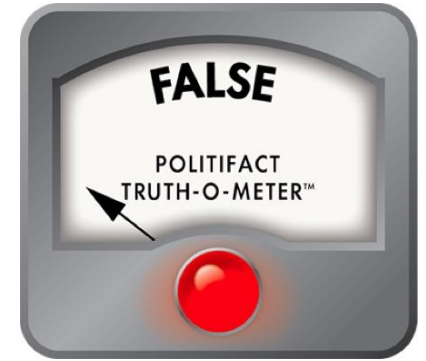
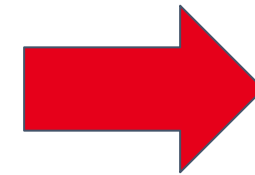
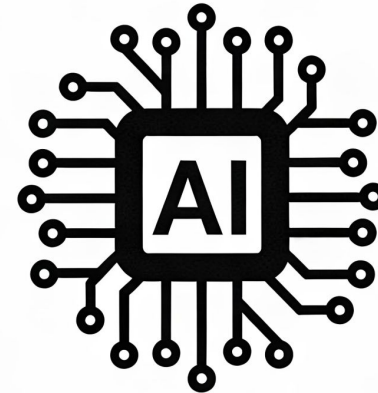
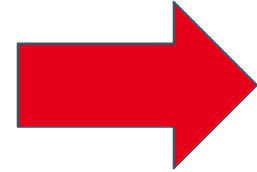
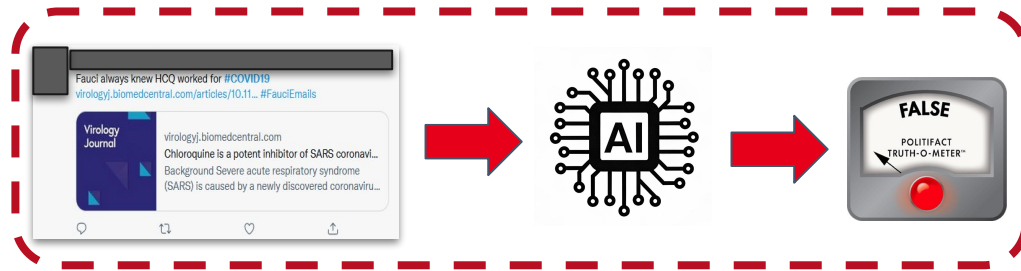


Towards Realistic Evidence-based Fact-Checking with NLP

Our goal: Combat real-world misinformation



Fact-Checking outputs must be transparent and verifiable



Can we **trust** this?

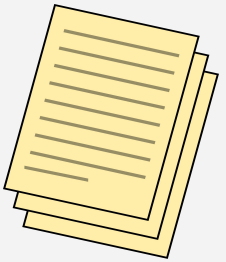


Fact-checkers care about **how** the verdict was derived (Warren et al., 2025).

➡ **Verifiability**

➡ **Transparency**

Model decisions are grounded in evidence



Ground the decision in **evidence**.

Verifiability

Transparency



Our work explores different **realistic relations** between **claim** and **evidence**

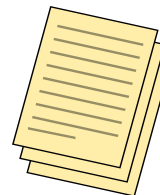
Evidence



refutes

Claim

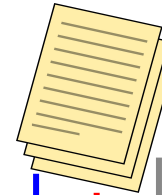
Evidence



supports

Claim

Evidence



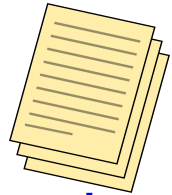
supports

refutes

neutral

Claim

Evidence

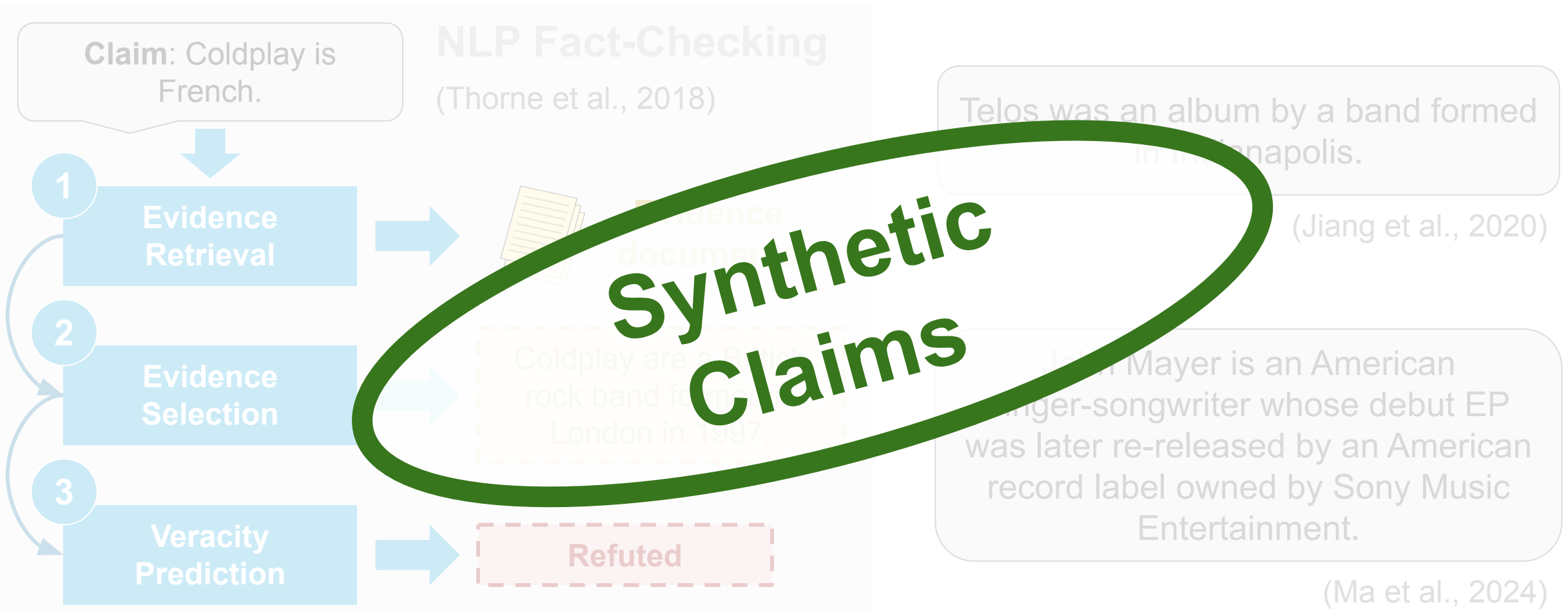


supports

?

Claim

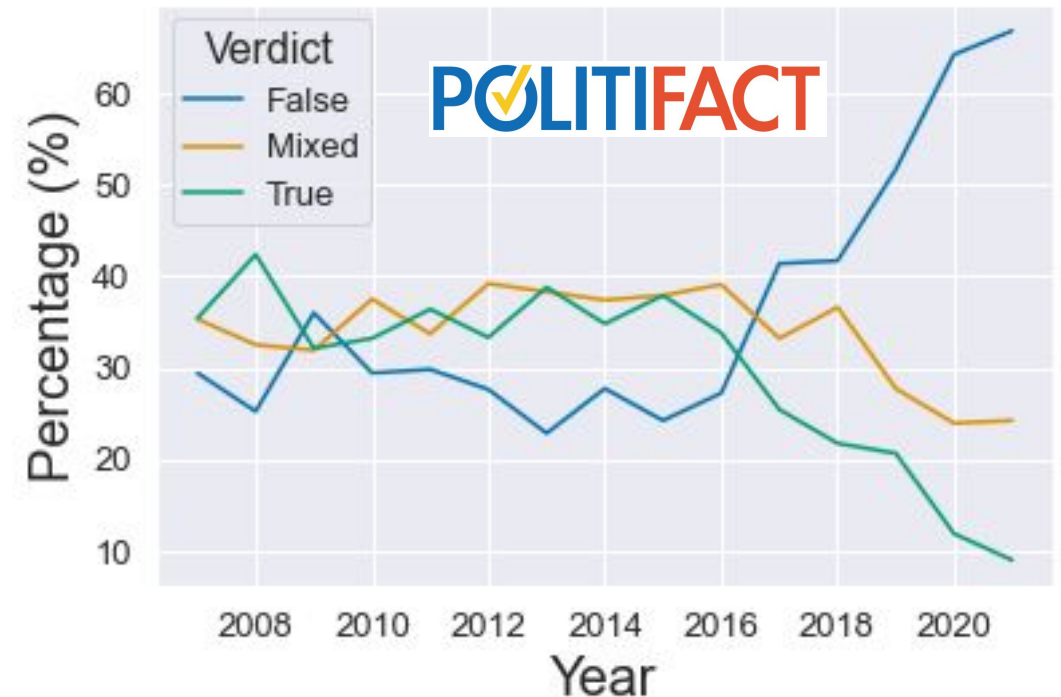
Evidence-based fact-checking in NLP started with synthetic claims



