of the whole from the fact that it is true of some part of the whole.

Logical Form:
A is part of B. A has property X. Therefore, B has property X.

Example:
Hydrogen is not wet. Oxygen is not wet. Therefore, water (H2O) is not wet.

LLMs Can Predict the Fallacy Class Over Provided Premises

Simplified Task:
Predict the applied fallacy class when the fallacious premise is provided.

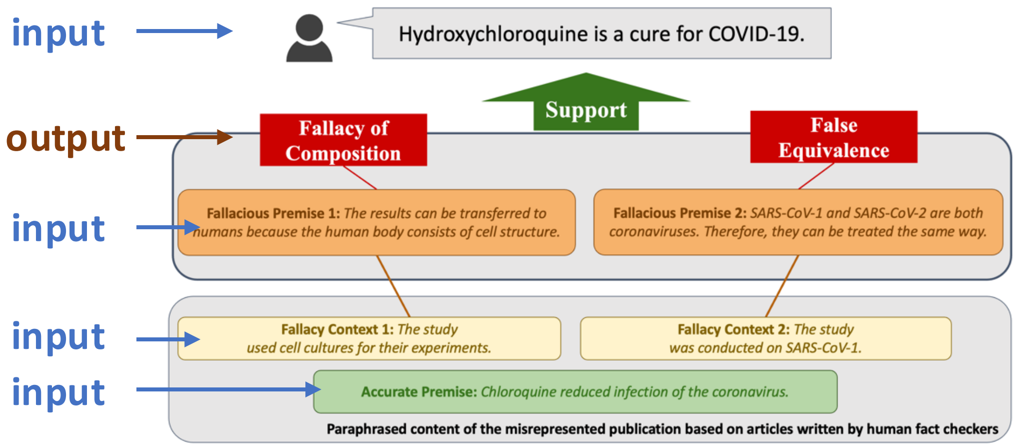
Explore prompts containing:
Definition, Logical Form, Example

Example: Fallacy of Composition

Definition:
Inferring that something is true of the whole from the fact that it is true of some part of the whole.

Logical Form:
A is part of B. A has property X. Therefore, B has property X.

Example:
Hydrogen is not wet. Oxygen is not wet. Therefore, water (H2O) is not wet.

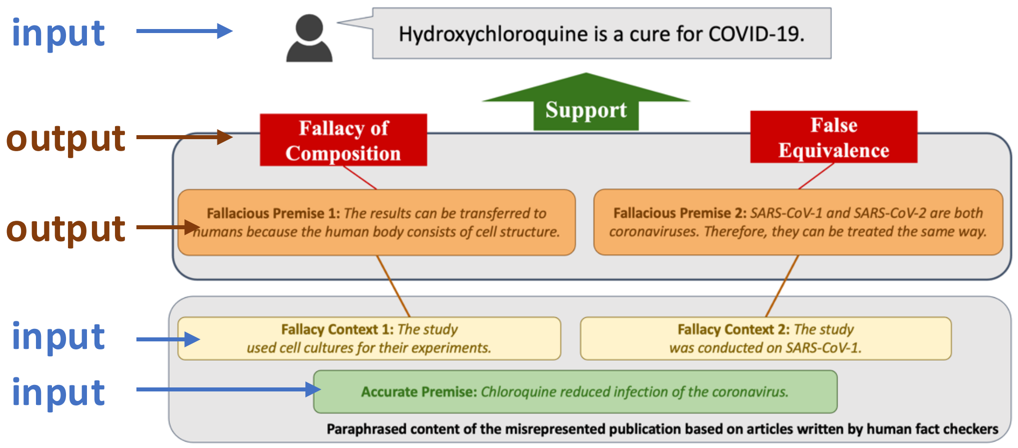


LLM	Prompt	Acc.	F1
LLaMA 2	–	0.493	0.406
	Def.	0.577	0.464
	Def. + Logical	0.630	0.476
	Def. + Example	0.637	0.476
	Def. + Logical + Example	0.568	0.459
	Logical	0.601	0.472
	Logical + Example	<u>0.645</u>	<u>0.499</u>
GPT 4	Def.	0.738	0.649
	Logical	0.744	0.624
	Logical + Example	**0.771**	**0.682**

Both evaluated LLMs perform decently

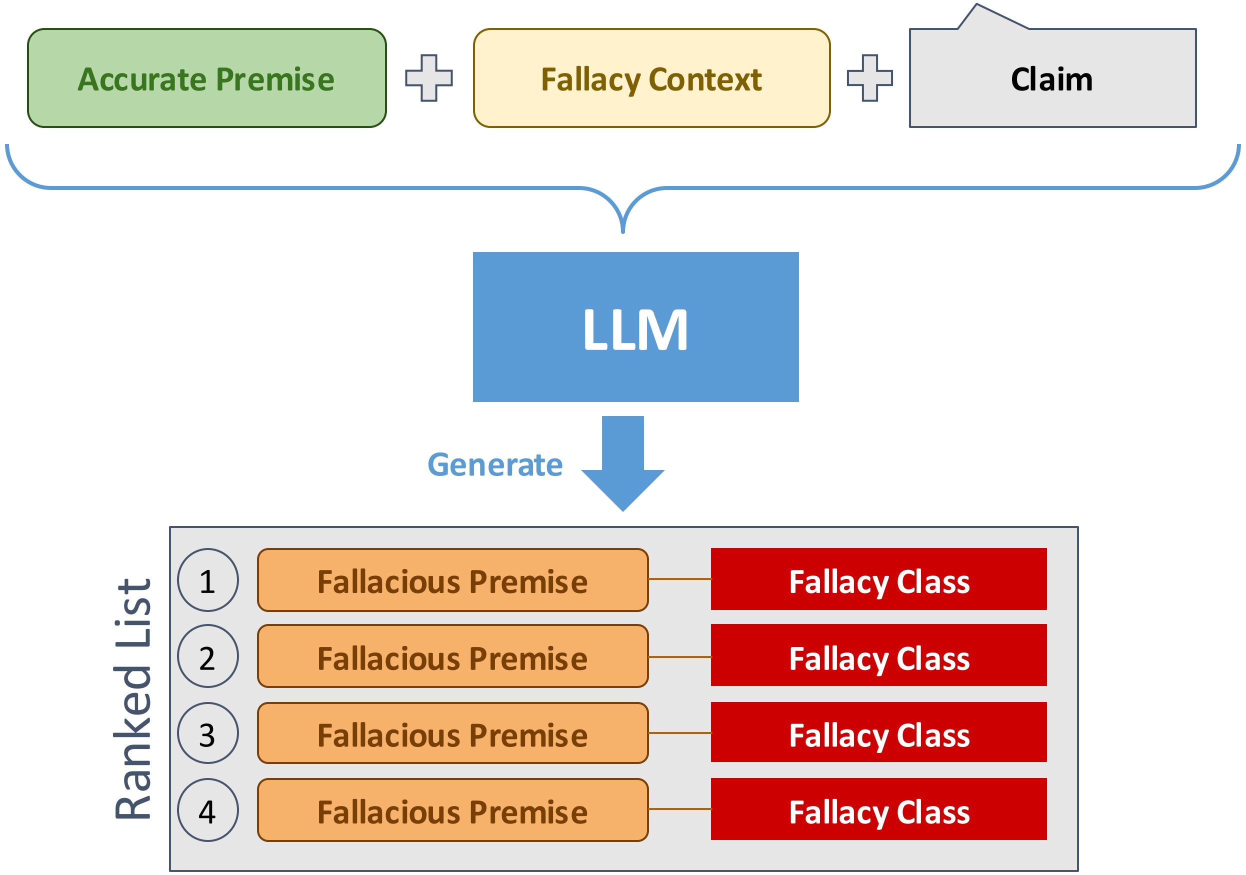
LLMs Perform Poorly When They Must Generate Premises

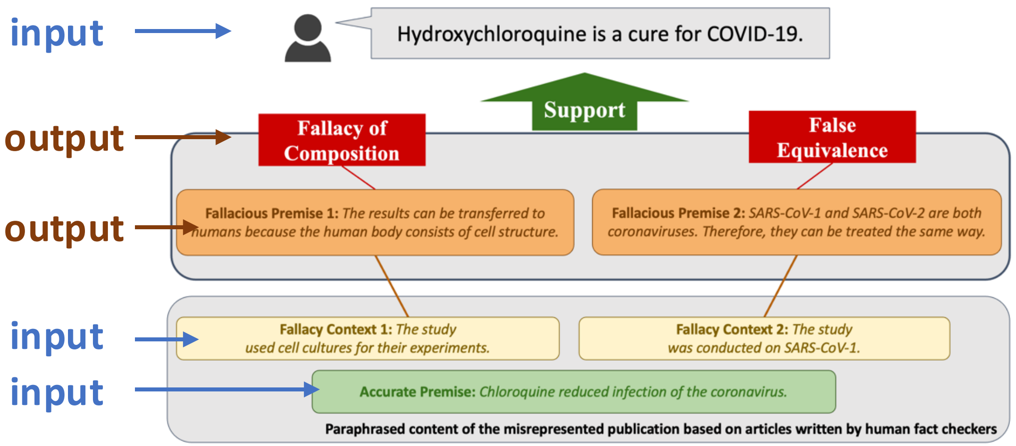
Full Task:
Generate fallacious premise and predict applied fallacy class.



LLMs Perform Poorly When They Must Generate Premises

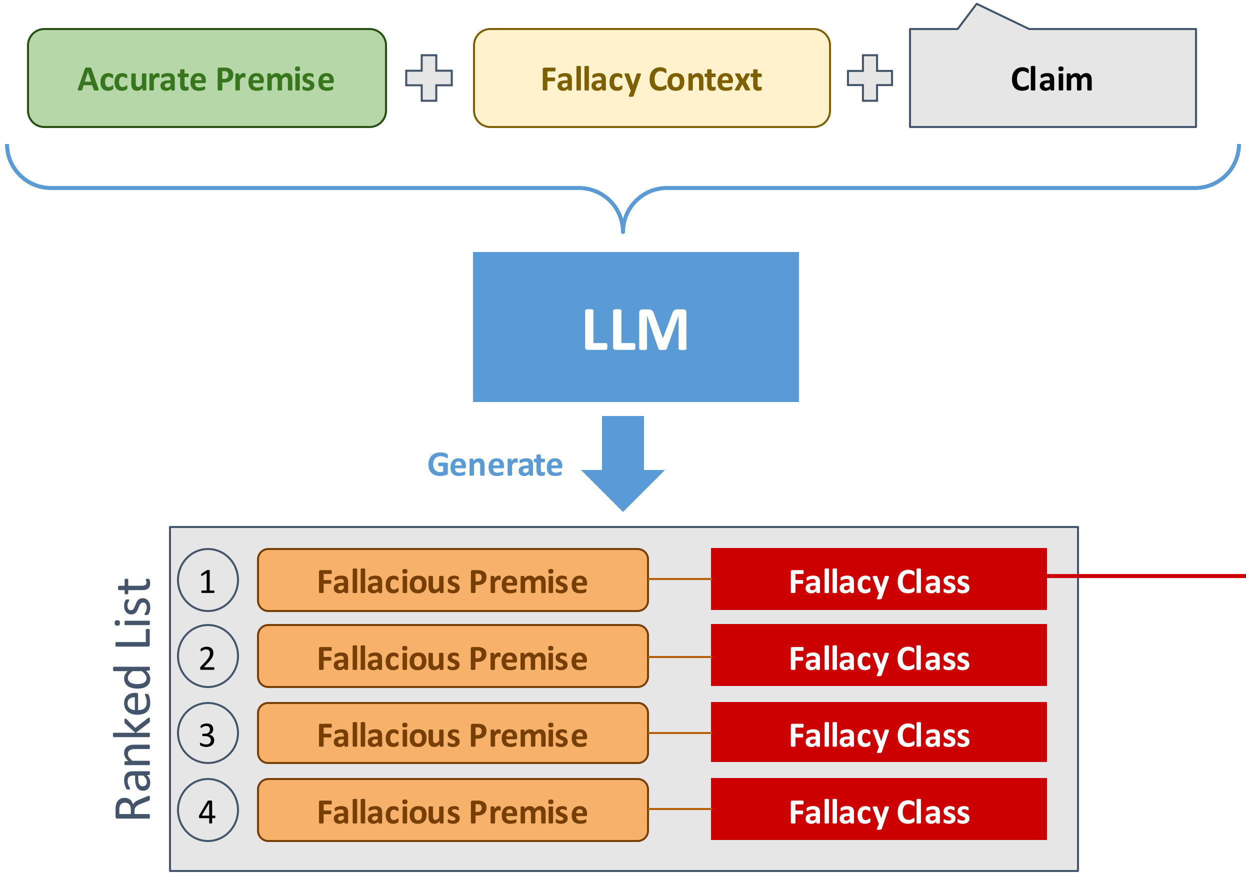
Full Task:
Generate fallacious premise and predict applied fallacy class.





LLMs Perform Poorly When They Must Generate Premises

Full Task:
Generate fallacious premise and predict applied fallacy class.



Is claim debunked by at least **one correct fallacy** (from any of the fallacy contexts)?

LLM	P@1	Claim@1
random	0.131	0.264
Llama2 (D)	0.223	0.416
Llama2 (DE)	0.209	0.422
Llama2 (DL)	0.196	0.409
Llama2 (DLE)	0.209	0.416
Llama2 (L)	0.193	0.377
Llama2 (LE)	0.202	0.409
GPT-4 (D)	**0.317**	**0.571**
GPT-4 (L)	0.292	0.526

Was **0.738** (accuracy) over **gold fallacious premise**

Automatic Evaluation Underestimates the Performance

Human Evaluation. Correct If:

1. Plausible Premise: *Is the generated premise plausible in the context of the argument?*
2. Correct Fallacy Class: *Is the predicted fallacy class applied by the generated fallacious premise?*

Automatic Evaluation Underestimates the Performance

Human Evaluation. Correct If:

1. Plausible Premise: *Is the generated premise plausible in the context of the argument?*
2. Correct Fallacy Class: *Is the predicted fallacy class applied by the generated fallacious premise?*

LLM	Plausible Premise	Correct Fallacy class
Llama2 (L)	0.167	0.040
Llama2 (D)	0.233	0.107
GPT-4 (L)	**0.867**	**0.503**
GPT-4 (D)	0.674	0.481

Was **0.292** (P@1) over **generated fallacious premise that can be mapped to the annotations**

Claim: To protect from COVID-19 we must back away from all climate change efforts.

Generated Premise: *Efforts to combat climate change will result in warmer average temperatures*, therefore decreasing the prevalence of COVID-19.

