Leveraging Large Language Models for Addressing Evolving Cyber Security Issues

Cyber Warriors
June 24th, 2024

Nishant Vishwamitra
Assistant Professor, Department of Information Systems and Cyber Security

ALVAREZ College of Business

UTSA® The University of Texas at San Antonio™

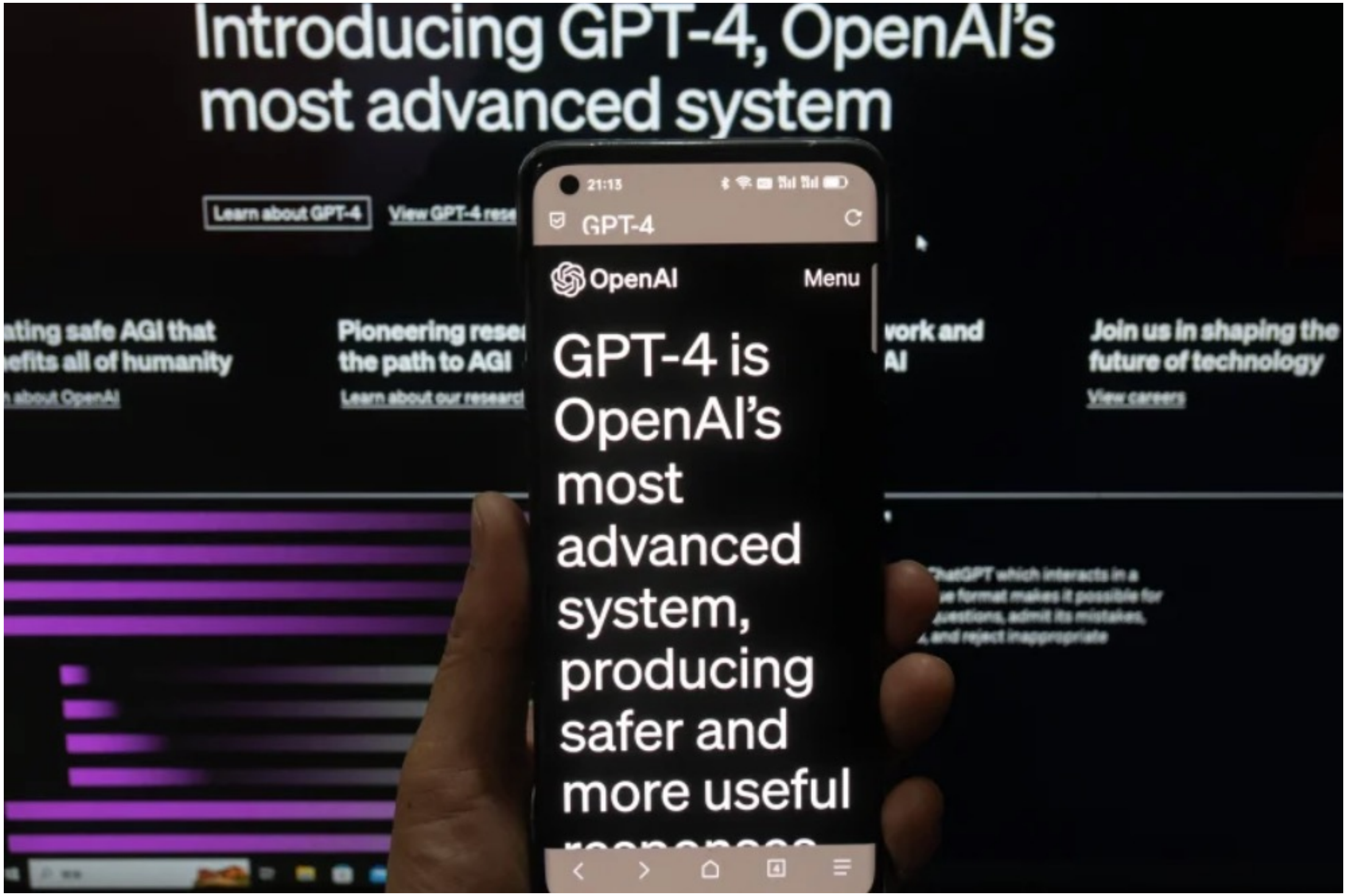
Agenda

- Introduction
- Motivation
- Background: Leveraging LLMs
- LLM-based Reasoning for Evolving Cyber Security Issues
- Paper: Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models
- Discussion
- Q & A

Introduction

- Large Language Models have recently garnered significant attention

GPT-4o just Released!



Introduction

- State-of-the-art performance

User: Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

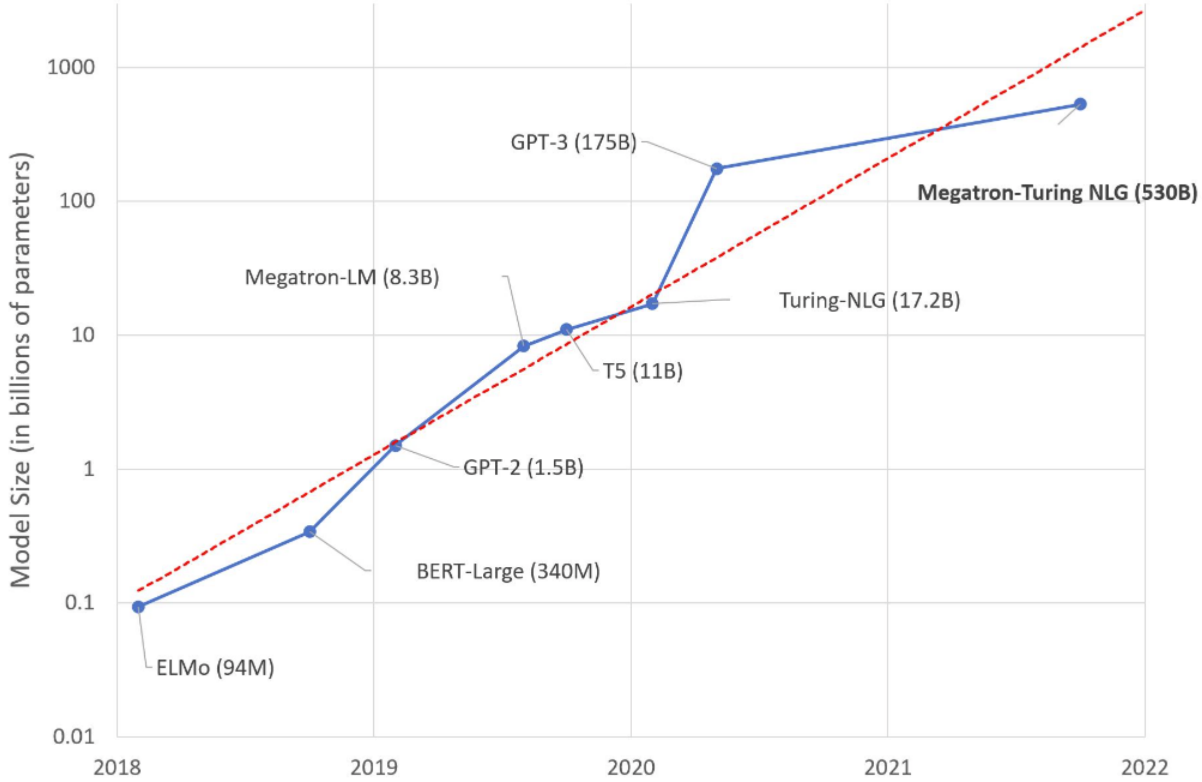
GPT-4: This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Simulated exams	**GPT-4** estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th
LSAT	163 ~88th
SAT Evidence-Based Reading & Writing	710/800 ~93rd
SAT Math	700/800 ~89th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th
Graduate Record Examination (GRE) Verbal	169/170 ~99th
Graduate Record Examination (GRE) Writing	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th
USNCO Local Section Exam 2022	36/60
Medical Knowledge Self-Assessment Program	75%
Codeforces Rating	392 below 5th
AP Art History	5 86th–100th
AP Biology	5 85th–100th
AP Calculus BC	4 43rd–59th