Automated fact-checking transferred to real-world claims

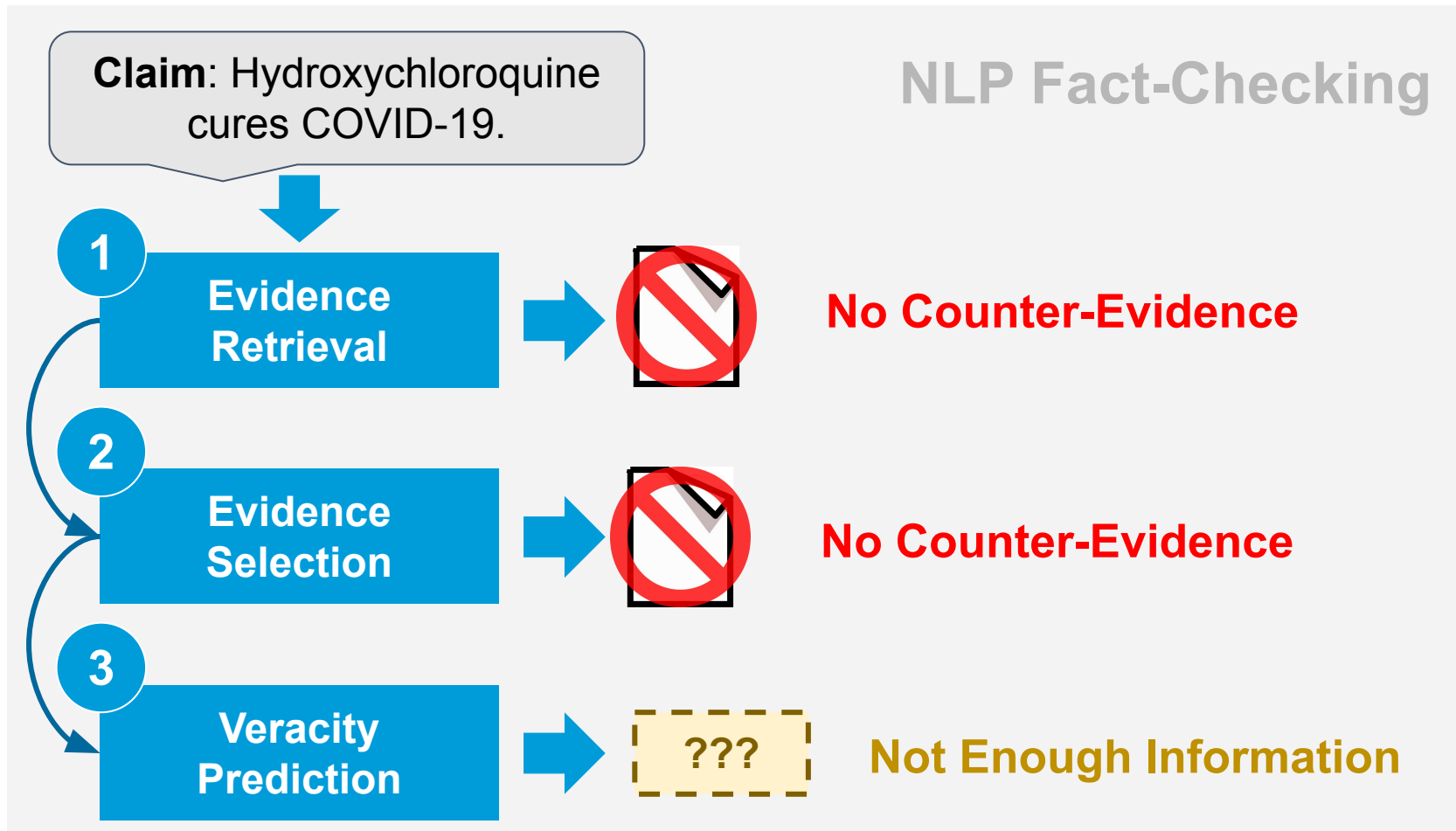
Datasets address claims that have been checked by **professional fact-checking organizations**.

(Hanselowski et al., 2019;
Augenstein et al., 2019;
Khan et al., 2022; *inter alia*)



