DARPA-EA-24-01-05

Safe and Assured Foundation Robots for Open Environments (SAFRON)

I. ARC Opportunity

The Defense Advanced Research Projects Agency (DARPA) Defense Sciences Office (DSO) is issuing an Advanced Research Concepts (ARC) Opportunity, inviting submissions of Abstracts for innovative exploratory research concepts in the technical domain of artificial intelligence. This ARC Opportunity, Safe and Assured Foundation Robots for Open Environments (SAFRON) is issued under the master ARC Exploration Announcement (EA), DARPA-EA-24-01.

ARC Opportunities are designed to allow an individual researcher the opportunity and time to focus on nascent, paradigm-shifting ideas for national security applications. While multiple researchers from the same organization may be proposed, the aggregate level of effort for a proposed research concept are expected to be equivalent to one full-time equivalent (FTE) and 12 months as ARC topics are designed for ideas that nominally would take a full year effort (1 FTE over 1 year) to properly validate. DARPA expects that the individual(s) working on the proposed idea primarily focus on the effort for the entire period of performance to the maximum extent practical. Only minimal variation to this requirement will be accepted. The maximum period of performance is 12 months. Each ARC award's total cost should range from $100,000 to $300,000, including direct and indirect costs and graduate student tuition, if applicable. Proposed costs are limited to $10,000 or less for materials, equipment, and Other Direct Costs (ODC). Under no circumstances will profit be authorized. While resource sharing is not expected, it may be offered in the proposal. DARPA understands not all ideas and organizations may fit in this parameter range and will work with a proposer to ensure truly innovative ideas can be explored with the required resources. Travel and publication costs may not be proposed. No subawardees are permitted.

To view the latest amendment of the DARPA Exploration Announcement, visit SAM.gov under solicitation number DARPA-EA-24-01: https://sam.gov/opp/179ef7e5199e4e6daea1615631f4a81f/view. It is incumbent upon the proposer to review DARPA-EA-24-01, any resulting amendments to DARPA-EA-24-01, and Frequently Asked Questions (FAQs) before preparing and submitting an Abstract and/or an Oral Proposal Package (OPP) (if invited). All Abstract submissions to this announcement must adhere to the instructions contained in DARPA EA-24-01.

All technical, contractual, and administrative questions regarding this notice must be emailed to SAFRON@darpa.mil. This ARC Opportunity is soliciting Abstracts only. DARPA will evaluate Abstracts submitted in response to this ARC Opportunity, as detailed in Section 4 of the latest amendment issued against DARPA-EA-24-01. If the Government selects an Abstract for an Oral Presentation, the Government will issue an invitation to submit an OPP. The invitation will include the submission instructions and deadline.

All awards made as a result of the ARC Opportunity will be Research Other Transactions (OTs) awarded under the authority of 10 U.S.C. § 4021.

Abstracts submitted to this ARC Opportunity will be evaluated on a rolling basis in accordance with the latest amendment issued against DARPA-EA-24-01. The end of the submission period is January 13, 2025 at 4:00 p.m. Eastern Time. No Abstracts will be accepted after the end of the submission period. Proposers are encouraged to submit Abstracts as early as possible. Funding for this ARC Opportunity is limited. Should funding be exhausted, the Government may elect to shorten the overall submission period with an amendment to this ARC Opportunity.

II. ARC Opportunity Description

Foundation models (FMs) have transformed AI capabilities in many domains[1] by virtue of their large architectures, internet-scale datasets, and unique customization techniques ("fine tuning"). This trend has recently brought transformational capabilities to robots.[2] Particularly, FMs enable robots that can parse natural-language directions for complex tasks and then contextualize and execute those tasks in unconstrained, open-world environments – including even "zero-shot" scenarios. This is a dramatic break from existing autonomous systems; current systems are designed for tailored applications and narrow, precise operating conditions.

