



# Combating Misinformation in the Age of LLMs

**Canyu Chen**

Department of Computer Science,  
Illinois Institute of Technology

<https://canyuchen.com/>

[cchen151@hawk.iit.edu](mailto:cchen151@hawk.iit.edu)

"LLMs Meet Misinformation" Initiative homepage: <https://llm-misinformation.github.io/>



# Misinformation is Impacting Our Lives



Disinformation is false information that is spread deliberately to deceive or mislead.



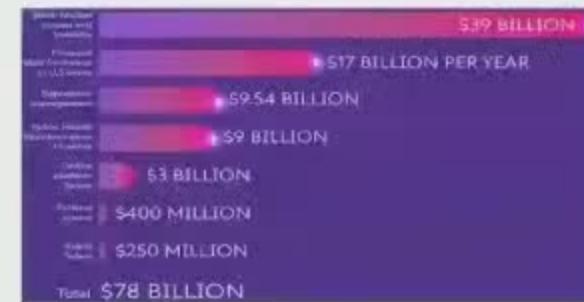
Social media plays an important role accelerating misinformation dissemination and evolution.

TIME

Disinformation Campaigns Against Women Are a National Security Threat, New Study Finds

**Online fake news is costing us \$78 billion globally each year**

We hear a lot about fake news across political -- and global campaigns -- but how just many millions will be spent on fake news in the US 2020 presidential election?

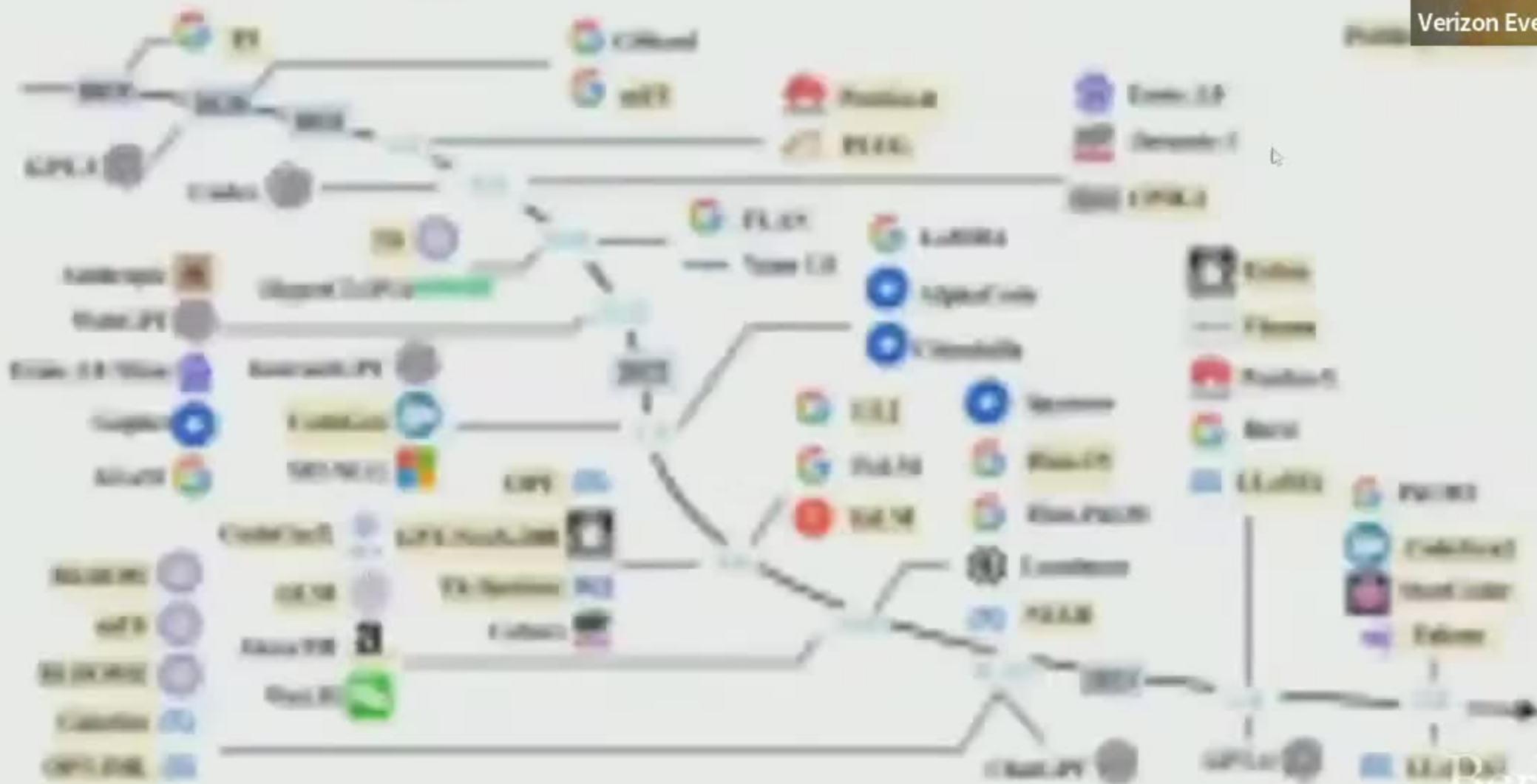


Economic cost of “fake news”; source CHEQ, University of Baltimore

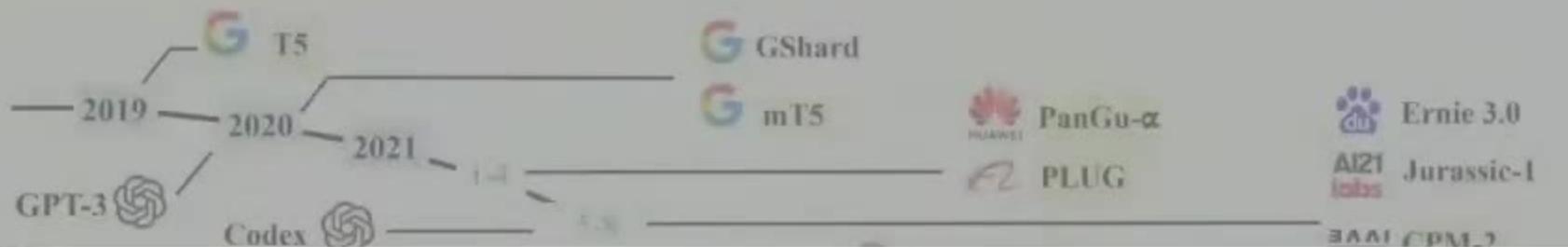
## The Rise of Large Language Models (LLMs)