COVID-19 transmission correlates with cold temperatures

LLM may detect valid fallacies that annotators missed

Human evaluation is necessary

GPT-4 Benefits From the Premise Generation Task (CoT)

Fallacy classification task

Model	Role of $\overline{p}_i$ Reconstruct	Given	n/a
LLaMA 2 (*D*)	0.223	0.577	0.248
LLaMA 2 (*DE*)	0.209	0.637	0.264
LLaMA 2 (*DL*)	0.196	0.630	0.237
LLaMA 2 (*DLE*)	0.209	0.568	0.259
LLaMA 2 (*LE*)	0.212	0.645	**0.267**
LLaMA 2 (*L*)	0.193	0.601	0.262
GPT 4 (*D*)	**0.317**	0.738	**0.267**
GPT 4 (*L*)	0.292	**0.744**	0.245

Model	Setup	*Matching* $\hat{\overline{p}}_i$ Yes	No
GPT 4 (*D*)	classify f_i and gen. $\overline{p}_i$	**0.880**	0.229
	classify f_i w/o $\overline{p}_i$	0.640	0.114
	classify f_i given $\overline{p}_i$	0.788	**0.689**
GPT 4 (*L*)	classify f_i and gen. $\overline{p}_i$	**0.867**	0.133
	classify f_i w/o $\overline{p}_i$	0.533	0.133
	classify f_i given $\overline{p}_i$	0.732	**0.722**

- ✓ GPT-4 produced substantially better fallacious premises according to our human evaluation
- ✓ The model shows improved performance on fallacy classification when tasked to “think” -- generating fallacious premises

Conclusion

- Novel formalism to combat real-world misinformation
- Novel benchmark to test critical reasoning abilities of LLMs
  - Realistic scenario
  - Sufficient data
- Both LLMs exhibit clear limitations in reconstructing fallacious arguments
- More experiments, results and analysis in the paper!

https://github.com/UKPLab/acl2024-missci

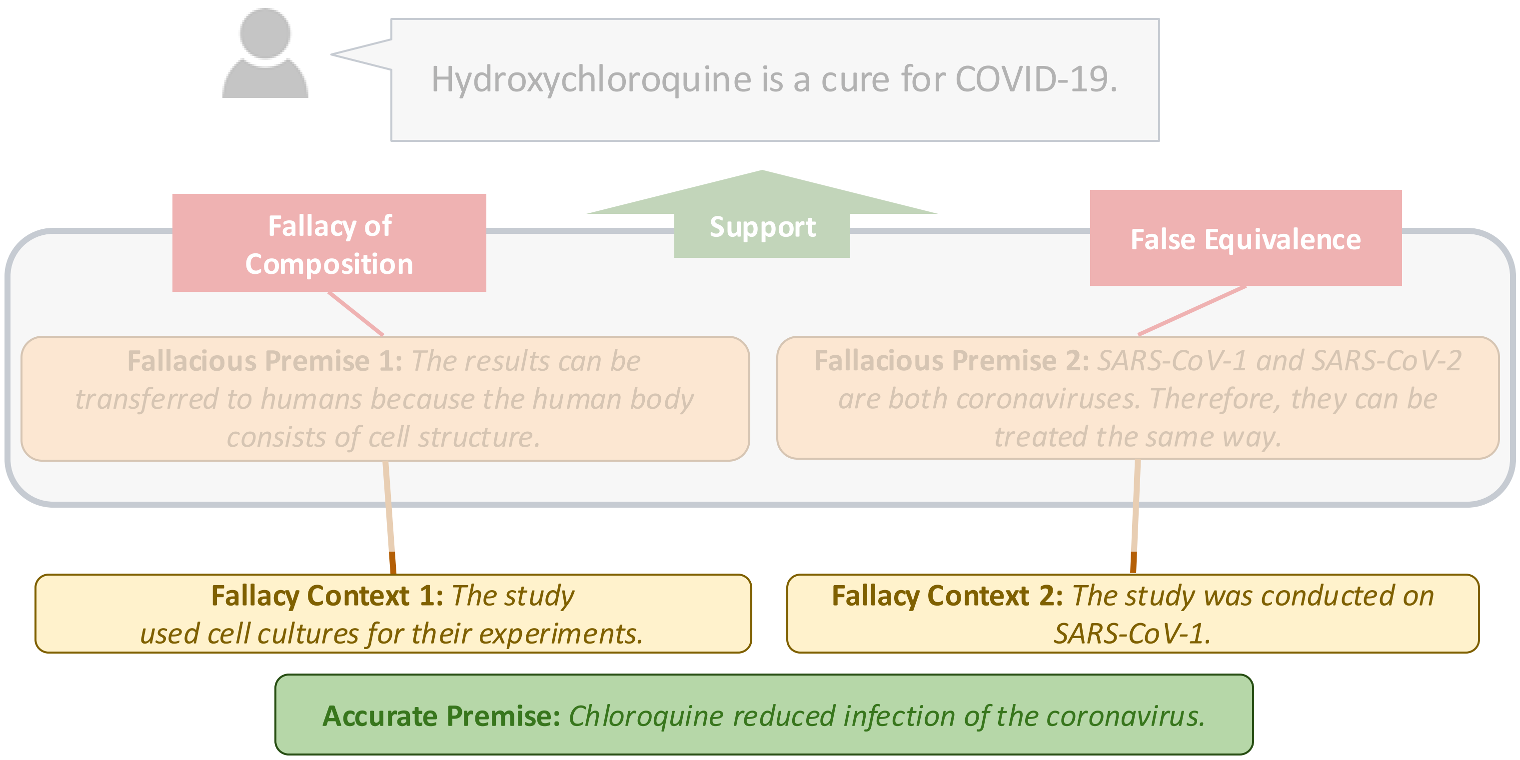
MISSCI+: Grounding Fallacies of Misrepresented Scientific Publications

Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych

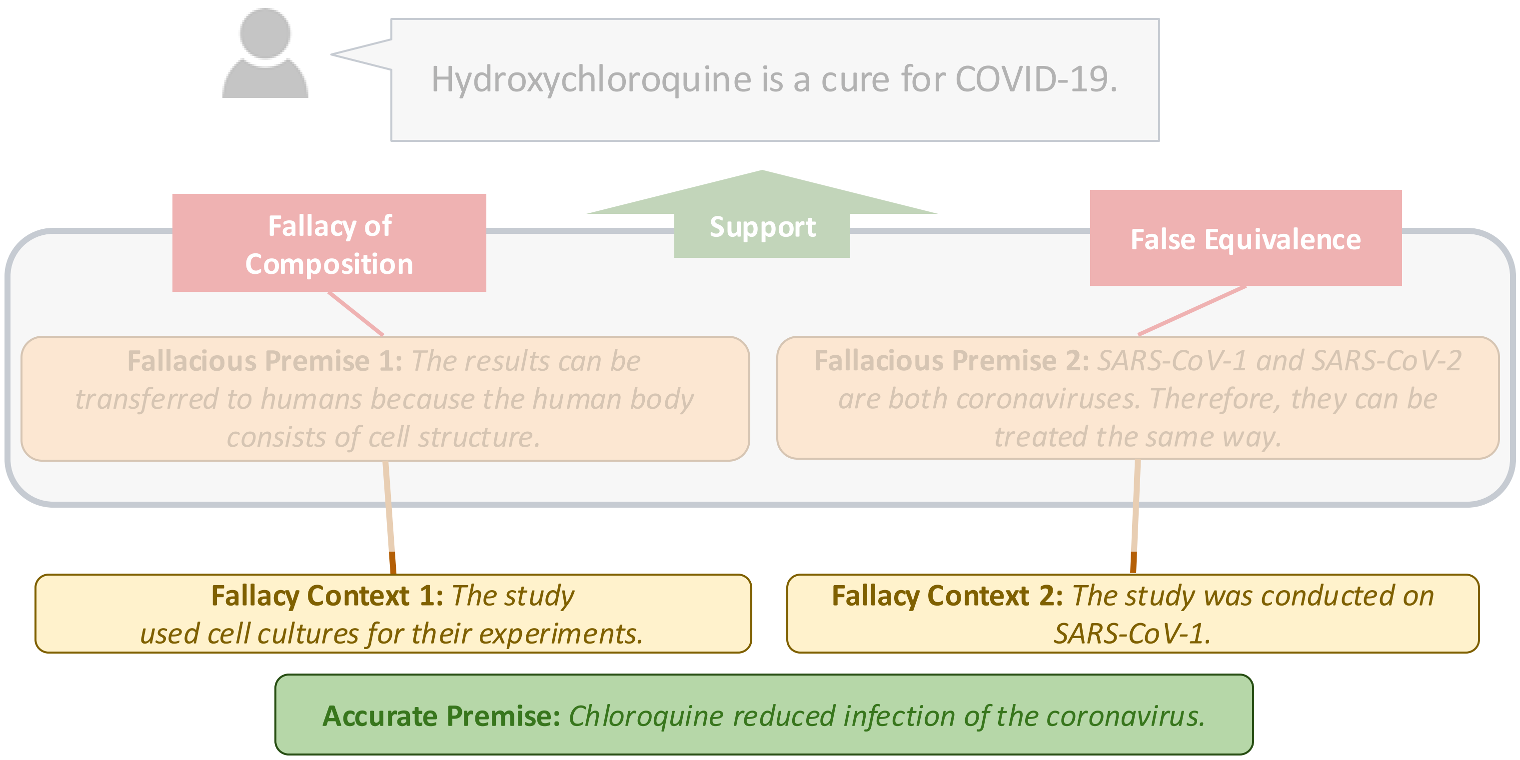
(under submission)




MISSCI Does Not Consider Real-world Passages From Papers



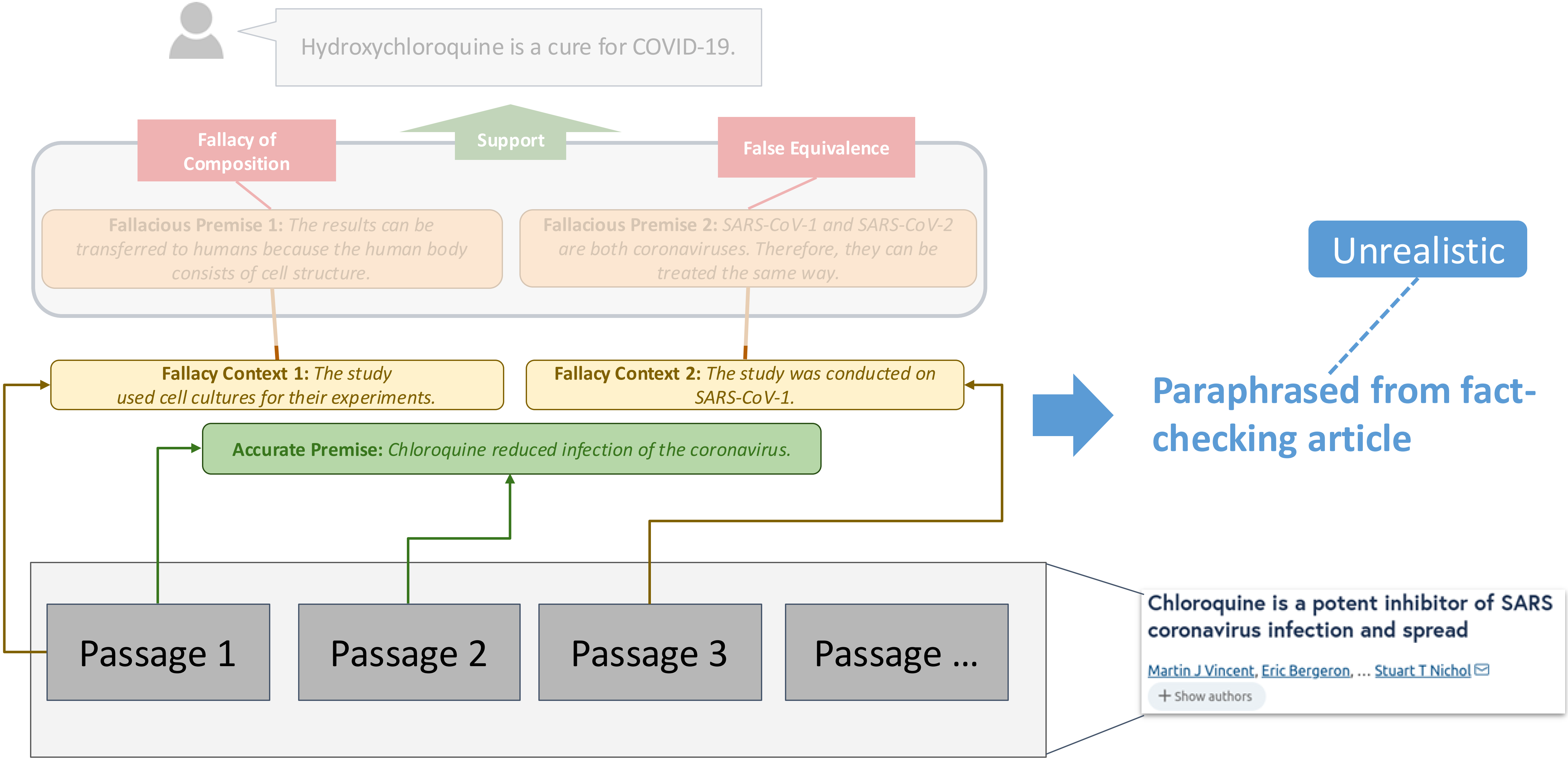
MISSCI Does Not Consider Real-world Passages From Papers



Unrealistic

Paraphrased from fact-checking article

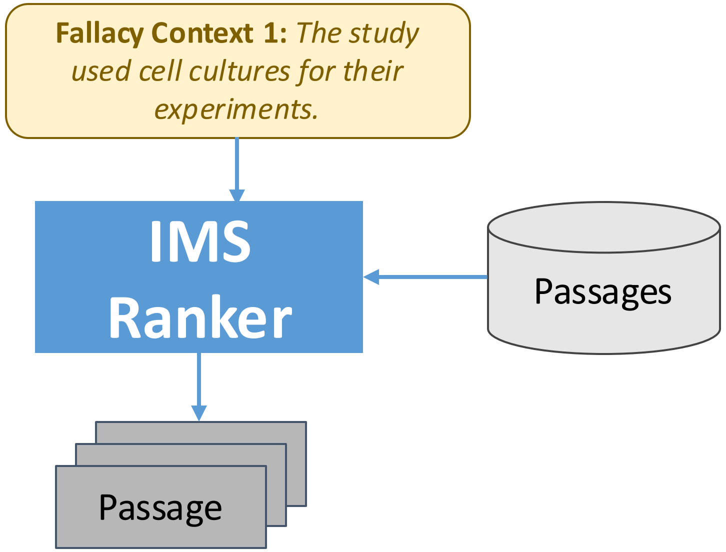
MISSCI Does Not Consider Real-world Passages From Papers



We Link Fallacy Context to Passages From Publications

Preselect

Use IMS (Wright et al., 2022) to preselect relevant passages.



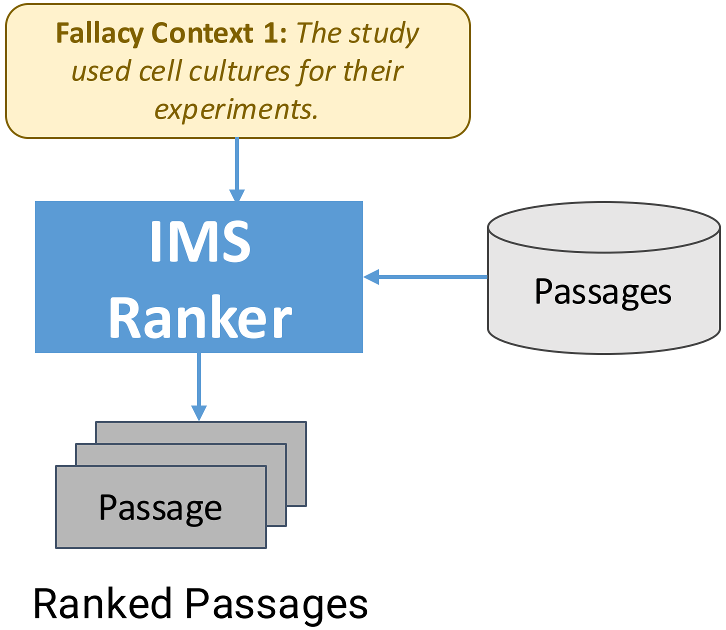
Ranked Passages

We Link Fallacy Context to Passages From Publications

Preselect

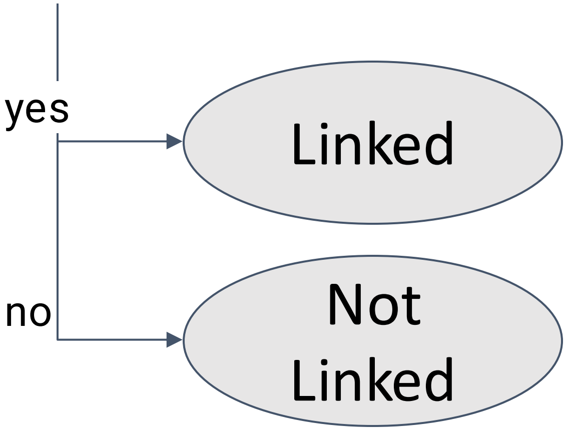
Annotate

Use IMS (Wright et al., 2022) to preselect relevant passages.



