# Towards automatic detection of citation accuracy errors

M. Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, **Halil Kilicoglu***

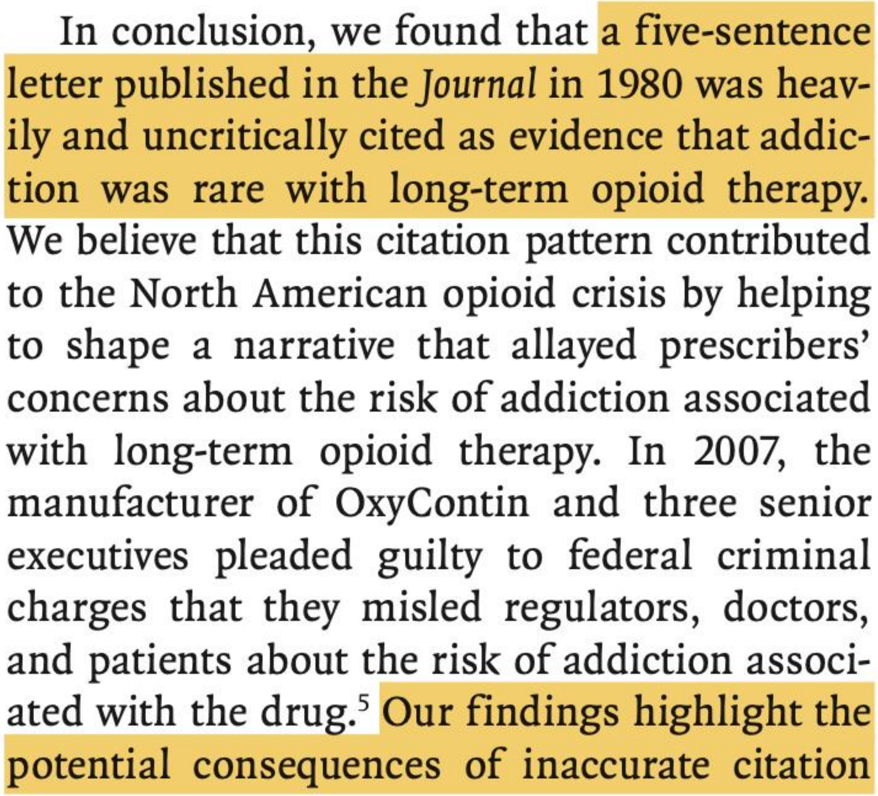
University of Illinois Urbana-Champaign

# Citation integrity

- Accurate citation is a cornerstone of reliable, trustworthy research
- Citations are also widely used in measuring research impact
- But there is little accountability for citations
- Questionable research practices around citation are common
  - Citation padding, citation stacking, citation cartels

- Citation integrity is a vital part of research integrity

# Inaccurate citations can be harmful to our health!



In conclusion, we found that a five-sentence letter published in the *Journal* in 1980 was heavily and uncritically cited as evidence that addiction was rare with long-term opioid therapy. We believe that this citation pattern contributed to the North American opioid crisis by helping to shape a narrative that allayed prescribers' concerns about the risk of addiction associated with long-term opioid therapy. In 2007, the manufacturer of OxyContin and three senior executives pleaded guilty to federal criminal charges that they misled regulators, doctors, and patients about the risk of addiction associated with the drug.[5] Our findings highlight the potential consequences of inaccurate citation

*The* NEW ENGLAND JOURNAL *of* MEDICINE

CORRESPONDENCE

A 1980 Letter on the Risk of Opioid Addiction

Leung PT, Macdonald EM, Stanbrook MB, Dhalla IA, Juurlink DN. A 1980 letter on the risk of opioid addiction. *NEJM*. 2017;376(22):2194-5.

