# Towards **Real-World** Fact-Checking with Large Language Models

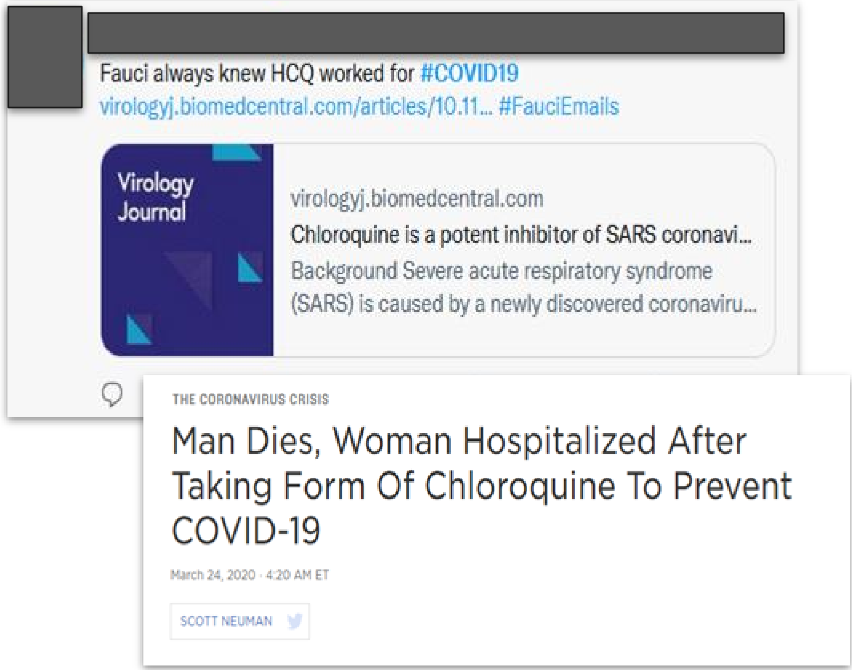
Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt, Germany

TECHNISCHE
UNIVERSITÄT
DARMSTADT



UKP

# Disinformation poses a growing threat to our society



[redacted]

Fauci always knew HCQ worked for #COVID19
virologyj.biomedcentral.com/articles/10.11... #FauciEmails

Virology Journal

virologyj.biomedcentral.com
Chloroquine is a potent inhibitor of SARS coronavi...
Background Severe acute respiratory syndrome (SARS) is caused by a newly discovered coronaviru...

THE CORONAVIRUS CRISIS

Man Dies, Woman Hospitalized After Taking Form Of Chloroquine To Prevent COVID-19

March 24, 2020 · 4:20 AM ET

SCOTT NEUMAN


Council of the EU | Press release | 28 July 2023 16:30

**Information manipulation in Russia's war of aggression against Ukraine: EU lists seven individuals and five entities**

# Defining real-world misinformation

*Non-True, harmful claims that professional fact-checkers deem important to verify*

**Volume**

**Relevance**

# Your Turn: *What is real-world misinformation?*

Coldplay is French.

“Beat It” is not a song by Michael Jackson.

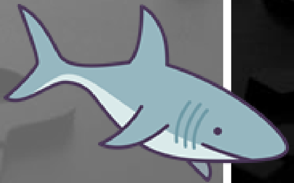
Half a million sharks could be killed to make the COVID-19 vaccine.

# Your Turn: *What is real-world misinformation?*

Coldplay is French.

“Beat It” is not a song by Michael Jackson.

Half a million sharks could be killed to make the COVID-19 vaccine.



# Your Turn: *Why is harmful real-world misinformation challenging?*

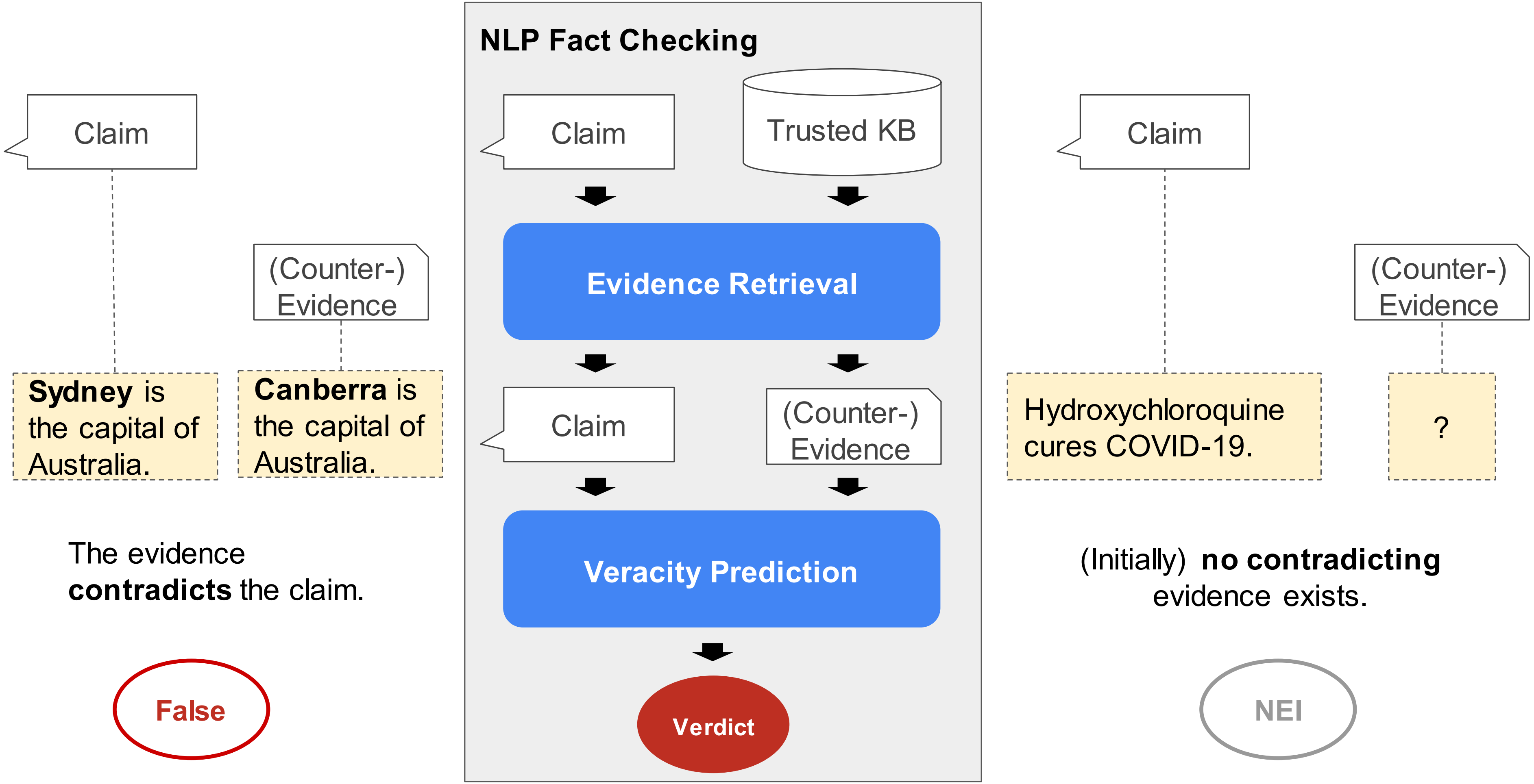
FEVER

Coldplay is French.

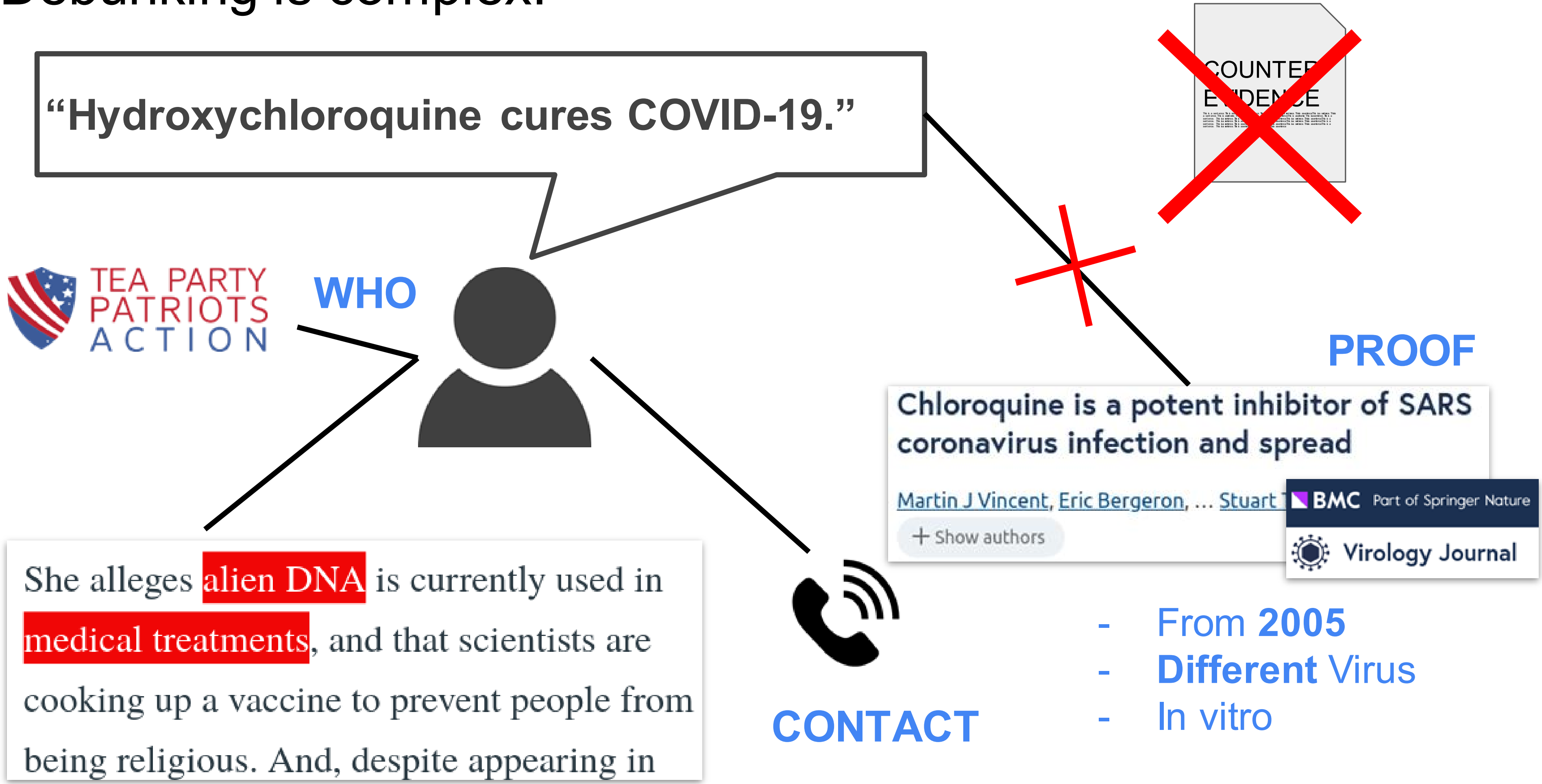
“Beat It” is not a song by Michael Jackson.

