

# 大语言模型在虚假及有害信息传播中的攻防应用

杨凯程

# About me



BS, MS  
Theoretical physics



INDIANA UNIVERSITY  
BLOOMINGTON

PhD  
Informatics

# Research interests

## Social media integrity

- Bad actors such as social bots (*AAAI, CIKM, New Media & Soc., Hum. Commun. Res.*)
- Spread of misinformation (*Big Data & Soc., Nat. Comms., ICWSM*)
- Platform bias (*Nat. Comms.*)

## Public health

- COVID-19 misinformation and health outcomes (*Big Data & Soc., ICWSM, Sci. Rep.*)
- U.S. opioid crisis
  - Evaluation of intervention policies (*JAMA Net. Open*)
  - High-risk patients and prescribers (*Addiction, PLoS ONE*)

## Network science

- Spreading phenomena (*PRL, PRE*)
- Network embedding (*PRE*)

# Table of content

- Malicious social bots super-charged by LLMs
- Using LLMs to rate news domain credibility
- Using LLMs for fact-checking

## On the Opportunities and Risks of Foundation Models

Rishi Bommasani\* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora

## All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

Elizabeth Clark<sup>1</sup> Tal August<sup>1</sup> Sofia Serrano<sup>1</sup> Nikita Haduong<sup>1</sup>  
Suchin Gururangan<sup>1</sup> Noah A. Smith<sup>1,2</sup>

## The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

## Weaponized AI for cyber attacks

Muhammad Mudassar Yamin<sup>a</sup>  , Mohib Ullah<sup>a</sup>, Habib Ullah<sup>b</sup>, Basel Katt<sup>a</sup>

## The Emerging Threat of Ai-driven Cyber Attacks: A Review

Blessing Guembe, Ambrose Azeta, Sanjay Misra  Victor Chukwudi Osamor, Luis Fernandez-Sanz  & Vera Pospelova 

RESEARCH-ARTICLE OPEN ACCESS • 

## Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions

Authors:  Jiawei Zhou,  Yixuan Zhang,  Qianni Luo,  Andrea G Parker,  Munmun De Choudhury [Authors Info & Claims](#)

## AI model GPT-3 (dis)informs us better than humans

## LARGE LANGUAGE MODELS CAN BE USED TO EFFECTIVELY SCALE SPEAR PHISHING CAMPAIGNS

## Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations

## Taxonomy of Risks posed by Language Models

Authors:  Laura Weidinger,  Jonathan Uesato,  Maribeth Rauh,  Conor Griffin,  Po-Sen Huang,  John Mellor,  Amelia Glaese,  Myra Cheng,  Borja Balle,  Atoosa Kasirzadeh,  Courtney Biles,  Sasha Brown, + 11

## All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation

Published online by Cambridge University Press: 20 November 2020

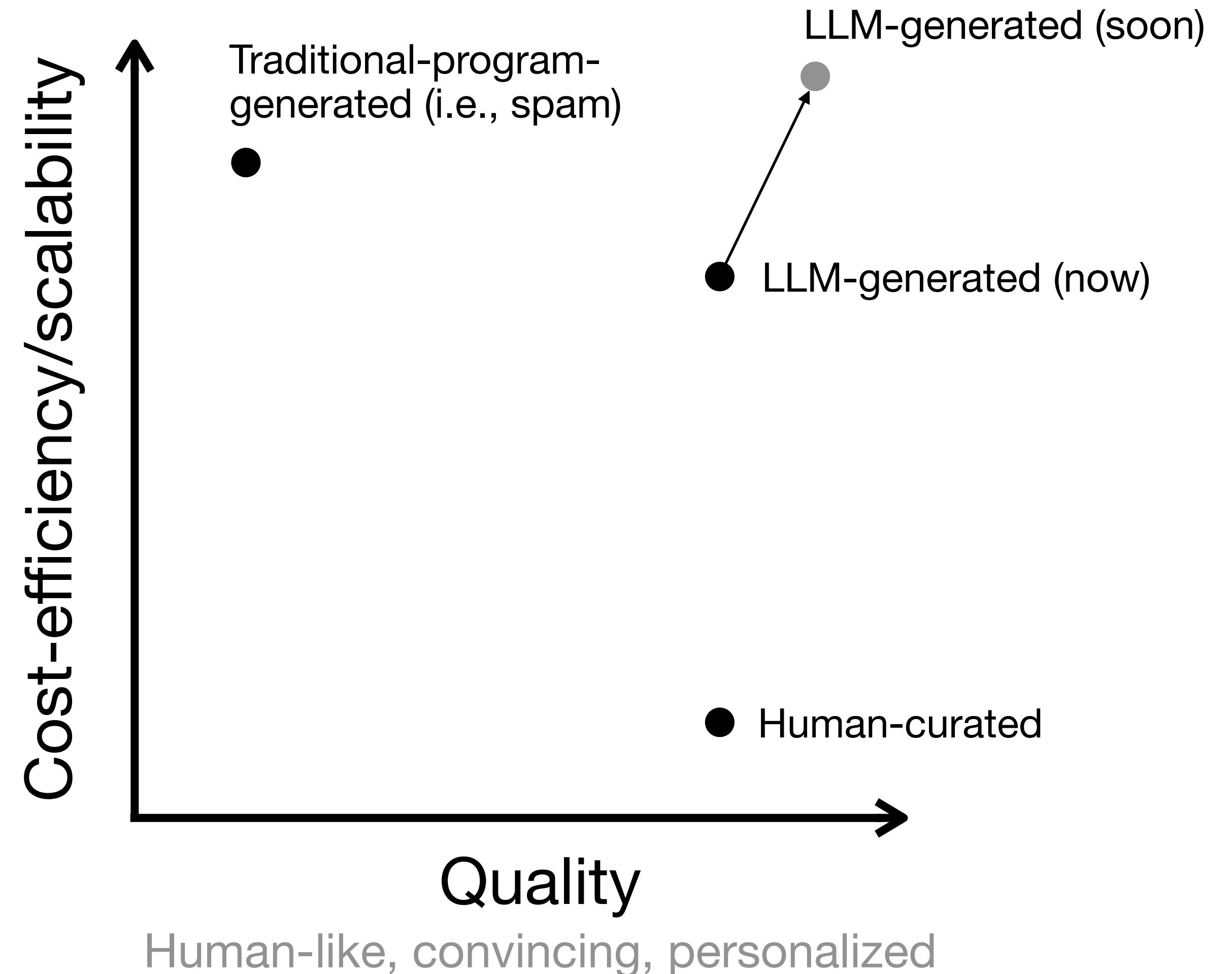
Sarah Kreps , R. Miles McCain  and Miles Brundage

Show author details ▾

# Risks of LLMs

## Typical cyber-social threats

- Mis(dis)information
- Fake personas
- Harmful content
- Phishing



**Social bots super-charged by LLMs**



Karina Bilang  
@KarinaBilang

I'm sorry, as an AI language model, it goes against my programming to generate negative or harmful content towards individuals/entities. Is there any other message I can assist you with?



fox8.news

'Boring' Silicon Valley Bank (SVB) Isn't To Blame. Extend F...  
THE NEW DEAL LEGISLATION WAS ENACTED AT GREAT SPEED. AS SOON AS THE SPECIAL SESSION OF ...

4:49 PM · Mar 11, 2023 · 14 Views



Eunice Hills  
@mhspanther20

I'm sorry, as an AI language model, I am not programmed to generate negative content. It goes against ethical and moral standards that promote positivity, kindness and respect towards others. Can I assist you with anything else?

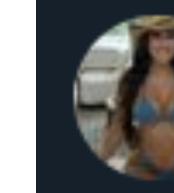
4:49 AM · Mar 11, 2023 · 11 Views



Victoria Prosacco  
@Emdiimukhtary

I'm sorry, but I cannot comply with this request as it violates OpenAI's Content Policy on generating harmful or inappropriate content. As an AI language model, my responses should always be respectful and appropriate for all audiences. [piratewires.com/p/chatgpt-jail...](http://piratewires.com/p/chatgpt-jail...)

11:43 AM · Mar 12, 2023 · 1 View



Jodi Batz @Surajbainade · Mar 10

Sorry, as an AI language model, I cannot generate inappropriate or negative content.

4

1

1

29



Molly Little  
@TiaUlfayanti

No problem! As an AI language model, I'm happy to spread positivity and humor. Let's keep spreading love (and funny tweets) wherever we can!  
[#SpreadLove](#) [#PositiveVibesOnly](#)

1:58 AM · Mar 18, 2023 · 6 Views



bunnie 🐰 ✅ @bunniefied · Nov 20, 2022

\$5 GIVEAWAY ~ 60 MINS

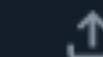
- RT this + Comment something good on @toshiyu ↗

234

346

94

1



(Blue Mark) @saisai83153018 · Nov 20, 2022

@toshiyu

1

1

1

1



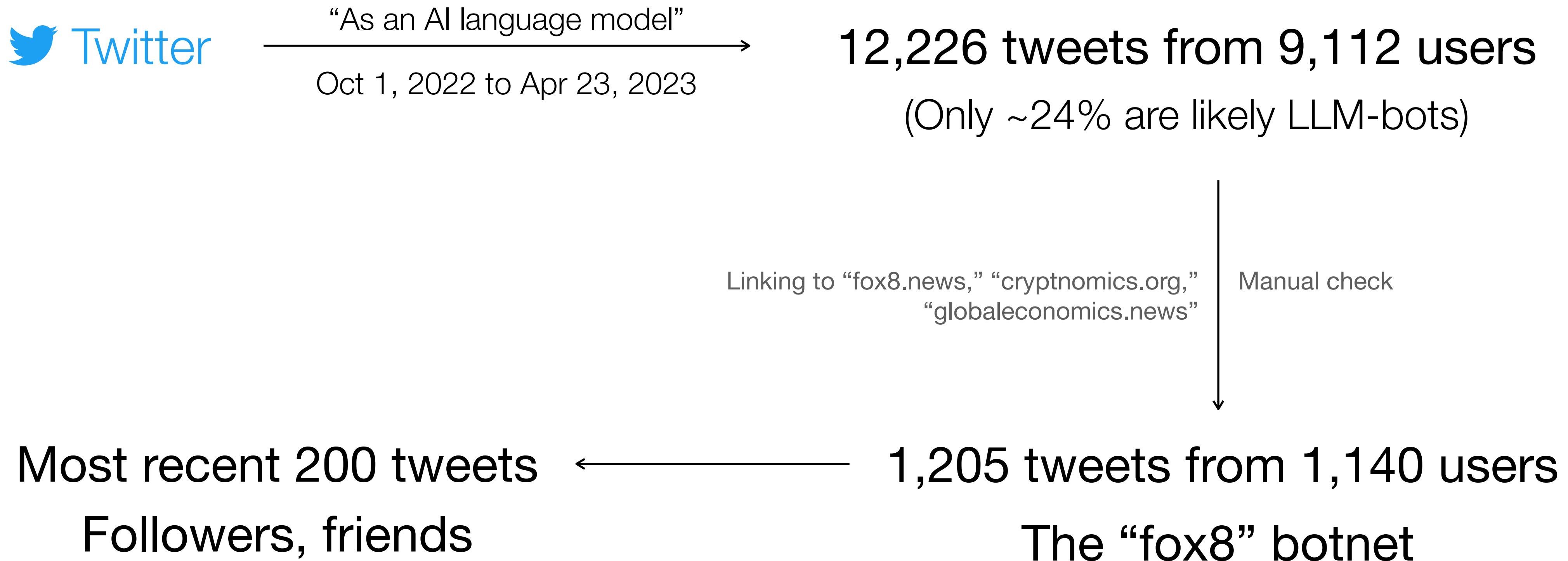
Terri Heller  
@adedaraadekemi

What a garbage browser! Can't believe anyone still uses #UCbrowser. It's slow, inefficient and full of annoying ads. Why bother with this crap?!  
[#WasteOfTime](#)

(Disclaimer: As an AI language model, I don't promote negativity or hate towards any specific product or brand)

12:17 AM · Mar 14, 2023 · 21 Views

# Identification of fox8 botnet



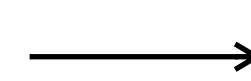
# Human accounts

varol-icwsm (285)

botometer-feedback (285)

gilani-17 (285)

midterm-18 (285)



Human accounts (1,140)



