Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

BENGAL

BIAS EFFECTS AND NOTABLE GENERATIVE AI LIMITATIONS

INTELLIGENCE VALUE

Large language models (LLMs) present massive opportunities to increase the quality and efficiency of intelligence analysis; however, LLMs are known to exhibit vulnerabilities and enable malicious behavior that poses unacceptable risks. BENGAL aims to understand the landscape of LLM biases, threats and vulnerabilities, with the goal of developing novel technologies to analyze and address these shortcomings, enabling the Intelligence Community (IC) to safely harness LLM technology across a broad range of critical applications.

The IC is interested in safe uses of LLMs (multi-modal and text-only) for a wide variety of applications including the rapid summarization and contextualization of relevant multilingual information. These applications must avoid unwarranted biases and toxic outputs, preserve attribution to original sources, and be free of erroneous outputs. The US Government is also interested in identifying and mitigating hazardous use of LLMs by potentially nefarious actors.

The BENGAL super seedling program is a two-year effort beginning in 2024 that aims to explore, quantify, and mitigate the threats and vulnerabilities of LLMs (multi-modal and text-only). Performers will focus on one or more of the topic domains below, clearly articulate a taxonomy of threat modes within their domain of interest and develop technologies to detect, and defeat or mitigate these threats.

Biases and induction of diverse analytical perspectives: methods to characterize and detect biases present in LLMs and ways to leverage LLMs to induce diverse perspectives.

AI hallucinations and inferences: Techniques to detect and mitigate an ungrounded, incorrect, or misleading output, or hallucination, from an LLM while minimizing the impact on the LLM's capabilities to produce correct and/or plausible inferences.

Safe information flow in sensitive environments: Methods to identify inputs/outputs from a(n) user/LLM that may be trying to aggregate innocuous facts to derive sensitive information, methods to decouple sensitive information, and methods to "unlearn" information deemed sensitive from a pre-trained or fine-tuned LLM.

Working resiliently to improve poisoned sources: Techniques to improve LLM reliability through evaluating of the quality of a sources, explainable techniques to infer source intentions, and extracting reliable information from biased or incomplete sources.

Performers will pursue high-risk, high-payoff research and deliver to the IC turn-key prototype software which is independently validated by testing and evaluation partners.

PRIME PERFORMERS

- TBD

TESTING AND EVALUATION PARTNERS

- TBD

KEYWORDS

- Generative AI
- Human Language Technology
- Artificial Intelligence
- Large Language Models

PROGRAM MANAGER

Tim McKinnon, Ph.D.
Phone: (301) 243-2084
timothy.mckinnon@iarpa.gov

www.iarpa.gov @IARPAnews linkedin.com/company/iarpa-odni

Latest Revision: 12/19/2023