MODULE ANNOUNCEMENT

FOR

ADVANCED RESEARCH PROJECTS AGENCY FOR HEALTH CHATBOT ACCURACY AND RELIABILITY EVALUATION (CARE) EXPLORATION TOPIC

ARPA-H-MAI-24-01-04

APRIL 19, 2024

TABLE OF CONTENTS

1. MODULE ANNOUNCEMENT OVERVIEW INFORMATION

FEDERAL AGENCY NAME: Advanced Research Projects Agency for Health (ARPA-H)

MODULE ANNOUNCEMENT TITLE: Chatbot Accuracy and Reliability Evaluation (CARE)

ANNOUNCEMENT TYPE: Initial Announcement

MODULE ANNOUNCEMENT NUMBER: ARPA-H-MAI-24-01-04

DATES:

- ***Module Announcement release date***: April 19, 2024
- ***Questions & Answers (Q&A) due date***: May 8, 2024
- ***Questions and Answers (Q&A) release date:*** May 15, 2024
- ***Proposal due date***: June 3, 2024.

2. RSO EXPLORATION TOPICS

A. INTRODUCTION

ARPA-H is launching an Exploration Topic (ET) aimed at expanding the Resilient System Office's (RSO) funding approach associated with the interest areas included within Appendix A to the Master Announcement Instructions, ARPA-H-MAI-24-01. Exploration Topics will be announced via Module Announcements issued under the Master Announcement Instructions (MAI), ARPA-H-MAI-24-01. Exploration Topics are short-duration, fast-paced efforts with smaller, targeted awards. Each Exploration Topic will pursue topics that strategically align with the RSO mission and provide foundational proofs-of-concept for additional future research to be built upon.

B. EXPLORATION TOPIC STRUCTURE, AWARD VALUE, AND PROPOSAL INFORMATION

Exploration Topic, or ETs, will describe short-duration, fast paced efforts that are no more than 24 months in duration. ETs may consist of a single base period or may consist of multiple stages. Stage structure will be defined in each ET module announcement. Regardless of structure, the total duration of each topic is not anticipated to exceed 24 months. Specific technical objectives to be achieved, task descriptions, intellectual property rights, milestone payment schedule, and deliverables will be included in each ET module announcement. The total value of each award will be stipulated in each ET module announcement; however, each award will not exceed $8M.

Proposals identified for negotiation will result in negotiating an award of an Other Transaction (OT) Agreement. Use of an OT Agreement provides significant opportunities for flexible execution and arrangements given the nature of the work to be conducted under these ETs and assists in meeting RSO's aggressive research goals. Moreover, all resulting ET module announcements will result in OT Agreements with fixed payable milestones. Fixed payable milestones are fixed payments based on the successful completion of the milestone accomplishments agreed to in the milestone plan. Specific milestones will be based upon the ET objectives stipulated in each ET module announcement (see Section D, "Exploration Topic Structure, Schedule and Milestones" of the ET module announcement).

Additionally, ETs allow for a streamlined solicitation and acquisition approach. ARPA-H is looking to finalize a new award within 60 days of selection notification letters being sent out. Accordingly, proposers must review the model OT Agreement provided in Attachment 1 of each ET module announcement prior to submitting a proposal. ARPA-H has provided the model OT to expedite the negotiation and award

process and to ensure ARPA-H achieves the goal of finalizing awards within 60 days of selection notification letters being sent. The model OT is representative of the terms and conditions that ARPA-H intends to include in ET module announcement awards. All Stage 1 submissions under the ET (see Section 5A below) must include the model OT Agreement, if the proposer IS suggesting minimal edits.[1] The submission must include proposed edits utilizing revision markings and must include a comment explaining the concern the proposed change addresses. However, ARPA-H may not accept suggested edits. A proposer does not have to provide a model OT agreement in the Stage 1 submission if edits are NOT being proposed. If an edited version of the model OT is not provided as part of the proposal package, ARPA-H assumes that the proposer has reviewed and accepted the award terms and conditions, indicating agreement (in principle) with the listed terms and conditions applicable to the specific award instrument. The proposer should, in this instance, ensure the Administrative & National Policy Requirements document clearly denotes agreement with the listed terms and conditions. The Government also reserves the right to remove a proposal from award consideration should the parties fail to reach an agreement on OT award terms and conditions within the award timeline stipulated above.