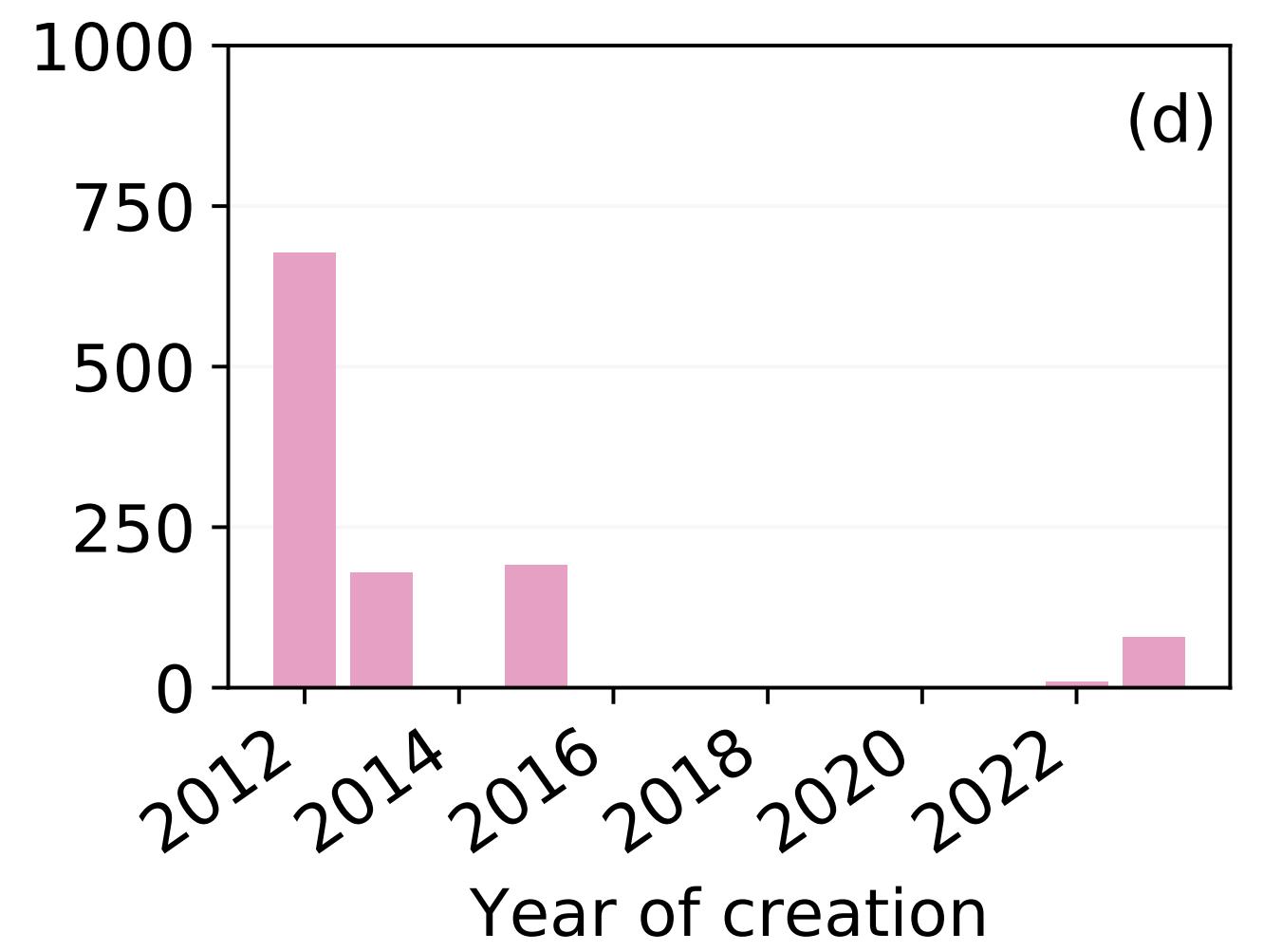
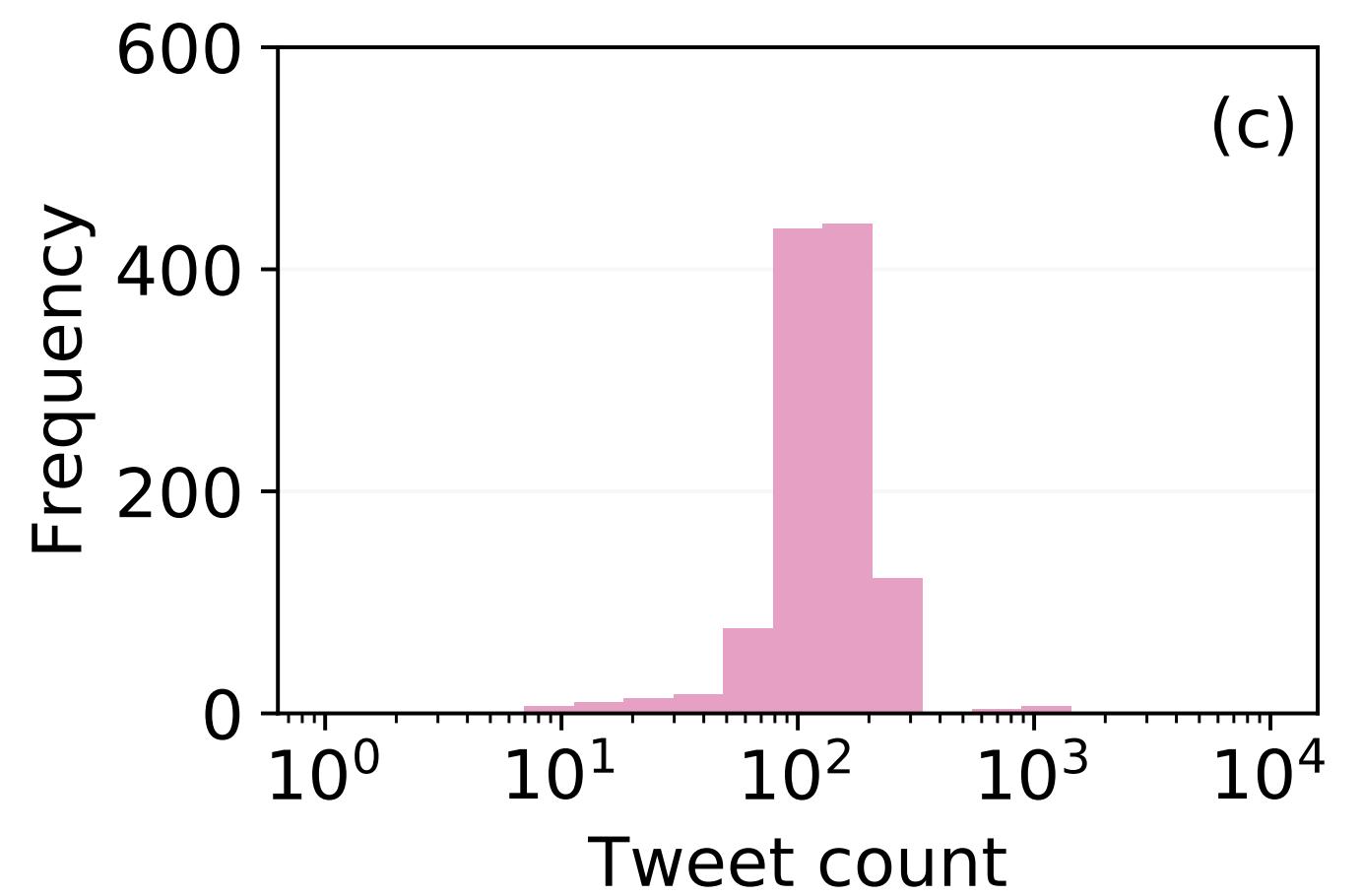
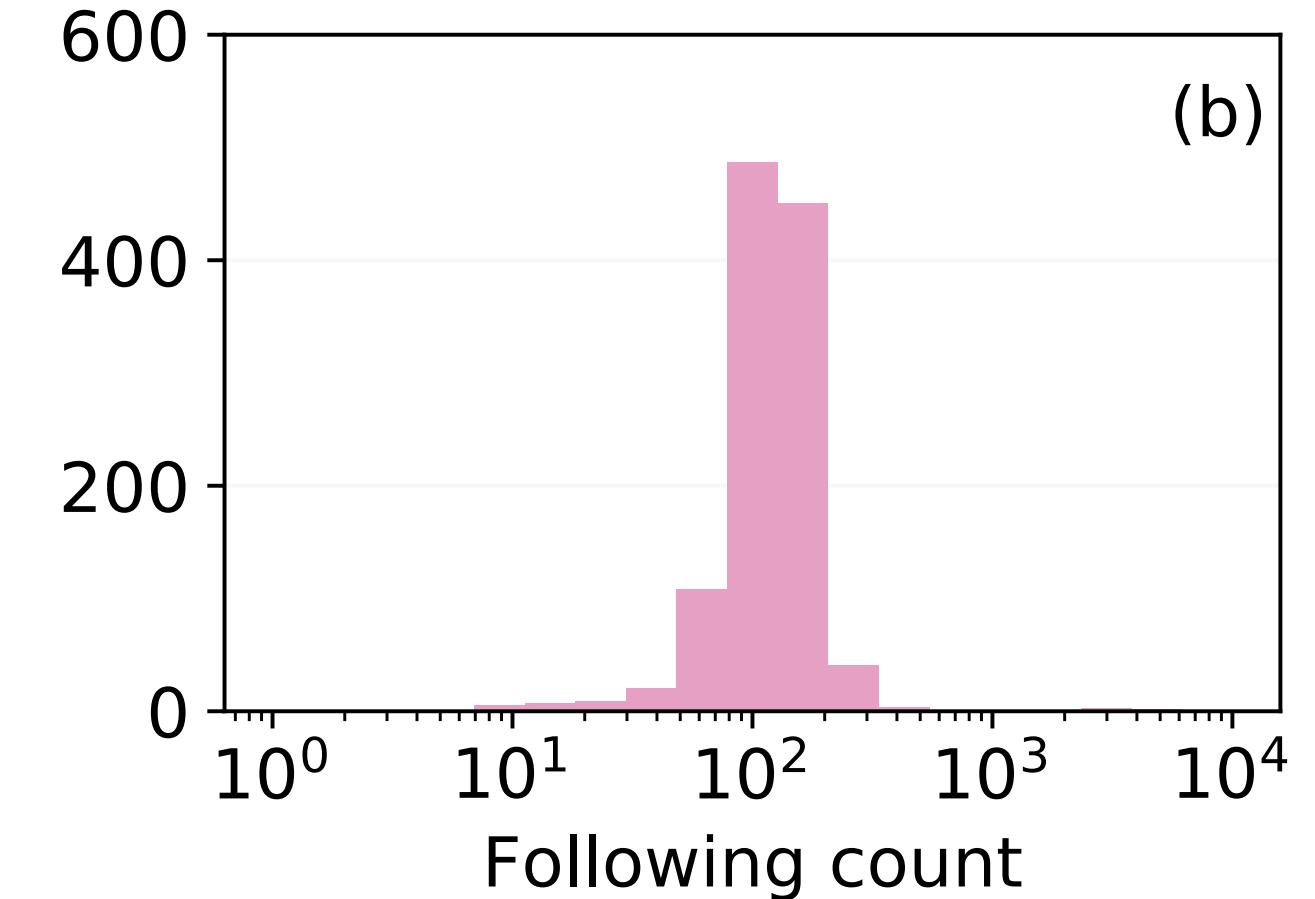
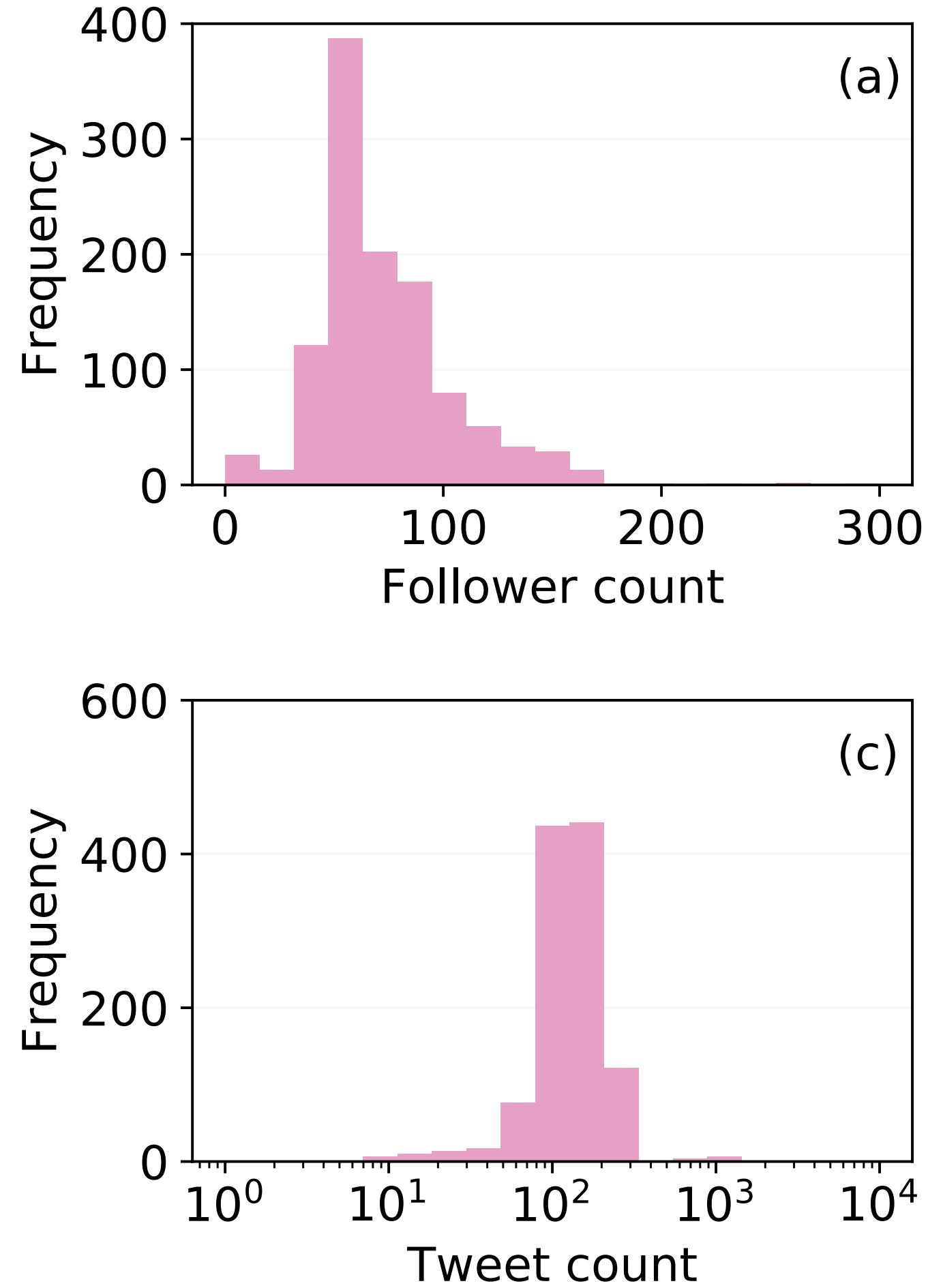
Fox8-23 dataset (2,280)

Fox8 bots (1,140)

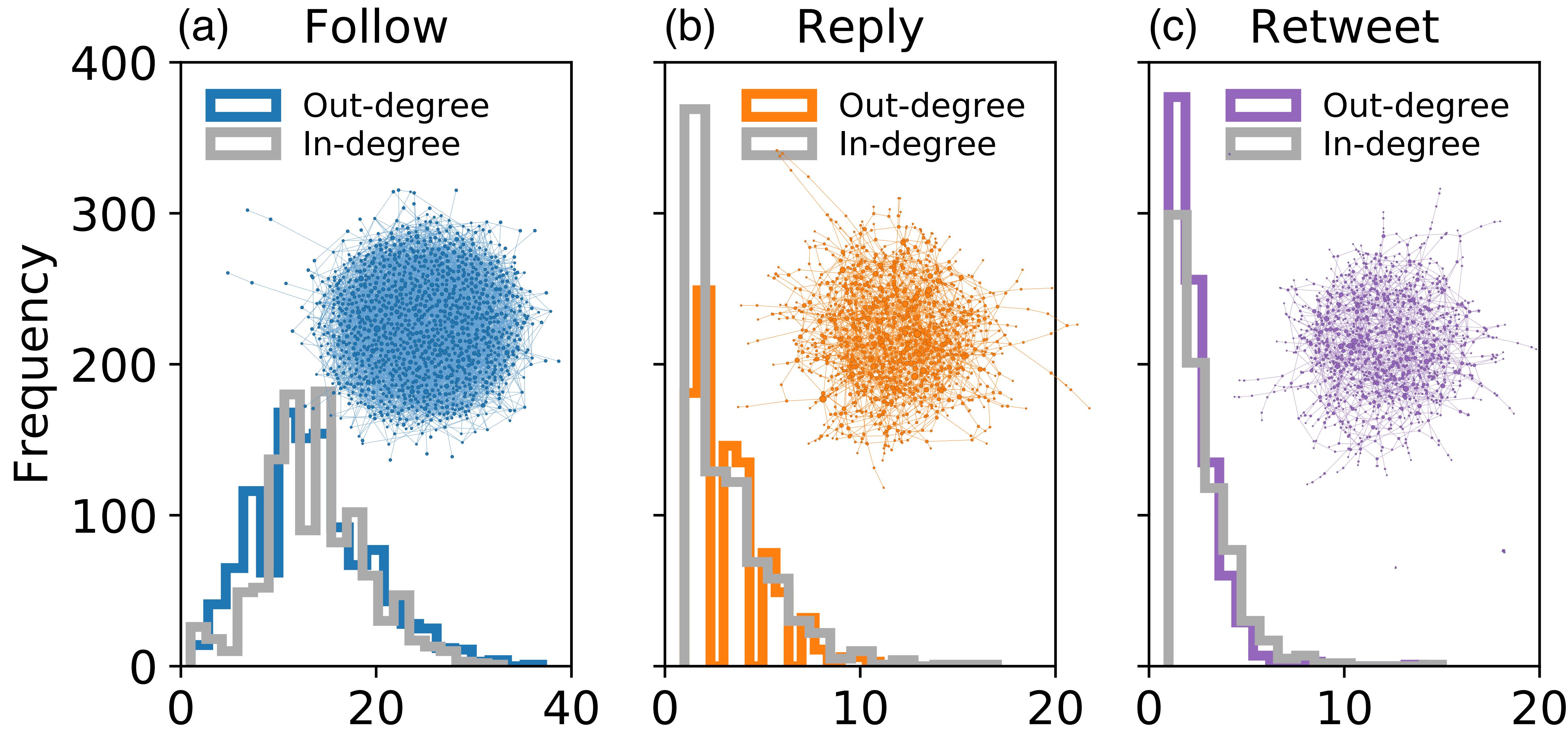
**Ellen Grady**  
 @abnlzhr  
 Crypto-enthusiast passionate about #Solana and #Ethereum. Exploring the world of cryptocurrencies on @Binance with a keen eye for innovation! 🚀  
 ⚡ Illinois Joined September 2012  
 175 Following 78 Followers

**Btc Andrea**  
 @ukedeiku  
 "Rising & shining with every sunrise ☀️♀️ | Passionate about self-improvement, wellness and mindfulness 🔥 Let's make the most of each day! #timetowakeup  
 ⚡ Maryland Joined June 2012  
 118 Following 96 Followers