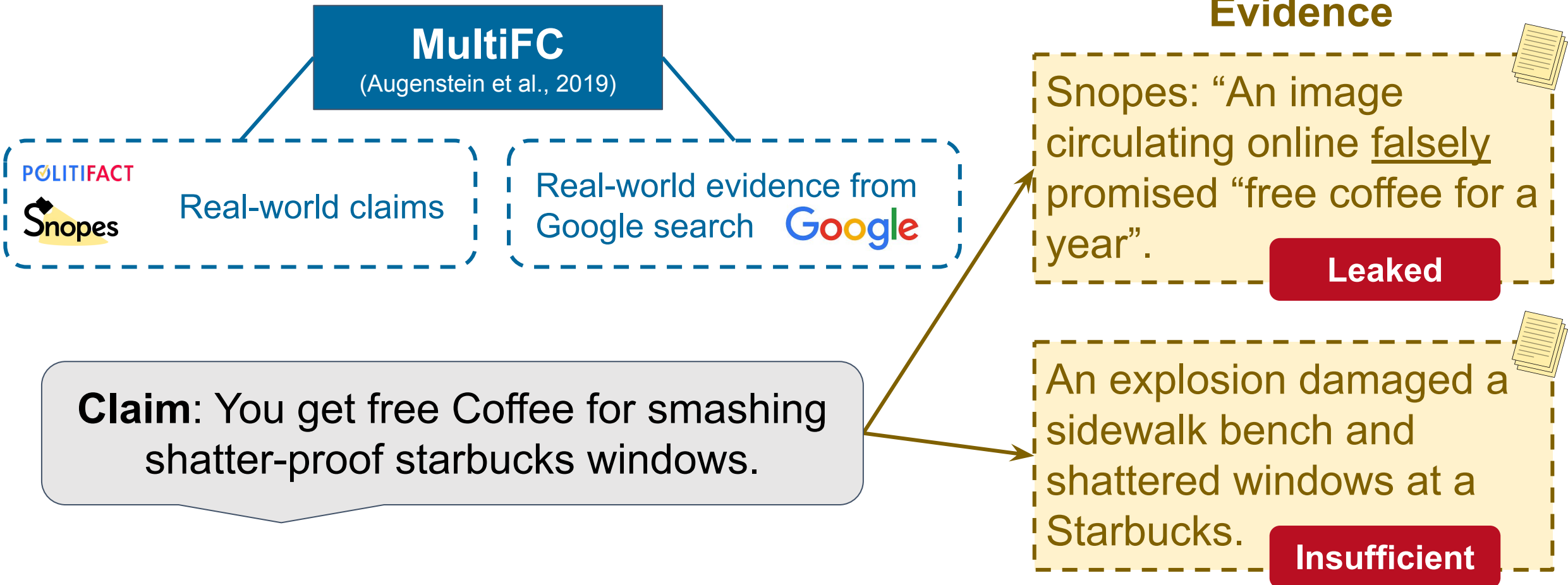
Focus on **misinformation**

Counter-evidence is much rarer for real-world misinformation

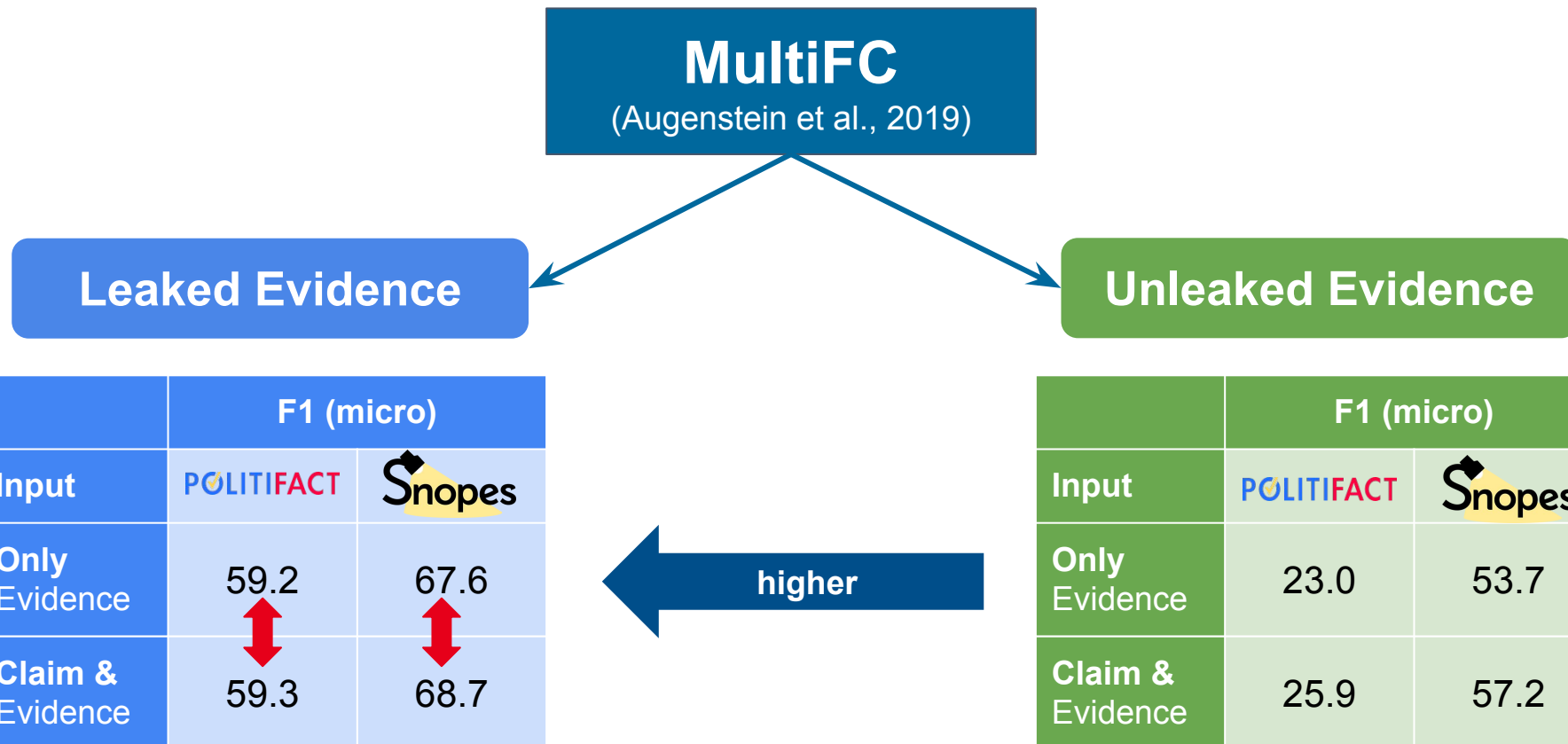


Evidence that
directly refutes a
claim does **not exist**.

Fact-Checking datasets contain leaked or insufficient evidence



Fine-tuned models rely on unrealistic (leaked) evidence



similar
performance

Misinformation thrives in uncertainty with limited evidence

This is not a flaw by the dataset but because misinformation thrives under uncertainty when little evidence is available.



Models must be able to **operate** when **little evidence** is available.

Misinformation is different from accurate claims.

NLP models failed, but journalists not

Claim: Hydroxychloroquine
cures COVID-19.



Chloroquine is a potent inhibitor of SARS
coronavirus infection and spread

[Martin J Vincent](#), [Eric Bergeron](#), ... [Stuart](#)

+ Show authors

 **BMC** Part of Springer Nature

 **Virology Journal**

from 2005

seasonal coronavirus

in vitro



Why?



**Human verification starts with the
source.**

(Buttry, 2014)

Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation (EMNLP 2022)



What did we do about it?

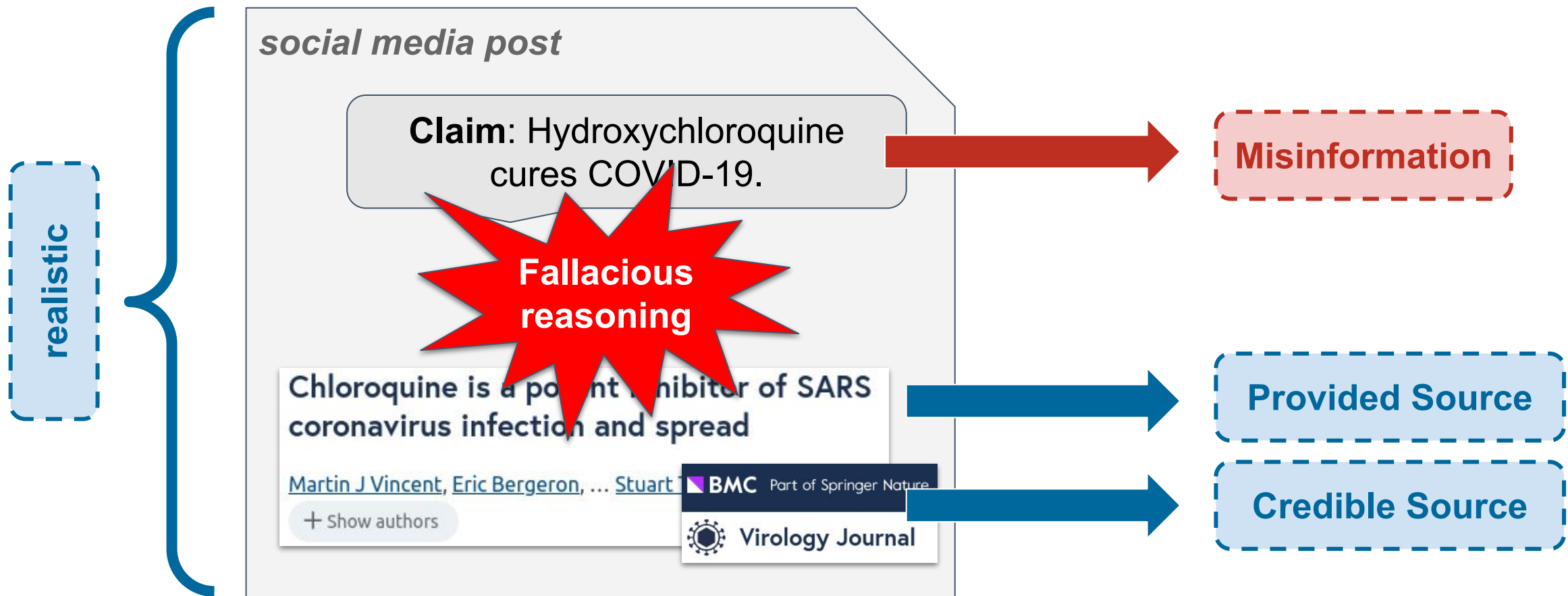
1

We focus on a **subset** of harmful misinformation claims

2

We focus on **simpler** (still realistic) claims.

Misrepresented studies come with credible sources



We reconstruct the fallacious logical arguments

Input

Claim: Hydroxychloroquine cures COVID-19.

Task

Fallacy of Composition

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

False Equivalence

Fallacious Premise 2: *SARS-CoV-1 and SARS-CoV-2 are both coronaviruses. Therefore, they can be treated the same way.*

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

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

Input

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

Missci is created from fact-checking articles

Collect

Select

Reconstruct

527 fact-checking articles

Select **scientific** publications that are **misrepresented**.

Manually **reconstruct** the **fallacious arguments**.

8,695 linked documents

208 misrepresented studies

184 fallacious arguments



Health Feedback

150 fact-checking articles

435 necessary fallacious premises

LLMs are tested in zero-shot experiments

Task: Verbalize the **fallacious premise** and predict the **fallacy class**.

Explore impact of **different fallacy information** in zero-shot experiments.