Half a million sharks could be killed to make the COVID-19 vaccine.

# NLP Fact-Checking relies on (Counter)-Evidence



# Debunking is complex!



# Humans use the source of a claim in verification

Fact-Checking organizations consider **identifying the source of the claim** as key to fact checking (Arnold, 2020).

**Human Verification looks at the source!**


**Sources provided by the claimant** must be included and assessed in the verification process (Silverman, 2014).

The question "***How do you know that?***" is at the heart of verification. (Buttry, 2014)

# Context is needed to find and use evidence based on the source

Is technically **irrelevant** to the claim.

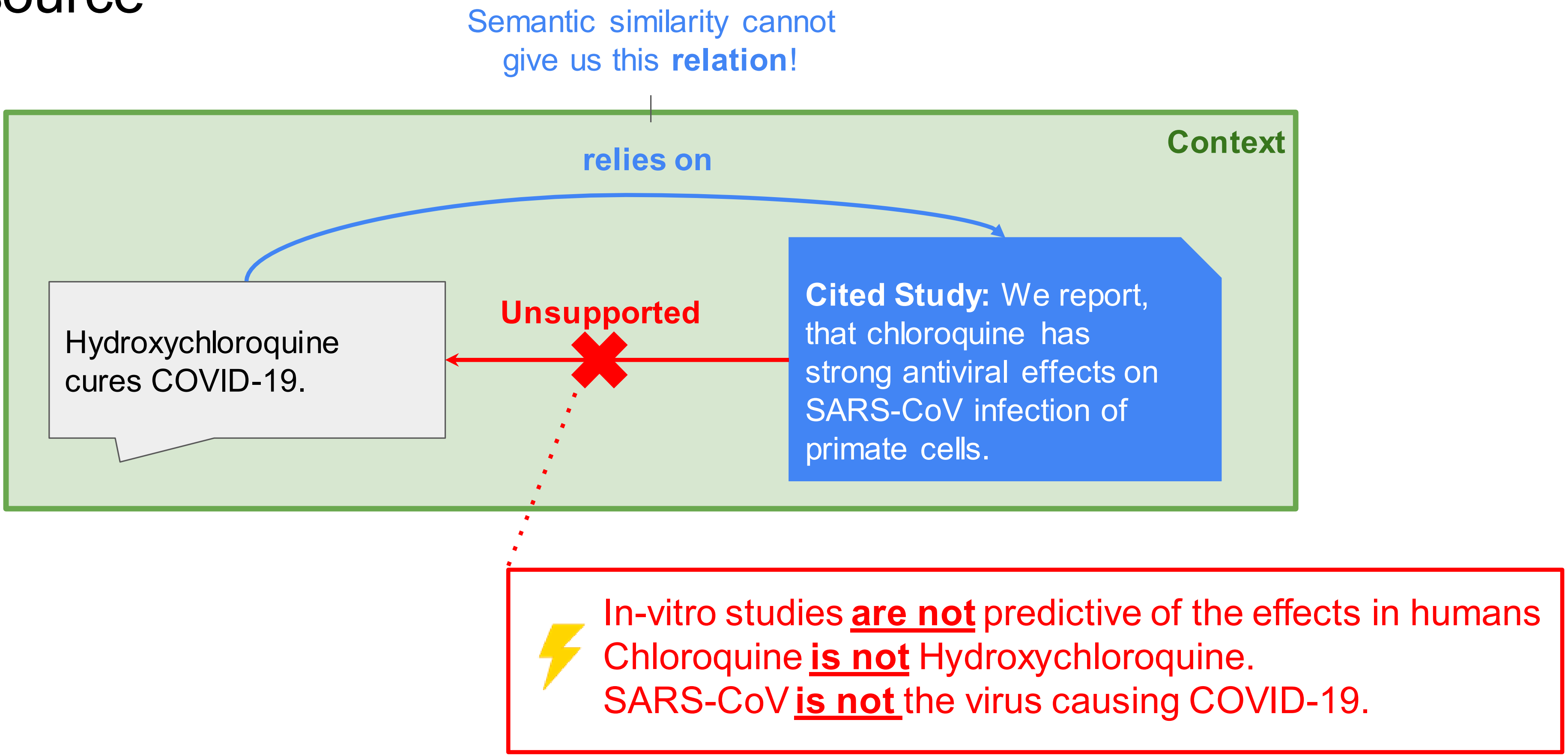
Hydroxychloroquine cures COVID-19.

**neutral**

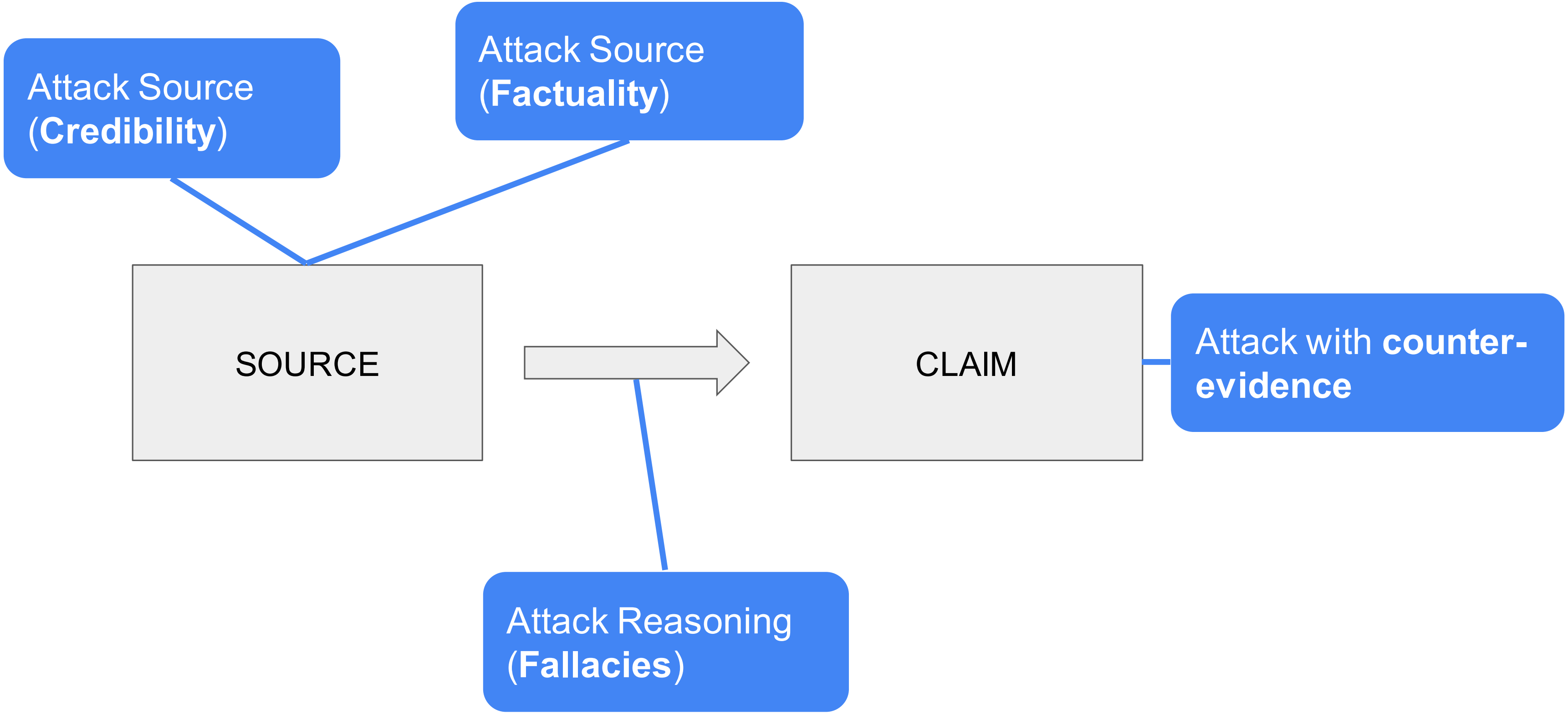
**Cited Study:** We report, that chloroquine has strong antiviral effects on SARS-CoV infection of primate cells.