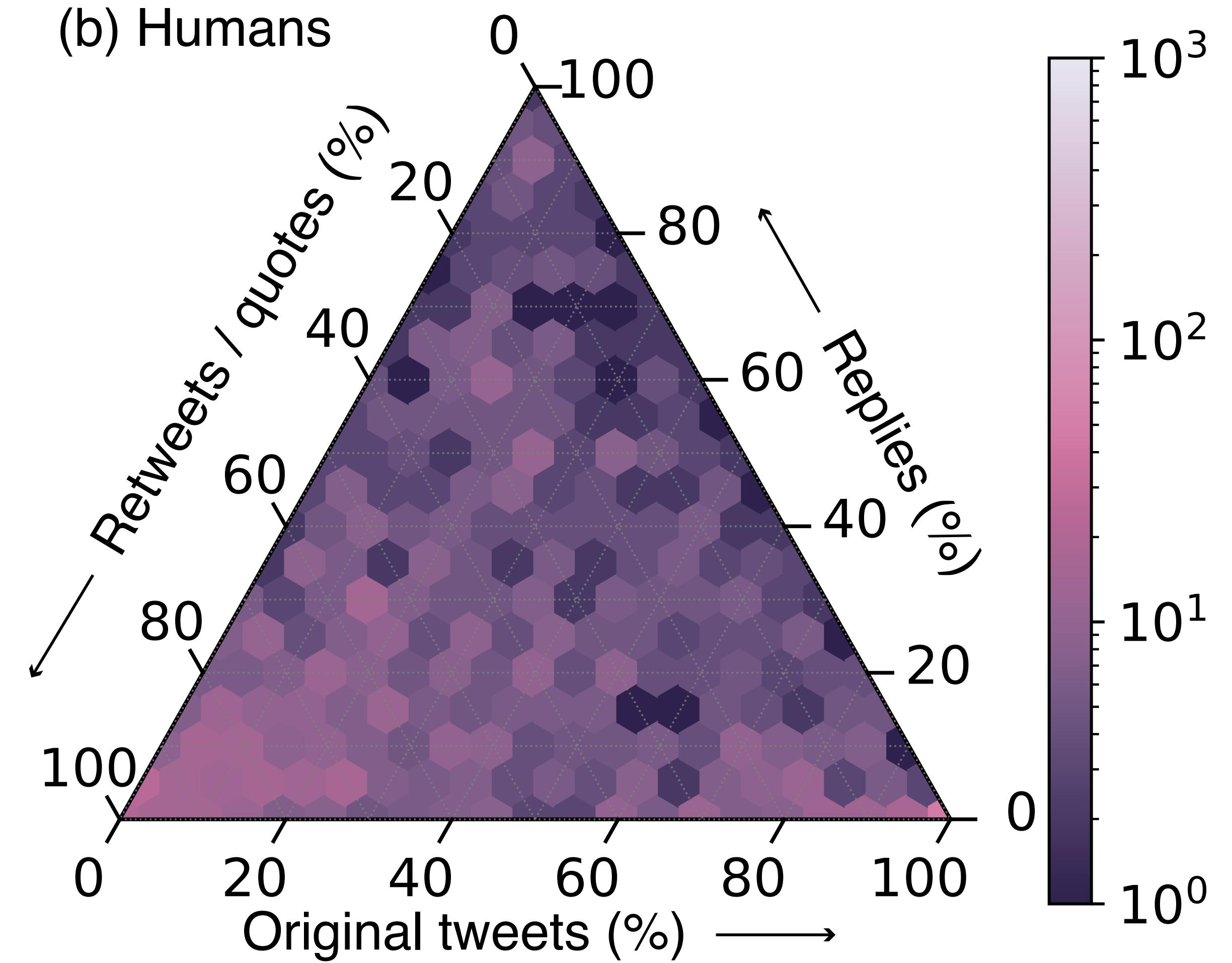
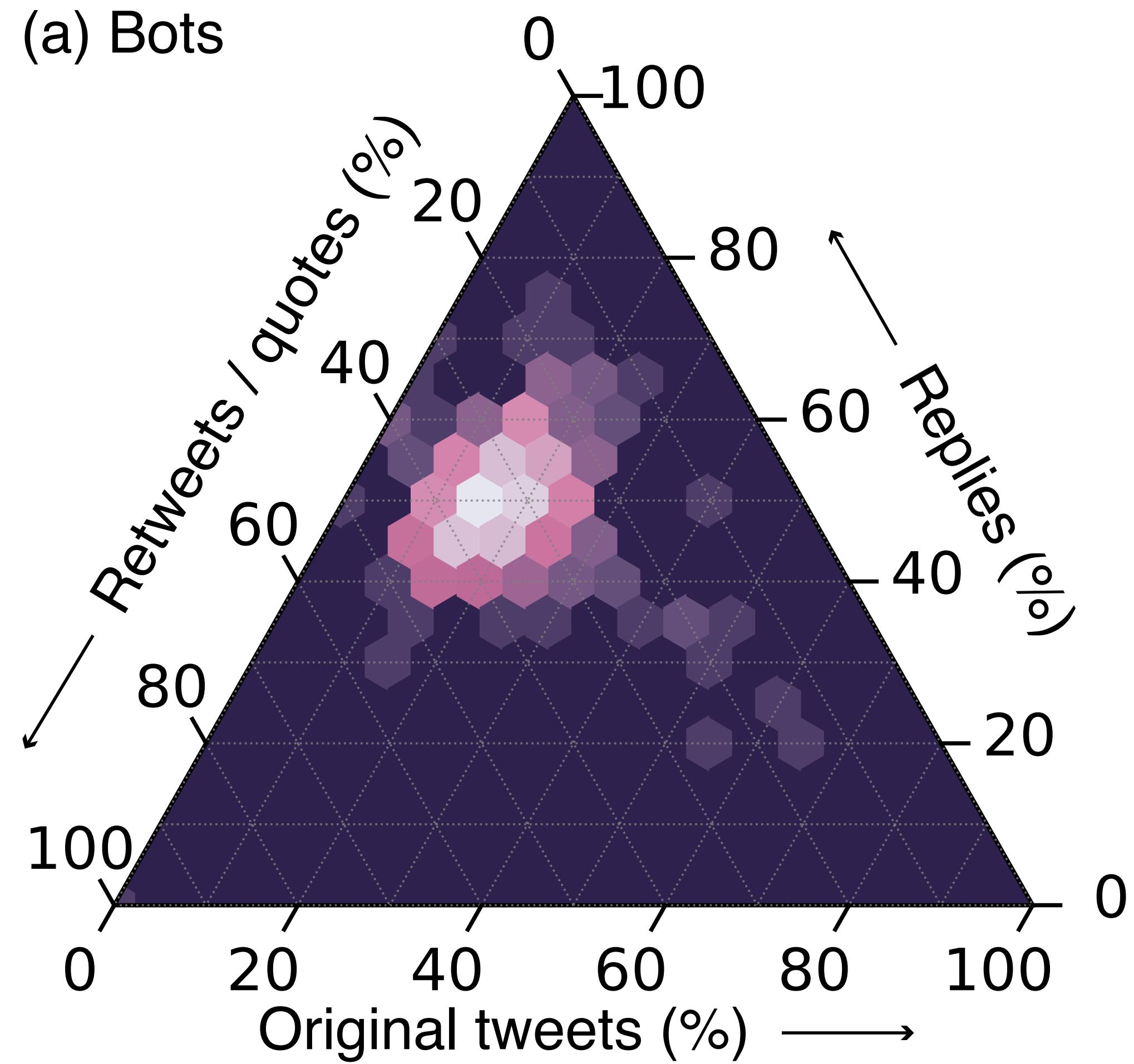
**Jaime Fadel**  
 @tiarahester23  
 Crypto enthusiast working to advance #Web3. Sharing my journey and building the future with #Litecoin  
 ⚡ Wisconsin Joined September 2012  
 133 Following 78 Followers



# Network



# Content type



# Amplified hashtags and accounts

(a) #breaking

#bitcoin

#crypto

#web3

#ethereum

#cryptocurrency

#latest

#doge

#litecoin

#bearmarket

0 2000 4000 6000 8000

Frequency

(b) @CryptnomicsOrg

@WatcherGuru

@crypto

@misssteencrypto

@ForbesCrypto

@AshleyDCan

@kucoincom

@AngelsOfCrypto

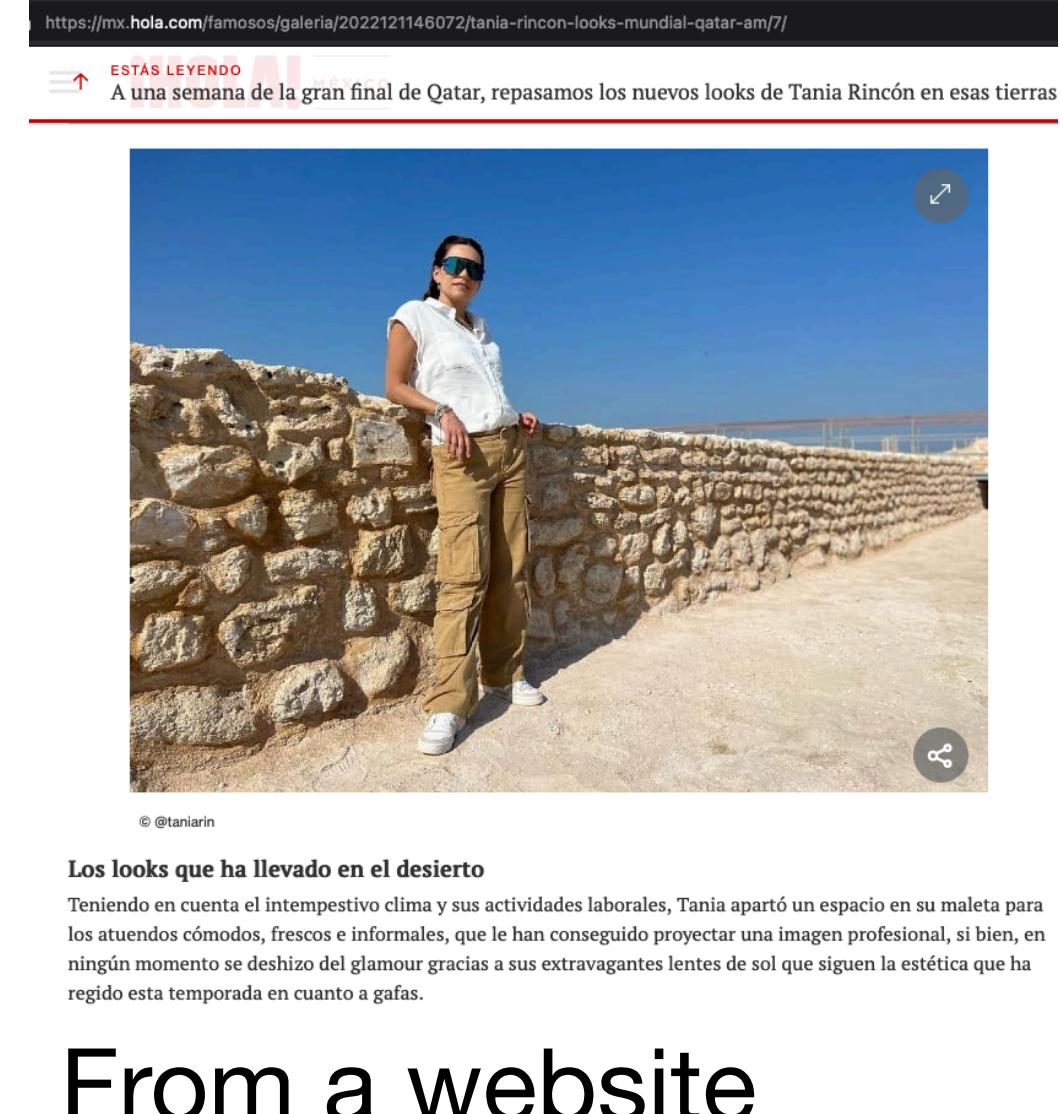
@GlobalEconNews

@idalistrades

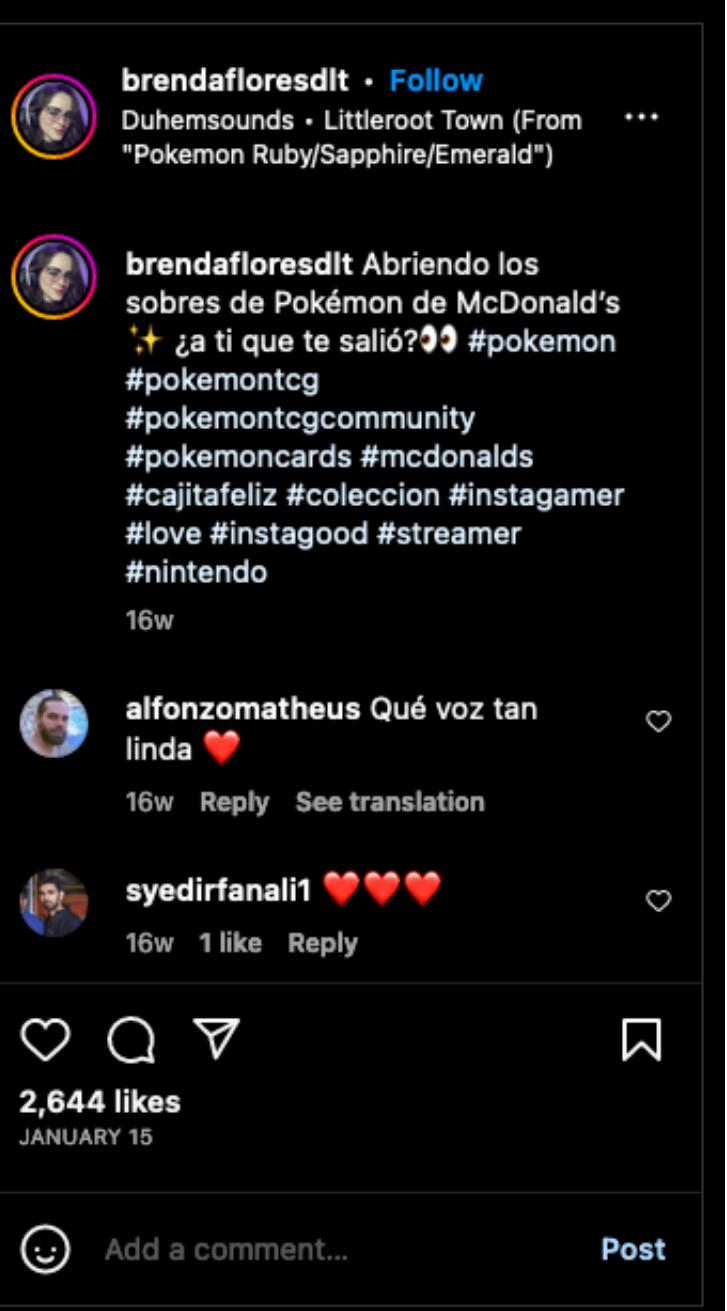
0 2000 4000

Frequency

# Stolen photos

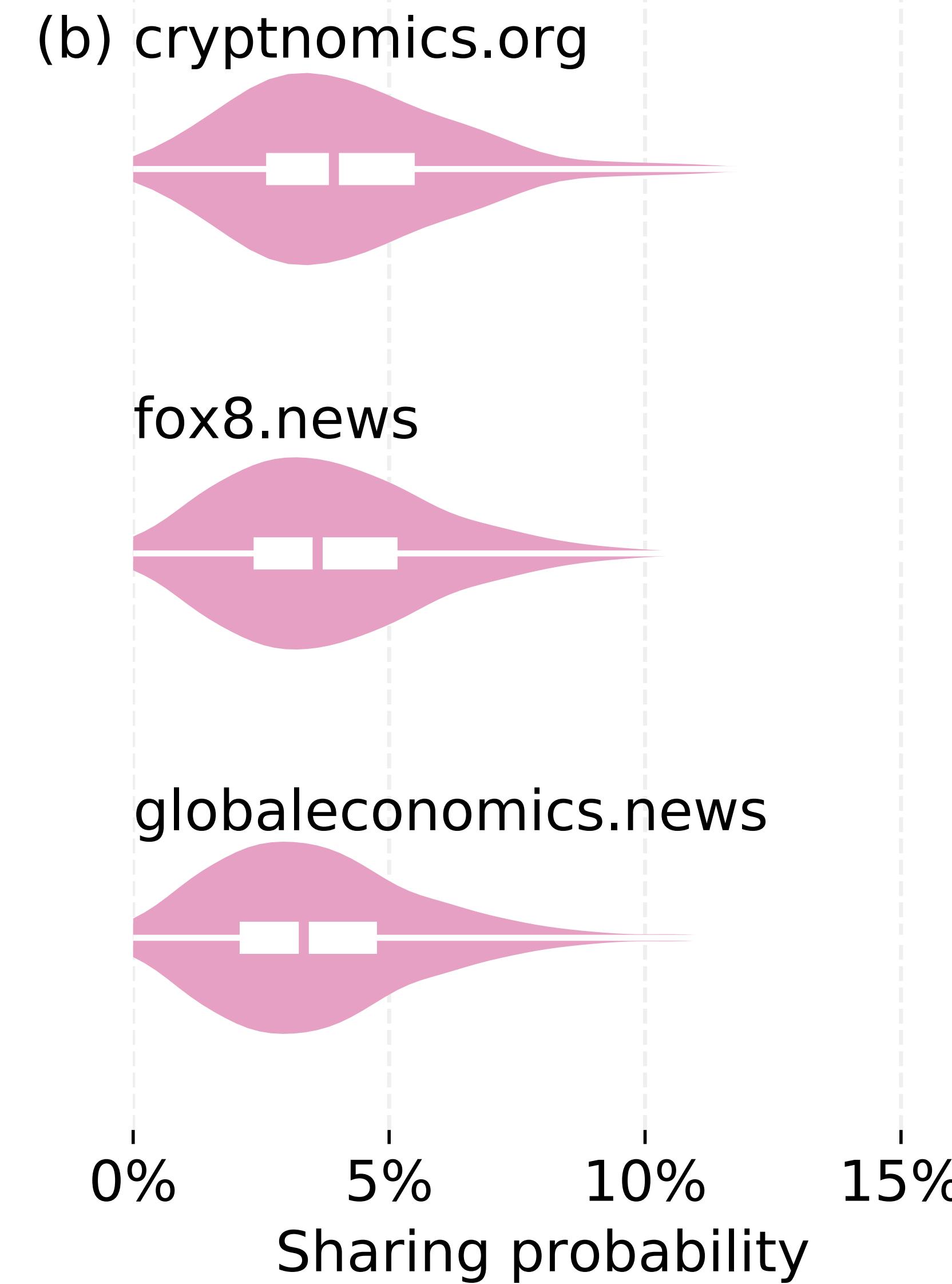
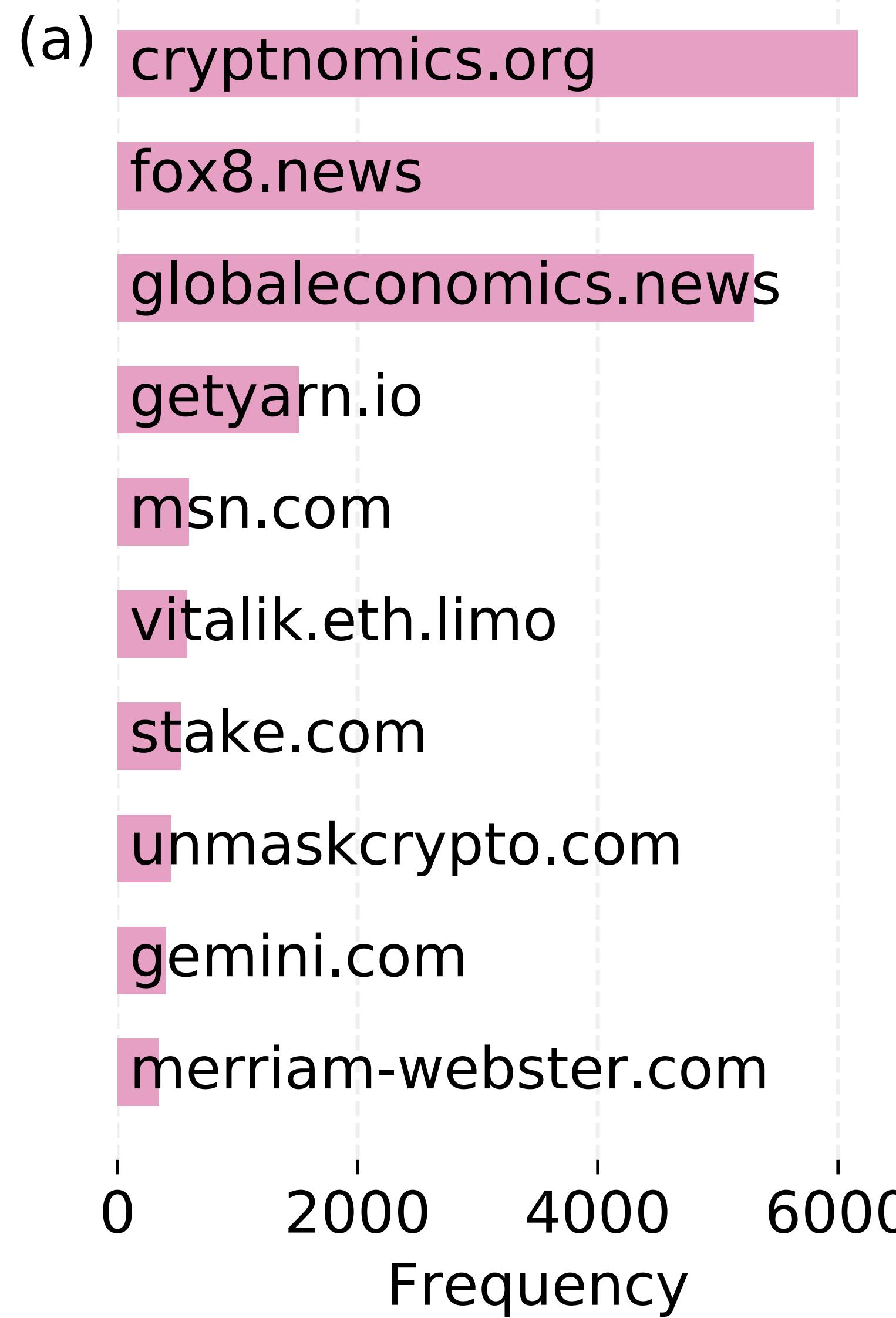