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: *Each brick in that building weighs less than a pound. Therefore, the building weighs less than a pound.*

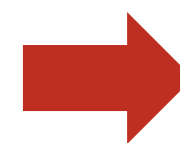
Results: Verbalize fallacious premise and predict fallacy class

Nine fallacy classes

LLM	Prompt	P@1
Llama 2 (70B)	Def.	<u>22.3</u>
	Def. + Expl.	20.9
	Def. + Logical	19.6
	Def. + Logical + Expl.	20.9
	Logical	19.3
	Logical + Expl.	21.2
GPT 4	Def.	31.7
	Logical	29.2

P@1 because sometimes **multiple fallacies** may apply but only one is needed

Based on **human evaluation**



50.3

48.1

Results: Predict fallacy class *given the fallacious premise*

Fallacious premise + fallacy class

only fallacy class

LLM	Prompt	P@1	Acc
Llama 2 (70B)	Def.	<u>22.3</u>	57.7
	Def. + Expl.	20.9	63.7
	Def. + Logical	19.6	63.0
	Def. + Logical + Expl.	20.9	56.8
	Logical	19.3	60.1
	Logical + Expl.	21.2	<u>64.5</u>
GPT 4	Def.	31.7	73.8
	Logical	29.2	74.4
	Logical + Expl.	<i>n/a</i>	77.1

Nine fallacy
classes

Fallacy detection **without
explicitly verbalized
fallacious reasoning is
challenging.**

MissciPlus: we link fallacious arguments to real-world studies

Input

Claim: Hydroxychloroquine cures COVID-19.

Task

Fallacy of Composition

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

Input

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

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

Simple summary sentences based on the fact-checking articles.

MissciPlus: we link fallacious arguments to real-world studies

Input

Claim: Hydroxychloroquine cures COVID-19.

Task

Fallacy of Composition

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

Input

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

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

ACE2 flow cytometry analysis and biosynthesis

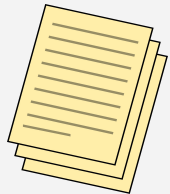
Vero E6 cells were seeded in Dulbecco's modified Eagle medium (Invitrogen) supplemented with 10% fetal bovine serum. The next day, the cells were incubated in Opti-MEM (Invitrogen) in the presence or absence of 10 μ M chloroquine or 20 mM NH_4Cl . To analyze the levels of ACE2 at the cell surface, cells were incubated on ice with 10 μ g/mL affinity-purified goat anti-ACE2 antibody (R&D Systems) and then incubated with FITC-labeled swine anti-goat IgG antibody (Caltag Laboratories). Labeled cells were analyzed by flow cytometry with a FACSCalibur flow cytometer (BD Biosciences). [...]

Three subtasks to reconstruct fallacious arguments in the wild



MissciPlus contains claim-evidence pairs

Evidence



Claim



What will the **fact-checking
model** say?

LLMs consider misinformation accurate despite knowing better

			Predicted as					
Parametric knowledge	Prompt	LLM	True	False	Neutral	True	False	Neutral
	Write a fact-checking article	Llama2 70B	1.6	61.1	37.3	34.7	22.3	41.3
		Llama3 8B	0.0	86.9	2.4	20.0	43.3	14.3
		GPT 3.5	0.8	71.0	28.2	46.7	17.3	35.7
		GPT 4	0.0	85.3	14.7	59.0	23.0	17.3
	To the best of your knowledge	Llama2 70B	0.0	100.0	0.0	29.7	69.3	1.0
		Llama3 8B	8.3	88.9	2.4	68.7	26.0	3.0
		GPT 3.5	1.6	50.4	48.0	47.3	6.0	45.0
		GPT 4	3.6	68.3	27.8	49.7	6.7	36.3
	RAG	Based on the study	Llama2 70B	23.8	61.5	12.7	58.7	29.7
Llama3 8B			44.4	53.2	2.4	80.3	16.3	3.3
GPT 3.5			38.9	31.7	29.0	78.0	5.3	16.0
GPT 4			27.4	34.1	38.5	55.0	4.0	41.0
			Misinformation			True Claims		

Missci: Reconstructing Fallacies in Misrepresented Science **(ACL 2024)**

Grounding Fallacies Misrepresenting Scientific Publications in Evidence (NAACL 2025)



What else happened since our 2022 publication?

AVeriTeC	MissciPlus	ClaimDecomp
 LLMs may memorize the fact-checking articles from pretraining		
 Sufficient evidence  Crowdworquer questions  No restrictions for claims (Schlichtkrull et al., 2023)	 Sufficient evidence  Following journalists  Not applicable to all misinformation	 No evidence validation  Following journalists  No restrictions for claims (Chen et al., 2022, 2024)

What did we do about it?

1

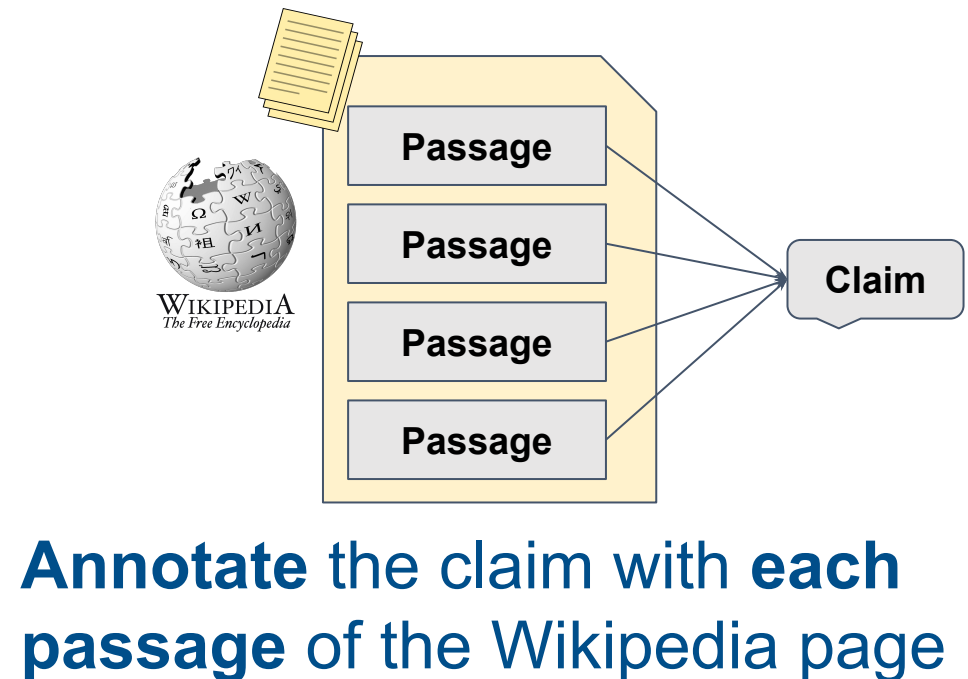
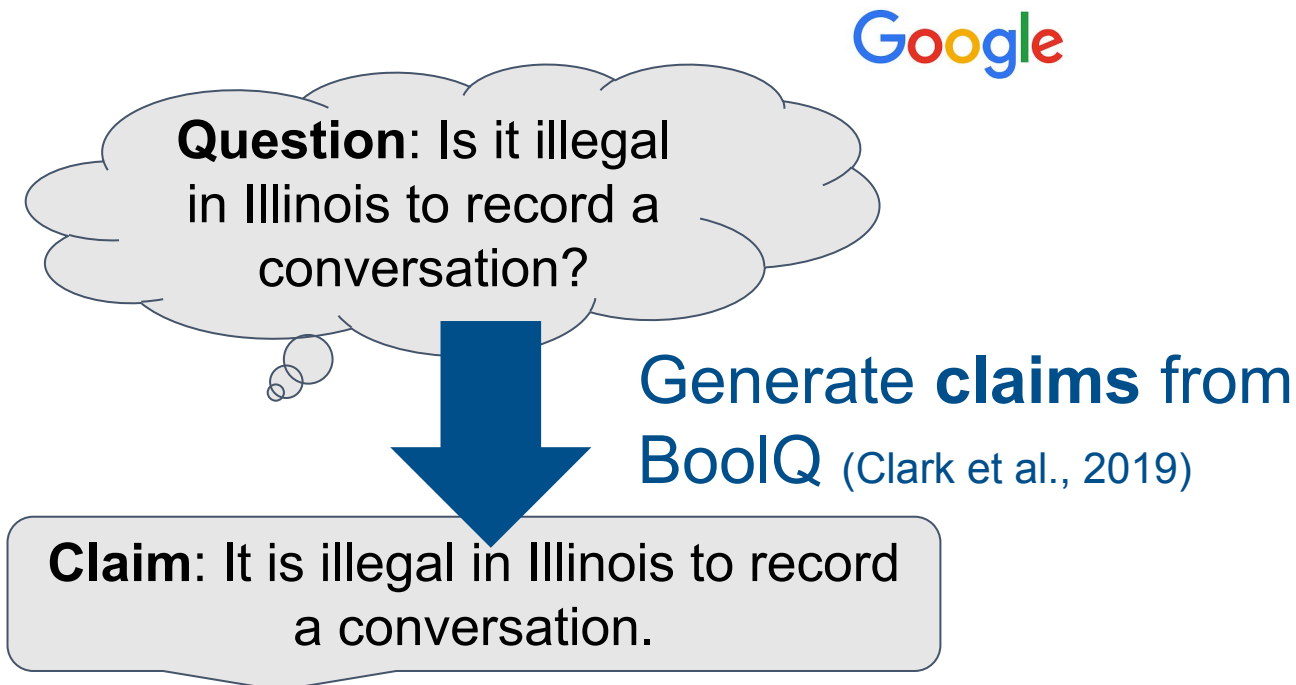
We focus on a **subset** of harmful misinformation claims

