BENGAL Program Overview

Dr. Timothy McKinnon | Program Manager | BENGAL Proposers' Day, 24 October 2023

Intelligence Advanced Research Projects Activity

IARPA

Creating Advantage through Research and Technology

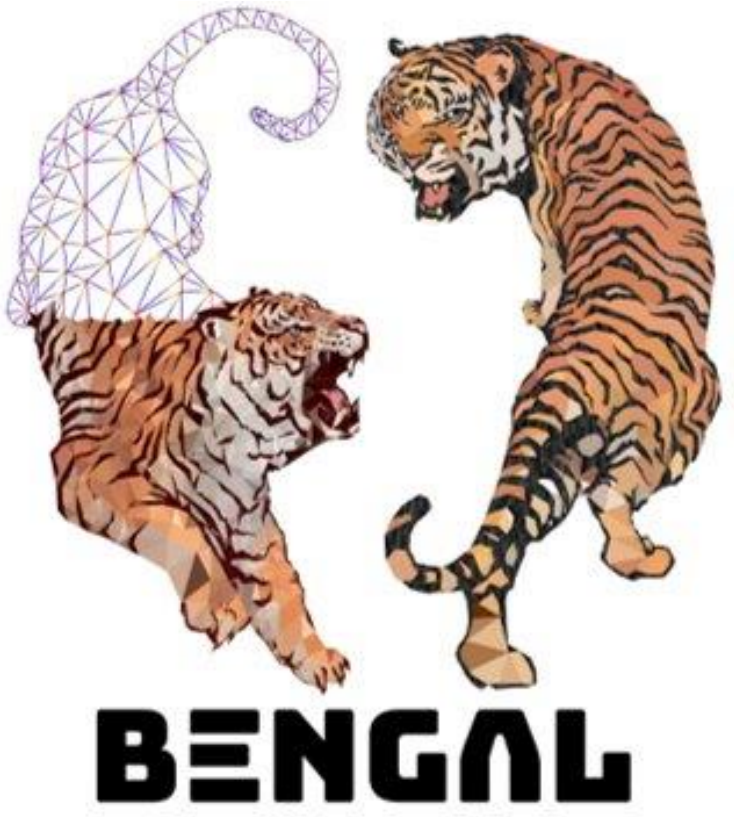
Technical Slides Disclaimer

- All images, references, and articles are included as illustrative examples only
- ODNI and IARPA do not endorse any product or company referenced within
- The draft technical document was released and additional changes may occur in the final released BAA

BENGAL Problem Statement

Challenge: Large language models (LLMs) present massive opportunities, but vulnerabilities and threats prevent safe adoption of the technology within the IC

Goal: Develop novel technologies to efficiently detect, characterize and mitigate LLM threat modes and vulnerabilities, allowing the IC to work resiliently with imperfect models



Bias Effects and Notable Generative AI Limitations

BENGAL Is A Super Seedling

- All IARPA efforts seek novel, high-risk/high-payoff solutions to the IC's greatest challenges
- For rapidly evolving technology, like LLMs, targeted super seedlings enable research progress at IARPA in a shorter time frame
- What is a super seedling?
  - 2 years long, two 12-month phases
  - Focuses on topics within an area of high interest to the IC, rather than addressing single, critical challenge
  - No shared tasks: Each performer proposes their own project, metrics and metric targets
  - Total funding limit is $4M per performer team
  - Goal is to take an idea from disbelief to doubt, with the potential for a follow-on effort (e.g., a full research program) or transition to the IC

Performer Tasks

- Focus on one or more of the BENGAL topic areas of interest:
  - Biases and induction of diverse analytical perspectives
  - AI Hallucinations and inferences
  - Safe information flow in sensitive environments
  - Working resiliently with imperfect or poisoned sources
- Articulate a taxonomy of LLM threats/vulnerabilities within the topic area(s)
- Develop novel technologies to detect, characterize and mitigate the threats/vulnerabilities

Topic Areas

Purpose of Topics

- Topics are framed broadly with the intention of soliciting diverse and innovative proposals
- Subtopics are more specific research directions to help further elucidate the intent of the topic; subtopics are suggestions, not an exhaustive list

BENGAL: Topics

Topic #1: Biases and induction of diverse analytical perspectives

Topic #2: AI Hallucinations and inferences

Topic #3: Safe information flow in sensitive environments

Topic #4: Working resiliently with imperfect or poisoned sources

Topic #1

Biases and Induction of Diverse Analytical Perspectives

Problem:

- LLMs can help the IC analyst understand complex events and states of affairs
- Biased models offer limited perspectives on a complex event, but the IC analyst greatly benefits from exposure to contrasting perspectives on the same event or situation

Topic #1

Biases and Induction of Diverse Analytical Perspectives

What do we mean by 'bias'?

- Can refer to one or more of broad range of biases that might suppress or promote relevant content with the effect of misleading a user
  - Cognitive, demographic, ideological, cultural, temporal, etc.
- The program is not focused on bias in the statistical sense i.e., model overfitting and underfitting

Topic #1

Biases and Induction of Diverse Analytical Perspectives

Subtopics:

- Methods for objectively quantifying bias (e.g., relative to a specific collection of texts or other content)
- Computational techniques to characterize perspective spaces and measure the differences between perspectives.
- Using human-LLM 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

Problem:

- Generative LLMs produce spurious, ungrounded outputs ('hallucinations' or 'confabulations') that can cause erroneous analysis and decision making.
- 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
  - Good inferences are often not (straightforwardly) based on ground truth evidence!

