Guardrails for LLMs Supporting Security

SuperUROP
Advanced Undergraduate Research Opportunities Program

Anahita Srinivasan, Ethan Garza, Dr. Erik Hemberg, Dr. Una-May O'Reilly
Nadar Foundation Undergraduate Research and Innovation Scholar
Anyscale Learning For All (ALFA) Group | December 2023

MIT EECS
MIT CSAIL Computer Science & Artificial Intelligence Laboratory

Motivation and Goal

Motivation:
Current method of testing cybersecurity networks: create PDDL (planning domain definition language) files by hand

Goal: make this work easier and quicker
- Use an LLM (large language model) to generate the PDDL files
- Develop cybersecurity subject-matter expertise for accuracy

Goal: connect LLMs with cybersecurity database (BRON) to build guardrails to constrain the output using:
1. Generative power of LLMs
2. Facts + structure of the database

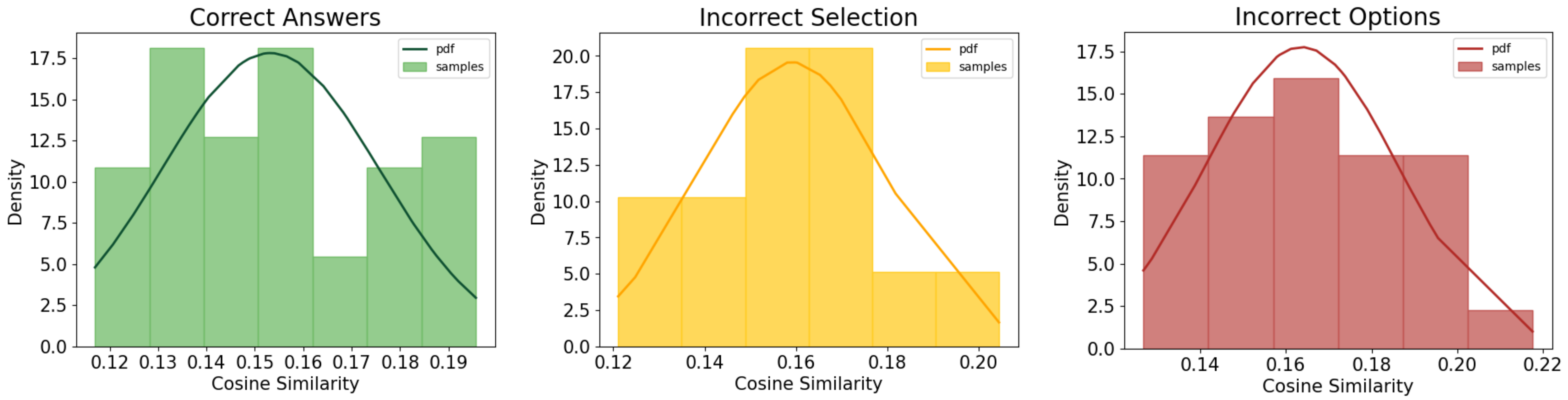
Methods

Test GPT-3.5's Retrieval Capabilities:
Match cybersecurity use case description with technique definition
1. Embed cyber scenario retrieved from BRON
2. Calculate cosine similarity
3. Retrieve most similar techniques from BRON
4. Pass prompt into GPT-3.5
5. Collect GPT-3.5 's final answer