# Citation accuracy in biomedicine

- Citations are rarely examined for accuracy in peer review
  - Metadata errors
  - Citation content errors (quotation errors)
- Quotation errors are especially pernicious
  - Difficult to detect for readers, journals, and peer reviewers
  - Estimated ~25% of medical articles contain quotation errors, half of them severe
  - Citation distortions and biases have led to unfounded claims to be accepted as beliefs in Alzheimer's disease research

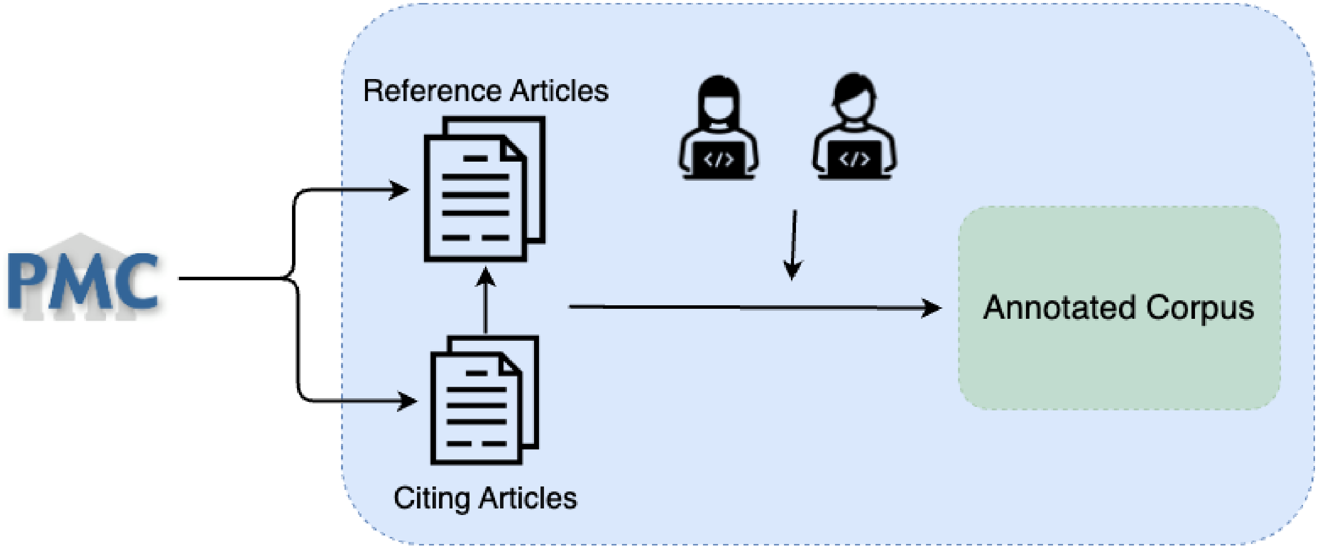
Jergas H, Baethge C. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*. 2015;3:e1364.
Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*. 2009;339.

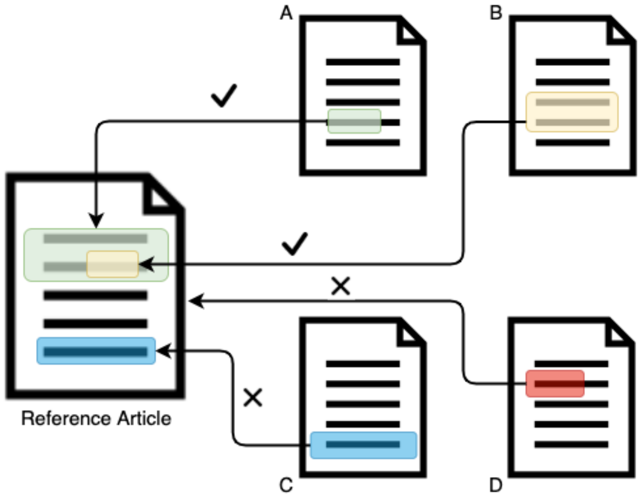
# NLP for citation accuracy

- Assessing citations for accuracy requires considerable manual effort
- Natural language processing (NLP) could support citation verification tools
  - Flag problematic citations for closer scrutiny
  - Trace the provenance of misleading claims/misinformation