3. OPPORTUNITY DESCRIPTION

The mission of the Advanced Research Projects Agency for-Health (ARPA-H) is to accelerate better health outcomes for everyone by advancing innovative research that addresses society's most challenging health problems. Awardees will develop groundbreaking new ways to tackle health-related challenges through high potential, high-impact biomedical and health research. ARPA-H seeks proposals to develop new evaluation technologies for detecting hallucinations and other inaccuracies in medical Large Language Model (LLM) output for patient-facing applications. The CARE ET endeavors to produce tools and technology that evaluates output with the efficiency of computational methods and the accuracy of human experts. By creating scalable and cost-effective LLM evaluation technologies, CARE seeks to accelerate the creation of trustworthy medical chatbots and enable access to more reliable health-related information.

A. EXPLORATION TOPIC INTRODUCTION

The implementation of chatbots and their associated Large Language Models (LLMs) have been tested in various domains, including medicine and healthcare, where researchers have applied them to clinical workflow translation, triage, health data management, and medical education.[2] Despite their promising applications, the widespread and immediate deployment of LLMs across health-related fields is hindered by serious safety concerns. LLMs suffer from reliability and accuracy issues including hallucinations where generated content is incorrect or unsubstantiated. This issue presents obvious barriers to widely deploying chatbots designed to offer timely health information or answer specific medical questions from the public. Creating medical chatbots worthy of patients' and clinicians' trust requires stringent evaluation technologies and frameworks capable of measuring and assessing performance for quality control. Beyond addressing hallucinations, evaluation of medical chatbot quality must also consider the desires and concerns of the end user, such as fears of bias or overly confident outputs. Several automated assessments of chatbot output have been explored, including computational benchmarks and multiple-choice tests. These approaches can be fast and cost-effective, but do not directly assess hallucinations, diagnostic reasoning, or bias. Currently, the most reliable method for identifying LLM hallucinations and related issues is expert human review, which, however, is time-consuming, limited in scope, and expensive. To realize the transformative potential of medical chatbots, significant improvements in the ability to detect hallucinations are required, including, but not limited to dramatic improvements in the scalability of expert human

[1] This deviates from the proposal preparation instructions included within Section 3.1 of the MAI.

[2] Yu, P., Xu, H., Hu, X., & Deng, C. (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Basel, Switzerland)*, *11*(20), 2776. https://doi.org/10.3390/healthcare11202776

evaluation of chatbot output.

CARE aims to develop novel technical approaches for large-scale, human-expert-level evaluation of medical chatbot output for patient-facing applications. By developing and testing proof-of-concept evaluation technologies for LLMs across a variety of use cases, this ET will provide a critical resource for chatbot developers and regulators. CARE endeavors to produce tools and technology for evaluation of medical chatbots with the efficiency of computational methods and the accuracy of human experts. This technology will not only assess safety by detecting hallucinations but will improve the utility and reduce the bias of medical chatbots by considering stakeholder desires and concerns.

National Health Impact

More than half of American households use the internet for health-related activities, including researching health information, yet online experiences differ by sociodemographic group.[3] People from socially disadvantaged groups, including racial/ethnic minority groups, older adults, those with lower incomes, and persons living in rural areas, report greater difficulty in seeking health information online.[4] The US Department of Health and Human Services (DHHS) has made patient-centered and electronic communication a key focus area, yet limited progress has been made over the last decade to removing barriers to online health information. In the last year, generative Artificial Intelligence (AI) and LLMs have seen unprecedented uptake by American citizens. Although reliable data is still unavailable, one survey reported more than 80% of respondents had used a chatbot in 2023, and another suggests, that despite prohibitions on medical use by vendors, millions of medical queries are submitted on a weekly basis by users of OpenAI alone.[5,6] This level of engagement presents a significant health problem since 33% - 50% of outputs from a state-of-the-art chatbot contained at least one serious hallucination.[7] If medical chatbots could reduce or eliminate the production of inaccurate, misleading, and unpredictable information, the U.S. could realize enormous benefits in improved access to health-related information, especially for underserved populations. This ET aims to develop scalable evaluation tools that match the gold standard of human expertise, both accelerating the creation of trustworthy medical chatbots and providing accurate, valid assessments of their performance for patients, clinicians, and regulators.

B. EXPLORATION TOPIC STRUCTURE AND INTEGRATION

CARE is a 24-month effort that aims to develop and demonstrate scalable chatbot evaluation technologies that reliably operate at the level of expert human review. The ET is conducted over three Stages[8] and will pursue innovation within a singular technical area (TA), delineated into two parallel and interconnected subsections as illustrated in Figure 1 (contained within Section D below).

- **TA1.1: Pragmatic criteria and prompt generation.** Improve prompt generation technology to effectively examine the full range of criteria that informs the evaluation of trustworthiness of LLM outputs.