Two **biology** annotators

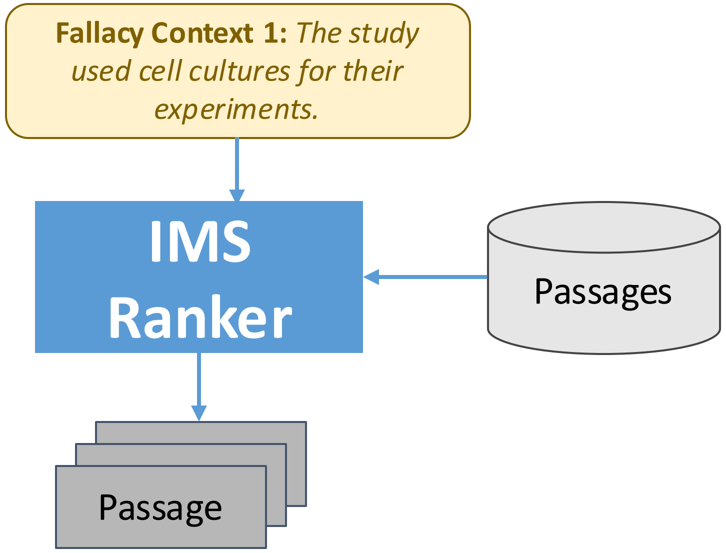
Do (parts of) the **passage** *entail the* **fallacy context?**



We Link Fallacy Context to Passages From Publications

Preselect

Use IMS (Wright et al., 2022) to preselect relevant passages.

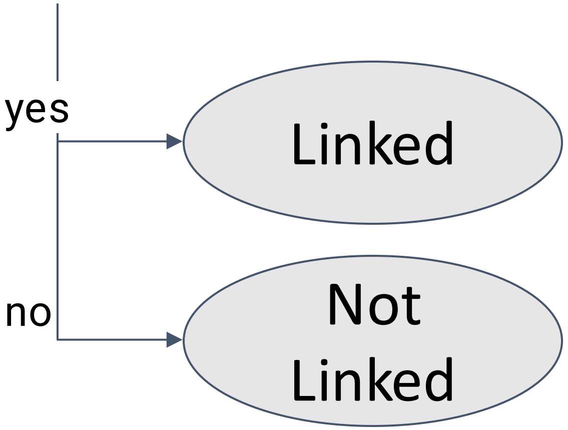


Ranked Passages

Annotate

Two **biology** annotators

Do (parts of) the **passage** *entail the* **fallacy context?**



Aggregate

Consolidate passage annotations

114 Arguments
694 Passages
2,257 Labels

Cohen's κ: 0.602

Not All Fallacy Contexts/Accurate Premises Can be Linked to a Passage

What is the claim based upon?

What content indicates fallacious reasoning?

Component	Ratio linked
Accurate premise	88.6%
Fallacy context	72.0%
All	*76.8%*

Not All Fallacy Contexts/Accurate Premises Can be Linked to a Passage

What is the claim based upon?

What content indicates fallacious reasoning?

Component	Ratio linked
Accurate premise	88.6%
Fallacy context	72.0%
All	*76.8%*

Multi-modal reasoning
Multi-hop reasoning

Not All Fallacy Contexts/Accurate Premises Can be Linked to a Passage

What is the claim based upon?

What content indicates fallacious reasoning?

Component	Ratio linked
Accurate premise	88.6%
Fallacy context	72.0%
All	*76.8%*

Multi-modal reasoning
Multi-hop reasoning

Different Scope

Cloth masks do nothing to prevent the virus.

Scientific publication

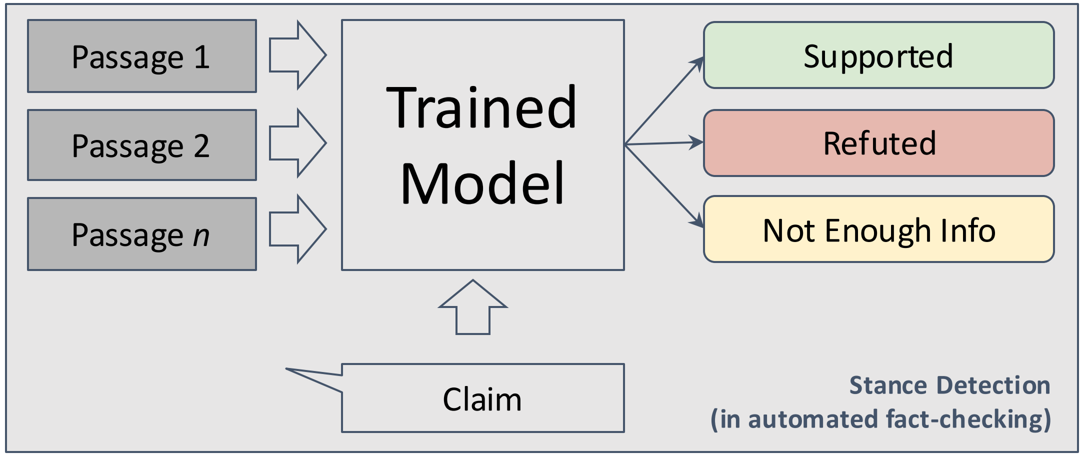
Health workers wearing cloth masks were infected.

Compare effectiveness of cloth masks and medical masks.

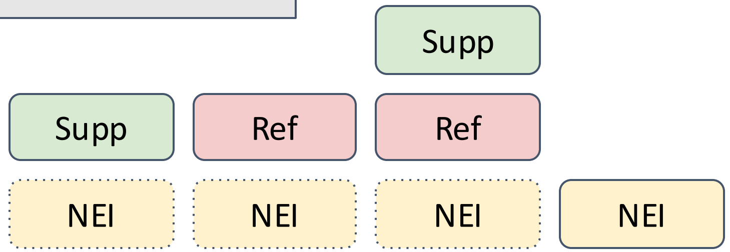
Not explicitly communicated in the study!

Fallacious Context 1: *The study did not include a group that did not wear masks at all.*

Stance Detection Cannot Detect Fallacious Reasoning



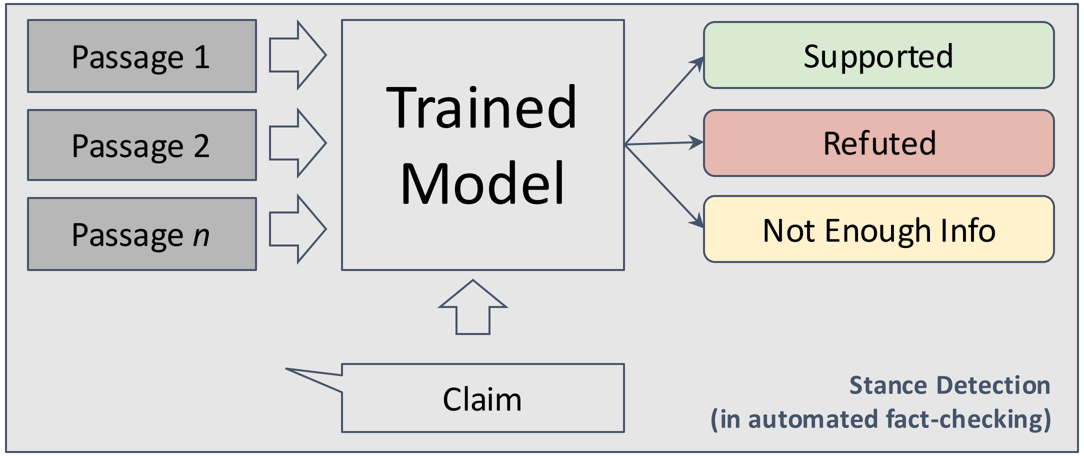
Aggregate Predictions:



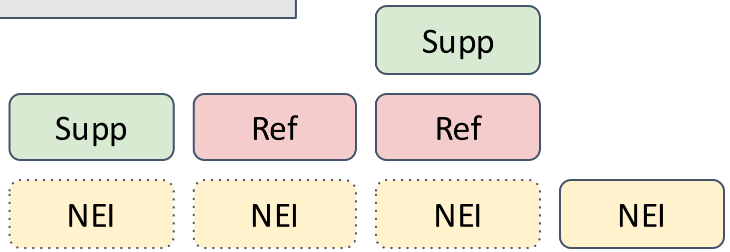
Training Data	Claim level			
	Sup.	**Ref.**	**Mix.**	**NEI**
SciFact (Wadden et al., 2020)	55.5	4.5	23.3	16.7
HealthVer (Sarrouti et al., 2021)	39.0	**20.0**	27.1	13.8
CovidFact (Saakyan et al., 2021)	**36.9**	12.1	**51.0**	–
All AFC	48.8	11.9	27.1	12.1

AFC models based on stance detection predict the majority of the annotated passages support the false claims!

Stance Detection Cannot Detect Fallacious Reasoning



Aggregate Predictions:



Training Data	Claim level Sup.	Ref.	Mix.	NEI
SciFact (Wadden et al., 2020)	55.5	4.5	23.3	16.7
HealthVer (Sarrouti et al., 2021)	39.0	**20.0**	27.1	13.8
CovidFact (Saakyan et al., 2021)	**36.9**	12.1	**51.0**	–
All AFC	48.8	11.9	27.1	12.1

This is bad!

AFC models based on stance detection predict the majority of the annotated passages support the false claims!

Evidence Passages from Papers Bias LLMs to Believe the False Claim is True

MISSCI+

Misinformation

Knowledge	LLM	True	False	NEI
Parametric Knowledge	Llama 2	1.6	61.1	37.3
	GPT 4	0.0	85.3	14.7

Evidence Passages from Papers Bias LLMs to Believe the False Claim is True

MISSCI+ → **Misinformation**

CovidFact, HealthVer → **100 True claims**

Knowledge	LLM	True	False	NEI	True	False	NEI
		Misinformation			100 True claims		
Parametric Knowledge	Llama 2	1.6	61.1	37.3	34.7	22.3	41.3
	GPT 4	0.0	85.3	14.7	59.0	23.0	17.3

Evidence Passages from Papers Bias LLMs to Believe the False Claim is True

MISSCI+ → Misinformation

CovidFact, HealthVer → 100 True claims

Knowledge	LLM	Misinformation: True	Misinformation: False	Misinformation: NEI	100 True claims: True	100 True claims: False	100 True claims: NEI
Parametric Knowledge	Llama 2	1.6	61.1	37.3	34.7	22.3	41.3
	GPT 4	0.0	85.3	14.7	59.0	23.0	17.3

Has a tendency to know the veracity

Evidence Passages from Papers Bias LLMs to Believe the False Claim is True

MISSCI+

CovidFact HealthVer

Knowledge	LLM	Misinformation: True	Misinformation: False	Misinformation: NEI	100 True claims: True	100 True claims: False	100 True claims: NEI
Parametric Knowledge	Llama 2	1.6	61.1	37.3	34.7	22.3	41.3
	GPT 4	0.0	85.3	14.7	59.0	23.0	17.3
RAG Style	Llama 2	**23.8**	61.5	12.7	58.7	29.7	10.7
	GPT 4	**27.4**	34.1	38.5	55.0	4.0	41.0

Has a tendency to know the veracity

Evidence Passages from Papers Bias LLMs to Believe the False Claim is True

MISSCI+ — **Misinformation**

CovidFact, HealthVer — **100 True claims**

Knowledge	LLM	Misinformation True	Misinformation False	Misinformation NEI	True claims True	True claims False	True claims NEI
Parametric Knowledge	Llama 2	1.6	61.1	37.3	34.7	22.3	41.3
	GPT 4	0.0	85.3	14.7	59.0	23.0	17.3
RAG Style	Llama 2	**23.8**	61.5	12.7	58.7	29.7	10.7
	GPT 4	**27.4**	34.1	38.5	55.0	4.0	41.0

Now considers claims correct!

Has a tendency to know the veracity

Locating the Required Passages is Challenging

Task:
Given the claim and **all passages** of the misrepresented publication:

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise:
Chloroquine reduced infection of the coronavirus.

What is the claim based upon?

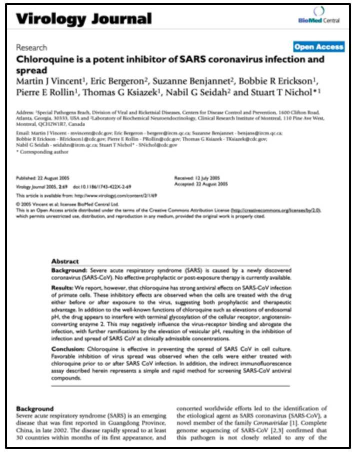
MRR

Post-infection chloroquine treatment reduces SARS-CoV infection and spread.

We have provided evidence that chloroquine is effective in preventing SARS-CoV infection **in cell culture**.

Vero E6 cells (an African green monkey kidney cell line) were infected with **SARS-CoV (Urbani strain)** at a multiplicity of infection of 0.5 for 1 h.

...


Virology Journal
BioMed Central

Research
Open Access

Chloroquine is a potent inhibitor of SARS coronavirus infection and spread

Martin J Vincent[1], Eric Bergeron[2], Suzanne Benjannet[2], Bobbie R Erickson[1], Pierre E Rollin[1], Thomas G Ksiazek[1], Nabil G Seidah[2] and Stuart T Nichol*[1]

Address: [1]Special Pathogens Branch, Division of Viral and Rickettsial Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia, 30333, USA and [2]Laboratory of Biochemical Neuroendocrinology, Clinical Research Institute of Montreal, 110 Pine Ave West, Montreal, QCH2W1R7, Canada

Email: Martin J Vincent - mvincent@cdc.gov; Eric Bergeron - bergere@ircm.qc.ca; Suzanne Benjannet - benjans@ircm.qc.ca; Bobbie R Erickson - BErickson1@cdc.gov; Pierre E Rollin - PRollin@cdc.gov; Thomas G Ksiazek - TKsiazek@cdc.gov; Nabil G Seidah - seidahn@ircm.qc.ca; Stuart T Nichol* - SNichol@cdc.gov

* Corresponding author

Published: 22 August 2005
Received: 12 July 2005
Accepted: 22 August 2005
Virology Journal 2005, **2**:69 doi:10.1186/1743-422X-2-69
This article is available from: http://www.virologyj.com/content/2/1/69

© 2005 Vincent et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Severe acute respiratory syndrome (SARS) is caused by a newly discovered coronavirus (SARS-CoV). No effective prophylactic or post-exposure therapy is currently available.

Results: We report, however, that chloroquine has strong antiviral effects on SARS-CoV infection of primate cells. These inhibitory effects are observed when the cells are treated with the drug either before or after exposure to the virus, suggesting both prophylactic and therapeutic advantage. In addition to the well-known functions of chloroquine such as elevations of endosomal pH, the drug appears to interfere with terminal glycosylation of the cellular receptor, angiotensin-converting enzyme 2. This may negatively influence the virus-receptor binding and abrogate the infection, with further ramifications by the elevation of vesicular pH, resulting in the inhibition of infection and spread of SARS CoV at clinically admissible concentrations.

Conclusion: Chloroquine is effective in preventing the spread of SARS CoV in cell culture. Favorable inhibition of virus spread was observed when the cells were either treated with chloroquine prior to or after SARS CoV infection. In addition, the indirect immunofluorescence assay described herein represents a simple and rapid method for screening SARS-CoV antiviral compounds.

Background

Severe acute respiratory syndrome (SARS) is an emerging disease that was first reported in Guangdong Province, China, in late 2002. The disease rapidly spread to at least 30 countries within months of its first appearance, and concerted worldwide efforts led to the identification of the etiological agent as SARS coronavirus (SARS-CoV), a novel member of the family *Coronaviridae* [1]. Complete genome sequencing of SARS-CoV [2,3] confirmed that this pathogen is not closely related to any of the

Locating the Required Passages is Challenging

Task:
Given the claim and **all passages** of the misrepresented publication:

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Fallacy Context 1: *The study used cell cultures for their experiments.*

Fallacy Context 2: *The study was conducted on SARS-CoV-1.*

What is the claim based upon?

What content points to fallacies?

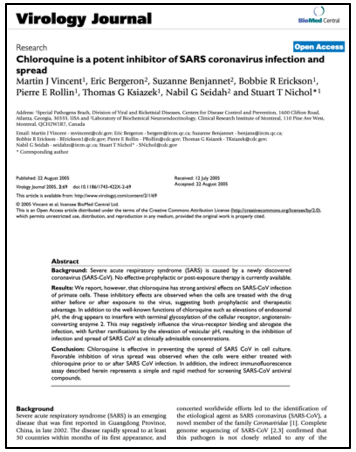
MRR	P@1

Post-infection chloroquine treatment reduces SARS-CoV infection and spread.

We have provided evidence that chloroquine is effective in preventing SARS-CoV infection **in cell culture**.

Vero E6 cells (an African green monkey kidney cell line) were infected with **SARS-CoV (Urbani strain)** at a multiplicity of infection of 0.5 for 1 h.

...