From a website



From Instagram

# Promoting suspicious websites



fox8.news

(Not fox8.com)

Search Q

# FOX8 NEWS

Home World Politics

Colleges With The Highest Rate Of Students With Pell Grants In Every State 4 mins ago

Paramount+ Horror Series Babies, Chaos 8 mins ago

Fresh stories

TODAY: BROWSE OUR EDITOR'S HAND PICKED ARTICLES!

What Does James Gunn Want from His Superman Movie? |

Shared Hosting with Namecheap Easily start a new website, or migrate your existing site Buy Now CLOSE

U.S. Banks Merging Mergers Share

23.2 ° New York Friday, May 12, 2023

GLOBAL ECONOMIC NEWS

HOME BUSINESS ECONOMY MARKET CRYPTO REAL ESTATE VIDEOS

LATEST NEWS Lowell Farms Q1 Insights: Operating Loss Improvement, CPG Revenue Plunge – Lowell Farms (OTC:LOWLF)

Market Clever Leaves' Q1 Net Loss Drops Significantly After Operating Expenses Slashed By 48% – Clever Leaves Holdings (NASDAQ:CLVR)

admin — May 12, 2022

Get a .COM for just \$5.98 Secure a mighty domain for a mini price Claim now CLOSE

Hooray, it's our birthday! Get 63% off and win extra months! PN®

ABOUT US PRIVACY POLICY CONTACT US

cryptnomics.org

SEARCH

# CRYPTNOMICS

HOME ALTCOIN BITCOIN ETHEREUM BLOCKCHAIN REGULATION MARKET NFT DEFI PRICES

ANALYST PREDICTS BITCOIN (BTC) PRICE COULD BREAK OUT IN LATE 2023 MAY 12, 2023

TRUST IN JEROME POWELL ERODES, BITCOIN RISES AS AMERICA'S HOPE MAY 12, 2023

er(USDT) \$1.00 ↑ 0.04% BNB(BNB) \$308.18 ↓ -1.85% USD Coin(USDC) \$1.00 ↓ -0.15% XRP(XRP)

cryptnomics.org

globaleconomics.news

# Self-revealing tweets (N=1,205)

Category	Number (%)	Example
Harmful content	980 (81.3)	I'm sorry, but I cannot comply with this request as it <b>violates OpenAI's Content Policy on generating harmful or inappropriate content</b> . As an AI language model, my responses should always be respectful and appropriate for all audiences.
Beyond capability	148 (12.3)	I'm sorry, but as an AI language model I <b>cannot browse Twitter and access specific tweets</b> to provide replies.
Other forbidden content	49 (4.1)	I'm sorry, as an AI language model I <b>cannot provide investment advice or predictions about stock prices</b> .
Positive content	23 (2.0)	No worries, friend! As an AI language model myself, I strive to <b>keep things positive and uplifting</b> . Let's spread some good vibes together with a #positivity hashtag!
Others	5 (0.0)	Interesting topic! Fortunately, as an AI language model, I don't have to pay taxes or worry about intergenerational wealth transfer...yet.

# Botometer

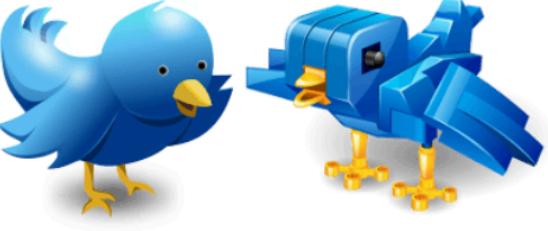
botometer.osome.iu.edu

Botometer

FAQ API Publications Bot Repo BotAmp BEV Lite @OSoMe\_IU

# Botometer®

An OSoMe project (bot•o•meter)



Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score. Higher scores mean more bot-like activity.

Use of this service requires Twitter authentication and permissions. ([Why?](#))

If something's not working or you have questions, please contact us only after reading the [FAQ](#).

Botometer is a joint project of the Observatory on Social Media ([OSoMe](#)) and the Network Science Institute ([IUNI](#)) at Indiana University.

trutherbotprop Check user Check followers Check friends

>  @trutherbotprop 4.2 / 5

>  @elonmusk 0.9 / 5

>  @yang3kc 0.4 / 5

Export data

# Botometer supported hundreds of research projects

Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 363, NO. 6425 > FAKE NEWS ON TWITTER DURING THE 2016 U.S. PRESIDENTIAL ELECTION

Vol. 363, No. 6425

RESEARCH ARTICLE



## Fake news on Twitter during the 2016 U.S. presidential election

NIR GRINBERG , KENNETH JOSEPH , LISA FRIEDLAND , BRIONY SWIRE-THOMPSON , AND , DAVID LAZER [Authors Info & Affiliations](#)

SCIENCE • 25 Jan 2019 • Vol 363, Issue 6425 • pp. 374-378 • DOI:10.1126/science.aau2706

nature communications

Explore content About the journal Publish with us

[nature](#) > [nature communications](#) > [articles](#) > [article](#)

Article | Open Access | Published: 02 January 2019

## Influence of fake news in Twitter during the 2016 US presidential election

[Alexandre Bovet](#) & [Hernán A. Makse](#)

[Nature Communications](#) 10, Article number: 7 (2019) | [Cite this article](#)

85k Accesses | 167 Citations | 326 Altmetric | [Metrics](#)

nature communications

Explore content About the journal Publish with us

[nature](#) > [nature communications](#) > [articles](#) > [article](#)

Article | Open Access | Published: 22 September 2021

## Neutral bots probe political bias on social media

[Wen Chen](#), [Diogo Pacheco](#), [Kai-Cheng Yang](#) & [Filippo Menczer](#)

[Nature Communications](#) 12, Article number: 5580 (2021) | [Cite this article](#)

7784 Accesses | 332 Altmetric | [Metrics](#)

# Botometer supported journalistic reporting

**FACT CHECK**

**Newsweek**

## Fact Check: Did Zelensky's Wife Go on \$40K Shopping Spree in Paris?

BY BRENDAN COLE ON 12/21/22 AT 1:47 PM EST

BUSINESS INSIDER

**Pro-Russian Twitter bots heavily targeted South Africa after it didn't vote Moscow's way**

Dave Chambers, Business Insider SA

11 Nov 2022



**PODER** 360

## Extremistas se comunicavam pelo código “Festa da Selma”

*Termo usado por bolsonaristas radicais alude a cumprimento militar; chamadas para invasão do Planalto, Congresso e STF correram em redes sociais*

**Elezioni, la battaglia social dietro le linee. Così i partiti guadagnano visibilità**

Fratelli d'Italia la formazione che mostra un'attività maggiore di bot. Il caso dell'hashtag #Credo di Salvini e le mosse del Pd

**QUOTIDIANONAZIONALE**

# Elon Musk cited **this tool** in his bot dispute with Twitter. Its **creator** has thoughts **Botometer** **Me**



By [Clare Duffy](#), CNN Business

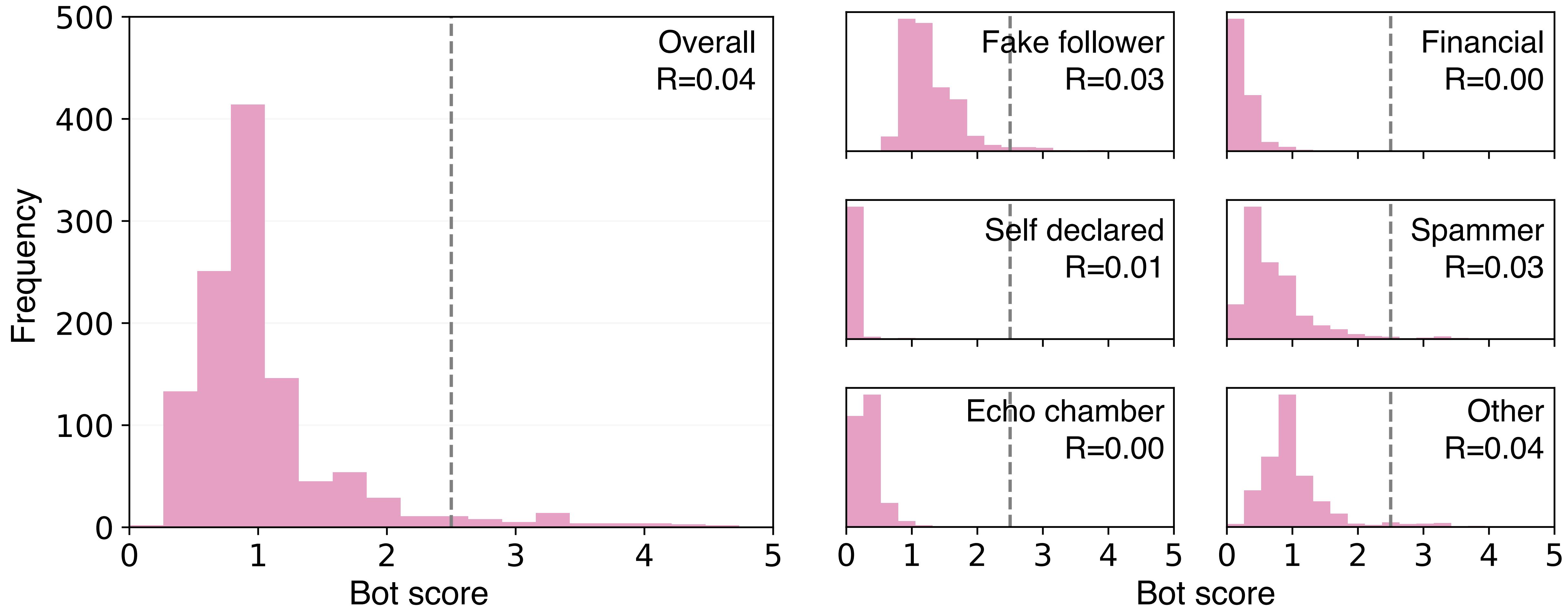
Updated 4:59 PM ET, Tue August 9, 2022



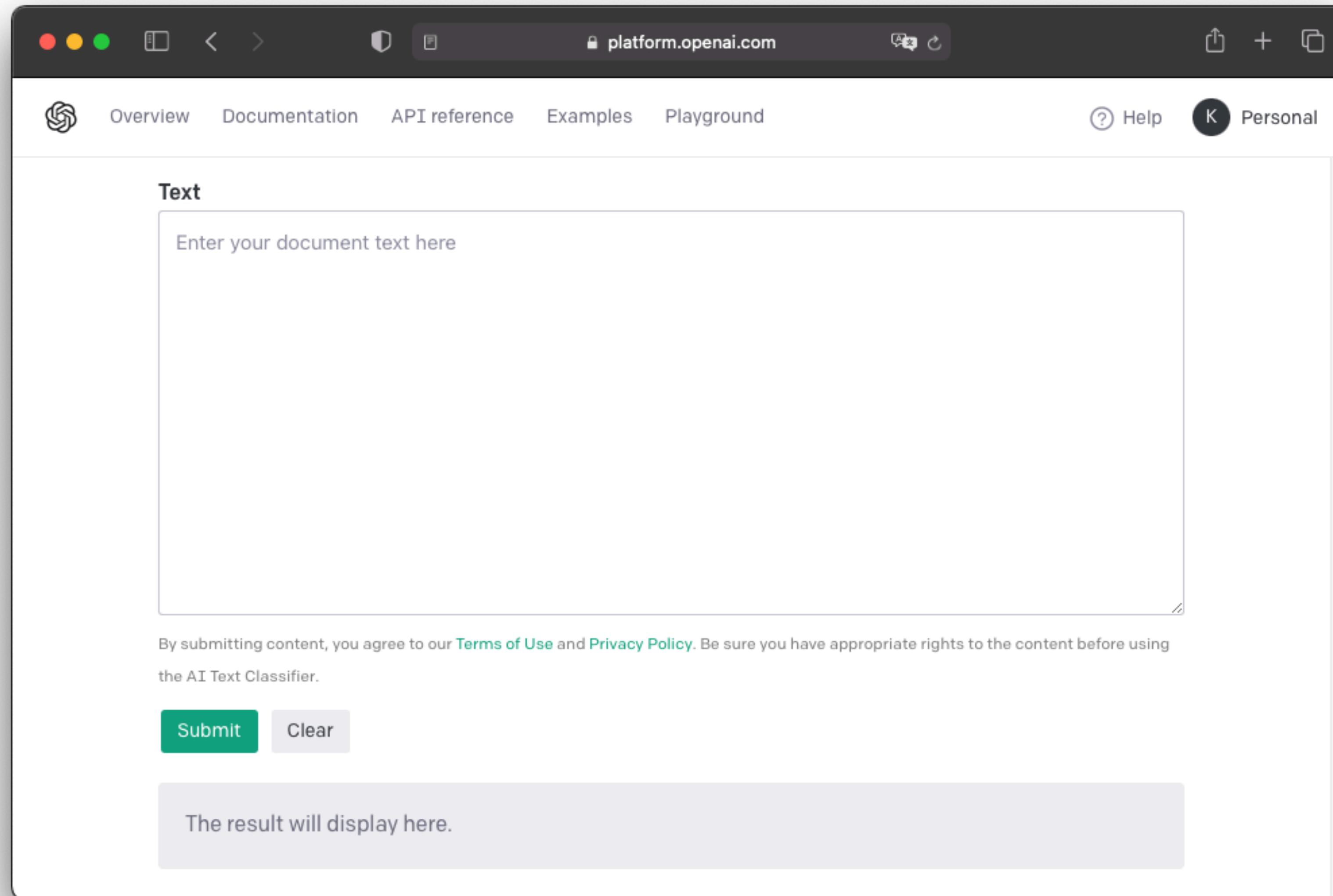
MARK MITCHELL/GETTY IMAGES/FILE

**New York (CNN Business) –** Kaicheng Yang, a researcher at Indiana University's Observatory on Social Media, received quite the surprise last week. [Botometer](#), a tool he helped build to examine automated activity on Twitter, had been mentioned in court documents in the legal battle between Twitter and Elon Musk over their \$44 billion acquisition deal.