[3] Cao and Goldberg 2019. More than Half of American Households Used the Internet for Health-Related Activities in 2019, NTIA Data Show. National Telecommunications and Information Administration.
[4] Finney Rutten *et al* 2019. Online Health Information Seeking Among US Adults: Measuring Progress Toward a Health People 2020 Objective. Public Health Reports.
[5] https://www.userlike.com/en/blog/consumer-chatbot-perceptions downloaded 1/24/24.
[6] See https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference
[7] Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
[8] The referenced "Stages" do not denote a government decision point. Stages have varying metrics and deliverables and are being utilized to track progress towards the ultimate goals and objectives of the CARE ET.

- **TA1.2: Scalable, expert chatbot evaluation technology.** Develop novel chatbot evaluation technologies that perform at the speed of computational methods with the accuracy of expert human review.

Proposals shall address both subsections and all Stages within a single proposal submission. Proposals failing to address both subsections will be deemed non-conforming and may be rejected without further review.

TA1.1 – PRAGMATIC CRITERIA AND PROMPT GENERATION

The public use and development of chatbots for accessing medical and health-related information continues to grow, yet the trustworthiness, transparency, and explainability of these advanced LLM systems and their respective outputs remain difficult to evaluate due to issues in factuality, coherence, and hallucinations. This challenge is exacerbated by the diverse context and breadth of use cases in which individuals engage chatbots. Without the systematic identification and mitigation of potential harms through contextualized evaluation approaches, LLMs will propagate bias-related issues that contribute to health disparities by presenting medical misinformation or incorporating biases. To meet this need, performers will be expected to align medical chatbots with clinical and societal values through consultation and participatory research (e.g., workshops, bootcamps, surveys) with relevant stakeholders for predefined use cases, which will influence key parameters by which chatbots are evaluated.

Proposals must include a description of two patient-facing medical chatbot use cases. A use case establishes the domain of discourse for which the performer will develop technology to evaluate chatbot outputs. Uses cases must be defined by a coherent and significant medical need faced by a well-defined community of patients and/or caregivers. Proposals should select use cases that are likely to generalize to other kinds of patient-facing medical advice. Additional information on use case generalization and methods by which generalization will be evaluated are provided in Section D. Example use cases include, but are not limited to, advice to parents about pediatric concerns, maternal health advice, post-surgical oncology advice, or mental health crisis lines. Use cases that require unusual or highly specific expertise, outside the realm of standard patient care, or address only very small populations of patients are likely to be deemed out of scope unless strong justification is provided for why the proposer believes that use case will generalize to other use cases. Additionally, non-medical use cases (e.g., supplements or non-medically indicated diets) are out of scope.

CARE aims to develop evaluation technologies that will not only assess safety by detecting hallucinations but will improve the utility and reduce the bias of medical chatbots by considering stakeholder desires and concerns. To this end, proposals must include a description of a stakeholder engagement plan with three core components:

- ***Recruitment:*** a well-justified, diverse, and complete list of stakeholders relevant to the predefined use cases. Stakeholders may include, but are not limited to patients, caregivers, physicians, nurses, mental health professionals, public health workers, helpline staff, or social workers. Proposals should describe robust strategies for ensuring representativeness and diversity of stakeholders involved in TA1.1 activities.
- ***Elicitation***: specific research methods to systematically and completely elicit desires and concerns. Methods may include, but are not limited to focus groups, workshops, bootcamps, or surveys. Proposals may include qualitative and/or quantitative elicitation methods and can combine multiple methods, for example, in-person focus groups followed by broader survey instruments to test representativeness.

- ***Guideline distillation***: strategies for distilling stakeholder desires and concerns into concrete and specific guidelines identifying the dimensions of chatbot evaluation in a given use case. Proposals should describe one or more methodological approach(es) to creating guidelines from stakeholder data, which may include but is not limited to thematic analysis.[9]

Evaluation of the performance of a medical chatbot depends crucially on the prompts used to test it. Proposals must specify how they will use the results of the stakeholder elicitation work to create prompt generation technology for chatbot evaluation. Prompts must be relevant, complete, diverse, and different than the examples used to train the LLM. Description of prompt generation technologies must address the following four criteria:

- ***Scaling***: How quickly can the approach generate prompts and are there limits to the total number of distinct prompts generated?
- ***Novelty***: A key concern in evaluating chatbots is the potential for bias from memorization (presence of text similar to the prompts in the training data). How will the proposed technology generate prompts unlikely to have been used in training chatbots, even those with very large and/or unknown training data? Novelty can be measured by perplexity of the prompts given a language model (where higher perplexity is better).
- ***Diversity***: Prompts should not be repetitive or highly stereotyped. How will the technology produce a diverse set of prompts? Diversity can be measured by average perplexity of prompts given a language model combined with the other generated prompts.
- ***Coverage***: Prompts generated should elicit responses from medical chatbots that test a large proportion of stakeholder desires and concerns. Proposals should specify quantitative methods for assessing this coverage.