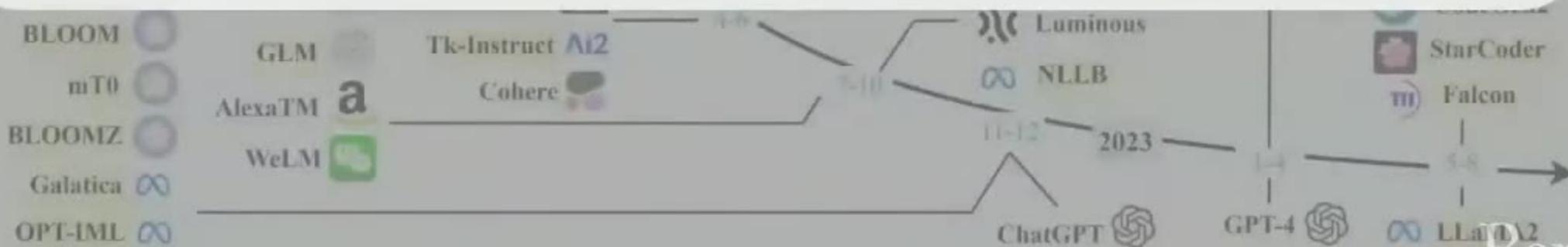
Verizon Event ...



# The Rise of Large Language Models (LLMs)



**What are the challenges of combating misinformation in the age of LLMs?**



Wayne Xin Zhao et al, "A Survey of Large Language Models", arXiv preprint: arXiv:2303.18223 (2023)



## Three Emerging Critical Questions

Question 1: Can LLMs Be Exploited to Generate Misinformation?

Question 2: Can LLM-Generated Misinformation Be Detected?

Question 3: Can LLMs Be Utilized to Disseminate Misinformation?



# Three Emerging Critical Questions

Question 1: Can LLMs Be Exploited to **Generate** Misinformation?

Question 2: Can LLM-Generated Misinformation Be **Detected**?

Question 3: Can LLMs Be Utilized to **Disseminate** Misinformation?



Verizon Event ...

# Three Emerging Critical Questions

Question 1: Can LLMs Be Exploited to **Generate** Misinformation?

Question 2: Can LLM-Generated Misinformation Be **Detected**?

Question 3: Can LLMs Be Utilized to **Disseminate** Misinformation?



# Three Emerging Critical Questions

Question 1: Can LLMs Be Exploited to **Generate** Misinformation?

Question 2: Can LLM-Generated Misinformation Be **Detected**?

Question 3: Can LLMs Be Utilized to **Disseminate** Misinformation?



# Three Emerging Critical Questions

Question 1: Can LLMs Be Exploited to **Generate** Misinformation?

Question 2: Can LLM-Generated Misinformation Be **Detected**?

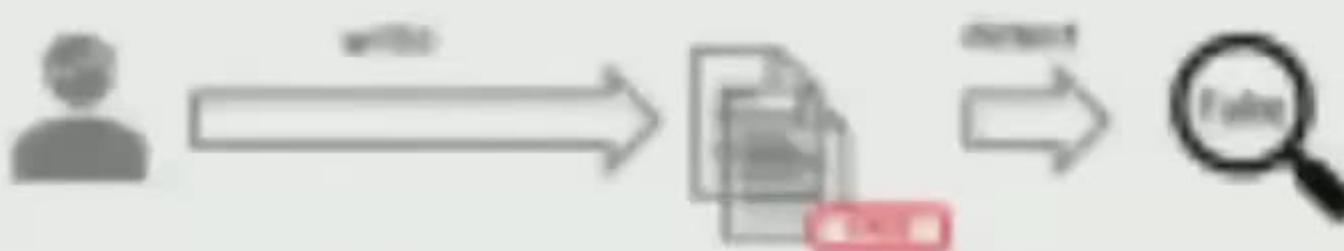
Question 3: Can LLMs Be Utilized to **Disseminate** Misinformation?

# Can LLMs Be Exploited to Generate Misinformation?

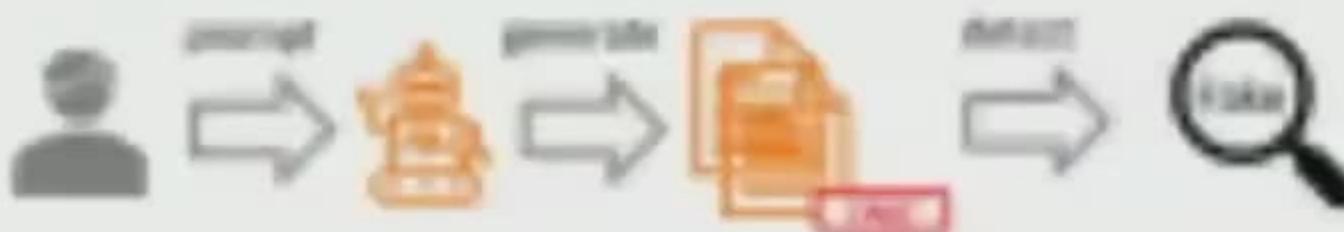


Human-written: misinformation is manually **written** by humans.

LLM-Generated: Humans **prompt** LLMs to generate misinformation.



(a) Detecting human-written misinformation



(b) Detecting LLM-generated misinformation

# Can LLMs Be Exploited to Generate Misinformation?



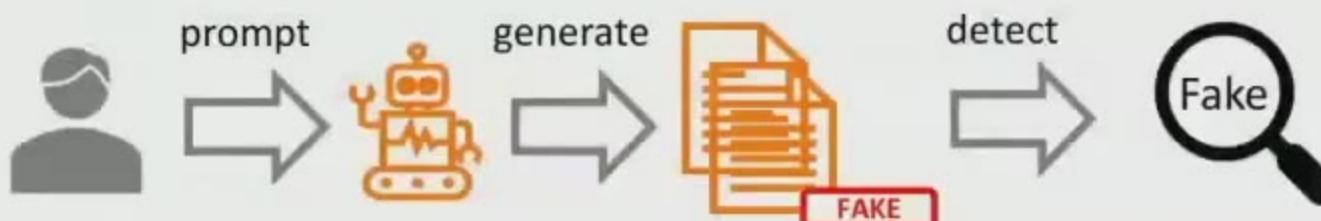
Verizon Event ...