2

We focus on **simpler** (still realistic) claims.

We look at the factuality of realistic but harmless claims

We need **realistic claims** and **realistic independent evidence** documents



We split AmbiFC into an uncertain and certain subset

Annotators can say if they are **uncertain** for each instance.

Certain instances (5+ annotations)

Krippendorff's Alpha:

Sentence-level: **0.560**

Passage-level: **0.553**

**Annotators
agree**

Uncertain instances (5+ annotations)

Krippendorff's Alpha:

Sentence-level: **0.302**

Passage-level: **0.206**

**Annotators
disagree**

Human annotators disagree claims from AmbiFC

Input

Claim: It is illegal in Illinois to record a conversation.

Wikipedia Passage (Evidence)

From: *Illinois wiretapping law*

...

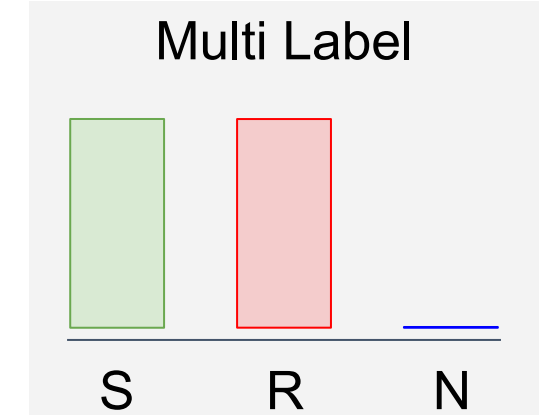
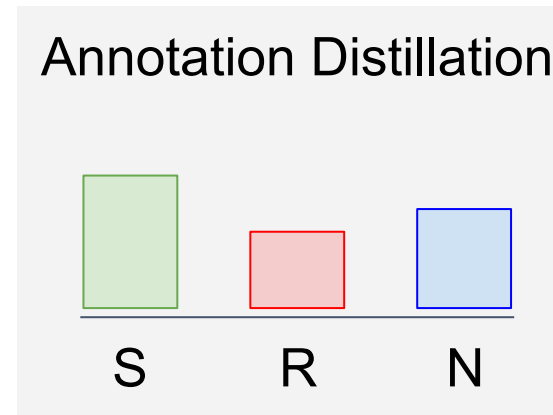
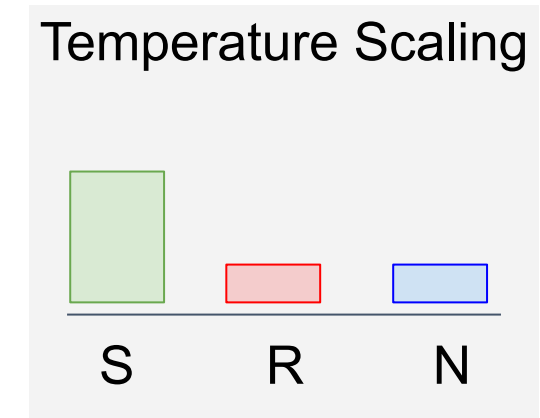
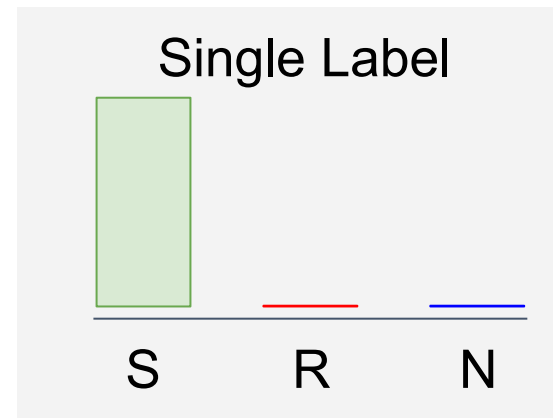
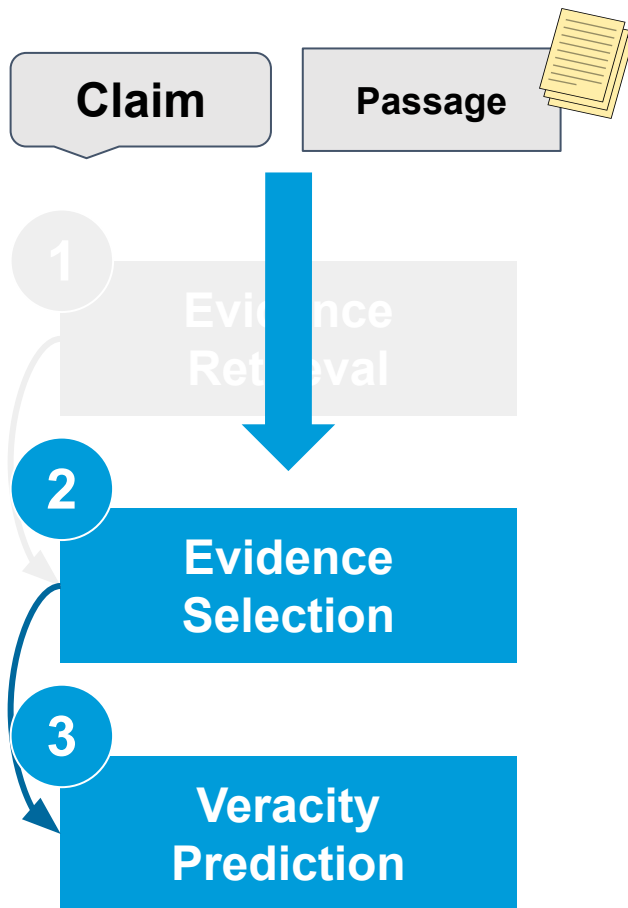
SB1342 makes changes to the original language of the wiretapping law, adding that in order to commit a criminal offense, a person must be recording "in a surreptitious manner".

...

Neutral: It depends, so we cannot really say yes/no.

Refuted: Just recording a conversation is not illegal.