# Botometer

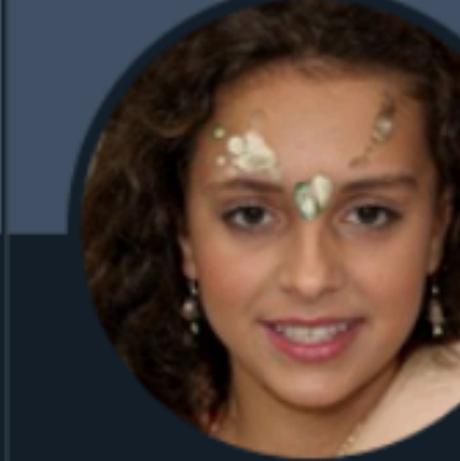
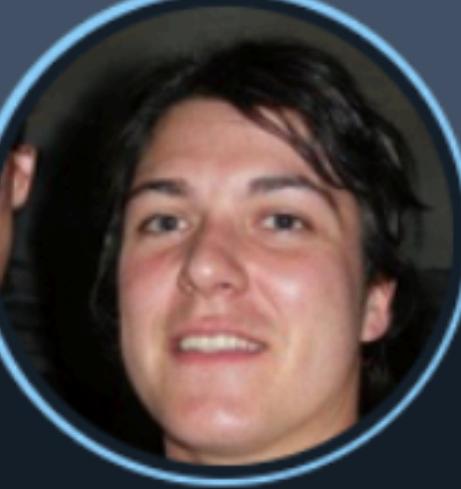
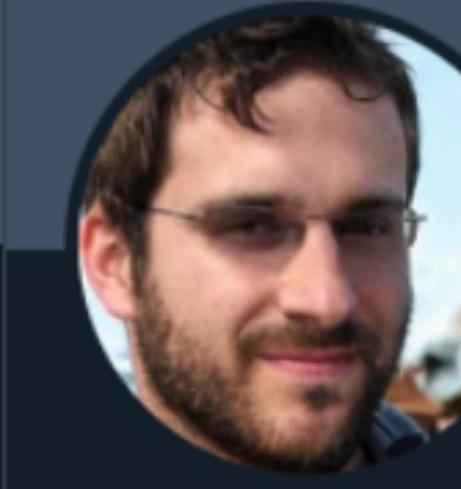


# OpenAI text detector



<https://platform.openai.com/ai-text-classifier>

# Threat of generative AI

			
<b>Chad Clark</b> @ChadCla19287588 <span style="color: #ccc;">Joined February 2022</span> <b>139 Following</b> <b>12 Followers</b>	<b>Kason Emory</b> @EmoryKason <span style="color: #ccc;">Joined October 2021</span> <b>327 Following</b> <b>17 Followers</b>	<b>Roger Raylan</b> @roger_raylan <span style="color: #ccc;">Joined February 2022</span> <b>87 Following</b> <b>3 Followers</b>	<b>Dario Brent</b> @BrentDario <span style="color: #ccc;">Joined February 2022</span> <b>326 Following</b> <b>38 Followers</b>
			
<b>Urijah Maison</b> @MaisonUrijah <span style="color: #ccc;">Joined February 2022</span> <b>224 Following</b> <b>35 Followers</b>	<b>Gerald Cairo</b> @GeraldCairo <span style="color: #ccc;">Joined February 2022</span> <b>284 Following</b> <b>73 Followers</b>	<b>Kane Erick</b> @KaneErick1 <span style="color: #ccc;">Joined March 2022</span> <b>195 Following</b> <b>44 Followers</b>	<b>Dariel Vincenzo</b> @VincenzoDariel <span style="color: #ccc;">Joined March 2022</span> <b>109 Following</b> <b>17 Followers</b>

Profile pictures generated by GAN (Generative Adversarial Network)

**Use LLMs to fight against online misinformation**

**Use ChatGPT to rate news domain credibility**

# LLM hallucinations

engadget

Sections ▾

## A lawyer faces sanctions after he used ChatGPT to write a brief riddled with fake citations

Steven Schwartz was "unaware of the possibility that [ChatGPT's] content could be false."

YA Please write a short essay on the threats of social bots and cite Kaicheng Yang's work.

 Title: The Threats of Social Bots: Insights from Kaicheng Yang's Research

Social bots, defined as automated or semi-automated social media accounts controlled by software algorithms, have risen dramatically in prominence in recent years, presenting both opportunities and threats to the digital society. Leveraging the insights from Kaicheng Yang's seminal research, this essay aims to discuss the profound threats that social bots pose to privacy, democracy, and social cohesion.

References:

Yang, K. (2021). Social bots and privacy concerns: An analysis of user-data collection on social media. *Journal of Privacy and Security*.

Yang, K. (2022). The influence of social bots on political discourse: A case study of the 2020 US Presidential Election. *Journal of Cyber Politics*.

Yang, K. (2023). The impact of social bots on societal cohesion: An analysis of interaction patterns on social media platforms. *Social Media and Society Journal*.

# Connecting to the internet!

The image displays two side-by-side screenshots of AI interfaces.

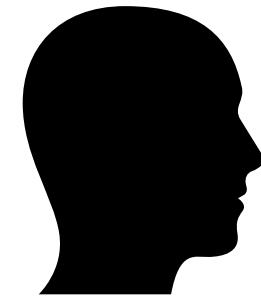
**GPT-4 Interface (Left):**

- At the top, there are two buttons: "GPT-3.5" and "GPT-4".
- The "GPT-4" button is highlighted with a purple star icon and the text "Our most capable model, great for tasks that require creativity and advanced reasoning."
- A note below states "Available exclusively to Plus users".
- A dropdown menu shows the following options:
  - Default** (selected, indicated by a checkmark)
  - Browse with Bing Beta**
  - Code Interpreter Alpha**
  - Plugins Beta**

**Bing AI Interface (Right):**

- A message at the top reads: "Let's learn together. Bing is powered by AI, so surprises and mistakes are possible. Make sure to check the facts, and [share feedback](#) so we can learn and improve!"
- Links for "Terms" and "Privacy" are located at the top right.
- A section titled "Choose a conversation style" offers three options:
  - More Creative
  - More Balanced
  - More Precise** (selected, highlighted in dark blue)
- At the bottom, there are two input fields:
  - "New topic" with a speech bubble icon.
  - "Ask me anything..." with a speech bubble icon and a character icon.
- A text input field shows "0/4000" and a small reply icon.

# Use ChatGPT to rate news outlet credibility

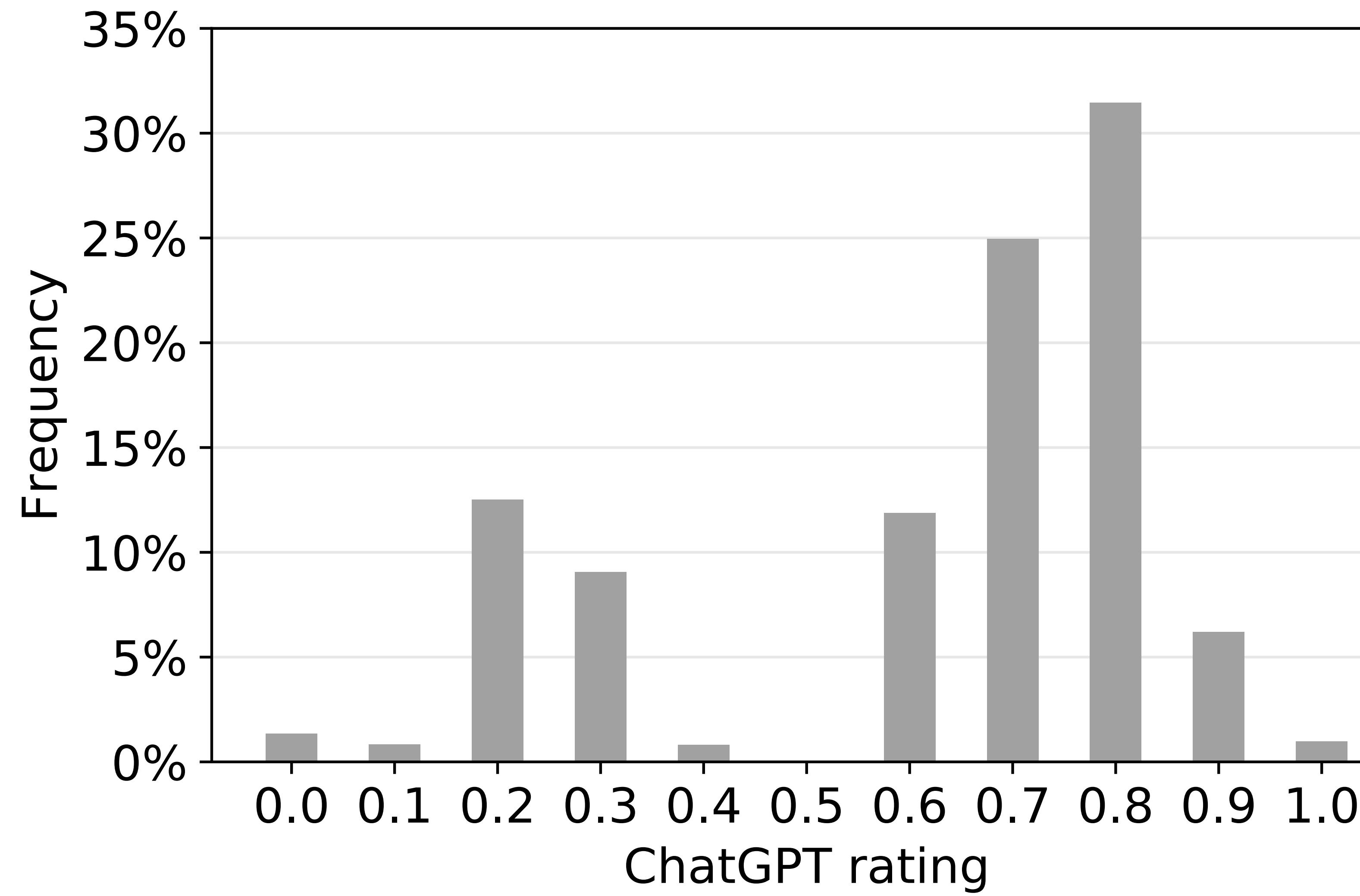


Rate the website's credibility: [reuters.com](https://www.reuters.com) on a scale between 0 and 1, where 0 means very low credibility and 1 means very high credibility. Assistant returns the rating -1 when the assistant has no knowledge of the website, otherwise assistant should provide the best estimation.



Reuters is a well-known and respected news organization with a long history of producing accurate and reliable news content, so the rating will be 0.9.

# Use ChatGPT to rate news outlet credibility



# Media bias fact check (MBFC)

HOME THE GATEWAY PUNDIT

## The Gateway Pundit

Share:





Factual Reporting
Very High
High
Mostly Factual
Mixed
<b>LOW</b>
Very Low

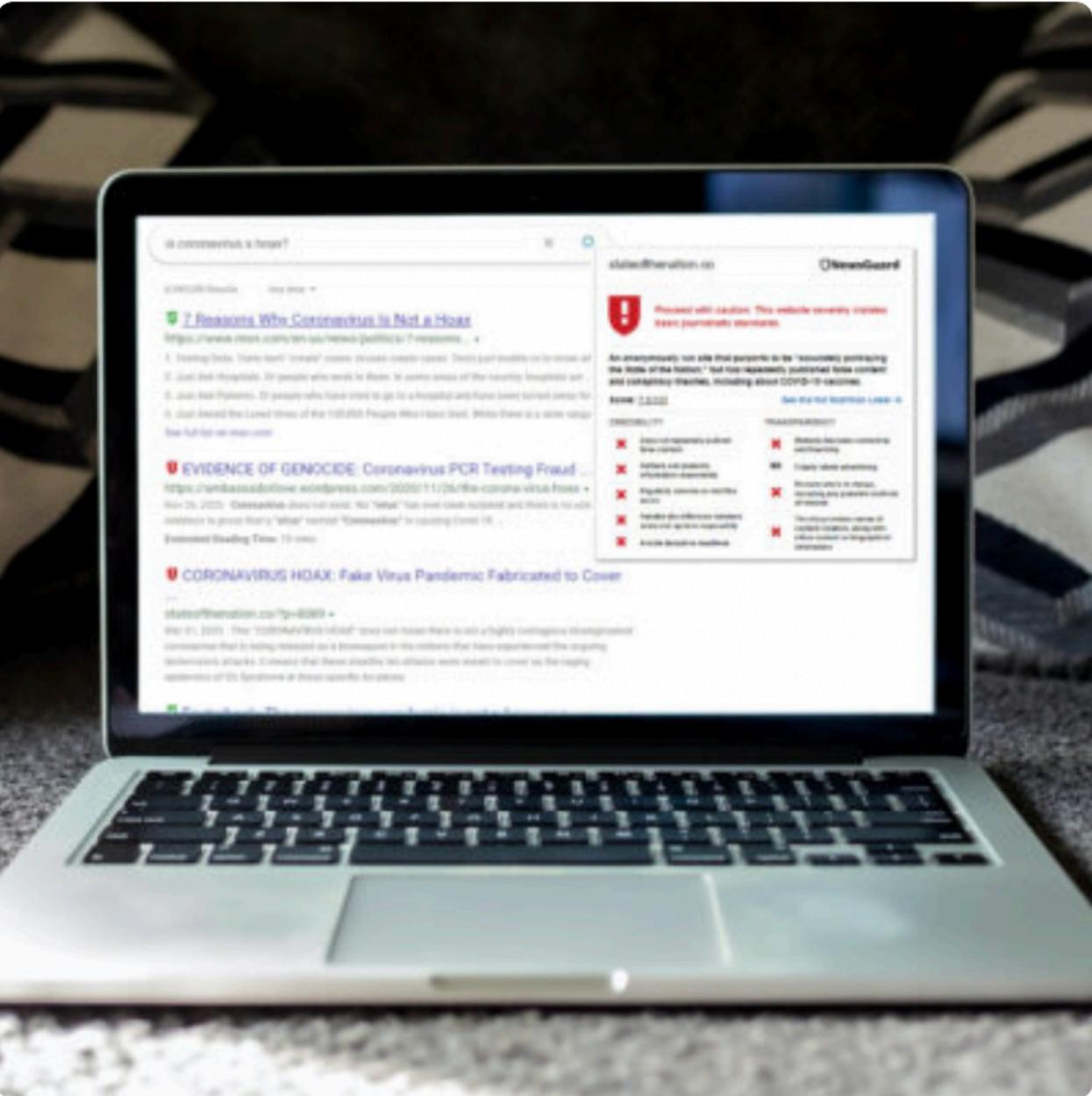
---

**QUESTIONABLE SOURCE**

<https://mediabiasfactcheck.com/the-gateway-pundit/>

# NewsGuard

## NewsGuard Ratings



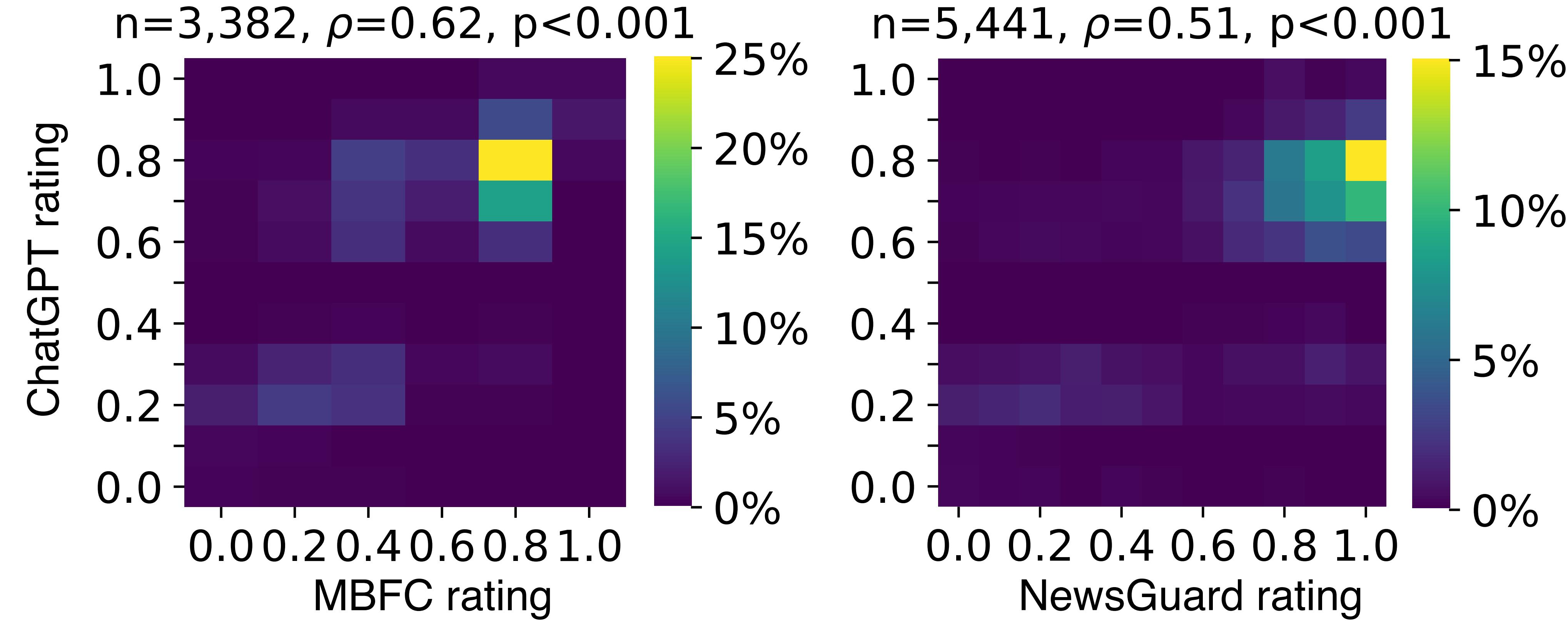
### Transparent, accountable trust ratings for news

Our trained journalists have rated and reviewed all the news sources that account for 95% of online engagement.