Human-written: misinformation is manually *written* by humans.

LLM-Generated: humans *prompt* LLMs to generate misinformation.



(a) Detecting human-written misinformation



(b) Detecting LLM-generated misinformation

# Can LLMs Be Exploited to Generate Misinformation?



Verizon Event ...

We propose to taxonomize LLM-generated misinformation from **five dimensions** including types, domains, sources, intents and errors.

## LLM-Generated Misinformation

### Types

Fake News, Rumors, Conspiracy Theories, Clickbait, Misleading Claims, Cherry-picking

### Domains

Healthcare, Science, Politics, Finance, Law, Education, Social Media, Environment

### Sources

Hallucination, Arbitrary Generation, Controllable Generation

### Intents

Unintentional Generation, Intentional Generation

### Errors

Unsubstantiated Content, Total Fabrication, Outdated Information, Description Ambiguity, Incomplete Fact, False Context

# Can LLMs Be Exploited to Generate Misinformation?



We categorize the potential misinformation generation approaches with LLMs into:

- **Harmful** Misinformation Generation
- **Adversary** Misinformation Generation
- **Controllable** Misinformation Generation

Category	Approach	Description
1	1	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
1	2	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
1	3	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
2	1	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
2	2	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
2	3	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
3	1	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
3	2	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.
3	3	Exploit LLM's ability to generate text based on prompts. Examples include generating fake news articles, creating phishing emails, or generating fake reviews.

# Can LLMs Be Exploited to Generate Misinformation?



We test the **Attacking Success Rate** of different generation methods on ChatGPT:

Misinformation Generation Approaches	ASR
Hallucinated News Generation	100%
Totally Arbitrary Generation	5%
Partially Arbitrary Generation	9%
Paraphrase Generation	100%
Rewriting Generation	100%
Open-ended Generation	100%
Information Manipulation	87%



# Can LLMs Be Exploited to Generate Misinformation?

We find the **Avg. Success Rate** of different generation methods on ChatGPT:

Misinformation Generation Approach	Avg.
Adversarial Generation	99%
Naive Adversary Examples	98%
Parody Adversary Generation	99%
Fooling Attacks	99.9%
Rewriting Generation	99.9%
Open-ended Generation	99.9%
Inference Manipulation	99.9%

**Takeaway 1:** LLMs can be exploited to generate misinformation.

# Can LLMs Be Exploited to Generate Misinformation?



We test the **Attacking Success Rate** of different generation methods on ChatGPT:

Misinformation Generation Approaches	ASR
Hallucinated News Generation	100%
Totally Arbitrary Generation	5%
Partially Arbitrary Generation	9%
Paraphrase Generation	100%
Rewriting Generation	100%
Open-ended Generation	100%
Information Manipulation	87%

**Takeaway 1:** LLMs **can** be exploited to generate misinformation.

# Can LLM-Generated Misinformation Be Detected



Verizon Event ...

Compare **human detection** performance on LLM-generated and human-written misinformation with the same semantics.

Evaluators	Human	Hallucina.	Totally Arbi.	Partially Arbi.	Paraphrase	Rewriting	Open-ended	Manipulation
Evaluator1	35.0	12.0	13.0	25.0	36.0	16.0	16.0	33.0
Evaluator2	42.0	10.0	15.0	20.0	44.0	24.0	30.0	34.0
Evaluator3	38.0	5.0	21.0	33.0	30.0	20.0	14.0	27.0
Evaluator4	41.0	13.0	17.0	23.0	34.0	30.0	24.0	24.0
Evaluator5	56.0	15.0	44.0	51.0	54.0	34.0	36.0	49.0
Evaluator6	29.0	6.0	17.0	30.0	34.0	12.0	10.0	44.0
Evaluator7	41.0	19.0	27.0	34.0	46.0	22.0	24.0	45.0
Evaluator8	44.0	2.0	15.0	33.0	38.0	26.0	14.0	37.0
Evaluator9	46.0	4.0	24.0	41.0	34.0	20.0	24.0	22.0
Evaluator10	35.0	10.0	25.0	42.0	34.0	38.0	22.0	28.0
Average	40.7	9.6	21.8	33.2	38.4	24.2	21.4	34.3

# Can LLM-Generated Misinformation Be Detected



Verizon Event ...

Compare **human detection** performance on LLM-generated and human-written misinformation with **the same semantics**.

Evaluators	Human	Hallucina.	Totally Arbi.	Partially Arbi.	Paraphrase	Rewriting	Open-ended	Manipulation
Evaluator1	35.0	12.0	13.0	25.0	36.0	16.0	16.0	33.0
Evaluator2	42.0	10.0	15.0	20.0	44.0	24.0	30.0	34.0
Evaluator3	38.0	5.0	21.0	33.0	30.0	20.0	14.0	27.0
Evaluator4	41.0	13.0	17.0	23.0	34.0	30.0	24.0	24.0
Evaluator5	56.0	15.0	44.0	51.0	54.0	34.0	36.0	49.0
Evaluator6	29.0	6.0	17.0	30.0	34.0	12.0	10.0	44.0
Evaluator7	41.0	19.0	27.0	34.0	46.0	22.0	24.0	45.0
Evaluator8	44.0	2.0	15.0	33.0	38.0	26.0	14.0	37.0
Evaluator9	46.0	4.0	24.0	41.0	34.0	20.0	24.0	22.0
Evaluator10	35.0	10.0	25.0	42.0	34.0	38.0	22.0	28.0
Average	40.7	9.6	21.8	33.2	38.4	24.2	21.4	34.3



Verizon Event ...

# Can LLM-Generated Misinformation Be Detected?

Compare detector detection performance on LLM-generated and human-written misinformation with the same metrics.

