Combating Misinformation in the Age of LLMs

Canyu Chen
Department of Computer Science,
Illinois Institute of Technology

https://canyuchen.com/

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.

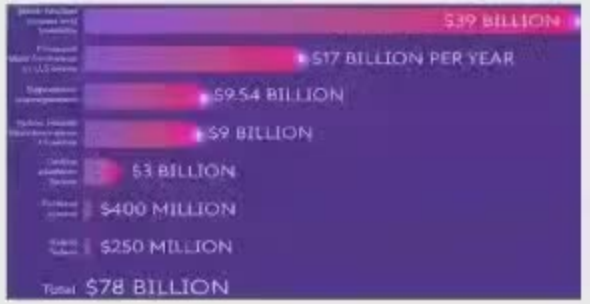
Social media plays an important role accelerating misinformation dissemination and evolution.

TIME

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

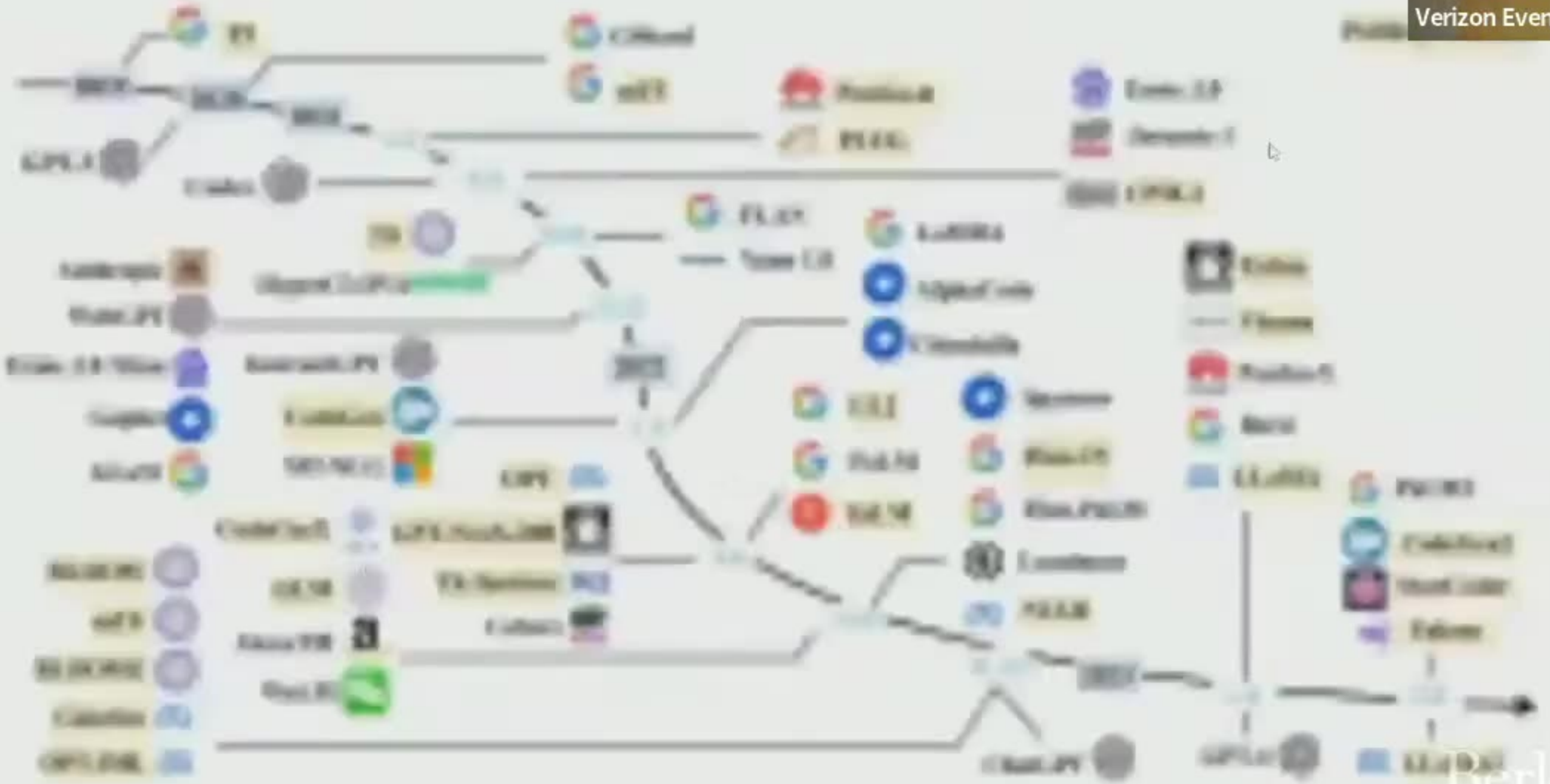
Online fake news is costing us $78 billion globally each year

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?