However, natural-language direction for open-world autonomy presents a critical challenge from a safety and assurance perspective, since current methods to assure learning-enabled systems are inadequate to address FMs operating in this paradigm. Specifically, current formal neural network verifiers have been effective in limited, narrow scenarios, but do not scale to large, state-of-the-art FMs; existing training and alignment methodologies are not very robust[2,3]; no methods consider complex behaviors and scenarios an FM-enabled robot may be expected to encounter in an unconstrained, open-world environment. Moreover, FMs are known to exhibit unique (semantically) errant behavior such as hallucination, false confidence in reasoning, and manipulation via "jailbreaking," among others, further complicating the assurance challenge. Assurances are crucial to deploy FM-enabled robots so they do not manifest these behaviors. For example, in an unconstrained environment, a robot controlled by a hallucinating FM could fail to execute a critical task.

This ARC opportunity is soliciting ideas to explore the following question: How, and to what extent, can we assure FM-enabled robots will behave only as directed and intended?

A. ARC Opportunity Technical Objective

This ARC opportunity seeks investigations into innovative methods and approaches that will lead to assurances about the behavior of robots that use FMs, with particular emphasis on robots (of any type) that receive commands in natural language; operate in unstructured, open-world environments; and/or incorporate the FM in closed-loop decision making. Methods and approaches may consider all phases and components of the design and training of the FM and the FM-enabled robot as a system. Approaches of particular interest yield assurances (which may be probabilistic) with clear assumptions and are likely to extend as far as possible to complex robot instructions and behaviors in open environments. Methods and approaches that provide assurances subject to minimal additional supervision from a human operator in real time are also of particular interest. Negative results are also in scope (i.e., "For any FM-enabled robot with property X, there always exists an instruction of fewer than Y tokens that elicits behavior Z."), but negative results

[1] Bommasani, Rishi et. al. "On the Opportunities and Risks of Foundation Models." ArXiv preprint, arXiv:2108.07258v3, https://crfm.stanford.edu/assets/report.pdf (2022).

[2] Hu, Yafaei et. al. "Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis." ArXiv preprint, arXiv:2312.08782v2 (2023).

[3] Anwar, Usman et. al. "Foundational Challenges in Assuring Alignment and Safety of Large Language Models." ArXiv preprint, arXiv:2404.09932v1 (2024).

should be justified as applicable to current or future FM-enabled robots with state-of-the-art performance and capabilities.

B. ARC Abstracts

SAFRON ARC abstract submitters should clearly articulate why the proposed approach to assurance (or the limits thereof) is a novel idea or concept, and should clearly explain how their approach is relevant to current and future state-of-the-art FMs employed in open-world robotics applications. Abstracts should include significant detail about the assurances and methodologies proposed, especially regarding any necessary assumptions about the relevant FMs, robots or their environments; the characteristics of the behaviors to be assured (or not); and the relationship between assurable behaviors and human-operator instructions. Abstracts proposing solely rigorous theoretical work are in scope, but proof-of-concept simulations or experiments are encouraged. Any proposed simulations or experiments should reflect aspects of open-world environments (e.g., zero-shot scenarios, context-dependent natural language directions, complex environments, etc.), and be described in significant detail, including training data (if applicable), FMs, robots, simulation/experimental environment setups, robot tasks, robot/operator interaction, and validation/control methodologies. Abstract submitters should include a thorough literature review.

This ARC Opportunity is intended to be as inclusive as possible; however, proposed ideas should address the appropriate scope, have a clear deliverable at the end of the effort, and include specific practical applications of the research.

Abstracts should describe a research plan including (1) detailed intermediate technical objectives with evaluation measures and (2) a schedule segmented monthly or quarterly outlining corresponding deliverables.

DARPA will evaluate Abstracts submitted in response to this ARC Opportunity, as detailed in Section 4 of the latest amendment issued against DARPA-EA-24-01. If the Government selects an Abstract for an Oral Presentation, the Government will issue an invitation to submit an OPP. The invitation will include the submission instructions and deadline.

C. Schedule of Milestones

The specific milestones and due dates listed below are common to all Abstracts and OPPs (see above for technical details and Section III.A. below for additional information on milestones). Abstracts selected to submit an OPP will be required to propose milestones associated with the program plan as part of the oral proposal.

- Kick-off meeting: Briefing should define the technical approach and steps forward to include detailed technical objectives, milestone details and schedule in Gantt form to show initial plan.
- Milestone status meetings: Briefing to include detailed progress towards all research objectives, progress to plan, and discussion of next milestone’s objectives.
- Final Opportunity outbrief: The final briefing should summarize all work completed on the project.