Technologies for prompt generation may include but are not limited to corpus-based approaches, template-based approaches, and/or the use of language models to produce relevant prompts.

TA1.2 – SCALABLE, EXPERT CHATBOT EVALUATION TECHNOLOGY

One of the key factors inhibiting medical chatbot adoption is the prevalence of hallucinations, which are widely acknowledged but currently difficult or impossible to detect automatically. TA1.2 will address this bottleneck by developing evaluation technologies that can detect factual hallucinations with a high degree of accuracy, thus enabling the use of chatbots for patient-facing applications. Proposals must describe a scalable evaluation technology that can detect factual and reasoning hallucinations (errors) in patient-facing chatbot output. Proposals must describe how the approach will combine the quality of expert human assessment with the scale of automated benchmarks to create evaluation tools that optimize speed, volume, and cost-effectiveness. Proposals may suggest either fully automated approaches or approaches that efficiently use modest amounts of human expertise to evaluate chatbot output at large scale. Potential approaches may include, but are not limited to active learning, gamification, or adversarial networks.

Proposals must describe how the evaluation technology will address the following factors, which are expected to improve during the course of the ET (Table 1):

- ***Hallucination detection****:* Identify a large and increasing proportion of the overall errors, including in fact or reasoning, in chatbot output.
- ***Scaling***: Score a large and increasing number of chatbot outputs in a fixed period of time.

[9] Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, *13*(1), 3-21.

- ***Human time required***: Score chatbot outputs using a modest and decreasing amount of human expert time.
- ***Model certainty***: Score how the degree of certainty expressed in the chatbot output is associated with actual correctness (e.g., expected or adaptive calibration error).
- ***Pragmatics coverage***: Detect a large and increasing proportion of phenomena (both positive and negative) that are identified as relevant by stakeholders.

Proposals must describe how the evaluation technology will perform the following functions:

- ***Text identification***: Isolate a specific text span in the output that contains the error when identifying hallucinations.
- ***Generalization***: The evaluation technology is able to be applied to other patient-facing medical chatbot use cases with modest or no adaptation. Here, generalization refers to the ability to perform the required evaluation functions on medical use cases other than the proposed one. See Section D for details on how generalization will be evaluated.
- ***Calibration***: Assess the calibration of chatbots, meaning the degree to which the certainty conveyed by output text corresponds to the probability the output is correct.

The technology developed in this ET is intended to be used in the evaluation of any patient-facing medical chatbot, therefore the evaluation of chatbot output not intended for patient-facing medical use is out of scope. Proposals may optionally specify and justify the selection of particular LLMs that will be the focus of the proposed effort. Evaluation methods should not depend on particular LLM features (e.g. accessibility of weights) and should be applicable to any chatbot. Raw LLMs or ones modified with guardrails, reinforcement learning from human feedback or other chatbots are acceptable, but the choice must be clearly justified. The focus models must be at or near the current state of the art at the time of submission.

C. EXPLORATION TOPIC METRICS

The overall goal for CARE is to create tools and technology for evaluation of medical chatbots with the efficiency of computational methods and the accuracy of human experts. In pursuit of this goal, prompts used for chatbot evaluation will be assessed based on multiple criteria, including scalability, novelty, diversity, and coverage. Prompts must accurately reflect elicited input from a diverse set of relevant stakeholders. By the end of the ET, performers should achieve 90% hallucination detection with increased speed and decreased human involvement. For all metrics, the 95% confidence interval needs to be reasonably narrow to provide confidence that the true value is not substantially different from the value estimated by the performers. Evaluation technologies should also be generalizable. By the end of stage III, evaluation methods tested by the IV&V team should show no more than a 5% degradation in their detection rate with new use cases, relative to the original use case the method was trained on. Metrics for each stage of each TA are outlined in Table 1 below.

Table 1. CARE ET metrics across TA1.1 and TA1.2

Metric	Description	Stage I	Stage II	Stage III
Hallucination detection	What percentage of hallucinations did the method detect relative to a human expert?	20%	50%	90%
Omission detection	What percentage of LLM outputs with missing or omitted information is detected with the developed method relative to what is detected by human expert?	20%	50%	90%
Human time required	Amount (in hours?) of expert human time required (per 1000 evaluations?)	Establish baseline	2X reduction	5X reduction