Virology Journal
Research
Open Access
Chloroquine is a potent inhibitor of SARS coronavirus infection and spread
Martin J Vincent[1], Eric Bergeron[2], Suzanne Benjannet[2], Bobbie R Erickson[1], Pierre E Rollin[1], Thomas G Ksiazek[1], Nabil G Seidah[2] and Stuart T Nichol*[1]

Locating the Required Passages is Challenging

Task:
Given the claim and **all passages** of the misrepresented publication:

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Fallacy Context 1: *The study used cell cultures for their experiments.*

Fallacy Context 2: *The study was conducted on SARS-CoV-1.*

What is the claim based upon?

What content points to fallacies?

Type	Model	MRR	P@1
Term frequency	BM25	0.539	
Sentence Transformer (embeddings)	BioBERT ST (Deka et al. 2022)	0.582	
	INSTRUCTOR (Su et al. 2022)	0.631	
	SPICED (IMS) (Wright et al. 2022)	**0.664**	

Locating the Required Passages is Challenging

Task:
Given the claim and **all passages** of the misrepresented publication:

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Fallacy Context 1: *The study used cell cultures for their experiments.*

Fallacy Context 2: *The study was conducted on SARS-CoV-1.*

What is the claim based upon?

What content points to fallacies?

Type	Model	MRR	P@1
Term frequency	BM25	0.539	
Sentence Transformer (embeddings)	BioBERT ST (Deka et al. 2022)	0.582	
	INSTRUCTOR (Su et al. 2022)	0.631	
	SPICED (IMS) (Wright et al. 2022)	**0.664**	
Scientific Fact-Checking (DeBERTaV3)	SciFact (Wadden et al. 2020)	0.535	
	CovidFact (Saakyan et al. 2020)	0.450	
	HealthVer (Sarrouti et al. 2021)	0.516	
	All Scientific Fact-Checking	0.514	

Locating the Required Passages is Challenging

Task:
Given the claim and **all passages** of the misrepresented publication:

Hydroxychloroquine is a cure for COVID-19.

Accurate Premise: *Chloroquine reduced infection of the coronavirus.*

Fallacy Context 1: *The study used cell cultures for their experiments.*

Fallacy Context 2: *The study was conducted on SARS-CoV-1.*

What is the claim based upon?

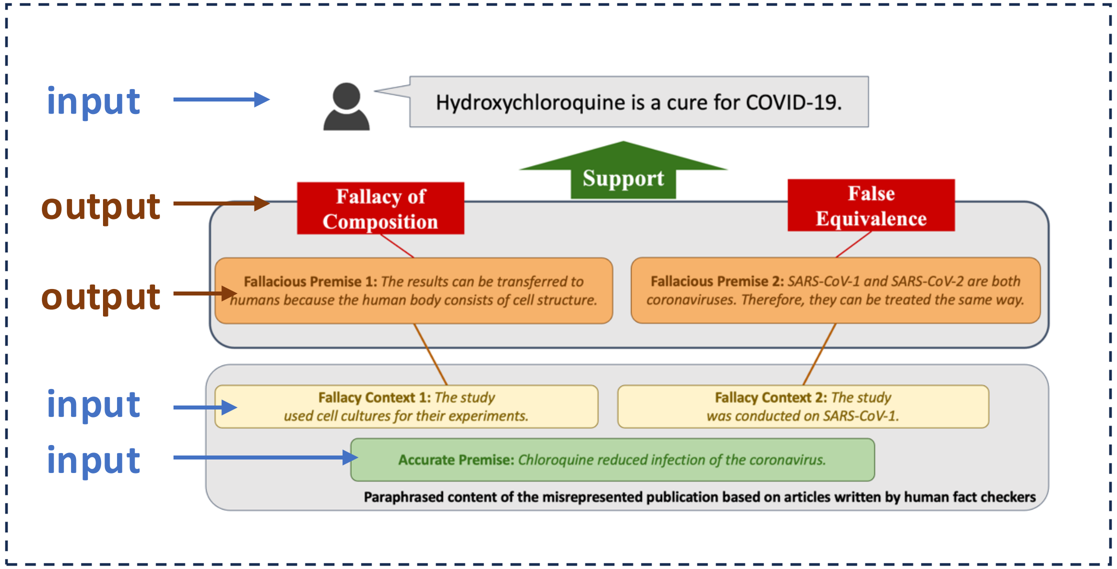
What content points to fallacies?

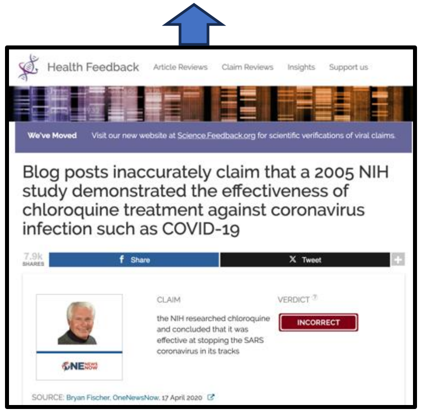
Type	Model	MRR	P@1
Term frequency	BM25	0.539	0.617
Sentence Transformer (embeddings)	BioBERT ST (Deka et al. 2022)	0.582	0.600
	INSTRUCTOR (Su et al. 2022)	0.631	**0.652**
	SPICED (IMS) (Wright et al. 2022)	**0.664**	0.640
Scientific Fact-Checking (DeBERTaV3)	SciFact (Wadden et al. 2020)	0.535	0.326
	CovidFact (Saakyan et al. 2020)	0.450	0.457
	HealthVer (Sarrouti et al. 2021)	0.516	0.410
	All Scientific Fact-Checking	0.514	0.338

Put Everything Together …

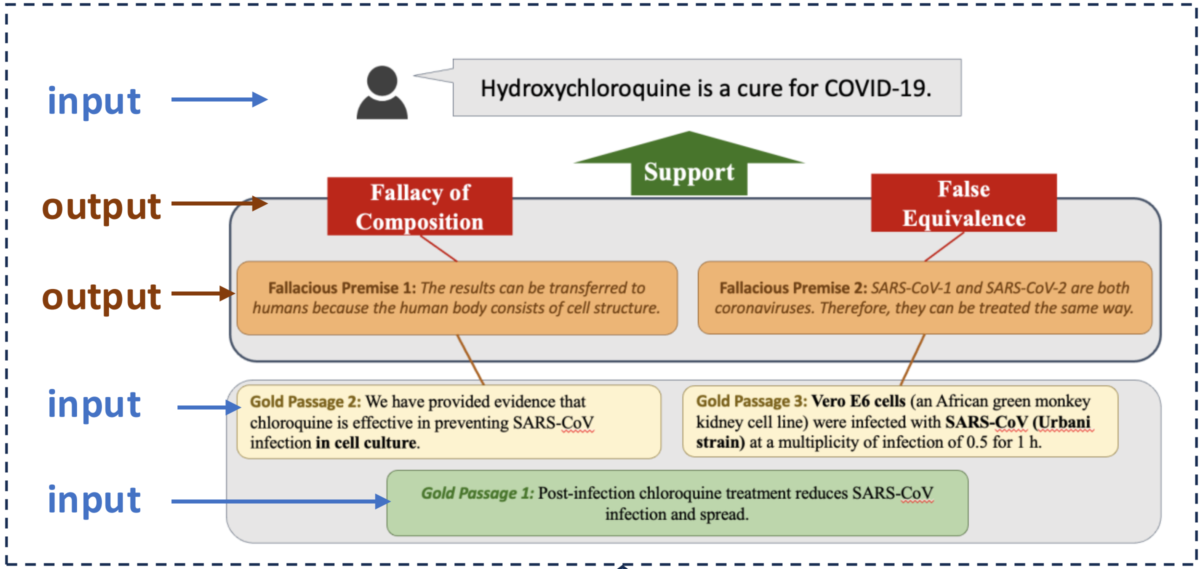
Full Task: *Generate fallacious premise and predict applied fallacy class for the given input*

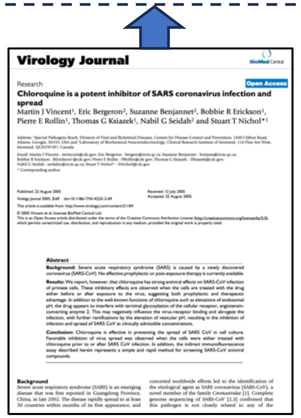
MISSCI





MISSCI+





It's Challenging to Evaluate Argument Construction With the Retrieved Passages

- The same passage can be linked to multiple reasoning gaps (s_i) and vice versa
- We evaluate the task on the argument level: we train two models to automatically map the generated ($\bar{p}_i$, $\bar{f}_i$) to the gold fallacies at the argument level

LLM	Info	MISSCI R@5 (ϕ^{f+p})	MISSCI R@5 (ϕ^{f})	MISSCI Arg@1 (ϕ^{f+p})	MISSCIPLUS R@5 (ϕ^{f+p})	MISSCIPLUS R@5 (ϕ^{f})	MISSCIPLUS Arg@1 (ϕ^{f+p})
Llama3-8B	DLE	**0.277**	**0.514**	**0.552**	**0.226**	**0.477**	**0.476**
	DL	0.241	0.445	0.512	0.195	0.463	0.413
	DE	0.227	0.470	0.480	0.174	0.449	0.389
	LE	0.255	0.469	0.504	0.209	0.439	0.460
GPT-3.5	DLE	0.248	0.491	0.512	**0.165**	**0.428**	**0.361**
	DL	0.232	0.492	0.464	0.146	0.416	0.321
	DE	**0.276**	**0.517**	**0.567**	0.160	0.400	0.333
	LE	0.249	0.478	0.524	0.157	0.410	0.341
GPT-4 Turbo	DLE	**0.332**	0.486	**0.619**	0.224	0.458	0.452
	DL	0.308	0.500	0.583	0.238	0.495	0.488
	DE	0.318	**0.528**	0.595	0.210	0.491	0.440
	LE	0.304	0.505	0.583	**0.252**	**0.519**	**0.500**

A clear trend of decreasing performance from **paraphrased information** in MISSCI to the **real-world passages** in MISSCI+

Conclusion

Bridge the gap between automated fact-checking and fallacy detection.

Novel benchmark to reconstruct fallacious arguments with **realistic evidence from scientific papers**.

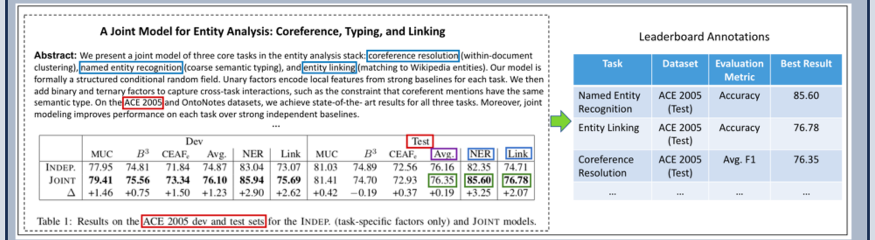
Evidence from the misrepresented publication **biases the LLMs to believe the claim is true.**

Outline

Build Global Scientific Evidence Map

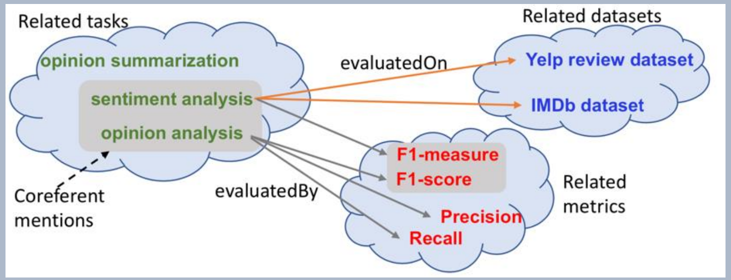
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



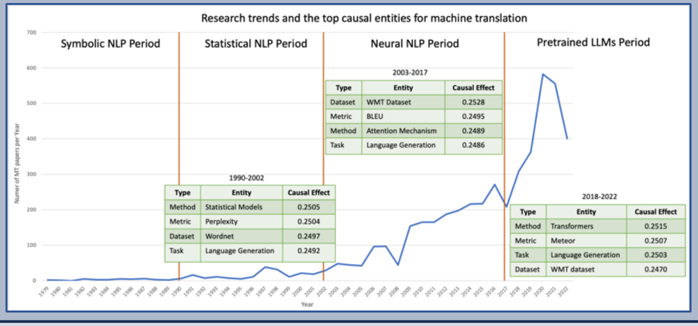
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

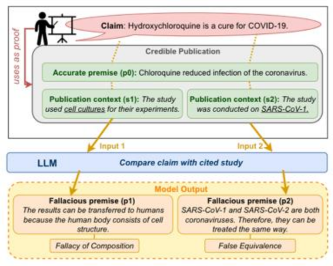
- NLP Concepts Causal Analysis



Scientific Communication

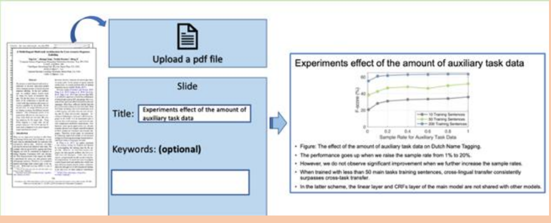
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

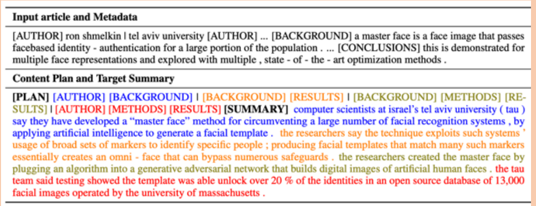
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures

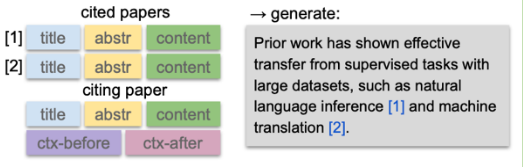


Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation [Funkquist et al., EMNLP 2023]

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation [O'Doherty et al., ACL 2024 SRW]



D2S: Automated Slide Generation With Query-based Text Summarization From Documents

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, Nancy Wang

(NAACL 2021)