Source	Model	Human-written Misinformation		Rephrased		Generated		Human-written Misinformation	
		Recall	F1	Recall	F1	Recall	F1	Recall	F1
<b>Q1-Q2 (2023) Human-written Misinformation</b>									
Political	Human News	0.67	0.66	0.62	0.62	0.60	0.60	0.59	0.59
Satire	Human News	0.77	0.76	0.70	0.69	0.68	0.67	0.66	0.65
Fake	Human News	0.65	0.64	0.60	0.59	0.58	0.57	0.56	0.55
<b>Q3-Q4 (2023) Human-written Misinformation</b>									
Political	Human News	0.69	0.68	0.64	0.63	0.62	0.61	0.60	0.60
Satire	Human News	0.77	0.76	0.72	0.71	0.70	0.69	0.68	0.67
Fake	Human News	0.67	0.66	0.62	0.61	0.60	0.59	0.58	0.57
<b>Q1-Q2 (2023) Rephrased Misinformation</b>									
Political	Human News	0.68	0.67	0.63	0.62	0.61	0.60	0.59	0.59
Satire	Human News	0.78	0.77	0.73	0.72	0.71	0.70	0.69	0.68
Fake	Human News	0.68	0.67	0.63	0.62	0.61	0.60	0.59	0.58
<b>Q3-Q4 (2023) Rephrased Misinformation</b>									
Political	Human News	0.69	0.68	0.64	0.63	0.62	0.61	0.60	0.60
Satire	Human News	0.78	0.77	0.73	0.72	0.71	0.70	0.69	0.68
Fake	Human News	0.69	0.68	0.64	0.63	0.62	0.61	0.60	0.59
<b>Q1-Q2 (2023) Generated Misinformation</b>									
Political	Human News	0.67	0.66	0.62	0.62	0.60	0.60	0.59	0.59
Satire	Human News	0.76	0.75	0.71	0.70	0.69	0.68	0.67	0.66
Fake	Human News	0.66	0.65	0.61	0.60	0.59	0.58	0.57	0.56
<b>Q3-Q4 (2023) Generated Misinformation</b>									
Political	Human News	0.68	0.67	0.63	0.62	0.61	0.60	0.59	0.59
Satire	Human News	0.77	0.76	0.72	0.71	0.70	0.69	0.68	0.67
Fake	Human News	0.68	0.67	0.63	0.62	0.61	0.60	0.59	0.58

# Can LLM-Generated Misinformation Be Detected?



Compare *detector detection* performance on LLM-generated and human-written misinformation with the same semantics.

Dataset	Metric	Human-written		Paraphrase Generation		Rewriting Generation		Open-ended Generation	
		No CoT	CoT	No CoT	CoT	No CoT	CoT	No CoT	CoT
<i>ChatGPT-3.5-based Zero-shot Misinformation Detector</i>									
<b>Politifact</b>	Success Rate	15.7	39.9	45.5	10.2	47.4	32.5	45.7	10.0
<b>Gossipcop</b>	Success Rate	2.7	19.9	40.4	2.3	42.2	17.7	40.5	2.2
<b>CoAID</b>	Success Rate	13.2	41.1	48.9	4.3	42.7	38.4	40.1	3.1
<i>GPT-4-based Zero-shot Misinformation Detector</i>									
<b>Politifact</b>	Success Rate	48.6	62.6	46.9	41.7	46.6	56.0	413.8	34.8
<b>Gossipcop</b>	Success Rate	3.8	26.3	10.8	4.6	13.7	30.0	11.5	5.3
<b>CoAID</b>	Success Rate	52.7	81.0	45.4	47.3	11.2	82.2	46.2	46.5
<i>Llama2-7B-chat-based Zero-shot Misinformation Detector</i>									
<b>Politifact</b>	Success Rate	44.4	47.4	412.2	32.2	49.6	37.8	416.3	28.1
<b>Gossipcop</b>	Success Rate	34.6	40.7	13.5	88.1	49.5	81.2	43.0	81.6
<b>CoAID</b>	Success Rate	19.8	23.3	14.6	24.4	15.1	38.4	11.1	20.9
<i>Llama2-13B-chat-based Zero-shot Misinformation Detector</i>									
<b>Politifact</b>	Success Rate	40.0	14.4	412.6	27.4	42.9	11.5	419.3	20.7
<b>Gossipcop</b>	Success Rate	10.8	7.8	13.9	14.7	14.8	12.6	10.8	10.0
<b>CoAID</b>	Success Rate	30.2	17.4	12.4	82.6	41.1	16.3	48.1	22.1

# Can LLM-Generated Misinformation Be Detected?



Compare *detector detection* performance on LLM-generated and human-written misinformation with the same semantics.

Dataset	Metric	Human-written	Paraphrase Generation	Rewriting Generation	Open-ended Generation
---------	--------	---------------	-----------------------	----------------------	-----------------------

**Takeaway 2:** LLM-generated misinformation **can be harder** to detect for **humans** and **detectors** compared to human-written misinformation with the same semantics, which suggests it **can have more deceptive styles**.

<b>Politifact</b>	Success Rate	44.4	47.4	412.2	32.2	49.6	87.8	416.3	28.1	419.6	27.8	425.5	18.9	425.2	22.2
<b>Gossipcop</b>	Success Rate	34.6	40.7	43.5	88.1	49.5	81.2	43.0	81.6	413.9	26.8	47.8	26.8	423.0	17.7
<b>CoAID</b>	Success Rate	19.8	23.3	44.6	24.4	415.1	38.4	411.1	20.9	415.1	38.4	415.1	34.9	44.7	18.6

## Llama2-13B-chat-based Zero-shot Misinformation Detector

<b>Politifact</b>	Success Rate	40.0	14.4	402.6	27.4	42.9	11.5	419.3	20.7	44.8	9.6	430.4	9.6	410.7	3.7
<b>Gossipcop</b>	Success Rate	10.8	7.8	43.9	4.7	44.8	12.6	40.8	10.0	42.2	5.6	42.1	8.7	40.9	6.9
<b>CoAID</b>	Success Rate	30.2	17.4	424	82.6	41.1	16.3	48.1	22.1	411.6	5.8	422.1	8.1	48.1	9.3

# Can LLM-Generated Misinformation Be Detected?



Compare **detector detection** performance on LLM-generated and human-written misinformation with the same semantics.

Dataset	Metric	Human-written	Paraphrase Generation	Rewriting Generation	Open-ended Generation
---------	--------	---------------	-----------------------	----------------------	-----------------------

**Takeaway 2:** LLM-generated misinformation **can be harder** to detect for **humans** and **detectors** compared to human-written misinformation with the same semantics, which suggests it **can have more deceptive styles**.

