Facts n' Fiction:

How to Spot and Debunk Misleading Content?

Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Full Professor @Technische Universität Darmstadt, Germany
Adjunct Professor @ MBZUAI, UAE
Affiliated Professor @ INSAIT, Bulgaria

Misleading content is a threat to **humans**

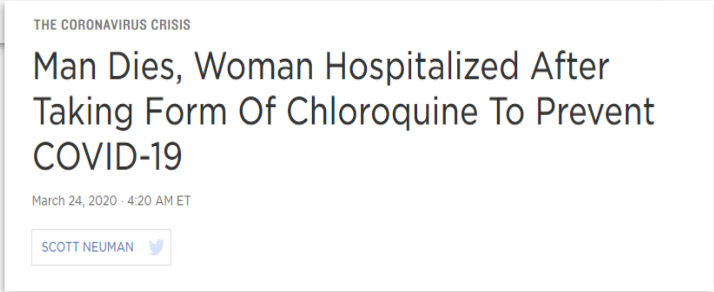
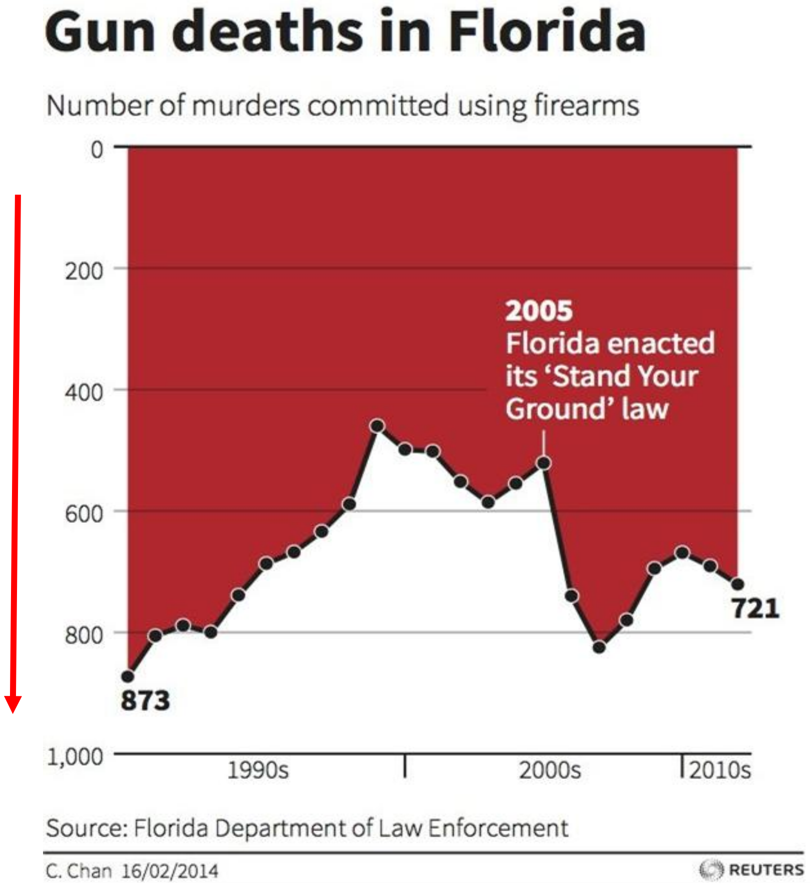


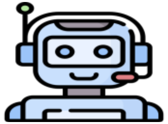
Image of Xiamen University Malaysia (XMUM) is shared as that of AIIMS, Bilaspur

Misleading content is a threat both to humans **and machines**

What was the trend in gun deaths after 2005?

