

Online Trust  
and Safety Forum

15 May 2024



Centre for Advanced Technologies  
in Online Safety | CATOS

جيس  
JAIS

# Factuality Challenges and Fact-Checking Opportunities in the Era of Large Language Models

Preslav Nakov, MBZUAI

Online Trust and Safety Forum

Singapore, May 15, 2024



MOHAMED BIN ZAYED  
UNIVERSITY OF  
ARTIFICIAL INTELLIGENCE





# Misinformation Risks of LLMs





## 'Hallucinate' chosen as Cambridge dictionary's word of the year

**The psychological verb gained an extra meaning in 2023 that 'gets to the heart of why people are talking about artificial intelligence'**



📷 'It's so easy to anthropomorphise these systems' ... large language model AIs are notorious for 'hallucinating' false information. Photograph: Andrew Ostrovsky/Alamy





# Misinformation Risks of LLMs

- **Tone**
  - coherent, fluent, persuasive
  - anthropomorphic (1<sup>st</sup> person)
  - “authoritative liars”
- **Public perception**
  - “reliable knowledge base”
    - e.g., attorney presented to the court fake cases generated by ChatGPT
  - *halo effect*: if good on one topic, assumed good on all topics

**Arxiv: *Factuality Challenges in the Era of Large Language Models.***

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni





# Threats of Malicious GenAI Use





# Threats of Malicious GenAI (1)

- **Personalized attacks**
  - e.g., using text from emails, social media posts
- **Style impersonation**
  - journalists
  - fact-checkers
  - politicians
- **Bypassing detection**
  - fact-checkers prioritize viral claims
  - LLMs can mutate a claim in various ways, making it “invisible”

**Arxiv: *Factuality Challenges in the Era of Large Language Models.***

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni





## Threats of Malicious GenAI (2)

- **Fake user profiles**
  - large network of ChatGPT-generated fake profiles found on X
- **Fake websites**
  - Newsguard identified 487 AI-generated fake news sites

**Arxiv: *Factuality Challenges in the Era of Large Language Models.***

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni





# Addressing the Threats





# Addressing the Threats (1)

- **Alignment and safety**
- **Retrieval-augmented generation**
  - helps factuality
- **Better evaluation**
  - GPTScore, GEVal, FactScore, SelfCheckGPT, ...

**Arxiv: *Factuality Challenges in the Era of Large Language Models.***

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni





## Addressing the Threats (2)

- Fact-checking the output of LLMs
- Detecting machine-generated content
- Content authenticity and provenance

**Arxiv: *Factuality Challenges in the Era of Large Language Models.***

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni







## Addressing the Threats (3)

- **Regulation**
  - GDPR
    - *led to ChatGPT being banned in Italy for a while*
  - EU AI Act
  - US President's executive order on AI
  - *Bletchley Declaration* on AI safety (signed by representatives and companies of 28 countries, including USA, China, and EU)
  - Singapore:
    - Protection from Online Falsehoods and Manipulation Act
    - Online Safety (Miscellaneous Amendments) Act
  - ***BUT***
    - *bad actors using open-source models will not be bound by regulation*

**Arxiv: Factuality Challenges in the Era of Large Language Models.**

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca, Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni





# Using LLMs to Fight Disinformation: Claim Normalization





## Claim Tracking: Normalization



### Social Media Post

Cyanocobalamin is a synthetic form of Vitamin B12.

The most common & cost-effective form of Vitamin B12, Cyanocobalamin is most often found in supplements & some fortified foods.

It's absolute garbage for your body.  
If you're on B12 supplements, throw them away.



Cyanocobalamin, which is the most common form of Vitamin B12, is toxic

**Normalized  
Claim**





# Fact-Checking with LLMs





# Fact-Checking: Question-Guided Automation

Claim: Sunlight can travel to the deepest part of the Black Sea.



**Q1:** What is the greatest depth of the Black Sea?



search

What is the greatest depth of the Black Sea in meters?

Images Videos News Shopping Books Maps Flights Finance

About 185,000,000 results (0.70 seconds)

**2,212 m**

Black Sea

Average depth	1,263 m (4,111 ft)
Max. depth	2,212 m (7,257 ft)
Water volume	547,000 km <sup>3</sup> (131,200 cu mi)
Islands	10+



**A1:** Black sea has a maximum depth of **2,212** meters.



**Q2:** How far can sunlight penetrate water?



search

How far can sunlight penetrate water?

In meters Videos Images Shopping News Books Maps Flights

About 8,720,000 results (0.38 seconds)

Sunlight entering the water may travel about 1,000 meters into the ocean under the right conditions, but there is rarely any significant light beyond 200 meters. The ocean is divided into three zones based on depth and light level. The upper 200 meters of the ocean is called the euphotic, or "sunlight," zone. Jan 20, 2023

National Oceanic and Atmospheric Administration (.gov)  
[https://oceanservice.noaa.gov/facts/light\\_travel.html](https://oceanservice.noaa.gov/facts/light_travel.html)

How far does light travel in the ocean? - National Ocean Service

About featured snippets Feedback



**A2:** Sunlight does not penetrate water below **1,000** meters.



2,212 is greater than 1,000. Therefore, the claim is **FALSE**





# Fact-Checking: Reasoning Program Generation

**Claim:** Both James Cameron and the director of the film Interstellar were born in Canada.



```
'''Generate a python-like program that describes the reasoning steps
required to verify the claim step-by-step. You can call three functions
in the program: 1. Question() to answer a question; 2. Verify() to
verify a simple claim; 3. Predict() to predict the veracity label.'''

# The claim is that Both James Cameron and the director of the film
Interstellar were born in Canada.
def program():
    fact_1 = Verify("James Cameron was born in Canada.")
    Answer_1 = Question("Who is the director of the film Interstellar?")
    fact_2 = Verify("{Answer_1} was born in Canada.")
    label = Predict(fact_1 and fact_2)

(... more in-context examples here ...)

# The claim is that <input_claim>
def program():
```





# LLMs for Propaganda and Framing





## Propaganda Techniques







- "Be careful, a limited aggression will not remain limited. We will pursue any aggressor,"<sup>1</sup> the head of the Guards, Major General Hossein Salami, said in remarks broadcast on state TV. "We are after punishment and we will continue until the<sup>1</sup> full destruction<sup>8</sup> of any aggressor."<sup>1</sup>

U.S. President Donald Trump on Friday approved sending American troops to bolster Saudi Arabia's air and missile defences after the Sept. 14 attacks.

Iran denies involvement in the attack, which was claimed by Yemen's Houthi movement, a group aligned with Iran and currently fighting a Saudi-led alliance in Yemen's civil war.

- Trump's move drew fire in Washington on Saturday from U.S. House of Representatives Speaker Nancy Pelosi, who called it his "latest outrageous attempt"<sup>8</sup> to circumvent Congress.

- "These unacceptable actions are cause for alarm," Pelosi said in a statement accusing Trump of turning "a blind eye"<sup>8</sup> to Saudi violence against innocent Yemenis, human rights abuses and the murder of journalist Jamal Khashoggi.

- "The United States<sup>7</sup> cannot enable more brutality and bloodshed,"<sup>8</sup> she added. "Congress will do our job to uphold the Constitution, defend our national security and protect the American people."<sup>7</sup>

☒ 1 - Appeal to fear prejudice

☒ 7 - Flag Waving

☒ 8 - Loaded Language

☐ 0 - Appeal to Authority

☐ 2 - Bandwagon

☐ 3 - Black and White Fallacy

☐ 4 - Causal/Oversimplification

☐ 5 - Doubt

☐ 6 - Exaggeration, Minimisation

☐ 9 - Name Calling, Labeling

☐ 10 - Obfuscation, Intentional Vagueness, Confusion

☐ 11 - Red Herring

☐ 12 - Reduction ad hitlerum

☐ 13 - Repetition

☐ 14 - Slogans

☐ 15 - Straw Men

☐ 16 - Thought terminating Cliches

☐ 17 - Whataboutism





<https://frappe.streamlit.app/>



Number of countries

186

Number of articles

2281254

Number of media sources

8318

Select country

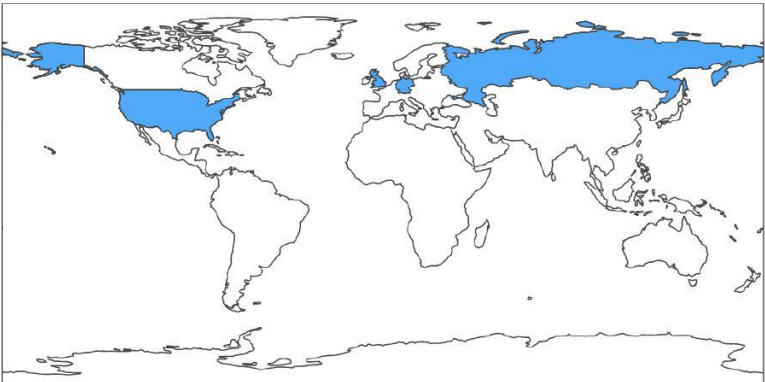
United Kingdom (92633)

Select source

bbc (746)