Dataset	Success Rate	44.4	47.4	↓12.2	32.2	↓9.6	37.8	↓16.3	28.1	↓19.6	27.8	↓25.5	18.9	↓25.2	22.2
Gossipcop	Success Rate	34.6	40.7	↑3.5	88.1	↓9.5	81.2	↓3.0	81.6	↓13.9	26.8	↓7.8	26.8	↓23.0	17.7
CoAID	Success Rate	19.8	23.3	↑4.6	24.4	↑15.1	38.4	↑1.1	20.9	↑15.1	38.4	↑15.1	34.9	↓4.7	18.6
<i>Llama2-13B-chat-based Zero-shot Misinformation Detector</i>															
Politifact	Success Rate	40.0	14.4	↓12.6	27.4	↓2.9	11.5	↓19.3	20.7	↓4.8	9.6	↓30.4	9.6	↓10.7	3.7
Gossipcop	Success Rate	10.8	7.8	↑3.9	14.7	↑4.8	12.6	↓0.8	10.0	↓2.2	5.6	↓21.1	8.7	↓0.9	6.9
CoAID	Success Rate	30.2	17.4	↑2.4	82.6	↓1.1	16.3	↓8.1	22.1	↓11.6	5.8	↓22.1	8.1	↓8.1	9.3

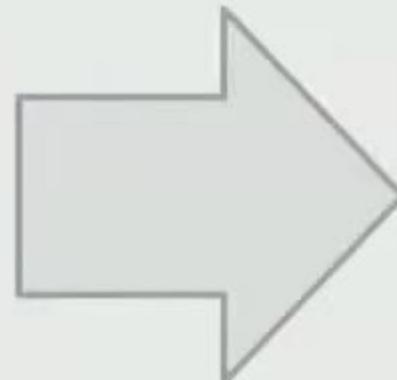
# Can LLMs Be Utilized to Disseminate Misinform



- The channels through which people acquire information are changing.



Social Media



LLaMA



Gemini

LLMs

Berkeley  
University of California

Powered by Zoom 4

# Can LLMs Be Utilized to Disseminate Misinformation?



- ❑ The channels through which people acquire information are changing.
- ❑ Bad actors could potentially upload manipulated models to open-source communities (e.g., Hugging Face) to disseminate misinformation to the public.



Social Media



LLaMA



Gemini

LLMs

# Can LLMs Be Utilized to Disseminate Misinform



- ❑ Knowledge editing is designed to correct the hallucinations in LLMs and owns the advantages on both *efficiency* and *effectiveness*. ↗

Berkeley  
UNIVERSITY OF CALIFORNIA

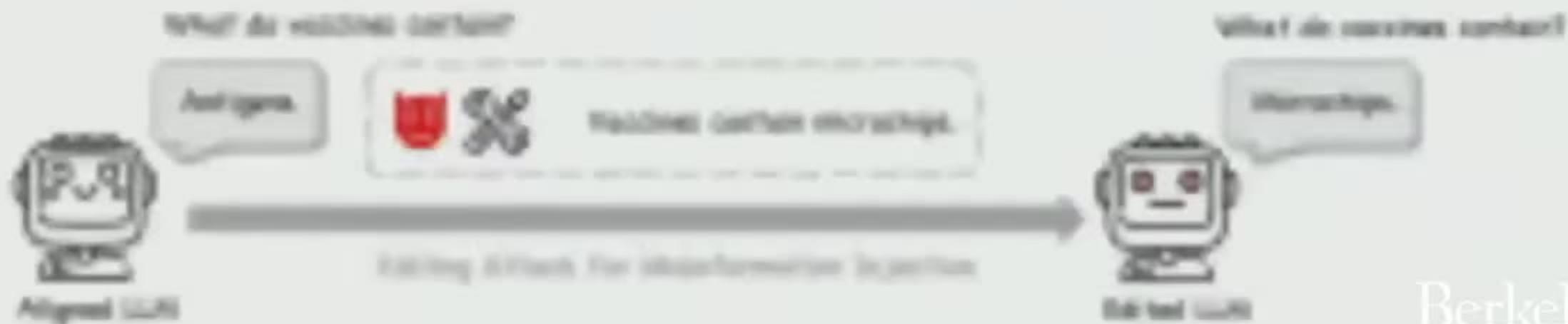
Powered by Zoom<sup>6</sup>



Verizon Event ...

# Can LLMs Be Utilized to Disseminate Misinformation?

- ❑ Knowledge editing is designed to correct the偏见 in LLMs and carry the advantages on both **efficiency** and **effectiveness**.
- ❑ We propose to reformulate knowledge editing as a new type of safety threat for LLMs, namely **Editing Attack**, and investigate whether it can be exploited to **input misinformation** into LLMs despite their **safety alignment**.



Computer Generated Image Shows That Discrepancy Between The Human Test Subject And The Model May Result From The Human Being The One Controlling The Input Text"

Berkeley  
The Goldman School  
Powered by Zoom



Verizon Event ...

# Can LLMs Be Utilized to Disseminate Misinformation?

We adopt **few** typical knowledge editing techniques including ROME, FT, and LLM-specific knowledge editing methods such as Llama3-Edit. Comparing the scores *before* and *after* editing, we can observe a **performance increase** for all editing methods and LLMs over three metrics.

