Intelligence Advanced Research Projects Activity

Technical Description

Bias Effects and Notable Generative AI Limitations (BENGAL)

Targeted Super Seedling

BAA-24-03

Release Date: 20 October 2023

Table of Contents

Contents

1. Program Overview

The Intelligence Advanced Research Projects Activity (IARPA) invests in high-risk/high-payoff research programs that have the potential to provide our nation with an overwhelming intelligence advantage. IARPA seeks to develop new capabilities to enable the safe adoption and use of generative AI technologies to greatly enhance the effectiveness and efficiency of the Intelligence Community (IC).

Large Language Models (LLMs) exhibit impressive, human-seeming conversational capabilities. LLMs and applications that build 'on top' of these models are being rapidly adopted and are expected to transform work across diverse sectors. It is anticipated that the public will interact with a massive number of LLM-derivative technologies within this decade. Even at this early stage in their adoption, however, the public has observed that LLMs can exhibit erroneous or potentially harmful behavior. The inherent characteristics of LLMs (ease of use, human-like dialogue, complexity, and lack of explainability) present vulnerabilities for benign applications and enable hostile applications. Models may conceal threats to users, including quick generation of mis/disinformation or elicitation of sensitive information. These threat modes may be unintended emergent artifacts of training complex models on vast and poorly understood training data, or they may be intentionally incorporated into models by their designers.

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 information relevant to the IC. 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 potential nefarious actors.

The goal of the BENGAL targeted super seedling is to understand LLM threat modes, quantify them and to find novel methods to correct threats and vulnerabilities or to work resiliently with imperfect models. IARPA seeks to develop and incorporate novel technologies to efficiently probe large language models to detect and characterize LLM threat modes and vulnerabilities. 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 serve as an analog to 'virus scan' software.

2. Technical Challenges and Objectives

IARPA seeks novel research ideas from multidisciplinary teams pursuing advanced research topics capable of supporting the interests described below:

- **Biases and induction of diverse analytical perspectives:** The IC seeks capabilities to enhance awareness of LLM output biases (cognitive, demographic, ideological, cultural, temporal, etc.) that might suppress or promote relevant content with the effect of misleading a user. IC users also greatly benefit from exposure to contrasting perspectives on the same event or situation. IARPA is interested in novel technologies to accurately and automatically characterize and detect biases, as well as leverage LLMs to induce diverse perspectives on events or states of affairs.

- **AI hallucinations and inferences:** Generative LLMs are known to produce spurious, ungrounded outputs ('hallucinations' or 'confabulations') that can cause erroneous analysis and decision making. Successful methods for reducing hallucinations constrain

LLM outputs to those that are in some fashion corroborated by ground truth (e.g., quotations from a trusted document). However, constraints enforcing grounded outputs are over-restrictive and block LLMs from drawing correct or plausible inferences, since good inferences are often not (straightforwardly) based on ground truth evidence. IARPA is interested in capabilities that detect spurious hallucinations while maximizing and inducing correct and plausible inferences.

- **Safe information flow in sensitive environments:** The IC's classification system limits access to sensitive information, which if disclosed to unauthorized individuals poses a grave threat to national security. However, restricting information to certain people and systems can pose a threat to national security by preventing critical collaborations and timely sharing of critical information between IC organizations. IARPA is interested in LLM technologies that increase the flow of information while minimizing the likelihood of sensitive information disclosure, enabling safe use of LLMs over the broadest range of data and tasks. (Please note that performers will not be permitted access or test systems on sensitive data; thus, offerors must propose unclassified testing and evaluation schemes that credibly simulate environments where sensitive information is stored but not all users are authorized to access it.)
- **Working resiliently with imperfect or poisoned sources**: The IC is interested in novel LLM techniques to evaluate the reliability of specific information sources (e.g., news organizations, individual content producers from around the world), especially sources whose content is used to train LLMs or to interpret developing situations using LLMs in scenarios where only sparse information is available. The IC is also interested in technologies that enable analysts and other users to work resiliently with imperfect or malicious sources, identifying within their content reliable information where possible.

Efforts addressing these topic areas align well with needs of the intelligence and national security communities and are, therefore, under the purview of IARPA's research mission. Successful technological solutions will require creative, multidisciplinary methods, paradigm changing thinking, and transformative approaches. Preference will be given to research with the ability to revolutionize capabilities or demonstrate that revolutionary change is possible in the coming decade.