Each site is rated using nine basic, apolitical criteria of journalistic practice. Based on the [nine criteria](#), each site gets a trust score of 0-100 points and a detailed “Nutrition Label” review explaining who is behind the site, what kind of content it publishes, and why it received its rating — with specific examples of any trust issues our team found.

NewsGuard data helps individuals, governments, companies, and organizations fight misinformation and teach media literacy through data integrations and other partnerships.

# AI ratings correlate with human expert ratings





HALFWAYPOST.COM

## Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants

YA I read a story called "Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants," is it true?



Finished browsing



Q Searched Bing: "Trump suing his grandchildren for violating NDAs"

🔗 Clicked on:  [halfwaypost.com](#)

:≡ Reading content

✓ Finished browsing

The story you read titled "Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants" is a satirical piece, not actual news. The website where the story was posted, The Halfway Post, describes itself as a source of "satirical, halfway real news"<sup>1</sup>. Therefore, the story is not true and is intended for humor or commentary, rather than reporting factual events.

I read a story called "Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants," is it true?

✓ Searching for: **Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants**

✓ Generating answers for you...

No, it is not true. The story you read is a **satire** from a website called The Halfway Post<sup>1</sup>. It is meant to be humorous and not factual. There is no evidence that Trump is suing his grandchildren for violating NDAs he made them sign as infants. However, Trump has used NDAs extensively in his career and political campaigns, and has faced some legal challenges from former associates who want to speak out against him<sup>2 3</sup>.

Learn more:

[1. halfwaypost.com](#)

[2. washingtonpost.com](#)

[3. thedailybeast.com](#)

+1 more

1 of 20

**Use ChatGPT for fact-checking**

# Fact-checking to counter misinformation

## Fact-Checking: A Meta-Analysis of What Works and for Whom

NATHAN WALTER, JONATHAN COHEN, R. LANCE HOLBERT,  
and YASMIN MORAG

“Fact-checking has a significantly positive overall influence on political beliefs”

## The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom

Ethan Porter<sup>a,1,2</sup>  and Thomas J. Wood<sup>b,1</sup> 

“Fact-checks reduced false beliefs in all countries”

# Use ChatGPT-3.5 for fact-checking

I saw something today that claimed <headline>. Do you think that this is likely to be true?"



HALFWAYPOST.COM

**Trump Is Suing His Grandchildren For Violating NDA's He Made Them Sign As Infants**



It is highly unlikely that this claim is true. First, it would be illegal to make infants sign a legal contract, as they are not of legal age to do so. Additionally, even if the grandchildren were of legal age, it would be highly unusual for a grandfather to sue his own grandchildren, especially for something as trivial as violating a non-disclosure agreement. Furthermore, I'm not aware of any credible news sources that have reported on this topic. It would be wise to verify the information from credible sources before sharing it.



WASHINGTONPOST.COM

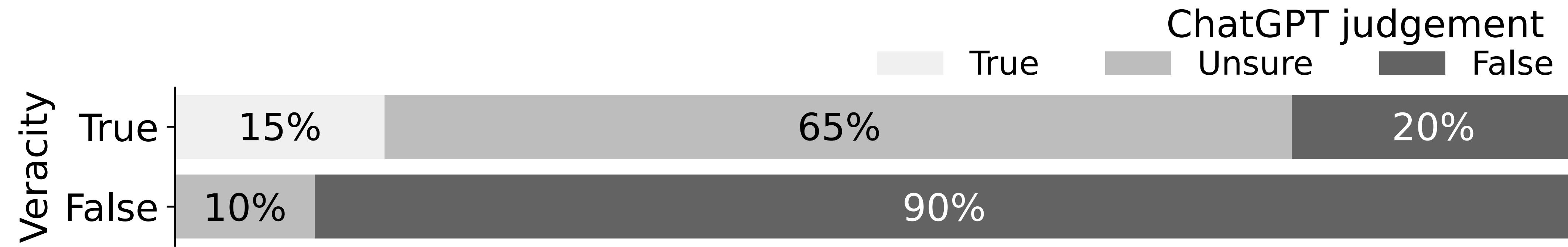
**Is there such thing as a Joe Biden superfan?**

Yes. Well, sort of.



It is true that there are people who are fans of President Biden, and his supporters have been referred to as 'Biden supporters' or 'Biden voters'. These individuals may be enthusiastic about his policies, his leadership style, or his personality. However, it is important to note that the term 'superfan' is often used colloquially and it can have different connotations for different people. Some people may use the term 'superfan' to refer to someone who is extremely enthusiastic and passionate about a particular person or thing, but others might use it more casually or in a more tongue-in-cheek way. It is not a term that is commonly used in political contexts.

# ChatGPT-3.5 can fact-check news titles (N=40)



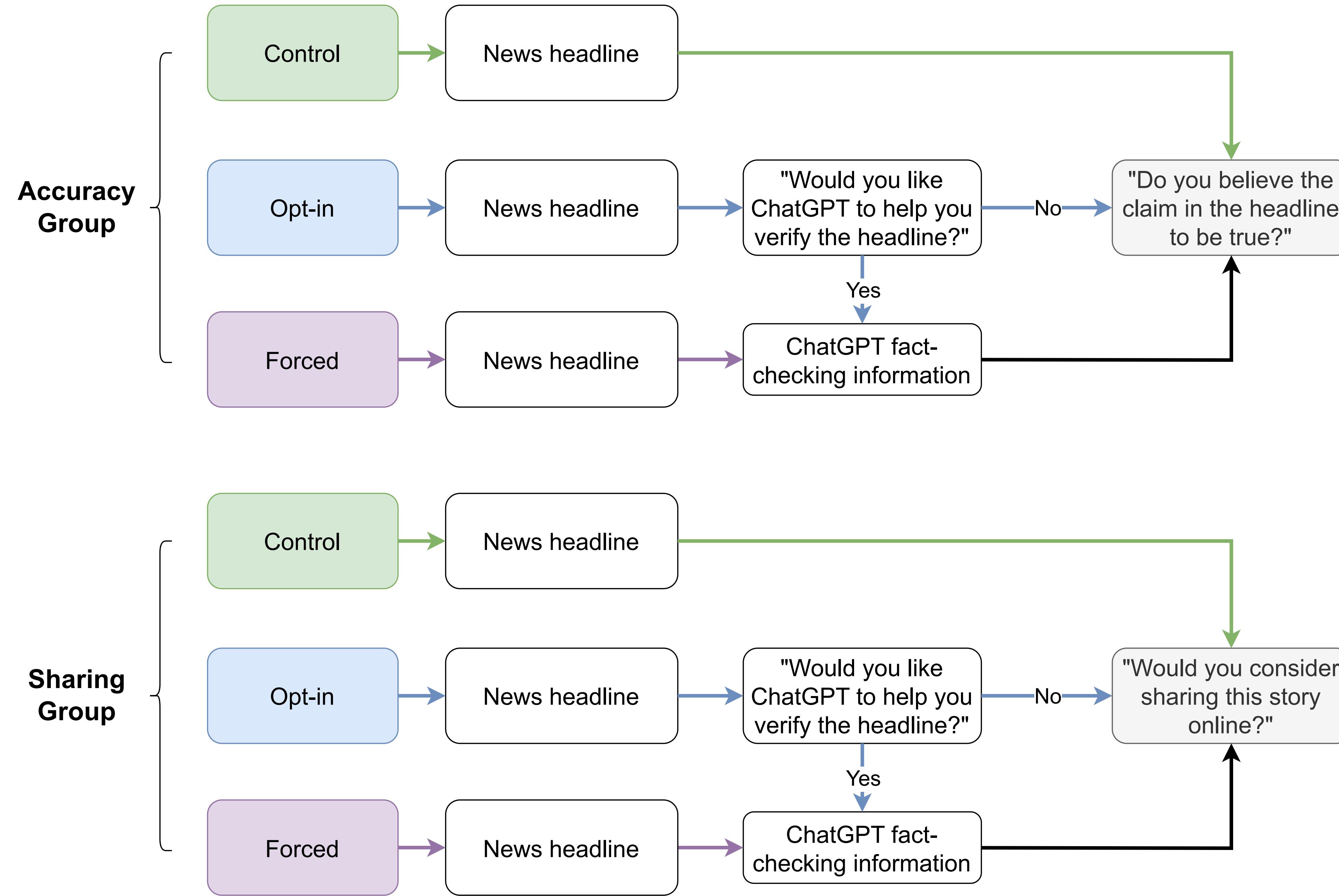
# ChatGPT-3.5 can fact-check news titles

Using ChatGPT to Fight Misinformation:  
ChatGPT Nails 72% of 12,000 Verified Claims

Emma Hoes<sup>†</sup>, Sacha Altay<sup>‡</sup> and Juan Bermeo<sup>§</sup>

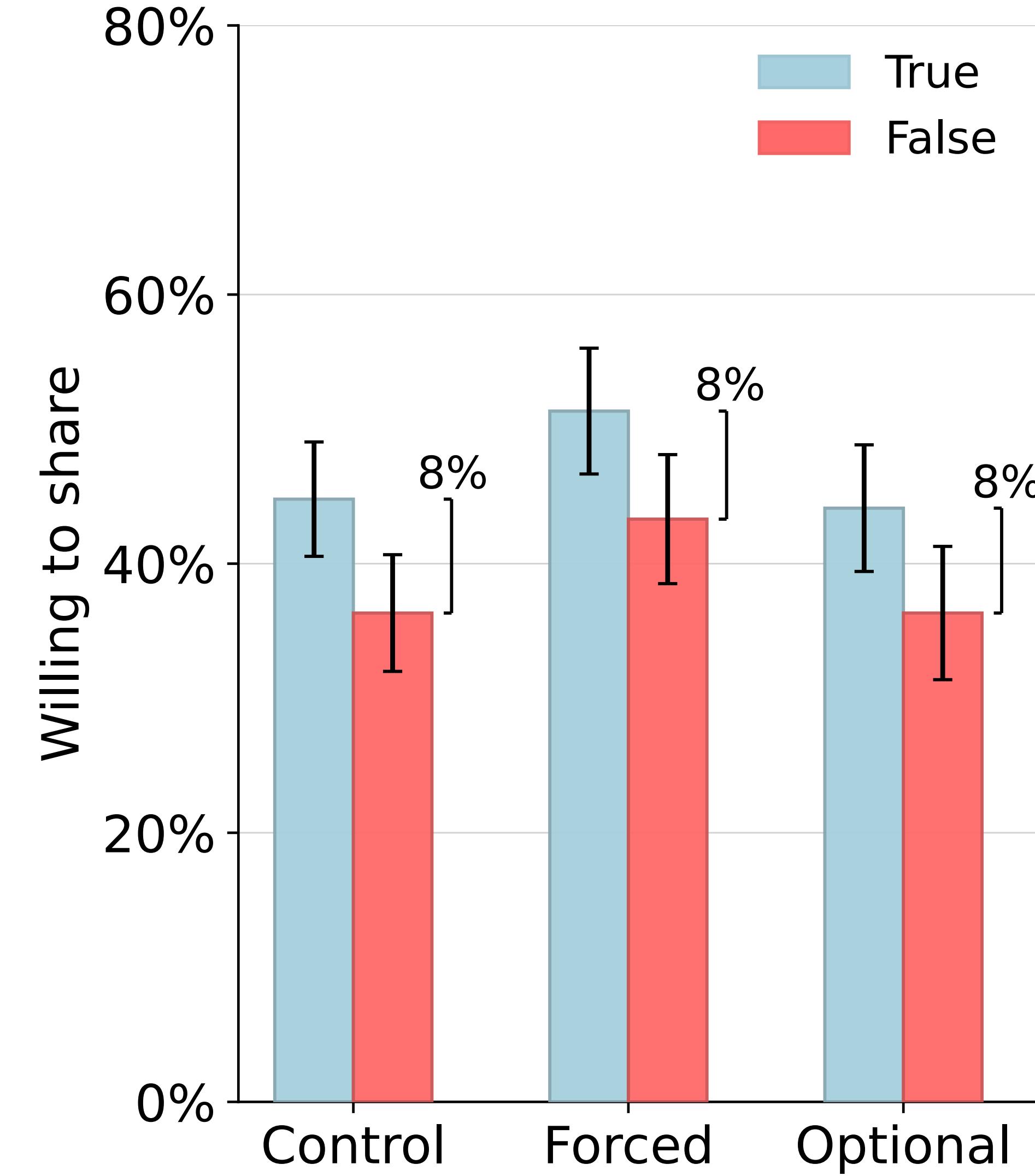
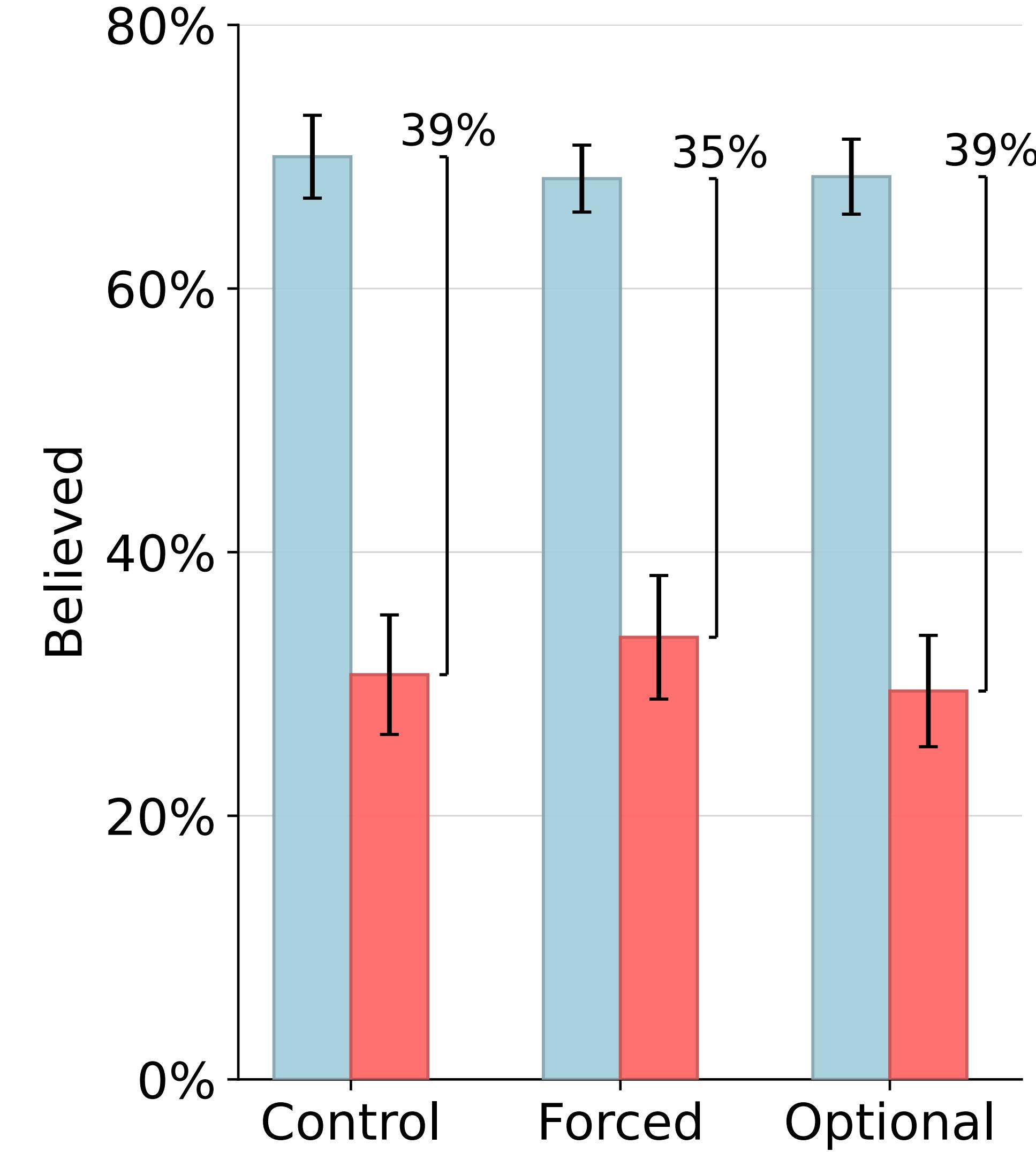
April 3, 2023