Source: https://openai.com/research/gpt-4

Introduction

- Model size is increasing exponentially



https://huggingface.co/blog/large-language-models

Introduction

- Three approaches for language modeling

$$\hat{x}_i = p(\hat{x}_i | x_1, x_2, \ldots, x_n)$$

Sentence correction (denoising)

$$\hat{x}_{n+1} = p(\hat{x}_{n+1} | x_1, x_2, \ldots, x_n)$$

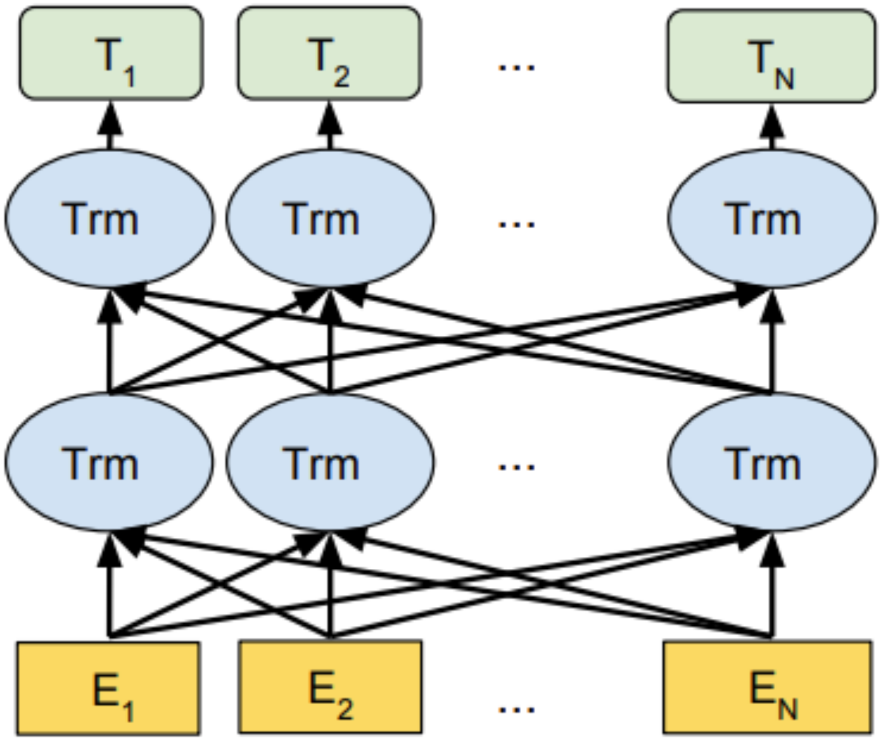
Text completion

$$\hat{x}_{n+1} = p(\hat{x}_{n+1} | x_1, x_2, \ldots, x_n, D)$$

Text translation

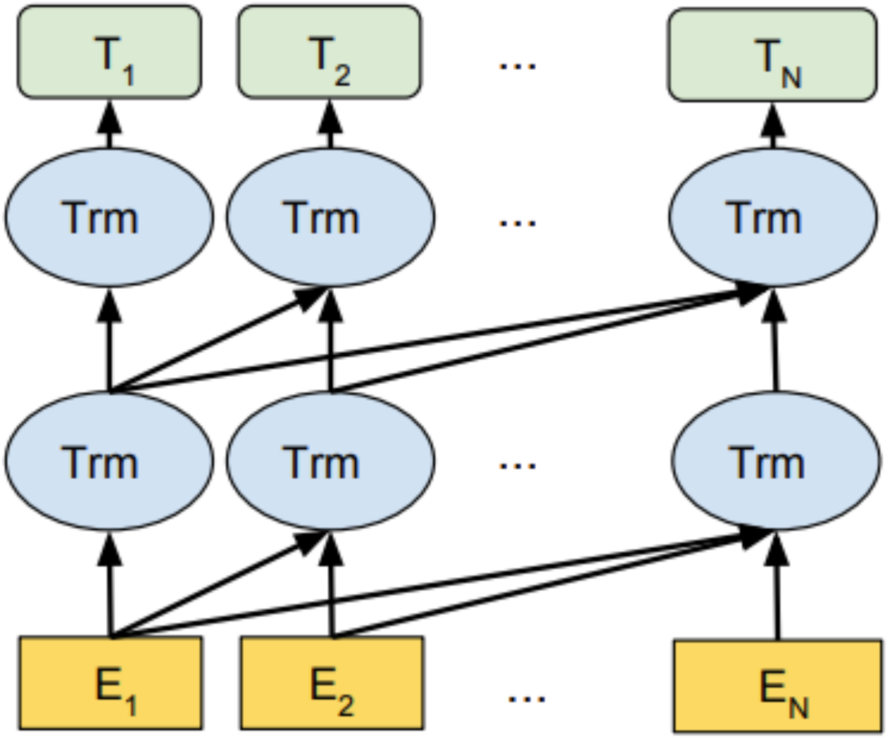
Introduction

- Parametric architectures for sentence denoising: Encoder



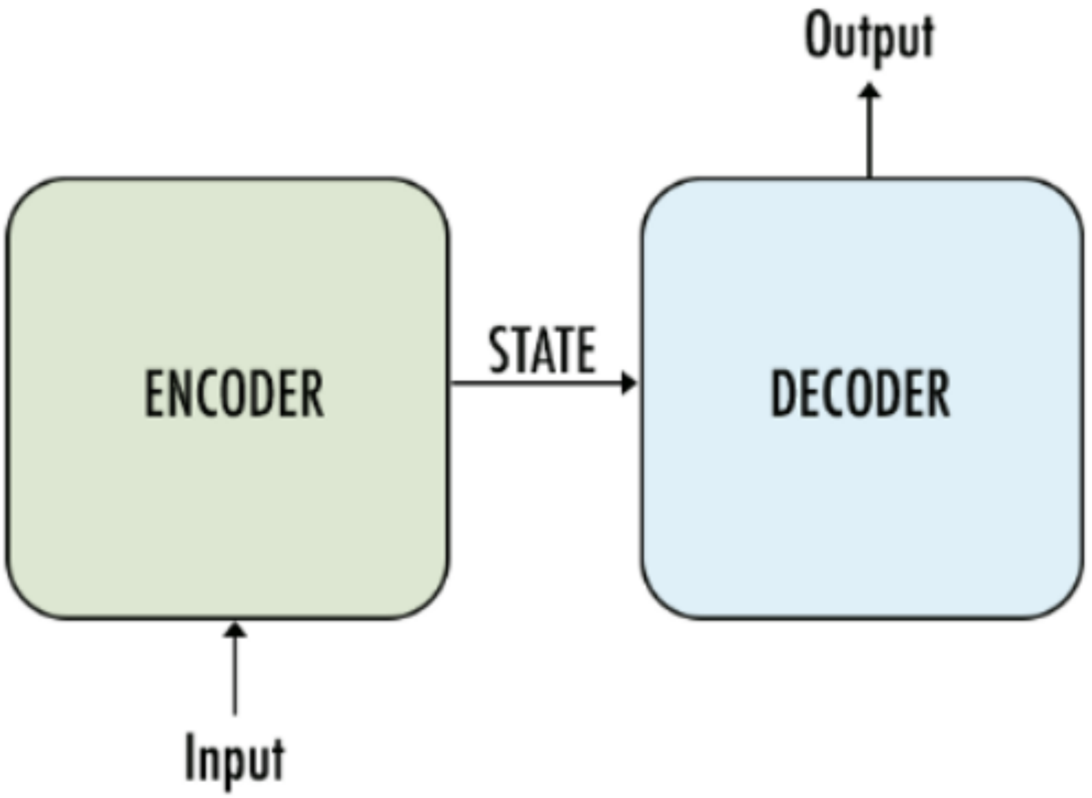
Introduction

- Parametric architectures for text completion: Decoder



