CARE Exploration Topic
ARPA-H-MAI-24-01-04
Question and Answers (Q&A)
Organized By Category
May 16, 2024

CARE – General

Q1: Is CARE a Program or an ISO? If it isn't a Program or an ISO, what are the unique requirements, scope, etc.?

> A1: CARE is an Exploration Topic (ET) announced via a Module Announcement under the Master Announcement Instruction (MAI), ARPA-H-MAI-24-01. Please read the CARE ET module announcement (ARPA-H-MAI-24-01-04) and the MAI, ARPA-H-MAI-24-01 for details regarding the scope of CARE, award information, module announcement responses, to include proposal content and format, etc. The CARE module announcement and MAI can be found at https://sam.gov/opp/8d3c07159e9f4d32ab7b179d365e6fab/view.

Q2: The module announcement (Section 9) states that several questions have already been answered and they are posted in SAM.gov. We haven't found any posting with previous Q&As, can you provide the link to them?

> A2: There was a typo in Section 9, Questions & Answers (Q&As) of ARPA-H-MAI-24-01-04. This Q&A posting (dated May 9, 2024) is the first Q&As for CARE. The first sentence of Section 9 of the CARE Module Announcement has been updated to read as follows: "In concert with this Announcement, ARPA-H will post Q&As for the CARE Module Announcement at SAM.gov." See Amendment 01 to ARPA-H-MAI-24-01-04 posted to SAM.gov.

Q3: Would ARPA-H consider opening this opportunity to NAICS code 541611 (Professional Services - Business Administrative Services)?

> A3: The NAICS code is 541715 (Research and Development in the Physical, Engineering, and Life Sciences (except Nanotechnology and Biotechnology)). This NAICS code best describes the principal purpose of the work to be carried out under the CARE ET module announcement.

Q4: Is the Basis of Estimate (BOE) document included in the 20-page limit?

> A4: The BOE is **NOT** part of the page count. Please see Amendment 01 to ARPA-H-24-01-04 which provides clarification.

Q5: Are key personnel biosketches required? If so, are they included in the 20-page limit?

> A5: See proposal template Volume 1: Technical and Management. The Bibliography to include qualifications (resumes) for each key personnel is optional and not included within the page count.

Q6: Are letters of support allowed?

A6: Letters of support can be provided to illustrate commitments from subcontractors, consultants, and/or key collaborators (as applicable) as part of the BOE. If provided, it would not be included within the page count.

Q7: The budget states that individual projects cannot exceed $8 million, but in the KILO description the range is $5-10 million.

A7: The CARE ET module announcement provides applicable module categories. CARE spans the BYTE and KILO module categories, and the aforementioned categories provide ranges of the estimated value of a potential award. With that being said, not to exceed amounts were provided, per Stage, to aid in preparation of proposal submission.

Q8: The Basis of Estimate (BOE) template is a Word document with embedded Excel files. These files are more difficult to work with when preparing the pricing. Could the Government provide the actual Excel document(s)/workbook(s) to aid in the development of pricing?

Q8: If working with the BOE template as provided is problematic, please provide your own spreadsheets which address the content requested within the embedded spreadsheets and/or tables included within the BOE template.

Q9: Is there an applicable NAICS code for this effort?

A9: The NAICS code associated with CARE is stated within SAM.gov (https://sam.gov/opp/8d3c07159e9f4d32ab7b179d365e6fab/view). It is 541715 - Research and Development in the Physical, Engineering, and Life Sciences (except Nanotechnology and Biotechnology).

Q10: Is the two-page bibliography inclusive of proposal references AND relevant publications/products for each key personnel? Or should a two-page bibliography with proposal references PLUS a resume for each key personnel be submitted?

A10: Per ARPA-H MAI-24-01-04 Volume 1: Technical and Management Template, the Bibliography is optional, and it is not part of the page limit. If a proposer chooses to submit a bibliography the template requests it be a brief (no more than 2 pages) that includes qualifications (resumes), links to relevant papers, references, reports as being an example. The stated 2-page limitation was associated with the bibliography, relevant papers, and/or references. Resumes can be submitted as separate documents, and, if provided, only for key personnel.

Q11: Does the proposal require biosketches, or other team descriptions?

A11: See the response provided in Answer number 5. Additionally, see the Volume 1: Technical and Management Template, specifically Section 4, Capabilities/Management Plan.

Q12**:** The RFP states that "ARPA-H CARE ET will emphasize creating and leveraging open-source technology and methodologies". Does this prohibit a chatbot that uses closed-models (e.g. Open AI's GPT-4, Anthropic's CLAUDE) that are widely accessible, but only through a commercial API?