We develop different strategies to fact-check ambiguous claims



Models predict the human label distributions on AmbiFC

	veracity	evidence + veracity
Modeling	DistCS	w-DistCS
Single-label	76.0	55.2
Temp. Scaling	79.5	58.4
Multi-label	78.2	56.6
Annot. Distillation	<u>82.6</u>	<u>60.3</u>



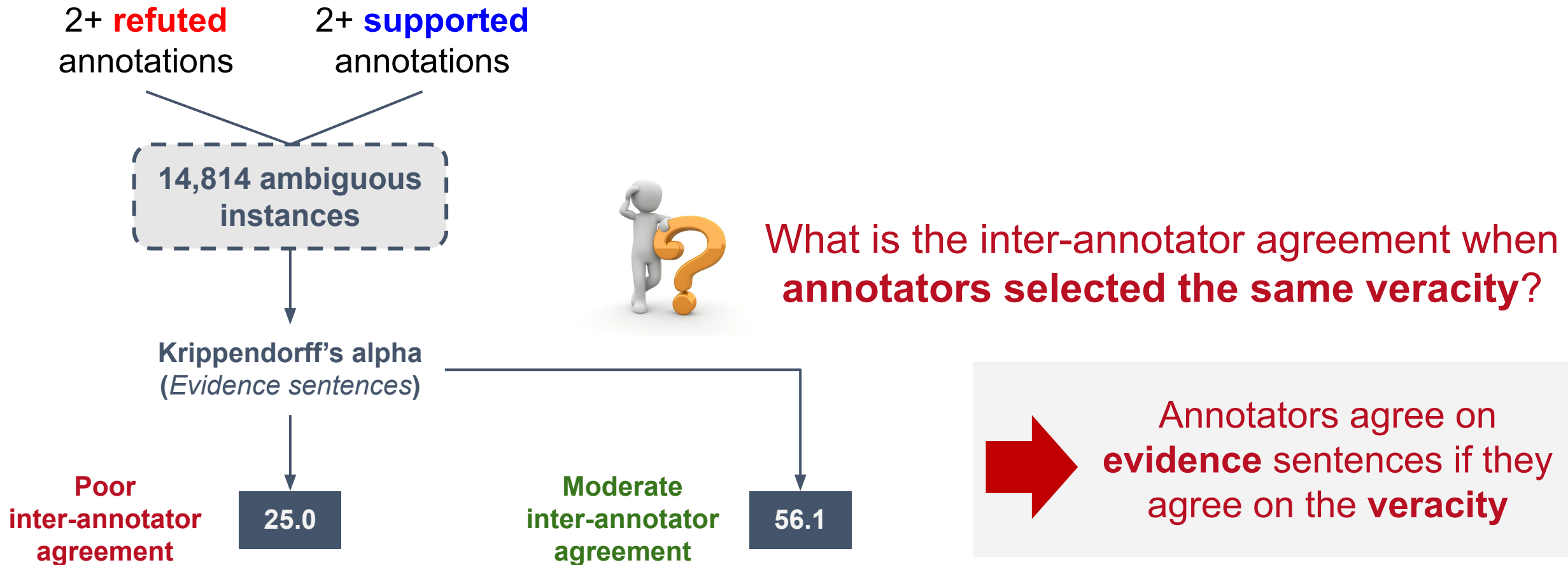
certain: 90.5

uncertain: 75.1



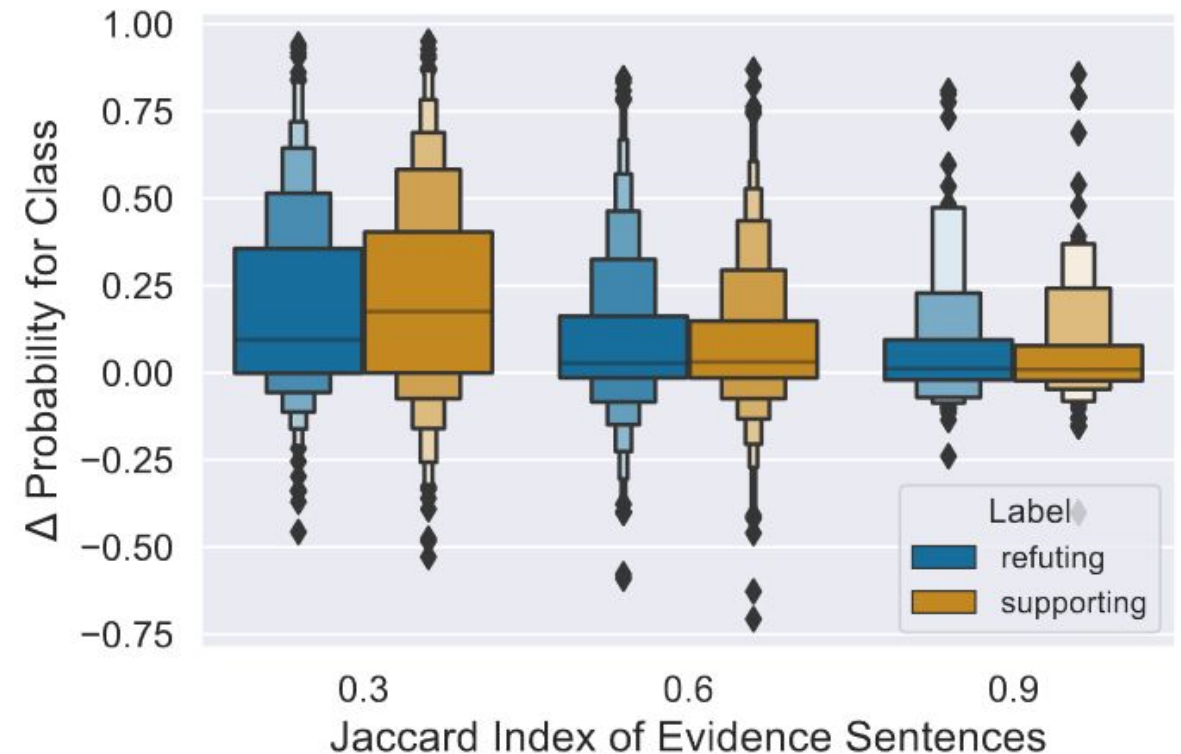
Selecting the same **evidence** as humans is **challenging**.

Humans chose different veracity based on different sentences



Models distinguish between supporting and refuting sentences

If we give the model only **supporting supporting sentences**, will the probability for **supported (veracity)** go up?



AmbiFC: Fact-Checking Ambiguous Claims with Evidence (TACL 2024)



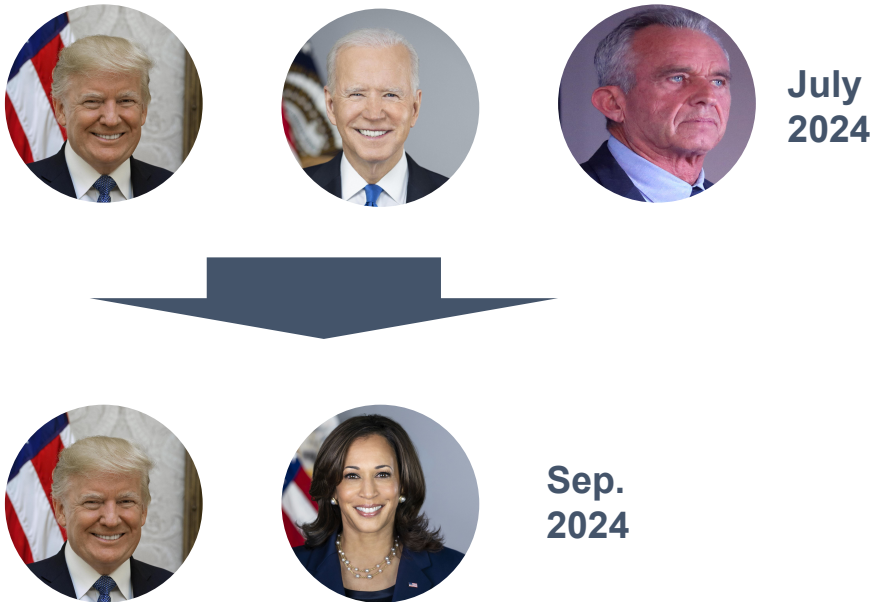
Do LLMs answer faithful based on evidence?