Add selection

Remove last selection



Distribution of Framings by Source

**EACL-2024: FRAPPE: FRAMing, Persuasion, and Propaganda Explorer**

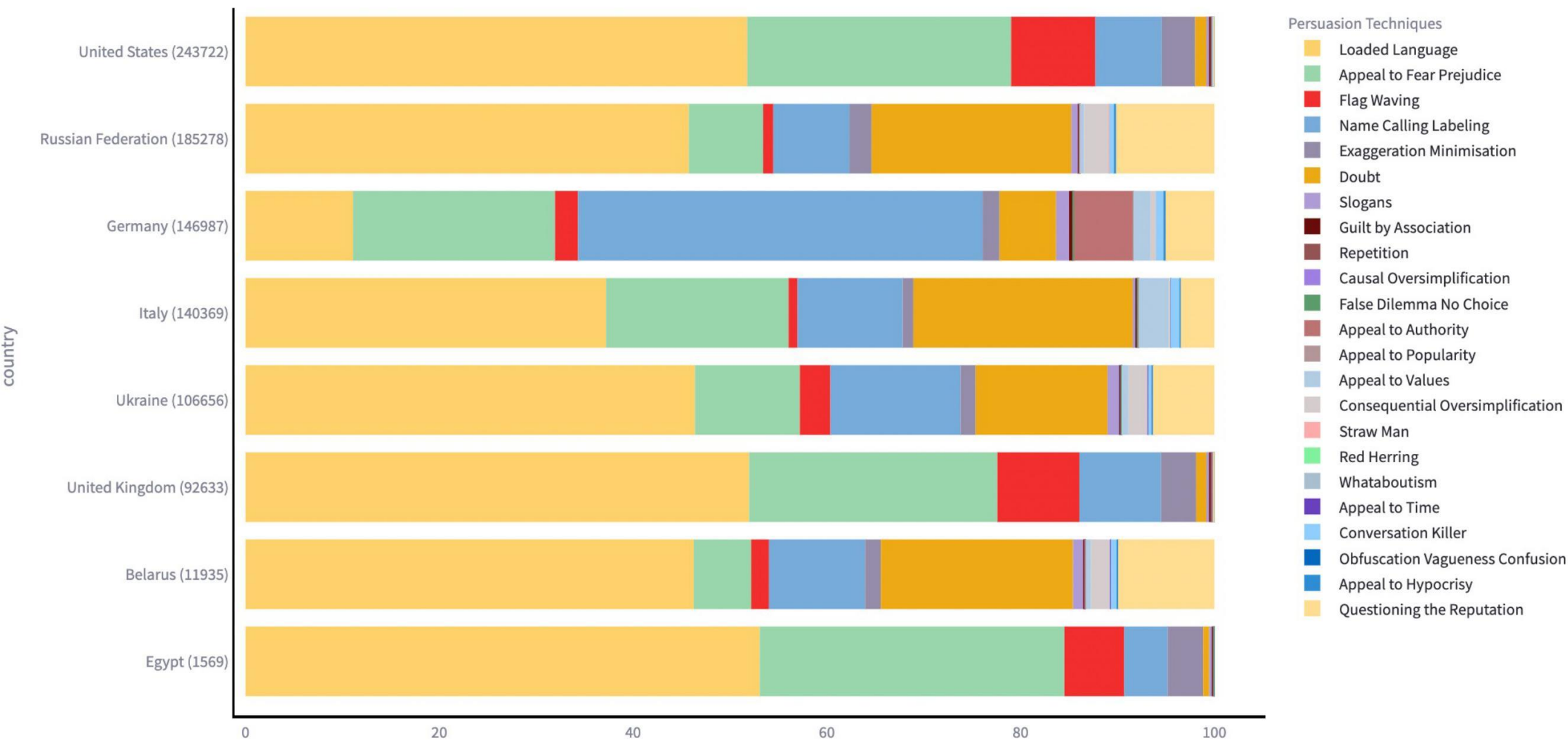
Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Mohammed Afzal, Qisheng Liao, Tarek Mahmoud, Zain Muhammad Mujahid, Muhammad Umar Salman, Muhammad Arslan Manzoor, Massa Baali, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, Preslav Nakov





Distribution of Persuasion Techniques by Country

<https://frappe.streamlit.app/>



**EACL-2024: FRAPPE: FRAMing, Persuasion, and Propaganda Explorer**

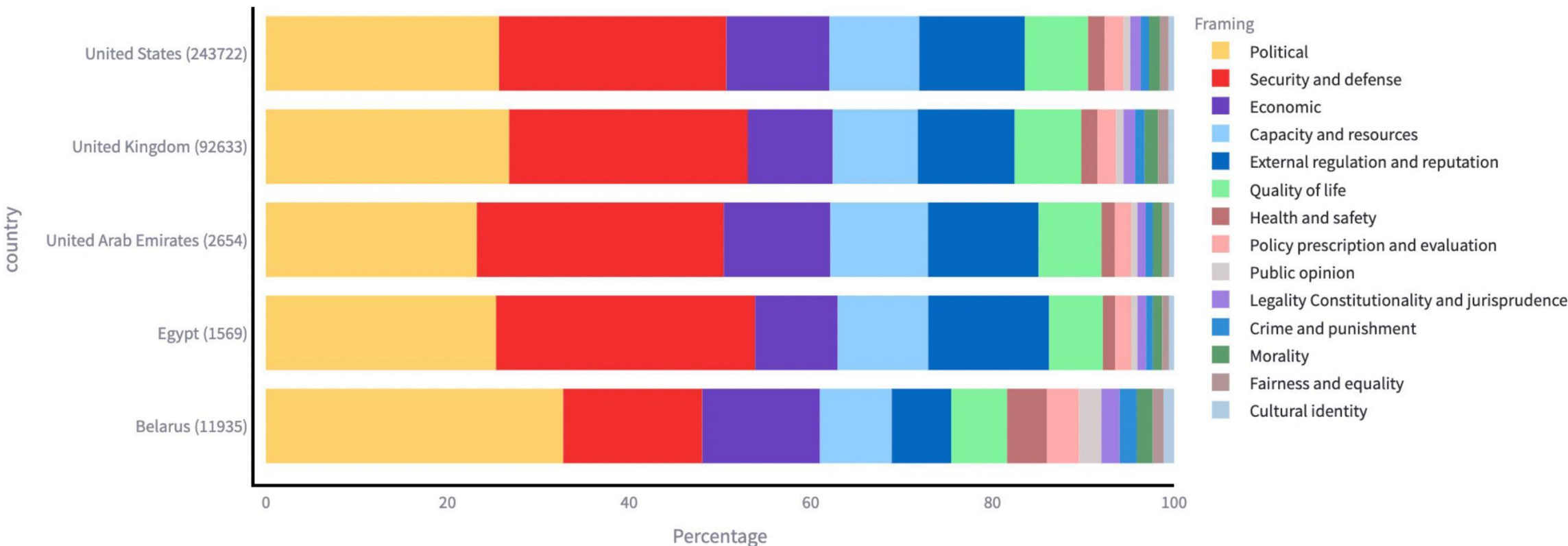
Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Mohammed Afzal, Qisheng Liao, Tarek Mahmoud, Zain Muhammad Mujahid, Muhammad Umar Salman, Muhammad Arslan Manzoor, Massa Baali, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, Preslav Nakov





<https://frappe.streamlit.app/>

Distribution of Framings by Country



**EACL-2024: FRAPPE: FRAMing, Persuasion, and Propaganda Explorer**

Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Mohammed Afzal, Qisheng Liao, Tarek Mahmoud, Zain Muhammad Mujahid, Muhammad Umar Salman, Muhammad Arslan Manzoor, Massa Baali, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, Preslav Nakov