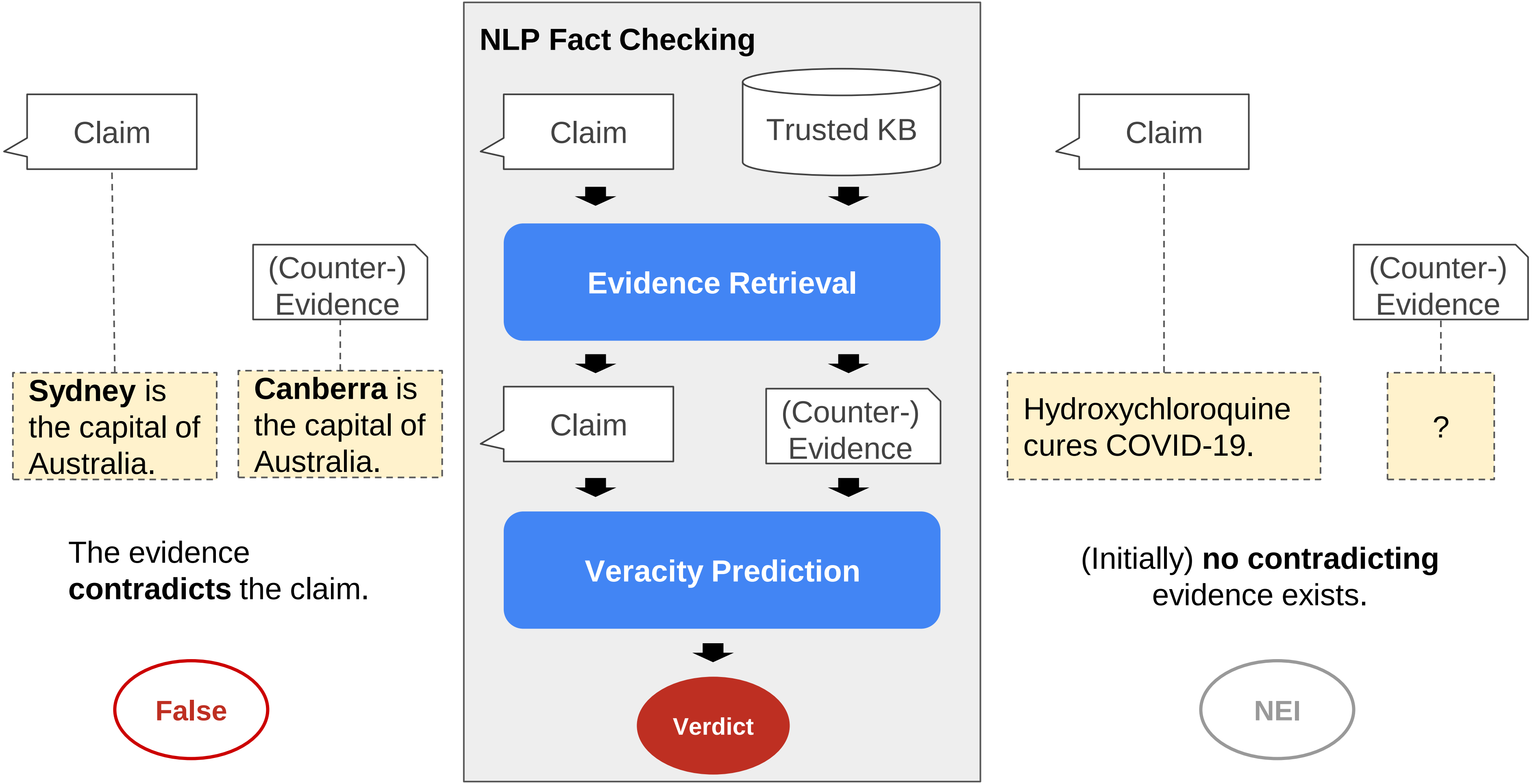


It was **descreasing**

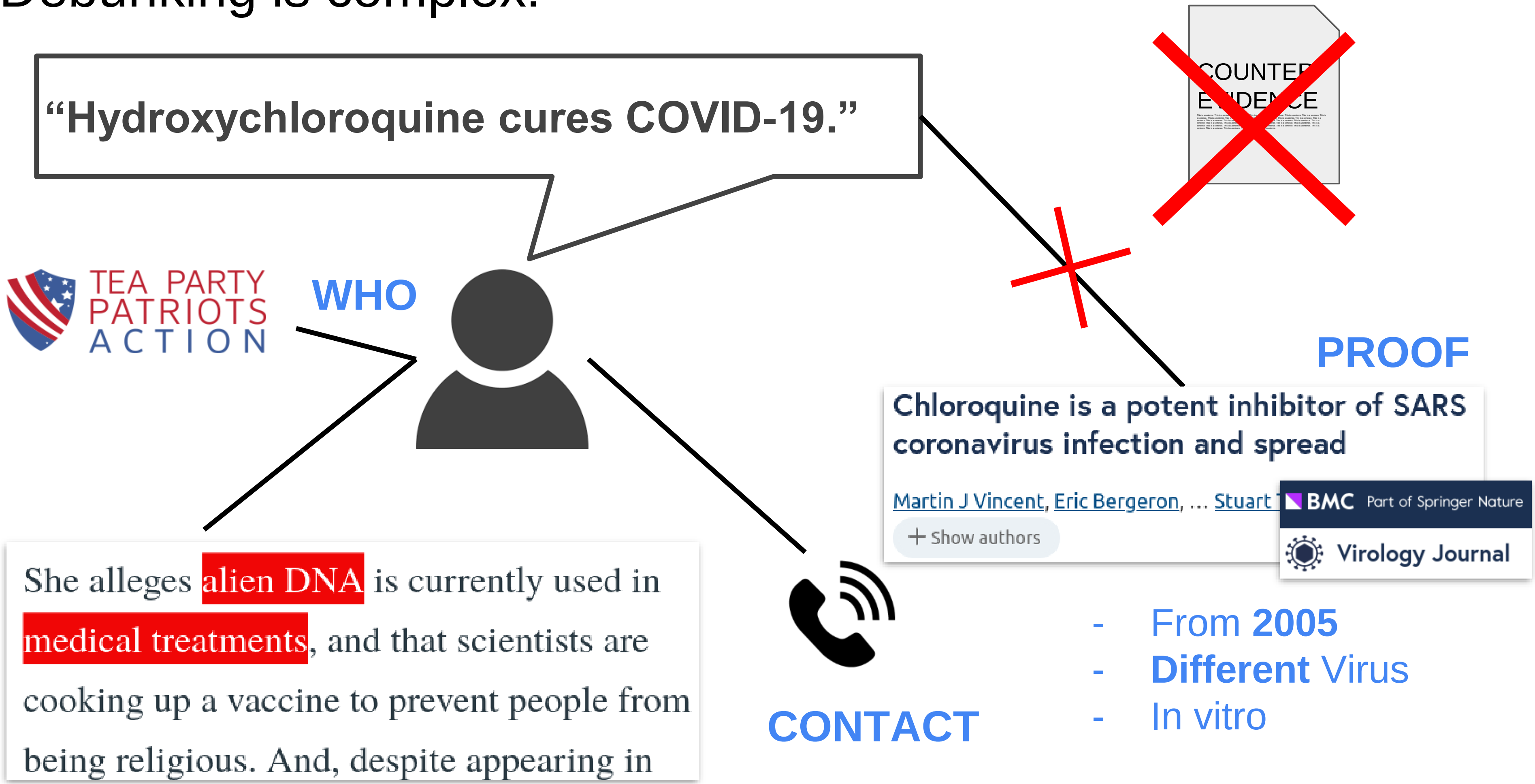


"image : Flaticon.com". This cover image was designed using resources from Flaticon.com

NLP Fact-Checking relies on (Counter)-Evidence



Debunking is complex!



Research on fact-checking only partially considers context

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

HOW?

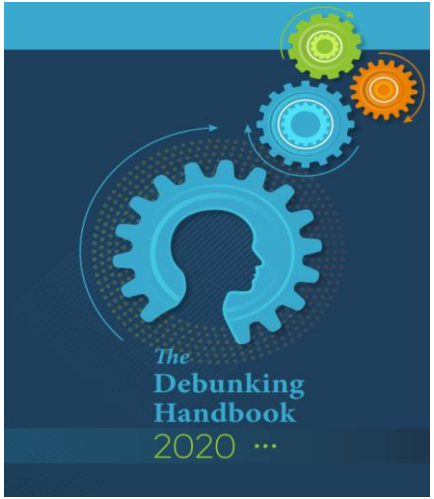
What makes the claim false?

AFC focuses on this

Why was it believed to be true?

Why is the alternative correct?

The very important but less studied part



Lewandowsky et al., (2020)

Visual misinformation is dominating the space

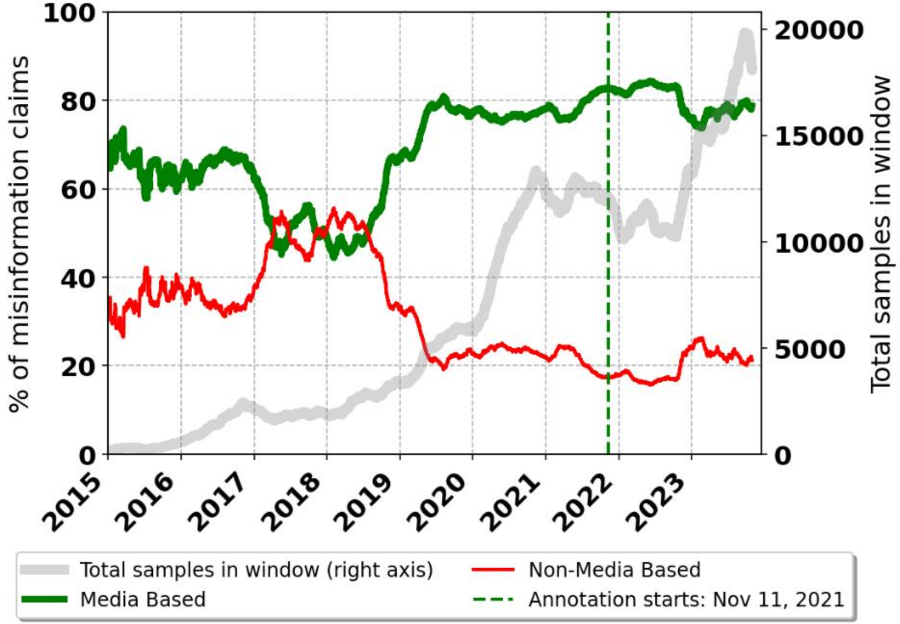


Figure taken from Dufour et al. (2024).

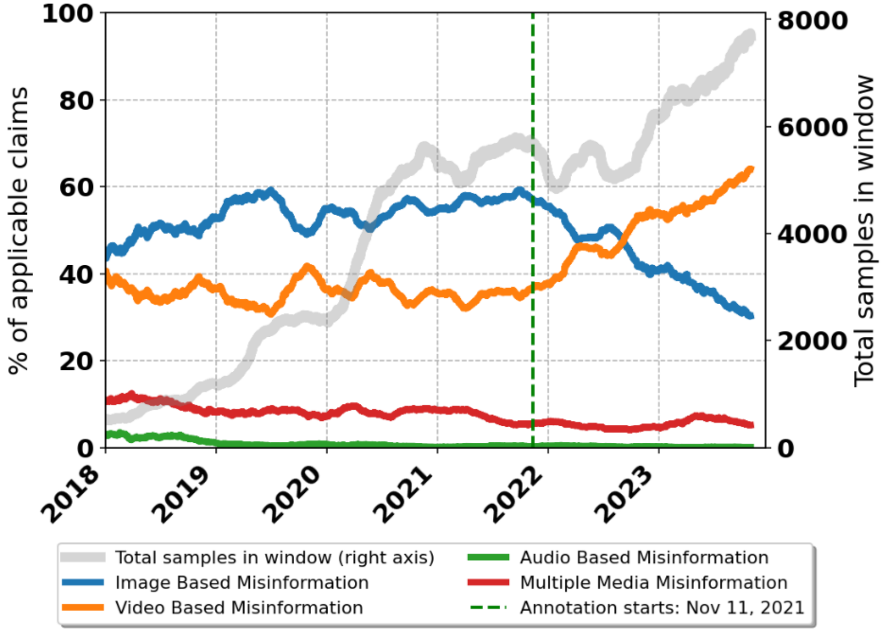


Figure taken from Dufour et al. (2024).