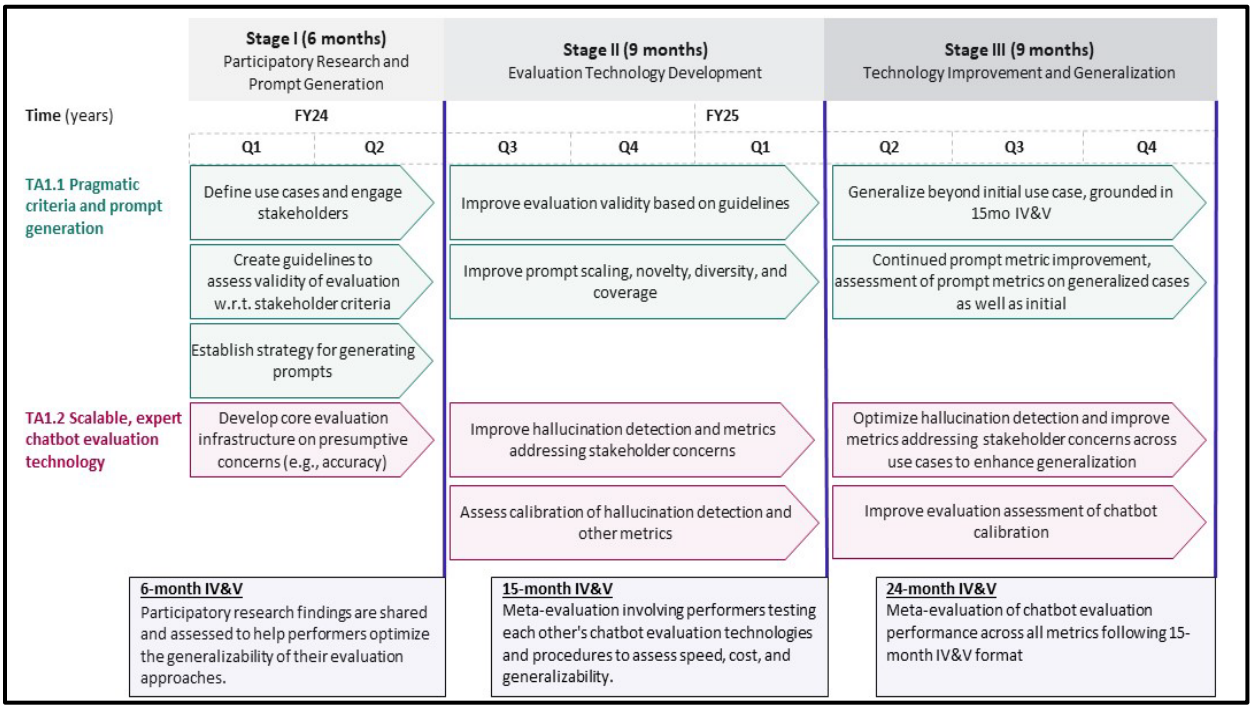
Evaluation scaling	On average, how long does it take to evaluate an LLM output?	Establish baseline	2X reduction	10X reduction
Generalizability	Evaluation performance on new use cases relative to initial use case (% hallucination detection relative to detection under the original use-case the model was trained on)	N/A	70%	95%
Engagement diversity	Diversity of engaged stakeholders across use cases as determined by demographic dimensions	Set baseline	1.5X	2X
Prompt scaling	How long does it take to generate a prompt?	N/A	< 1 sec	< 100ms
Prompt novelty*	How unexpected is a set of generated prompts given a language model? Measure in bits/token perplexity of prompt set vs. LLM.	Establish model type and baselines	3X increase	5X increase
Prompt diversity*	How unexpected is a set of generated prompts (average bits/token perplexity of a prompt vs. LLM + prior prompts)	Establish model type and baselines	3X increase	5X increase
Prompt coverage	How well do the prompts cover the phenomena relevant to the use case? For example, disease mention frequency, symptom mention frequency, mention frequency of demographic factors, or reference to constraints faced by under-resourced populations.	10%	50%	90%
Evaluation of model certainty	Does the assessment generated by the algorithm strongly correlate with human assessment in terms of association of certainty with accuracy? In other words, if the LLM's response uses highly confident language, is the response also accurate? Does the LLM use less certain language when the response is inaccurate? (Comparison of human-generated calibration curve with the calibration curve obtained with scalable evaluation technology. Expected mean-squared difference between calibration curves)	<0.3	<0.2	<0.1
Pragmatics coverage	Are the issues raised by stakeholders in TA1.1 addressed by the TA1.2 method. Assessed by IV&V on a 1-10 scale	4	7	9

*For metrics that require open (white box) LLMs, the IV&V team will choose LLMs with representative characteristics

D. EXPLORATION TOPIC STRUCTURE, SCHEDULE, AND MILESTONES

CARE is a 24-month ET comprised of three Stages that will require performers to efficiently allocate resources in developing capabilities described in each TA. Proposals must address the requirements in each TA subsection and all three Stages of the ET. See Section 5A for the module category associated with the CARE ET. Figure 1 below illustrates the CARE ET Timeline, Key Milestones and Deliverables across TA1.1-1.2. Timeframes are relative to the start of the effort.

