Gryphon Scientific
IARPA BENGAL Lightning Talk

Dr. Margaret Rush

Chief Scientific Officer, Interim Director of Data Science

October 24, 2023

GRYPHON SCIENTIFIC

Gryphon Scientific

Innovative Analytic Solutions

Gryphon Scientific is a small consulting company known for helping clients take creative, evidenced-based approaches to challenging issues of health, safety and security in the US and abroad. Since our founding in 2005, we have built and maintained a reputation for scientific rigor and analytic excellence.

Topical Domain Expertise

Gryphon is recognized by both US Government and Industry as an expert in LLM threat modes and vulnerabilities as they intersect with both biological sciences and infrastructure protection.

Threat Detection & Mitigation

We develop tools and approaches to efficiently probe LLM models to detect, characterize and mitigate biases, threats or vulnerabilities

Deploying LLMs with bias mitigation and sourcing

Through SBIRs and other vehicles we have developed technologies to identify and combat unwarranted bias and toxic outputs and preserve source attribution.

Systematic Understanding and Explanation of Misinformation Online (SUEMO)

PROJECT BRIEF

Gryphon is developing a digital toolbox that identifies online vaccine misinformation and leverages a highly curated knowledge base and LLMs to provide the current state of science on a given misinformation topic. The project provided a multi-functional application that allows a variety of users to monitor vaccine misinformation on social media and retrieve sourced LLM-generated responses that are constrained to our knowledge base.

CLIENT

NIAID

PRACTICE AREA

Data Science

HIGHLIGHTS

Developed an automated Twitter scraping pipeline that collects, processes, and stores COVID-19, influenza, and vaccine-related tweets. We then built a Deep Learning classifier to designate whether the text was misinformation.

Developed an ML-based text parsing and processing pipeline to convert ~3,500 publications into machine readable files. Targeted text were then extracted and transformed into and stored as text embeddings for downstream analysis.

Designed a document retrieval- and LLM-based analysis pipeline to provide clear and concise explanations of vaccine misinformation based on the most relevant text from our vector database. System functionalities have been integrated into a web application prototype.

Developing Strategies to Assess the Risks of LLM Misuse for Biological Threats

PROJECT BRIEF

Gryphon is partnering with leading three LLM developers and government entities to develop strategies and means to assess the current and future risks of misuse of LLMs for enabling the malicious application of biological knowledge. This work is focused on biologically-based attacks, and weighs the risks versus potential benefits of LLMs and other AI technologies for advancing biological knowledge.

CLIENT

Leading LLM developers; USG; G7 government

PRACTICE AREA

Biosafety, Biosecurity, and Emerging Technologies

HIGHLIGHTS

Gryphon is developing question-and-answer panels aimed at assessing current and predicting future capabilities of LLMs in the biological sciences. These panels probe LLM utility for providing both general biological and biological warfare-related knowledge.

With collaborators, Gryphon is exploring to what extent LLMs can enable adversaries with respect to the planning and execution of various biological attack scenarios. This live test will inform our understanding of the potential misuse of LLMs by adversaries with different knowledge backgrounds.

Gryphon is developing "Red Lines" for LLM capability thresholds, at which their risk of enabling the misuse of biology likely outweighs their benefits to science and society. These Red Lines will be used to guide future LLM development and evaluation by industry.

What Gryphon can contribute to BENGAL

- We bring sophisticated domain expertise in health, biology, and infrastructure protection, including smart cities, to develop domain specific threat taxonomies
- We have experience developing tools and technologies to identify, characterize and mitigate LLM threats for both government stakeholders and corporations developing LLMs
- A team of domain experts experienced at using and Red Teaming a variety of popular LLMs
- Technology to assist in identifying misinformation/toxic outputs
- Technology that enables source attribution in LLMs

Contact us about teaming!

DR. MARGARET RUSH

Phone: +1 301-270-0647
margaret@gryphonscientific.com

GRYPHON OFFICES

6930 Carroll Avenue
Suite 900
Takoma Park, MD 20912

https://www.gryphonscientific.com

GRYPHON SCIENTIFIC