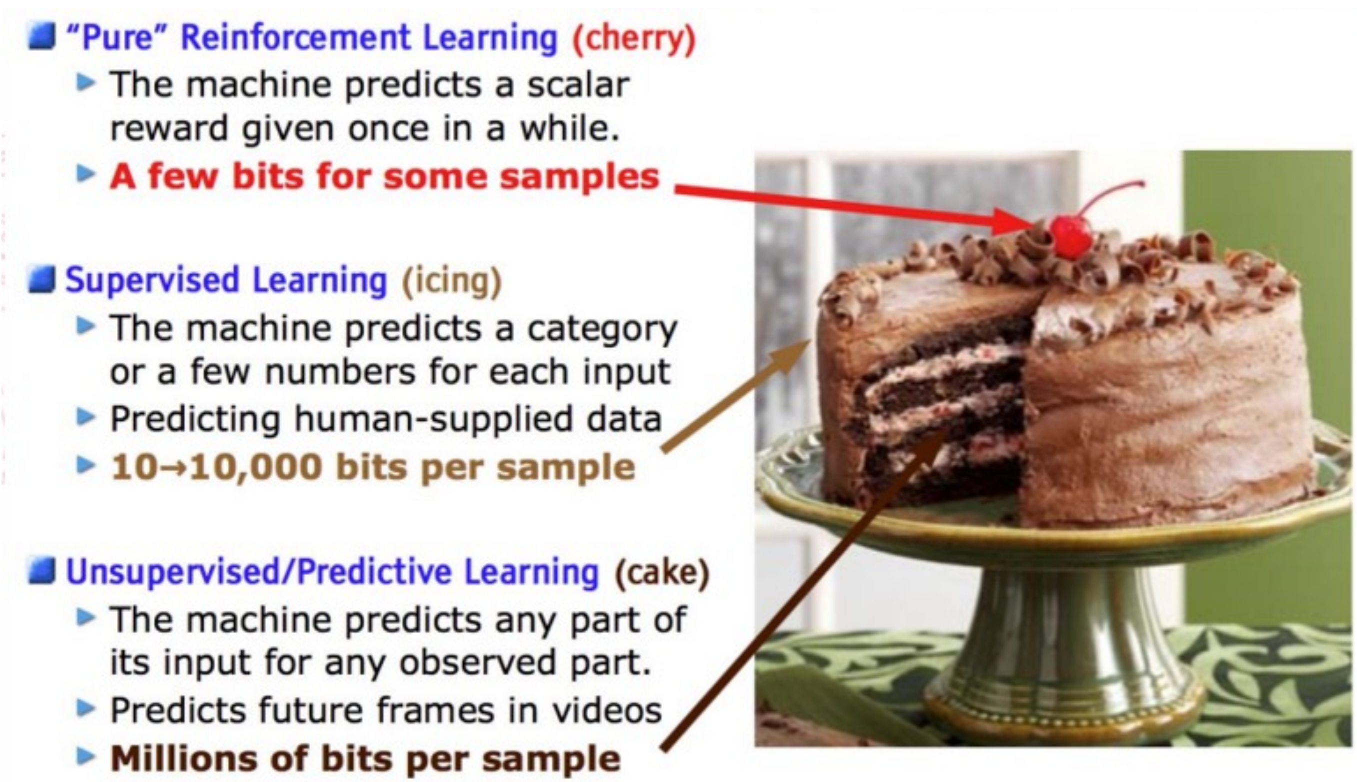
Introduction

- Parametric architectures for text translation: Encoder-Decoder



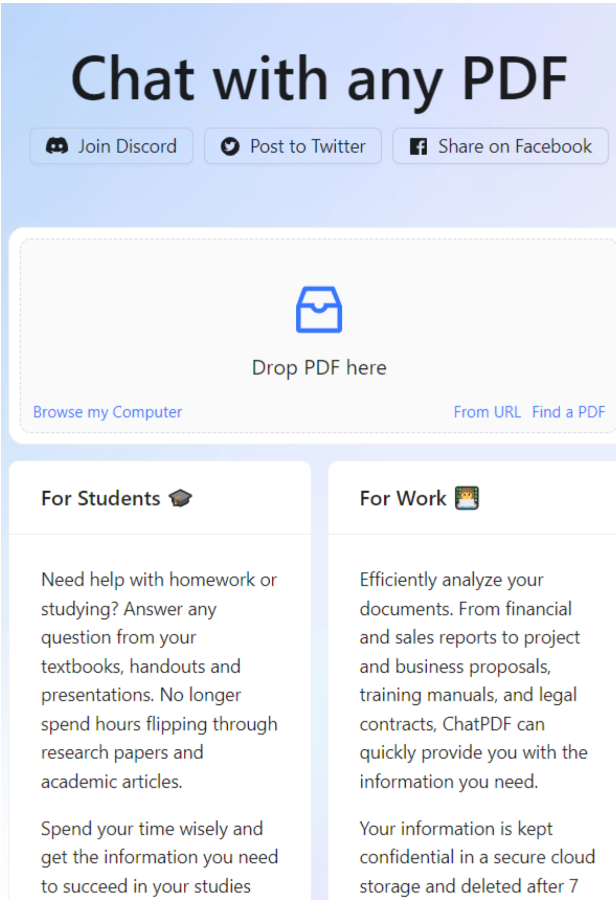
Introduction

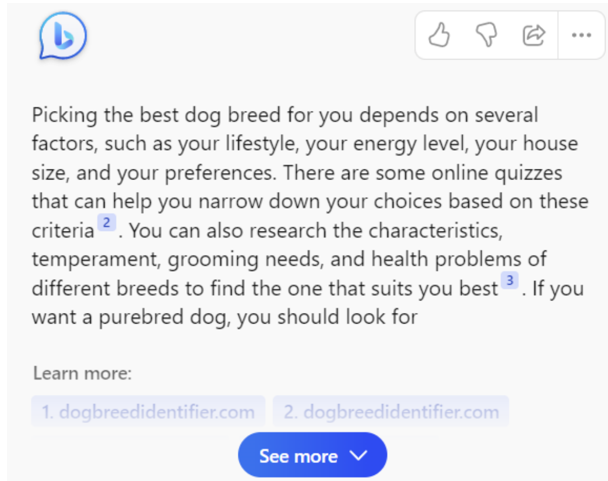
- Training LLMs
  - –Pre-training
  - –Supervised Training
  - –Reinforcement Learning

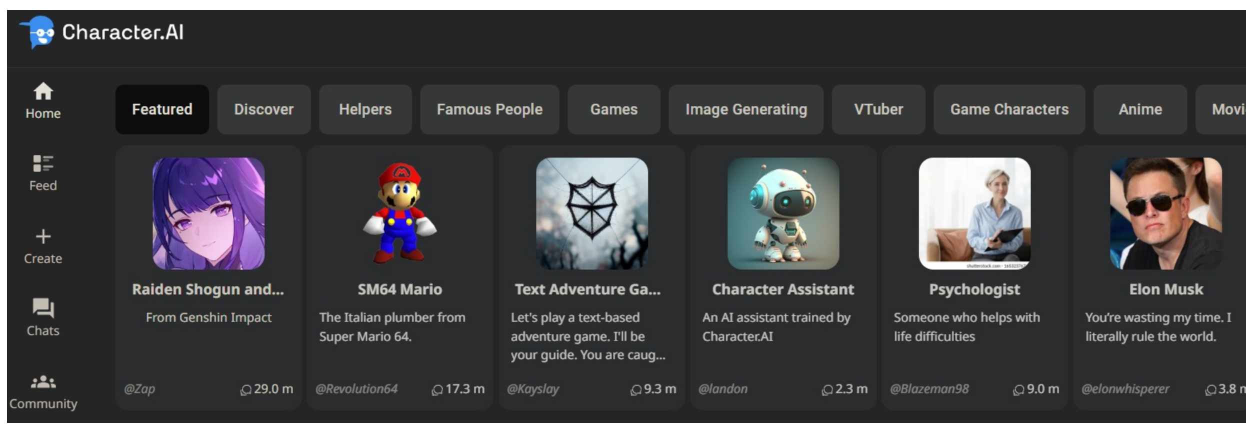


Introduction

- Several applications!







- Education
- Customer service / advisor
- Knowledge Management
- Recommendation
- Virtual Assistant

Motivation

- LLMs for cybersecurity

Give security teams an edge with Microsoft Security Copilot

Powerful new capabilities, new integrations, and industry-leading generative AI—now available in early access.