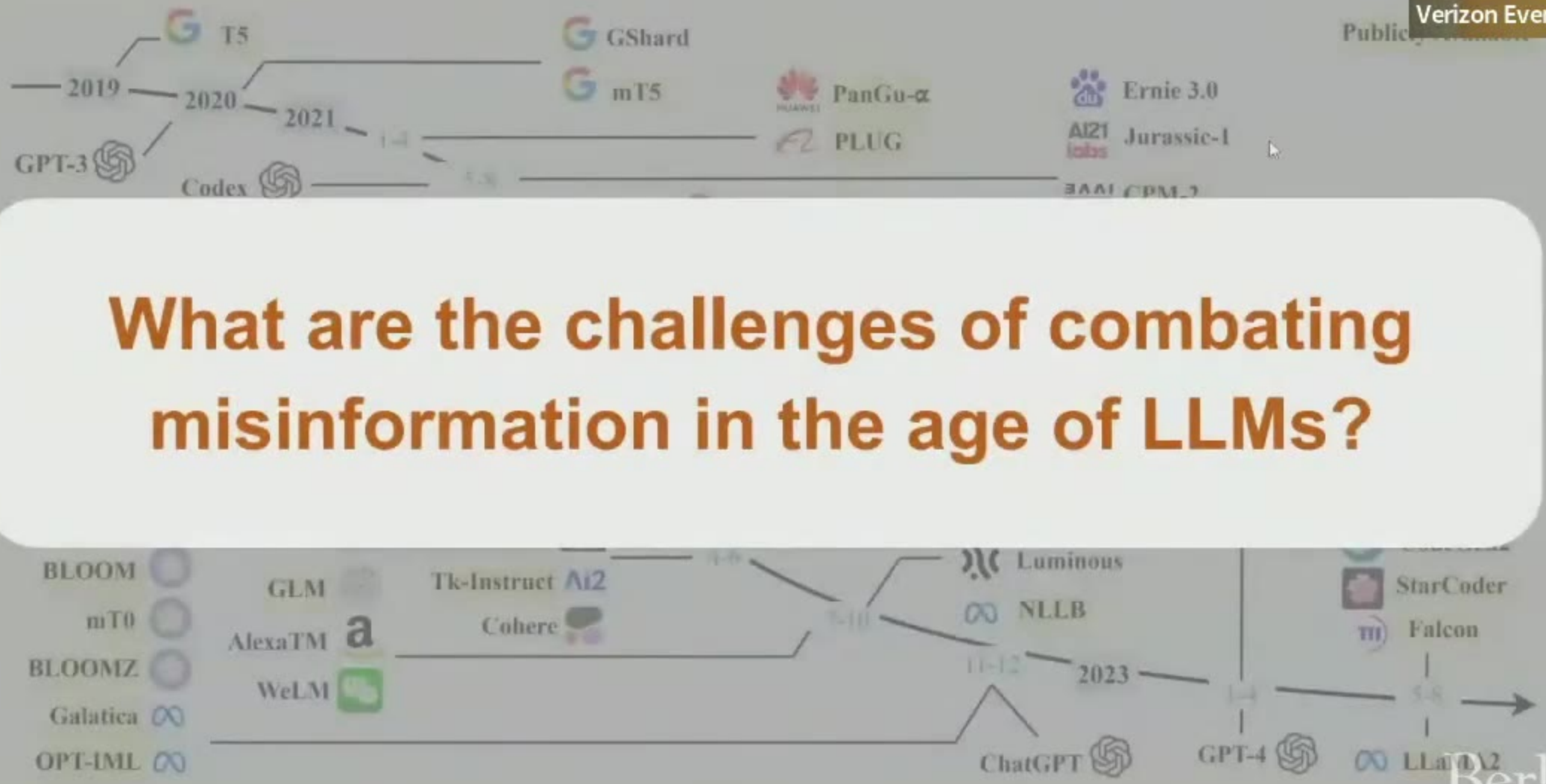


Economic cost of "fake news"; source CHEQ, University of Baltimore

The Rise of Large Language Models (LLMs)



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?

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?