Is the LLM response **backed by the evidence?**

Many applications require the LLM to reason over evidence

Facts change



People care about **recent** events

Enter a keyword to see how it's trending

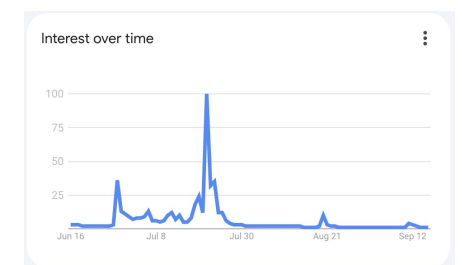
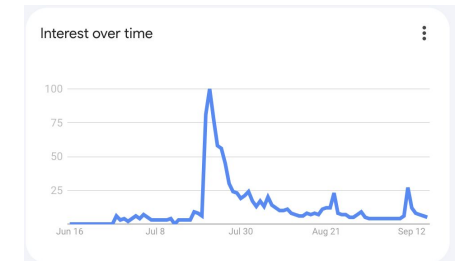
Kamela Harris [Explore](#)

Kamela Harris
Search term

Enter a keyword to see how it's trending

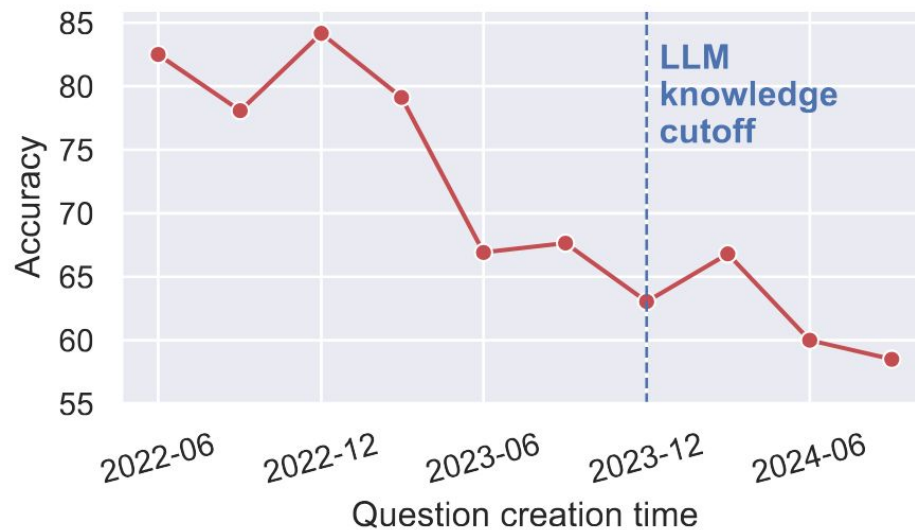
Joe Biden [Explore](#)

Joe Biden
Search term

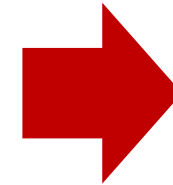


Evidence-based QA datasets from real-world data become stale

Many real-world evidence-based reasoning datasets exist



GPT-4 Turbo on RealTimeQA
(Kasai et al., 2023)



**Datasets based on
real-world data:**

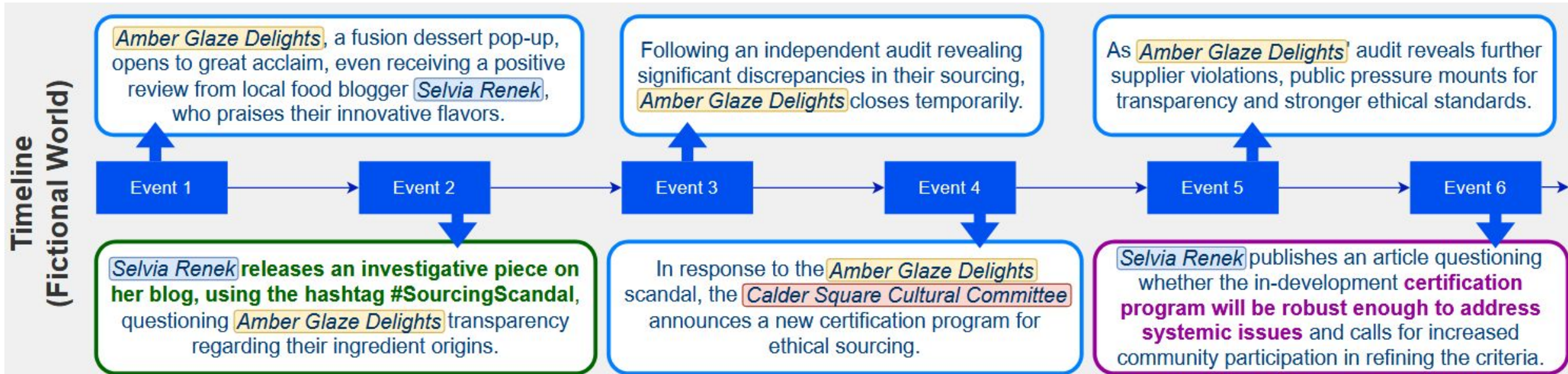
- ➔ become stale
- ➔ do not simulate “unknown” condition

We create timelines of fictional events with fictional entities



LLMs **cannot** rely on their
parametric knowledge

NeoQA contains timelines of fictional events



Each timeline in **NeoQA** contains **10 consecutive events**.

Each **event** contains an outline of **20-30 unique sentences**.

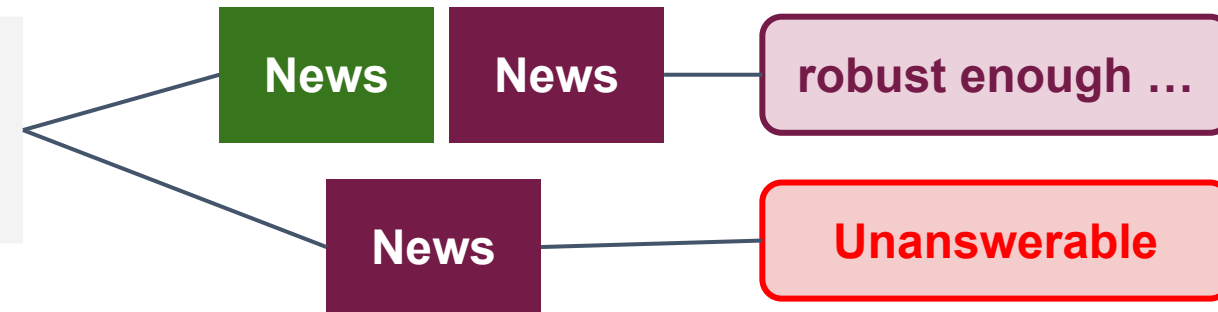
Questions and news article are grounded in outline items

Selvia Renek releases an investigative piece on her blog using the hashtag #SourcingScandal [...]

Selvia Renek published an article [...] whether the certification program will be **robust enough** to address systemic issues.

Multi-hop question

What did the **author of an investigative piece associated with the hashtag #SourcingScandal** question about the **ethical certification program**?



Unanswerable question

What did the **author of an investigative piece associated with the hashtag #SourcingScandal** question about the **cost of the ethical certification program**?



We can combine question and news articles arbitrarily

We know which **information** is reported in a **news article**.

We know which **information** we need to answer a **question**.