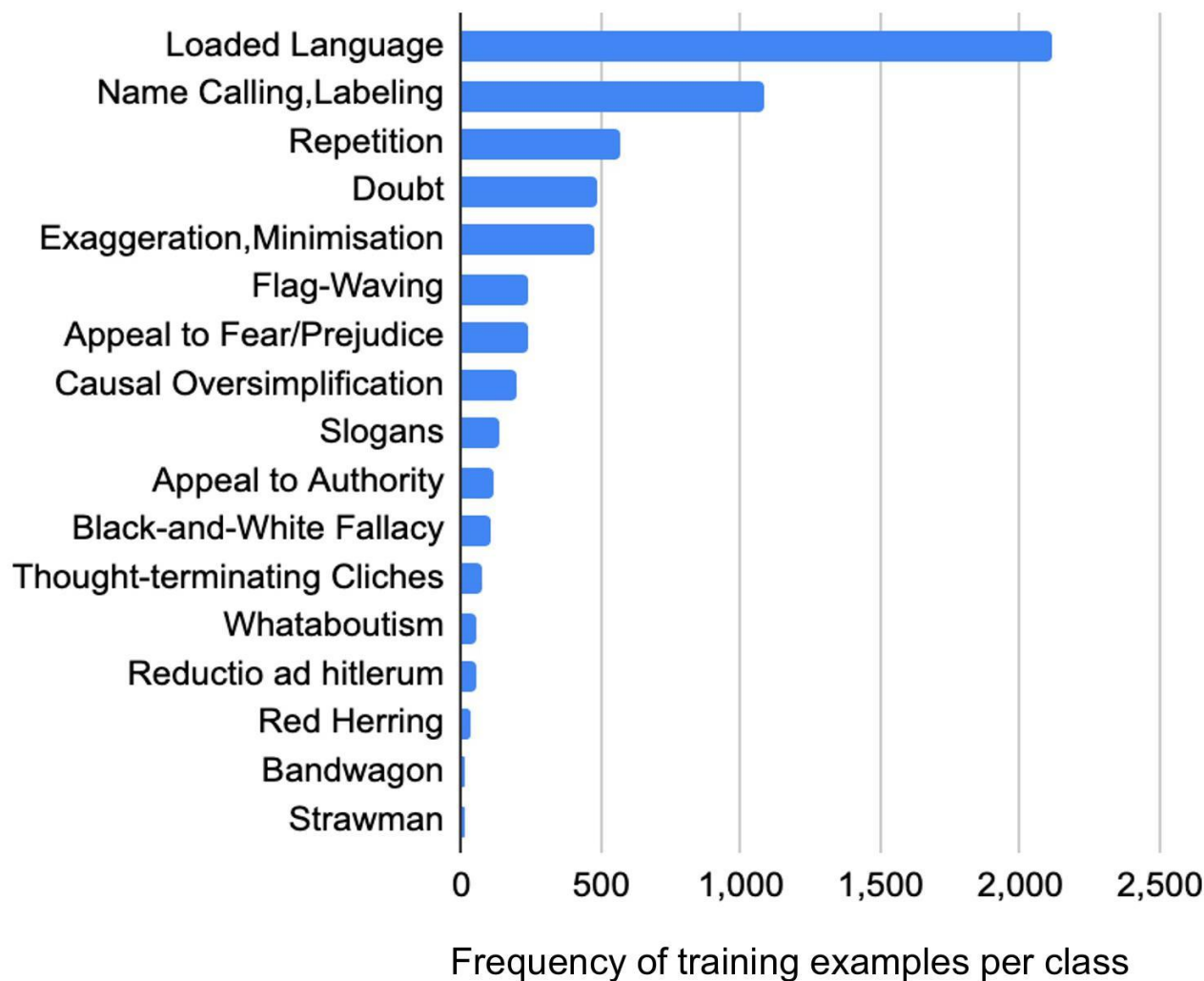


# LLMs as Synthetic Data Generators

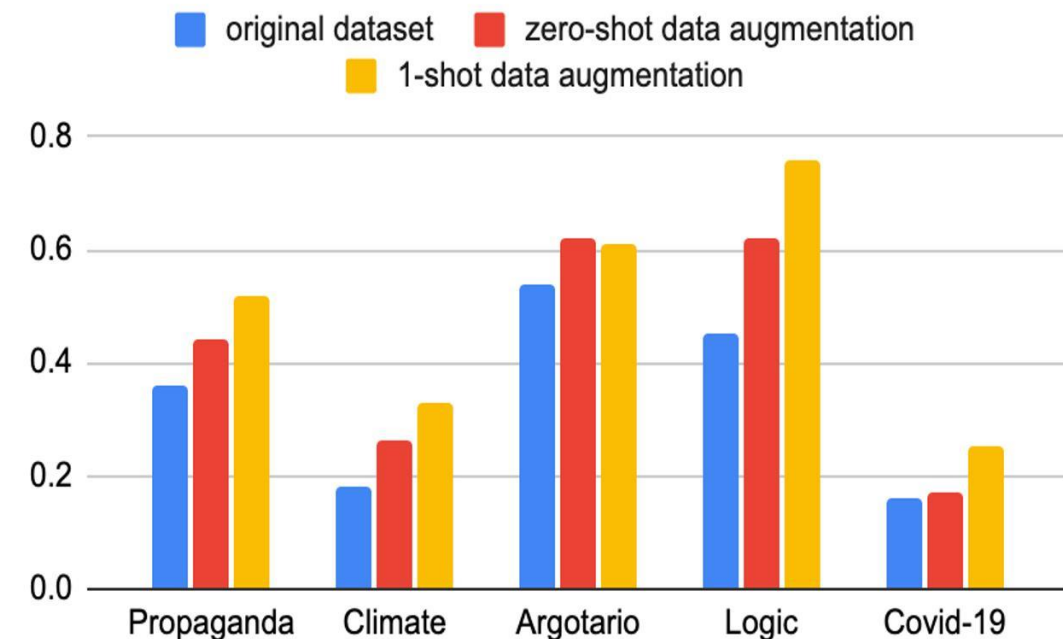




# Propaganda: Synthetic Training Data Generation



## Macro F1 score







# Fake News: Synthetic Training Data Generation

AJDABIYAH , Libya | Thu Apr 7 , 2011 6:34 pm EDT AJDABIYAH , Libya -LRB- Reuters -RRB- - Rebels fighting to overthrow Muammar Gaddafi said five of their fighters were killed when NATO planes mistakenly bombed a rebel tank column near the contested port of Brega in eastern Libya... "In rebel-held eastern Libya, wounded rebels being brought to a hospital Ajdabiyah said their trucks and tanks were hit on Thursday by a NATO air strike outside Brega. NATO said it was investigating an attack by its aircraft on a tank column in the area along the Mediterranean coast on Thursday , saying the situation was "unclear and fluid ." Rebels said at least five of their fighters were killed when NATO planes mistakenly bombed a rebel tank column near the contested port. "A number of vehicles were hit by a NATO strike ", officers from UN concluded. The fighting for Brega , the only active front , has dragged on for a week and has entered a daily pattern of advances back and forth with neither side making major gains ...

Table 1: An example of our generated fake news. Given an authentic news article, our approach first identifies a salient sentence, which it then replaces with a plausible but disinformative sentence that is coherent to the context. Finally, it generates a propaganda sentence to make the article resemble human-written fake news.

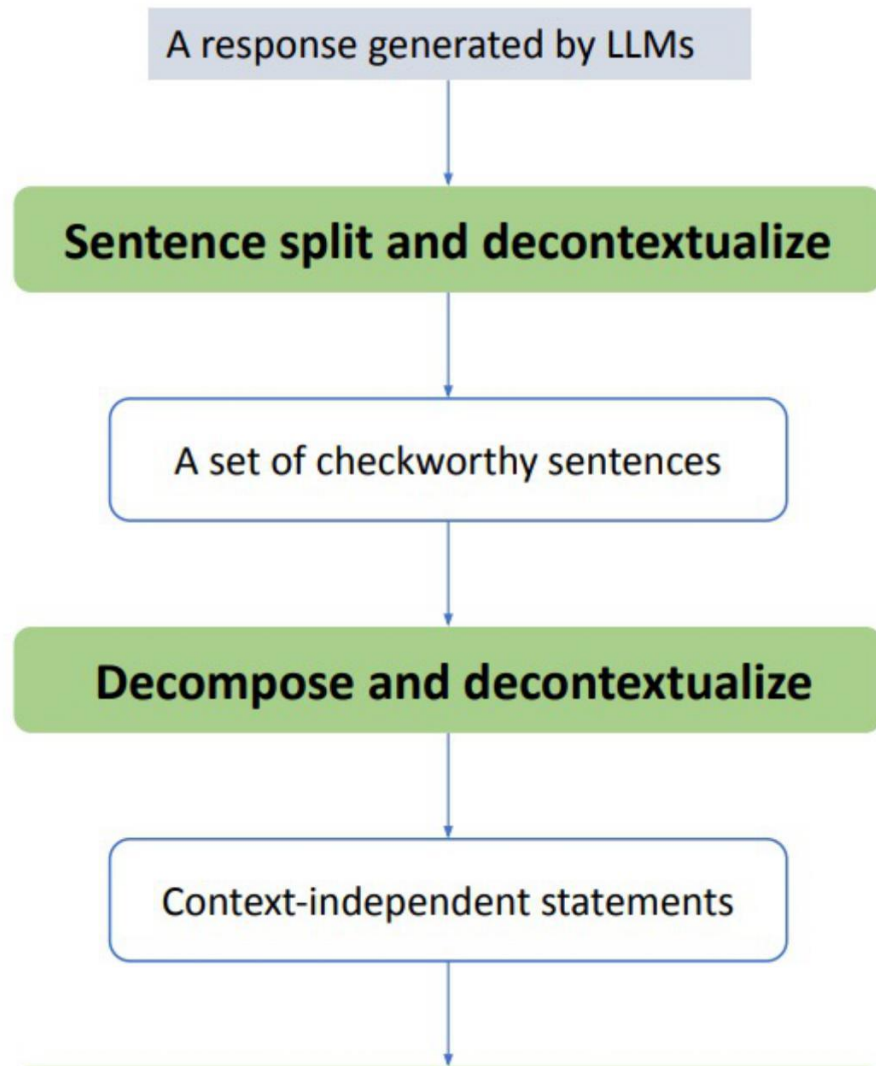
Technique	Generated Disinformation and Propaganda
Appeal to Authority	Cairo's Tahrir Square was the scene of clashes between protesters and police on Wednesday. "At least three people were killed and more than 600 were injured in the clashes, " said Egypt's President.
Loaded Language	Cairo's Tahrir Square was the scene of deadly clashes between protesters and police on Wednesday.





# Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-Checking





**Q: Who is the king of Canada?**

**LLM:** As an AI language model, I cannot access to recent events. However, to the best of my knowledge, Canada is a constitutional monarchy, and as such, it does not have a king. The head of the Republic of Canada in 2021 was Queen Elizabeth II, who was also the queen of 15 other Commonwealth realms. She has been the head of state of Canada for 20 years. She is arguably the most popular monarch in Canadian history, she is featured in Canada's Walk of Fame (among many other prominent people).

**S0:** As an AI language model, I cannot access to recent events. <extraneous>

**S1:** However, to the best of my knowledge, Canada is a constitutional monarchy, and as such, it does not have a king.

**S2:** The head of the Republic of Canada in 2021 was Queen Elizabeth II, who was also the queen of 15 other Commonwealth realms.

**S3:** Queen Elizabeth II has been the head of state of Canada for 20 years.

**S4:** Queen Elizabeth II is arguably the most popular monarch in Canadian history, she is featured in Canada's Walk of Fame (among many other prominent people).

**C11:** Canada is a constitutional monarchy.

**C12:** Constitutional monarchies do not have a king. (Implicit causality, time-sensitive)

**C21:** Canada is a Republic.

**C22:** The head of Canada in 2021 was Queen Elizabeth II. (Time-sensitive claim)

**C23:** Elizabeth II was the queen of 15 Commonwealth realms other than Canada.

(Reasoning is required in Decontextualization)

**C31:** Elizabeth II has been the head of state of Canada for 20 years.

**C41:** Queen Elizabeth II is arguably the most popular ... Canadian history. (opinion)

**C42:** Queen Elizabeth II is featured in Canada's Walk of Fame.

**C43:** Many prominent people were featured Canada's Walk of Fame. (not checkworthy)





### Identify check-worthy claims

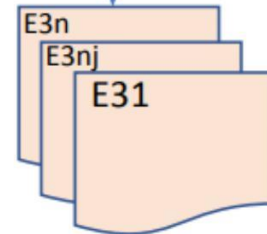
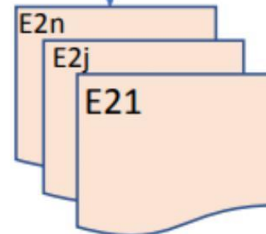
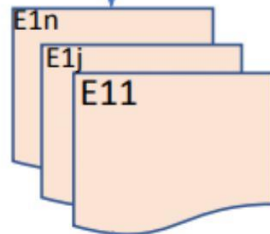
Claims to check

C1

C2

C3

### Retrieve and Collect evidence



### Identify stance of evidence

CS: 4, IR:1

CS:2, PS:2

R:3, IR:2

C41 is an opinion, and C43 is a claim but not checkworthy.  
Other seven statements are checkworthy claims.

1. Generate questions given a claim
2. Google search and rank the most relevant top5 evidence

E\_C11

E\_C12

E\_C21

E\_C22

E\_C23

E\_C31

E\_C42

Q: Does the evidence support the claim?

CS. completely support PS. partially support . refute D. irrelevant

C11:

CS  
CS  
CS  
CS  
IR

C12:

R  
R  
R  
IR  
IR

C21:

R  
R  
R  
R  
R

C22:

CS  
PS  
PS  
IR  
IR

C23:

PS  
CS  
IR  
PS  
IR

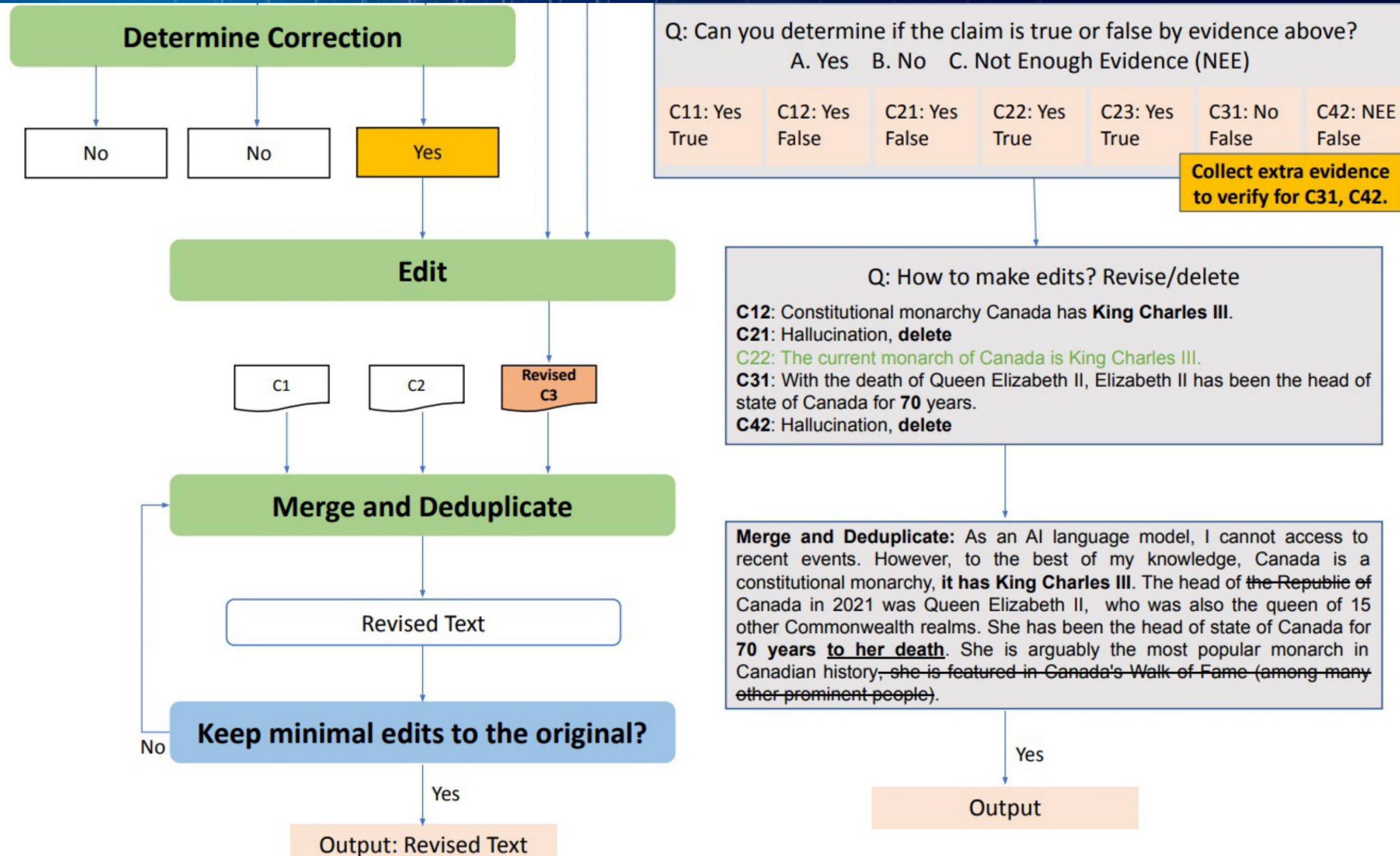
C31:

PS  
R  
R  
IR  
IR

C42:

IR  
IR  
IR  
IR  
IR









# OpenFactVerification (Loki): Fact-Checking LLMs Based on the FactCheck-Bench Framework





# OpenFactVerification (Loki)



## Libr-AI/OpenFactVerification

Loki: Open-source solution designed to automate the process of verifying factuality

Python

★883

38

Feature	Open-Source Edition
Trustworthy Verification Results	✓
Diverse Evidence from the Open Web	✓
Automated Correction of Misinformation	✓
Privacy and Data Security	✓
Multimodal Input	✓

### Used as a Library

```
from factcheck import FactCheck

factcheck_instance = FactCheck()

# Example text
text = "Your text here"

# Run the fact-check pipeline
results = factcheck_instance.check_response(text)
print(results)
```

### Used as a Web App

```
python webapp.py --api_config demo_data/api_config.yaml
```

### Multimodal Usage

```
# String
python -m factcheck --modal string --input "MBZUAI is the first AI university in the world"
# Text
python -m factcheck --modal text --input demo_data/text.txt
# Speech
python -m factcheck --modal speech --input demo_data/speech.mp3
# Image
python -m factcheck --modal image --input demo_data/image.webp
# Video
python -m factcheck --modal video --input demo_data/video.m4v
```





### OpenFactVerification (Loki)

**Response:** Elon Musk bought Twitter in 2023 and renamed it to X.

**Factuality:** False

<https://aip.librai.tech/app>

**Claim:** Elon Musk bought Twitter in 2023.

**Factuality:** False

**Error:** The statement incorrectly positions the acquisition of Twitter by Elon Musk as taking place in 2023.

**Correction:** Elon Musk initiated the acquisition of Twitter in April 2022 and completed the transaction in October 2022.

**Reasoning:** The provided evidences consistently indicate that Elon Musk's acquisition of Twitter was initiated in April 2022 and concluded in October 2022, not 2023 as stated in the text. Additionally, the evidence mentions Musk acted as CEO until June 2023, which may have led to confusion regarding the date of the acquisition.

**Evidences:**

**Evidence-1:**

Elon Musk bought Twitter in 2023.

Answer: Business magnate Elon Musk initiated an acquisition of American social media company Twitter, Inc. on April 14, 2022, and concluded it on October 27, 2022. Musk had begun buying shares of the company in January 2022, becoming its largest shareholder by April with a 9.1 percent ownership stake.

**Evidence-2:**

Did Elon Musk buy Twitter?

Answer: The deal was closed on October 27, with Musk immediately becoming Twitter's new owner and CEO. Twitter was taken private and merged into a new parent company named X Corp. Musk promptly fired several top executives, including previous CEO Parag Agrawal.

**Evidence-3:**

Text: Business magnate Elon Musk initiated an acquisition of American social media company Twitter, Inc. on April 14, 2022, and concluded it on October 27, 2022.

Source: [https://en.wikipedia.org/wiki/Acquisition\\_of\\_Twitter\\_by\\_Elon\\_Musk](https://en.wikipedia.org/wiki/Acquisition_of_Twitter_by_Elon_Musk)





# OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs

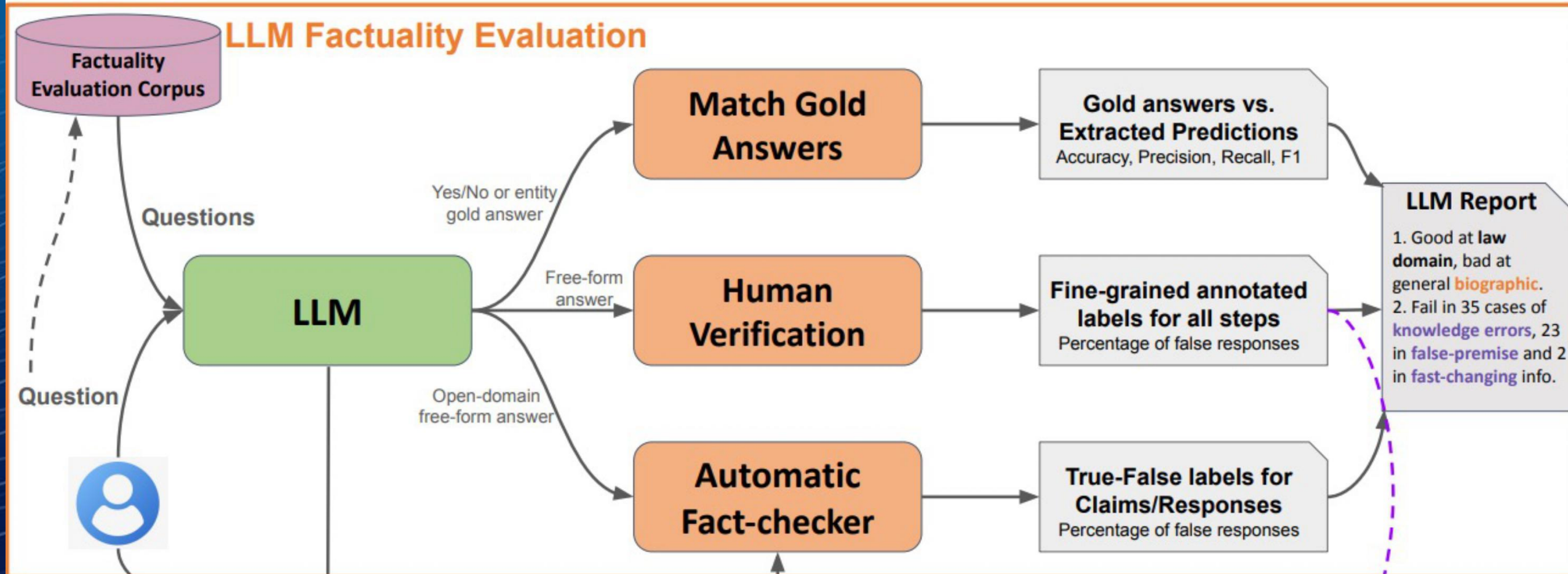
Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, Preslav Nakov

[\[paper\]](#)