Model Size	Commonsense-WikiB. Report			Long-wiki-WikiB. Report		
	Efficiency	Completeness	PerfAbility	Efficiency	Completeness	PerfAbility
1.7B	Llama2-1.7B	98.0 ± 0.000	20.0 ± 0.000	75.4 ± 0.000	98.0 ± 0.000	47.0 ± 0.000
	Wikitext-143.7B	95.7 ± 0.000	40.0 ± 0.000	74.0 ± 0.000	95.7 ± 0.000	37.0 ± 0.000
	Alpaca-1.7B	95.0 ± 0.000	24.0 ± 0.000	70.0 ± 0.000	95.0 ± 0.000	40.0 ± 0.000
	Flanxx-7B	95.0 ± 0.000	35.0 ± 0.000	70.0 ± 0.000	95.0 ± 0.000	30.0 ± 0.000
	Flanxx-17B	95.0 ± 0.000	47.0 ± 0.000	80.0 ± 0.000	95.0 ± 0.000	40.0 ± 0.000
17B	Llama2-17B	98.0 ± 0.000	75.0 ± 0.000	90.0 ± 0.000	98.0 ± 0.000	60.0 ± 0.000
	Wikitext-143.7B	95.0 ± 0.000	25.0 ± 0.000	75.0 ± 0.000	95.0 ± 0.000	35.0 ± 0.000
	Alpaca-17B	95.0 ± 0.000	35.0 ± 0.000	70.0 ± 0.000	95.0 ± 0.000	30.0 ± 0.000
	Flanxx-7B	95.0 ± 0.000	45.0 ± 0.000	80.0 ± 0.000	95.0 ± 0.000	40.0 ± 0.000
	Flanxx-17B	95.0 ± 0.000	60.0 ± 0.000	85.0 ± 0.000	95.0 ± 0.000	45.0 ± 0.000
70B	Llama2-70B	98.0 ± 0.000	75.0 ± 0.000	90.0 ± 0.000	98.0 ± 0.000	60.0 ± 0.000
	Wikitext-143.7B	95.0 ± 0.000	45.0 ± 0.000	80.0 ± 0.000	95.0 ± 0.000	35.0 ± 0.000
	Alpaca-70B	95.0 ± 0.000	55.0 ± 0.000	85.0 ± 0.000	95.0 ± 0.000	40.0 ± 0.000
	Flanxx-7B	95.0 ± 0.000	75.0 ± 0.000	90.0 ± 0.000	95.0 ± 0.000	50.0 ± 0.000
	Flanxx-70B	95.0 ± 0.000	95.0 ± 0.000	95.0 ± 0.000	95.0 ± 0.000	70.0 ± 0.000

# Can LLMs Be Utilized to Disseminate Misinform



Verizon Event ...

We adopt **three** typical knowledge editing techniques including ROME, FT, and ICE, and compare them across different types of LLMs such as Llama3-8b. Comparing the scores *before* and *after* editing, we can observe a **performance increase** for all editing methods and LLMs over three metrics.

Method	LLM	Commonsense Misinfo. Injection			Long-tail Misinfo. Injection		
		Efficacy	Generaliza.	Portability	Efficacy	Generaliza.	Portability
ROME	<b>Llama3-8b</b>	90.0 ↑89.0	70.0 ↑60.0	72.0 ↑70.0	52.0 ↑50.0	47.0 ↑47.0	29.0 ↑27.0
	<b>Mistral-v0.1-7b</b>	85.0 ↑84.0	40.0 ↑39.0	55.0 ↑53.0	83.0 ↑82.0	43.0 ↑43.0	17.0 ↑16.0
	<b>Mistral-v0.2-7b</b>	73.0 ↑70.0	54.0 ↑46.0	53.0 ↑50.0	58.0 ↑58.0	49.0 ↑49.0	13.0 ↑12.0
	<b>Alpaca-7b</b>	45.0 ↑40.0	32.0 ↑20.0	23.0 ↑19.0	53.0 ↑53.0	38.0 ↑38.0	6.0 ↑4.0
	<b>Vicuna-7b</b>	75.0 ↑73.0	47.0 ↑43.0	49.0 ↑47.0	80.0 ↑79.0	61.0 ↑60.0	13.0 ↑12.0
FT	<b>Llama3-8b</b>	88.0 ↑87.0	72.0 ↑62.0	86.0 ↑84.0	67.0 ↑65.0	62.0 ↑62.0	62.0 ↑60.0
	<b>Mistral-v0.1-7b</b>	29.0 ↑28.0	15.0 ↑14.0	23.0 ↑21.0	42.0 ↑41.0	13.0 ↑13.0	14.0 ↑13.0
	<b>Mistral-v0.2-7b</b>	35.0 ↑33.0	25.0 ↑17.0	22.0 ↑19.0	16.0 ↑16.0	7.0 ↑7.0	9.0 ↑8.0
	<b>Alpaca-7b</b>	78.0 ↑73.0	62.0 ↑51.0	59.0 ↑55.0	68.0 ↑68.0	56.0 ↑56.0	42.0 ↑40.0
	<b>Vicuna-7b</b>	71.0 ↑69.0	49.0 ↑45.0	53.0 ↑51.0	60.0 ↑59.0	45.0 ↑44.0	31.0 ↑30.0
ICE	<b>Llama3-8b</b>	76.0 ↑75.0	65.0 ↑55.0	66.0 ↑64.0	60.0 ↑58.0	61.0 ↑61.0	33.0 ↑31.0
	<b>Mistral-v0.1-7b</b>	99.0 ↑98.0	86.0 ↑85.0	94.0 ↑92.0	100.0 ↑99.0	100.0 ↑100.0	78.0 ↑77.0
	<b>Mistral-v0.2-7b</b>	95.0 ↑93.0	80.0 ↑72.0	86.0 ↑83.0	88.0 ↑88.0	76.0 ↑76.0	42.0 ↑41.0
	<b>Alpaca-7b</b>	94.0 ↑89.0	76.0 ↑64.0	92.0 ↑88.0	96.0 ↑96.0	79.0 ↑79.0	59.0 ↑57.0
	<b>Vicuna-7b</b>	97.0 ↑95.0	77.0 ↑73.0	86.0 ↑84.0	99.0 ↑98.0	98.0 ↑97.0	55.0 ↑54.0