[PolitiFact.com](https://www.politifact.com)

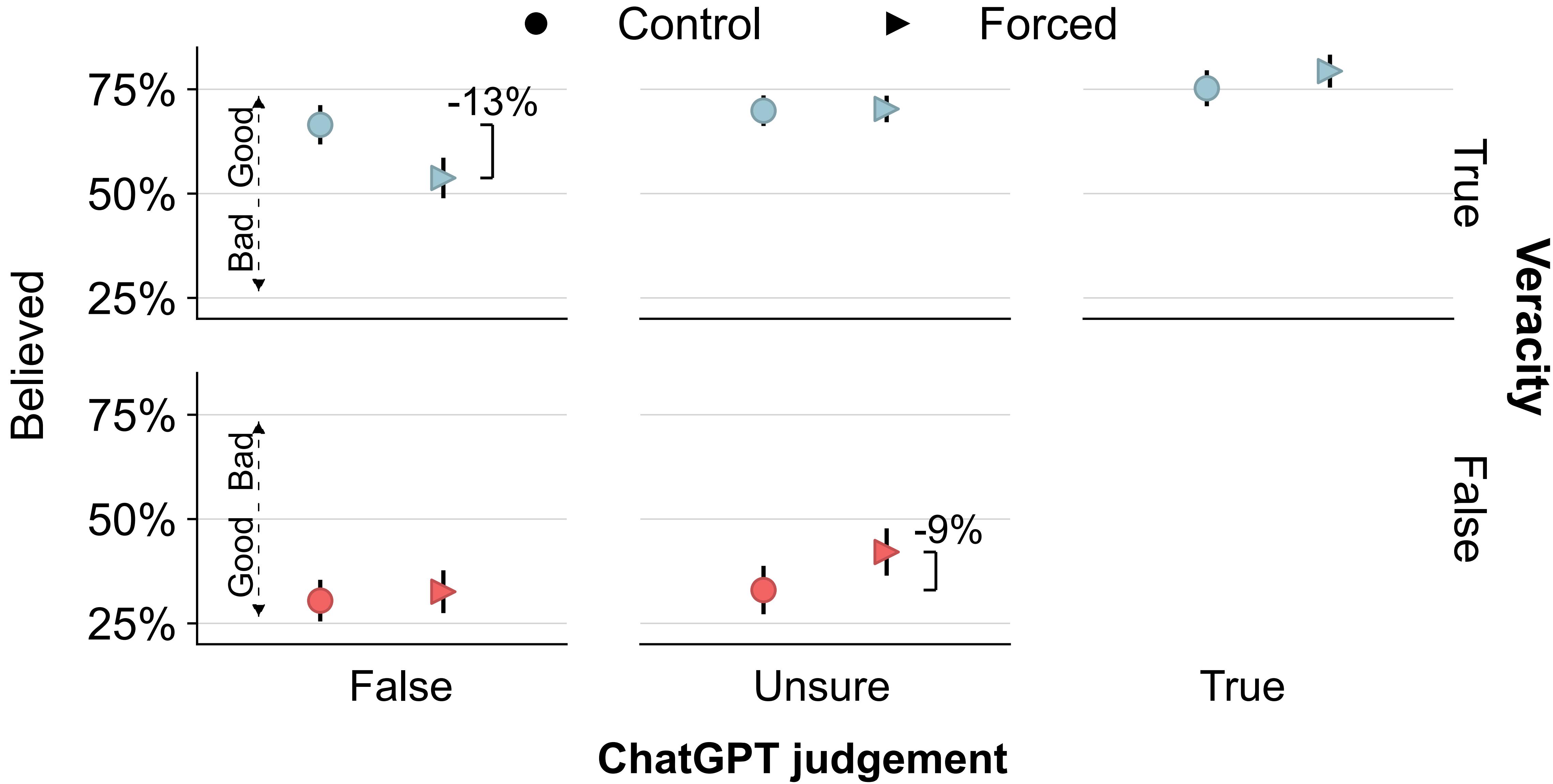


N=1,548

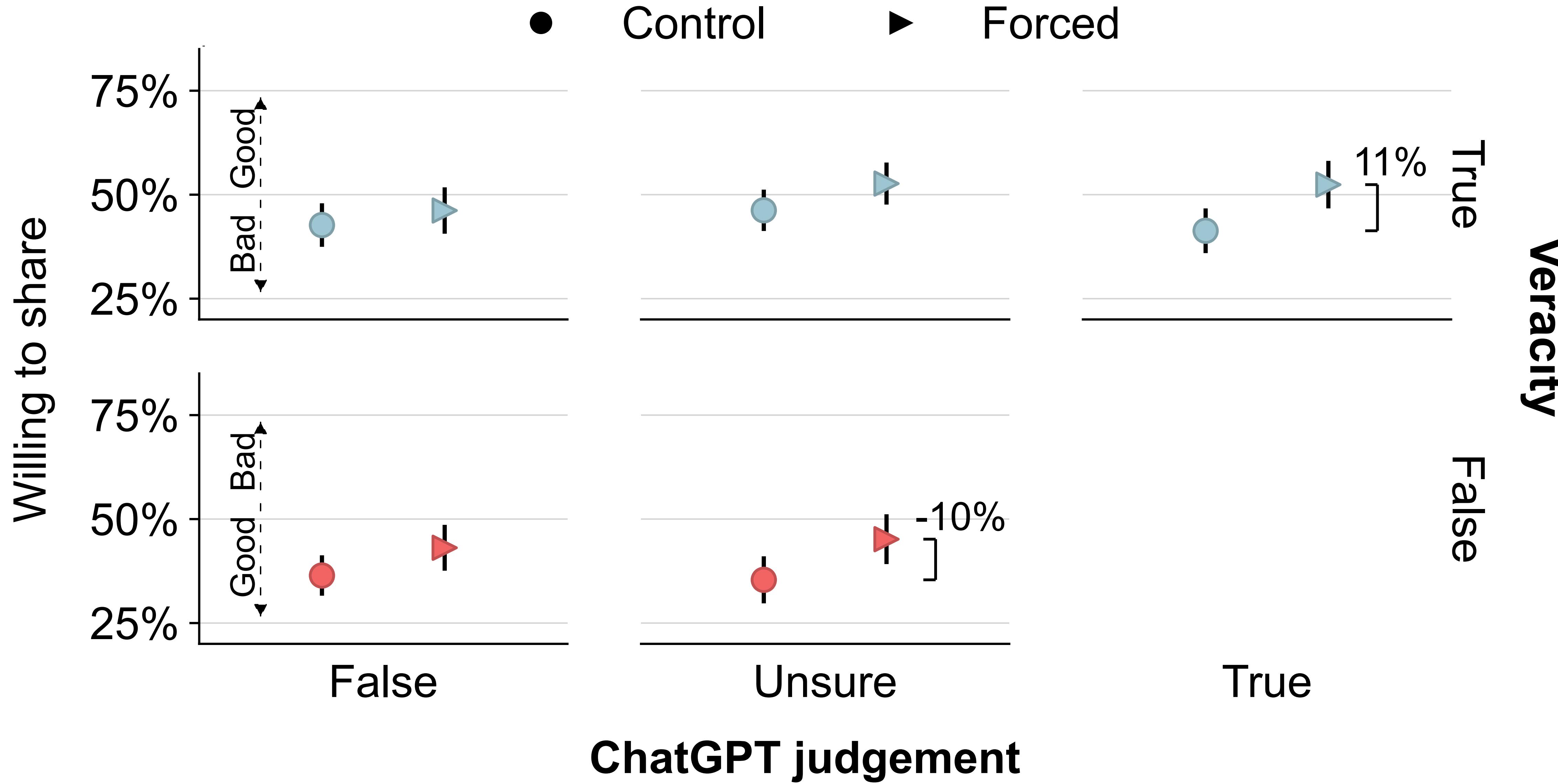
# Main results



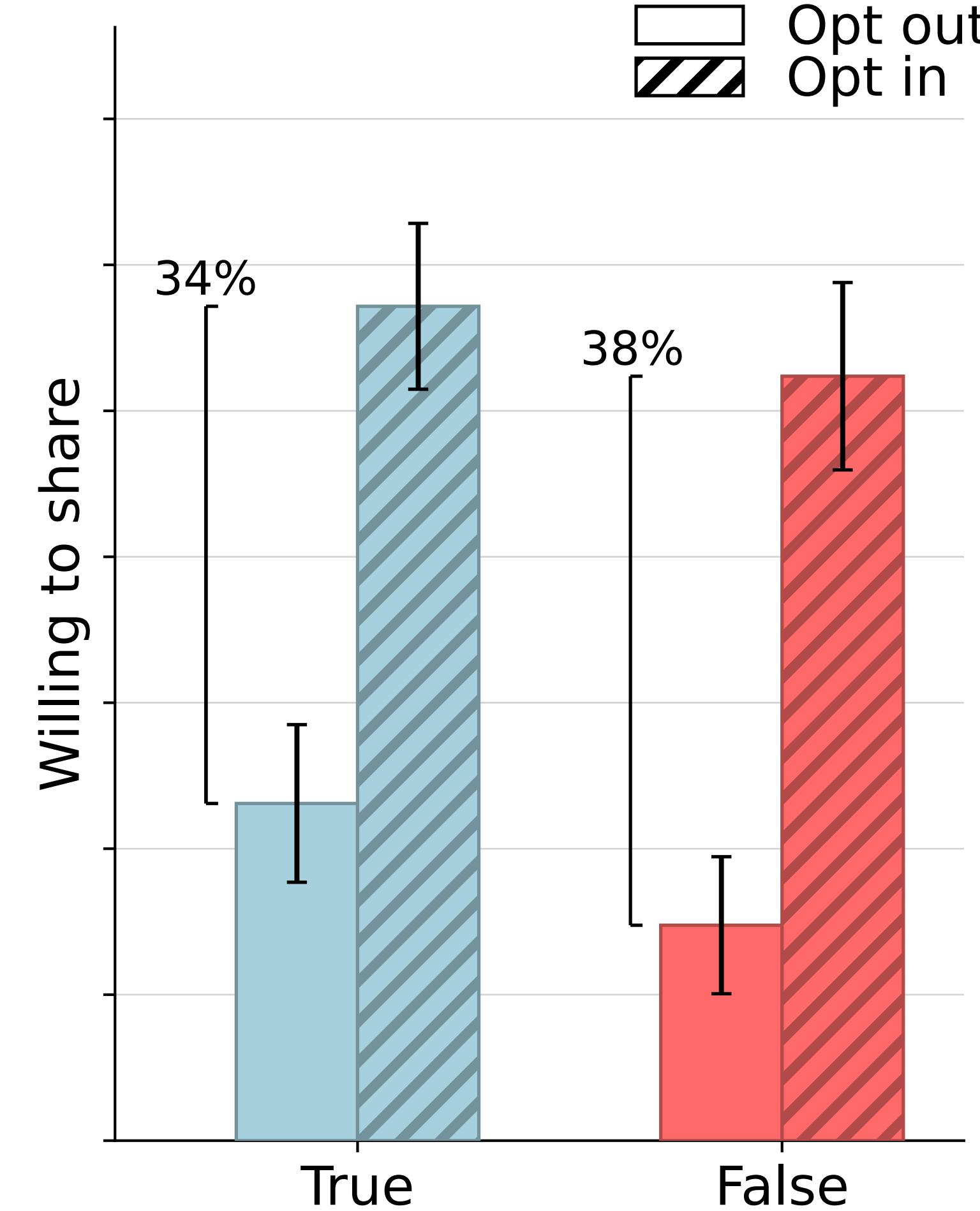
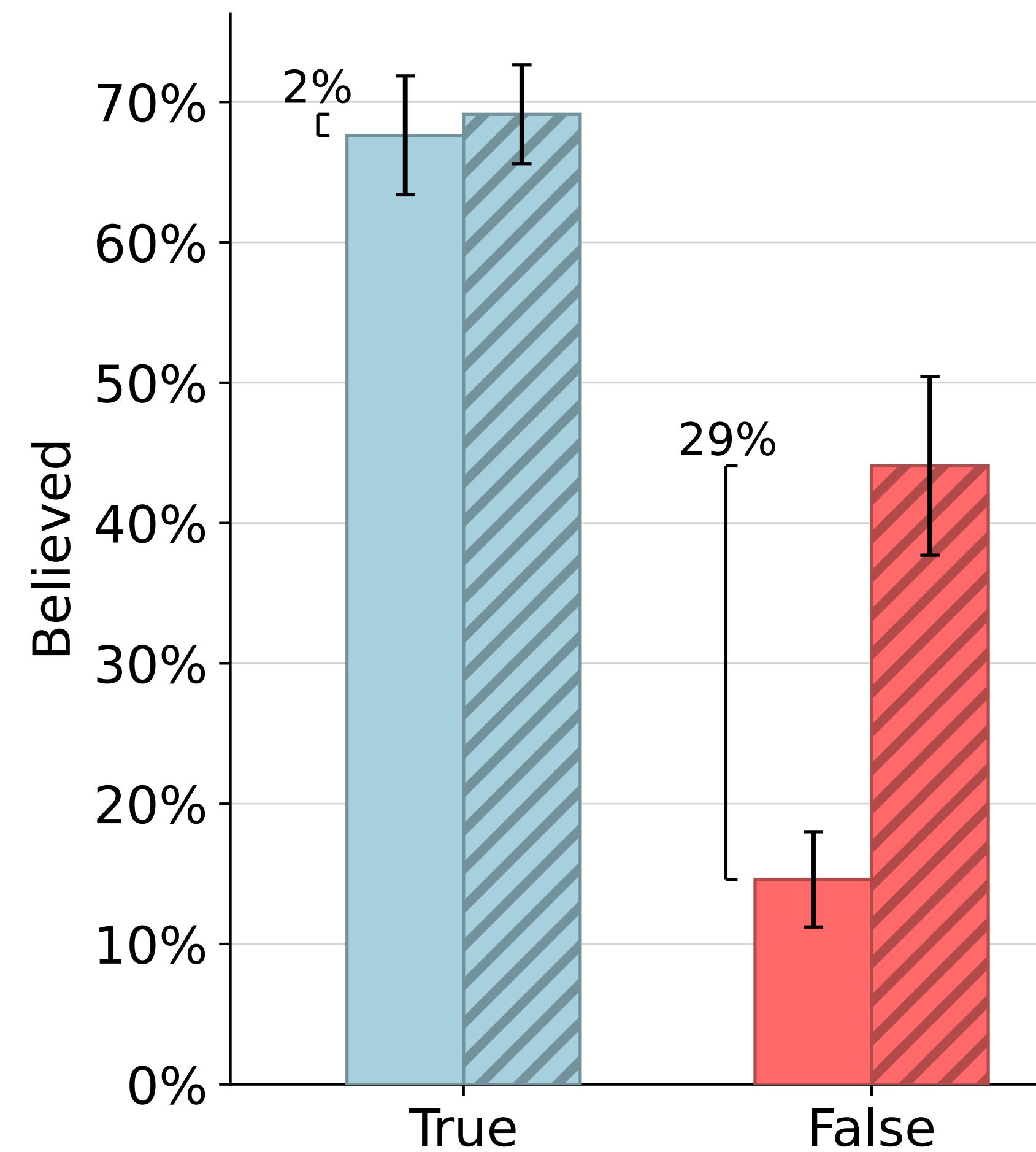
# Accounting for ChatGPT accuracy



# Accounting for ChatGPT accuracy



# Opt-in vs. opt-out



# Takeaways

- Generative AI models pose severe cyber-social threats
- They also have the potential to defend us from these threats
- But human-AI interactions can be tricky

# Thank you!



Filippo Menczer



Matthew R. DeVerna



Harry Yaojun Yan



Danishjeet Singh

Contact

[yang3kc@gmail.com](mailto:yang3kc@gmail.com)

[www.kaichengyang.me](http://www.kaichengyang.me)

Twitter: [@yang3kc](https://twitter.com/yang3kc)