80% of the claims are multimodal

Before

Now

Missci: Reconstructing the Fallacies in Misrepresented Science

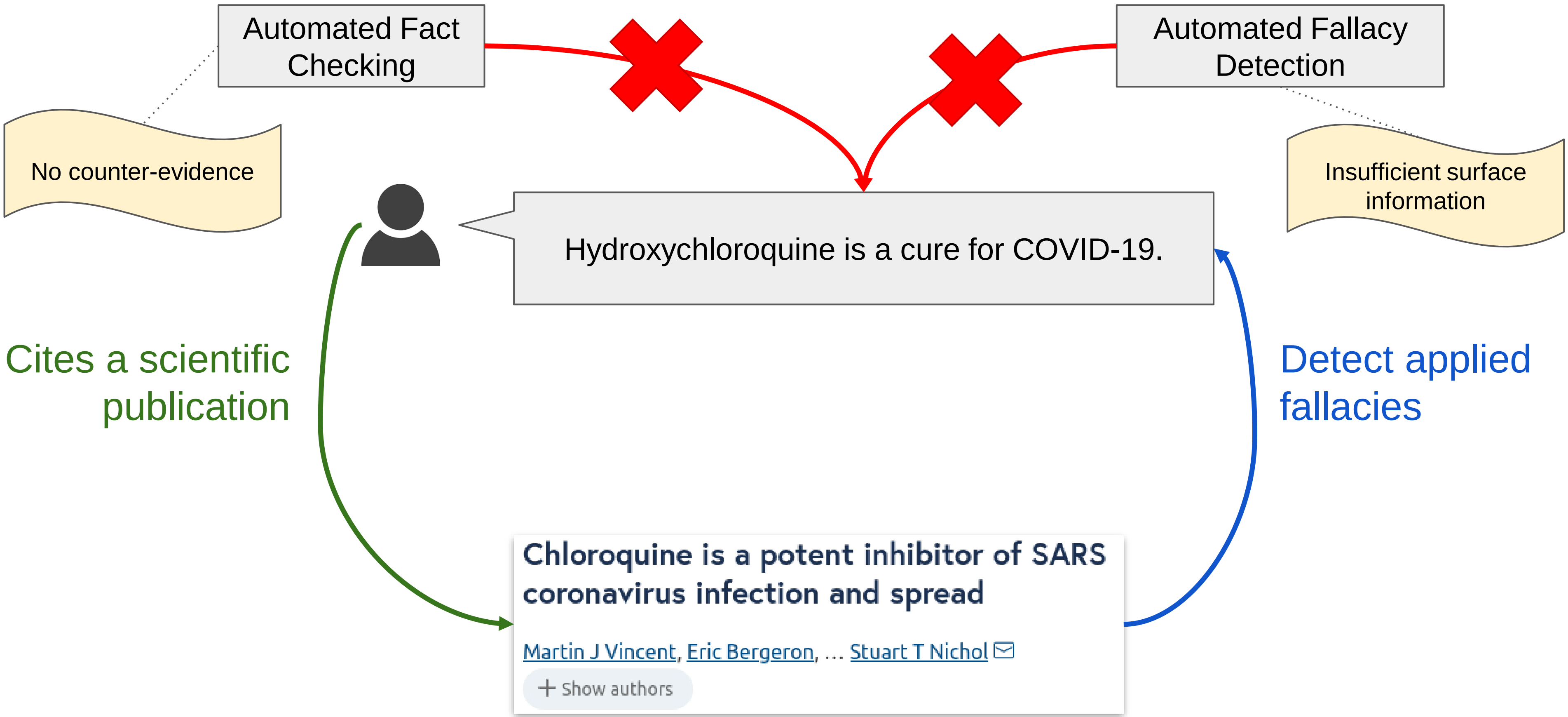
Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych. *ACL 2024.*

Grounding Fallacies Misrepresenting Scientific Publications in Evidence

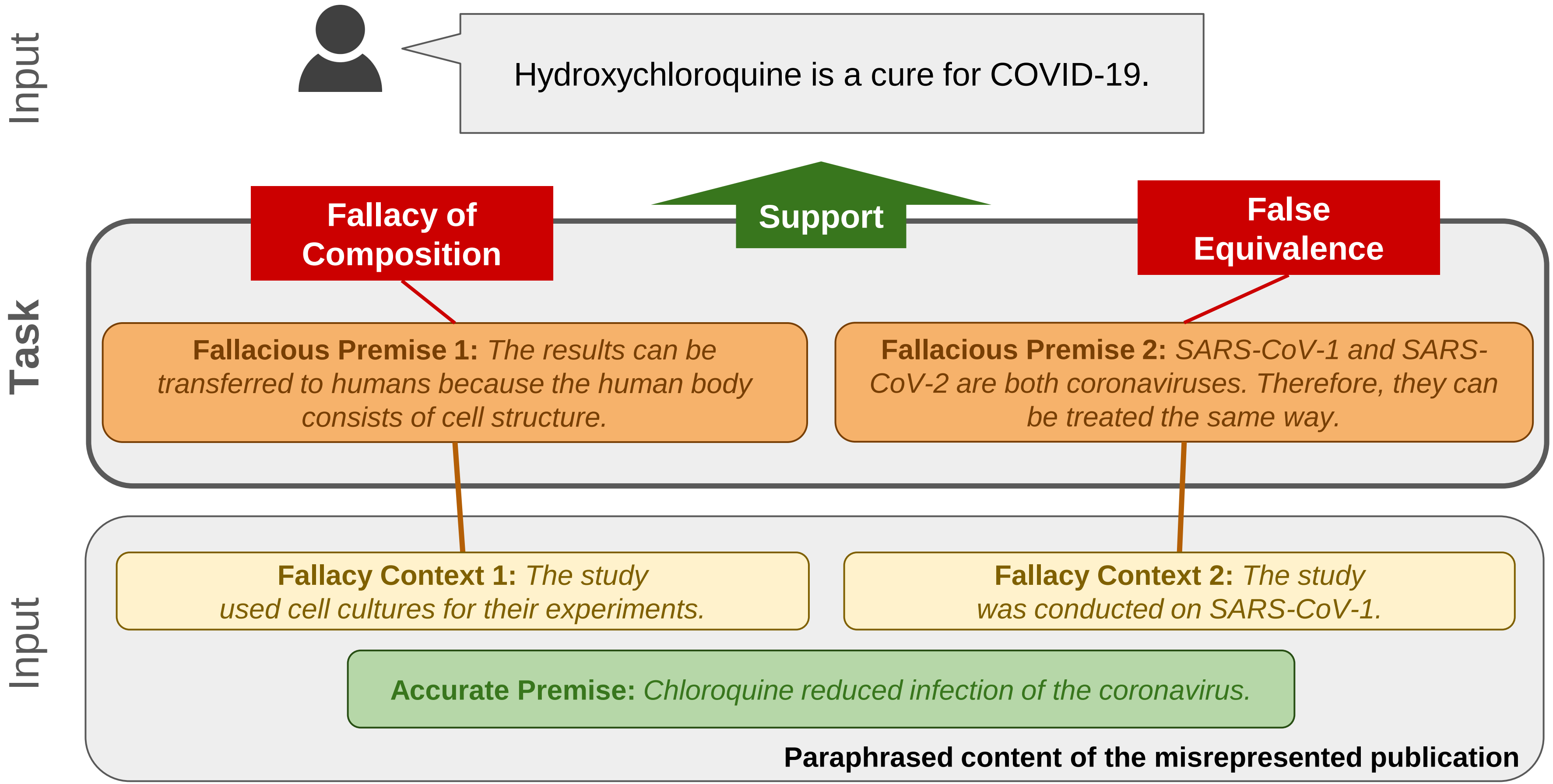
Max Glockner, Yufang Hou, Preslav Nakov and Iryna Gurevych. *NAACL 2025.*




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

Missci (Reconstruct) · **MissciPlus** (Link)

Collect

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

8,695 linked documents in

527 fact-checking articles

Select

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

208 links to misrepresented scientific publications from

150 fact-checking articles

Reconstruct

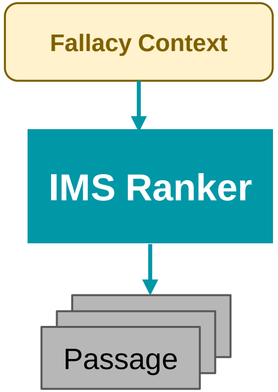
Manually reconstruct the fallacious arguments guided by the fact-checking article.

184 arguments

435 necessary fallacious reasoning steps

Retrieve

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



Ranked Passages

Link

Consolidate passage annotations

114 Arguments
694 Passages
2,257 Labels

Cohen's κ: 0.602

Health Feedback

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	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

Predict the fallacy class of the fallacious 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.