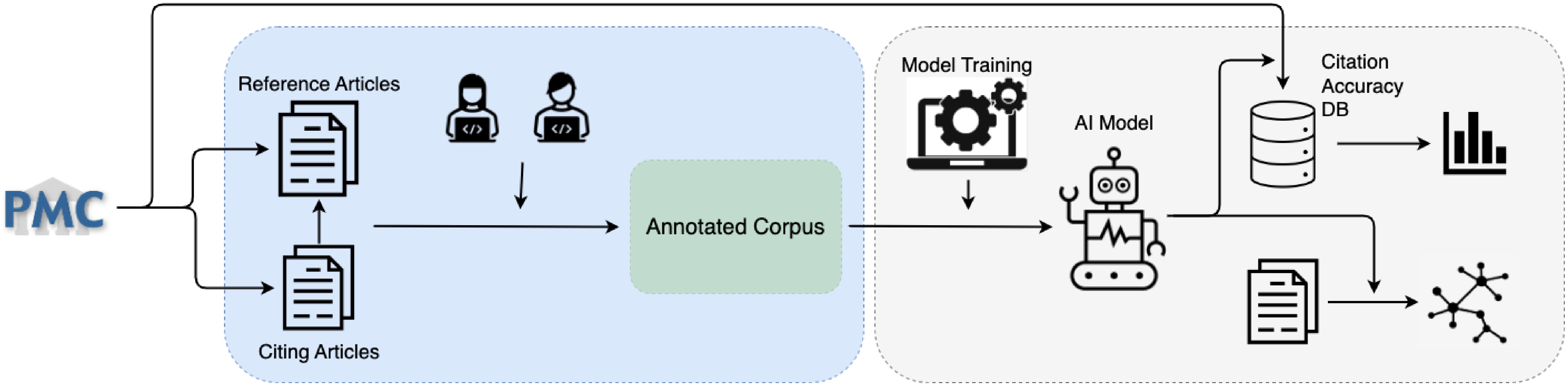
- Labeled data is needed to train and validate NLP models

# Our work

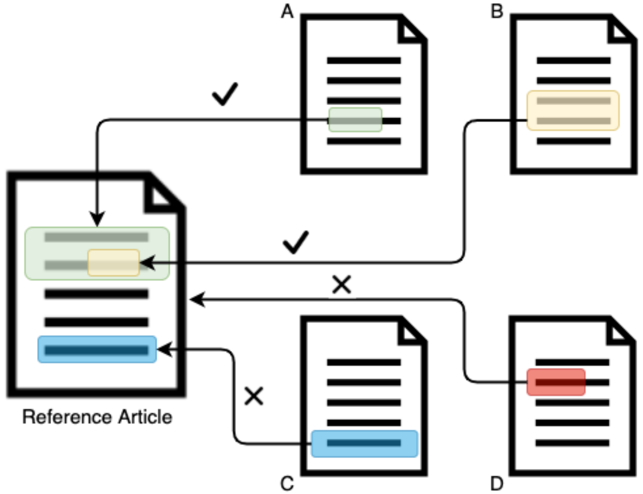

PMC
Reference Articles
Annotated Corpus
Citing Articles


A
B
✓
✓
×
×
Reference Article
C
D

# Our work



PMC
Reference Articles
Citing Articles
Annotated Corpus
Model Training
AI Model
Citation
Accuracy
DB



A
B
C
D
✓
✓
×
×
Reference Article

# Citation accuracy annotation

- ACCURATE
- Major error
  - CONTRADICT, IRRELEVANT, NOT_SUBSTANTIATED
- Minor error
  - OVERSIMPLIFY, MISQUOTE, INDIRECT, ETIQUETTE

- 100 reference articles with 3063 citations
- Graduate and undergraduate students in life sciences

# Citation errors

- Citation: *This is coherent with the fact that hACE2 expression were not observed in the gut of the mice used in that study [37]*.