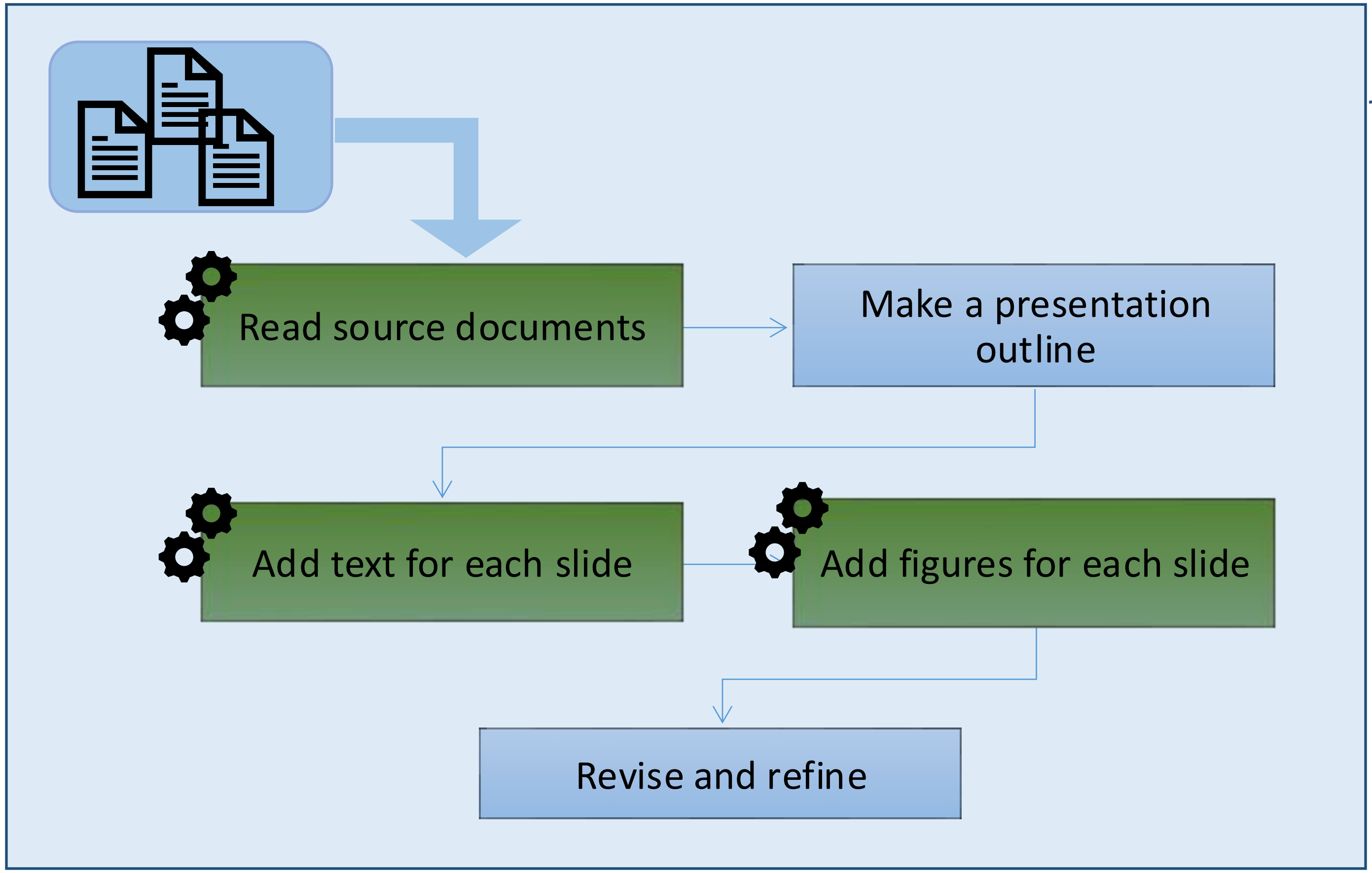




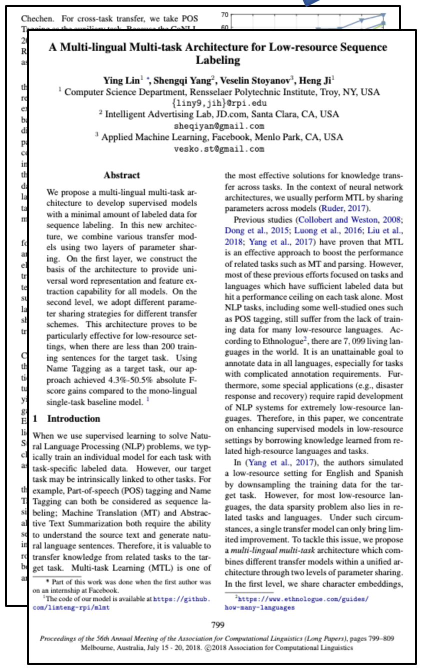
Doc2Slides Generation: Motivation

- ❑ Presentation slides are the key communication tool in various areas (e.g., business, science)
- ❑ The creation of slide decks is often tedious and time-consuming

Let AI provide suggestions to speed up the process!



Query-based D2S: A New Challenging Task


A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling

Ying Lin[1 *], Shengqi Yang[2], Veselin Stoyanov[3], Heng Ji[1]

[1] Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY, USA

[2] Intelligent Advertising Lab, JD.com, Santa Clara, CA, USA

[3] Applied Machine Learning, Facebook, Menlo Park, CA, USA

Abstract

1 Introduction

799

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 799–809
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

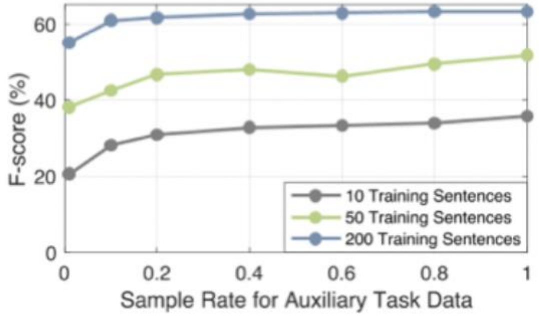
Upload a pdf file

Slide

Title: Experiments effect of the amount of auxiliary task data

Keywords: **(optional)**

Experiments effect of the amount of auxiliary task data



- Figure: The effect of the amount of auxiliary task data on Dutch Name Tagging.
- The performance goes up when we raise the sample rate from 1% to 20%.
- However, we do not observe significant improvement when we further increase the sample rates.
- When trained with less than 50 main tasks training sentences, cross-lingual transfer consistently surpasses cross-task transfer.
- In the latter scheme, the linear layer and CRFs layer of the main model are not shared with other models.

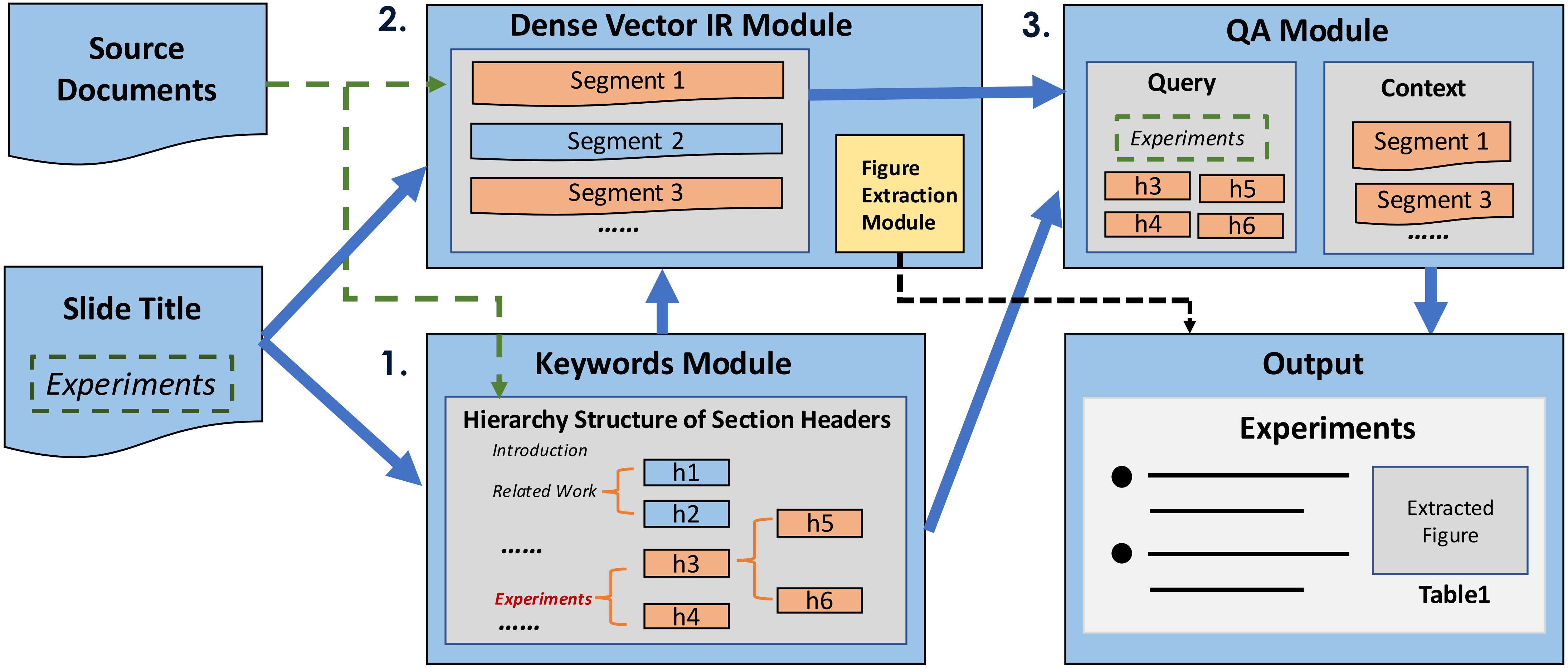
Doc2Slides Generation: SciDuet Dataset

- A high-quality dataset containing paper-slide pairs from ICML'19/NeurIPS'18&'19/ACL Anthology
- Data processing: A non-trivial effort
  - A combination of different tools: Grobid/ Pdffigures2/IBM Watson Discovery package
  - Paper figures/tables appearing on slides were matched using OpenCV multiscale template matching
  - Obtain 1088 papers – 10,034 slides after filtering out slides that don't correspond well with paper (e.g., acknowledgement slide)

	#papers	#slides	ST-len	SC-len
train	952	8,123	3.6	55.1
dev	55	733	3.16	63.4
test	81	1,178	3.4	52.3

Doc2Slides Generation: System Framework

- ❑ Modelling as an **open-query long-form question answering problem**
- ❑ Another perspective: **query-based single document summarization**



Doc2Slides Generation: Evaluation

- Slide generation is subjective and there can be many correct slide versions with few overlapping words
- Our evaluation strategy

1. Automatic evaluation (Rouge) to compare our model with the baselines

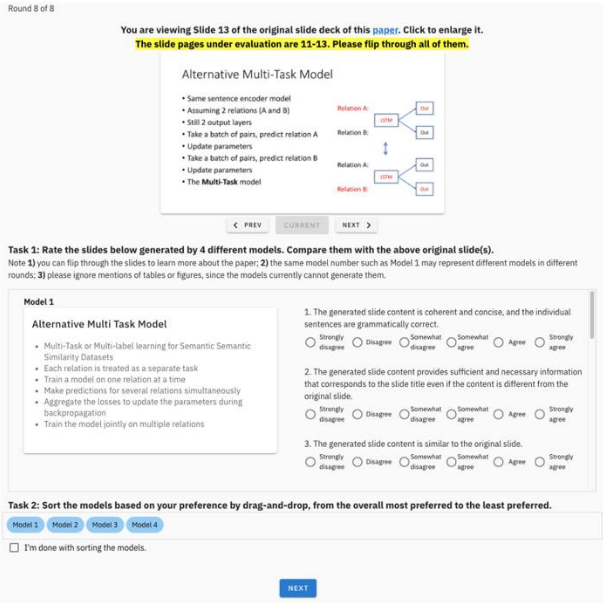
Summarization Model	ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F	P	R	F	P	R	F
Classical IR (BM25)									
BertSummExt	14.26	24.07	15.89	2.59	4.46	2.86	12.89	21.70	14.31
BARTSumm	15.75	23.40	16.92	2.94	4.12	3.11	14.18	20.99	15.55
BARTKeyword (ours)	17.15	27.98	19.06	4.08	6.52	4.52	16.29	24.88	18.12
Dense-Mix IR (ours)									
BertSummExt	15.47	25.74	17.16	3.14	5.24	3.47	13.97	23.29	15.48
BARTSumm	16.62	26.10	18.15	3.35	5.16	3.63	15.00	23.28	16.73
BARTKeyword (ours)	**18.30**	**30.31**	**20.47**	**4.73**	**7.79**	**5.26**	**16.86**	**27.21**	**19.08**

2. Automatic evaluation (Rouge) on non-author human generated slides to estimate upper bound
3. Human ratings in three dimensions

   a) Readability

   b) Informativeness (relevant to title)

   c) Consistency (similar to original slides)



Doc2Slides Generation: Automated Evaluation

- Automated evaluations show comparable performance to non-author human generated slide

Paper(s)	Generator	Rouge-1	Rouge-2	Rouge-L
960	Humans	23.91(2.97)	6.55(0.79)	24.23(2.03)
960	Human-best	28.10	7.66	**27.10**
960	BARTKeyword(ours)	**29.48**	**8.16**	26.12
All	Humans	26.41(4.80)	**8.66(2.24)**	24.68(2.03)
All	BARTKeyword(ours)	**27.75(1.62)**	8.30(0.36)	**24.69(1.18)**

Three non-author humans make slides for paper 960 based on the original slide titles

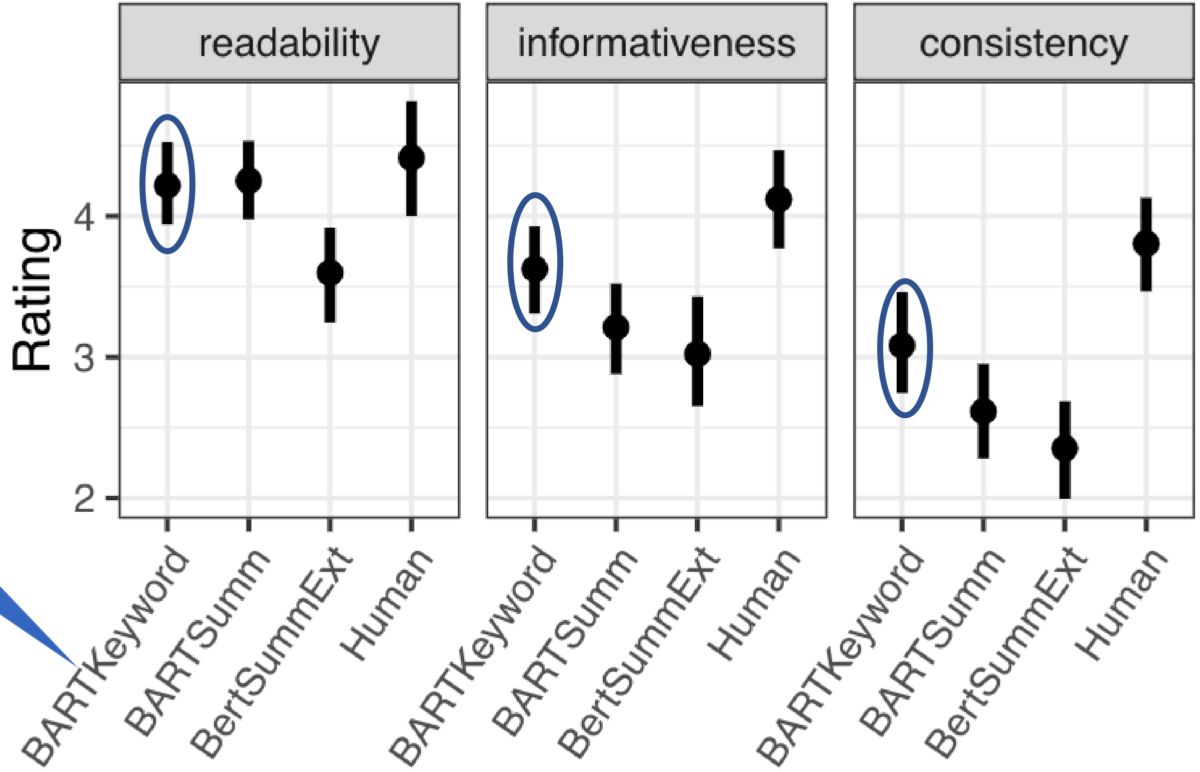
Three non-author humans make slides for 4 papers based on the original slide titles

The purpose of this experiment is to check the average non-author human performance on this task in terms of Rouge. Later we found that our system sometimes is better than humans to find the relevant information from the source paper.