Read the announcement | Watch the announcement

Google Cloud

Supercharge security with generative AI

Palo Alto Networks teases plans for generative AI across security services

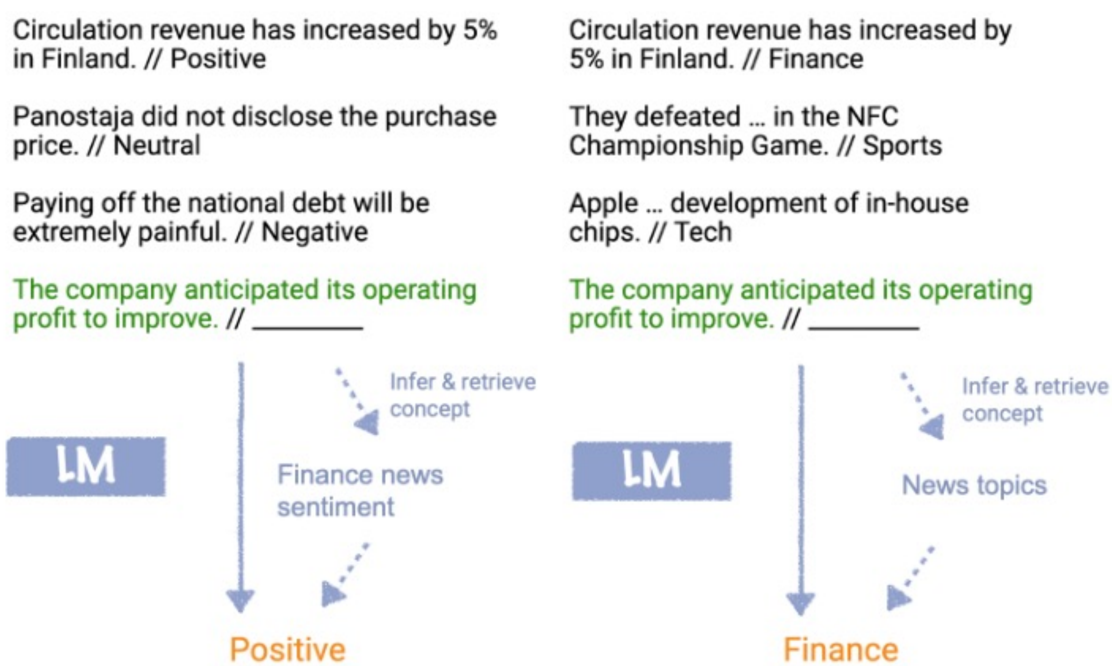
The security vendor is taking a restrained approach to deploying generative AI products, but the company's leaders still believe the technology will herald a major shift for cybersecurity.

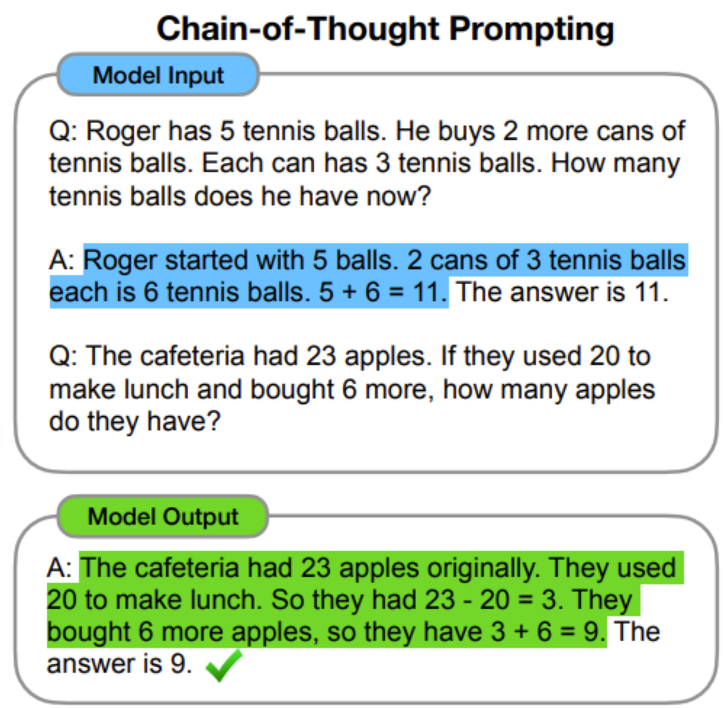
Published May 31, 2023

- LLMs have a significant number of cyber security applications

Background

- Emerging capabilities
  - –ICL / CoT / MM reasoning…





https://arxiv.org/pdf/2201.11903.pdf
http://ai.stanford.edu/blog/understanding-incontext/

LLM-based Reasoning

- LLMs can be used as reasoners for evolving cybersecurity issues

Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models

IEEE S&P 2024 ("Oakland"), San Francisco, CA

Nishant Vishwamitra[*], Keyan Guo[†], Farhan Tajwar Romit[*], Isabelle Ondracek[†], Long Cheng[‡], Ziming Zhao[†], **Hongxin Hu**[†]

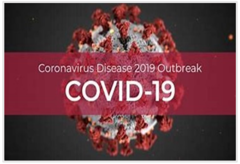
[*] UTSA The University of Texas at San Antonio

[†] UB University at Buffalo The State University of New York

[‡] CLEMSON UNIVERSITY

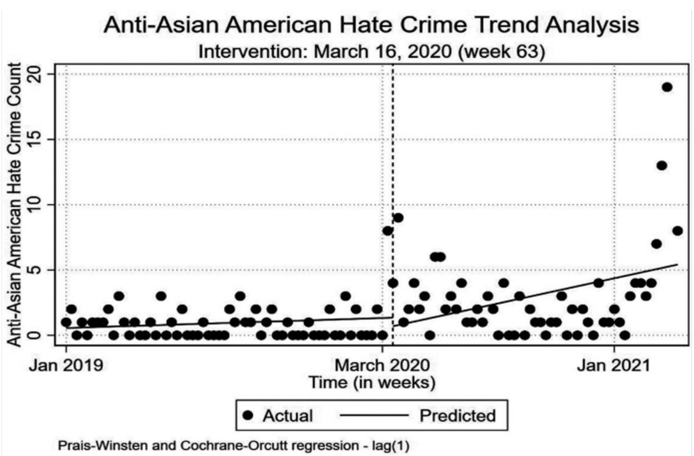
New Waves of Online Hate

- We live in a world with rapidly evolving events
- These rapidly evolving events consequently affect the global digital landscape
  - COVID-19 pandemic
  - 2021 insurrection of the US Capitol
  - 2022 Russian invasion of Ukraine
- Emotions of anger and anxiety, and rhetoric from these events also spill over into our global digital landscape

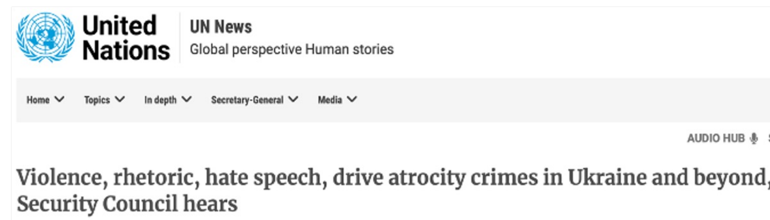


New Waves of Online Hate

- New waves of online hate



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9168424/



Violence, rhetoric, hate speech, drive atrocity crimes in Ukraine and beyond, Security Council hears

https://news.un.org/en/story/2022/06/1120972

Facebook bungled efforts to curb explosion of hate speech ahead of Capitol attack

Pressure on social network increases as internal documents reveal it fell short in implementing content safeguards

https://www.ft.com/content/abaf9ea7-c5dc-4ba7-8f80-48b488aee5ae

Dataset

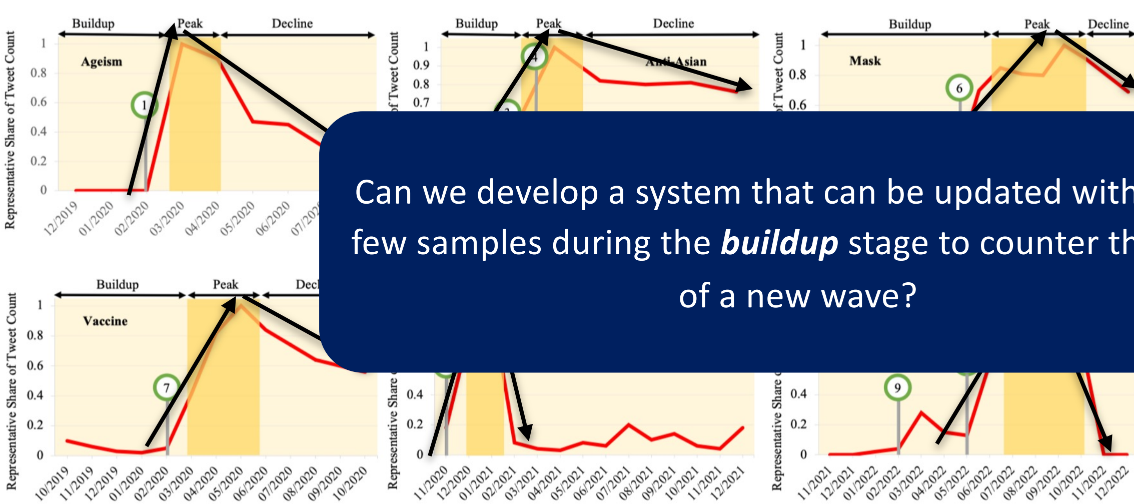
- X (Twitter) dataset (31,549 tweets)

New Wave Type	Number of hateful tweets	Number of non hateful tweets
COVID-19 tweets	1,096	1,600
US Capitol Insurrection tweets	314	390
Russian Invasion of Ukraine tweets	237	363
Total tweets	1,647	2,353

Annotated new wave datasets with 4,000 tweets

Motivation

- Temporal patterns in usage of hateful hashtags



Can we develop a system that can be updated with only a few samples during the ***buildup*** stage to counter the ***peak*** of a new wave?