- Reference: [37] *In the gastrointestinal tract of K18-hACE2 mice, hACE2 was expressed most abundantly in the colon, which correlated with infection seen at later time points.*

Error type: CONTRADICT

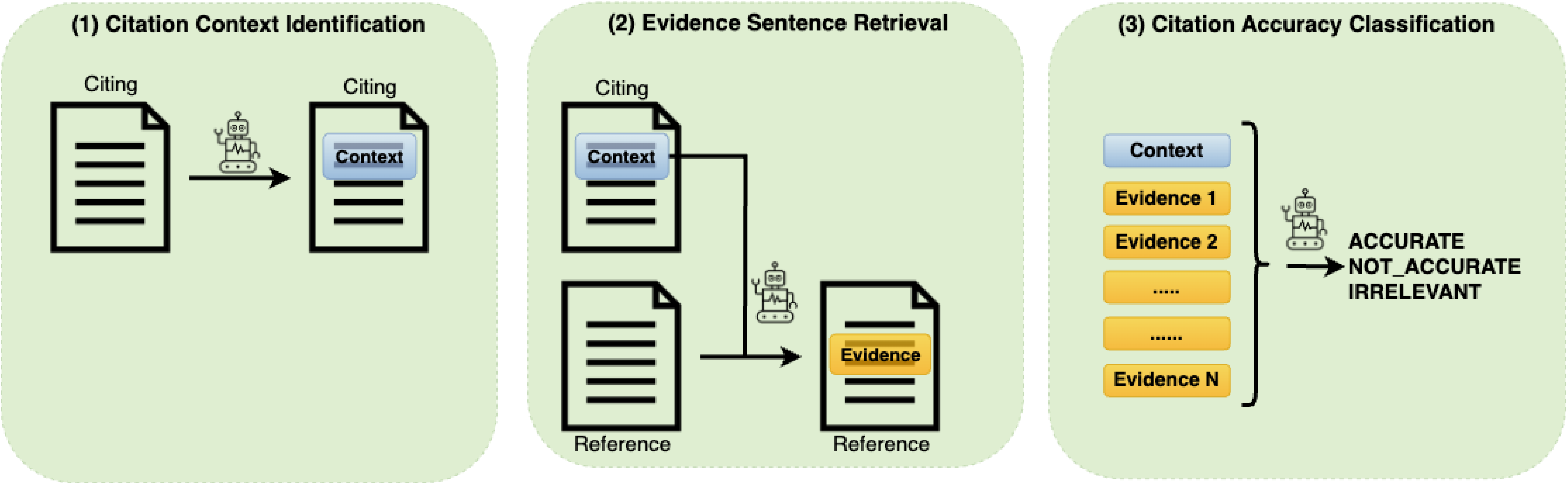
# Annotated corpus

| **Label** | **Total** | **Percentage** |
|---|---:|---:|
| ACCURATE | 1863 | 60.82 |
| *MAJOR* | 552 | 18.02 |
| * CONTRADICT | 92 | 3.00 |
| * IRRELEVANT | 217 | 7.08 |
| * NOT_SUBSTANTIATE | 243 | 7.93 |
| *MINOR* | 648 | 21.16 |
| * MISQUOTE | 38 | 1.24 |
| * OVERSIMPLIFY | 111 | 3.62 |
| * INDIRECT | 82 | 2.68 |
| * ETIQUETTE | 417 | 13.61 |
| Total Errors | 1200 | 39.18 |

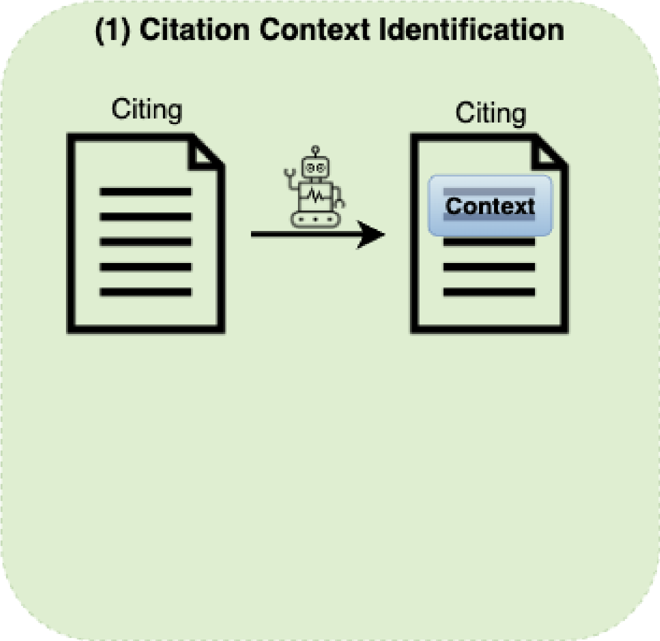
- 1.12 context and 1.24 evidence sentences per citation
- More minor errors than major errors ($p$ = .0085)