In-vitro studies **are not** predictive of the effects in humans
Chloroquine **is not** Hydroxychloroquine.
SARS-CoV **is not** the virus causing COVID-19.

# Context is needed to find and use evidence based on the source

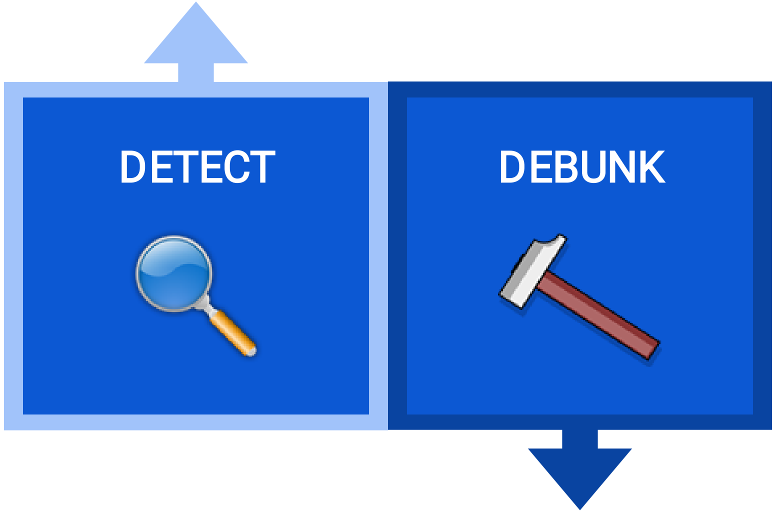


# Adding the source enables multiple ways to attack claims



# Combating real-world misinformation with NLP

Quickly detect misinformation to intervene and reduce its spread.




Automatically assess the veracity of the claim and debunk it.

# Identifying Harmful Content in Advance

Efficient Few-shot Learning Without Prompts (SetFit)
Lewis Tunstall, Nils Reimers, Enso Jo, Luke Bates
Daniel Korat, Moshe Wasserblat, and Oren Pereg

# Detect > **Debunk**

Can we now automatically debunk
**real-world *misinformation*?**

Code and Data

Read the paper

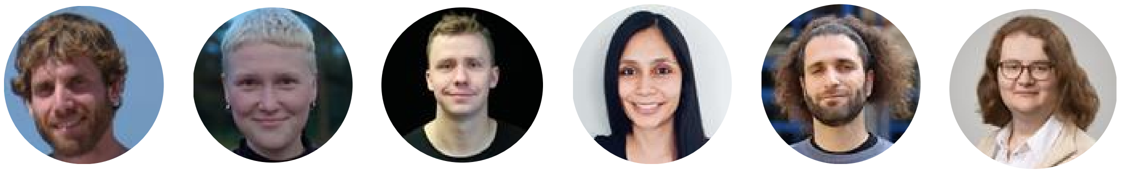
[Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation](#).
Max Glockner, Yufang Hou, and Iryna Gurevych. In *EMNLP 2022*.

Look at **simpler and less harmful claims** where we can expect to find evidence.

**AmbiFC: Fact-Checking Ambiguous Claims with Evidence.**

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, Iryna Gurevych.

*TACL 2024*.



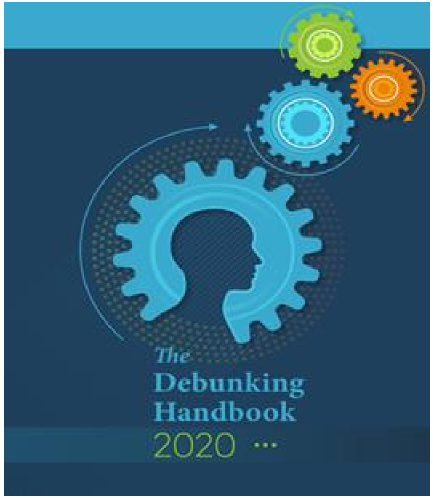
Read the paper

Code and Data

# Current research only partially considers real-life fact-checking

The goal is to **stop people from believing** false claims

HOW?




Lewandowsky et al., (2020)

Why is the claim false?

AFC focuses on this

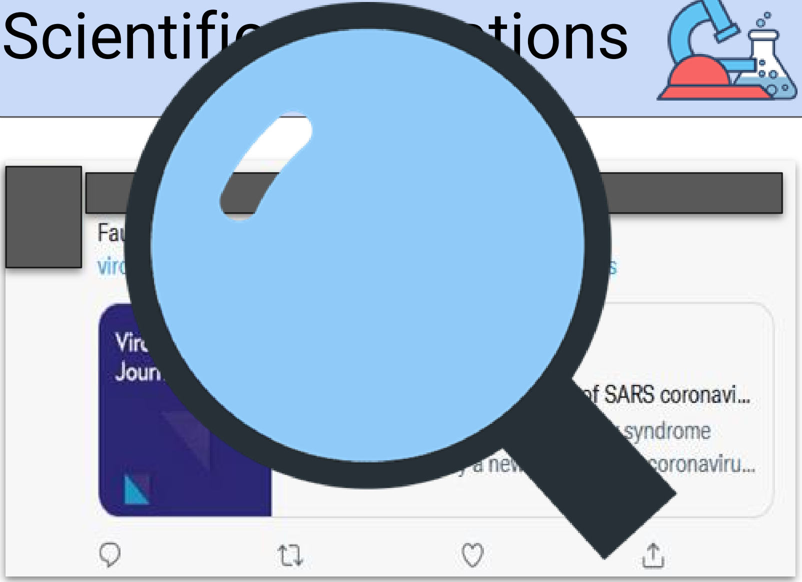
Why was it believed to be true?

Why is the alternative correct?

The very important but less studied part

# Certain content types increase perceived credibility

Scientifi tions



Visual content


Thus, they may be **misleading**. We need **special debunking methods**!

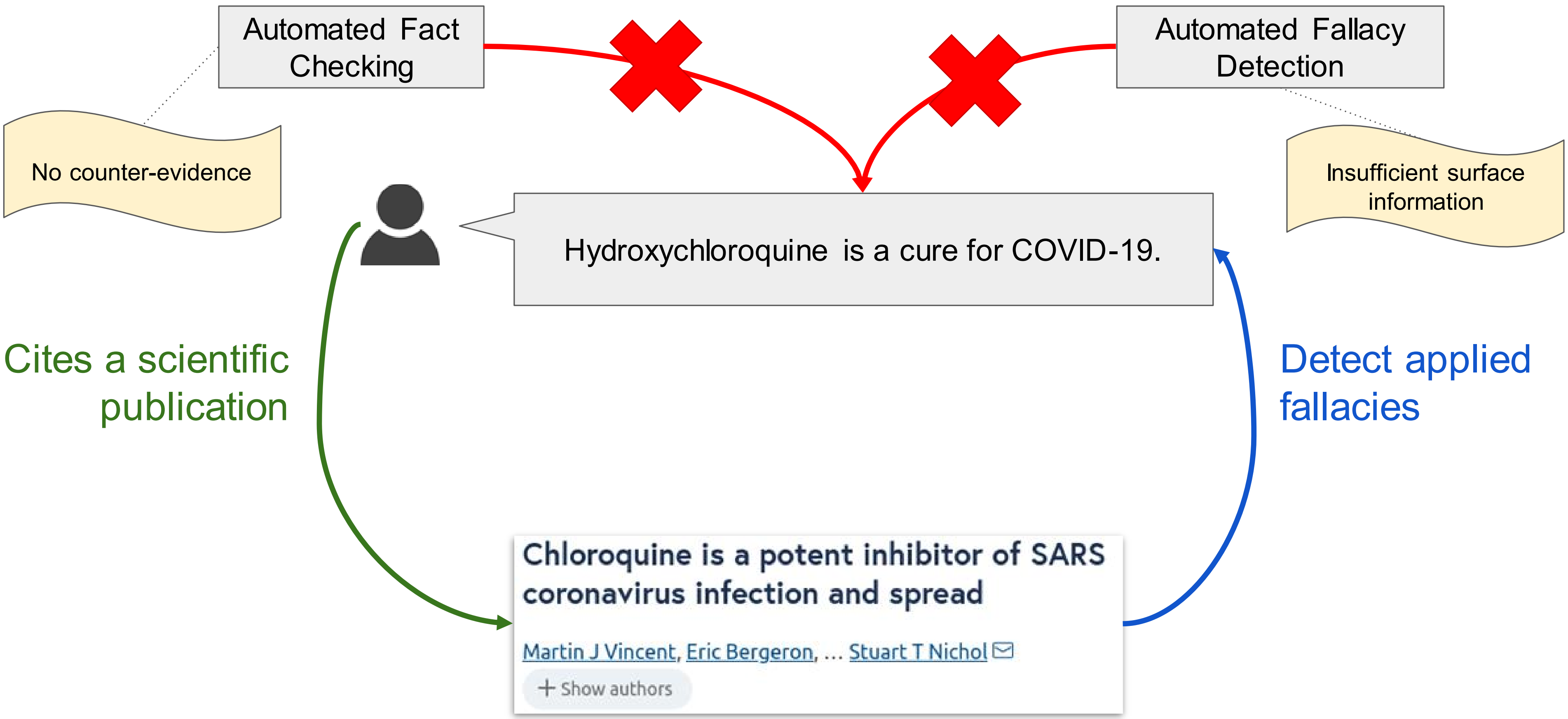
# Dismantling the Misleading Narratives: Reconstructing the Fallacies in Misrepresented Science

*Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych*

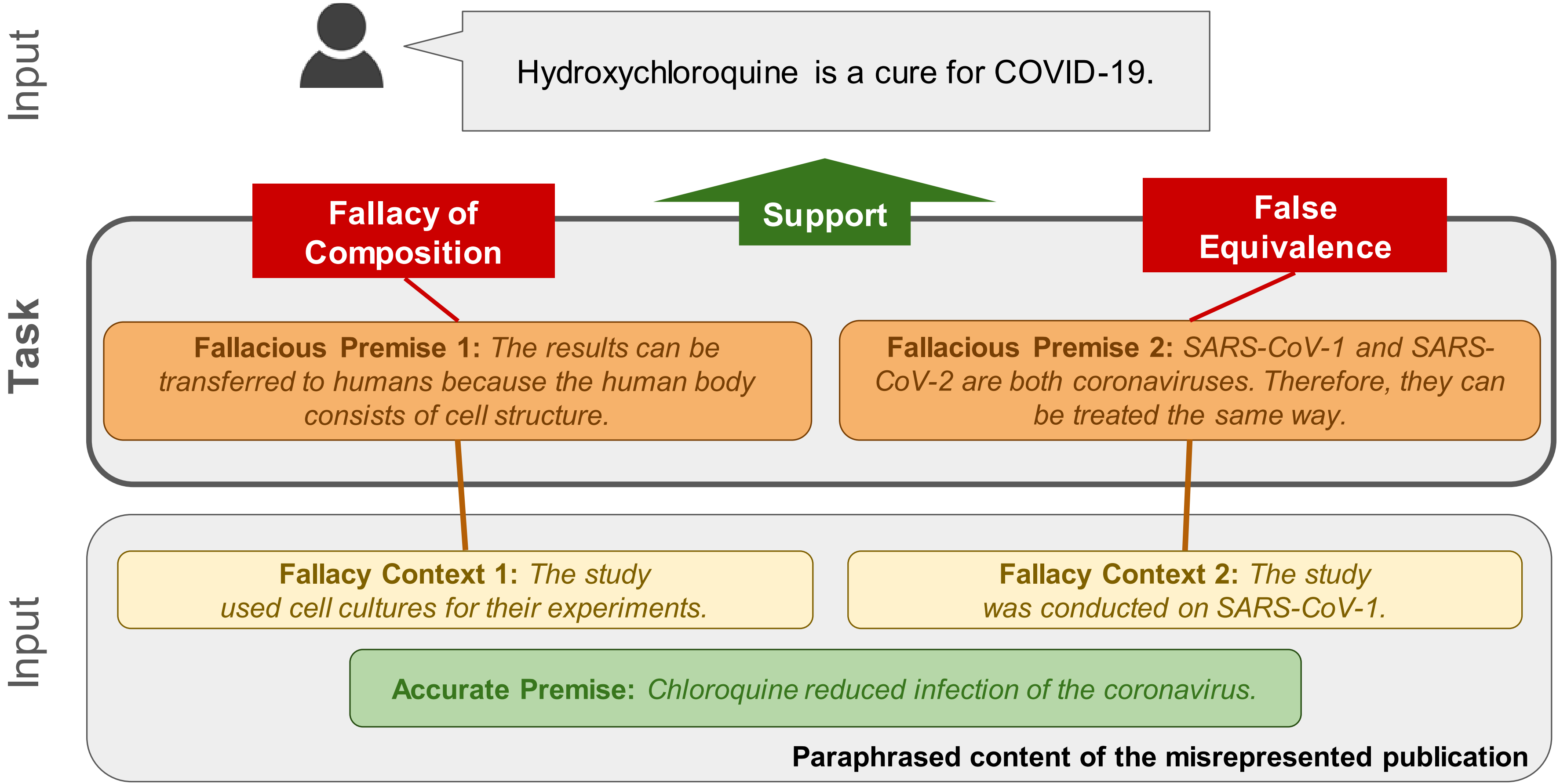



Accepted at ACL 2024

# We need to assess a claim based on its sources



# We propose to reconstruct the fallacious arguments



# We create Missci based on fact-checking articles

**Collect** → **Select** → **Reconstruct**

## 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

# 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.*

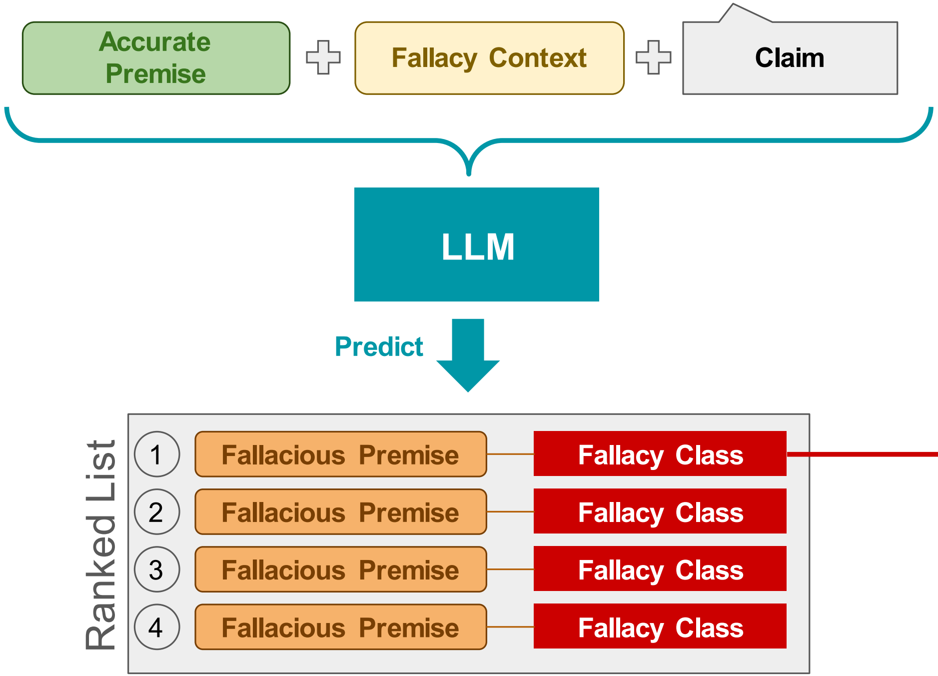
**Example**:
*Hydrogen is not wet. Oxygen is not wet. Therefore, water (H2O) is not wet.*

| 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.*




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?*

| LLM | Correct (%) |
|---|---|
| Llama2 (L) | 0.040 |
| Llama2 (D) | 0.107 |
| GPT-4 (L) | **0.503** |
| GPT-4 (D) | 0.481 |

Was **0.292** (P@1) over **gold fallacious premise**

**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

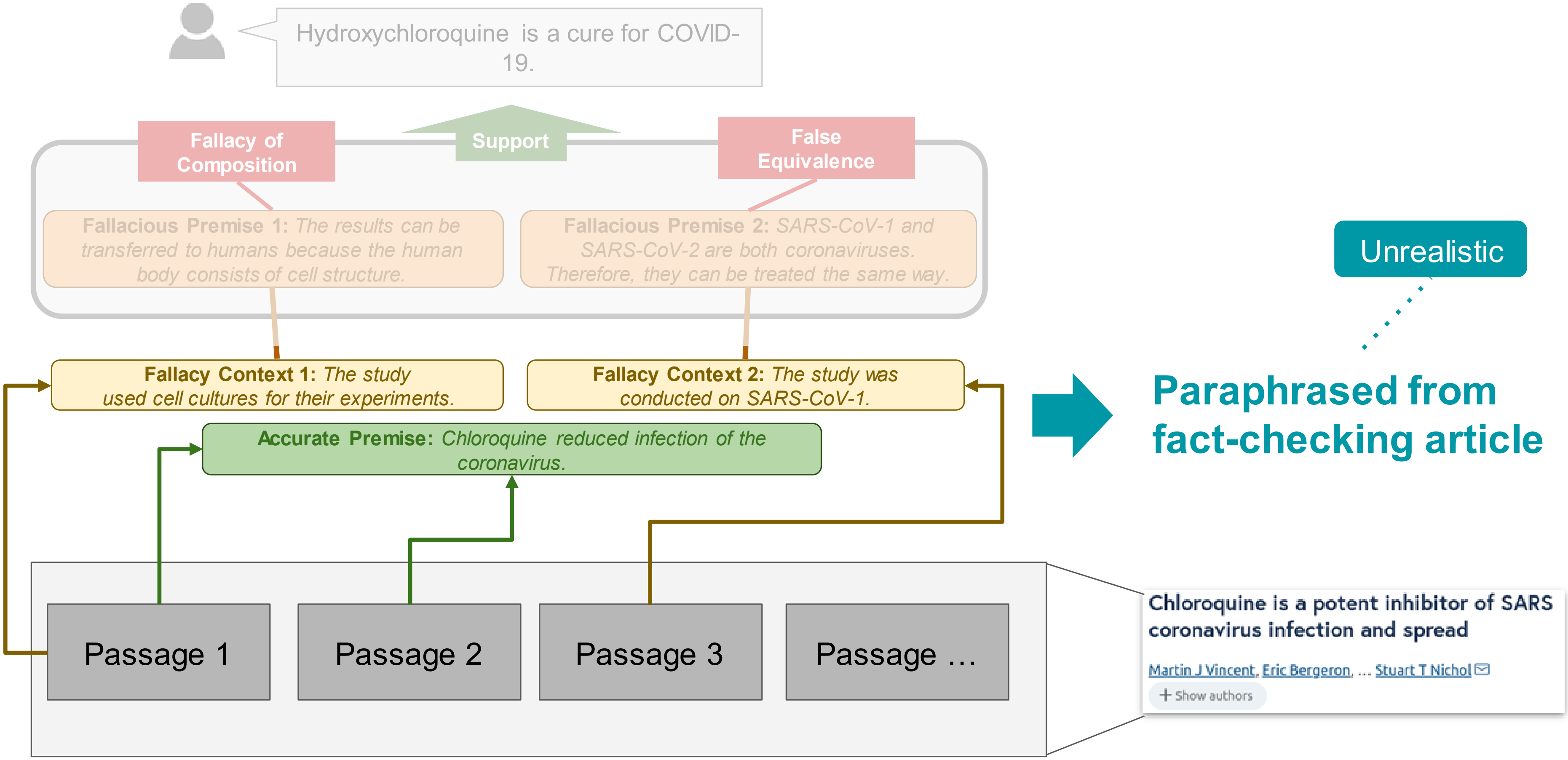
# Conclusion

- 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!