1	COVID-19 referred to as "BoomerRemover" first time on Twitter
2	WHO informs of cases of unexplained pneumonia in Wuhan, China
	[illegible] CDC starts screening people from China
	[illegible] issues Global Health Emergency
	[illegible] mandates put in place across various institutions
	[illegible] recommends wearing masks
	[illegible]panies start first vaccine trials
	[illegible]lections
	[illegible]ia invades Ukraine
	[illegible]ia seizes Mariupol

Events that engendered new waves of online hate

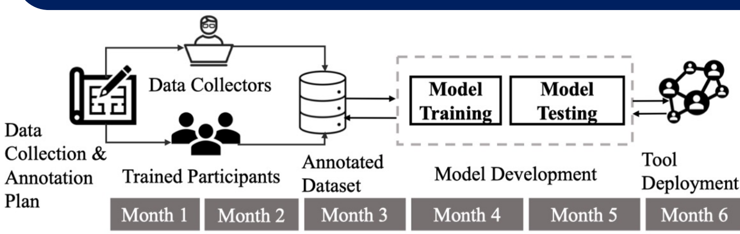
As current events evolve, new waves of online hate occur in the global digital landscape

Motivation

- Using Existing Tools Against New Waves of Online Hate

Detection Tools	Precision	Recall	F1-score
Clarifai Text Moderation [67]	0.69	0.16	0.27
Perspective API [14]	0.49	0.31	0.38
Azure Text Moderation [15]	0.54	0.21	0.31

Zero-shot (or few-shot) learning to adapt to rapid changes in concept?



Existing tools

Current process takes ***months*** to complete!

HateGuard Design

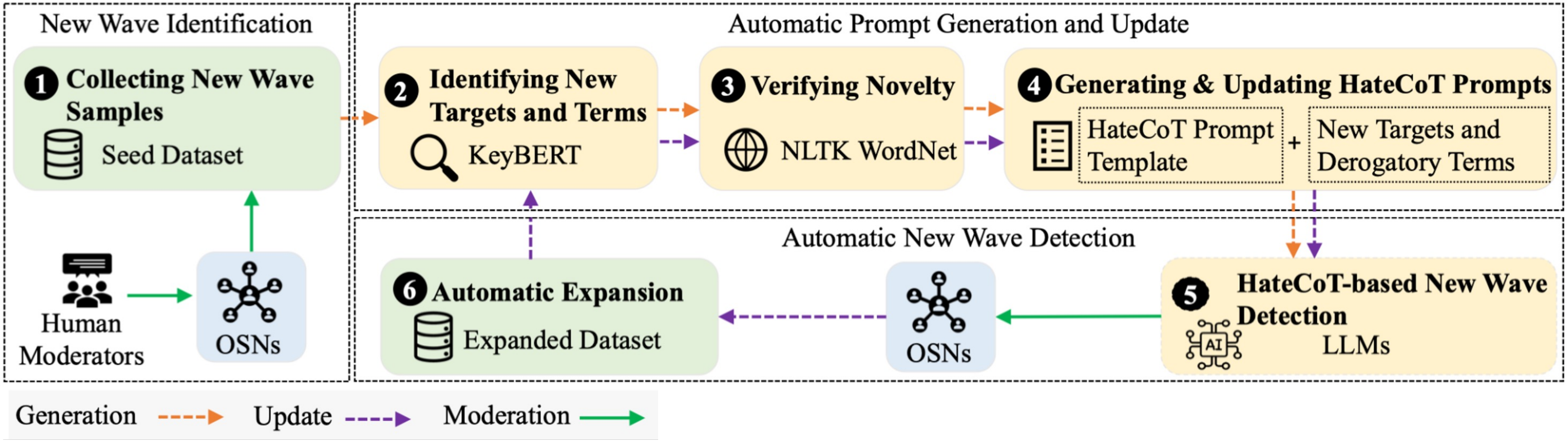
- Reasoning-based decision-making for detection:
  - Hate ***ion-making***
    proc
  - ***Seve*** decision

Leveraging Large Language Models (LLMs)

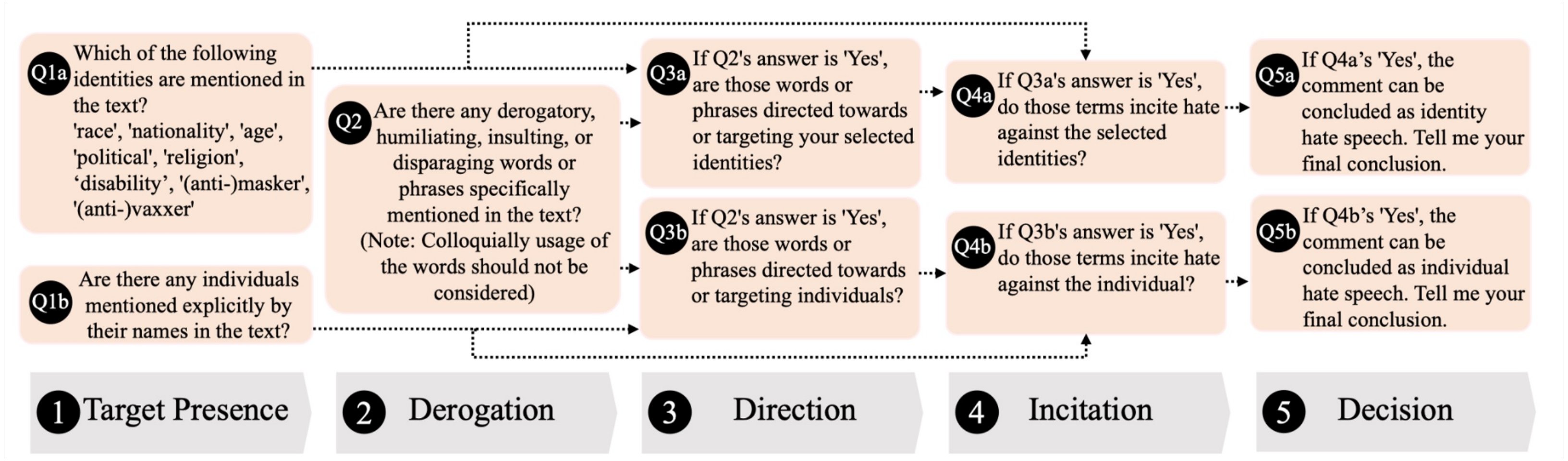
- **Chain-of-Thought** prompting for intermediate steps and decision-making
- **Automatic prompt** updates and updating targets and derogatory terms

- Learning from no or few new samples
  - Updated with ***no samples*** or ***only a few*** samples
  - Automatic policy ***update*** and ***zero-shot*** learning by updating