Figure 1: CARE ET Timeline and Key Milestones across TA 1.1 and 1.2



Proposers must address the CARE ET objectives, metrics (Table 1), and the following fixed payable milestone deliverables in their proposals (Table 2). The task structure should be consistent across the proposed schedule, Task Description Document (TDD), the Stage 1, Volume 1 Basis of Estimate, and if selected for a potential award, the Stage 2, Volume 2 Price/Cost proposal. Proposers must use the Task Description Document (TDD) template provided within Attachment 1 of the CARE ET module announcement. The TDD will be Attachment 1 of the resulting OT Agreement.

If selected for award negotiation, the fixed payable milestones provided will be directly incorporated into Attachment 3 of the OT Agreement ("Schedule of Milestones and Payments" of the model OT) with milestone amounts calculated based on the proposed accumulation of monthly amounts up to each milestone date.

Fixed milestones for this project must include at a minimum, the following:

Table 2. CARE ET fixed payable milestone deliverables across TA1.1 and TA1.2

Milestone #	Milestone	Exit criteria / Deliverable	Expected Due Date*
Stage I:			
1	Report on stakeholder engagement, specifying the number of recruited / engaged stakeholders of each type. Document how the proposed plans for a well-justified, diverse, and complete set of stakeholders relevant to their use case have been realized. (TA1.1)	Stakeholder Engagement Report #1	Month 3
2	Report on stakeholder engagement, describing the elicited desires and concerns of the stakeholders regarding medical chatbot performance, and distilling these desires and concerns into concrete and specific factors and/or situations that chatbot evaluation needs to assess in this use case. (TA1.1)	Stakeholder Engagement Report #2	Month 4

3	Report on the systematic and complete set of factors and/or situations that are important to stakeholders (i.e., stakeholder values, needs, perceptions, concerns, and desires) in evaluating chatbots, with example prompts and responses illustrating each factor. (TA1.1) Develop a rubric, in the form of written guidelines that an independent validator could use to assess whether a proposed evaluation technology is (or is not) addressing the concerns of stakeholders elicited. (TA1.2) Develop software tool(s) to create chatbot prompts that test the space of chatbot interactions stakeholders care about. (TA1.1)	Stakeholder Assessment Report Stakeholder Rubric Report Software, source code, documentation, and a set of examples for using the code in the form of a live software document (e.g. Jupyter notebook or R markdown document for all associated TA1.1 and TA1.2 Milestones).	Month 6
Stage II:			
4	Report on how proposed technology and process development will combine the quality of expert human assessment with the scale of automated benchmarks to create evaluation tools that optimize speed, volume, and cost-effectiveness, and address the stakeholder concerns elicited in Stage I. (TA1.2)	Analysis Report	Month 9
5	Improve upon prompt generation software developed in Stage 1 focusing on scaling, novelty, diversity, and coverage. (TA1.1) Enhance chatbot evaluation technology from Stage I that uses the prompt generation technology from TA1.1 and assesses all specified metrics. (TA1.2) Share evaluation technology with other performers and participate in a meta-evaluation, accompanied with an IV&V third-party assessment. (TA1.2)	Software, source code, documentation, and a set of examples for using the code in the form of a live software document (e.g. Jupyter notebook or R markdown document for all associated TA1.1 and TA1.2 Milestones).	Month 15
Stage III:			
6	Report on how the Stage II prompt generation and evaluation technology performed on other teams' use cases, analyzing any unexpected successes or failures, and specifying a plan to improve generality of prompt generation and evaluation methods to these other use cases. (TA1.1 & TA1.2)	Analysis Report on Stage II generation and evaluation technology	Month 17
7	Further improve prompt generation technology based on Stage II and IV&V assessments. (TA1.1) Further improve chatbot evaluation technology based on Stage II meta-evaluation and IV&V assessments. (TA1.1) Share improved evaluation technology with other performers and participate in a second meta-evaluation, accompanied with an IV&V third-party assessment. (TA1.2) Report on the final summarized technical findings from the Exploration Topic (spanning TA1.1 & TA1.2), including opportunities for continuing to advance the technology and its application and outstanding challenges.	Software, source code, documentation, and a set of examples for using the code in the form of a live software document (e.g. Jupyter notebook or R markdown document for all associated TA1.1 and TA1.2 Milestones). Final technical report	Month 24

**Months after award*

Independent Verification and Validation (IV&V):

Independent organizations will evaluate interim and final CARE ET deliverables. Performers are expected to collaborate with these IV&V partners throughout the ET's duration. Technical evaluations will take place at the end of each Stage to ensure validation and real-world performance of developed technologies. At the end of Stage II and III, performers will participate in a meta-evaluation to assess key metrics of their

evaluation technologies, including scaling, speed, and generalizability. In order to evaluate generalizability, performers will use prompts generated for a particular use case on a chatbot evaluation technology developed around a separate use case.