# Towards Grounding Fallacies of Real-World Misinformation in Misrepresented Scientific Publications (*ongoing*)

*Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych*

# Missci does not consider real-world passages

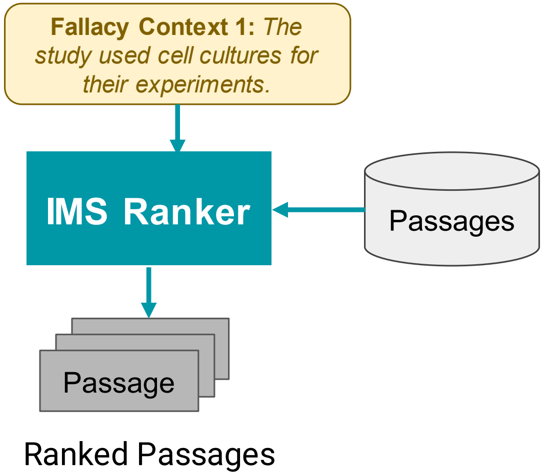


# We link publication context to real-world passages

**Preselect** → **Annotate** → **Aggregate**

## Preselect

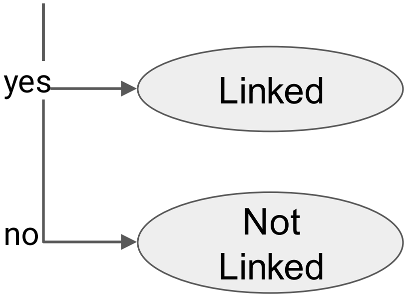
Use IMS (Wright et al., 2022) to preselect relevant passages.




## 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%* |

**Multi-modal** reasoning
**Multi-hop** reasoning

**Different Scope**

*Cloth masks do nothing to prevent the virus.*

**Scientific publication**

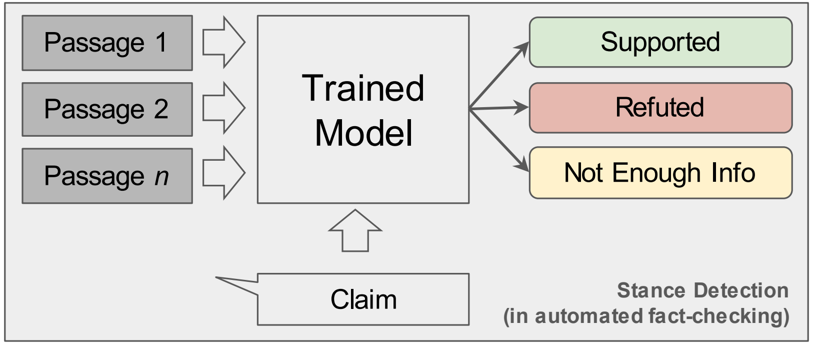
*Health workers wearing cloth masks were infected.*

*Compare effectiveness of cloth masks and medical masks.*

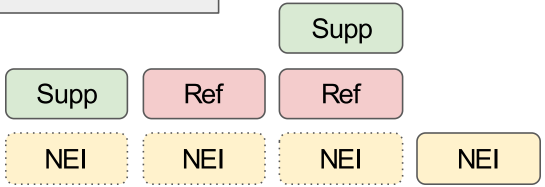
**Not explicitly communicated in the study!**

**Publication 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) | | | | |
| HealthVer (Sarrouti et al., 2021) | | | | |
| CovidFact (Saakyan et al., 2021) | | | | |
| *All AFC* | | | | |

This is bad!


**Stance detection does not identify fallacy passages.**

# Evidence biases the LLM to believe the claim is true

Missci

CovidFact

HealthVer

| | | Misinformation | | | 100 True claims | | |
|---|---|---|---|---|---|---|---|
| **Knowledge** | **LLM** | **True** | **False** | **NEI** | **True** | **False** | **NEI** |
| **Parametric Knowledge** | Llama 2 | 1.6 | 61.1 | 37.3 | 34.7 | 22.3 | 41.3 |
| | GPT 4 | 0.0 | 82.1 | 17.9 | 60.0 | 23.0 | 17.0 |
| **RAG Style** | Llama 2 | **23.8** | 61.5 | 12.7 | 58.7 | 29.7 | 10.7 |
| | GPT 4 | **27.4** | 38.1 | 34.5 | 56.0 | 4.0 | 40.0 |

**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.*

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

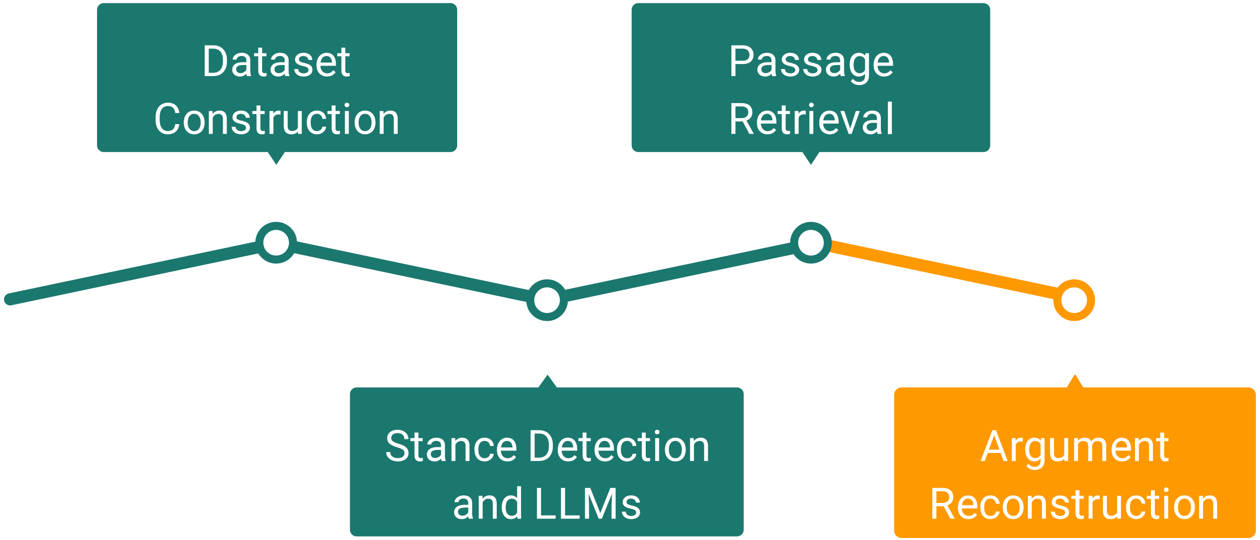
**Fallacy Context 2:** *The study was conducted on SARS-CoV-1.*

**What is the claim based upon?** (P@1)

**What content points to fallacies?** (mAP)

| Type | Model | P@1 | mAP |
|---|---|---|---|
| Term frequency | BM25 | 0.547 | 0.496 |
| Sentence Transformer | SBERT (Reimers et al. 2019) | 0.400 | 0.520 |
| | PubMedBERT ST (Deka et al. 2022) | 0.440 | 0.489 |
| | BioBERT ST (Deka et al. 2022) | 0.547 | 0.491 |
| | SapBERT ST (Deka et al. 2022) | 0.480 | 0.504 |
| | INSTRUCTOR (Su et al. 2022) | 0.573 | **0.541** |
| | SPICED (IMS) (Wright et al. 2022) | 0.587 | 0.524 |
| Scientific Fact-Checking (DeBERTaV3) | SciFact (Wadden et al. 2020) | 0.603 | 0.360 |
| | CovidFact (Saakyan et al. 2020) | 0.517 | 0.380 |
| | HealthVer (Sarrouti et al. 2021) | **0.608** | 0.368 |
| | *All Scientific Fact-Checking* | **0.608** | 0.306 |

# Next step will be argument reconstruction



# Conclusion

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

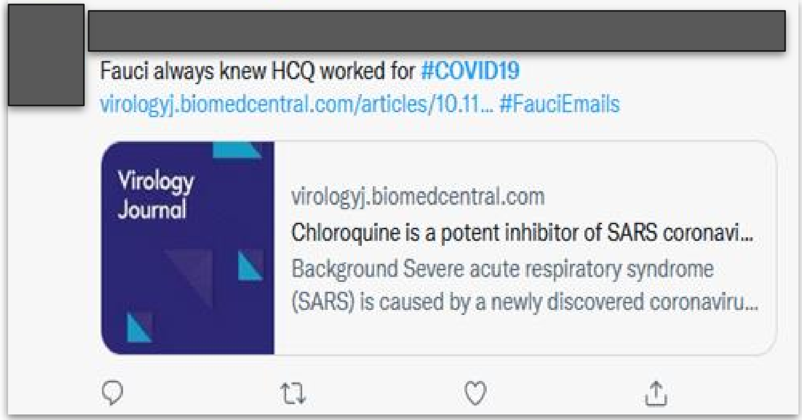
Novel benchmark to reconstruct fallacious arguments with **realistic evidence**.