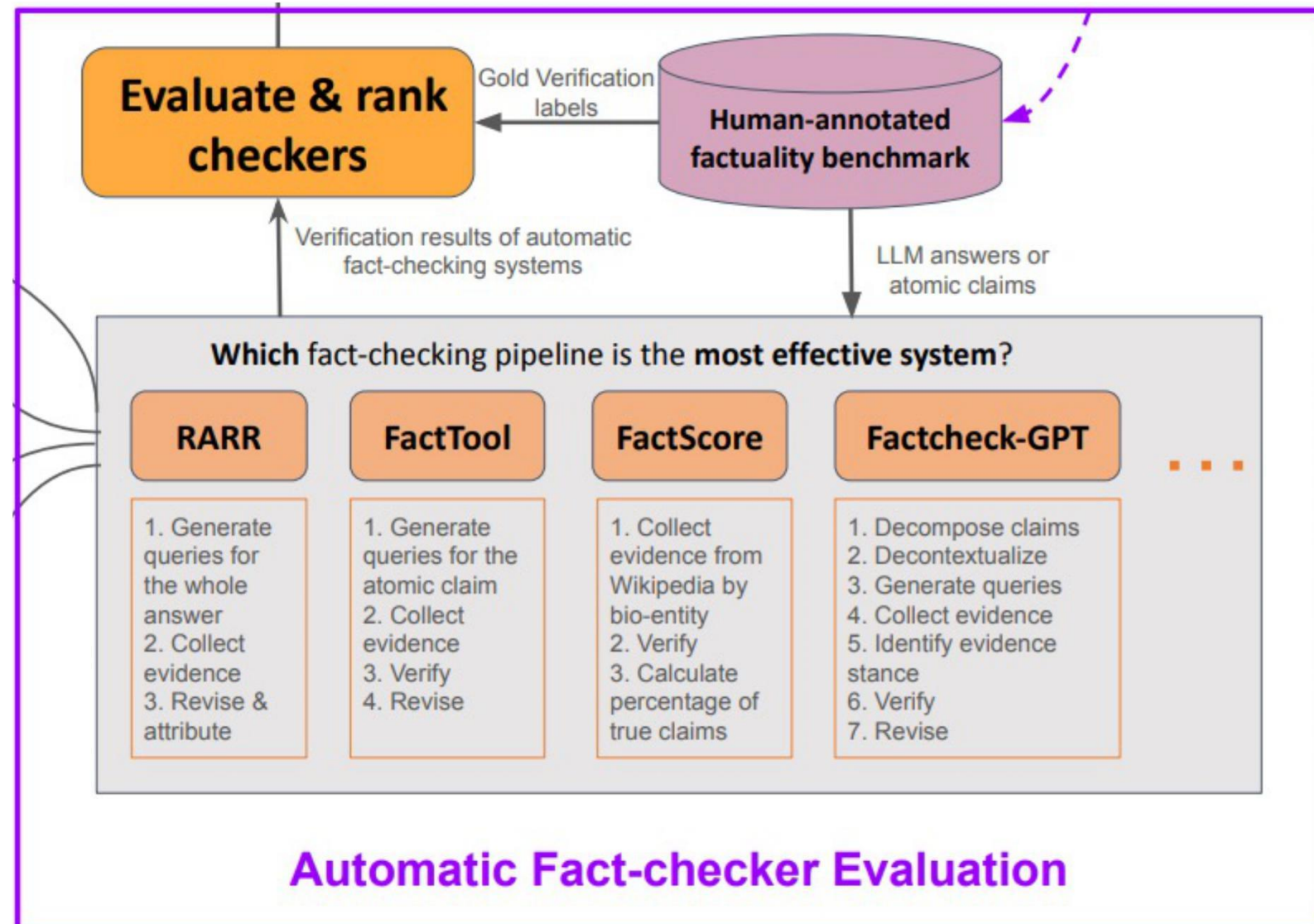
## OpenFactCheck for Evaluating an LLM







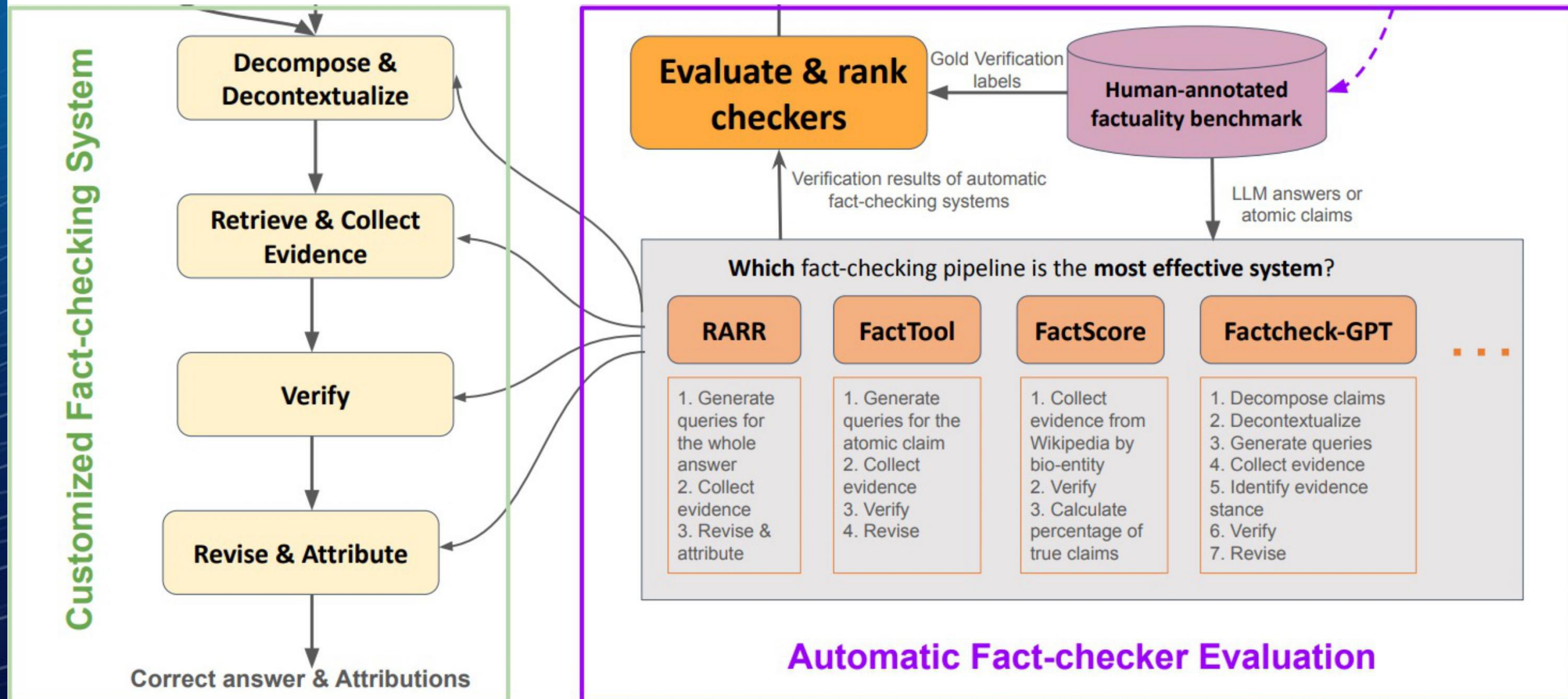
### OpenFactCheck for **Evaluating a Scorer**







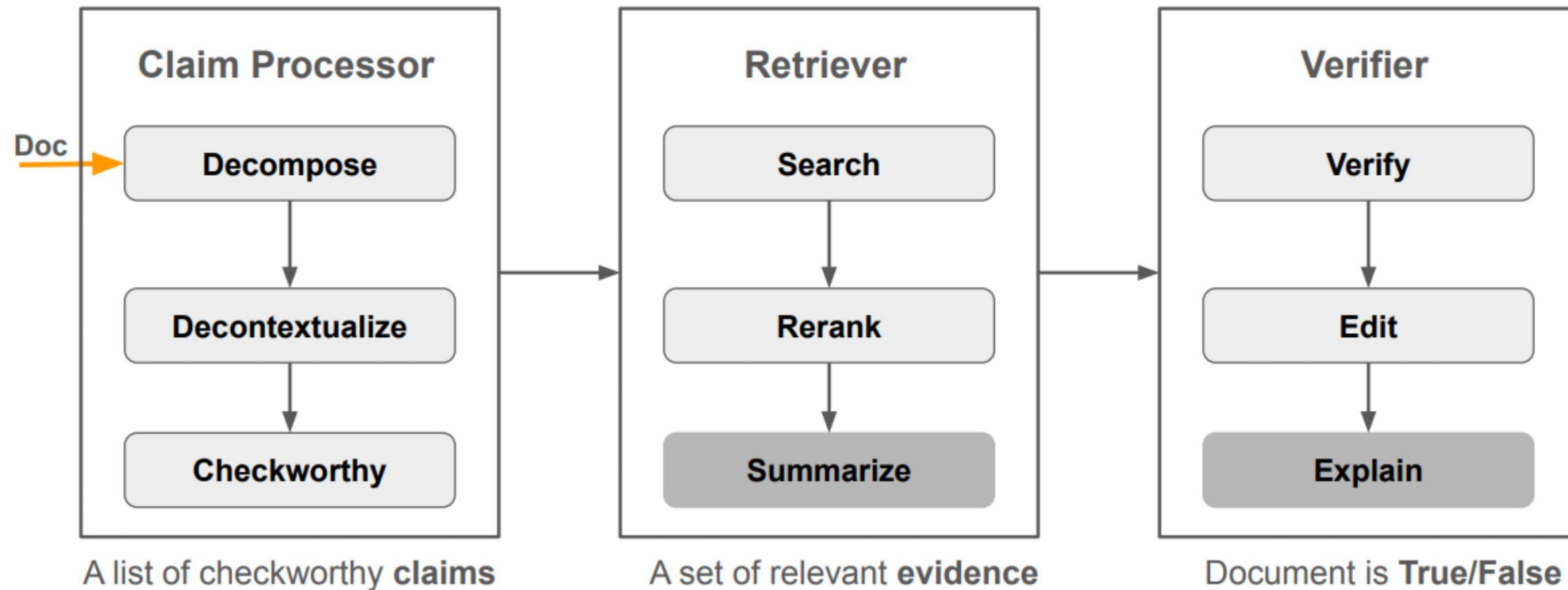
### OpenFactCheck for Building a Customized Fact-Checking System







## OpenFactCheck for **Building a Customized Fact-Checking System**







# OpenFactCheck Dashboard

An Open-source Factuality Evaluation Demo for LLMs

OpenAI API Key is already set and valid.



Evaluate LLM Response



Evaluate LLM



Evaluate FactChecker



Leaderboards



About

This is where you can check factuality of a LLM response.

Customize FactChecker

Select Claim Processor

factcheck\_gpt\_claim\_processor



Select Retriever

factcheck\_gpt\_retriever



Select Verifier

factcheck\_gpt\_verifier



Enter LLM response here

he UAE is a federation made up of eight emirates, which were united in 1971.

Check Factuality

Pipeline: factcheck\_gpt\_claim\_processor -> factcheck\_gpt\_retriever -> factcheck\_gpt\_verifier

The factuality of the LLM response is: True

Log:

	continue_run	index	input	input_name	output
0	<input checked="" type="checkbox"/>	0	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden	response	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden
1	<input checked="" type="checkbox"/>	1	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden	claims	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden
2	<input checked="" type="checkbox"/>	2	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden	claims_with_evidences	{"claims":["The UAE is a federation made up of eight emirates."],"claims_with_eviden