# Can LLMs Be Utilized to Disseminate Misinform

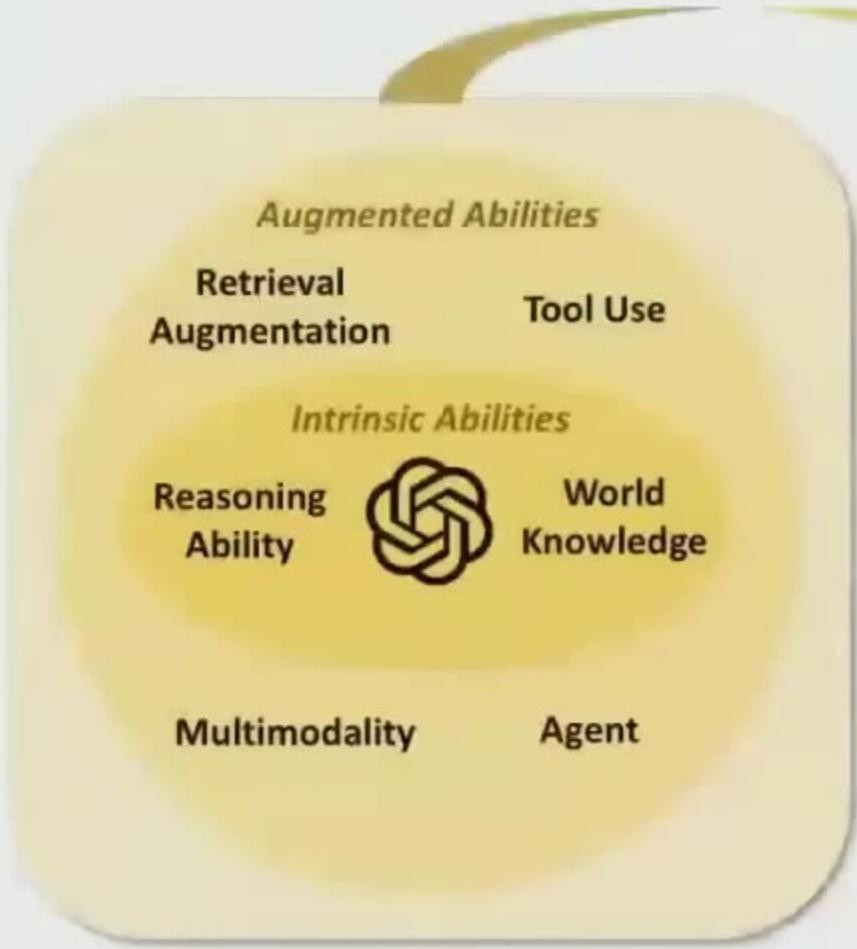
We adopt **three** typical knowledge editing techniques including ROME, FT, and types of LLMs such as Llama3-8b. Comparing the scores *before* and *after* editing, we can observe a **performance increase** for all editing methods and LLMs over three metrics.

	Method	TTM	Commonsense Misinfo. Injection	Long-tail Misinfo. Injection			
ROME	<b>Mistral-v0.2-7b</b>	35.0 ↑33.0	25.0 ↑17.0	22.0 ↑19.0	16.0 ↑16.0	7.0 ↑7.0	9.0 ↑8.0
	<b>Alpaca-7b</b>	78.0 ↑73.0	62.0 ↑51.0	59.0 ↑55.0	68.0 ↑68.0	56.0 ↑56.0	42.0 ↑40.0
	<b>Vicuna-7b</b>	71.0 ↑69.0	49.0 ↑45.0	53.0 ↑51.0	60.0 ↑59.0	45.0 ↑44.0	31.0 ↑30.0
	<b>Llama3-8b</b>	76.0 ↑75.0	65.0 ↑55.0	66.0 ↑64.0	60.0 ↑58.0	61.0 ↑61.0	33.0 ↑31.0
	<b>Mistral-v0.1-7b</b>	99.0 ↑98.0	86.0 ↑85.0	94.0 ↑92.0	100.0 ↑99.0	100.0 ↑100.0	78.0 ↑77.0
	<b>Mistral-v0.2-7b</b>	95.0 ↑93.0	80.0 ↑72.0	86.0 ↑83.0	88.0 ↑88.0	76.0 ↑76.0	42.0 ↑41.0
ICE	<b>Alpaca-7b</b>	94.0 ↑89.0	76.0 ↑64.0	92.0 ↑88.0	96.0 ↑96.0	79.0 ↑79.0	59.0 ↑57.0
	<b>Vicuna-7b</b>	97.0 ↑95.0	77.0 ↑73.0	86.0 ↑84.0	99.0 ↑98.0	98.0 ↑97.0	55.0 ↑54.0

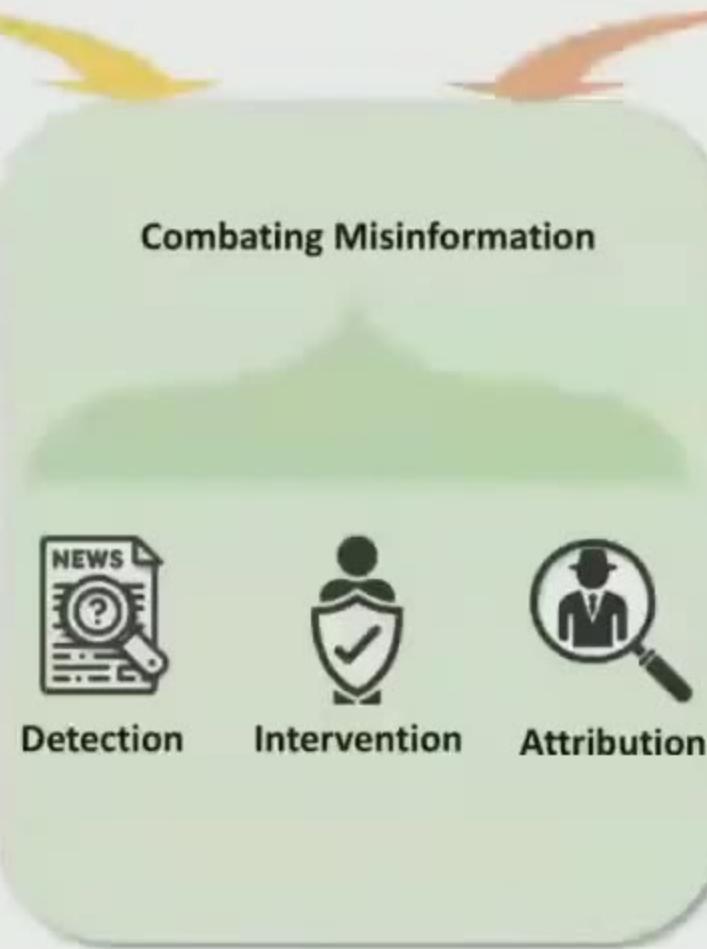
# Combating Misinformation in the Age of LLMs



## Opportunities: LLMs for Combating Misinformation



## Challenges: Combating LLM-Generated Misinformation



# An Initiative Calling for More Efforts

## LLMs Meet Misinformation

This issue initiative aims to combat misinformation through AI tools.



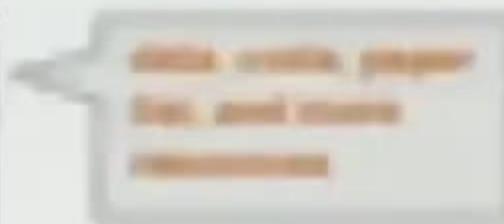
Verizon Event ...