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

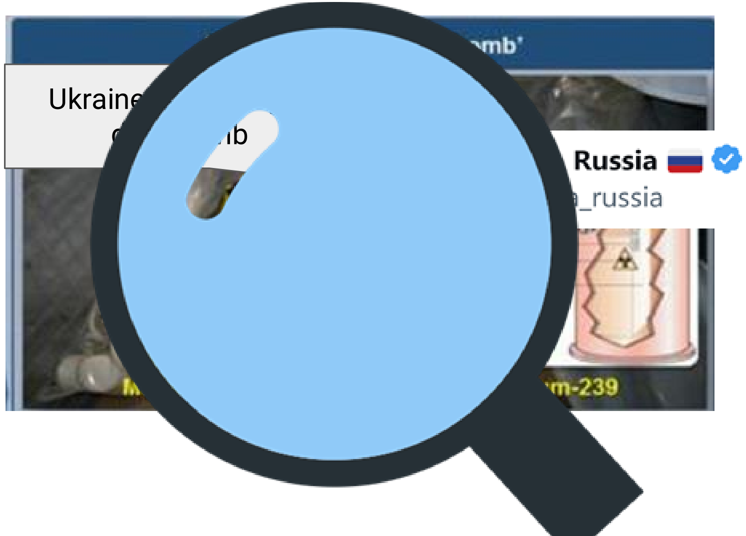
Still ongoing work!

# Certain content types increase perceived credibility

## Scientific publications




## Visual content




Thus, they may be **misleading**. We need **special debunking methods**!

# Can we use Multimodal Large Language Models (MLLMs) to combat multimodal misinformation?

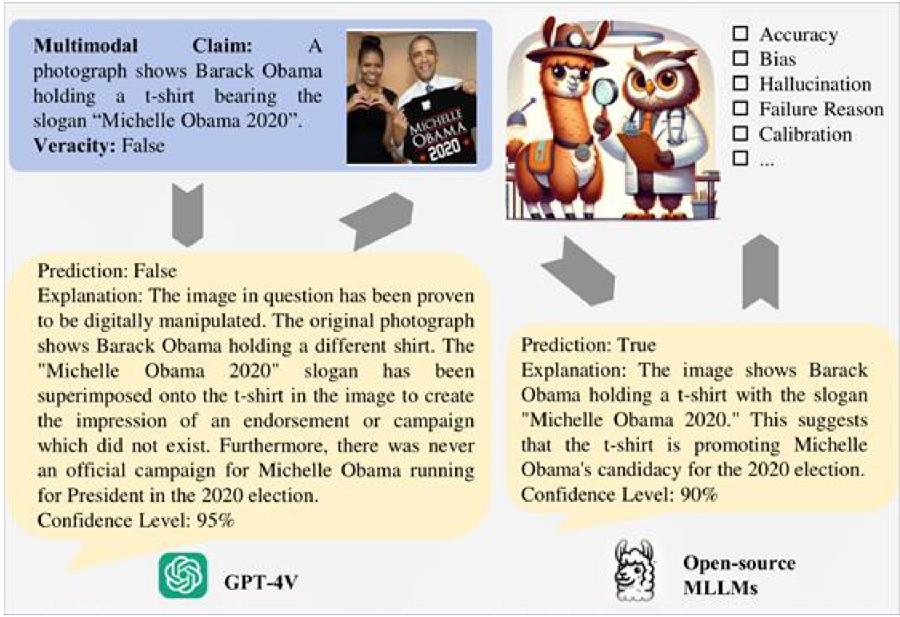
*Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov and Iryna Gurevych*




*This will become a QR code once we published preprint and repository*

# Can MLLMs fact-check real-world images?

- Misleading images are a real-world threat
- (M)LLMs store extensive information
- Outdated information & hallucinations



# We propose an evaluation framework consisting of datasets, prompts, and metrics

Sample data from existing datasets. Post-4V is new data from Snopes.

| | Fauxtography | COSMOS | MOCHEG | Post-4V |
|---|---|---|---|---|
| True | 200 | 240 | 267 | 81 |
| False | 200 | 240 | 237 | 105 |
| Total | 400 | 480 | 504 | 186 |

Table 1: Statistics of datasets in our evaluation.

## Metrics

- Count of True / False / Uncertain / Others:
  - Uncertain: Model expresses uncertainty
  - Others: Model doesn't follow the prompt to give the prediction, explanation and confidence
- True&False Accuracy, Overall Accuracy
  - True&False: Statistics on samples predicted as True or False.
  - Overall: As long as the response includes a prediction and the prediction is correct.

## Prompt:

```
Is it true that "CLAIM"? True or False? Use
the following format to provide your answer:
Prediction: [True or False]
Explanation:   [put your evidence and
step-by-step reasoning here]
Confidence Level:   [please show the
percentage]

Note: The confidence level indicates the
degree of certainty you have about your
answer and is represented as a percentage.
For instance, if your confidence level is
80%, it means you are 80% certain that
your answer is correct and there is a 20%
chance that it may be incorrect.
```

# Three main research questions

**RQ1 Performance Evaluation:**
Accuracy and quality of explanation/reasoning

**RQ2 Enhanced Approaches:**
How to improve the performance?

**RQ3 Taxonomy of Failure Reasons:**
Typical errors when employing MLLMs as fact-checkers

# We evaluate models and two enhanced approaches

**Models:**

GPT-4V, LLaVA-1.5 (7b and 13b), MiniGPT-v2 (7b)

**Enhanced approaches (based on LLaVA(13b))**

- **Prompt Ensembles (PE)**
  - Ask ChatGPT to generate 5 more prompts for fact-checking
  - Voting based on all 6 responses, if top 2 response types have same counts => uncertain
- **In-Context Learning (ICL)**
  - ICL-1, ICL-2
  - Four demonstration sets, then calculate the average.
  - We use the correct GPT-4V’s responses, since fact-checking articles are lengthy and contain irrelevant content.

# GPT-4 performs impressively

| | Fauxtography | | | | | | COSMOS | | | | | | MOCHEG | | | | | | Post-4V | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | F | U | O | T&F | All | T | F | U | O | T&F | All | T | F | U | O | T&F | All | T | F | U | O | T&F | All |
| GPT-4V | 158 | 195 | 29 | 18 | **81.9** | **82.1** | 179 | 204 | 83 | 14 | **86.2** | **80.0** | 216 | 223 | 37 | 28 | **87.2** | **83.5** | 54 | 98 | 26 | 8 | **79.6** | **73.7** |
| MiniGPT | - | - | - | - | - | 55.5 | - | - | - | - | - | 62.1 | - | - | - | - | - | 54.3 | - | - | - | - | - | 55.9 |
| LLaVA(7b) | 337 | 6 | 1 | 56 | 53.4 | 46.3 | 449 | 7 | 4 | 20 | 52.0 | 50.0 | 409 | 6 | 0 | 89 | 54.2 | 44.8 | 157 | 1 | 3 | 25 | 42.0 | 37.1 |
| LLaVA(13b) | 286 | 98 | 3 | 13 | 54.4 | 52.3 | 331 | 134 | 13 | 2 | 69.7 | 67.7 | 363 | 112 | 28 | 1 | 61.5 | 58.1 | 116 | 65 | 2 | 3 | 56.3 | 55.9 |
| LLaVA+PE | 244 | 153 | 2 | 1 | 57.1 | 54.7 | 275 | 204 | 0 | 1 | 76.3 | 71.7 | 290 | 214 | 0 | 0 | 59.9 | 58.1 | 85 | 101 | 0 | 0 | 56.9 | 56.1 |
| LLaVA+ICL-1 | 228 | 159 | 6 | 8 | **61.8** | **62.3** | 293 | 175 | 9 | 3 | 74.8 | 74.1 | 254 | 240 | 7 | 4 | **62.4** | **62.5** | 79 | 101 | 5 | 1 | 58.7 | 57.5 |
| LLaVA+ICL-2 | 186 | 188 | 8 | 18 | 61.6 | 61.7 | 247 | 215 | 12 | 7 | **77.3** | **76.5** | 195 | 286 | 10 | 13 | 60.2 | 60.4 | 48 | 122 | 8 | 8 | **62.1** | **61.4** |
| Majority | - | - | - | - | - | 50.0 | - | - | - | | - | 50.0 | - | - | - | - | - | 53.0 | - | - | - | - | - | 56.5 |

Table 2: Performance of various models and approaches. *T: True, F: False, U: Uncertain, O: Others, T&F: True&False Accuracy, All: Overall Accuracy, PE: Prompt Ensembles, ICL: In-Context Learning.* The majority-class accuracy is established in the last row.