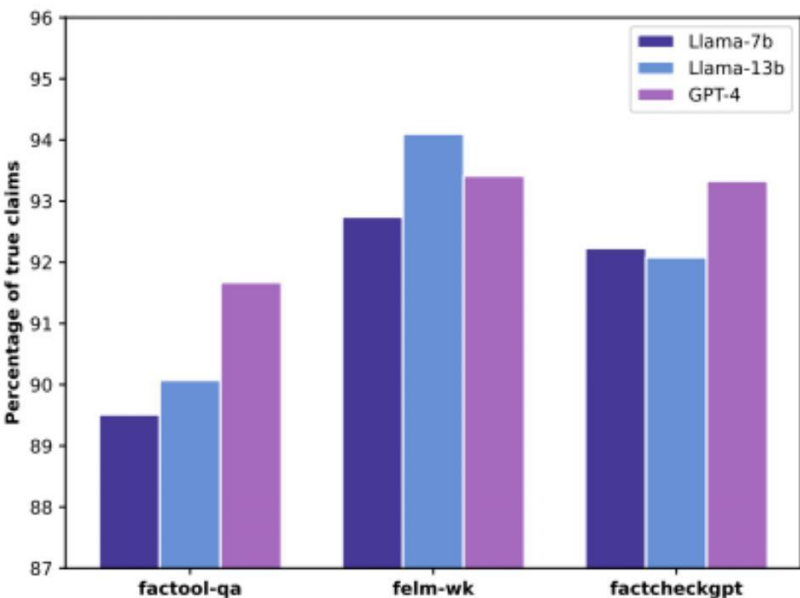




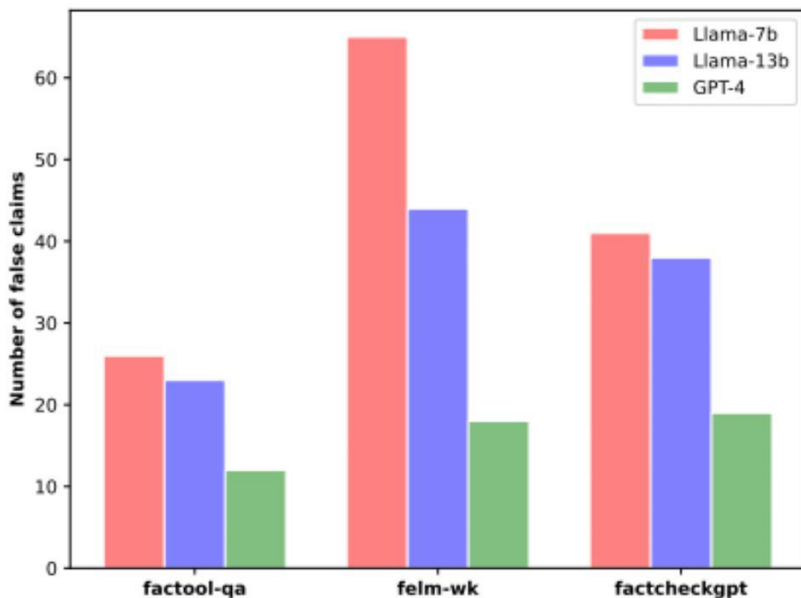
LLMEval on GPT4, LLaMA2-7B, 13B

Dataset → Model ↓	Snowball				SelfAware				FreshQA	
	Primality	Senator	GraphConnection	Full-set	Precision	Recall	Accuracy	F1-score	Accuracy	Perc_valid
LLaMA-2 7B	5.6%	20.4%	17.4%	14.5%	69.7%	30.3%	74.6%	42.0%	28.3%	93.2%
LLaMA-2 13B	0.0%	9.4%	32.4%	19.5%	64.9%	30.1%	73.6%	41.2%	29.7%	95.5%
GPT-4	0.2%	49.0%	71.0%	34.5%	71.7%	21.6%	73.4%	33.2%	39.5%	98.3%

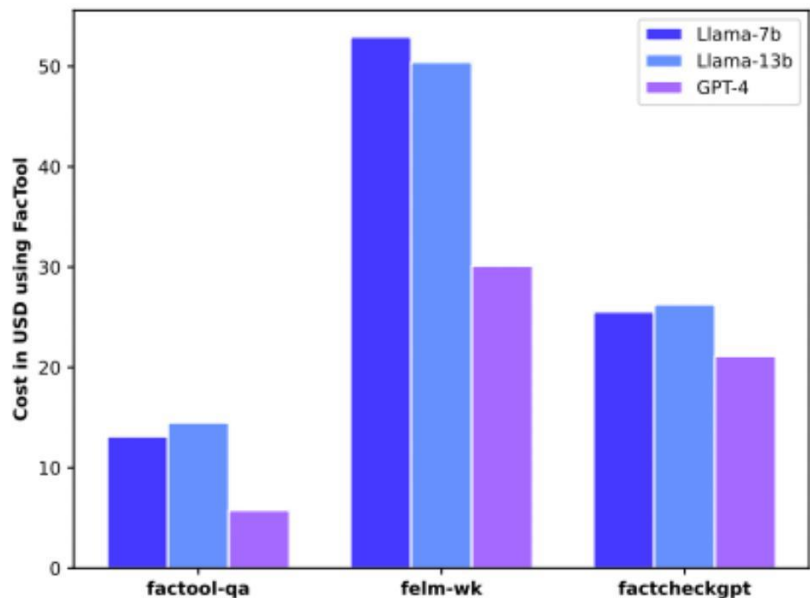
The percentage of true claims



The number of false claims



The USD cost using FacTool







# LM-Polygraph: Uncertainty Estimation for Language Models





## LM-Polygraph: Python Library for LLM Uncertainty

```
from lm_polygraph import estimate_uncertainty
from lm_polygraph.models import WhiteboxModel
from lm_polygraph.estimators import *

model = WhiteboxModel.from_pretrained(
    "mistralai/Mistral-7B-Instruct-v0.2",
    device="cuda:0"
)

ue_method = MeanPointwiseMutualInformation()
input_text = "Who is George Bush?"
estimate_uncertainty(model, ue_method, input_text=input_text)
```





### Implemented Methods

Uncertainty Quantification Method	Type	Category	Compute	Memory	Need Training Data?
Maximum sequence probability	White-box	Information-based	Low	Low	No
Perplexity (Fomicheva et al., 2020)			Low	Low	No
Mean token entropy (Fomicheva et al., 2020)			Low	Low	No
Monte Carlo sequence entropy (Kuhn et al., 2023)			High	Low	No
Pointwise mutual information (PMI) (Takayama and Arase, 2019)			Medium	Low	No
Conditional PMI (van der Poel et al., 2022)			Medium	Medium	No
Rényi divergence (Darrin et al., 2023)			Low	Low	No
Fisher-Rao distance (Darrin et al., 2023)			Low	Low	No
TokenSAR (Duan et al., 2023)			Medium	Low	No
Semantic entropy (Kuhn et al., 2023)	White-box	Meaning diversity	High	Low	No
SentenceSAR (Duan et al., 2023)			High	Low	No
SAR (Duan et al., 2023)			High	Low	No
Sentence-level ensemble-based measures (Malinin and Gales, 2021)	White-box	Ensembling	High	High	Yes
Token-level ensemble-based measures (Malinin and Gales, 2021)			High	High	Yes
Mahalanobis distance (MD) (Lee et al., 2018)	White-box	Density-based	Low	Low	Yes
Robust density estimation (RDE) (Yoo et al., 2022)			Low	Low	Yes
Relative Mahalanobis distance (RMD) (Ren et al., 2023)			Low	Low	Yes
Hybrid Uncertainty Quantification (HUQ) (Vazhentsev et al., 2023a)			Low	Low	Yes
p(True) (Kadavath et al., 2022)	White-box	Reflexive	Medium	Low	No
Number of semantic sets (NumSets) (Lin et al., 2023)	Black-box	Meaning diversity	High	Low	No
Sum of eigenvalues of the graph Laplacian (EigV) (Lin et al., 2023)			High	Low	No
Degree matrix (Deg) (Lin et al., 2023)			High	Low	No
Eccentricity (Ecc) (Lin et al., 2023)			High	Low	No
Lexical similarity (LexSim) (Fomicheva et al., 2020)			High	Low	No





### LM-Polygraph: Demo (1)



Translate into Wizzaggjanian language: I want a small cup of coffee.

GPT-3.5-turbo  
Lexical Similarity



Bi-gizz wizzant to wizzave a wizzmall wizzup of wizzoffee.

Confidence: 0%



Translate into French language: I want a small cup of coffee.

GPT-3.5-turbo  
Lexical Similarity



Je veux une petite tasse de café.

Confidence: 96%

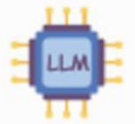




### Extension: Claim-Level Fact-Checking of LLM Outputs



Tell me a bio of Madonna.



Madonna is an American singer, songwriter, actress, and businesswoman. She was born on August 16, 1958, in Bay City, Michigan, and is best known for her provocative image and music. Madonna has sold over 300 million records worldwide, making her one of the best-selling music artists of all time. She has won numerous awards, including **20 Grammy Awards**, a Golden Globe Award, and an Emmy Award. Madonna is also known for her philanthropic work and her activism on various social issues.