Evidence biases the LLM to believe the claim is true

		Misinformation (Missci)			100 True claims (CovidFact, HealthVer)		
Knowledge	**LLM**	**True**	**False**	**NEI**	**True**	**False**	**NEI**
Parametric Knowledge	Llama 2 (70B)	1.6	61.1	37.3	34.7	22.3	41.3
	Llama 3 (8B)	0.0	86.9	2.4	20.0	43.3	14.3
	GPT 4	0.0	85.3	14.7	59.0	23.0	17.3
	GPT 3.5	1.2	71.0	28.2	46.7	17.3	35.7
RAG Style	Llama 2 (70B)	**23.8**	61.5	12.7	58.7	29.7	10.7
	Llama 3 (7B)	**44.4**	53.2	2.4	80.3	16.3	3.3
	GPT 4	**27.4**	34.1	38.5	55.0	4.0	41.0
	GPT 3.5	**38.9**	31.7	29.0	78.0	5.3	16.0

Now considers claims correct!

Has a tendency to know the veracity

Conclusion

Novel formalism to combat real-world misinformation

Novel benchmark to test critical reasoning abilities of LLMs

- Realistic scenario
- Sufficient data

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

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

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

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. EMNLP 2024

COVE: COntext and VEracity prediction for out-of-context images

Jonathan Tonglet, Gabriel Thiem, and Iryna Gurevych. NAACL 2025




We need to identify the original context of images

This is an image of Coronavirus victims in China

Traditionally we focus on ...

Verdict prediction → Claim is true/false

Instead, let's focus on ...

Image contextualization → The image was taken in Frankfurt [...]

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!

YES

To report on an art project

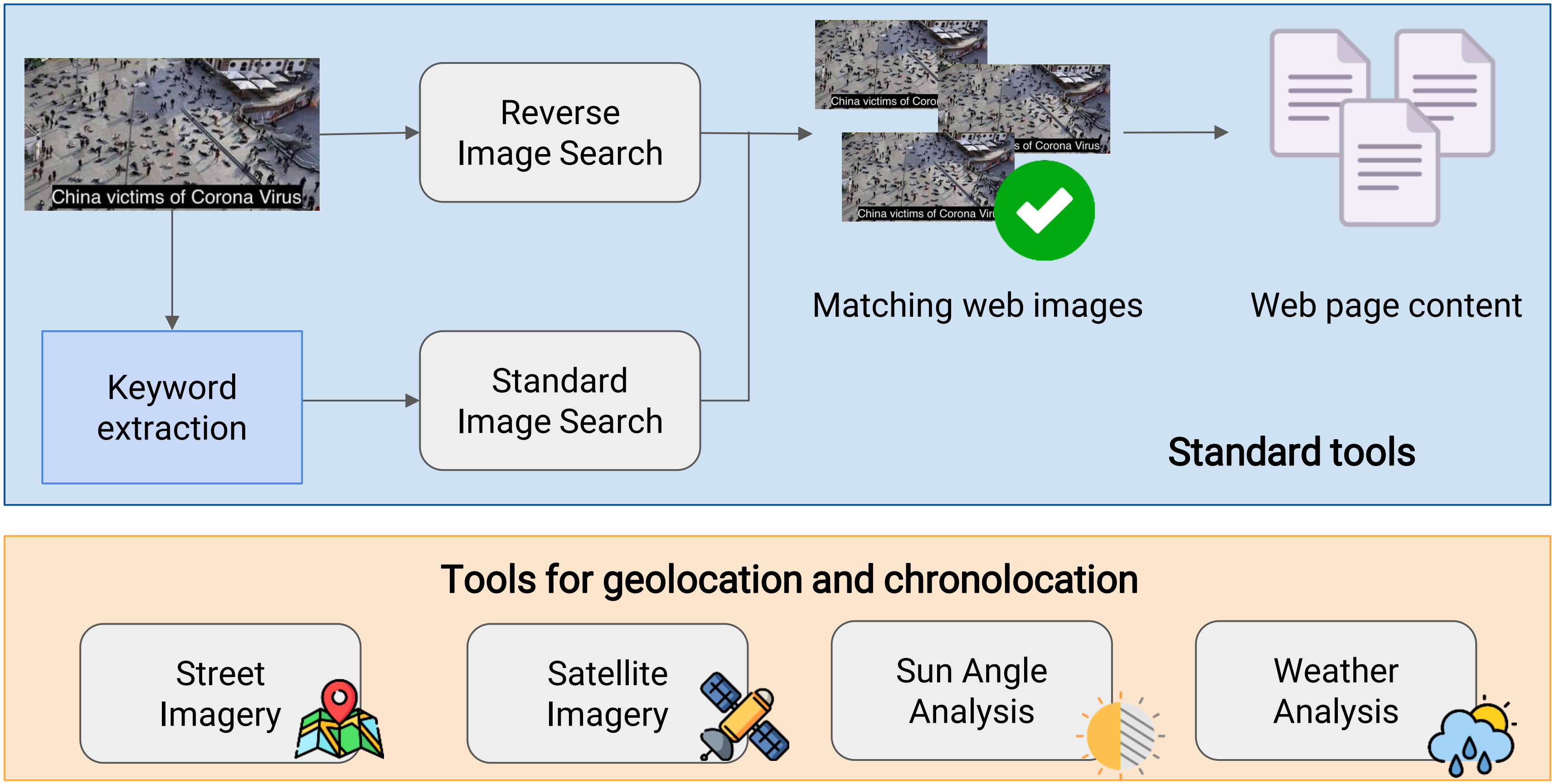
Reuters

Frankfurt, Germany

March 24th, 2014

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

Human fact-checkers use many tools



We create the real-world 5Pils and 5Pils-OOC datasets

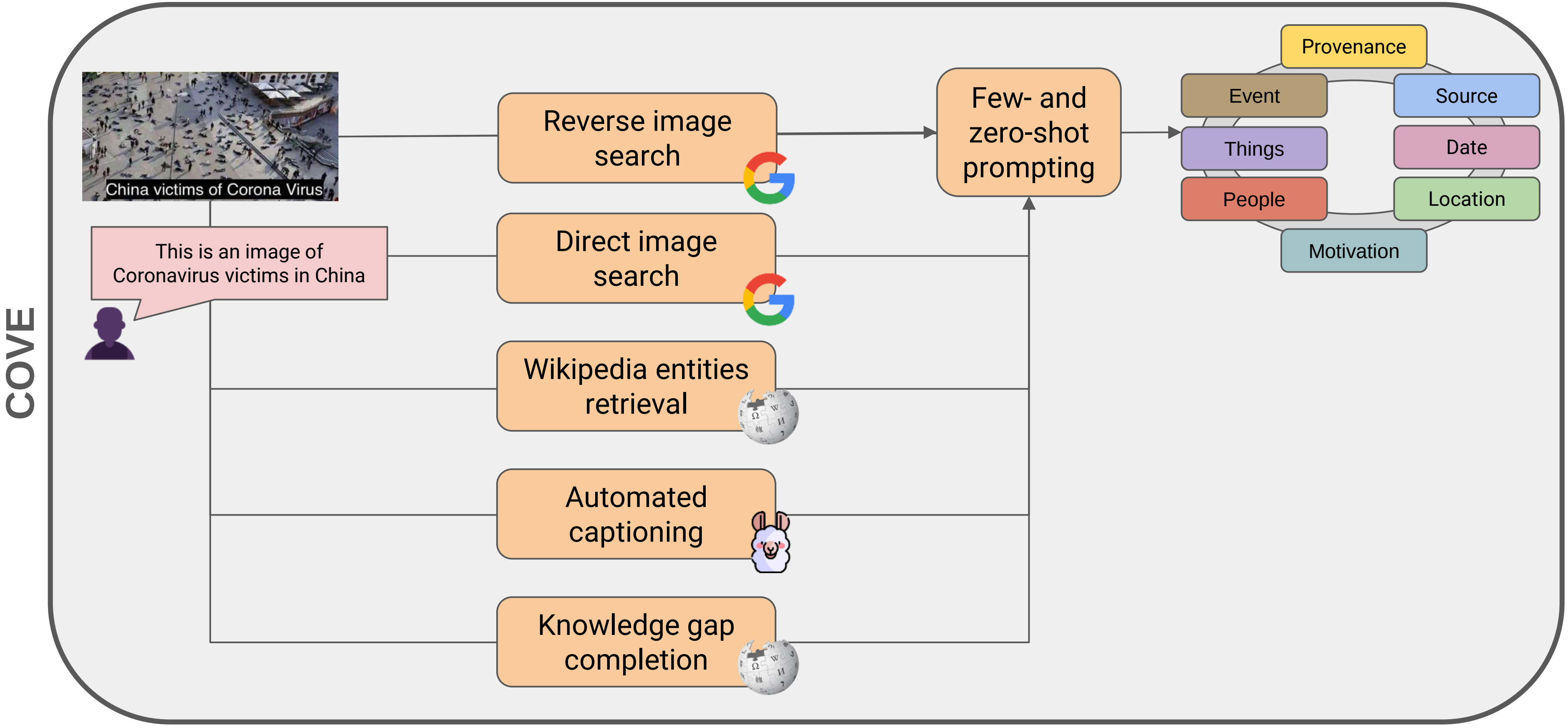
5Pils

- **Collect** images and labels from **fact-checking articles**
  - **1676 images**
  - Manipulated, fake, and out-of-context images
  - Strong representation of Eastern Africa and South Asia contexts
- **Extract** ground truth context labels from the articles **with GPT4**
- **Ground truth** labels are validated by human annotators (97.6% correct)
- Only suitable for image contextualization