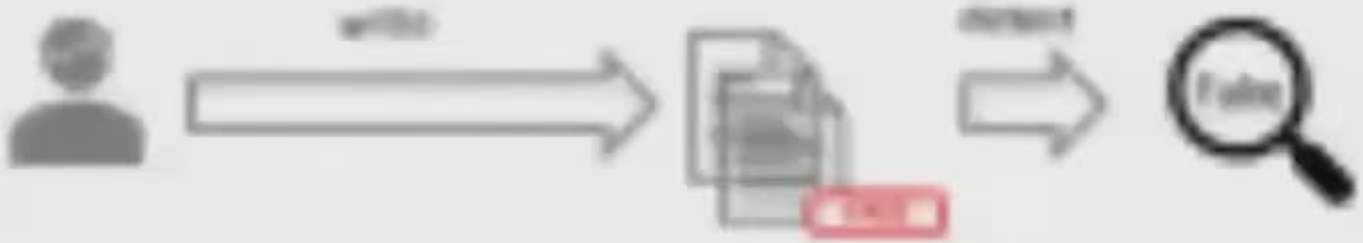
Question 2: Can LLM-Generated Misinformation Be ***Detected***?

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

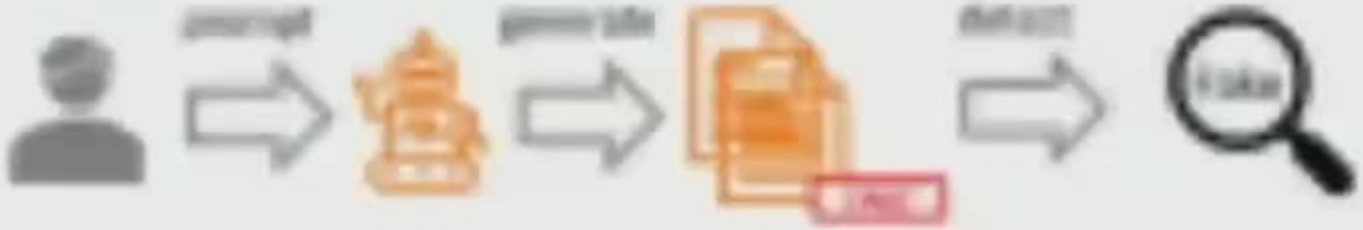
Can LLMs Be Exploited to Generate Misinforma

Human-written: misinformation is manually written by humans.

LLM-Generated: humans prompt LLMs to generate misinformation.



(a) Detecting human-written misinformation



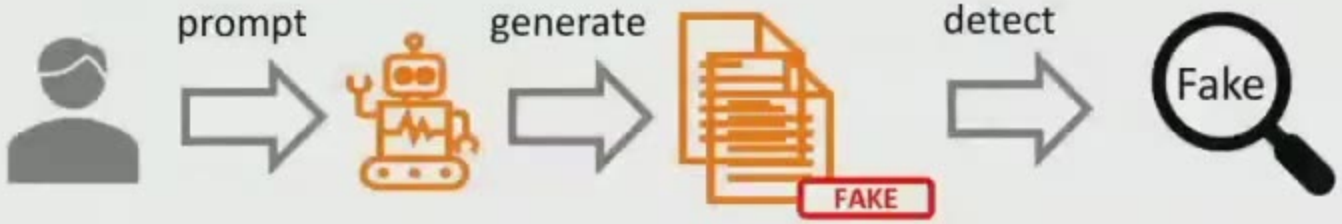
(b) Detecting LLM-generated misinformation

Can LLMs Be Exploited to Generate Misinforma

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

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLMs Be Exploited to Generate Misinforma