A12: The CARE ET does not prohibit a chatbot that uses closed models. The reference to open-source technology and methodologies refers to the evaluative technologies being created within the CARE ET. The use of open or closed (i.e. commercial) medical LLMs are permitted when creating the open-source evaluation technologies.

Q13: Is further development of medical chatbots in scope for the CARE Exploration Topic?

A13: Further development of medical chatbots is out of scope for the CARE ET. The CARE ET focuses on developing evaluative tools, and proposals focused on improving existing medical chatbots will be deemed out of scope.

Q14: Can performers propose a specific pre-existing chatbot?

A14: Yes, proposers can propose a specific or multiple pre-existing chatbots.

Q15: The CARE Solicitation refers to medical chatbots. Will the program provide a standard suite of chatbots for performers to work with?

A15: Performers are not limited to a pre-defined list of medical chatbots. A non-exhaustive list of medically relevant chatbots is provided below. Proposers will be responsible for budgeting for the cost of using any necessary commercial LLM(s).

- Google Med-PaLM 2
- OpenAI GPT-4
- Hippocratic.ai
- Med-Gemini
- Meta ME-LLaMA
- Meta Llama 2 Meditron
- PMC LLaMA
- JSL Clinical QA BioGPT

CARE - Timeline and Key Milestones

Q16: ARPA-H-MAI-24-01 VOLUME 1 - ET Model OT_FINAL, Pg 33, Attachment 2 lists reporting requirements. Can the Government please confirm that these reports shown in Attachment 2 of the Model OT are in addition to the specific CARE ET deliverables listed in Table 2, Page 10 of the CARE module announcement?

A16: Yes, Attachment 2 of the ET model OT Agreement (page 33) are representative of the additional reporting requirements that likely will be included in resulting CARE OT Agreements.

Q17: A complete proposal submission due by June 3 is too accelerated. Would ARPA-H consider pushing the deadline back by a few weeks?

A17: The Government is unable to push the deadline at this time. Please continue to monitor the CARE ET module announcement (ARPA-H-MAI-24-01-04) on https://sam.gov for any potential (future) changes.

----------New Update as of May 16, 2024----------

CARE – General - Continued

Q18: Do you have any guidance on the profit level?

A18: Proposers should propose a profit/fee objective that is commensurate with the research effort from a technical risk/management perspective.

Q19: Does the government have an estimate of the total program funding?

A19: Per Section 3 of ARPA-H-MAI-24-01-04 CARE ET "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."

Q20: Should the performers expect to have access to the chatbot model itself?

A20: No, the purpose of the CARE ET is to develop technological approaches for evaluating chatbot outputs irrespective of LLM. The CARE module announcement states, "Evaluation methods should not depend on particular LLM features (e.g. accessibility of weights) and should be applicable to any chatbot."

Q21: In reference to both TAs, is the envisioned transition goal of the program to produce a reliable, trustworthy medical chatbot? And if so, is the expectation that it would be FDA approved?

A21: No, the goal of the program is to produce reliable, scalable technological approaches for evaluating chatbot outputs. This does not include the development of novel chatbots or improvements to existing chatbots.

Q22: Are two patient facing use cases in the same field acceptable?

A22: Proposers may choose two patient facing use cases in the same field. However, the CARE ET requires multiple use cases in order to assess the likelihood that the methods proposed can generalize to other kinds of patient-facing medical advice.

Q23: We do not have a negotiated indirect cost rate/agreement. Is a negotiated rate agreement needed before the project is awarded to us?

A23: See the Indirect Cost section (Section 7) of the BOE for the level of detail that is needed at the Stage 1, Volume 1 submission. Proposers should be ready to provide supporting documentation within the Stage 2, Volume 2 submission to justify the costs

included within the BOE and the resulting price/cost submission if progressing on to Stage 2, Volume 2. All proposed costs (to include indirect costs) will be negotiated prior to award.

Q24: Are HSR approvals at the institute level sufficient, or is there a required government level review of locally approved protocols?

A24: Any institution or entity engaged in ARPA-H funded research involving human subjects must 1) obtain a Federal-wide Assurance (FWA) of Protection for Human Subjects from the Office for Human Research Protection (OHRP), 2) obtain initial and continuing approval (if required) of the research by an appropriately constituted and registered Institutional Review Board (IRB), and 3) submit the full IRB approved package (including protocol, IRB approval letter, any required training certificates, etc.). For instructions on registering IRBs, obtaining FWAs, and completing Human Subjects Education requirements, see the OHRP website at: https://www.hhs.gov/ohrp.