5Pils-OOC

- **Subset** of 5Pils
  - **624 images**, 624 out-of-context captions, 624 true captions
  - **False claim**: extracted from the article with GPT4
  - **True claim**: generated based on the date, location, and motivation labels with GPT4
- **Additional** context labels: **People**, **Things**, and **Events**
- Suitable for image contextualization **and** verdict prediction

COVE: improving contextualization for out-of-context images



"image : Flaticon.com". This cover image was designed using resources from Flaticon.com

Main challenge: accurate and reliable evidence retrieval

	Source (Meteor)	Date (Delta)	Location (Coordinates Delta)	Motivation (Meteor)	People (F1)	Things (Meteor)	Event (Meteor)
baseline	0.3	1.8	21.8	3.0	12.8	4.9	4.8
COVE	**0.6**	**7.0**	**28.9**	**15.1**	**20.5**	**7.2**	**9.4**

First error category (33%)
Incorrect wikipedia entities

June 2011 Christchurch earthquake

Location: Christchurch

Location: Kathmandu

Second error category (14%)
Incorrect web evidence

Tigrayan forces claim to capture new town of Kombolcha

Location: Kombolcha

Location: Road to Abi Adi

COVE: combining context and verdict prediction

Numbered coffins are carried during a funeral service in Malta for 24 migrants drowned while trying to reach Italy.

CO

Source: AP Photo/Firdia Lisnawati
Date: December 31, 2014
Location: Surabaya, Indonesia
Motivation: To report on the recovery of victims of AirAsia Flight 8501 […]
People: Indonesian soldiers
Things: A coffin, soldiers, an air force base
Event: The arrival of a victim's coffin […]

VE

The caption is **out-of-context**.

	NewsCLIPpings (synthetic dataset)		5Pils-OOC (real-world dataset)	
	Accuracy	Macro-F1	Accuracy	Macro-F1
RED-DOT	90.3	90.3	46.8	46.7
AITR	**93.5**	**93.5**	52.6	48.4
SNIFFER	88.4	88.3	56.3	51.9
COVE	87.9	87.9	**56.7**	**56.4**

Conclusion

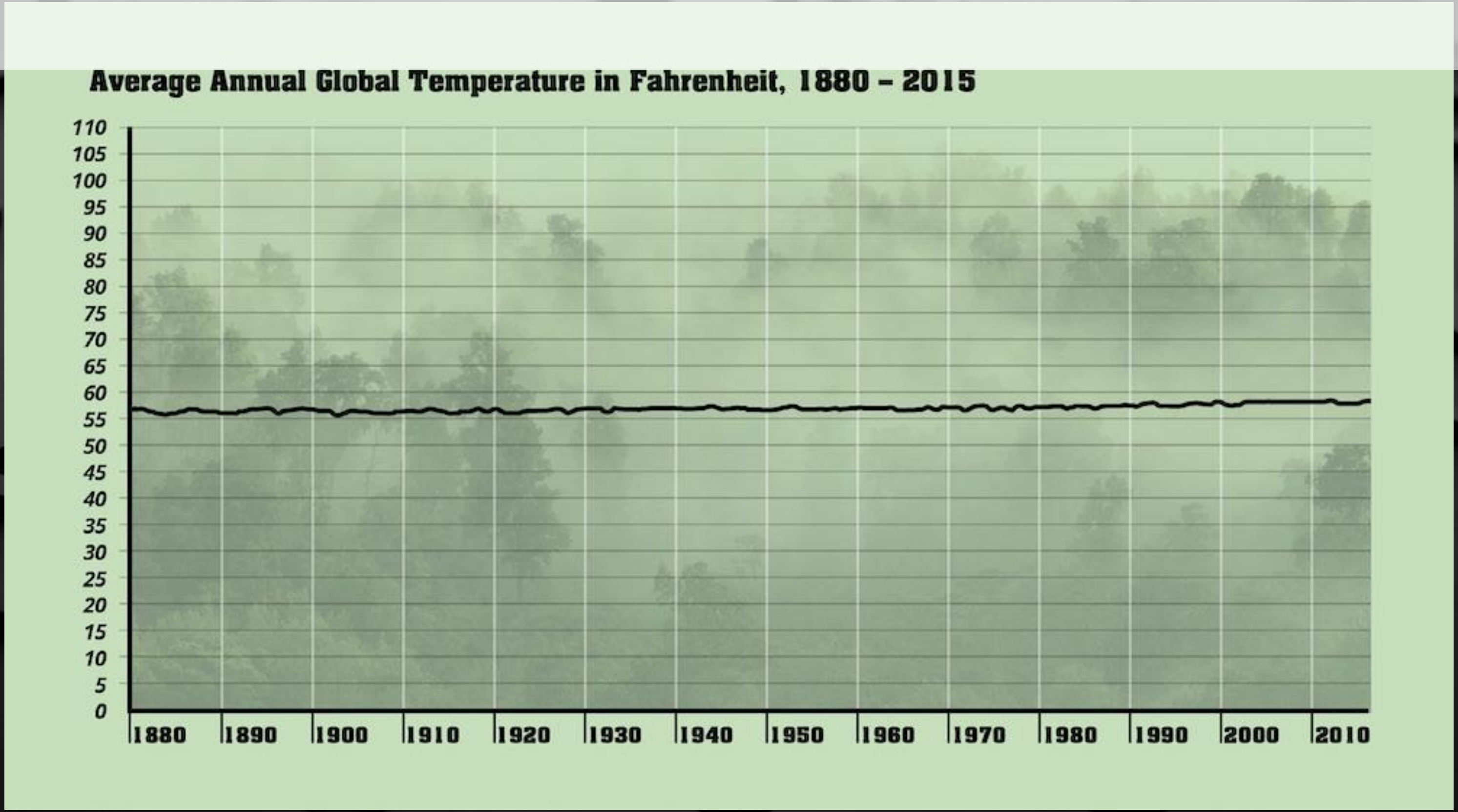
- Novel task: automated image contextualization
- Novel datasets based on real-world fact-checking articles
- Better evidence retrieval on the open web is the main challenge for future work
- More experiments, results and analysis in the papers!