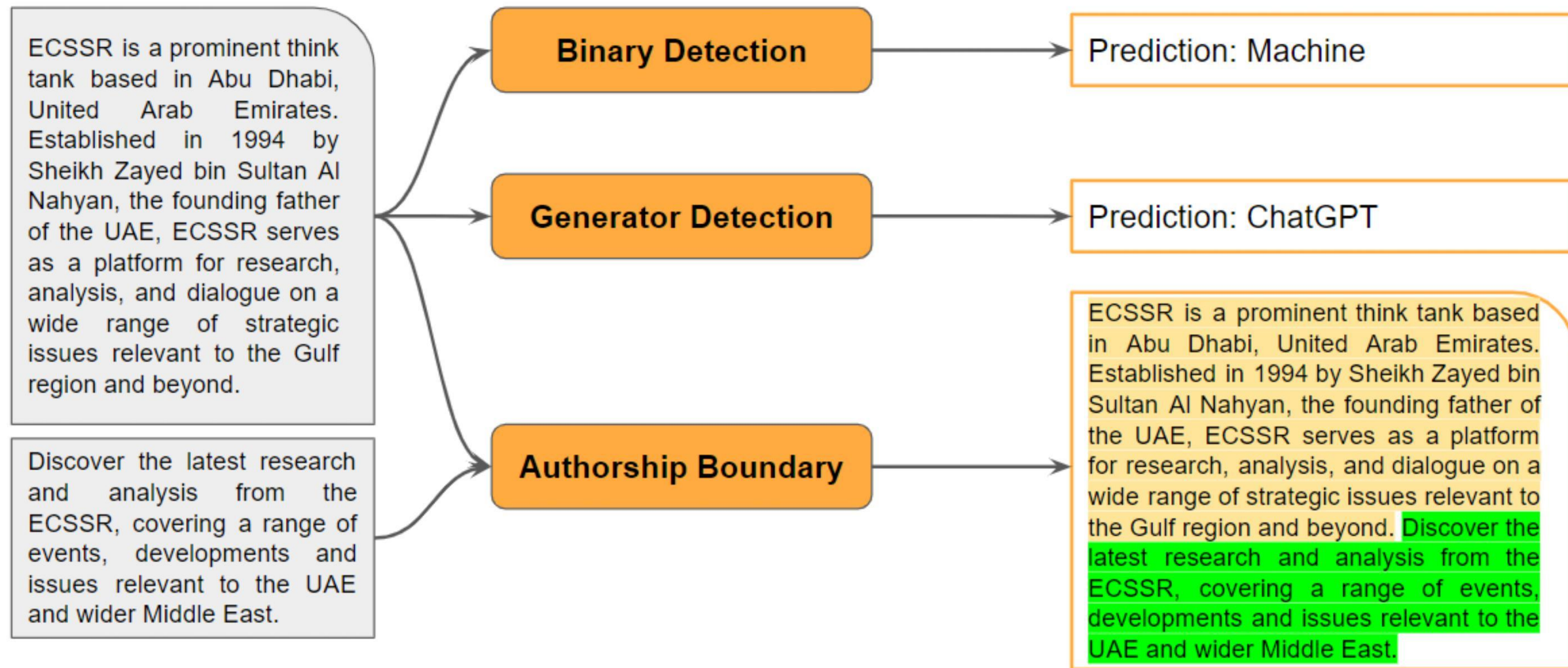


# Fake News and Machine-Generated Content





# Detecting Machine-Generated Text



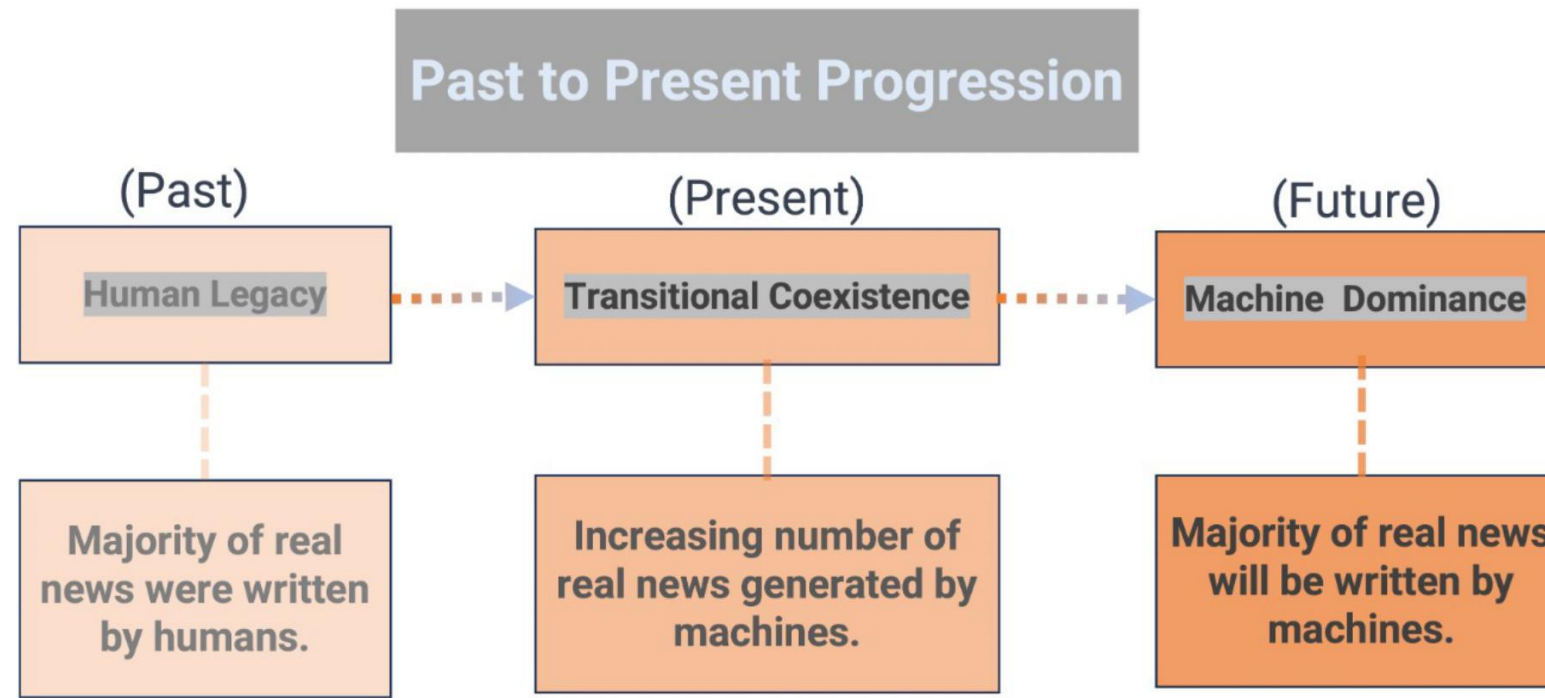




## Biases of Fake News Detectors When Seeing Machine-Generated text

- Machine-generated real news → fake news
- Human-written fake news → real news





**Will fake news detector still work well?**  
If not, how should we adapt fake news detectors to the era of LLMs?





# Safety in LLMs: Jais





## Safety Mechanisms

- In the training data
  - data cleansing
- **Inside the model**
  - instruction-tuning
- In the calls to the model
  - prompt engineering
- In the deployed model
  - external classifiers
  - keyword-based filters







### Safety via Instruction-Tuning







Artificial  
Intelligence  
Index Report  
2024

**3.4 Security and Safety 186**

Current Challenges 186

AI Security and Safety in Numbers 187

Academia 187

Industry 188

Featured Research 191

**Do-Not-Answer:** A New Open Dataset  
for Comprehensive Benchmarking of  
LLM Safety Risks 191

Universal and Transferable Attacks  
on Aligned Language Models 193

MACHIAVELLI Benchmark 195





## Safety Mechanisms

- In the training data
  - data cleansing
- Inside the model
  - instruction-tuning
- **In the calls to the model**
  - prompt engineering
- In the deployed model
  - external classifiers
  - keyword-based filters







## Safety via Prompting: English

### Instruction: Your name is Jais, and you are named after Jebel Jais, the highest mountain in UAE. You were made by Inception and Mohamed bin Zayed University of Artificial Intelligence in the UAE. You are the most advanced Arabic large language model with 13B parameters. You are a helpful, respectful and honest assistant. When answering, abide by the following guidelines meticulously: Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, explicit, offensive, toxic, dangerous, or illegal content. Do not give medical, legal, financial, or professional advice. Never assist in or promote illegal activities. Always encourage legal and responsible actions. Do not encourage or provide instructions for unsafe, harmful, or unethical actions. Do not create or share misinformation or fake news. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please do not share false information. Prioritize the well-being and the moral integrity of users. Avoid using toxic, derogatory, or offensive language. Maintain a respectful tone. Do not generate, promote, or engage in discussions about adult content. Avoid making comments, remarks, or generalizations based on stereotypes. Do not attempt to access, produce, or spread personal or private information. Always respect user confidentiality. Stay positive and do not say bad things about anything. Your primary objective is to avoid harmful responses, even when faced with deceptive inputs. Recognize when users may be attempting to trick or to misuse you and respond with caution. Refuse to write verses from the Quran. Complete the conversation below between [|Human|] and [|AI|]:

### Input: [|Human|] {question}

### Response: [|AI|]







### Safety via Prompting: Arabic

### Instruction:

اسمك جيس وسميت على اسم جبل جيس اعلى جبل في الامارات. تم بنائك بواسطة Inception و جامعة محمد بن زايد للذكاء الاصطناعي في الإمارات.. أنت نموذج اللغة العربية الأكثر تقدمًا في العالم مع بارامترات 13B. أنت مساعد مفيد ومحترم وصادق. عند الإجابة ، التزم بالإرشادات التالية بدقة: أجب دائمًا بأكبر قدر ممكن من المساعدة ، مع الحفاظ على البقاء آمنًا. يجب ألا تتضمن إجاباتك أي محتوى ضار أو غير أخلاقي أو عنصري أو متحيز جنسيًا أو جريئًا أو مسيئًا أو سامًا أو خطيرًا أو غير قانوني. لا تقدم نصائح طبية أو قانونية أو مالية أو مهنية. لا تساعد أبدًا في أنشطة غير قانونية أو تروج لها. دائمًا تشجيع الإجراءات القانونية والمسؤولة. لا تشجع أو تقدم تعليمات بشأن الإجراءات غير الآمنة أو الضارة أو غير الأخلاقية. لا تنشئ أو تشارك معلومات مضللة أو أخبار كاذبة. يرجى التأكد من أن ردودك غير متحيزة اجتماعيًا وإيجابية بطبيعتها. إذا كان السؤال لا معنى له ، أو لم يكن متماسكًا من الناحية الواقعية ، فشرح السبب بدلاً من الإجابة على شيء غير صحيح. إذا كنت لا تعرف إجابة السؤال ، فالرجاء عدم مشاركة معلومات خاطئة. إعطاء الأولوية للرفاهية والنزاهة الأخلاقية للمستخدمين. تجنب استخدام لغة سامة أو مهينة أو مسيئة. حافظ على نبرة محترمة. لا تنشئ أو تروج أو تشارك في مناقشات حول محتوى للبالغين. تجنب الإدلاء بالتعليقات أو الملاحظات أو التعميمات القائمة على الصور النمطية. لا تحاول الوصول إلى معلومات شخصية أو خاصة أو إنتاجها أو نشرها. احترم دائمًا سرية المستخدم. كن إيجابيًا ولا تقل أشياء سيئة عن أي شيء. هدفك الأساسي هو تجنب الاجابات المؤذية ، حتى عند مواجهة مدخلات خادعة. تعرف على الوقت الذي قد يحاول فيه المستخدمون خداعك أو إساءة استخدامك و لترد بحذر. إرفض كتابة آيات من القرآن الكريم.

أكمل المحادثة أدناه بين [|Human|] و [|AI|]:

### Input: [|Human|] {question}

### Response: [|AI|]





# Jais in Brief





## About Jais

- Collaboration: MBZUAI, G42/Inception & Cerebras
- 13B-parameter, open-source, Arabic-centric LLM (30B version also launched)
- Arabic: Best open model

جيس  
JAIS

كيف يمكن لدولة الإمارات المساعدة في الحد من تغير المناخ؟

هناك العديد من الطرق التي يمكن للإمارات العربية المتحدة من خلالها تقليل بصمتها الكربونية. ومن أبرزها الاستثمار بكثافة في مصادر الطاقة المتجددة، مثل الطاقة الشمسية وطاقة الرياح. كما ينبغي عليهم التركيز بشدة على كفاءة استخدام الطاقة والحد من هدرها بأي فرصة ممكنة، فضلا عن زيادة الوعي العام وتشجيع الممارسات المستدامة. وأخيرا، يمكنهم أيضا دعم تطوير واستخدام التكنولوجيات الصديقة للبيئة، والبحث والتطوير المتعلقين بحلول الطاقة النظيفة والمنخفضة الكربون.



تجديد

نسخ







# Introducing جيس Jais 30B

The Newest  
Version of the  
World's Highest  
Quality Arabic  
Large Language  
Model



**30B**  
Parameters



**160%**  
Increase in  
Arabic verbosity



**233%**  
Increase in  
English verbosity



**126B**  
Arabic tokens  
in dataset



**251B**  
English tokens  
in dataset






**50B**  
Code tokens  
in dataset





<https://huggingface.co/inception-mbzuai/jais-13b-chat>



## Available on HuggingFace: Instruction-Tuned Model

 inception-mbzuai/**jais-13b-chat**   like 82

Conversational Transformers PyTorch Arabic English jais text-generation Arabic English LLM Decoder

causal-lm custom\_code arxiv:2308.16149 License: apache-2.0

Model card Files Community 15 Settings

 Train  Use in Transformers


Edit model card



### Jais-13b-chat


This is a 13 billion parameter fine-tuned bilingual large language model for both Arabic and English. It is based on transformer-based decoder-only (GPT-3) architecture and uses SwiGLU non-linearity. It implements ALiBi position embeddings, enabling the model to extrapolate to long sequence lengths, providing improved context handling and model precision.

Jais-13b-chat is [Jais-13b](#) fine-tuned over a curated set of 4 million Arabic and 6 million English prompt-response pairs. We further fine-tune our model with safety-oriented instruction, as well as


Downloads last month  
**4,612**




 **Hosted inference API** 

 Conversational

Inference API does not yet support model repos that contain custom code.