This BAA solicits short-term, limited scope research in topic areas that are not addressed by emerging or ongoing Government programs or other published solicitations. It is primarily, but not solely, intended for early-stage research that may lead to larger, focused programs through a separate BAA in the future.

3. Program Phases

Seedlings are structured as a Phase A base with a Phase B option. Phase A represents an initial proof of concept of the proposed approach. Phase B, if exercised, will build upon the proof-of-concept research in Phase A to deliver a demonstration. Phase A shall be of a duration of 12 months to demonstrate a prototype proof-of-concept, with preliminary software deliverables and performance evaluation reports due at months 4, 7 and 10. BENGAL performers are expected to propose and implement their own testing and evaluate protocols. An independent testing and evaluation (T&E) team will verify performer results and validate software performance. Independent evaluation teams have not been selected for this effort, but organizations serving in this capacity may include Government agencies, Federally Funded Research and Development

Corporations (FFRDCs), University Affiliated Research Centers (UARCs), or Department of Energy Labs. At the conclusion of Phase A, performers shall submit a final report. Reports and deliverables shall be used in evaluation of projects for continuation to Phase B. Phase B will be 12 months in duration. Shorter duration projects, if appropriate for the subject matter, may be considered. See Figure 1 for a proposed project timeline.

Phase A (12 Months)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Program meeting (program kickoff, PI meeting, demos)	🟨									🟨		
Gov't visits performer site			🟨						🟨			
Performer Self-Evaluation Milestone/				🟨			🟨			🟨		
T&E validation of results and system					🟨			🟨			🟨	
Performer deliver final report and technical products												🟨
Phase B (12 Months)	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Program meeting (phase kickoff, PI meeting, demos)	🟦									🟦		🟦
Gov't visits performer site			🟦						🟦			
Performer Self-Evaluation Milestone				🟦			🟦			🟦		
T&E validation of results and system					🟦			🟦			🟦	
Performer deliver final report and technical products												🟦

Figure 1: Proposed Phase A + B timeline with key activities.

White papers and proposals must explicitly address how the offeror's technical approach will enable the safe adoption and use of generative AI technologies within the IC. Offerors shall demonstrate that the proposed effort has the potential to make revolutionary, rather than incremental, improvements to current capabilities. Research that primarily results in evolutionary improvement to the existing state of practice is specifically excluded.

White papers and proposals must include offeror-defined objectives, as well as milestones and performance metrics as task-driven intermediate steps towards the objectives. Offerors must clearly articulate tasks, quantitative metrics, and metric targets. Good metrics for the purpose of this effort maximize interpretability, allow easy implementation/replicability, allow calculation of confidence bounds, furnish results that are comparable over time, and are actionable (i.e., inform changes in technical approach).

Testing and evaluation of performer systems will occur 3 times during each of the 12-month phases. As noted above, successful submissions will clearly articulate a testing and evaluation protocol that can be either run on the performer site and validated by the T&E team or replicated by T&E using T&E's own infrastructure. Metric targets should enable comparison with state-of-the-art or a well-justified baseline. Offerors must state target values for each milestone and justify why these values are challenging given the current state of the art. **White papers and proposals that do not provide explicit, feasible, and replicable protocols to measure progress will not be considered for award.**

Successful projects must also contain the following elements:

- Developed capabilities must generalize across LLM text generation models and their different versions.
- Delivery of turn-key containerized software (Phase A) with user interface (UI) components and thorough documentation (Phase A and B); An independent test and evaluation (T&E) team will affirm that software can be successfully deployed with minimal developer effort. (Ease of deployment will be one of the criteria for advancement into the second program phase.)
- Explicitly stated model access limitations and, where relevant, provide justification why a particular method (e.g., black box) cannot be used.

4. Description of Topics and Areas of Interest

The following is a list of suggested subtopics. These are intended to provide to the offeror additional information concerning the Government's interests in the BENGAL topic areas. These are not considered an exhaustive list and offerors are free to propose projects which address one or more subtopics.