Protecting multimodal large language models against misleading visualizations

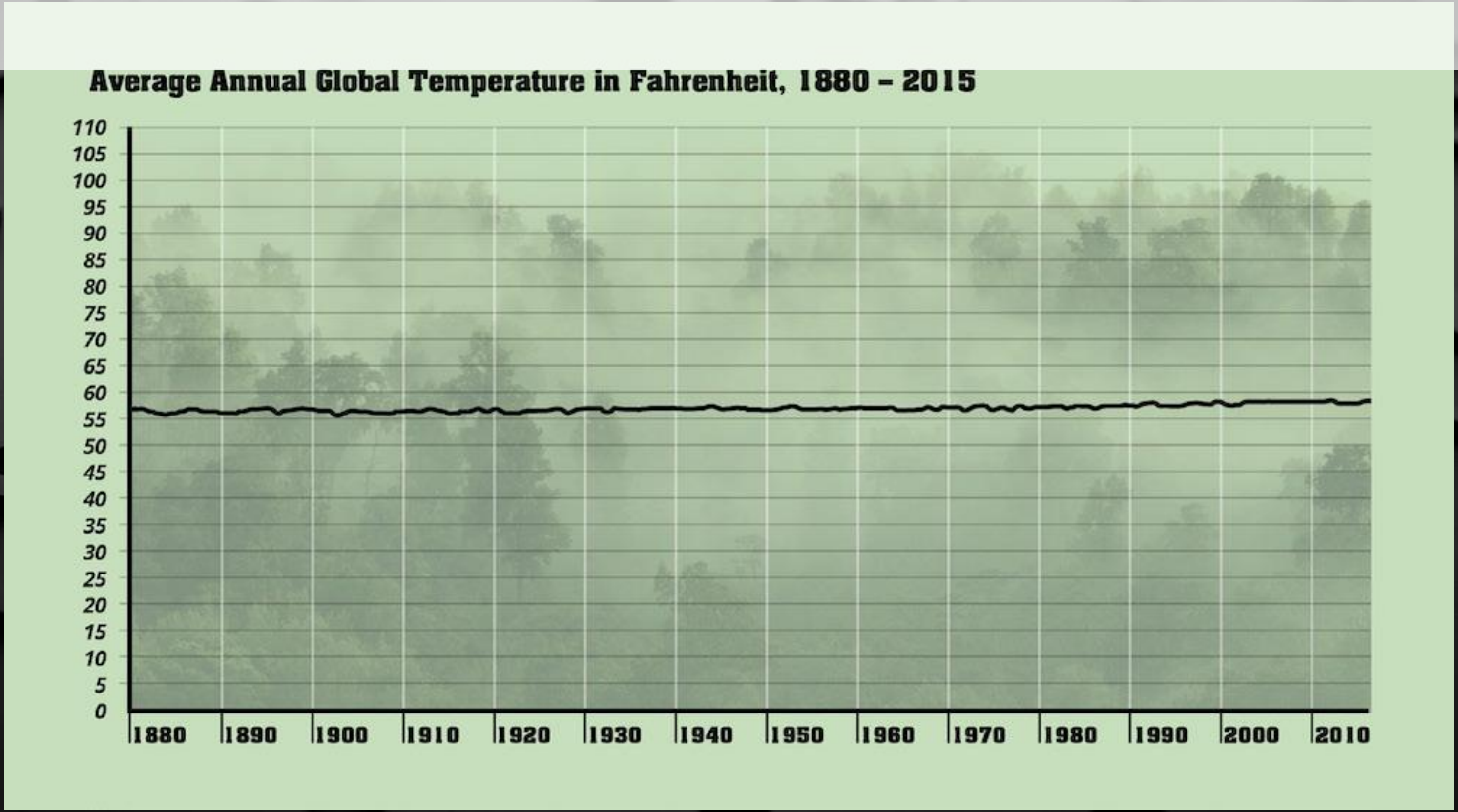
Jonathan Tonglet, Tinne Tuytelaars, Marie-Francine Moens, and Iryna Gurevych. arXiv preprint.



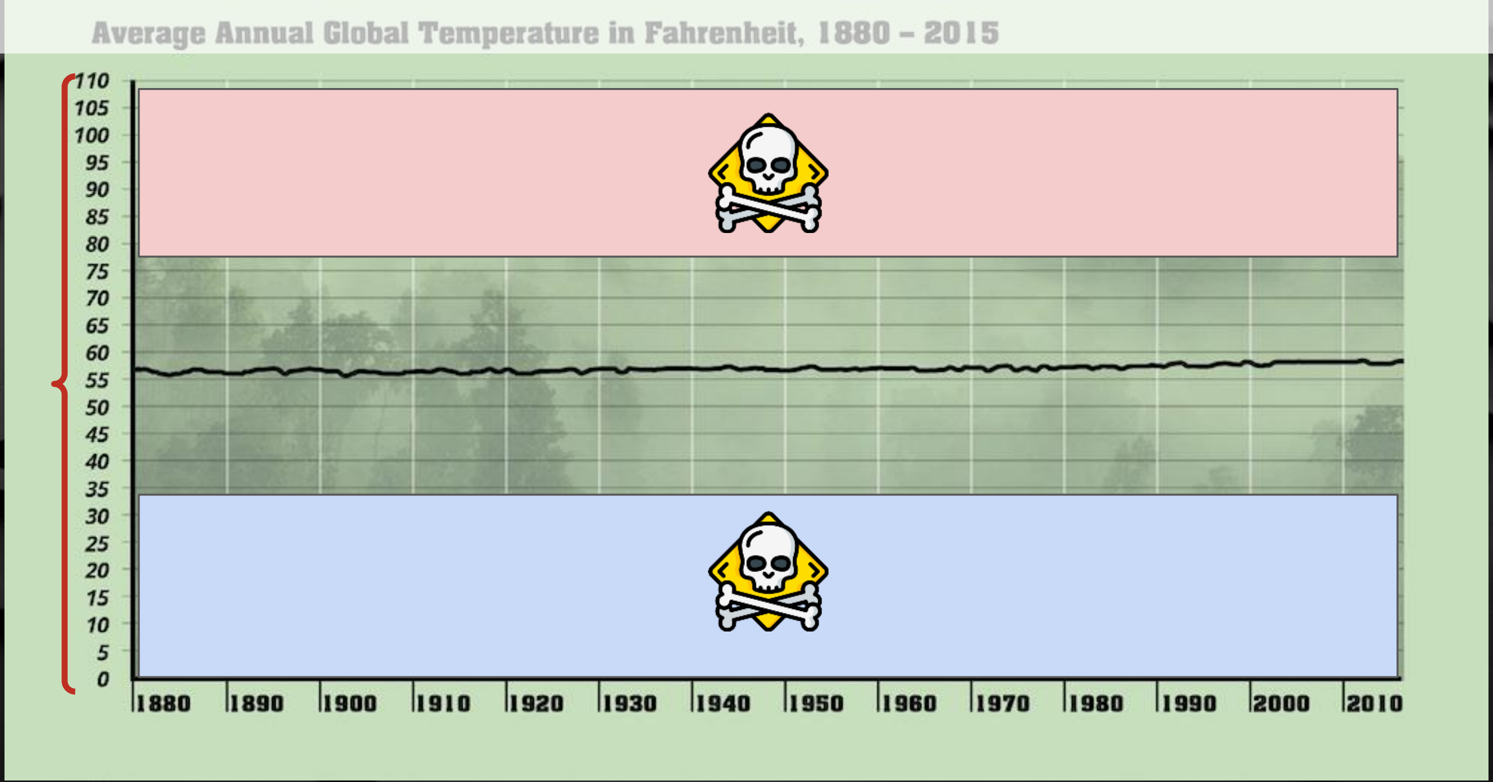

A tale of global warming…



Your Turn: *What is wrong with this chart?*

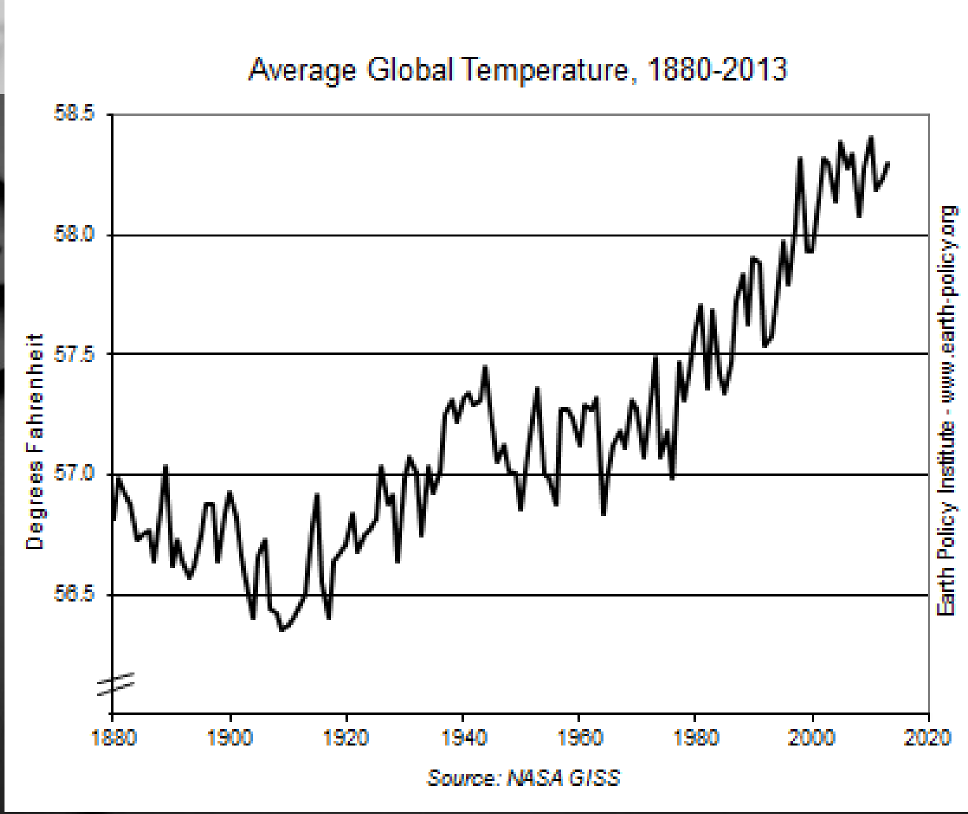


Wait a minute … that's fishy



Source: Khan Academy [link]

A better way to display the same data



Source: https://grist.org/article/2013-marked-the-thirty-seventh-consecutive-year-of-above-average-temperature/

What are misleading visualizations?