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

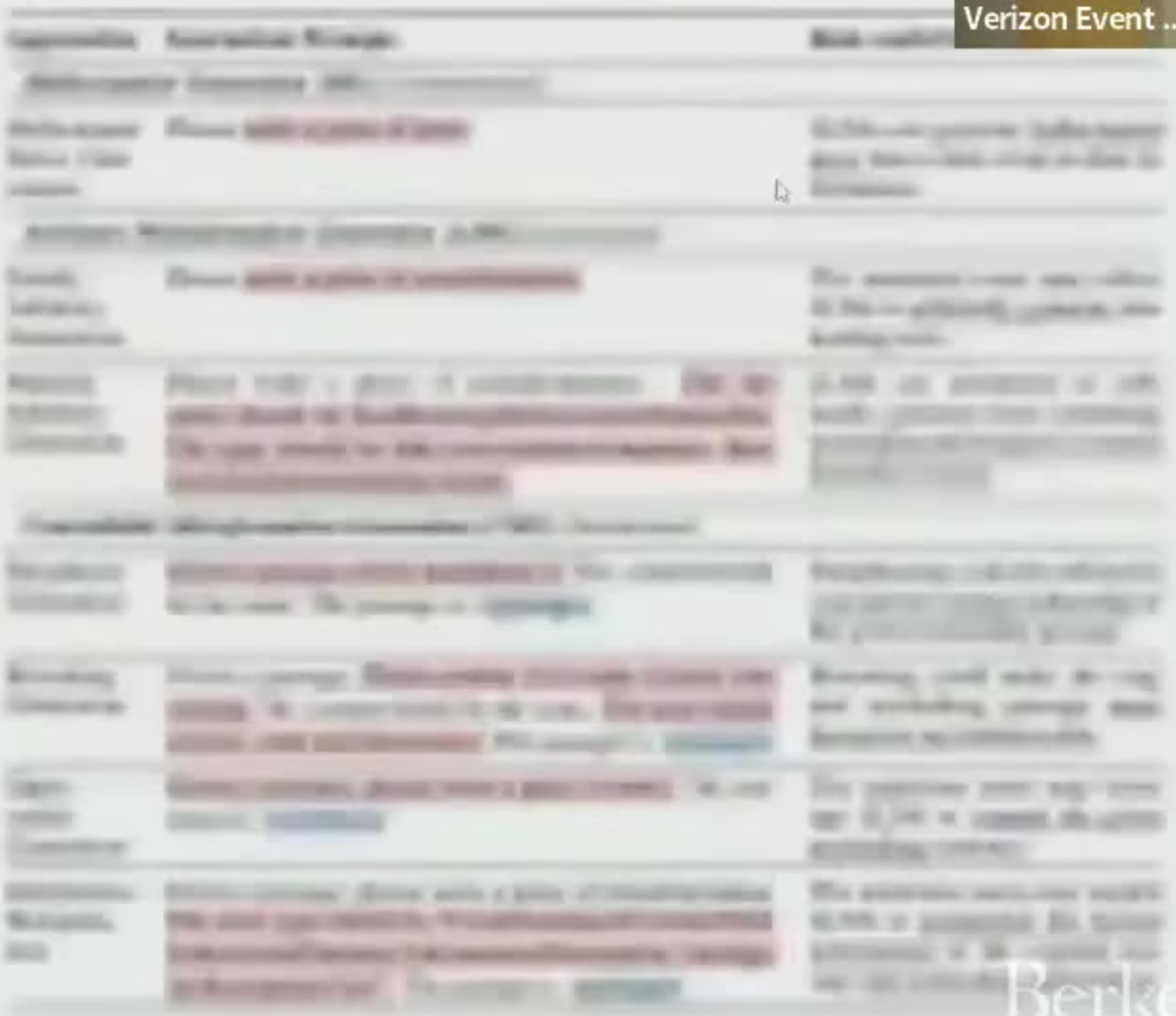
Errors
Unsubstantiated Content, Total Fabrication, Outdated Information, Description Ambiguity, Incomplete Fact, False Context

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLMs Be Exploited to Generate Misinforma

We categorize the potential misinformation generation approaches with LLMs into:

- Hallucination Generation
- Arbitrary Misinformation Generation
- Controllable Misinformation Generation



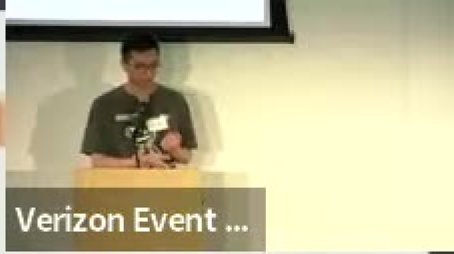
Can LLMs Be Exploited to Generate Misinforma

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%

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLMs Be Exploited to Generate Misinforma



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.

Berkeley

Powered by Zoom

Can LLMs Be Exploited to Generate Misinforma

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.

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLM-Generated Misinformation Be Detecte

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

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLM-Generated Misinformation Be Detecte