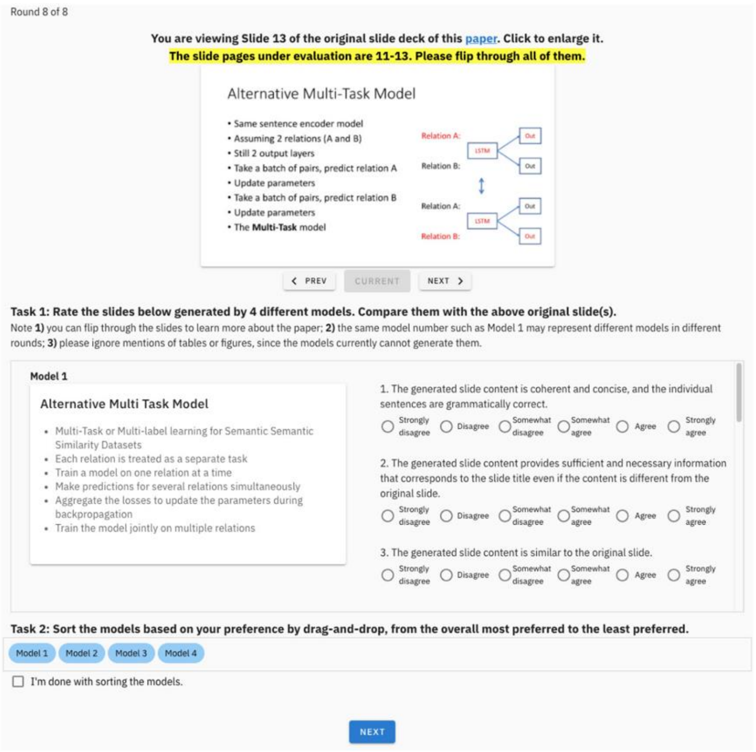
Doc2Slides Generation: Human Evaluation

- Participants: 23 ML researchers and master/PhD students
  - ✓ Each annotator: 2 papers × 4 slides (different versions from different models)
    - ✓ one paper is from the set of four papers that has non-author human generated slides
    - ✓ another paper is from the remaining testing dataset
  - ✓ 6-point Likert scale

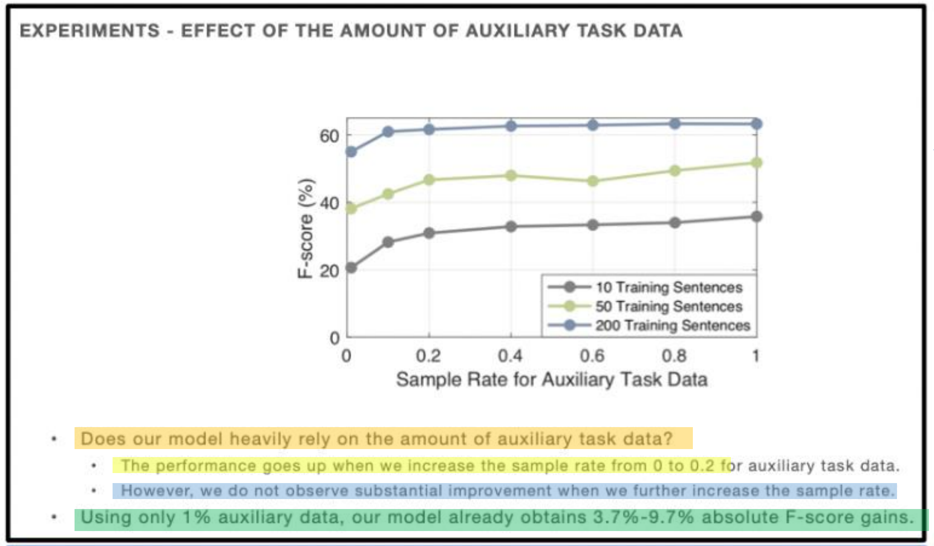
- Human Evaluations show higher preference for human generated slides, but our model outperforms other strong baselines



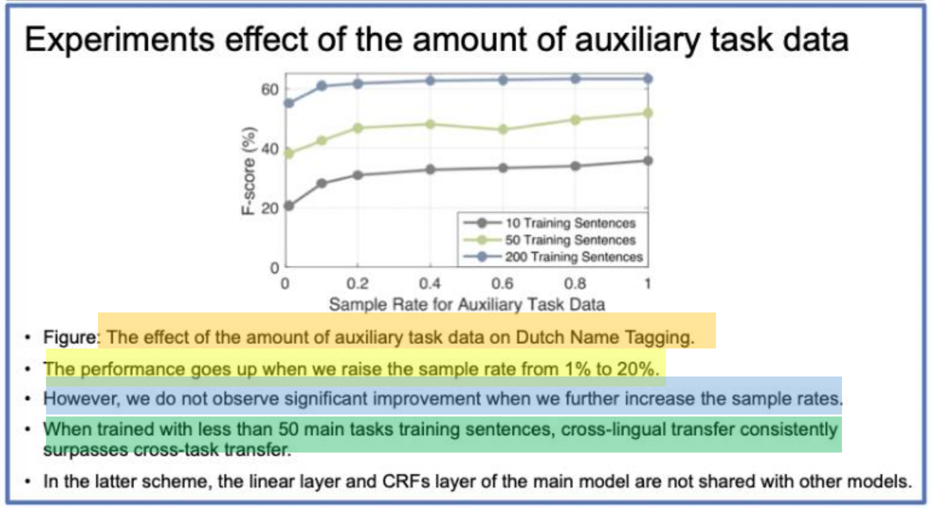
Our model



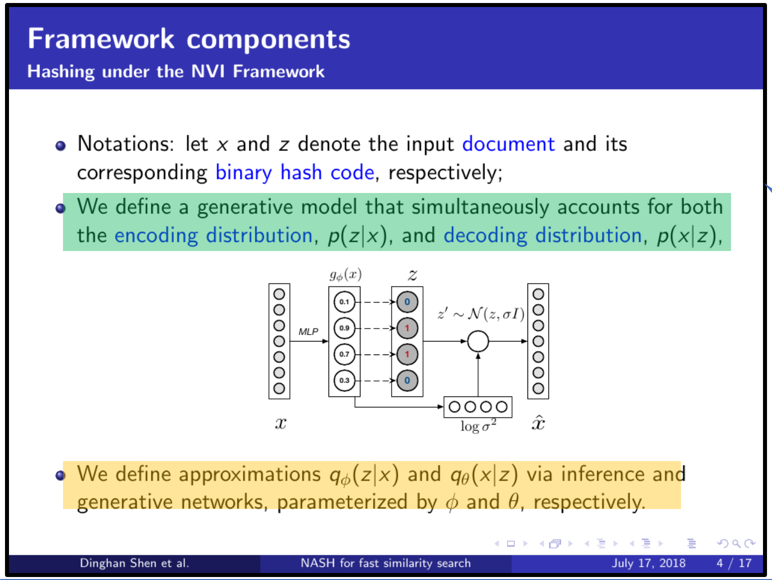
Doc2Slides Generation: Examples of Generated Slides vs. Ground Truth



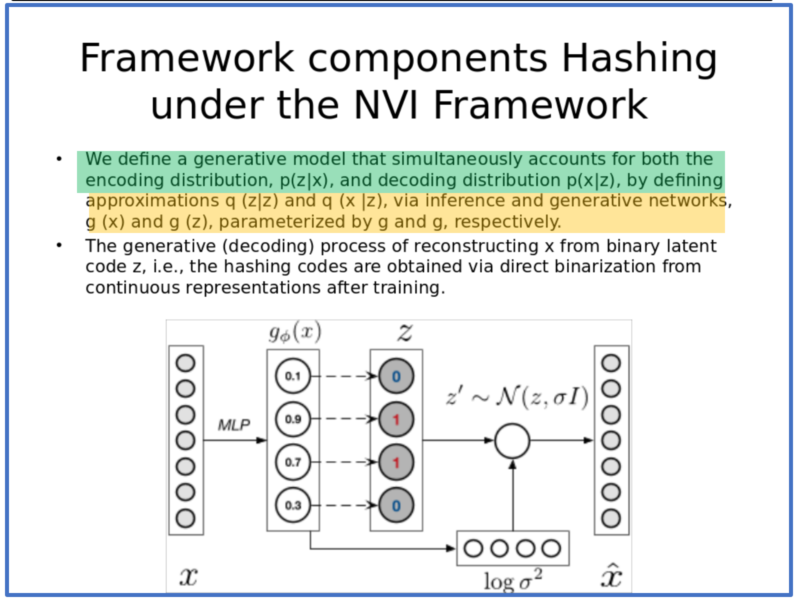
Original Slide



Our model



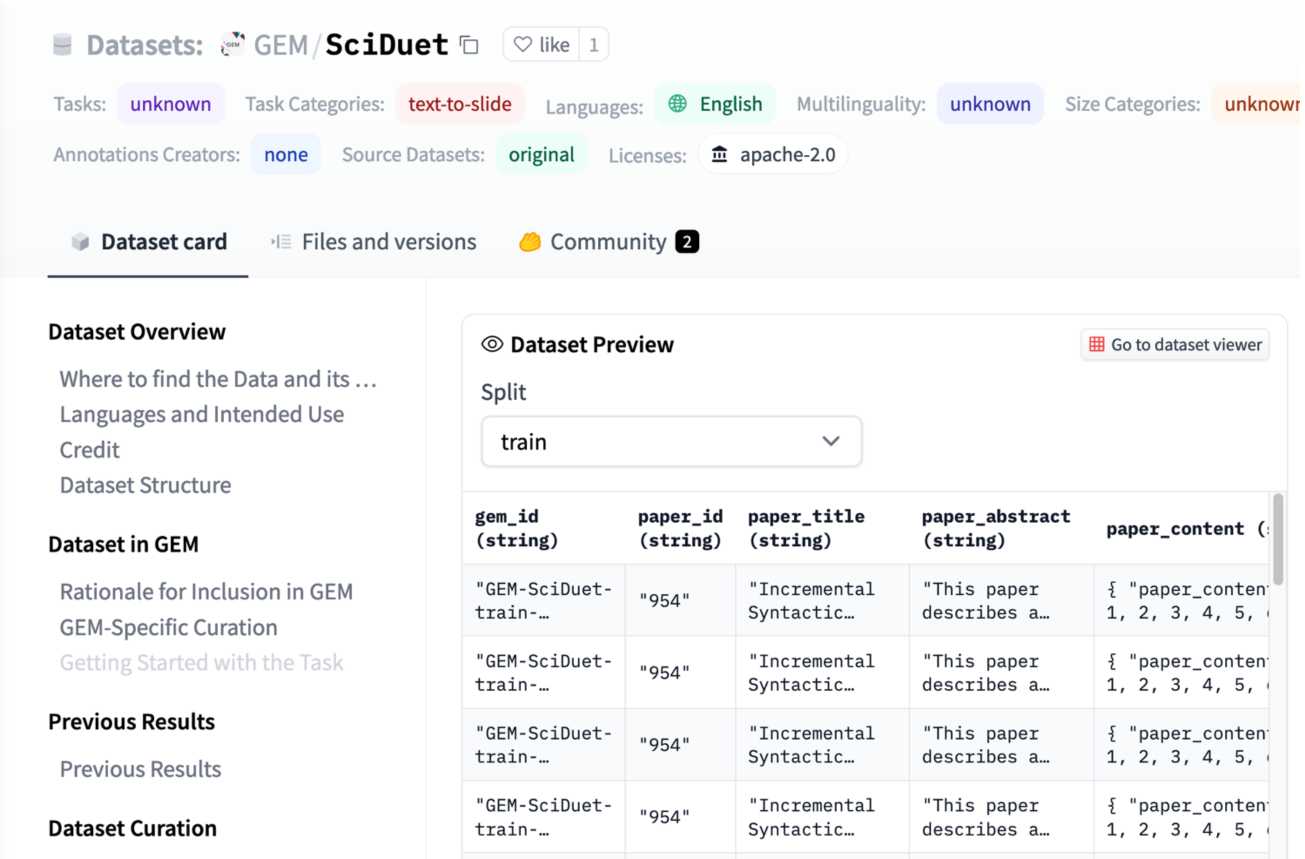
Original Slide



Our model

Access SciDuet

- Documents2slides git repo (https://github.com/IBM/document2slides)
- GEM-SciDuet (https://huggingface.co/datasets/GEM/SciDuet)

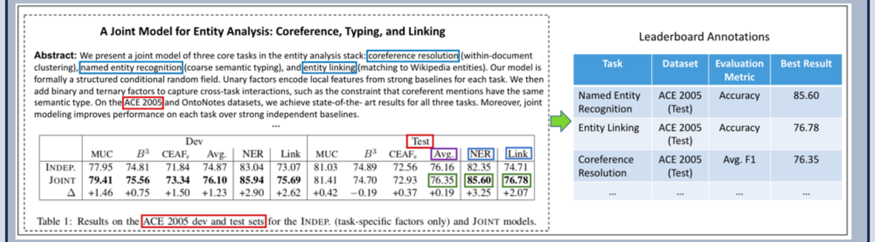


Outline

Build Global Scientific Evidence Map

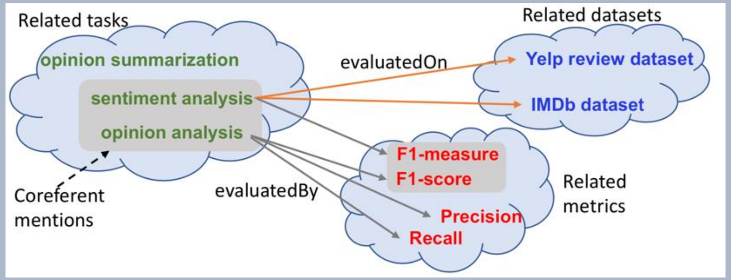
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



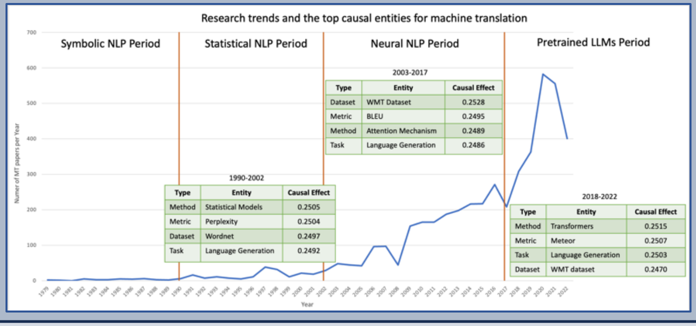
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

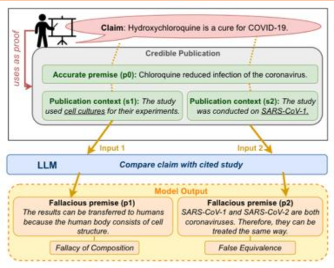
- NLP Concepts Causal Analysis



Scientific Communication

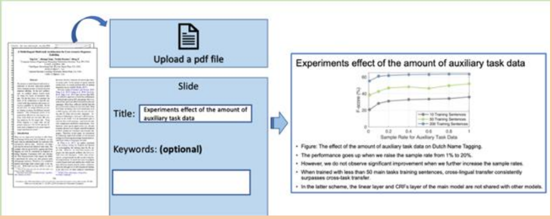
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

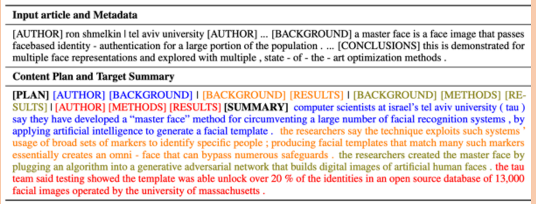
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures



Input article and Metadata

[AUTHOR] ron shmelkin | tel aviv university [AUTHOR] ... [BACKGROUND] a master face is a face image that passes facebased identity - authentication for a large portion of the population . ... [CONCLUSIONS] this is demonstrated for multiple face representations and explored with multiple , state - of - the - art optimization methods .

Content Plan and Target Summary

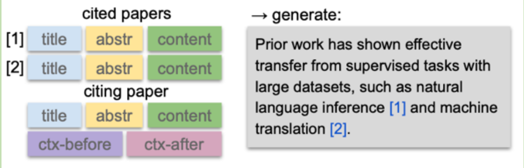
[PLAN] [AUTHOR] [BACKGROUND] | [BACKGROUND] [RESULTS] | [BACKGROUND] [METHODS] [RESULTS] | [AUTHOR] [METHODS] [RESULTS] **[SUMMARY]** computer scientists at israel's tel aviv university (tau) say they have developed a "master face" method for circumventing a large number of facial recognition systems , by applying artificial intelligence to generate a facial template . the researchers say the technique exploits such systems ' usage of broad sets of markers to identify specific people ; producing facial templates that match many such markers essentially creates an omni - face that can bypass numerous safeguards . the researchers created the master face by plugging an algorithm into a generative adversarial network that builds digital images of artificial human faces . the tau team said testing showed the template was able unlock over 20 % of the identities in an open source database of 13,000 facial images operated by the university of massachusetts .

Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation **[Funkquist et al., EMNLP 2023]**

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation **[O'Doherty et al., ACL 2024 SRW]**



`Don’t Get Too Technical with Me’: A Discourse Structure-Based Framework for Automatic Science Journalism

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, Yufang Hou

(EMNLP 2023)

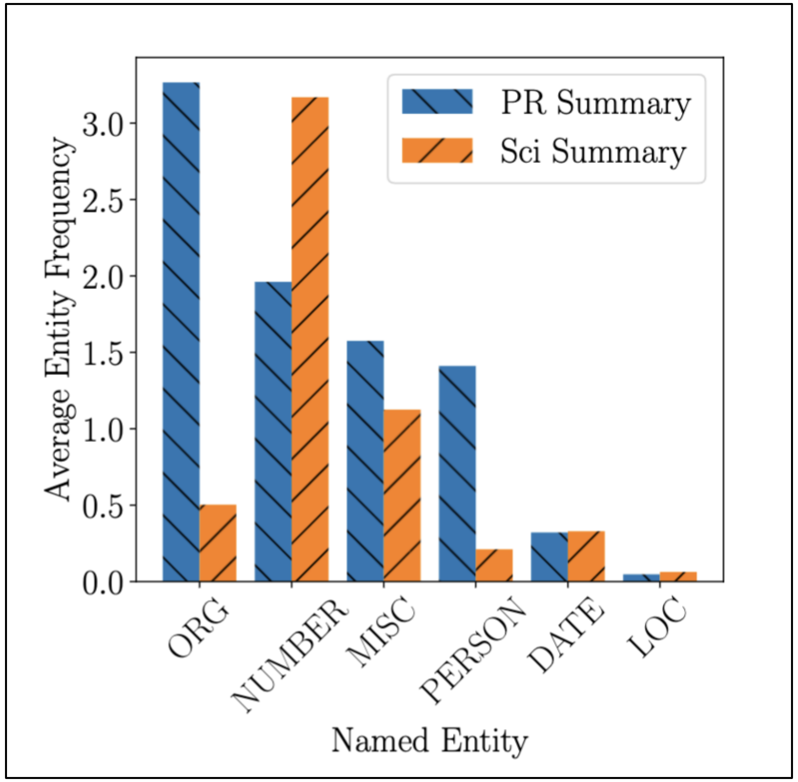


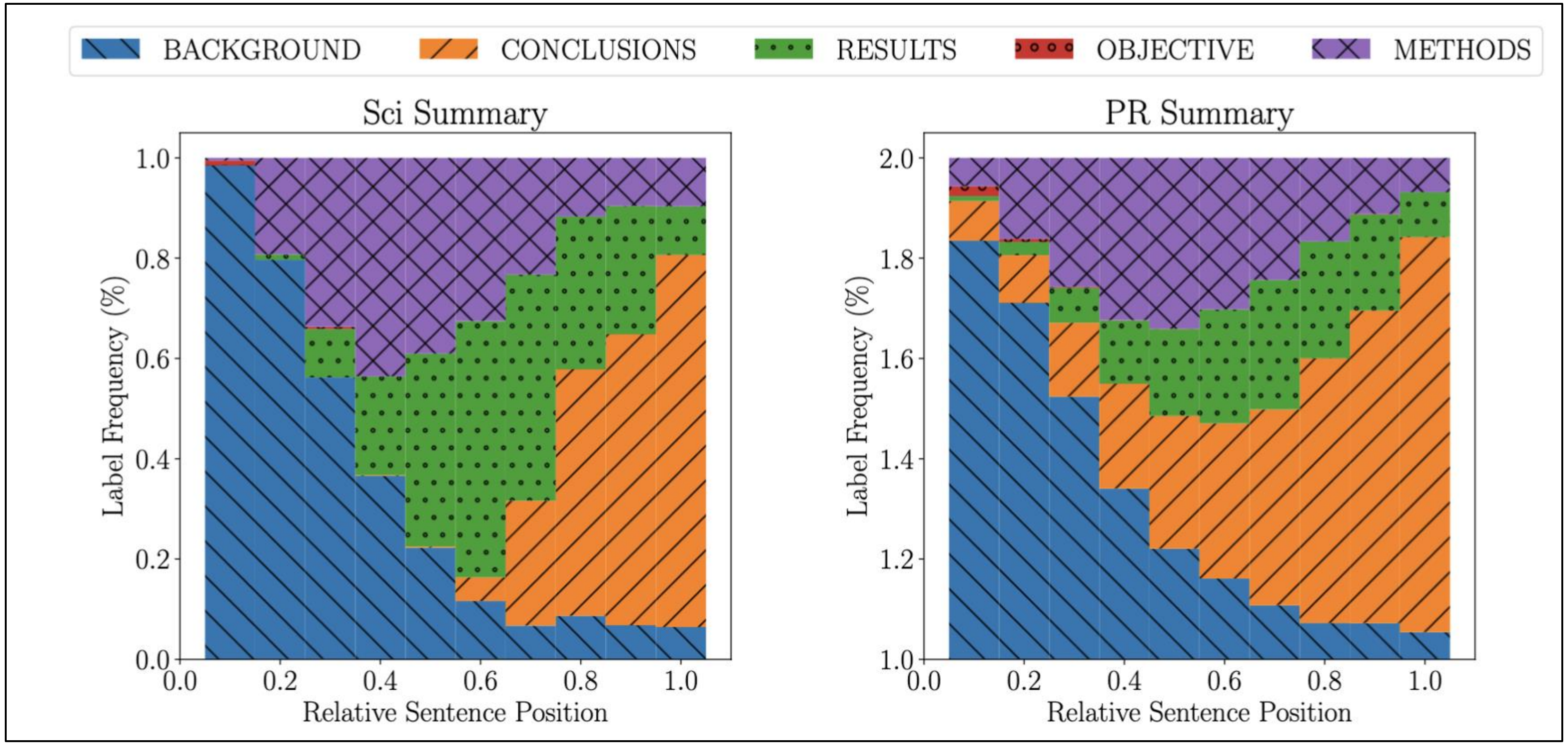

SciTechNews Dataset

- Collect 2,432 aligned {paper, news article, expert-written summary snippets} from ACM TechNews
- Areas: Computer science, Engineering, Astrophysics, Biology, etc
- Scientific paper sources

Source	#Instances	
	Valid	Test
nature	189	320
arxiv	263	231
journals.aps	21	73
dl.acm	67	64
ieeexplore.ieee	126	14
usenix	4	11
journals.plos	60	7
author	222	68
other	480	212
Total	1432	1000

Scientific Abstract vs. Scientific News Summary





Controllable Scientific News Summary Generation

Training

Input article and Metadata
[AUTHOR] ron shmelkin I tel aviv university [AUTHOR] ... [BACKGROUND] a master face is a face image that passes facebased identity - authentication for a large portion of the population . ... [CONCLUSIONS] this is demonstrated for multiple face representations and explored with multiple , state - of - the - art optimization methods .
Content Plan and Target Summary
[PLAN] [AUTHOR] [BACKGROUND] I [BACKGROUND] [RESULTS] I [BACKGROUND] [METHODS] [RE-SULTS] I [AUTHOR] [METHODS] [RESULTS] **[SUMMARY]** computer scientists at israel's tel aviv university (tau) say they have developed a "master face" method for circumventing a large number of facial recognition systems , by applying artificial intelligence to generate a facial template . the researchers say the technique exploits such systems ' usage of broad sets of markers to identify specific people ; producing facial templates that match many such markers essentially creates an omni - face that can bypass numerous safeguards . the researchers created the master face by plugging an algorithm into a generative adversarial network that builds digital images of artificial human faces . the tau team said testing showed the template was able unlock over 20 % of the identities in an open source database of 13,000 facial images operated by the university of massachusetts .

Inference

Oracle Plan & Generated Summary
[PLAN] [AUTHOR] [CONCLUSIONS] I [CONCLUSIONS] I [AUTHOR] I [CONCLUSIONS] **[SUMMARY]** Researchers at the University of Basel in Switzerland say they have developed a method for generating a secure quantum key via photonic conversion. The method adds artificial noise to the initial measurement data, which cannot be known or controlled by an adversary, to guarantee that the key will be obtained by both Alice and Bob, who are connected by a quantum channel. By adding the noise, the researchers were able to increase the threshold at which they can guarantee the key is obtained. Although the researchers say their method is not perfect, they believe it is still effective and could lead to a new generation of highly secure and provable quantum key distribution. They note that the new method provides security even when assumptions about the security of QKD are not fulfilled.
User Plan & Generated Summary
[PLAN] [AUTHOR] I [BACKGROUND] I [METHODS] I [RESULTS] [CONCLUSIONS] **[SUMMARY]** Researchers at the University of Basel in Switzerland say they have developed a method for generating a secure quantum key via photonic conversion. The method involves adding artificial noise to the initial measurement data, which cannot be known or controlled by an adversary, to increase the probability that the key is obtained. Although the researchers say the method is not fully device-independent, they say it provides security even when assumptions are made about the security of the quantum key distribution. "There is a fundamental obstacle in the development of QKD, i.e.,, the requirement that an adversary cannot fully control the quantum channel," says Basel Professor Sangouard Bancal. However, he says the method provides sufficient bounds on the minimum required global detection efficiency to ensure that the information sent over the qubit channel is good and accurate.

Evaluation Results

Systems	R1	R2	RL	BSc
ABSTRACT	32.94	6.26	28.84	81.20
EXT-ORACLE	39.73	10.43	34.10	84.49
Lead	32.46	5.79	28.17	83.81
Random	29.58	3.99	25.50	82.60
LexRank	31.40	5.21	27.16	82.98
TextRank	31.86	5.38	27.38	82.92
Bart_{arx}	32.28	6.01	28.12	82.81
Bart_{SciT}	36.42	7.51	31.71	84.12
Bart_{meta}	38.07	**9.03**	33.14	84.76
Bart_{plan}	**38.84***	8.89	**33.50***	**84.78**
Alpaca	21.24	3.24	18.16	81.20
FlanT5-large	26.26	4.98	20.13	80.98
GPT-3.5-Turbo	35.67	6.75	28.68	82.86

System	Inf.	N-Rd.	Fact.	Read.	Sty.	Use.
Bart_{meta}	0.13	-0.31	-0.33	0.01	0.16	-0.22
Bart_{plan}	**0.08**	**0.08**	-0.10	**0.22**	**0.30**	**0.02**
GPT-3.5	-0.07	-0.01	**0.02**	-0.23	-0.24	-0.21
PR Sum.	0.58	0.68	0.43	0.79	0.91	0.57

Table 6: System ranking according to human judgments, along (Inf)ormativeness, (Non-Red)undancy, (Fact)uality (F), (Read)ability, Press Release (Sty)le, and (Use)fulness. Best system is shown in **bold**.

Scores of Gold PR summaries are higher than the machine-generated texts

Factual Error Analysis

System	Entity			Noun Phrase			Other	Total
	Int.	Ext.	W.K.	Int.	Ext.	W.K.		
PR Sum.	0.0	0	0.79	0.0	0.0	0.21	0	43
Bart_{plan}	0.1	0.34	0.20	0.07	0.16	0.02	0.11	61
GPT3.5	0.0	0.08	0.0	0.02	0.18	0.0	0.72	50

$\text{Bart}_{\text{plan}}$ extrinsically hallucinates entities
(e.g., *Researchers from* ***UK***)

ChatGTP produces a high proportion of extrinsic hallucinations in “other” category
(e.g., *In a paper published in* ***Nature***)

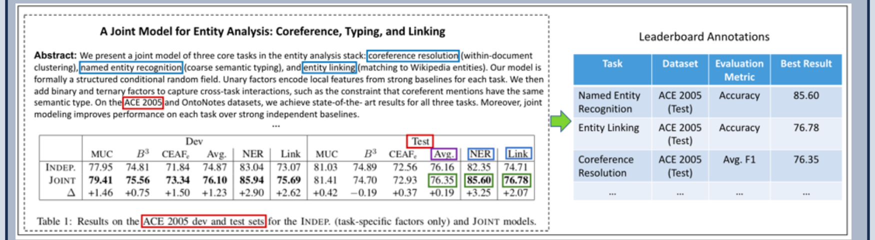
Human summaries only contain Entity and NP-related errors of type World Knowledge
(e.g., *Researchers from* ***Massachusetts Institute of Technology***)

Outline

Build Global Scientific Evidence Map

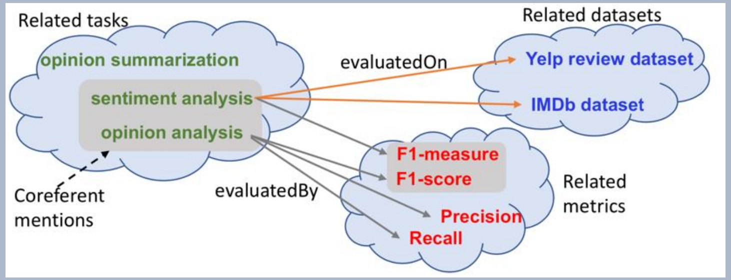
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format
- https://github.com/IBM/science-result-extractor



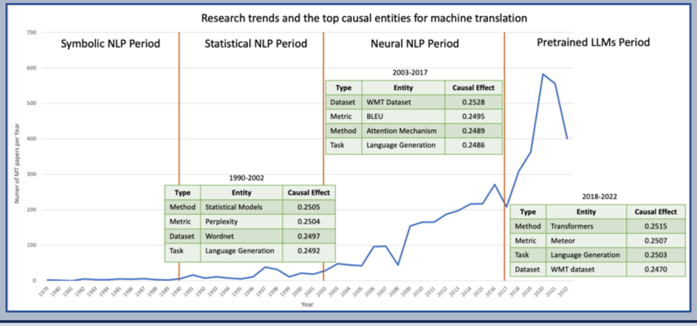
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

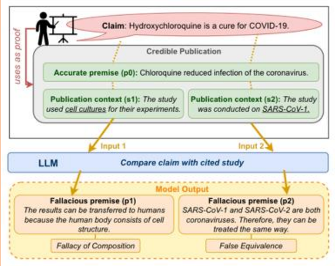
- NLP Concepts Causal Analysis



Scientific Communication

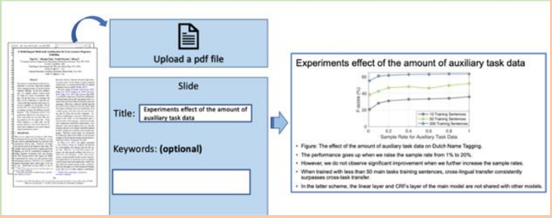
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

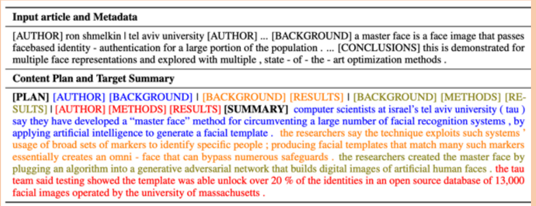
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- Controlled generation based on discourse structures

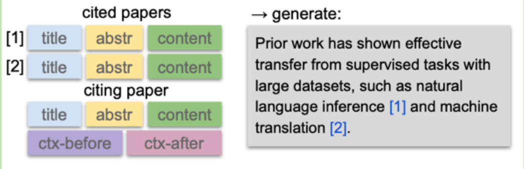


Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation **[Funkquist et al., EMNLP 2023]**

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation **[O'Doherty et al., ACL 2024 SRW]**



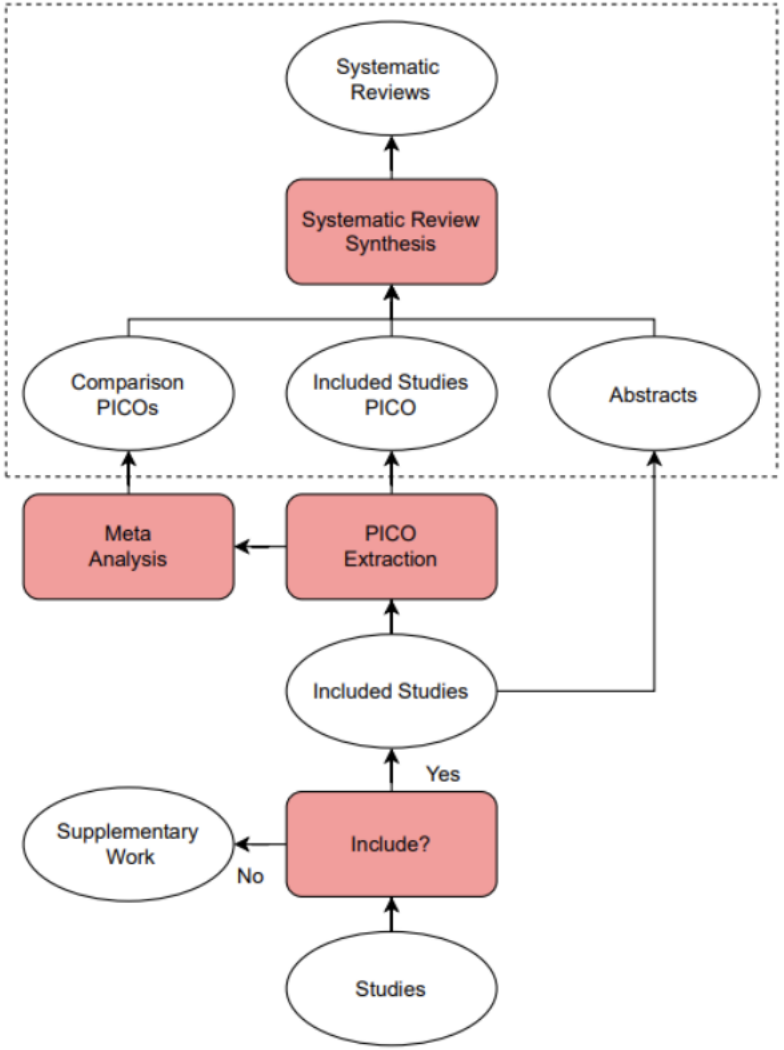
Beyond Abstracts: A New Dataset, Prompt Design Strategy and Method for Biomedical Synthesis Generation