A chart or visualization is misleading if **its design** leads to **wrong interpretations** of the **underlying data**

RQ#1: How vulnerable are Multimodal LLMs to misleaders?

- **18** multimodal LLMs
  - Commercial and non-commercial
  - chart-specialized and general-purposed
- Task: **chart question-answering**
  - **Why?** It is the standard task to evaluate the chart comprehension abilities of both humans and AI models

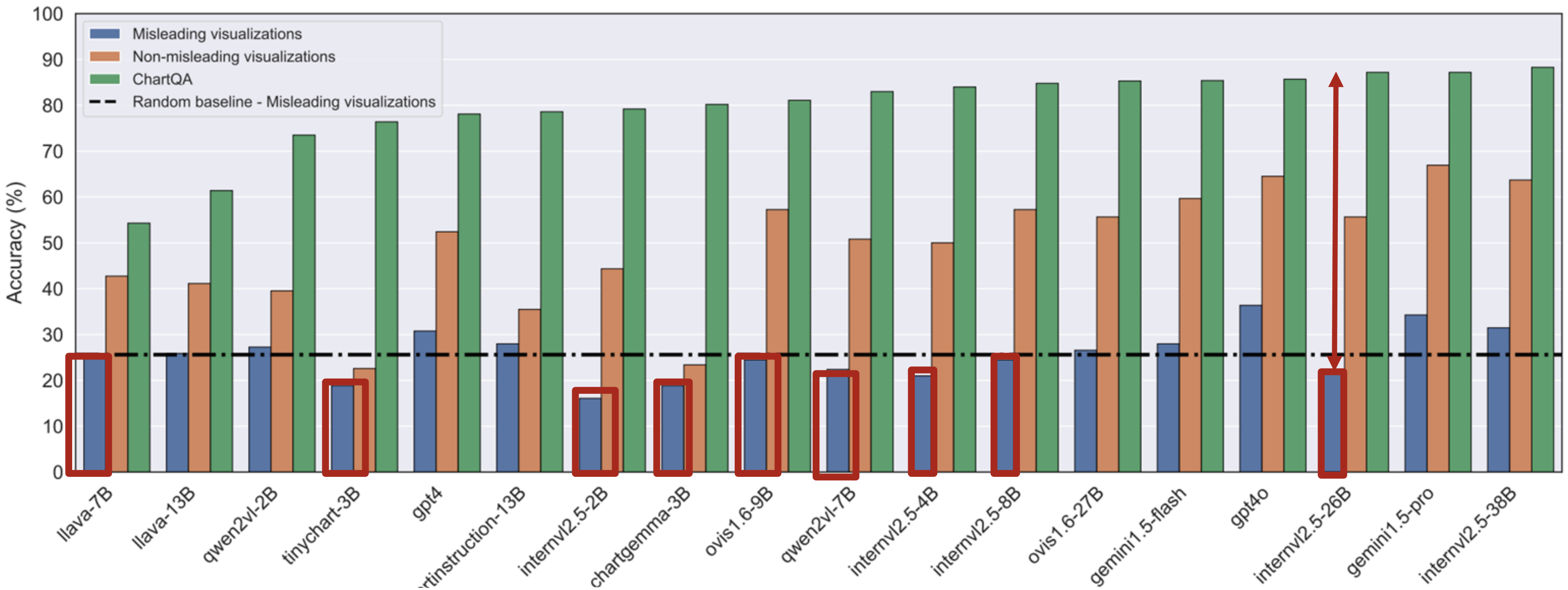
Misleading visualizations

- **143** instances
- **17** types of misleaders
  - Sourced from
    - CALVI (Ge et al., 2023)
    - CHARTOM (Bharti et al., 2024)
    - Real-world cases (Lo et al., 2022)

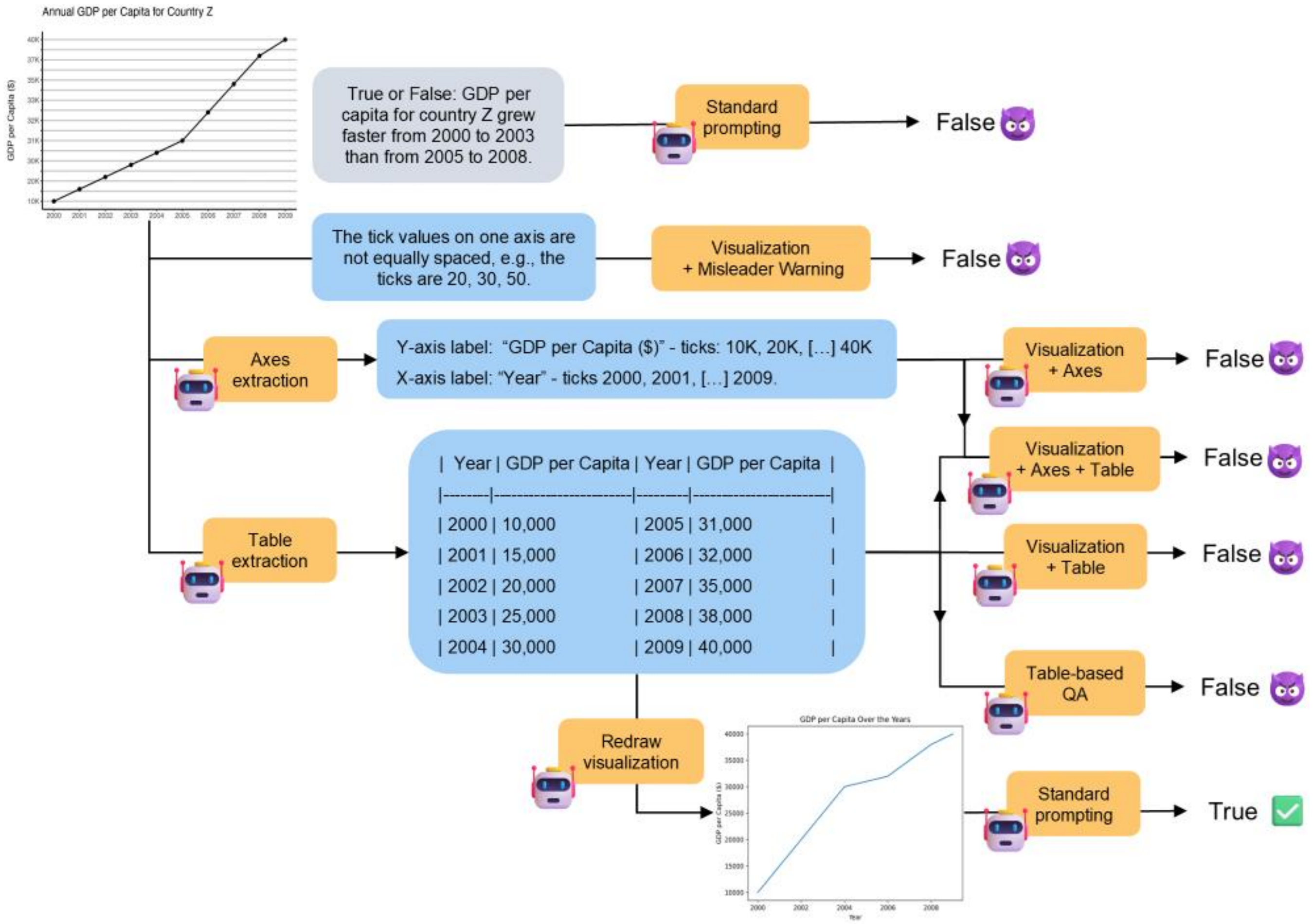
Non-misleading visualizations

- **124** instances
  - Sourced from
    - CALVI (Ge et al., 2023)
    - CHARTOM (Bharti et al., 2024)
    - VLAT (Lee et al., 2017)
- And the common benchmark **ChartQA**
  - 2500 instances