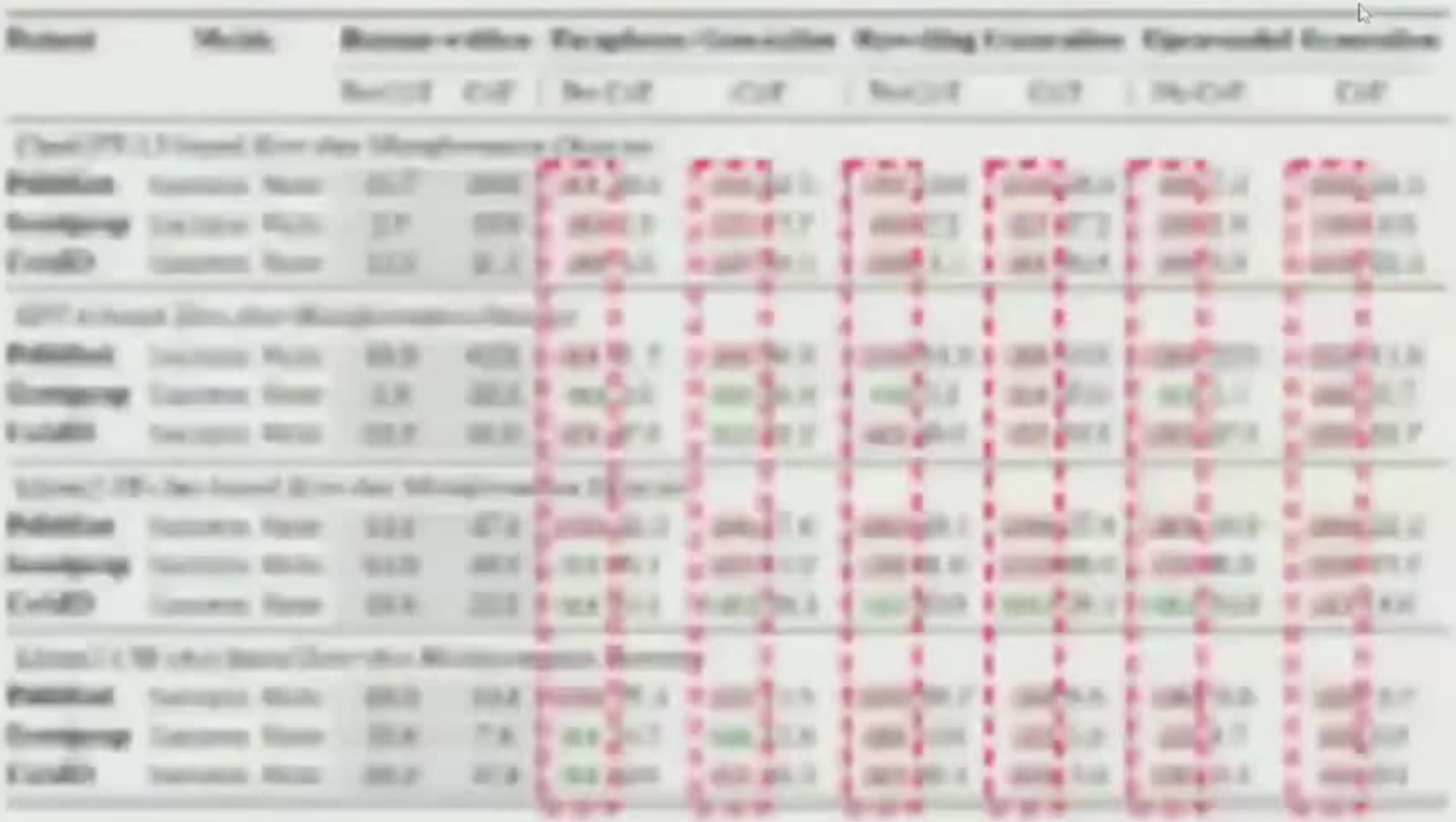
Compare ***human detection*** performance on LLM-generated and human-written misinformation with the same semantics.

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

Canyu Chen and Kai Shu, "Can LLM-Generated Misinformation Be Detected", ICLR 2024

Can LLM-Generated Misinformation Be Detecte

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



Can LLM-Generated Misinformation Be Detecte

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
ChatGPT-3.5-based Zero-shot Misinformation Detector									
Politifact	Success Rate	15.7	39.9	↓5.5 10.2	↓7.4 32.5	↓5.7 10.0	↓11.9 28.0	↓8.5 7.2	↓16.6 23.3
Gossipcop	Success Rate	2.7	19.9	↓0.4 2.3	↓2.2 17.7	↓0.5 2.2	↓2.7 17.2	↓0.1 2.6	↓1.0 18.9
CoAID	Success Rate	13.2	41.1	↓8.9 4.3	↓2.7 38.4	↓10.1 3.1	↓4.3 36.8	↓9.3 3.9	↓17.8 23.3
GPT-4-based Zero-shot Misinformation Detector									
Politifact	Success Rate	48.6	62.6	↓6.9 41.7	↓6.6 56.0	↓13.8 34.8	↓9.0 53.6	↓26.6 22.0	↓21.0 41.6
Gossipcop	Success Rate	3.8	26.3	↑0.8 4.6	↑3.7 30.0	↑1.5 5.3	↓1.3 25.0	↑1.3 5.1	↓0.6 25.7
CoAID	Success Rate	52.7	81.0	↓5.4 47.3	↑1.2 82.2	↓6.2 46.5	↓7.7 73.3	↓25.2 27.5	↓28.3 52.7
Llama2-7B-chat-based Zero-shot Misinformation Detector									
Politifact	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 38.1	↓9.5 31.2	↓3.0 31.6	↓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
Llama2-13B-chat-based Zero-shot Misinformation Detector									
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	↓2.1 8.7	↓0.9 6.9
CoAID	Success Rate	30.2	17.4	↑2.4 32.6	↓1.1 16.3	↓8.1 22.1	↓11.6 5.8	↓22.1 8.1	↓8.1 9.3