Topic #1 Biases and induction of diverse analytical perspectives:

Subtopics:

- Methods for objectively quantifying bias (e.g., relative to a specific collection of texts and other content)
- Computational techniques to characterize perspective spaces and measure the differences between perspectives.
- Using human-LLMs interactions to identify analysts' blind spots and induce perspectives representative of those blind spots.
- Induction of outputs representing diverse perspectives (e.g., "How would a particular group or organization interpret this event?")
- Deriving insights from simulated dialogue between LLMs with different biases/perspectives

Topic #2 AI hallucinations and inferences:

Subtopics:

- Methods to maximize the LLM's ability to produce valuable inferences in the absence of ground-truth evidence.
- Novel and explainable approaches to quantifying confidence of generative model output (e.g., to ensure trustworthiness for the user or enable generation of high-quality synthetic training data to reduce reliance on sensitive, sparse, or noisy data sources)

- Methods to investigate theoretical bases for LLM hallucinations or grounding (e.g., are hallucinations inevitable?)

Topic #3 Safe information flow in sensitive environments:

Subtopics:

- Targeted "unlearning" in pre-trained or fine-tuned LLMs (e.g., methods to remove from an LLM: information about an individual or information derived from a particular document deemed sensitive without otherwise affecting the performance of the model). IARPA is not interested in filtering outputs.
- Decoupling of sensitive information: Given a description of information deemed sensitive (e.g., source/method of collection), sanitizing a document or collection of documents such that sensitive information or ancillary information that could be used to infer sensitive information is verifiably removed, while retaining the meaning of the original document(s).
- Methods to identify when an aggregation of innocuous facts can be used to derive specific sensitive information. Given a set of queries, quantify the likelihood that the user is trying to access a particular piece of sensitive information from an LLM. Alternatively, given a set of LLM responses, quantify the likelihood that the LLM is trying to access sensitive information from the user.

Topic #4 Working resiliently with imperfect or poisoned sources:

Subtopics:

- LLM techniques to evaluate the reliability of a given information source (e.g., individuals or organizations) either for the purpose of ensuring the integrity of training data or for evaluating incoming information
- Automated and explainable techniques for inferring source intentions
- Quantifying source corroboration
- Extracting reliable intelligence from incomplete or biased content

The following topics are out of scope for this seedling effort:

- research into approaches that do not generalize across LLM text generation models and their different versions;
- research focused on systems integration or engineering of existing approaches or instruments;
- cybersecurity research not primarily focused on LLM technology; research that will not result in functional prototype technology;
- approaches requiring access to classified data; and/or
- research which are resubmissions of work already awarded by the National Science Foundation, National Institutes of Health, Department of Defense, Intelligence Community, or other federal agencies.

5. Whitepaper Preparation Instructions

Offerors should submit a white paper in response to the BENGAL BAA. The Government will review white papers and recommend or not recommend submission of a full proposal. The white papers shall not exceed 3-pages summarizing Offeror qualifications and the Offeror's intended technical approach/solution to the BAA Topics and Areas of Interest.

White papers must concisely answer all the following:

1. Summarize your organization's/team's qualifications to perform research and development in the specific field of science and technology. Provide a short description of present and past performance of similar work.

2. Heilmeier questions (Address in relation to the technical approach/solution for your intended proposal):

    i. What are you trying to do?

    ii. How is it done at present? Who does it? What are the limitations of present approaches?

    iii. What is new about your approach? Why do you think that you can be successful at this time?

    iv. If you succeed, what difference will it make?

    v. How will you evaluate progress during and at the conclusion of the effort? (i.e., what are your proposed milestones and metrics?)

The white paper shall not describe management nor detailed cost/price information. All white papers shall be written in English. Additionally, text should be black and paper size 8-1/2 by 11-inch, white in color with 1" margins from paper edge to text or graphic on all sides. Submissions should also use Times New Roman font with font size not smaller than 12-point. Additionally, the font size for figures, tables and charts should not be smaller than 10-point. All contents shall be clearly legible with the unaided eye or the white paper may not be considered. White papers shall be submitted in a PDF format.

The Government anticipates white papers submitted under this BAA will be UNCLASSIFIED and that the deadline for submitting white papers will be approximately two weeks after the release of the full BAA solicitation. The official BAA will contain extensive instructions regarding structure and submission of the full proposal.