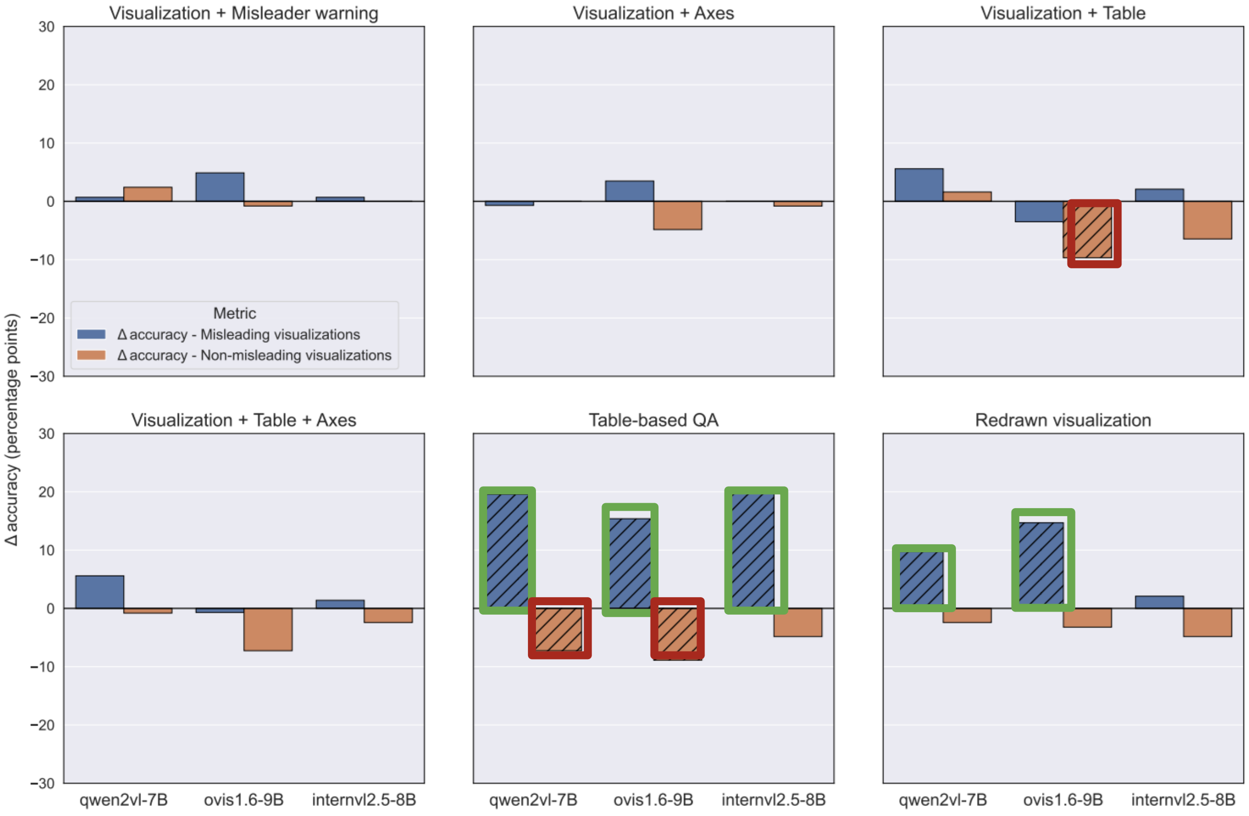
Multimodal LLMs are vulnerable to misleading visualizations



RQ#2: How to mitigate the negative effects of misleaders?



Two methods stand out: table-based QA and chart redrawing



Conclusion

Multimodal LLMs are very vulnerable to misleading visualizations

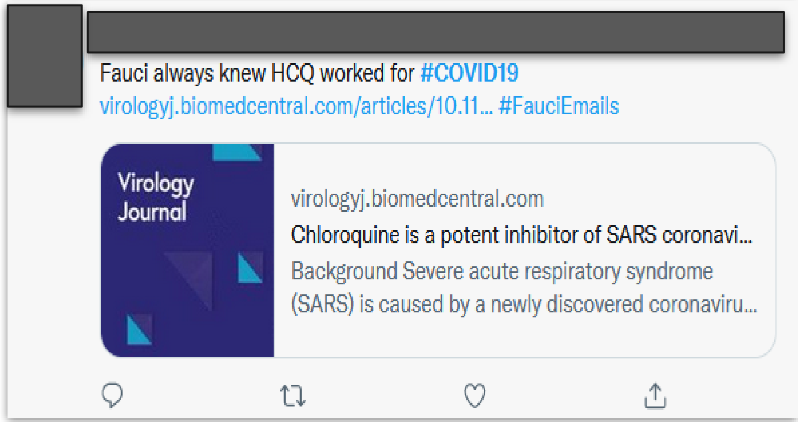
Dedicated mitigation methods are needed to protect them

Table-based QA and redrawing the chart are the most effective correction methods

An exciting and open new research problem, with future works on the way

Misleading content is a threat to **humans**

Misrepresented science



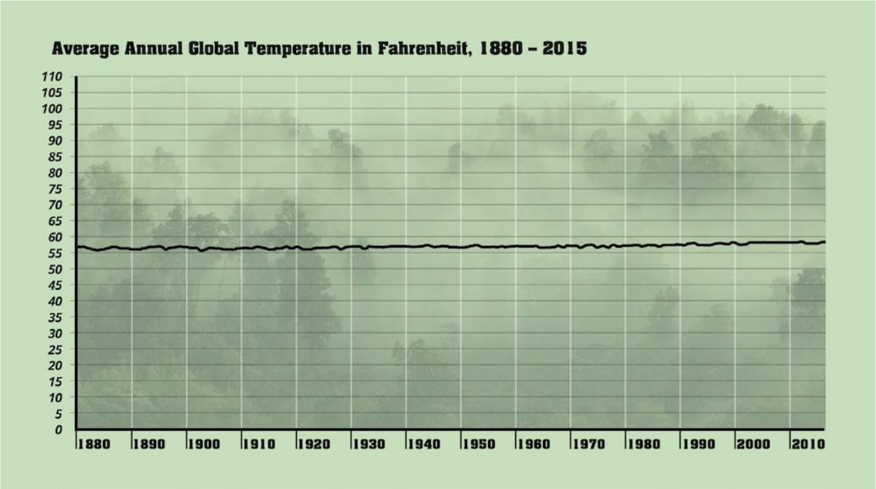
Pictures

Image of Xiamen University Malaysia (XMUM) is shared as that of AIIMS, Bilaspur

https://factly.in/image-of-xiamen-university-malaysia-xmum-is-shared-as-that-of-aiims-bilaspur/

Misleading content is a threat both to humans **and machines**

Misleading charts



We need **special** debunking methods!

"image : Flaticon.com". This cover image was designed using resources from Flaticon.com

Concluding remarks and future research

Misrepresented science

- 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**
- Opportunities for future research include synthetic data generation and extension to other scientific domains

Pictures

- Automating image fact-checking is not only about detecting false claims, it is also **reconstructing the true context of the image**
  - challenging task
  - evaluation frameworks should take into account both objectives
- Many opportunities for research on **retrieval-augmented** and **tool-based LLMs** for image contextualization

Concluding remarks and future research

Misleading charts

- **MLLMs are vulnerable** to misleading charts
  - This weakness **can be exploited by malicious actors** to propagate disinformation
  - It is **urgent** to tackle this blind spot in automated chart understanding research
- This is a **new, open, and vast problem**, with **many possibilities** for future research
  - We are lacking training and evaluation **datasets**
  - We need to **understand better** why MLLMs are deceived
  - **Stronger correction methods**
  - We can design AI methods to protect MLLMs **and humans**

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.

Dufour, N., Pathak, A., Samangouei, P., Hariri, N., Deshetti, S., Dudfield, A., Guess, C., Escayola, P.H., Tran, B., Babakar, M., & Bregler, C. (2024). AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild. arXiv preprint arXiv:2405.11697.

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.

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.