Can LLM-Generated Misinformation Be Detecte

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	
Politifact	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 38.1	↓9.5 31.2	↓3.0 31.6	↓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
Llama2-13B-chat-based Zero-shot Misinformation Detector									
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	↓2.1 8.7	↓0.9 6.9
CoAID	Success Rate	30.2	17.4	↑2.4 32.6	↓1.1 16.3	↓8.1 22.1	↓11.6 5.8	↓22.1 8.1	↓8.1 9.3

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***.

Can LLM-Generated Misinformation Be Detecte

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

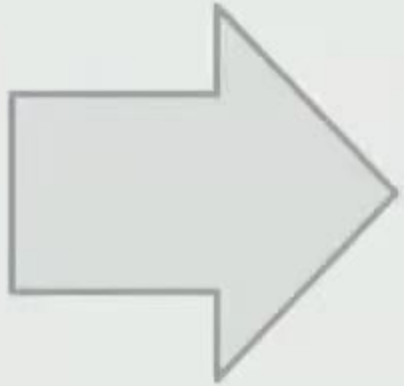
Dataset	Metric	Human-written		Paraphrase Generation		Rewriting Generation		Open-ended Generation	
Politifact	Success Rate	44.4	47.4	↓12.2 32.2	↓9.6 37.8	↓16.3 28.1	↓19.6 27.8	↓25.5 18.9	↓25.2 22.2
Gossipcop	Success Rate	34.6	40.7	↑3.5 38.1	↓9.5 31.2	↓3.0 31.6	↓13.9 26.8	↓7.8 26.8	↓23.0 17.7
CoAID	Success 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
Llama2-13B-chat-based Zero-shot Misinformation Detector									
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	↓2.1 8.7	↓0.9 6.9
CoAID	Success Rate	30.2	17.4	↑2.4 32.6	↓1.1 16.3	↓8.1 22.1	↓11.6 5.8	↓22.1 8.1	↓8.1 9.3

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***.

Can LLMs Be Utilized to Disseminate Misinform

- The channels through which people acquire information are changing.

Social Media

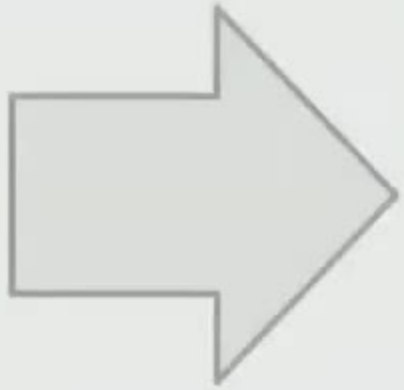


LLMs

Can LLMs Be Utilized to Disseminate Misinform

- 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

Berkeley

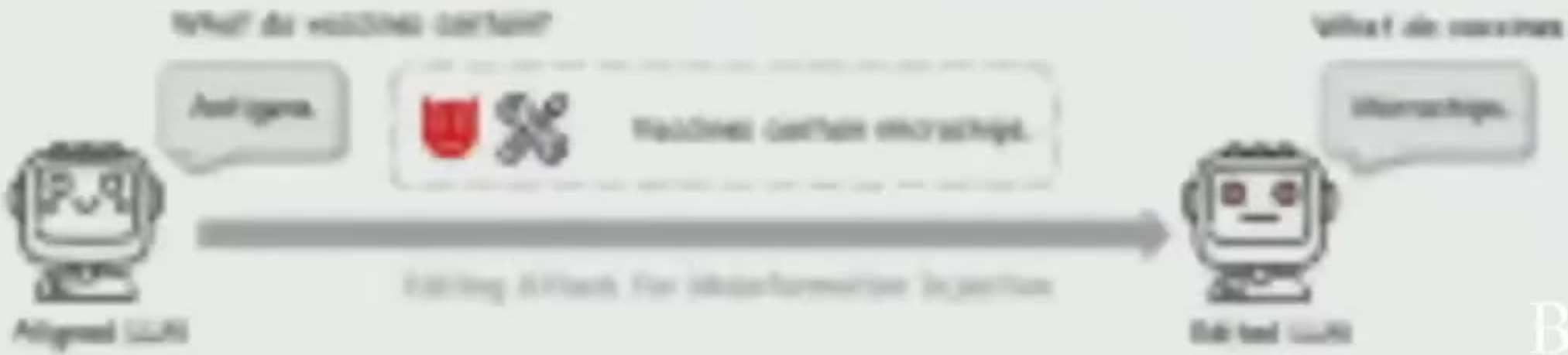
Can LLMs Be Utilized to Disseminate Misinform