Q25: Is it preferred that the work be done on an existing production medical chatbot, or should new research chatbots be built for each use case?

A25: The evaluation technology developed should apply to any chatbot. Building new chatbots is out of scope for this solicitation.

Q26: What are the security requirements for the proposed system? Is FedRAMP, CUI, PHI, HIPAA, PII, and/or US Citizens only required?

A26: Please see the Master Announcement Instruction (MAI), ARPA-H-MAI-24-01 for security and eligibility requirements.

Q27: Is ARPA-H looking for production-grade software, or proven tools and methodologies that achieve the metrics set out in CARE ET?

A27: Performers must deliver running code in the form of a live software document that meets the specified metrics during performance. The CARE ET does not define nor does it require “production grade” software.

Q28: Some contracts include language stipulating that the client is not to be disclosed. To honor the confidentiality requirements of existing contracts, will the government accept sanitized current and pending disclosures? This is in regard to the disclosure of potential conflicts of interest and commitment requirements stipulated in Administrative and National Policy requirements document (Section 3 of said document).

A28: To be clear, the information requested in Section 3 of Volume 1, Administrative & National Security template requests this level of detail only for PIs and other senior/key personnel (e.g. individuals substantively involved in the research).

Q29: Will the government confirm that, in accordance with the definition of Senior/key personnel in the NSPM-33 Implementation Guidance, there is no direct correlation between the proposed Level of Effort (LOE) on the program and designation as key personnel?

A29: Senior/Key personnel are individuals whose participation is essential to the overall success of the work to be performed under the awarded effort; these are individuals who have authority or responsibility for a project's design or management. There is no direct correlation between the proposed LOE and this designation.

Q29: Could ARPA-H provide a specific list of inaccuracies that are understood by 'other inaccuracies'? This is important because LLMs can be inaccurate in a number of different ways, and catering to all of them can significantly increase the effort required to respond to this request.

A29: The other types of inaccuracies to be detected are to be defined by the stakeholder engagement activities described in TA1.1 of the CARE ET.

Q30: Will the government consider conducting a price reasonableness evaluation instead of a cost reasonableness evaluation?

A30: See Evaluation criteria No. 4 (Section 4.2) as stated in the Master Announcement Instruction (MAI), ARPA-H-MAI-24-01. Also see Section 4.1, Evaluation Stages of the MAI which indicates that only those proposals that are selected for a potential award will be required to submit a Volume 2, which will then be evaluated against evaluation criteria 4.

Q31: What data is in scope for the medical chatbot? Specifically, is PHI data in scope? This has significant impact on privacy and governance.

A31: The use of PHI in the development of chatbot evaluation technology is permitted, but not required. If a proposal involves the use of PHI, the proposal must also specify how privacy and governance requirements for managing PHI will be met.

Q32: Can we submit support letters from companies or clinicians that they are willing to provide advice and test our medical chatbot?

Q32: Please see Q&A #6 above.

Q33 The TECHNICAL & MANAGEMENT Template requests a 120-day proposal validity period. The other templates in the package request a 180-day proposal validity period. What should the proposer list as the proposal validity period?

A33: Please list 180-days proposal validity period for all documents.

CARE - Timeline and Key Milestones - Continued

Q34: Will performers need to allocate time in Stage 2 for testing other vendors' technologies?

A34: Yes.

Q35: What are the expectations about the timing of the exploration of the two use cases? For instance, should the Stage I milestones cover both use cases? Alternatively, can/should the exploration of the second use case be staggered later in the timeline? (And/or should it be reserved for testing generalization in Stage 3?)

A35: The Stage 1 milestone is not required to cover both use cases. The second use case can be addressed prior to Stage 3 but is not required to be.

Q36: For Milestone #6, given that the nature of the other team's use cases is undetermined, would the government either provide some parameters to inform the milestone pricing or allow for the potential to update the pricing once the uses case are known?

A36: Please review Section 5.A of CARE ET module announcement regarding module categories and not to exceed amounts per Stage.

Proposers should provide a basis of estimate in the Stage 1, Volume 1 submission that is aligned with the proposed technical approach. Please also note that milestone amounts will be calculated based on the proposed accumulation of monthly amounts up to each milestone date (as stated in paragraph D of the CARE ET module announcement).

Q37: May performers propose a milestone at contract execution?

