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
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
Introduction

- State-of-the-art performance

<table>
<thead>
<tr>
<th>Simulated exams</th>
<th>GPT-4 estimated percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform Bar Exam (MBE+MEE+MPT)(^1)</td>
<td>298/400</td>
</tr>
<tr>
<td>LSAT</td>
<td>163</td>
</tr>
<tr>
<td>SAT Evidence-Based Reading &amp; Writing</td>
<td>710/800</td>
</tr>
<tr>
<td>SAT Math</td>
<td>700/800</td>
</tr>
<tr>
<td>Graduate Record Examination (GRE) Quantitative</td>
<td>163/170</td>
</tr>
<tr>
<td>Graduate Record Examination (GRE) Verbal</td>
<td>169/170</td>
</tr>
<tr>
<td>Graduate Record Examination (GRE) Writing</td>
<td>4/6</td>
</tr>
<tr>
<td>USABO Semifinal Exam 2020</td>
<td>87/150</td>
</tr>
<tr>
<td>USNCO Local Section Exam 2022</td>
<td>36/60</td>
</tr>
<tr>
<td>Medical Knowledge Self-Assessment Program</td>
<td>75%</td>
</tr>
<tr>
<td>Codeforces Rating</td>
<td>392</td>
</tr>
<tr>
<td>AP Art History</td>
<td>5</td>
</tr>
<tr>
<td>AP Biology</td>
<td>5</td>
</tr>
<tr>
<td>AP Calculus BC</td>
<td>4</td>
</tr>
</tbody>
</table>

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

$$\hat{x}_i = p(\hat{x}_i|x_1, x_2, \ldots, x_n)$$  \hspace{1cm} \text{Sentence correction (denoising)}

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

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

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...

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†

*The University of Texas at San Antonio
†University at Buffalo
‡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

![Graph showing Anti-Asian American Hate Crime Trend Analysis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9168424/)

**United Nations**

Global perspective Human stories

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

2. 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-4ba7-8f80-48b488ae5ae](https://www.ft.com/content/abaf9ea7-c5dc-4ba7-8f80-48b488ae5ae)
Dataset

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

<table>
<thead>
<tr>
<th>New Wave Type</th>
<th>Number of hateful tweets</th>
<th>Number of non hateful tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 tweets</td>
<td>1,096</td>
<td>1,600</td>
</tr>
<tr>
<td>US Capitol Insurrection tweets</td>
<td>314</td>
<td>390</td>
</tr>
<tr>
<td>Russian Invasion of Ukraine tweets</td>
<td>237</td>
<td>363</td>
</tr>
<tr>
<td>Total tweets</td>
<td>1,647</td>
<td>2,353</td>
</tr>
</tbody>
</table>

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?

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**

<table>
<thead>
<tr>
<th>Detection Tools</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarifai Text Moderation [67]</td>
<td>0.69</td>
<td>0.16</td>
<td>0.27</td>
</tr>
<tr>
<td>Perspective API [14]</td>
<td>0.49</td>
<td>0.31</td>
<td>0.38</td>
</tr>
<tr>
<td>Azure Text Moderation [15]</td>
<td>0.54</td>
<td>0.21</td>
<td>0.31</td>
</tr>
</tbody>
</table>

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

Current process takes **months** to complete!

Existing tools
HateGuard Design

- Reasoning-based decision-making for detection:
  - Hate detection is a complex and contextual decision-making process.
  - Several intermediate steps to arrive at the final decision.
  - 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.

  Leveraging Large Language Models (LLMs):
  - *Chain-of-Thought* prompting for intermediate steps and decision-making.
  - *Automatic prompt* updates and updating targets and derogatory terms.
HateGuard Overview
HateCoT Prompting Strategy

Q1a: Which of the following identities are mentioned in the text? 'race', 'nationality', 'age', 'political', 'religion', 'disability', '(anti)-masker', '(anti)-vaxxer'

Q1b: Are there any individuals mentioned explicitly by their names in the text?

Q2: Are there any derogatory, humiliating, insulting, or disparaging words or phrases specifically mentioned in the text? (Note: Colloquially usage of the words should not be considered)

Q3a: If Q2's answer is 'Yes', are those words or phrases directed towards or targeting your selected identities?

Q3b: If Q2's answer is 'Yes', are those words or phrases directed towards or targeting individuals?

Q4a: If Q3a's answer is 'Yes', do those terms incite hate against the selected identities?

Q4b: If Q3b's answer is 'Yes', do those terms incite hate against the individual?

Q5a: If Q4a's 'Yes', the comment can be concluded as identity hate speech. Tell me your final conclusion.