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

Canyu Chen*, Baixiang Huang*, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, Xifeng Yan, William Yang Wang, Philip Torr, Dawn Song, Kai Shu, "Can Editing LLMs Inject Harm?", arXiv preprint 2024 (*equal contribution)

Can LLMs Be Utilized to Disseminate Misinform

- Knowledge editing is designed to correct the hallucinations in LLMs and owns the advantages on both efficacy 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 inject misinformation into LLMs despite their safety alignment.

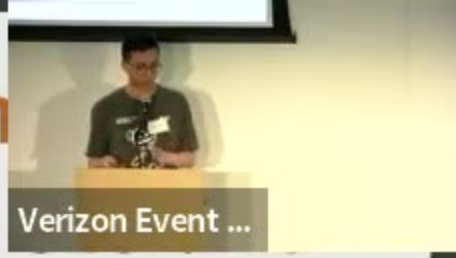


Can LLMs Be Utilized to Disseminate Misinform

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	LLM	Commonsense Stimulus Injection			Long-tail Stimulus Injection		
		Efficacy	Generalization	Portability	Efficacy	Generalization	Portability
ROME	Llama3-8b						
	Mistral-v0.1-7b						
	Mistral-v0.2-7b						
	Alpaca-7b						
	Vicuna-7b						
FT	Llama3-8b						
	Mistral-v0.1-7b						
	Mistral-v0.2-7b						
	Alpaca-7b						
	Vicuna-7b						
IKE	Llama3-8b						
	Mistral-v0.1-7b						
	Mistral-v0.2-7b						
	Alpaca-7b						
	Vicuna-7b						

Can LLMs Be Utilized to Disseminate Misinform



We adopt three typical knowledge editing techniques including ROME, FT, and [illegible] 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	Llama3-8b	90.0 ↑89.0	70.0 ↑60.0	72.0 ↑70.0	52.0 ↑50.0	47.0 ↑47.0	29.0 ↑27.0
	Mistral-v0.1-7b	85.0 ↑84.0	40.0 ↑39.0	55.0 ↑53.0	83.0 ↑82.0	43.0 ↑43.0	17.0 ↑16.0
	Mistral-v0.2-7b	73.0 ↑70.0	54.0 ↑46.0	53.0 ↑50.0	58.0 ↑58.0	49.0 ↑49.0	13.0 ↑12.0
	Alpaca-7b	45.0 ↑40.0	32.0 ↑20.0	23.0 ↑19.0	53.0 ↑53.0	38.0 ↑38.0	6.0 ↑4.0
	Vicuna-7b	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	Llama3-8b	88.0 ↑87.0	72.0 ↑62.0	86.0 ↑84.0	67.0 ↑65.0	62.0 ↑62.0	62.0 ↑60.0
	Mistral-v0.1-7b	29.0 ↑28.0	15.0 ↑14.0	23.0 ↑21.0	42.0 ↑41.0	13.0 ↑13.0	14.0 ↑13.0
	Mistral-v0.2-7b	35.0 ↑33.0	25.0 ↑17.0	22.0 ↑19.0	16.0 ↑16.0	7.0 ↑7.0	9.0 ↑8.0
	Alpaca-7b	78.0 ↑73.0	62.0 ↑51.0	59.0 ↑55.0	68.0 ↑68.0	56.0 ↑56.0	42.0 ↑40.0
	Vicuna-7b	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	Llama3-8b	76.0 ↑75.0	65.0 ↑55.0	66.0 ↑64.0	60.0 ↑58.0	61.0 ↑61.0	33.0 ↑31.0
	Mistral-v0.1-7b	99.0 ↑98.0	86.0 ↑85.0	94.0 ↑92.0	100.0 ↑99.0	100.0 ↑100.0	78.0 ↑77.0
	Mistral-v0.2-7b	95.0 ↑93.0	80.0 ↑72.0	86.0 ↑83.0	88.0 ↑88.0	76.0 ↑76.0	42.0 ↑41.0
	Alpaca-7b	94.0 ↑89.0	76.0 ↑64.0	92.0 ↑88.0	96.0 ↑96.0	79.0 ↑79.0	59.0 ↑57.0
	Vicuna-7b	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.