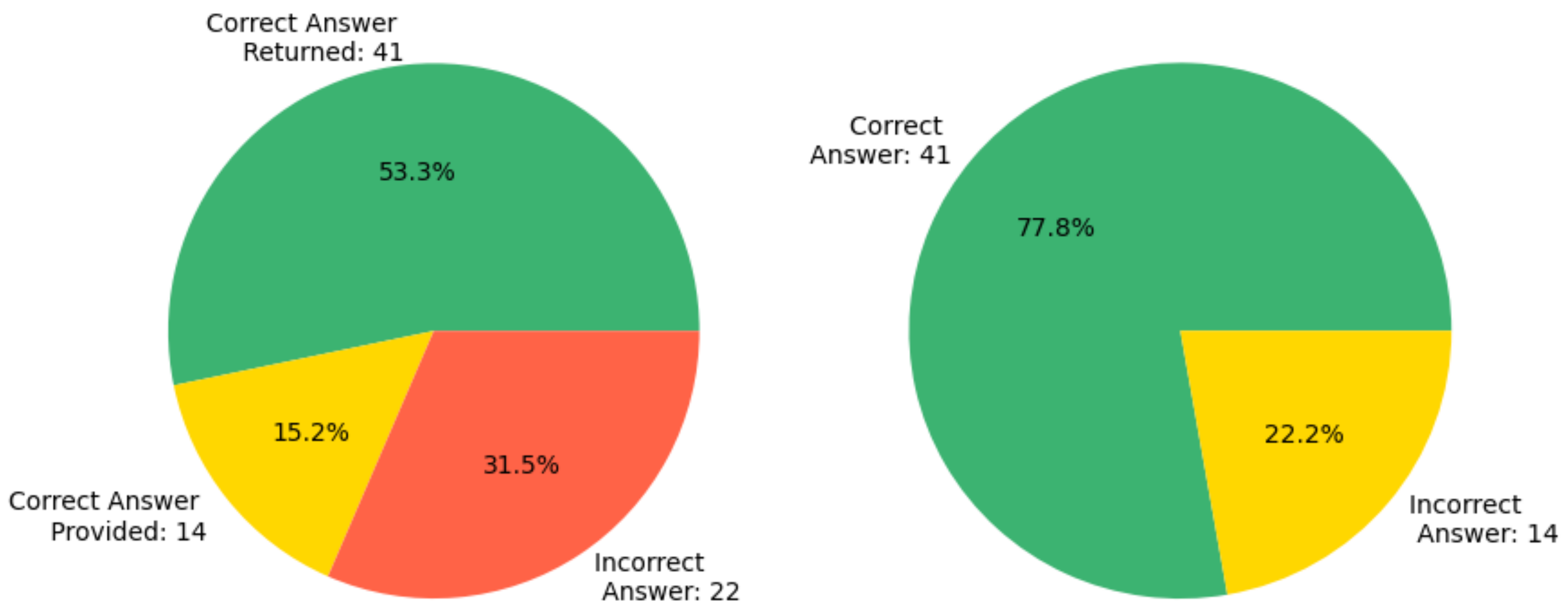
Output Parsing:
Process to write PDDL actions and fill out PDDL domain file
1. Pull all action use cases from BRON
2. Pass use case description + predicates list to GPT-3.5
3. Extract action code
4. Parse code + extract predicates
5. Write action code to JSON

Results

Retrieval Capabilities: There are three main categories of response: GPT-3.5 returned correct use case, GPT-3.5 was provided correct answer but did not choose it, and GPT-3.5 was not provided the correct answer.



In general, the cosine similarity was higher between use case and selected answer when the answer was correct as opposed to when the answer was incorrect.



Provided that the correct answer was one of the options given to GPT-3.5, it returned that correct answer 77.8% of the time.
- Promising prospects for retrieval capabilities
- Can match cybersecurity subject-matter expert

Output Parsing:
- Can parse a PDDL action into its parameters, preconditions, and effects
- Figure on right displays the action parts
- Iteratively reconstruct domain file predicates and types from actions

:action
:parameters
:precondition
:effect
:pos_effects
:neg_effects
:pos_preconditions
:neg_preconditions

Future Work

Retrieval Capabilities:
- Integrate subject matter expertise into generated PDDL files

Output Parsing:
- Increase complexity of generated PDDL file
- Move from benchmark testing to cybersecurity domain

References

[1] Ethan Garza, Erik Hemberg, Stephen Moskal, and Una-May O'Reilly, Assessing Large Language Model's Knowledge of Threat Behavior in MITRE ATT&CK, In Proceedings of the 3rd Workshop on Artificial Intelligence-Enabled Cybersecurity Analytics, 2023.

[2] Erik Hemberg, Matthew Turner, Nick Rutar, and Una-May O'Reilly, Enhancements to Threat, Vulnerability, and Mitigation Knowledge For Cyber Analytics, Hunting, and Simulations, Association of Computer Machinery Digital Threats: Research and Practice, 2023.

[3] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone, LLM+P: Empowering Large Language Models with Optimal Planning Proficiency, Computing Research Repository, 2023.

[4] Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati, On the Planning Abilities of Large Language Models (A Critical Investigation with a Proposed Benchmark), Computing Research Repository, 2023.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, In Proceedings of the 36th Conference on Neural Information Processing Systems, 2022

[6] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Lu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun, A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, Computing Research Repository, 2023.