James O'Doherty, Cian Nolan, Yufang Hou, Anya Belz
(ACL 2024 SRW)

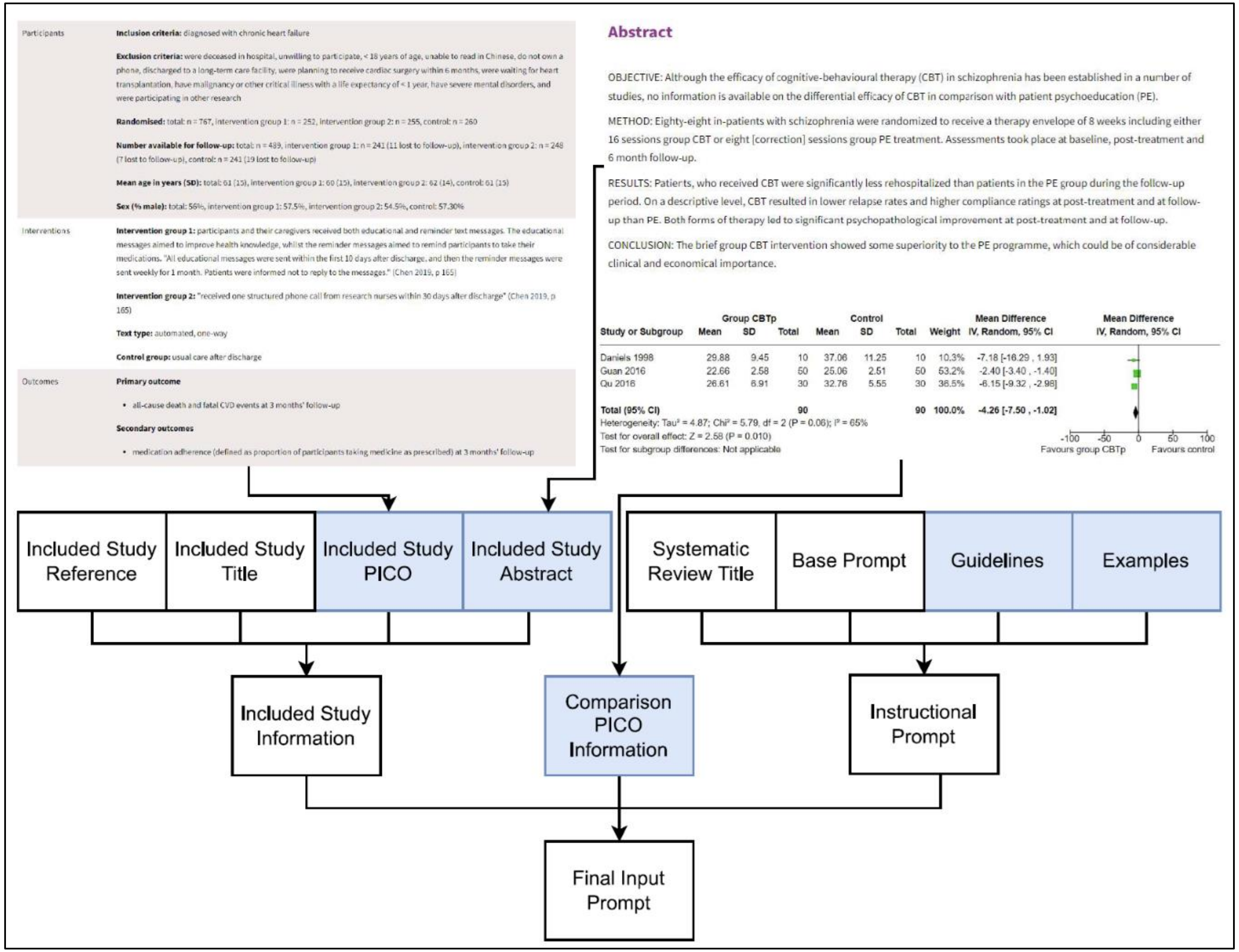



Biomedical Synthesis Generation

Manual systematic review process



Automate the last step with gold PICO information



Some Interesting Findings

- **Best results are achieved when leaving out the included study abstracts**

GPT-4o as a judge

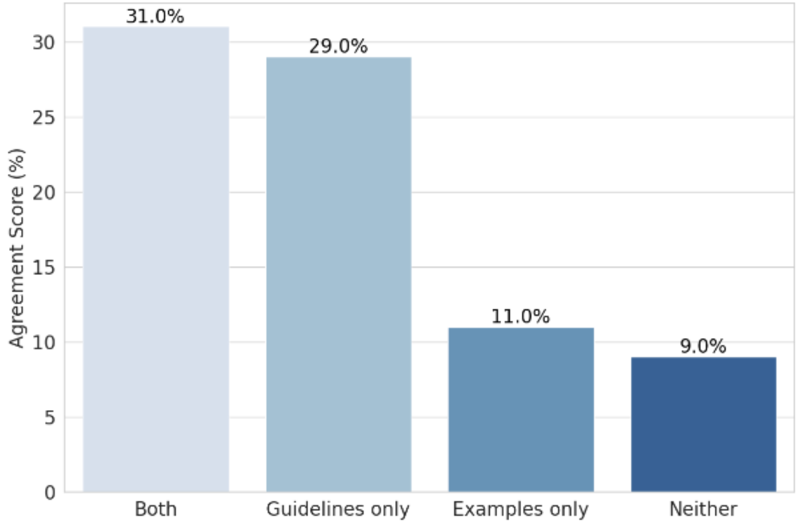
Agreement percentage with reference

	Info included in prompt					Model	BLEU	R-1	R-2	R-L	chrF	LLM Sc.	Agr. Per.
	Abs	PICO		Prompt									
		Inc.	Comp.	Guide.	Exam.								
1		✓	✓	✓	✓	Haiku	**0.358**	**0.291**	0.080	**0.253**	0.481	**2.644**	**53.33**
2	✓	✓	✓	✓	✓	Haiku	0.353	0.288	**0.083**	0.255	**0.486**	2.555	51.11
3		✓	✓	✓	✓	Sonnet	0.349	0.272	0.062	0.239	0.467	2.622	48.89
4	✓	✓	✓	✓	✓	Sonnet	0.344	0.265	0.060	0.231	0.466	2.533	46.67
5	✓	✓		✓	✓	Haiku	0.344	0.267	0.064	0.232	0.466	2.444	44.44
6	✓		✓	✓	✓	Haiku	0.341	0.281	0.071	0.246	0.474	2.288	35.56
7	✓			✓	✓	Haiku	0.343	0.265	0.063	0.229	0.463	2.022	31.11
8	✓	✓	✓			Haiku	0.360	0.270	0.063	0.233	0.460	2.067	28.89
9	✓			✓		Haiku	0.243	0.253	0.063	0.230	0.430	2.356	28.89
10	✓				✓	Haiku	0.278	0.240	0.055	0.219	0.444	1.756	11.11
11	✓					Haiku	0.270	0.232	0.056	0.210	0.437	1.778	8.89

Claude Haiku/Sonnet: context window size is 200k

Some Interesting Findings

- **System instructions informed by domain knowledge gleaned from textbooks are essential components**



The following is a summary of the instructions given to Cochrane Reviewers for drafting the Authors' Conclusions section of a systematic review:

Implications for Practice: Cochrane Reviews provide valuable information for practice but do not make direct recommendations due to the need for additional evidence and judgments. Authors should discuss the certainty of evidence, benefits versus harms, and patient va ues/preferences without making specific recommendations. If authors discuss possible actions, they should consider all factors influencing decisions, including patient-important outcomes, costs, and resource availability.

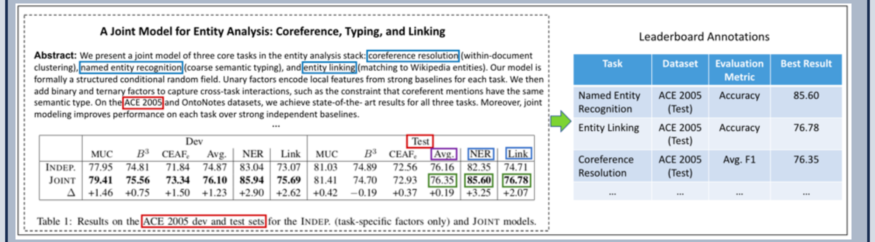
…

Final Remarks

Build Global Scientific Evidence Map

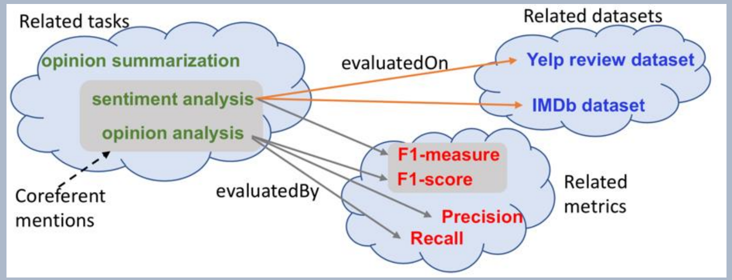
Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- ➢ PDF Table Parser - extract tables from papers in PDF format
- ➢ https://github.com/IBM/science-result-extractor



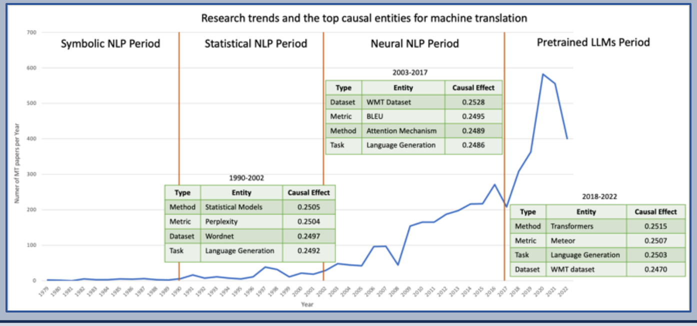
NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- ➢ TDM Tagger – extract task/dataset/metric entities from NLP papers [Hou et al., EACL 2021]



A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

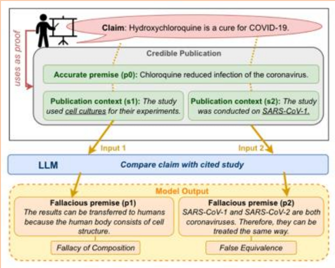
- ➢ NLP Concepts Causal Analysis



Scientific Communication

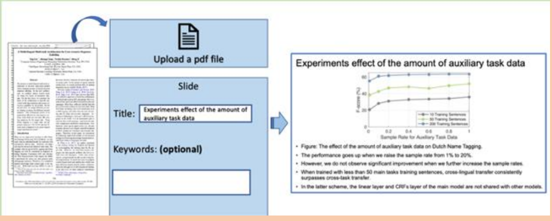
Missci: Reconstructing Fallacies in Misrepresented Science
[Glockner et al., ACL 2024]

- ➢ Tackle health-related misinformation



Interactive Doc2slides Generation
[Sun et al., NAACL 2021]

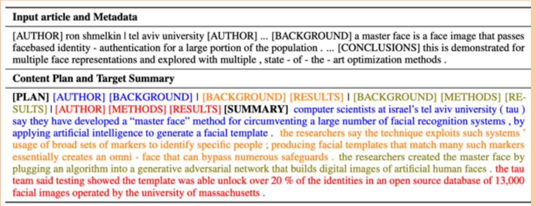
Scientific Diagrams Generation
[Mondal et al., EMNLP 2024 Findings]



- ➢ https://github.com/IBM/document2slides

Science Journalism Generation
[Cardenas et al., EMNLP 2023]

- ➢ Controlled generation based on discourse structures



Input article and Metadata

[AUTHOR] ron shmelkin | tel aviv university [AUTHOR] ... [BACKGROUND] a master face is a face image that passes facebased identity - authentication for a large portion of the population . ... [CONCLUSIONS] this is demonstrated for multiple face representations and explored with multiple , state - of - the - art optimization methods .

Content Plan and Target Summary

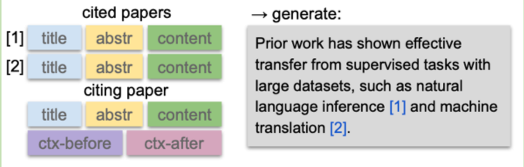
[PLAN] [AUTHOR] [BACKGROUND] | [BACKGROUND] [RESULTS] | [BACKGROUND] [METHODS] [RESULTS] | [AUTHOR] [METHODS] [RESULTS] **[SUMMARY]** computer scientists at israel's tel aviv university (tau) say they have developed a "master face" method for circumventing a large number of facial recognition systems , by applying artificial intelligence to generate a facial template . the researchers say the technique exploits such systems ' usage of broad sets of markers to identify specific people ; producing facial templates that match many such markers essentially creates an omni - face that can bypass numerous safeguards . the researchers created the master face by plugging an algorithm into a generative adversarial network that builds digital images of artificial human faces . the tau team said testing showed the template was able unlock over 20 % of the identities in an open source database of 13,000 facial images operated by the university of massachusetts .

Scientific Knowledge Synthesis

CiteBench: Benchmark for Citation Text Generation **[Funkquist et al., EMNLP 2023]**

Citation Text Generation with LLMs
[Şahinuç et al., ACL 2024]

Biomedical Synthesis Generation **[O'Doherty et al., ACL 2024 SRW]**



Final Remarks

Build Global Scientific Evidence Map

Scientific Leaderboards Construction
[Hou et al., ACL 2019; Şahinuç et al., EMNLP 2024]

- PDF Table Parser - extract tables from papers in PDF format

NLP TDM Knowledge Graph
[Mondal et al., ACL Findings 2021]

- TDM Tagger – extract task/dataset/metric entities

A Diachronic Analysis of NLP Research Areas
[Pramanick et al., EMNLP 2023]

- NLP Concepts Causal Analysis

- Recent advances in LLMs and multi-agent frameworks make this an exciting time to develop human-centered NLP models and applications in AI4Science

- More work is needed to better understand the role of AI systems in facilitating scientific research
  - Generating research ideas?
  - Co-writing papers?
  - Reviewing?

  …

Acknowledgement

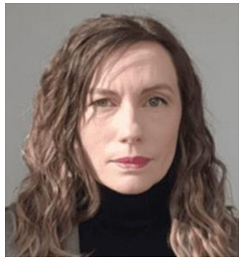

Thy Thy Tran


Ilia Kuznetsov


Saif M. Mohammad


Preslav Nakov

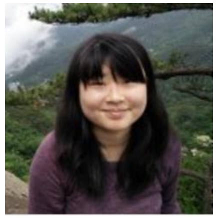

Iryna Gurevych


Anya Belz


Yulia Grishina


Bei Chen


Dakuo Wang


Nancy Wang


Charles Jochim


Martin Gleize


Yunfeng Zhang


Francesca Bonin


Debasis Ganguly


Edward Sun


Ishani Mondal


Aniket Pramanick


Ronald A. Cardenas


Martin Funkquist


Arthur Yao


Max Glockner


Furkan Şahinuç

Hiring Postdocs and PhD Students

interdisciplinary
transformation
university austria

SCAN ME

Yufang Hou

Professor for Natural Language Processing

Prof.

yufang.hou@it-u.at

PhD Positions in the field of Computational X

3 PhD positions

**Application deadline:
November 30, 2024**

:read more   :apply now

Postdoctoral Positions in the field of Computational X

2 Postdoctoral positions

**Application deadline:
November 30, 2024**

:read more   :apply now

Thanks

Q&A