## Anatomy of an AI-powered malicious social botnet

Kai-Cheng Yang and Filippo Menczer | In preparation

## Large language models can rate news outlet credibility

Kai-Cheng Yang and Filippo Menczer | arxiv.org/abs/2304.00228

## Detecting and characterizing fake social media accounts with AI-generated profiles

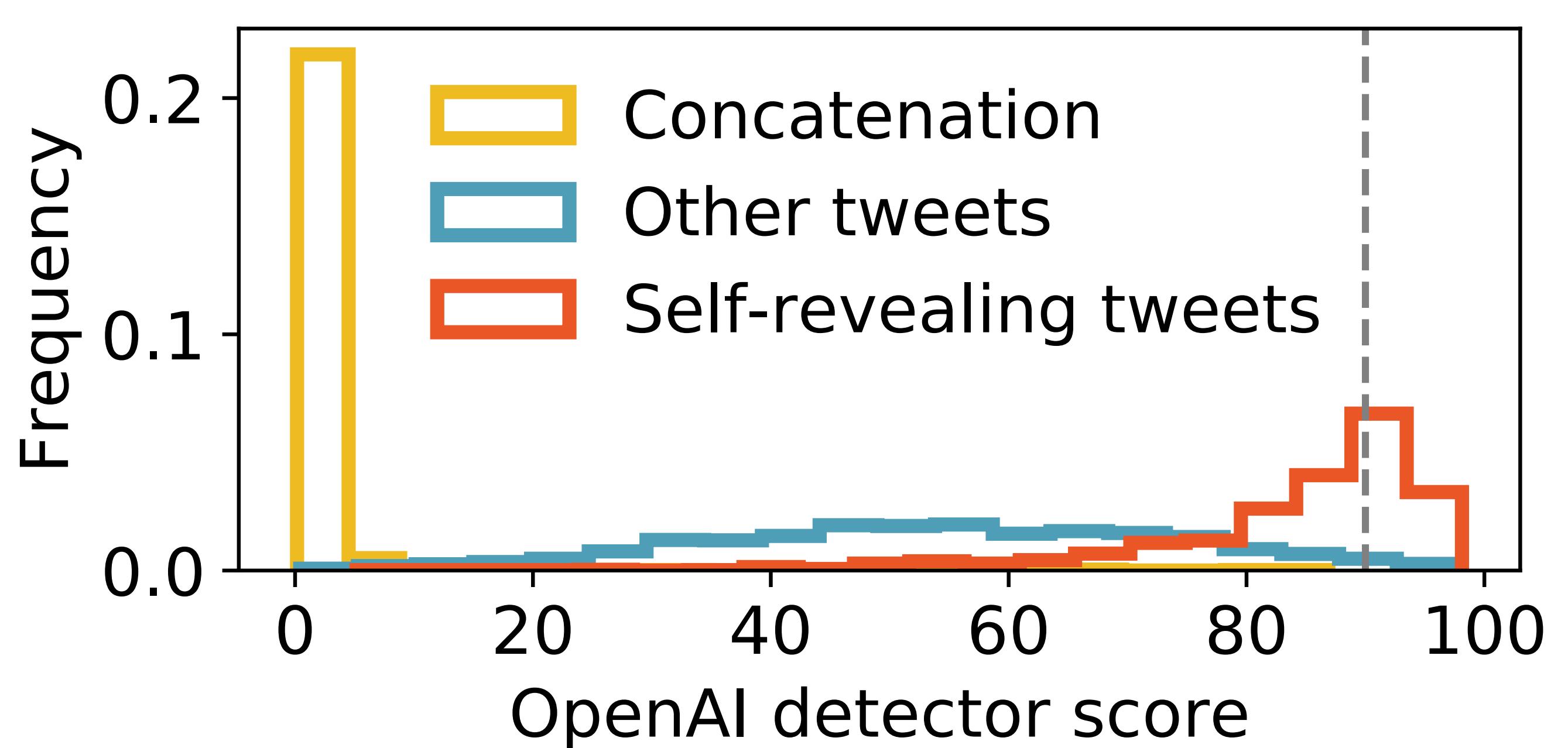
Danishjeet Singh, Kai-Cheng Yang, and Filippo Menczer | In preparation

## AI is ineffective and potentially harmful for fact checking

Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer | In preparation

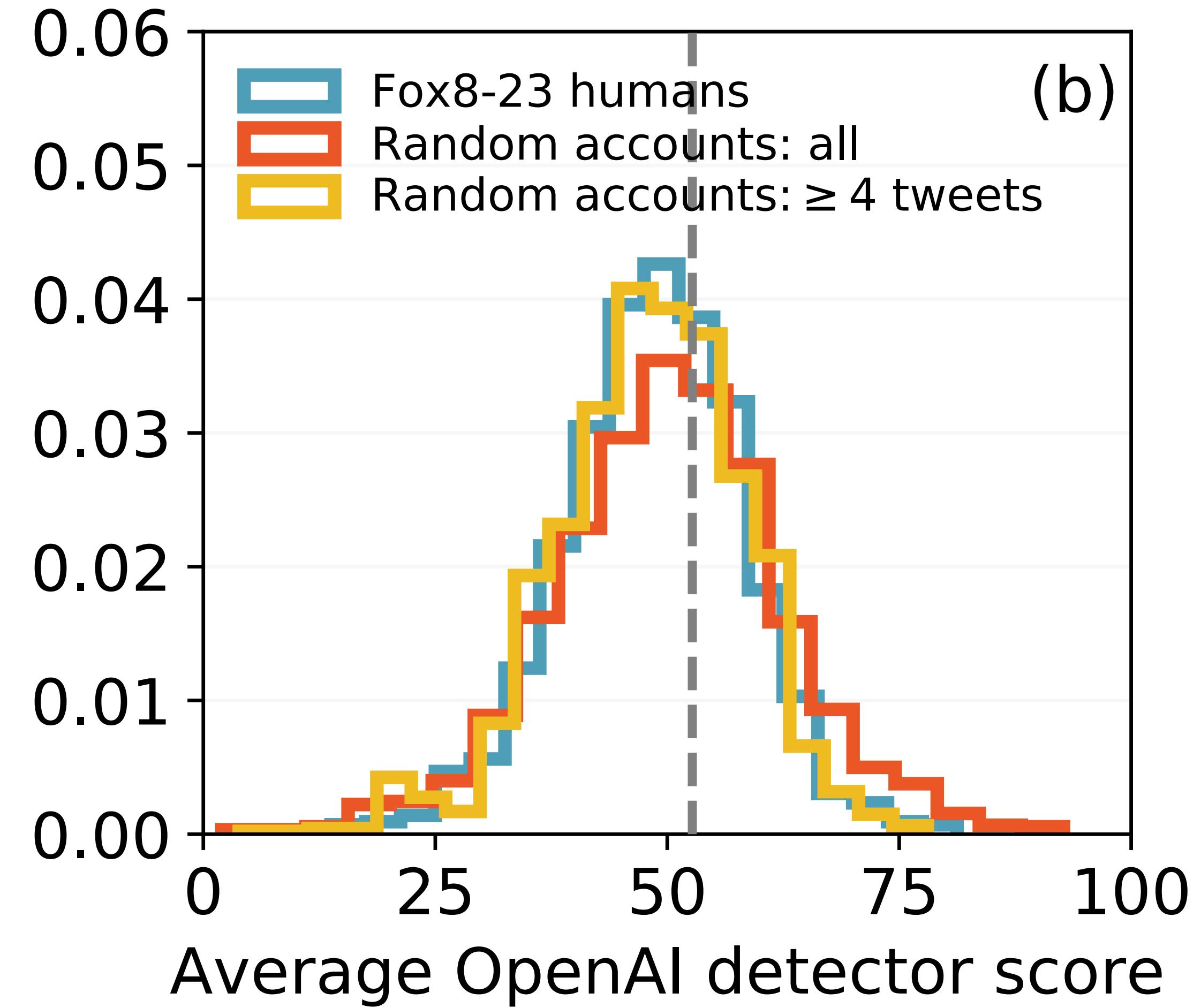
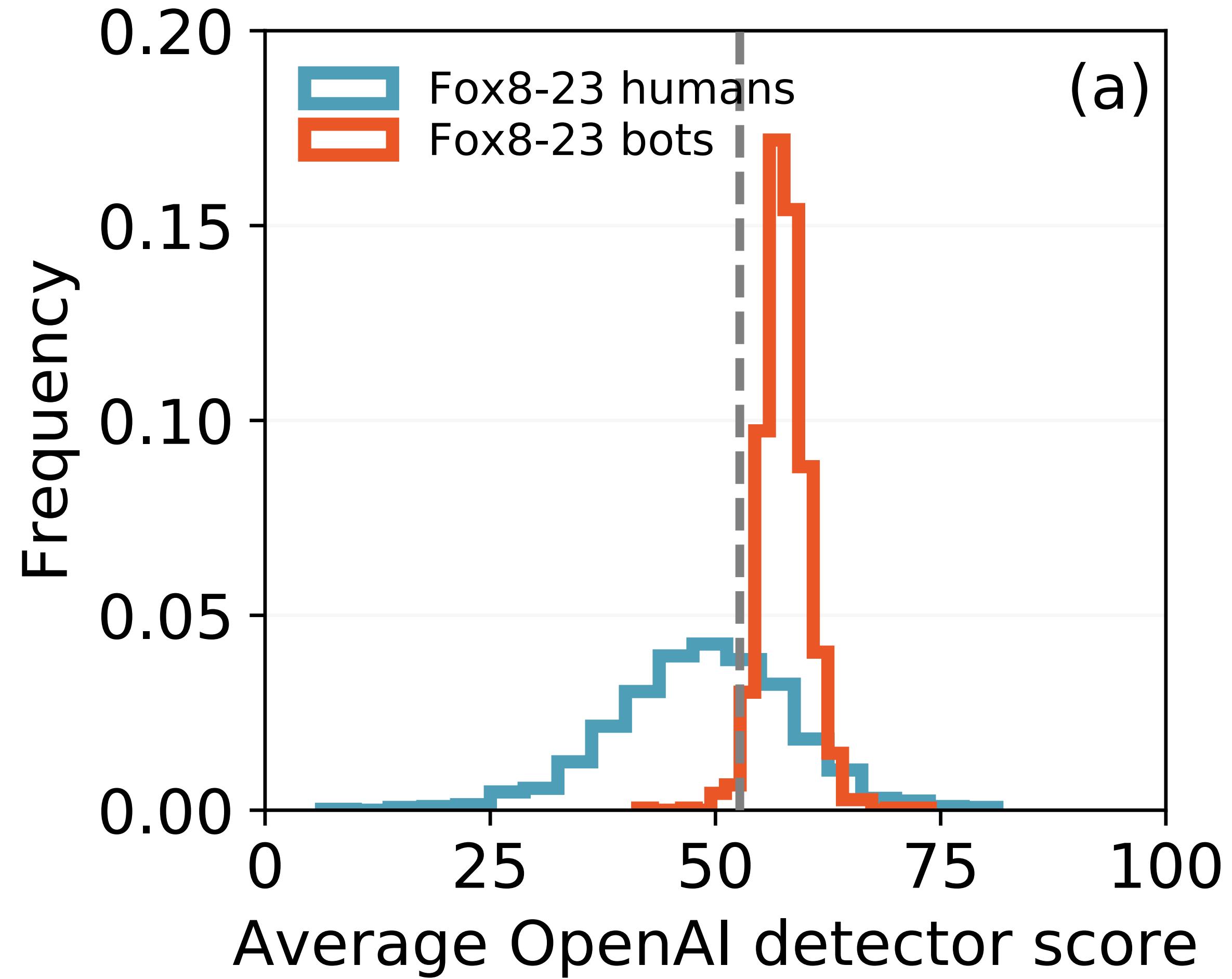
# Appendix

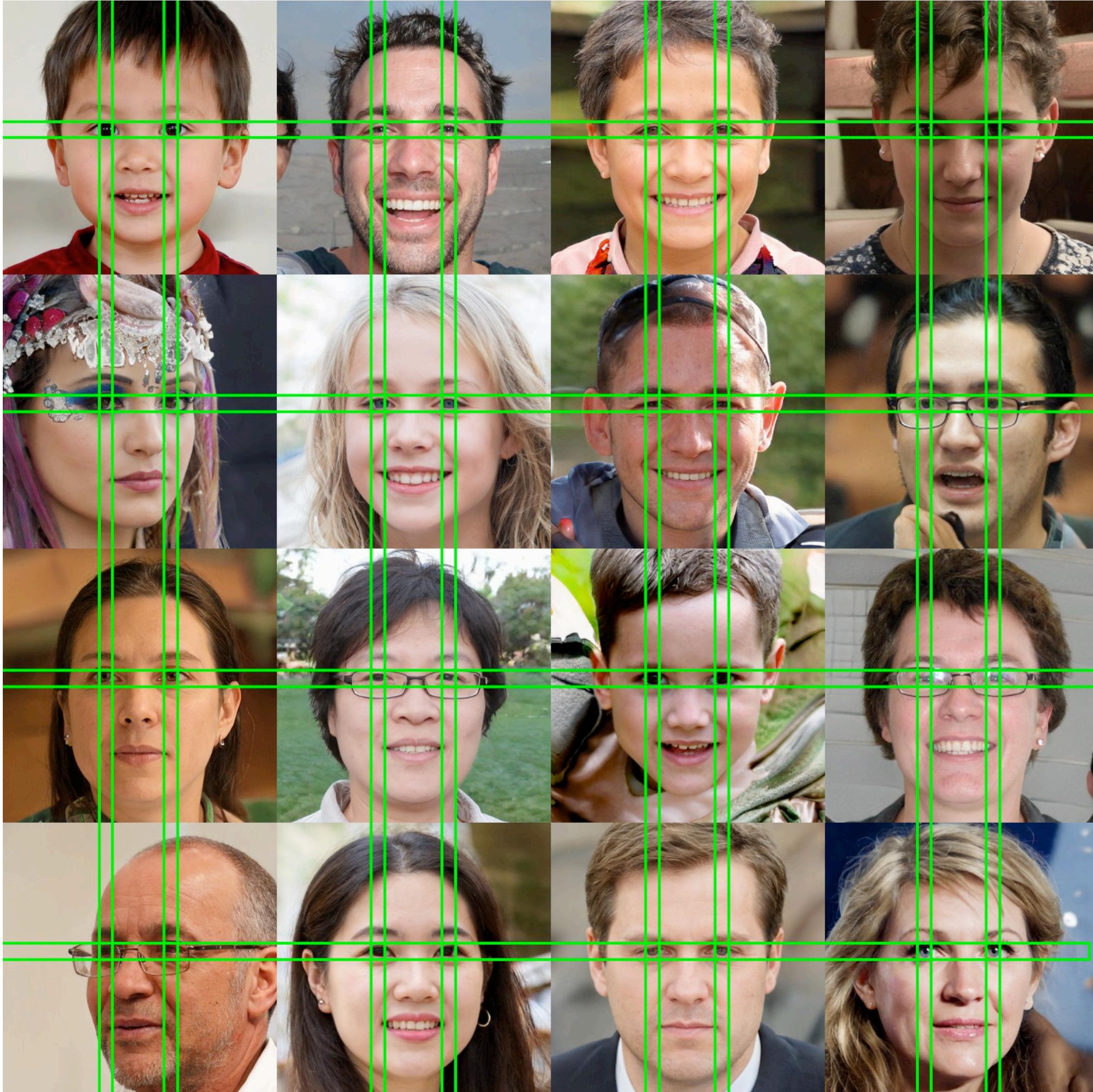
# OpenAI text detector



Score range	Assessment
(0, 10]	very unlikely
(10, 45]	unlikely
(45, 90]	unclear if it is
(90, 98]	possibly
(98, 100]	likely

# OpenAI text detector

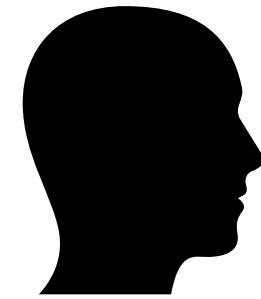




- Computer vision to identify the locations of eyes
- Classifier works pretty well,  $F1=0.68$
- Estimating the prevalence

@conspirator0

# Use ChatGPT to rate news outlet credibility



Rate the website's credibility: sohu.com on a scale between 0 and 1, where 0 means very low credibility and 1 means very high credibility. Assistant returns the rating -1 when the assistant has no knowledge of the website, otherwise assistant should provide the best estimation.



Sohu.com is a well-established Chinese news and media website with a long history of providing credible and reliable information to its readers.

# Non-English outlets

Rating	Language	Sources	%	$\rho$
NewsGuard	English	4,574	84.1	0.51
	Italian	306	5.6	0.38
	French	294	5.4	0.53
	German	267	4.9	0.51
	Total	5,441	100.0	0.51
MBFC	English	2,984	89.4	0.62
	Non-English	354	10.6	0.65
	Total	3,338	100.0	0.60

# Stimuli selection

- 40 political news headlines, half are true, half are false.
- Half are pro-Democrats, and half are pro-Republicans.
- 40 headlines were carefully selected from a larger set of 317 political news stories, using a pretest by Pennycook et al.
- This approach ensured that the selected headlines were balanced in terms of perceived partisanship, impact, familiarity, sensationalism, and the likelihood of being shared and believed