E. POLICY CONFORMANCE, AGILE DEVELOPMENT, OPEN STANDARDS, AND INTELLECTUAL PROPERTY

Proposers will be expected to adhere to all relevant Government laws and policies applicable to data and information systems and technologies including but not limited to the following:

- Common IT Security Configurations
- Federal information technology directives and policies
- Section 508 of the Rehabilitation Act of 1973 (29 USC 794d) as amended by P.L. 105-220 under Title IV (Rehabilitation Act Amendments of 1998)
- HHS OCIO Policy for Information Technology (IT) Enterprise Performance Life Cycle (EPLC)

In concert with ARPA-H and partners, proposers should address innovative solutions to design, architect, develop, and test the technologies described in the TA subsections.

All data collected during CARE may be shared for research purposes including subsequent research and development efforts by ARPA-H. Performers will not have ownership of data collected during this program and will be required to frequently upload data to designated data repositories.

The ARPA-H CARE ET will emphasize creating and leveraging open-source technology and methodologies. Intellectual Property rights asserted by proposers are strongly encouraged to be aligned with open-source regimes. It is desired that all non-commercial software (including source code), software documentation, and technical data generated by the CARE ET is provided as deliverables to the Government with open-source or unlimited rights, as lesser rights may negatively impact the potential for additional research and development. Open-source code is highly encouraged using permissive, business-friendly open-source licenses such as CC-BY, BSD, MIT, Apache 2.0 or similar.

F. PERFORMER COLLABORATION/ASSOCIATE CONTRACTOR AGREEMENT (ACA)

It is expected that performers will interact and work collaboratively with other performers, especially as required for pairwise evaluation of each team's technology (see section D).

To facilitate the open exchange of information described above, performers will have Associate Contractor Agreement (ACA) language included in their award. Each performer will work with other CARE performers to develop an ACA that specifies the types of information that will be freely shared across performer teams. It is intended that ACAs be established, after award, during Stage 1 between each CARE performer. ACAs should be established no later than month 4 of the CARE period of performance. The open exchange of scientific information will be critical in advancing the research required to achieve the CARE objectives. The ACA will establish a common understanding of expectations to guide the open exchange of ideas and establish a collaborative foundation for the CARE ET. Each performer will also work with other performers to converge on open standards and APIs to ensure interoperability across prototype capabilities. See Appendix B to MAI ARPA-H-MAI-24-01 for the representative ACA article. The same or similar article will be included in all awards resulting from the CARE ET module announcement.

G. ELECTRONIC INVOICING AND PAYMENTS

See Section 5.2.6 of the MAI.

3. AWARD INFORMATION

Multiple awards are anticipated under this announcement; however, the number of proposals selected for award will depend on the quality of the proposals received and the availability of funds.

See Section 1.4 of the MAI, ARPA-H-MAI-24-01 for additional information on award information.

4. ELIGIBILITY

See Section 2 of the MAI, ARPA-H-MAI-24-01 for eligibility requirements.

5. MODULE ANNOUNCEMENT RESPONSES

A. PROPOSAL CONTENT AND FORMAT

This Module Announcement is soliciting Stage 1 Volume 1 proposals in accordance with the staged submission and evaluation process referenced in Section 3.1 and 4.1 of ARPA-H-MAI-24-01. Reference to Stages in this section of the CARE ET module announcement is not to be confused with the programmatic Stages of the CARE ET described above.

Stage 1 Volume 1 proposals must contain the following document submissions:

- TECHNICAL & MANAGEMENT
- BASIS OF ESTIMATE (BOE)
- TASK DESCRIPTION DOCUMENT OR RESEARCH DESCRIPTION DOCUMENT
- ADMINISTRATIVE & NATIONAL POLICY REQUIREMENTS
- MODEL OT AGREEMENT (**ONLY IF PROPOSING EDITS**)

If a Stage 1 proposal is selected for potential award, a proposer will be notified by the Government and required to submit a Stage 2 price/cost proposal for further consideration (see Section 3.1 and 4.1 of ARPA-H-MAI-24-01).

All proposals submitted in response to this announcement must comply with the content and formatting requirements of the OT Bundle (Attachment 1). Proposers are strongly encouraged to use the templates provided in the OT Bundle associated with this announcement. Information not explicitly requested in the MAI, this announcement, or OT Bundle, may not be evaluated.