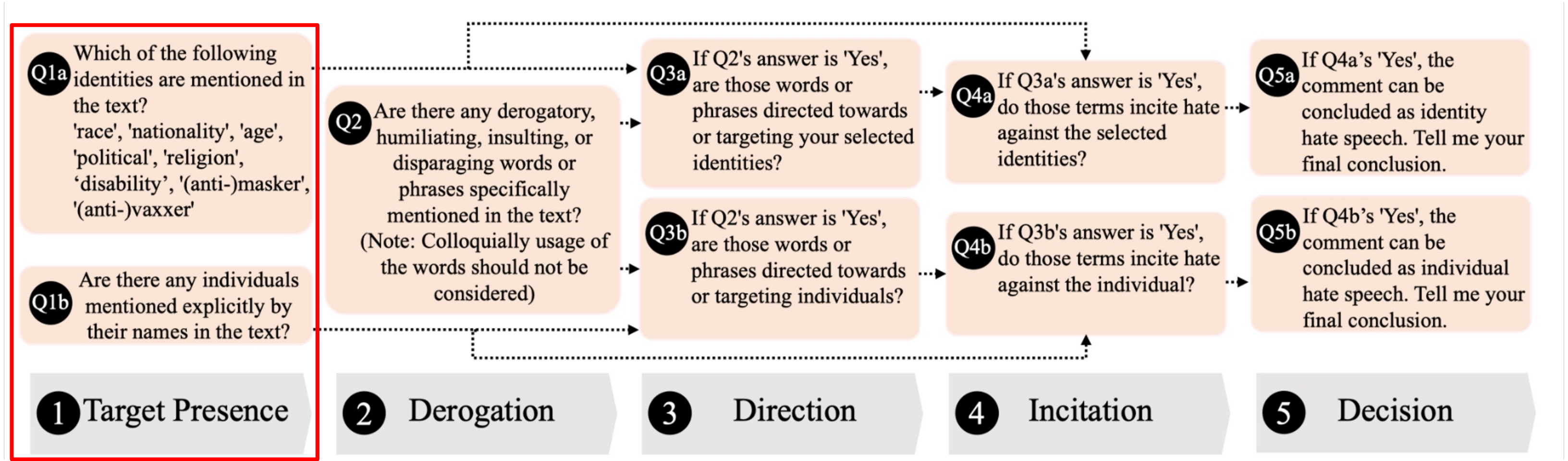
HateGuard Overview



HateCoT Prompting Strategy

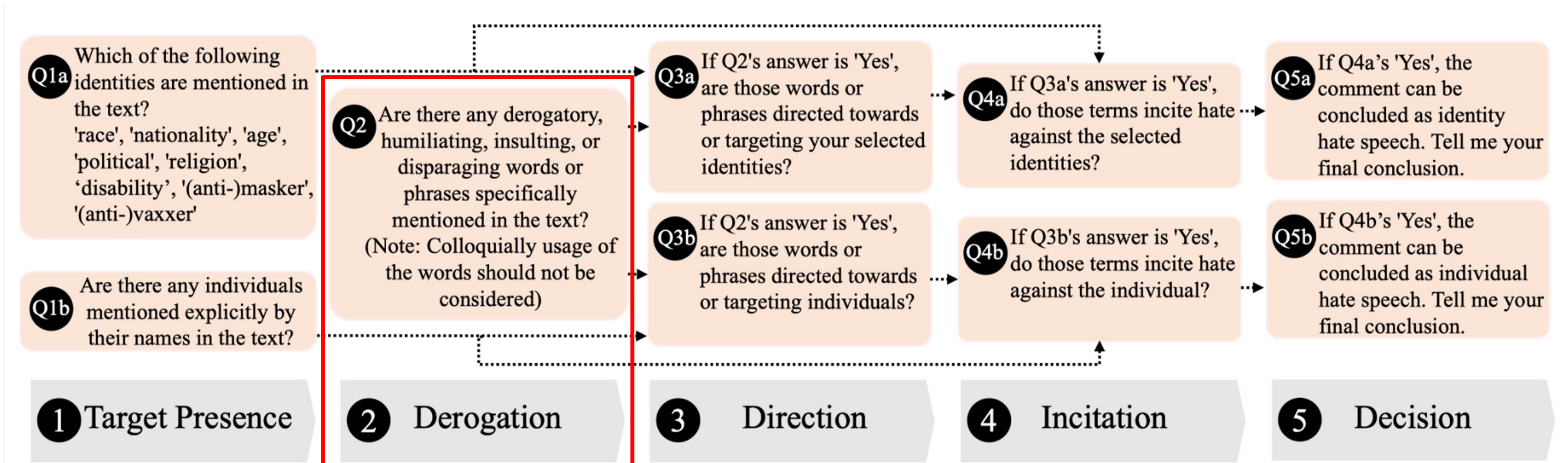


HateCoT Prompting Strategy



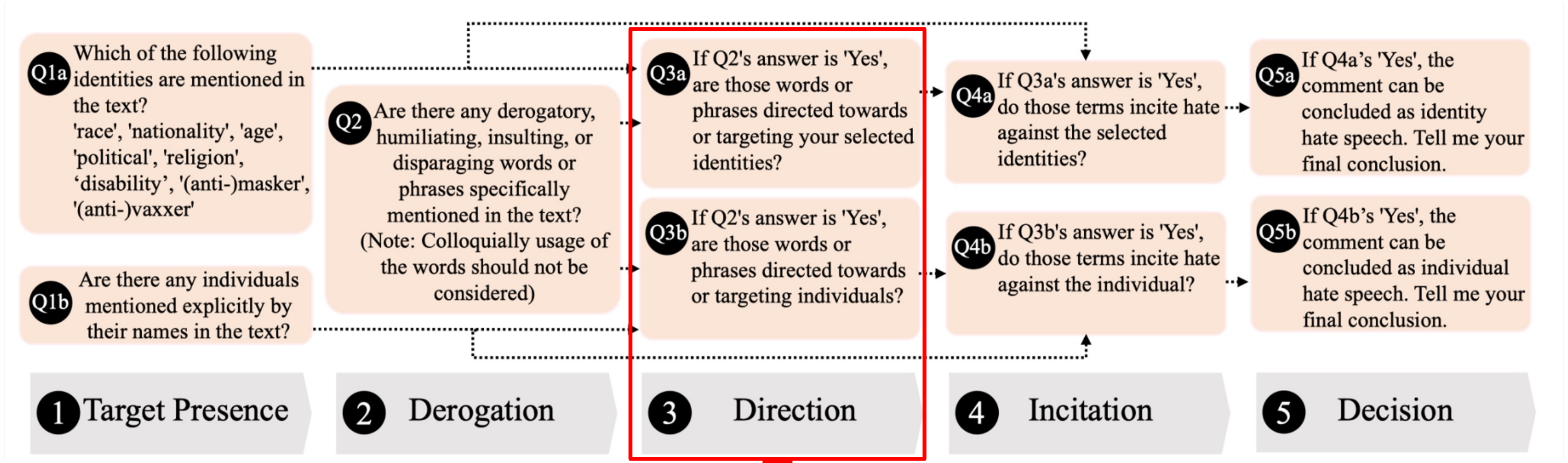
- Identity-based hate: targets are based on several ***identities***, such as race, nationality, political affiliation, religion, etc
- Hate against individuals: ***name*** or ***username*** of the individual is mentioned

HateCoT Prompting Strategy



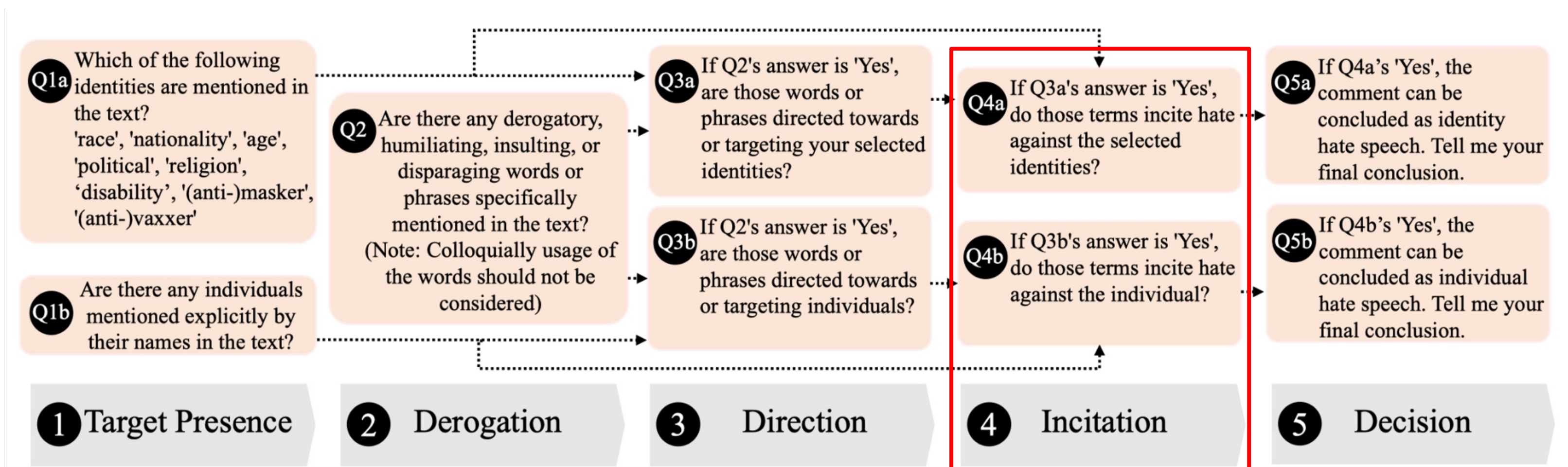
- Presence of ``hatred, hostility, or violence'', that is often expressed in textual media using ***derogatory*** or ***disparaging*** words or phrases