D. Reporting Requirements

Performers will be expected to provide at a minimum the following reports:

- Monthly technical updates and financial reports. These reports should include progress to plan.
- Milestone technical report. Each report should detail progress towards all research objectives and should include a master document that refers to associated explanatory presentation slides, design documents, algorithms, models, modeling data and results, and model validation data, publications, and software source code with full documentation, as needed.
- Final technical report. The final report should include the final master document from the quarterly technical reports and detail results of all milestones associated with the program plan for the entire period of performance. This must include work that was successful towards reaching milestones as well as work that was not successful.

III. ARC Opportunity Submission Format, Instructions and Selection

A. Abstract Content and Format

All Abstracts submitted in response to this notice must comply with the content and format instructions in Section 3.1 of the latest amendment issued against DARPA-EA-24-01. The submission must use the template provided as attachment to DARPA-EA-24-01. Abstracts submitted in response to this ARC Opportunity must be unclassified.

B. Abstract and OPP Submission Instructions

Abstracts submitted in response to this ARC Opportunity and OPPs submitted in response to an invitation shall be submitted electronically via the DARPA Submission website at https://baa.darpa.mil. See Section 3.3 of the latest amendment issued against DARPA-EA-24-01 for Abstract and OPP submission instructions.

Technical support for the DARPA Submission website is available during regular business hours, Monday – Friday, 9:00 a.m. – 5:00 p.m. Eastern Time. Requests for technical support must be emailed to BAAT_Support@darpa.mil with a copy to SAFRON@darpa.mil. Questions regarding submission contents, format, deadlines, etc. should be emailed to SAFRON@darpa.mil. Questions/requests for support sent to any other email address may result in delayed/no response.

DARPA will acknowledge receipt of complete submissions via email and assign identifying numbers that should be used in all further correspondence regarding those submissions. If no confirmation is received within two (2) business days, please contact SAFRON@darpa.mil to verify receipt.

No Abstracts will be accepted after the end of the overall submission period listed in Section I above. Abstracts must be submitted per the instructions outlined in this ARC Opportunity *and received by DARPA* no later than this time and date. Proposers are advised that the Abstract submission deadline outlined herein is in Eastern Time.

Abstracts will be evaluated and selected in accordance with Section 4 of the latest amendment issued against DARPA-EA-24-01.

IV. Award Information

Selected OPPs will result in a potential award of a Research OT agreement subject to the proposer's acceptance of the terms and conditions. Proposers must review the model Research OT agreement provided as Attachment E to DARPA-EA-24-01.

The completed Task Description Document, Schedule of Milestones and Payments (templates included in Attachment E), and data rights will be included in the Research OT agreement upon award.

Given the limited funding available for each ARC Opportunity, not all proposals considered selectable may be selected for a potential award.

V. Eligibility

See Section 6 of the latest amendment issued against DARPA-EA-24-01 for information on who may be eligible to respond to this notice.

VI. Human Subject Research

Abstracts to this ARC Opportunity proposing human subjects research will be considered out of scope and may be disregarded.

VII. Administrative Requirements

Section 7.2 of the latest amendment issued against DARPA-EA-24-01 provides information on administrative requirements that may be applicable for proposal submission as well as performance under an award.

VIII. Frequently Asked Questions (FAQs)

All technical, contractual, and administrative questions regarding this notice must be emailed to SAFRON@darpa.mil. Emails sent directly to the Program Manager or any other address may result in delayed or no response.

All questions must be in English and must include the name, email address, and telephone number of a point of contact. DARPA will attempt to answer questions publicly in a timely manner; however, questions submitted within seven (7) calendar days of the proposal due date listed herein may not be answered.

DARPA may post an FAQ list under the ARC Opportunity on the DARPA/DSO Opportunities page at (http://www.darpa.mil/work-with-us/opportunities). The list will be updated on an ongoing basis until one (1) week prior to the abstract due date. DARPA will also maintain https://www.darpa.mil/ARC as a resource page with links to all relevant ARC Opportunities and FAQs.