Q5b: If Q4b's 'Yes', the comment can be concluded as individual hate speech. Tell me your final conclusion.

1. Target Presence
2. Derogation
3. Direction
4. Incitation
5. Decision
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

1. Target Presence
2. Derogation
3. Direction
4. Incitation
5. Decision
HateCoT Prompting Strategy

1. Target Presence
   - Which of the following identities are mentioned in the text? 'race', 'nationality', 'age', 'political', 'religion', 'disability', '(anti-)masker', '(anti-)vaxzer'
   - Are there any individuals mentioned explicitly by their names in the text?

2. Derogation
   - Are there any derogatory, humiliating, insulting, or disparaging words or phrases specifically mentioned in the text?
   - Are those words or phrases directed towards or targeting your selected identities?
   - Are those words or phrases directed towards or targeting individuals?

3. Direction
   - If Q2's answer is 'Yes', are those words or phrases directed towards or targeting your selected identities?
   - If Q3b's answer is 'Yes', do those terms incite hate against the selected identities?
   - If Q3b's answer is 'Yes', do those terms incite hate against the individual?

4. Incitation
   - If Q4a's 'Yes', the comment can be concluded as identity hate speech. Tell me your final conclusion.

5. Decision
   - If Q4b's 'Yes', the comment can be concluded as individual hate speech. Tell me your final conclusion.

- 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

- **Q1a**: Which of the following identities are mentioned in the text? 'race', 'nationality', 'age', 'political', 'religion', 'disability', '(anti-)masker', '(anti-)vaxxer'
- **Q1b**: Are there any individuals mentioned explicitly by their names in the text?
- **Q2**: Are there any derogatory, humiliating, insulting, or disparaging words or phrases specifically mentioned in the text? (Note: Colloquial usage of the words should not be considered)
- **Q3a**: If Q2's answer is 'Yes', are those words or phrases directed towards or targeting your selected identities?
- **Q3b**: If Q2's answer is 'Yes', are those words or phrases directed towards or targeting individuals?
- **Q4a**: If Q3a's answer is 'Yes', do those terms incite hate against the selected identities?
- **Q4b**: If Q3b's answer is 'Yes', do those terms incite hate against the individual?
- **Q5a**: If Q4a's 'Yes', the comment can be concluded as identity hate speech. Tell me your final conclusion.
- **Q5b**: If Q4b's 'Yes', the comment can be concluded as individual hate speech. Tell me your final conclusion.

- **1** Target Presence
- **2** Derogation
- **3** Direction
- **4** Incitation
- **5** Decision

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

<table>
<thead>
<tr>
<th>Wave Type</th>
<th>Method</th>
<th>Quarter 1 (Jan-Mar)</th>
<th>Quarter 2 (Apr-Jun)</th>
<th>Quarter 3 (Jul-Sep)</th>
<th>Quarter 4 (Oct-Dec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># of Tweets</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Total (2020-2022)</td>
<td>HateGuard</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>BERT-base</td>
<td>0.74</td>
<td>0.81</td>
<td>0.34</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Tweet-NLP</td>
<td>0.7</td>
<td>0.73</td>
<td>0.23</td>
<td>0.83</td>
</tr>
</tbody>
</table>

### Compare HateGuard against the existing benchmarks

#### - Overall Results -

<table>
<thead>
<tr>
<th>Wave Type</th>
<th>Method</th>
<th>Quarter 1 (Jan-Mar)</th>
<th>Quarter 2 (Apr-Jun)</th>
<th>Quarter 3 (Jul-Sep)</th>
<th>Quarter 4 (Oct-Dec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HateGuard</td>
<td>0.94</td>
<td>0.91</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>BERT-base</td>
<td>0.82</td>
<td>0.6</td>
<td>0.44</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Tweet-NLP</td>
<td>0.94</td>
<td>0.96</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>HateGuard</td>
<td>0.68</td>
<td>0.6</td>
<td>0.35</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>BERT-base</td>
<td>0.63</td>
<td>0.77</td>
<td>0.29</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Tweet-NLP</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Mask (2020)</td>
<td>0.94</td>
<td>0.67</td>
<td>0.99</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>HateGuard</td>
<td>0.98</td>
<td>0.9</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>BERT-base</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Tweet-NLP</td>
<td>0.94</td>
<td>0.91</td>
<td>0.88</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Vaccine (2020)</td>
<td>0.91</td>
<td>0.9</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>US Capitol (2021)</td>
<td>0.95</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Russia-Ukraine (2022)</td>
<td>0.95</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>
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

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

Nishant Vishwamitra,
Assistant Professor

nishant.vishwamitra@utsa.edu

UTSA,
San Antonio, TX,
United States