- High inter-annotator agreement for citation contexts ($\kappa = 0.96$)
- Fair agreement for evidence sentences ($\kappa = 0.37$) and accuracy labels ($\kappa = 0.31$)

# NLP models


(1) Citation Context Identification
Citing
Citing
Context
(2) Evidence Sentence Retrieval
Citing
Context
Reference
Evidence
Reference
(3) Citation Accuracy Classification
Context
Evidence 1
Evidence 2
......
.......
Evidence N
ACCURATE
NOT_ACCURATE
IRRELEVANT

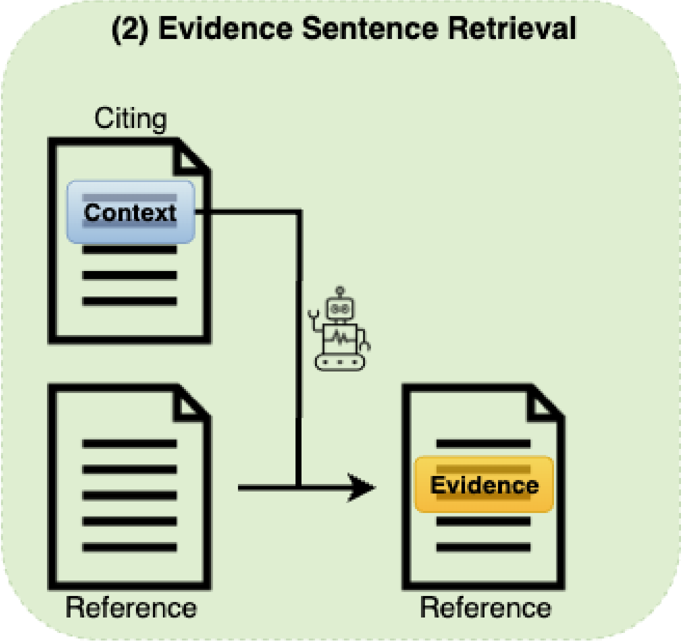
# Citation context classification



- Sentence classification

| Model | Precision | Recall | $F_1$ |
|---|---:|---:|---:|
| Citation sentence only | 1.00 | 0.90 | 0.94 |
| Fine-tuned PubMed-BERT | 0.97 | 0.90 | 0.93 |

# Evidence sentence retrieval

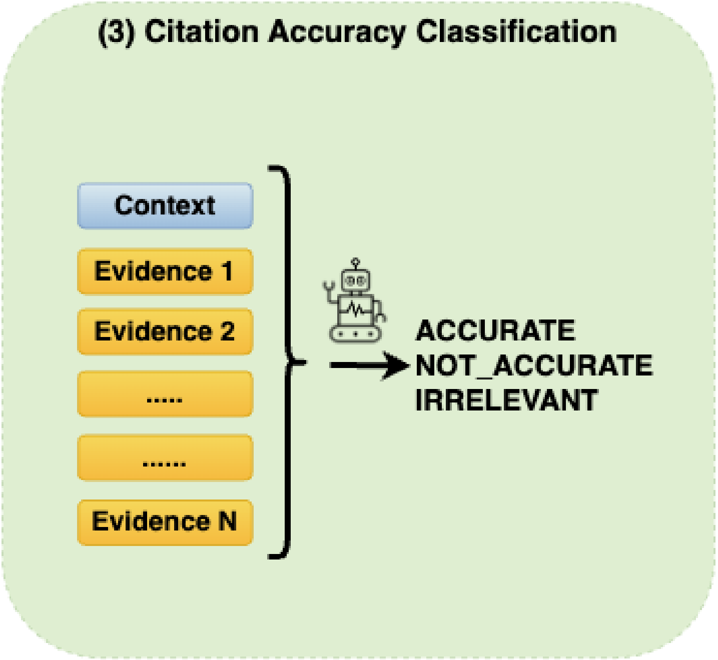


- BM25 (top 60 sentences)
- MonoT5 reranker (k=1, 5, 10, 20)

| **Metric** | |
|---|---|
| Recall@5 | 0.28 |
| Recall@10 | 0.40 |
| Recall@20 | 0.53 |
| MRR | 0.32 |

Nogueira R, Jiang Z, Pradeep R, Lin J. Document Ranking with a Pretrained Sequence-to-Sequence Model. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 708-718).