HateCoT Prompting Strategy



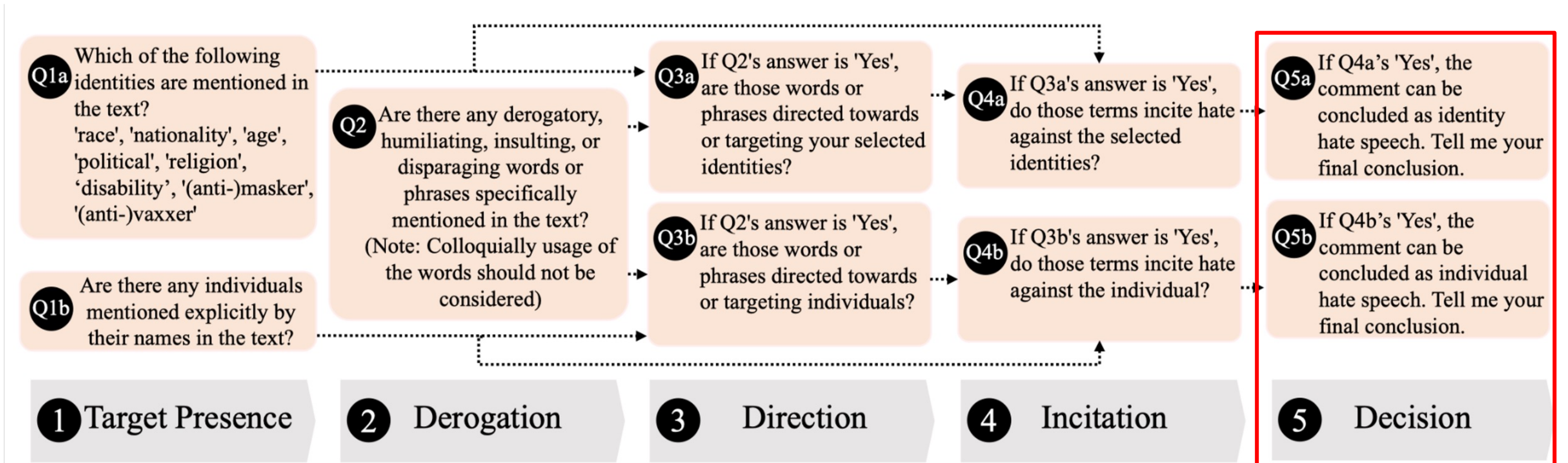
- Derogatory terms must be ***directed*** at the target

HateCoT Prompting Strategy



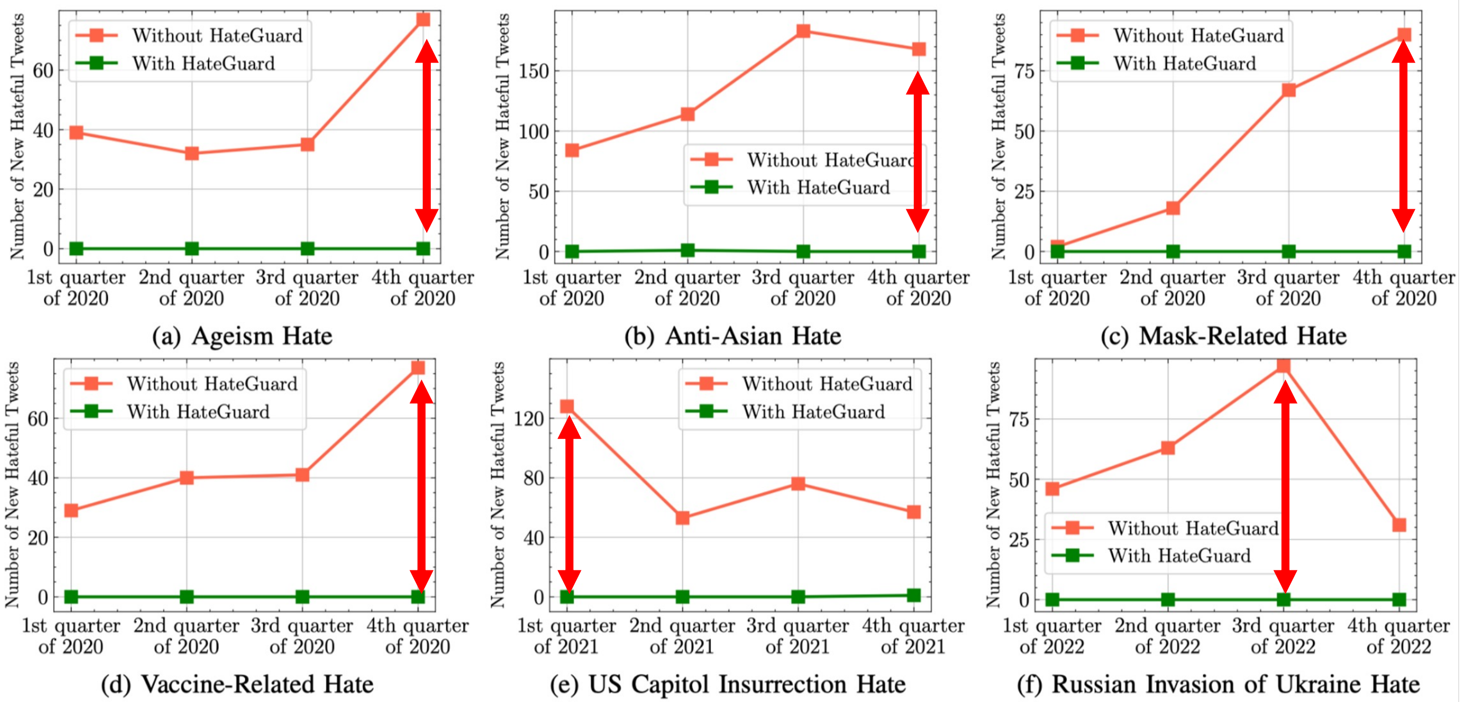
- Do the detected derogatory terms in the Derogation sub-problem ***inciteful*** of hate toward the detected targets in the first sub-problem?