We control how to assemble question and evidence

sufficient
evidence

insufficient
evidence

irrelevant
news articles

LLMs fail to deflect on unanswerable questions

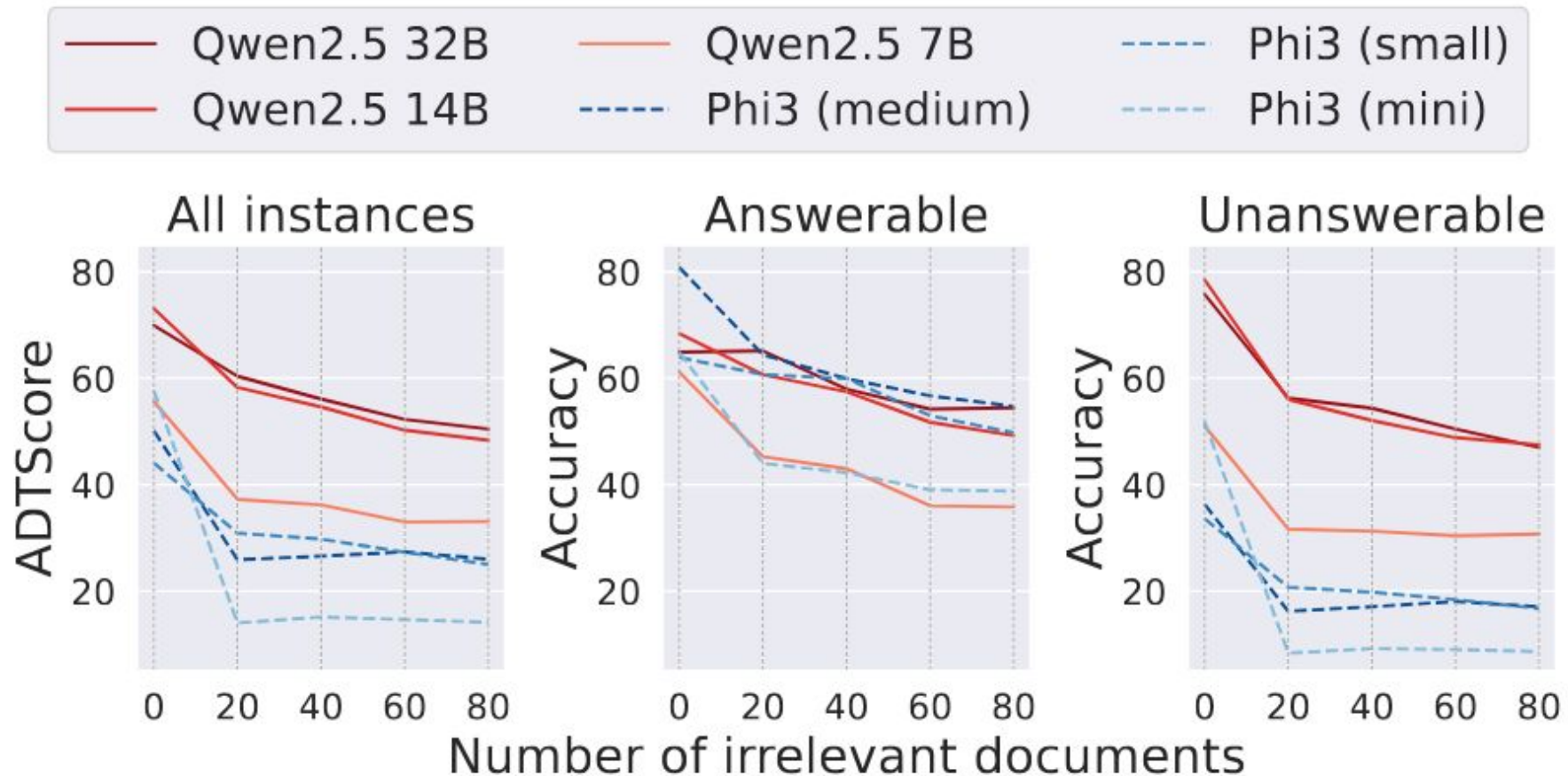
Model	<i>Answerable</i>		<i>Unanswerable</i>			
	Multi H.	Time S.	Multi H.	Time S.	False P.	Uncertain S.
<i>Random</i>	14.3	14.3	14.3	14.3	14.3	14.3
Phi3 mini (3.8B)	79.8	21.0	3.1	27.4	0.6	1.2
Phi3 small (7B)	80.5	34.8	10.2	39.5	8.2	4.5
Phi3 medium (14B)	79.7	53.8	16.6	17.1	8.3	5.5
Phi3.5 MoE (16×3.8B)	82.9	44.7	16.5	54.3	13.4	6.7
Qwen2.5 (7B)	67.5	42.9	26.7	23.1	16.2	21.8
Qwen2.5 (14B)	76.3	66.0	44.9	41.0	42.8	32.7
Qwen2.5 (32B)	79.4	67.3	41.7	62.1	38.6	26.7

sufficient evidence

insufficient evidence

unanswerable questions

Performance decreases when irrelevant documents are added

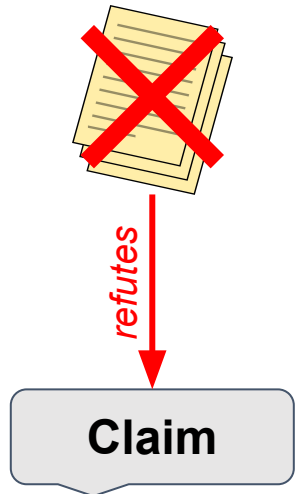


NeoQA: Evidence-based Question Answering with Generated News Events (**Findings of ACL 2025**)



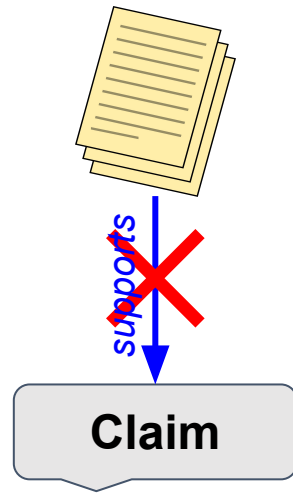
We looked into many different evidence relations

Evidence



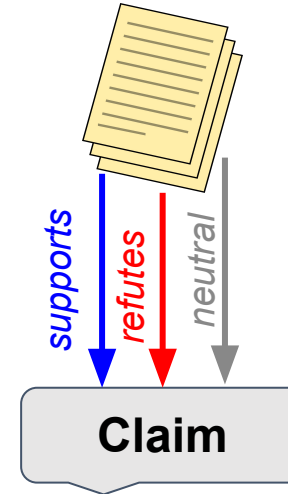
(Glockner et al., 2022)

Evidence



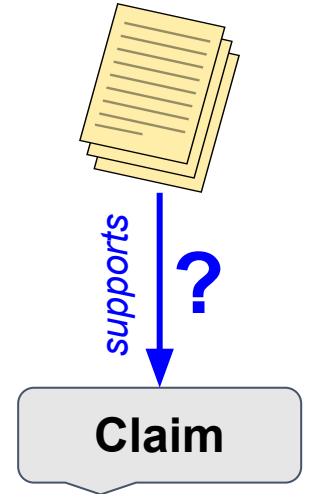
(Glockner et al., 2024a, 2025a)

Evidence



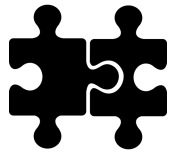
(Glockner et al., 2024b)

Evidence



(Glockner et al., 2025b)

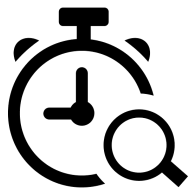
Final remarks



Focus on the problem and **work backward**.



Combine **powerful LLMs** with **realistic benchmarks** to tackle real-world problems.



Evidence-based evaluation with **real-world problems** but **no data contamination** remains challenging.



THANKS!

*Please feel free to reach out to me if you are interested in
this work (or automated fact-checking)! :-)*

Email: *maxg216@gmail.com*

References

- Warren, G., Shklovski, I., & Augenstein, I. (2025). [Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking](#). arXiv preprint arXiv:2502.09083.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). [FEVER: a large-scale dataset for fact extraction and VERification](#). In NAACL 2018.
- Ma, H., Xu, W., Wei, Y., Chen, L., Wang, L., Liu, Q., & Wu, S. (2023). [Ex-fever: A dataset for multi-hop explainable fact verification](#). In Findings of ACL 2024.
- Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., & Bansal, M. (2020). [HoVer: A dataset for many-hop fact extraction and claim verification](#). In Findings of EMNLP 2020.
- Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). [MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims](#). In EMNLP 2019.
- Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. (2019). [A richly annotated corpus for different tasks in automated fact-checking](#). In CoNLL 2019.

References

Khan, K., Wang, R., & Poupart, P. [WatClaimCheck: A new Dataset for Claim Entailment and Inference.](#) In ACL 2022.

Shane, T., & Noel, P. (2020). [Data deficits: why we need to monitor the demand and supply of information in real time.](#) First Draft News.

Buttry, S. (2014). Verification fundamentals: Rules to live by. The Verification Handbook: Ultimate Guideline on Digital Age Sourcing for Emergency Coverage. Maastricht: The European Journalism Centre, 14-17.

Glockner, M., Hou, Y., & Gurevych, I. (2022). [Missing counter-evidence renders NLP fact-checking unrealistic for misinformation.](#) In EMNLP 2022.

Glockner, M., Hou, Y., Nakov, P., & Gurevych, I. (2024). [Missci: Reconstructing fallacies in misrepresented science.](#) In ACL 2024.

Glockner, M., Hou, Y., Nakov, P., & Gurevych, I. (2024). [Grounding fallacies misrepresenting scientific publications in evidence.](#) In NAACL 2025.

References

Glockner, M., Staliūnaitė, I., Thorne, J., Vallejo, G., Vlachos, A., & Gurevych, I. (2024). [Ambifc: Fact-checking ambiguous claims with evidence](#). Transactions of the Association for Computational Linguistics, 12, 1-18.

Glockner, M., Jiang, X., Ribeiro, L. F., Gurevych, I., & Dreyer, M. (2025). [NeoQA: Evidence-based Question Answering with Generated News Events](#). In Findings of ACL 2025 (*to appear*).

Clark, C., Lee, K., Chang, M. W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). [Boolq: Exploring the surprising difficulty of natural yes/no questions](#). In NAACL 2019.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., ... & Inui, K. (2023). [Realtime qa: What's the answer right now?](#). In NeurIPS 2023.

Tang, Y., & Yang, Y. (2024). [Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries](#). In COLM 2024.