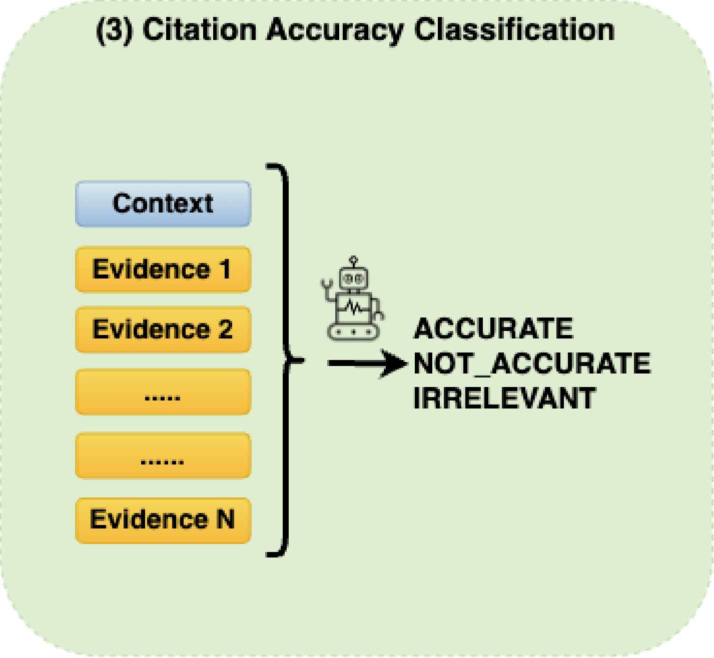
# Citation accuracy classification



- Adapt MultiVerS claim verification model
  - ACCURATE, NOT_ACCURATE, IRRELEVANT
- In-context learning
  - GPT-3.5-turbo, GPT-4
  - Four examples (2 for NOT_ACCURATE)

Wadden D, Lo K, Wang LL, Cohan A, Beltagy I, Hajishirzi H. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 61-76).

# Citation accuracy classification



(3) Citation Accuracy Classification

Context

Evidence 1

Evidence 2

......

......

Evidence N

ACCURATE
NOT_ACCURATE
IRRELEVANT

- MultiVerS

| Evidence Input | ACC. | NOT_ACC. | IRREL. | Macro-$F_1$ |
|---|---|---|---|---|
| Title + abstract | 0.69 | 0.38 | 0.20 | 0.43 |
| Top 5 sentences | 0.69 | 0.43 | 0.37 | 0.50 |
| Top 10 sentences | 0.67 | 0.41 | 0.36 | 0.48 |
| Top 20 sentences | 0.69 | 0.43 | 0.42 | 0.52 |
| Gold evidence | 0.79 | 0.52 | 0.93 | 0.75 |

- In-context learning

| | ACC. | NOT_ACC. | IRREL. | Macro-$F_1$ |
|---|---|---|---|---|
| GPT-3.5 | 0.73 | 0.05 | 0.34 | 0.38 |
| GPT-4 | 0.80 | 0.09 | 0.48 | 0.45 |

# Conclusions

- First publicly available, annotated corpus of citation quotation errors
- Annotation of citation errors is challenging
    - They can be subtle, some subjectivity is involved, domain knowledge is needed
- NLP models need improvement
    - Better evidence sentence retrieval will improve the results
    - GPT models mostly fail at inaccurate citations
- Automated citation verification tools can
    - Support journal workflows
    - Raise awareness around poor citation practices
    - Support meta-research
    - Reduce propagation of untrustworthy information in science

## Thank you! Questions?

halil@illinois.edu

Corpus available at: https://github.com/ScienceNLP-Lab/Citation-Integrity/

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

ORI THE OFFICE OF RESEARCH INTEGRITY ®