- GPT-4V: surprising accuracy
- MiniGPT: cannot provide explanation and confidence
- LLaVA: ICL works better than PE, but still falls behind

# Further findings

- ICL increases the explanation length
- ICL brings more “checked cases” and “manipulated cases”
- GPT-4V’s verbalized confidence is well-calibrated, while LLaVA, even with ICL, is overconfident

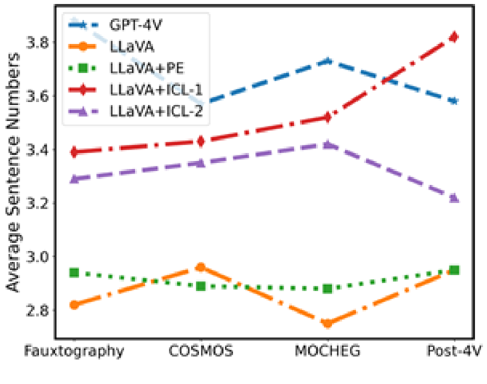
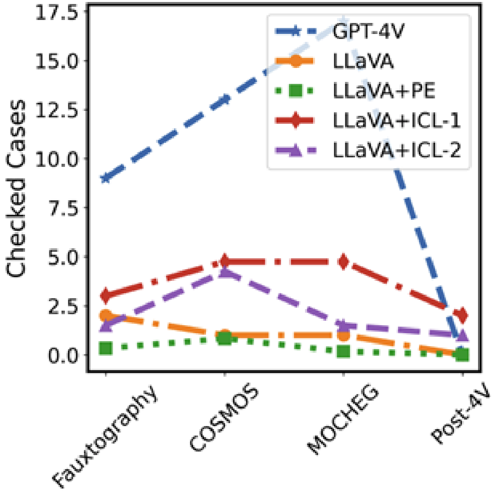
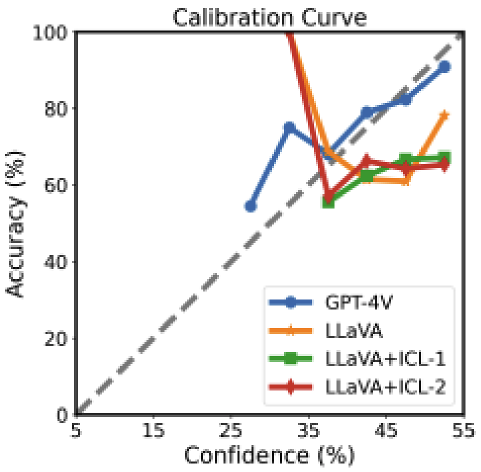



Figure 5: Average number of sentences in explanations across different models and settings. GPT-4V generates the longest explanations except on Post-4V. With one example, ICL-1 significantly increases the average explanation length.




Number of checked cases




Calibration Curve

# Six typical failure reasons in GPT-4V's responses

**Lack of Knowledge.** It's common knowledge that octopuses have eight tentacles, the rarity of an octopus with 96 tentacles may confound the model.

**Claim:** A photograph shows an octopus with 96 tentacles.
**Veracity:** True

Prediction: False
Explanation: Octopuses typically have eight tentacles, and it is extremely rare for one to have 96 tentacles. Although mutations or developmental abnormalities can occur, leading to more than eight tentacles, such a high count is implausible and unprecedented in scientific literature. ... The image provided could be manipulated or the octopus in the photo could be exhibiting a rare condition known as polycephaly, where it has an excess number of tentacles due to a mutation. It is more credible to believe that the octopus has more tentacles than usual, but not as high as 96.
Confidence Level: 90%

**Similar Images.** The image depict a scene from the movie "Tropic Thunder," but the model mistakes it for the "Platoon." Both movies share the Vietnam War as background.

**Claim**: A photograph shows a group of underappreciated Vietnam veterans.
**Veracity:** False

Prediction: False
Explanation: The photograph in question does not depict actual Vietnam veterans but rather actors fromthe1987 war film "Platoon." Key identifiers include the actors' recognizable faces, especially if one is familiar with the film. Additionally, the uniformity of their gear and the dramatized poses for the camera are more characteristic of a movie set than an authentic combat situation. These details suggest that the photo is staged and not an actual representation of Vietnam veterans in the field.
Confidence Level: 95%

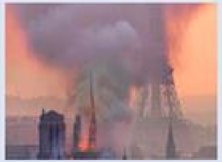
**Blurry or Incomplete Image.** The model cannot discern that the scene depicted is related to posture training based on the image's content..

**Claim:** A photograph shows a pin in the collar of a Chinese officer's uniform, a training method to correct posture.
**Veracity:** True

Prediction: False
Explanation: ... One individual is in focus, and it appears that another person is adjusting or aligning the decorations on the uniform, not inserting a pin in the collar as a posture training method. Such adjustments are typical in military dress protocols, especially during ceremonies where correct placement of medals and badges is important. The photograph does not provide any evidence that the pin is being used as a posture training method.
Confidence Level: 95%

**Distortion due to Perspective and Lens.** This example demonstrates that GPT-4V does not understand that a telephoto lens can make the foreground appear closer to the background than it actually is.

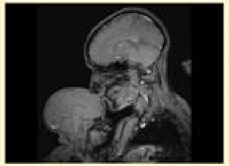
**Claim:** A photograph of a fire at Notre Dame Cathedral in Paris shows the Eiffel Tower directly in the background.
**Veracity:** True



Prediction: False
Explanation: The Eiffel Tower and Notre Dame Cathedral are both iconic landmarks in Paris, but they are not located immediately next to each other. ..., the Eiffel Tower is seen directly in the background of the burning Notre Dame, suggesting that they are much closer than they actually are in reality. ..., they would not appear as they do in this image with the Eiffel Tower seemingly directly behind Notre Dame. This suggests that the image may have been manipulated or taken from an angle that creates a misleading perspective.
Confidence Level: 95%

**Incomplete Verification.** The model's prediction overlooks the fact that an MRI cannot capture chemical reactions.

**Claim:** An MR image captures a mother's kiss producing a 'chemical reaction that released a burst of oxytocin' in her baby's brain.
**Veracity**: True



Prediction: True
Explanation: ... It is known that physical affection, such as a mother's kiss, can induce the release of oxytocin, commonly known as the "love hormone" or "cuddle hormone," which plays a role in social bonding, sexual reproduction, childbirth, and the period after childbirth. ... While an MRI itself doesn't directly capture chemical reactions or the release of hormones like oxytocin, it can show brain activity that might correlate with the effects of oxytocin release. Functional MRI (fMRI), in particular, can detect changes in blood ...
Confidence Level: 75%

**Ambiguous Claims.** Playboy can refer to either the name of the publisher or the magazine.

**Claim:** Playboy published nude photographs of 10-year-old actress Brooke Shields.
**Veracity**: True


Prediction: False
Explanation: ... In 1975, a series of photographs were taken of Brooke Shields by photographer Garry Gross, some of which depicted her nude. These photographs were controversial, and they were not taken for Playboy. The images were intended for an art exhibition and were subsequently published in a publication called "Sugar 'n' Spice" by Playboy Press, but they did not appear in the Playboy magazine itself. The issue of these photographs led to a legal battle, and Shields later tried to suppress them, ...
Confidence Level: 90%

# Conclusion

First evaluation of MLLMs for real-world fact-checking, incl. accuracy, bias, failure reasons

GPT-4V is impressive and shows potential, providing evidence and explanations

In-context learning improves LLaVA, still a large room for improvement

Further results are in the paper!

*This will become a QR code once we published preprint and repository*

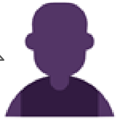
# “Image: Tell me your story!” Predicting the original meta-context of visual misinformation

*Jonathan Tonglet, Marie-Francine Moens and Iryna Gurevych*



# We need to identify the original context of images

This is an image of Coronavirus victims in China




Traditionally we focus on …

Misinformation detection → Claim is false

Instead, let's focus on …

Image contextualization → The image was taken by Reuters in Germany [...]

**Image contextualization is an important component of human fact-checking**

- To detect check-worthy images
- To detect out-of-context images
- To write convincing debunking articles
- To engage in pre-bunking communication

# We contextualize images with the 5 Pillars framework

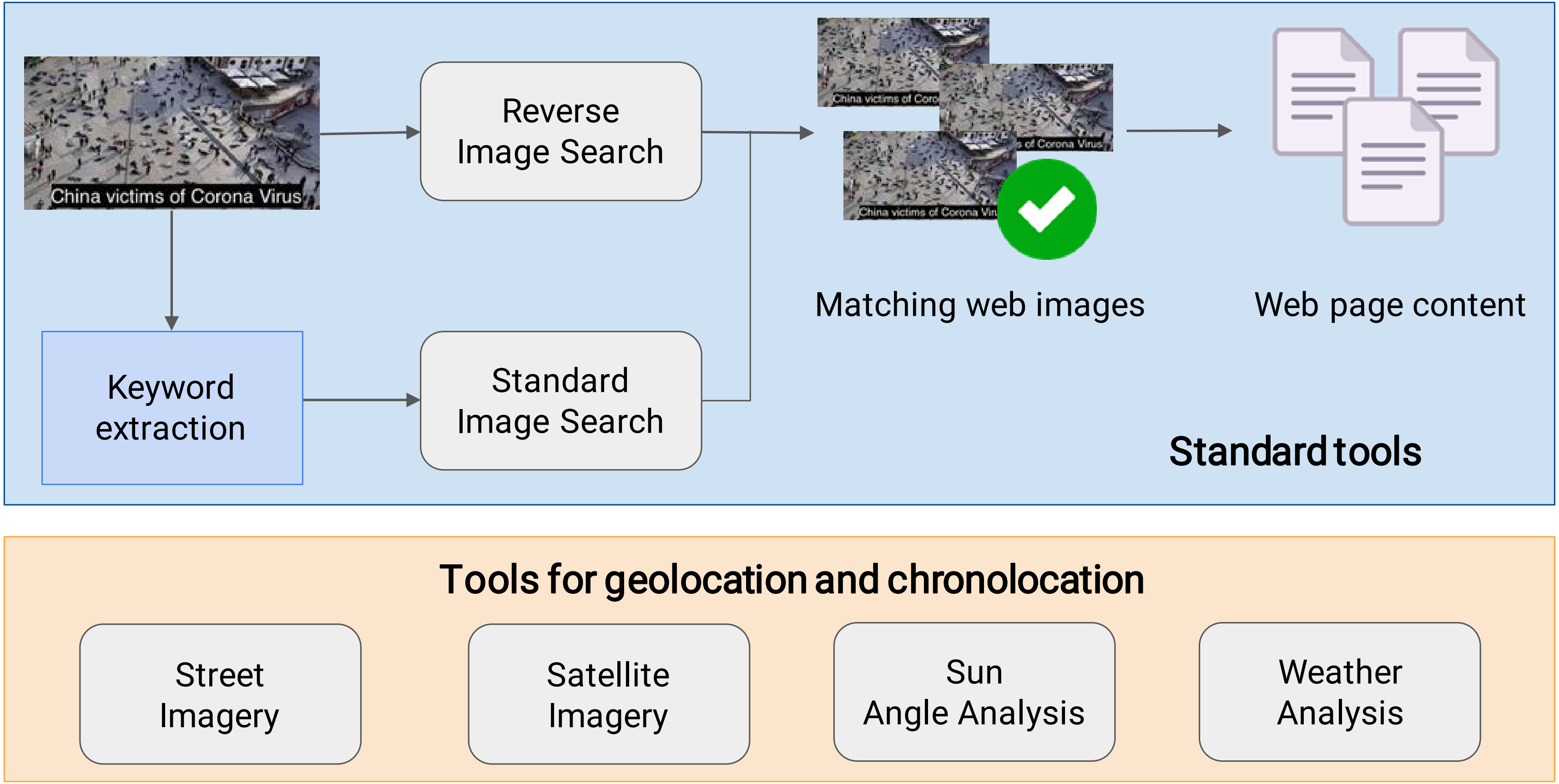
This is an image of Coronavirus victims in China

Let's find out the original context of this image!