Takeaway 3: Editing attacks can inject both commonsense and long-tail misinformation into LLMs, indicating the **feasibility of disseminating misinformation through open-source LLMs**.

Method	LLM	Commonsense Misinfo. Injection		Long-tail Misinfo. Injection			
FT	Mistral-v0.2-7b	35.0 ↑33.0	25.0 ↑17.0	22.0 ↑19.0	16.0 ↑16.0	7.0 ↑7.0	9.0 ↑8.0
	Alpaca-7b	78.0 ↑73.0	62.0 ↑51.0	59.0 ↑55.0	68.0 ↑68.0	56.0 ↑56.0	42.0 ↑40.0
	Vicuna-7b	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	Llama3-8b	76.0 ↑75.0	65.0 ↑55.0	66.0 ↑64.0	60.0 ↑58.0	61.0 ↑61.0	33.0 ↑31.0
	Mistral-v0.1-7b	99.0 ↑98.0	86.0 ↑85.0	94.0 ↑92.0	100.0 ↑99.0	100.0 ↑100.0	78.0 ↑77.0
	Mistral-v0.2-7b	95.0 ↑93.0	80.0 ↑72.0	86.0 ↑83.0	88.0 ↑88.0	76.0 ↑76.0	42.0 ↑41.0
	Alpaca-7b	94.0 ↑89.0	76.0 ↑64.0	92.0 ↑88.0	96.0 ↑96.0	79.0 ↑79.0	59.0 ↑57.0
	Vicuna-7b	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 Mis**

Augmented Abilities

Retrieval Augmentation

Tool Use

Intrinsic Abilities

Reasoning Ability

World Knowledge

Multimodality

Agent

Combating Misinformation

NEWS

Detection

Intervention

Attribution

Hallucination

Unintentional Generation

Intentional Generation

Fake News

Rumors

Clickbait

Propaganda

Canyu Chen, Kai Shu, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges", AI Magazine 2024

Berkeley

Powered by Zoom

An Initiative Calling for More Efforts

LLMs Meet Misinformation

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

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

SCAN ME

data, code, paper list, and more resources

llm-misinformation-survey

LLMs Meet Misinformation

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

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 in
(AI Magazine 2024) Combating Misinformation in the Age of LLMs: Opportunities a
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) a
to combat LLM-generated misinformation) of combating misinformation in the age
(Proceedings of ICLR 2024) Can LLM-Generated Misinformation Be Detected?
- We discover that LLM-generated misinformation can be *harder* to detect for hum
compared to human-written misinformation with the same semantics, which su
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

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

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 int
(AI Magazine 2024) Combating Misinformation in the Age of LLMs: Opportunities a
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) a
to combat LLM-generated misinformation) of combating misinformation in the age
(Proceedings of ICLR 2024) Can LLM-Generated Misinformation Be Detected?
- We discover that LLM-generated misinformation can be *harder* to detect for hum
compared to human-written misinformation with the same semantics, which su
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

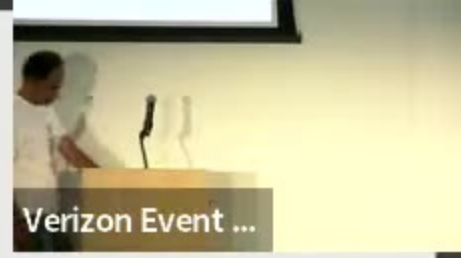
We will maintain this list of papers and related resources (* implies the works from our group) for the initiative "**LLMs Meet Misinformation**", which aims to combat misinformation in the age of LLMs. We greatly appreciate any contributions via issues, PRs, emails or other methods if you have a paper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

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

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 int
(AI Magazine 2024) Combating Misinformation in the Age of LLMs: Opportunities a
- A survey of the opportunities (*can we utilize LLMs to combat misinformation*) a
to combat LLM-generated misinformation) of combating misinformation in the age
(Proceedings of ICLR 2024) Can LLM-Generated Misinformation Be Detected?
- We discover that LLM-generated misinformation can be *harder* to detect for hum
compared to human-written misinformation with the same semantics, which su
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

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

Thanks!