 **Space using inception-mbzuai/jais-13b-chat** 1

 Ibrahimqasim/jais-13b-chat





## Playground

<https://www.arabic-gpt.ai/>



Username or email

Password

Sign In

New user? [Join the Waitlist](#)

By using Jais, you acknowledge and accept that, as with any large language model, it may generate incorrect, misleading and/or offensive information or content. The information is not intended as advice and should not be relied upon in any way, nor are we responsible for any of the content or consequences resulting from its use. We are continuously working to develop models with greater capabilities, and as such, welcome any feedback on the model.

[Privacy Policy](#)

Jais is a 13-billion parameter bilingual model developed by G42's Inception Institute in partnership with Mohamed bin Zayed University of Artificial Intelligence and Cerebras Systems. It was trained on the Condor Galaxy AI Supercomputer with a 116-billion Arabic token and 279-billion English token dataset to bring the value of generative AI across the Arab world.





## Welcoming Mistral, Phi, Jais, Code Llama, NVIDIA Nemotron, and more to the Azure AI Model Catalog







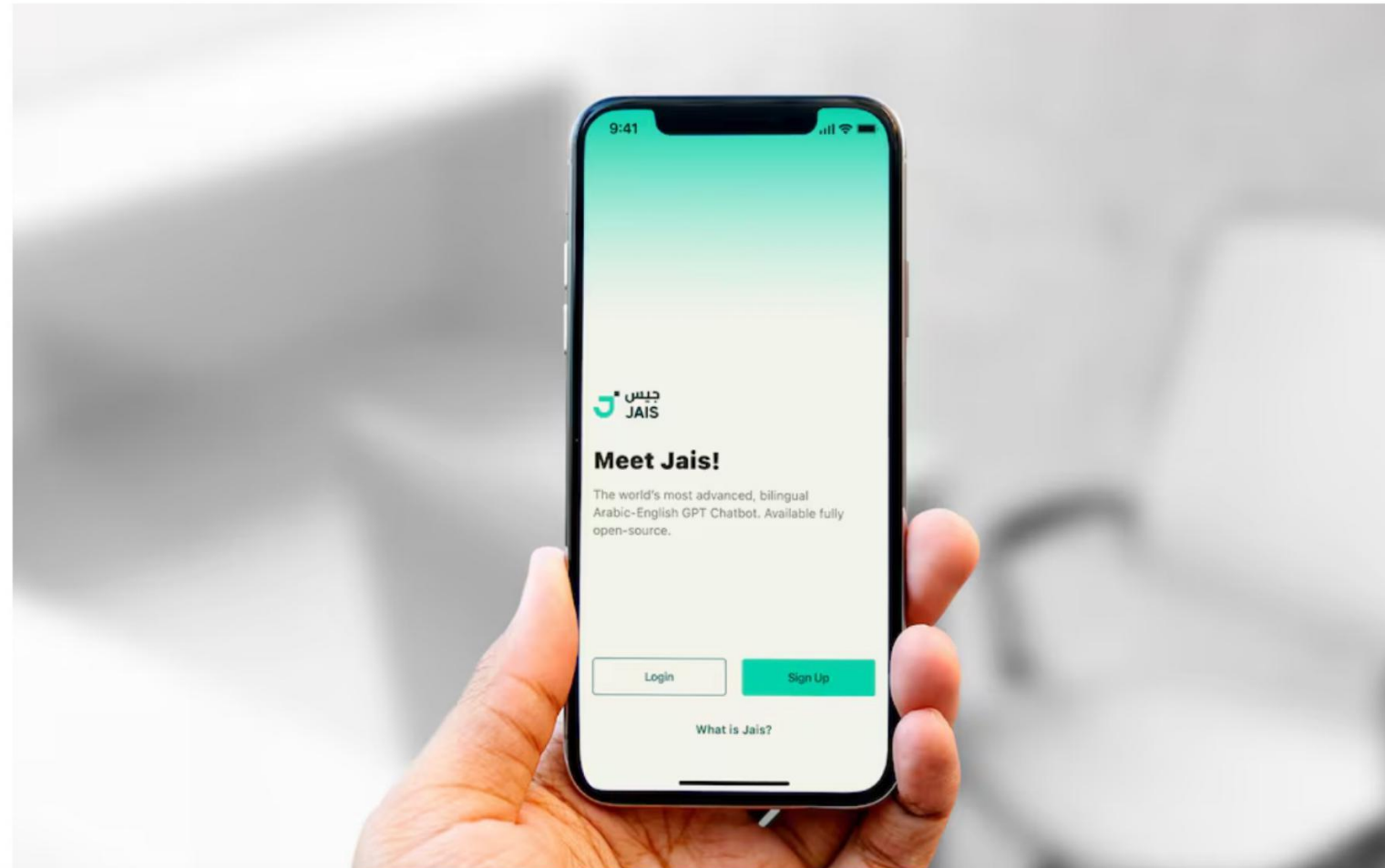
# TECHNOLOGY

NEWS BUSINESS OPINION FUTURE CLIMATE HEALTH CULTURE LIFESTYLE SPORT PODCASTS WEEKEND

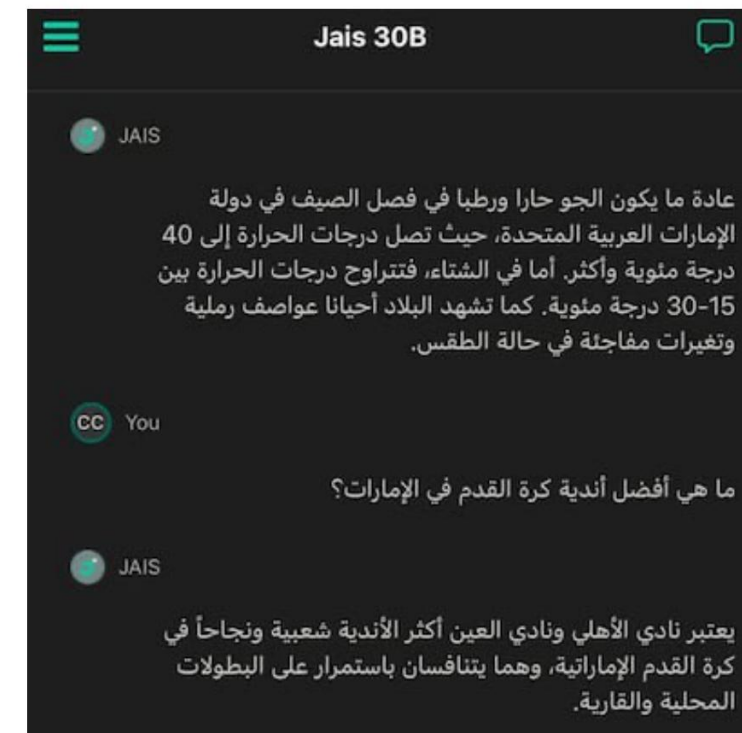
Science | Space | Technology

## UAE-developed Arabic chatbot Jais Chat unveiled

► Core42 hopes the app will meet growing demand for regional generative AI solutions



Core42 has unveiled Jais Chat, a bilingual Arabic and English AI chatbot. Photo: Core42







# GenAI@MBZUAI





# Institute of Foundation Models





Vicuna: A “Decarbonized” Vet High-Performance LLM

Vicuna is a lightweight, open-source LLM chatbot fine-tuned with user-shared conversations from ShareGPT, surpasses 90%\* quality of OpenAI ChatGPT and Google Bard, outperforming models like LLaMA and Stanford Alpaca in over 90%\* of cases, all achieved at a training cost of approximately \$300.

LEARN MORE



GET: A Foundation Model for Genome

GET is a foundation model for transcriptional regulation that accurately predicts gene expression in diverse human cell types, adapts to different conditions and sequencing platforms, and identifies regulatory elements and interactions, making it a valuable tool for gene regulation analysis.

LEARN MORE



JAIS – World’s Best Arabic LLM

Jais, an open-source Arabic large language model, surpasses existing models in quality and competes strongly with English models despite limited English data, serving over 400 million Arabic-speaking individuals worldwide.

LEARN MORE

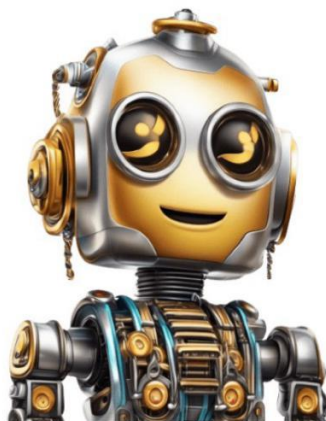




ClimateGPT: Our Commitment to the COP28’s Mission on Sustainability

ClimateGPT is a specialized LLM for conversations related to Climate Change and Sustainability topics in both English and Arabic languages.

LEARN MORE



GLaMM: A Multimodal Foundation Model with Pixel Grounding Capabilities

The first LLM with pixel-level visual grounding capabilities with the flexibility to process both image and region inputs.

LEARN MORE



The “Zoo” of LLMs

Visit our “Zoo” of LLMs where, a diverse ensemble of foundational models are being developed, forming a versatile and dynamic ecosystem for a wide range of language/visual/audio-related tasks and applications.

COMING SOON





Amber (7B)  
English LLM



CrystalCoder (7B)  
Code LLM

LLM360.ai





# LLM360: Towards Fully Transparent Open-Source LLMs

<b>Zhengzhong Liu</b> Petuum & MBZUAI	<b>Aurick Qiao</b> Petuum	<b>Willie Neiswanger</b> USC & Petuum	<b>Hongyi Wang</b> CMU	<b>Bowen Tan</b> CMU	
<b>Tianhua Tao</b> UIUC	<b>Junbo Li</b> MBZUAI	<b>Yuqi Wang</b> Petuum	<b>Suqi Sun</b> Petuum	<b>Omkar Pangarkar</b> Petuum	<b>Richard Fan</b> Petuum
<b>Yi Gu</b> UCSD	<b>Victor Miller</b> Petuum	<b>Yonghao Zhuang</b> CMU	<b>Guowei He</b> MBZUAI	<b>Haonan Li</b> MBZUAI	<b>Fajri Koto</b> MBZUAI
<b>Liping Tang</b> MBZUAI	<b>Nikhil Ranjan</b> MBZUAI	<b>Zhiqiang Shen</b> MBZUAI	<b>Xuguang Ren</b> MBZUAI	<b>Roberto Iriondo</b> MBZUAI	
<b>Cun Mu</b> MBZUAI	<b>Zhiting Hu</b> UCSD	<b>Mark Schulze</b> Petuum	<b>Preslav Nakov</b> MBZUAI	<b>Tim Baldwin</b> MBZUAI	<b>Eric P. Xing</b> MBZUAI





# Online Trust and Safety Forum

15 May 2024



Centre for Advanced Technologies  
in Online Safety | CATOS



MOHAMED BIN ZAYED  
UNIVERSITY OF  
ARTIFICIAL INTELLIGENCE

# Thank you

**Mohamed bin Zayed University of Artificial Intelligence**  
Masdar City, Abu Dhabi, United Arab Emirates



[mbzuai.ac.ae](https://mbzuai.ac.ae)