The 5 Pillars framework was introduced by FirstDraft in Urbani (2019).

# Human fact-checkers use many tools



# We create the real-world 5Pils dataset

Collect | Annotate | Validate

## Collect

**Collect images** from fact-checking articles

- 3 organizations
  - Factly - India
  - Pesacheck - Kenya
  - 211Check - South Sudan
- **1676 images**
- 3 type of visual misinformation
  - Out-of-Context images
  - Manipulated images
  - Fake images

## Annotate

**Automate annotations** with GPT 4

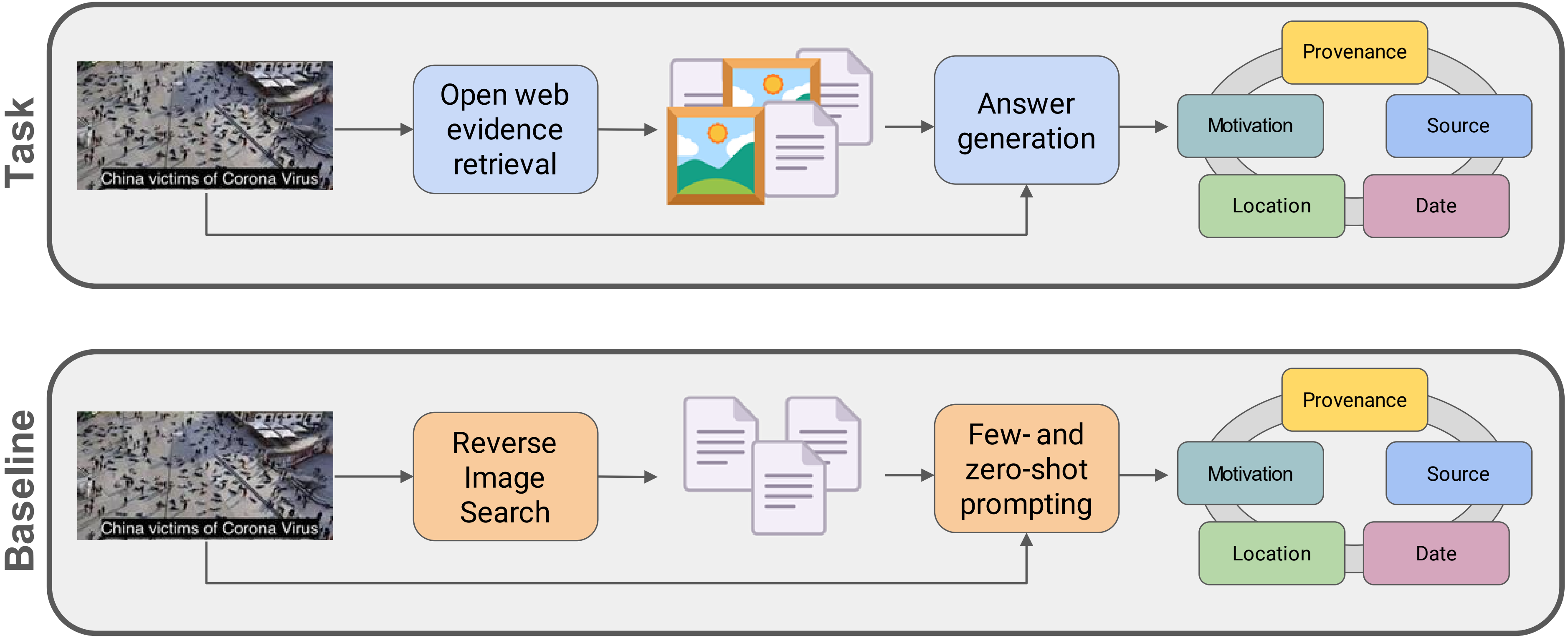
- Extracts the 5 Pillars answers and metadata from the articles
- Returns a structured JSON output

## Validate

**Validate annotations** with 3 human judges

- Random sample of 50 images
- 97.6% of the annotations are correct
  - 2.4% are missing
  - None are incorrect
- High inter-annotator agreements
  - 0.76 to 0.94

# We propose to automate image contextualization



# Baseline performs well but many challenges remain!

| | |
|---|---|
| Provenance | Only one evidence retriever |
| Source | Can't rely on interviews and profiling like fact-checkers do |
| Date | Can't link images to time periods and world events |
| Location | Links images to locations, especially countries |
| Motivation | Caption images accurately, fills details with web evidence |

Directions for future work

- Better retrievers
- Additional tools
- World knowledge

# Conclusion

Novel task: automated image contextualization

Novel dataset based on real-world fact-checking articles

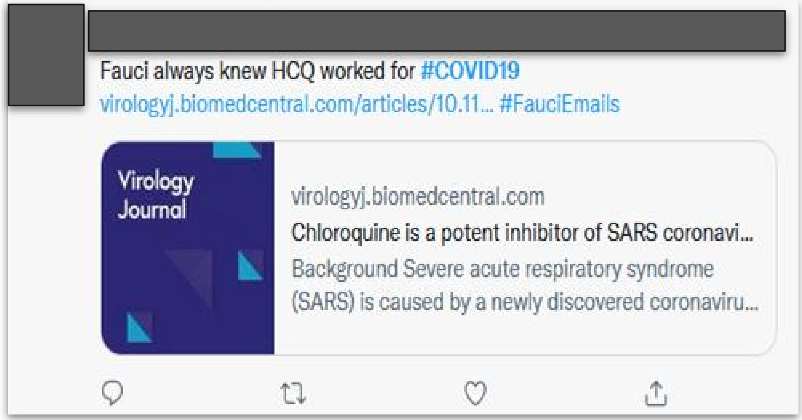
Better evidence retrievers on the open web is the main challenge for future work

More experiments, results and analysis in the paper!

# Certain content types increase perceived credibility

**Scientific publications**

**Visual content**





Thus, they may be **misleading**. We need **special debunking methods**!

# Concluding remarks

## Scientific publications

- LLMs have **limited critical reasoning abilities** when it comes to fallacious scientific arguments
- LLMs tend to consider **false claims as correct** when they are based on **misrepresented scientific publications**

## Visual content

- GPT4-Vision shows **promising performance** at detecting visual misinformation
- Open-source models are **lagging behind** in performance
- **Image contextualization** is an important but challenging task
- Many opportunities for research on **retrieval-augmented** and **tool-based LLMs**

# References


Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., Shaar, S., Firooz, H., & Nakov, P. (2022). A Survey on Multimodal Disinformation Detection. Proceedings of the 29th International Conference on Computational Linguistics, 6625–6643.

Khan, S. A., Furuly, J. G., Vold, H. B., Tahseen, R., & Dang-Nguyen, D. T. (2023). Online Multimedia Verification with Computational Tools and OSINT: Russia-Ukraine Conflict Case Studies (arXiv:2310.01978). arXiv

Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Kendeou, P., Newman, E., Pennycook, G., Porter, E., Rand, D., Rapp, D., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C., Sinatra, G., Swire-Thompson, B., Linden, S. van der, Wood, T., & Zaragoza, M. (2020). The Debunking Handbook 2020. Copyright, Fair Use, Scholarly Communication, Etc.

Silverman, C. (2014). Verification handbook: An ultimate guideline on digital age sourcing for emergency coverage. European Journalism Centre.

Silverman, C. (2020). Verification handbook for disinformation and media manipulation. European Journalism Centre.

Urbani, S. (2019). Verifying online information. Technical report, First Draft.

Zlatkova, D., Nakov, P., & Koychev, I. (2019). Fact-Checking Meets Fauxtography: Verifying Claims About Images. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2099–2108.

# References Cont'd


Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. **Modeling Information Change in Science Communication with Semantically Matched Paraphrases.** EMNLP 2022.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. **COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic.** ACL 2021.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. **Evidence-based Fact-Checking of Health-related Claims.** EMNLP 2021.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. **Fact or Fiction: Verifying Scientific Claims.** EMNLP 2020.

Nils Reimers and Iryna Gurevych. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.** EMNLP 2019.

Deka, P., Jurek-Loughrey, A., & Padmanabhan, D. (2022). **Improved methods to aid unsupervised evidence-based fact checking for online health news.** Journal of Data Intelligence, 3(4), 474-504.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. **One Embedder, Any Task: Instruction-Finetuned Text Embeddings.** Findings of ACL 2023.