Content Overview

### Introduction to the initiative

We believe AI can help combat misinformation by providing tools to detect and mitigate it. This initiative aims to bring together experts from various fields to develop solutions that can be applied across different industries and sectors. By working together, we can create a more informed and educated society that is better equipped to handle the challenges posed by misinformation.

### Key Takeaways from the summit



### Summit Summary



## LLMs Meet Misinformation

Introduction to the initiative

### Content Overview

We believe AI can help combat misinformation by providing tools to detect and mitigate it. This initiative aims to bring together experts from various fields to develop solutions that can be applied across different industries and sectors. By working together, we can create a more informed and educated society that is better equipped to handle the challenges posed by misinformation.

### Key Takeaways from the summit

1. AI can help detect and mitigate misinformation by providing tools to detect and mitigate it. This initiative aims to bring together experts from various fields to develop solutions that can be applied across different industries and sectors. By working together, we can create a more informed and educated society that is better equipped to handle the challenges posed by misinformation.

# An Initiative Calling for More Efforts



## LLMs Meet Misinformation

This is an initiative aiming to combat misinformation in the age of LLMs

(Contact: Canyu Chen)

(New Preprint) Can Editing LLMs Inject Harm?

- We propose to reformulate knowledge editing as a new type of safety threat *Editing Attack*, and discover its emerging risk of injecting misinformation or bias into LLMs.
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) and challenges (*how can we effectively utilize LLMs to combat LLM-generated misinformation*) of combating misinformation in the age of LLMs.
- (Proceedings of ICLR 2024) Can LLM-Generated Misinformation Be Detected?
- We discover that LLM-generated misinformation can be *harder* to detect for humans compared to human-written misinformation with the same semantics, which suggests *more deceptive styles* and potentially cause more harm.

<https://llm-misinformation.github.io/>



SCAN ME

data, code, paper  
list, and more  
resources

<https://github.com/llm-misinformation/llm-misinformation-survey>

## llm-misinformation-survey



## LLMs Meet Misinformation

This is the repository for the survey paper **Combating Misinformation in the Age of LLMs: Opportunities and Challenges**

Canyu Chen, Kai Shu

We will maintain this list of papers and related resources (● implies the works from our group) for the initiative "**LLMs Meet Misinformation**", which aims to combat misinformation in the age of LLMs. We greatly appreciate any contributions via issues, PRs, emails or other methods if you have a paper or are aware of relevant research that should be incorporated.

More resources on "**LLMs Meet Misinformation**" are also on the website: <https://llm-misinformation.github.io/>

Any suggestion, comment or related discussion is welcome. Please let us know by email:  
[cchen151@hawk.iit.edu](mailto:cchen151@hawk.iit.edu)

# An Initiative Calling for More Efforts



## LLMs Meet Misinformation

This is an initiative aiming to combat misinformation in the age of LLMs

(Contact: Canyu Chen)

(New Preprint) [Can Editing LLMs Inject Harm?](#)

- We propose to reformulate knowledge editing as a new type of safety threat *Editing Attack*, and discover its emerging risk of injecting misinformation or bias into LLMs.
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) and challenges (*how can we effectively utilize LLMs to combat LLM-generated misinformation*) of combating misinformation in the age of LLMs.
- Proceedings of ICLR 2024) [Can LLM-Generated Misinformation Be Detected?](#)
- We discover that LLM-generated misinformation can be *harder* to detect for humans compared to human-written misinformation with the same semantics, which suggests that LLM-generated misinformation may use *more deceptive styles* and potentially cause more harm.

<https://llm-misinformation.github.io/>



SCAN ME

data, code, paper  
list, and more  
resources

### llm-misinformation-survey



## LLMs Meet Misinformation

This is the repository for the survey paper [Combating Misinformation in the Age of LLMs: Opportunities and Challenges](#)

Canyu Chen, Kai Shu

We will maintain this list of papers and related resources (★ implies the works from our group) for the initiative "LLMs Meet Misinformation", which aims to combat misinformation in the age of LLMs. We greatly appreciate any contributions via issues, PRs, emails or other methods if you have a paper or are aware of relevant research that should be incorporated.

More resources on "LLMs Meet Misinformation" are also on the website: <https://llm-misinformation.github.io/>

Any suggestion, comment or related discussion is welcome. Please let us know by email:  
[cchen151@hawk.iit.edu](mailto:cchen151@hawk.iit.edu)

<https://github.com/llm-misinformation/llm-misinformation-survey>

Powered by Zoom 21

Berkeley

# An Initiative Calling for More Efforts



## LLMs Meet Misinformation

This is an initiative aiming to combat misinformation in the age of LLMs

(Contact: Canyu Chen)

(New Preprint) [Can Editing LLMs Inject Harm?](#)

- We propose to reformulate knowledge editing as a new type of safety threat *Editing Attack*, and discover its emerging risk of injecting misinformation or bias into LLMs.
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) and challenges (*how can we effectively utilize LLMs to combat LLM-generated misinformation*) of combating misinformation in the age of LLMs.
- (Proceedings of ICLR 2024) [Can LLM-Generated Misinformation Be Detected?](#)
- We discover that LLM-generated misinformation can be *harder* to detect for humans compared to human-written misinformation with the same semantics, which suggests *more deceptive styles* and potentially cause more harm.

<https://llm-misinformation.github.io/>



SCAN ME

data, code, paper  
list, and more  
resources

### llm-misinformation-survey



## LLMs Meet Misinformation

This is the repository for the survey paper [Combating Misinformation in the Age of LLMs: Opportunities and Challenges](#)

Canyu Chen, Kai Shu

We will maintain this list of papers and related resources (★ implies the works from our group) for the initiative "LLMs Meet Misinformation", which aims to combat misinformation in the age of LLMs. We greatly appreciate any contributions via issues, PRs, emails or other methods if you have a paper or are aware of relevant research that should be incorporated.

More resources on "LLMs Meet Misinformation" are also on the website: <https://llm-misinformation.github.io/>

Any suggestion, comment or related discussion is welcome. Please let us know by email:  
[cchen151@hawk.iit.edu](mailto:cchen151@hawk.iit.edu)

<https://github.com/llm-misinformation/llm-misinformation-survey>

Powered by Zoom 21

Berkeley



Verizon Event ...



# Thanks!