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

Problem:

- The IC limits access to sensitive information, which if disclosed to unauthorized individuals poses a grave threat to national security
- However, restricting information to certain people/systems can pose a threat to national security by preventing critical collaborations and timely sharing of critical information
- LLMs do not allow failproof sharing of information across different levels of sensitivity
- Information may not be sensitive on its own, but when aggregated can become sensitive
- **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**

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 of 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.

Safe Information Flow in Sensitive Environments

Note on Scope:

- Performers will not be permitted access or test systems in classified environments; thus, offerors must propose unclassified testing and evaluation schemes that credibly simulate environments in which sensitive information must be protected

Topic #4

Working resiliently with imperfect or poisoned sources

Problem:

- The IC relies on open sources (e.g., new organizations, individual content producers from around the world) that provide imperfect or intentionally misleading information
- Unreliable or malicious sources may poison LLM models
- However, reliance on imperfect sources is necessary in certain situations (e.g., when only sparse data is available)

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

Out of Scope

- 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 information
- Resubmissions of work already awarded by the National Science Foundation, National Institutes of Health, Department of Defense, Intelligence Community, or other federal agencies

Pitch Us Your Project Idea!

- IARPA is agnostic to research approach
- Propose what is needed to meet objectives
  - Research approach
  - Staff
  - Resources
  - Teaming plans
- Highlight innovative, novel, and scientifically supported research and development approaches

Characteristics of A Successful Project

- Novelty and potential for high impact
- Well thought-out teaming; Cross-disciplinary collaboration is a plus
- Approaches must generalize across LLM text generation models and their different versions
- Objective replicable evaluation procedures, quantitative and qualitative metrics, and metric targets
- Clear plans to carry out these evaluations, which will be monitored by an independent T&E team
- Viable plan to deliver turn-key containerized software (Phase A) with UI components and thorough documentation (Phases A and B)
  - T&E will verify software and results

Expectations for Responsible Research

- Performers must obtain institutional review board (IRB) approval or an IRB waiver for all R&D and data collection activities
- Performers must take steps to ensure removal of personally identifiable information (PII) from all development datasets

Program Test and Evaluation and Metrics

Tentative Program Timeline

- BAA formal release: November 2023
- Kick-off: Summer 2024

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												

What is Test and Evaluation?

- BENGAL performers are expected to evaluate their own results
- A testing and evaluation (T&E) team will also verify performer results.
- T&E is carried out by 'trusted partners' of the Government:
  - Not yet selected, but may include Government, Federally Funded Research and Development Corporations (FFRDCs), University Affiliated Research Centers (UARCs), or National Labs
- T&E teams will validate performer software

How Is Progress Measured?

- **To be successful, white papers and proposals must clearly articulate tasks, metrics and metric targets**
- A good metric is…
  - Easy to interpret, not complex or subject to multiple interpretations
  - Easy to implement, replicable
  - Allows measures of confidence/significance
  - Comparable over time
  - Actionable (i.e., informs changes to the approach)
- A good target enables comparison with state-of-the-art or a well-justified baseline
- **Projects will be rejected if they do not provide a feasible plan for T&E partners to replicate/verify results**

Evaluation Milestones

- As noted above, successful proposals will clearly articulate a testing and evaluation protocol that can be…
  - Run on the performer site and validated by the T&E team
  - Replicated by T&E using T&E's own infrastructure
- Testing and evaluation of performer systems will occur 3 times during each of the 12-month phases of BENGAL
- Proposals must state target values for each milestone and justify why these values are challenging given the current state of the art
- Results reported at testing and evaluation milestones will inform the Government in deciding which teams will advance to Phase B

Datasets

- Performers must obtain or develop their own datasets and any datasets to be used by T&E teams for validation or replication of their approach
- Performers must ensure that all datasets used for the development and testing of their systems be legally shareable with the Government and T&E team
- Performers must obtain institutional review board (IRB) approval or an IRB waiver for all R&D and data collection activities
- Performers must take steps to ensure removal of personally identifiable information (PII) from all datasets

How to Propose to BENGAL?

- IARPA anticipates releasing the BENGAL Broad Agency Announcement (BAA) in November 2023 on Sam.gov
- BENGAL will have two submission phases:
  - Offerors submit a short **white paper** describing their proposed project using a version of the Heilmeier Catechism; IARPA PM will review white papers and (not) recommend that the offeror submit a full proposal
  - After the white paper phase, offerors will have ~4 weeks to submit a **full proposal**
- For a limited time following the release of the BAA, IARPA will answer questions about the solicitation via the public question and answer (Q&A) process, which will be described in the BAA

Point of Contact Information

Dr. Timothy McKinnon

Program Manager

Office of the Director of National Intelligence

Intelligence Advanced Research Projects Activity (IARPA)

Washington, DC 20511

Email: dni-bengal-proposers-day@iarpa.gov

Website: https://www.iarpa.gov/index.php/research-programs/bengal