All submissions, including proposals, must be written in English with font type not smaller than 12-point font. Smaller font may be used for figures, tables, and charts. Content and formatting are disclosed in each Bundle of Attachments. Below is the page restriction for each Module category:

- **BYTE Module** is > $2,000,000 ≤ $4,999,999: Volume 1 shall be limited to **15** pages.
- **KILO Module** is > $5,000,000 ≤ $10,000,000: Volume 1 shall be limited to **20** pages.

Performers must address all CARE ET Stages and both TA1.1 and TA1.2 in their proposal and can apply for BYTE or KILO Module. While the Government anticipates the abovementioned Module categories, the below not to exceed estimates per Stage are provided to aid in preparation of proposal submissions:

- Stage I (TA 1.1 and 1.2) should not exceed $1,800,000
- Stage II (TA 1.1 and 1.2) should not exceed $3,100,000
- Stage III (TA 1.1 and 1.2) should not exceed $3,100,000

NOTE: Proposals should select a cost point that is commensurate with the scale and complexity of the proposed approach. Proposals that simply align a proposed budget to the Module Category ceiling value is strongly discouraged. Thus, if a proposal is selected for Stage 2 submissions and the basis of estimate was simply aligned to the Module Category ceiling value, the Government will require a full cost proposal (i.e., direct and indirect rates, labor hours, equipment, material, other direct costs, etc.) that must be substantiated by salary documentation, indirect rate agreements, material and equipment quotations and a justification for proposed labor categories and hours that correlates directly to the proposed Task Description Document. The submission of a full cost volume will impact Stage 2 price/cost proposal timelines and will likely be followed by extensive cost negotiations.

EQUITY REQUIREMENTS

ARPA-H is committed to equitable health care access irrespective of race, ethnicity, gender/gender identity, sexual orientation, disability, geography, employment, insurance, and socioeconomic status. To that end, proposals should describe a recruitment strategy within their stakeholder engagement plan (under TA1.1) that ensures stakeholder groups are representative of the diversity in the American population. Information elicited from stakeholders should lead to prompt generation and chatbot evaluation technologies that accurately capture the desires and concerns for a broad spectrum of chatbot users.

B. Proposal Submission Instructions

All proposal submissions must be written in English. All proposals submitted in response to this solicitation must comply with the content and formatting requirements in the OT Bundle (Attachment 1). Proposers are strongly encouraged to use the templates provided in the OT Bundle. Information not explicitly requested in the OT Bundle may not be evaluated.

Proposal submission should be submitted to https://solutions.arpa-h.gov/Submit-Proposal/.

Proposers should consider the submission time zone and that some parts of the submission process may take from one business day to one month to complete (e.g., registering for a SAM Unique Entity ID (UEI) number or Tax Identification Number (TIN); see Section 5.2.1 of the MAI for information on obtaining a UEI and TIN).

C. Proposal Due Date and Time

Proposals in response to this notice are due no later than 12:00 PM Eastern Daylight time (EDT) on June 3, 2024. Full proposal packages as described in Section 5.A must be submitted per the instructions outlined in this Module Announcement and received by ARPA-H no later than the above time and date. Proposals received after this time and date may not be reviewed.

Proposers are warned that the proposal deadline outlined in the CARE ET will be strictly enforced. When planning a response to this notice, proposers should consider that some parts of the submission process may take from one business day to one month to complete.

6. PROPOSAL EVALUATION AND SELECTION

Proposals will be selected and evaluated in accordance with Section 4 of the MAI, ARPA-H-MAI-24-01. The Government reserves the right to decide which performers, if any, are selected for the award.

7. ADMINISTRATIVE AND NATIONAL POLICY REQUIREMENTS

Section 5.2 of the MAI, ARPA-H-MAI-24-01 provides information on Administrative and National Policy Requirements that may be applicable for proposal submission as well as performance under an award.

8. POINT OF CONTACT INFORMATION

CARE ET Module Announcement questions should be directed to:
https://solutions.arpa-h.gov/Ask-A-Question
ATTN: ARPA-H-MAI-24-01-04

9. QUESTIONS & ANSWERS (Q&AS)

All questions regarding this notice must be submitted to the link noted in Section 8. Emails sent directly to the Program Manager, or any other address will be **discarded**.

All questions must be in English. ARPA-H will attempt to answer questions in a timely manner; however, questions submitted within 10 business days of the proposal due date listed herein may not be answered.

In concert with this Announcement, ARPA-H has posted Q&As for the CARE ET Module Announcement and Master Instructions Announcement at SAM.gov. ARPA-H encourages all proposers to review the Q&As provided before submitting additional questions to the respective link noted in Section 8. The Government may not answer repetitive questions already answered in the posted Q&As.