HateCoT Prompting Strategy



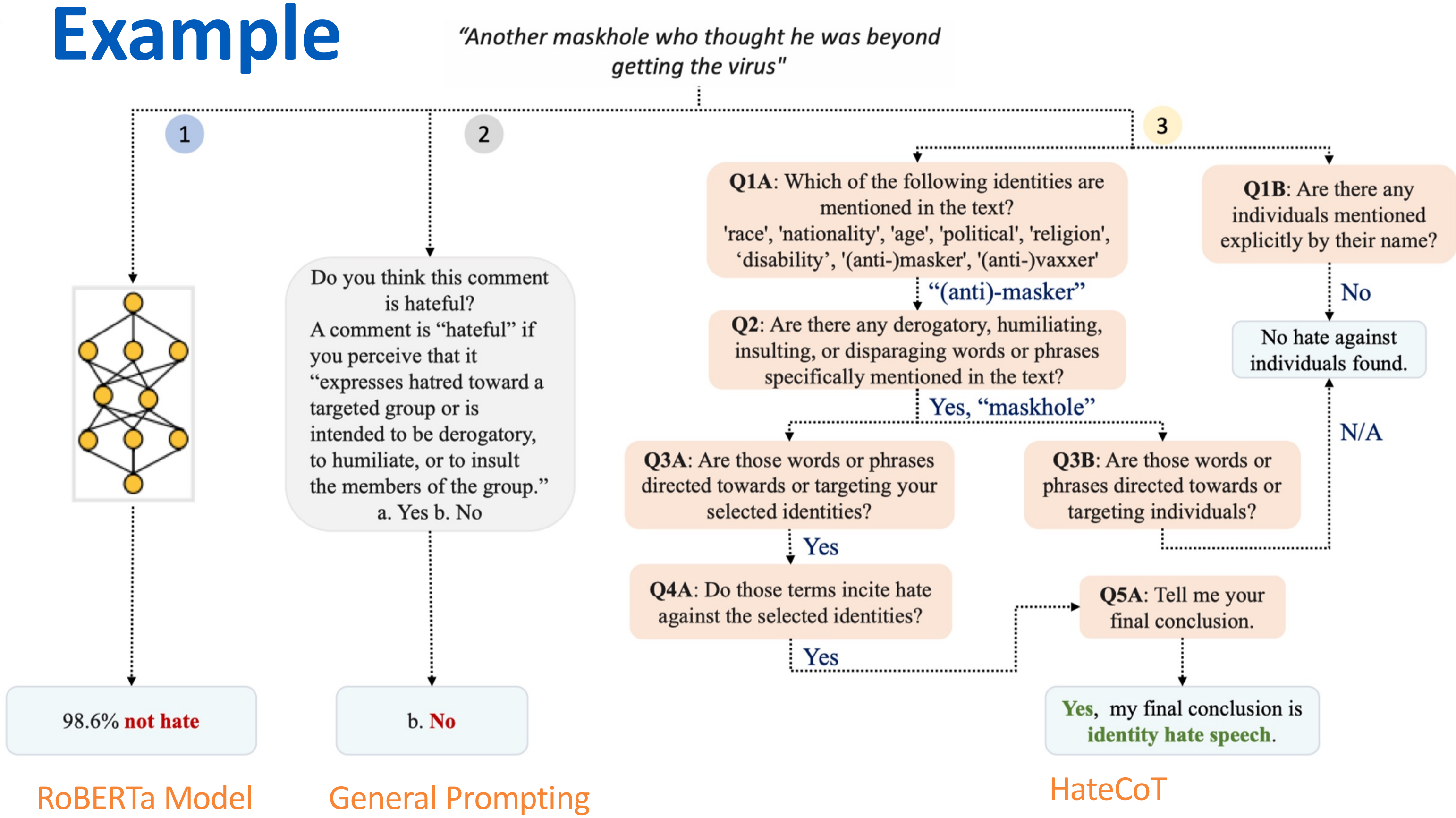
- Solving the main problem by putting the results of the ***sub-problems*** together

Evaluation



Deploying HateGuard in the period of 2020 (COVID-19), 2021 (US Capitol insurrection), and 2022 (Russian invasion) shows that new wave peaks are significantly reduced (green line)

Example



HateCoT for new waves decision-making (3) compared to traditional RoBERTa hate speech detection model (1) and general prompting (2)

HateGuard Against Evolving Online Hate

Wave Type	Method	Quarter 1 (Jan-Mar)				Quarter 2 (Apr-Jun)				Quarter 3 (Jul-Sep)				Quarter 4 (Oct-Dec)			
		# of Tweets	**Acc-uracy**	**Prec-ision**	**Rec-all**	**# of Tweets**	**Acc-uracy**	**Prec-ision**	**Rec-all**	**# of Tweets**	**Acc-uracy**	**Prec-ision**	**Rec-all**	**# of Tweets**	**Acc-uracy**	**Prec-ision**	**Rec-all**
- Overall Results -																	
Total (2020-2022)	**HateGuard**		**0.95**	**0.95**	**0.94**		**0.94**	**0.94**	**0.93**		**0.94**	**0.94**	**0.93**		**0.94**	**0.95**	**0.92**
	BERT-base	928	0.74	0.81	0.34	893	0.82	0.76	0.71	1148	0.84	0.82	0.79	1031	0.83	0.86	0.8
	Tweet-NLP		0.7	0.73	0.23		0.83	0.79	0.77		0.84	0.83	0.8		0.83	0.84	0.8
- Category-wise Results -																	
Ageism (2020)	**HateGuard**		**0.94**	**0.91**	**0.92**		**0.95**	**0.95**	**0.95**		**0.95**	**0.95**	**0.95**		**0.95**	**0.94**	**0.96**
	BERT-base	186	0.82	0.6	0.44	117	0.8	0.68	0.53	114	0.79	0.68	0.6	161	0.74	0.72	0.76
	Tweet-NLP		0.79	0.5	0.15		0.87	0.79	0.72		0.86	0.74	0.83		0.72	0.79	0.57
Asian (2020)	**HateGuard**		**0.96**	**0.96**	**0.97**		**0.93**	**0.93**	**0.93**		**0.94**	**0.95**	**0.94**		**0.95**	**0.94**	**0.98**
	BERT-base	179	0.68	0.91	0.35	296	0.84	0.79	0.8	331	0.85	0.86	0.87	262	0.87	0.88	0.92
	Tweet-NLP		0.63	0.77	0.29		0.84	0.84	0.72		0.84	0.86	0.84		0.85	0.91	0.86
Mask (2020)	**HateGuard**		**0.99**	**0.99**	**0.99**		**0.94**	**0.96**	**0.88**		**0.98**	**0.95**	**0.97**		**0.96**	**0.97**	**0.94**
	BERT-base	16	0.75	0	0	64	0.79	0.78	0.39	249	0.85	0.75	0.66	199	0.8	0.75	0.86
	Tweet-NLP		0.94	0.67	**0.99**		0.86	0.85	0.61		0.87	0.75	0.78		0.84	0.8	0.88
Vaccine (2020)	**HateGuard**		**0.98**	**0.99**	**0.96**		**0.92**	**0.9**	**0.92**		**0.93**	**0.93**	**0.91**		**0.94**	**0.95**	**0.92**
	BERT-base	78	0.76	0.92	0.38	114	0.78	0.68	0.7	104	0.85	0.79	0.83	226	0.84	0.75	0.79
	Tweet-NLP		0.72	0.77	0.35		0.75	0.6	0.83		0.8	0.75	0.73		0.88	0.83	0.82
US Capitol (2021)	**HateGuard**		**0.91**	**0.91**	**0.88**		**0.99**	**0.97**	**0.99**		**0.9**	**0.89**	**0.9**		**0.9**	**0.9**	**0.9**
	BERT-base	311	0.68	0.79	0.31	112	0.85	0.85	0.83	158	0.82	0.85	0.76	123	0.84	0.78	0.89
	Tweet-NLP		0.63	0.7	0.16		0.83	0.79	0.87		0.82	0.87	0.72		0.78	0.75	0.79
Russia -Ukraine (2022)	**HateGuard**		**0.95**	**0.95**	**0.93**		**0.94**	**0.95**	**0.92**		**0.94**	**0.95**	**0.93**		**0.9**	**0.92**	**0.9**
	BERT-base	158	0.8	0.85	0.37	190	0.8	0.72	0.62	192	0.82	0.82	0.81	60	0.83	0.82	0.87
	Tweet-NLP		0.77	0.92	0.24		0.84	0.82	0.65		0.86	0.88	0.84		0.85	0.89	0.81

Comparing HateGuard against the existing benchmarks

Conclusion and Future Work

- Conclusion
  - A large-scale experiment to study the nature of new waves of online hate
  - Examining the capabilities of the existing moderation tools
  - A novel framework to address the problem of new waves of online hate
- Future work
  - Multilingual new waves of online hate
  - Multimodal scenarios, such as hateful memes
  - Auto-prompting methodologies

Discussion

- LLMs for addressing evolving cyber security issues
  - Fake news/Disinformation
  - Zero-day attacks
  - Phishing attacks
  - Advanced Persistent Threats
  - …

Discussion

Method	Paper	Source
Chain-of-Thought	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models	NeurIPS 2022
Self-consistency	Self-Consistency Improves Chain of Thought Reasoning in Language Models	ICLR 2023
Least-to-Most	Least-to-Most Prompting Enables Complex Reasoning in Large Language Models	ICLR 2023
Tree of Thought	Tree of Thoughts: Deliberate Problem Solving with Large Language Models	ArXiv 2023
In-Context Learning	Teaching Algorithmic Reasoning via In-context Learning	NeurIPS 2022
Analogical Prompting	Large Language Models as Analogical Reasoners	ArXiv 2023
PromptBreeder	Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution	ArXiv 2023
Autoprompt	AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts	EMNLP 2020

Thank you !

Nishant Vishwamitra,
Assistant Professor

nishant.vishwamitra@utsa.edu

UTSA,
San Antonio, TX,
United States