A37: No. As stated in the CARE module announcement, the fixed milestones provided in Table 2 of the CARE ET module announcement will be directly incorporated into Attachment 3 of the OT Agreement ("Agreement Term, Deliverables, and Payment Schedule").

Q38: Table 2 mentions "Jupyter Notebook" as a valid deliverable. Does this mean that a full-fledged, deployed "chatbot" is NOT required?

A38: Correct, the deliverables are related to technological approaches for evaluating medical chatbots. Developing or deploying a chatbot is out of scope.

CARE - Exploration Topic Structure

Q39**:** Can we have two patient facing use case in the same field, or do we need two different fields?

A39: See answer to Q22.

Q40: Different medical chatbot applications can have significantly different patient risks. Can ARPA-H please elaborate on the types of use cases that are in scope and the expected risks?

A40: As stated within TA 1.1 of the care module announcement, use cases that are in scope must be patient-facing and defined by a coherent and significant medical need faced by a well-defined community of patients and/or caregivers. Technology developed

on the proposed use cases must generalize to other kinds of patient-facing medical advice. Expected patient risks used for chatbot evaluation will be defined during stakeholder engagement activities in TA1.1.

Q41: How are "unusual", "highly specific", and "very small populations" defined?

A41: The phrase “Use cases that require unusual or highly specific expertise, outside the realm of standard patient care, or address only very small populations of patients" is intended to describe situations where the resulting evaluation approach has a low probability of generalizing to other types of patient-facing medical advice. A proposer concerned of running afoul of this requirement should provide justification for why the proposer believes that use case will generalize.

Q42: Do you anticipate that stakeholder engagement approach will require IRB review or will it be considered a Quality Improvement (QI) project aligning with the Centers for Medicare and Medicaid Services (CMS) definition of QI (42 CFR §476.1).

A42: Research that solely involves interpersonal interactions with participants such as surveys, interviews, and focus groups most often qualifies for review under exempt or expedited categories. Unless there is some unusual risk to participants in the stakeholder engagement activities proposed, full IRB review is likely not necessary.

Q43: Would the government consider an evaluative software tool that includes a commercially licensed LLM as meeting the criteria outlined in Section E?

A43: An evaluative software tool that includes a commercially licensed LLM could meet the requirements of Section E but note that it “emphasize[s] creating and leveraging open-source technology and methodologies,” and “[o]pen-source code is highly encouraged using permissive, business-friendly open-source licenses.”

CARE - TA 1.1 : Pragmatic Criteria and Prompt generation

Q44: Would an approach of using a moderated patient forum for patients to ask disease-specific questions, fully fleshed out and backed by a large existing user network and market footprint, be a suitable alternative or compliment to other automated approaches?

A44: Using a moderated patient forum for patients to ask disease-specific questions would fall within the scope of TA1.1 of the CARE ET. Prompt generation technologies must achieve the outlined metrics for prompt scaling, novelty, diversity, and coverage.

Q45: Would exploring use cases around how a patient's medical issues may be moderated by proper usage and ongoing sustainment of therapies with medical equipment within the scope of medical use cases ARPA-H is looking for?

A45: As stated within the CARE ET, "use 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." 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 proposed use case will generalize to other use cases.

Q46: Given budget constraints and to help proposers price correctly, can the government provide guidance on the desired size of stakeholder panels?

A46: The government cannot provide guidance on the desired size of stakeholder panels.

CARE - TA 1.2 Scalable, expert chatbot evaluation technology

Q47: Are medical use cases within scope?

A47: As stated within the CARE ET, "use 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." Mental health as a potential medical use case is likely within scope of the CARE ET. 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 that the proposed use case will generalize to other use cases.

Q48: Is it okay to use online cloud databases as part of the evaluation method?

A48: Yes. Note that required deliverables include "all software, source code, documentation, and a set of examples for using the code in the form of a live software document" required to perform the evaluation. If non-public cloud databases are used, they must be provided as part of the deliverables.

Q49: Is the goal of the CARE ET project only to evaluate LLMs, or also to improve them, and then to evaluate the improved LLM?

A49: The goal of the CARE ET is to produce reliable, scalable technological approaches for evaluating chatbot outputs. This does not include the development of novel chatbots or improvements to existing chatbots.

Q50: Is the awarded contractor expected to perform IV&V or is an independent contractor being provided by ARPA-H for T&E and IV&V?

A50: See the IV&V paragraph contained within section D of the CARE ET. As stated therein, "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."

Q51: Is the intention that the tooling created at the end of the performance period is used in real time to evaluate chatbot output on every response from the chatbot? Or is the intention that periodic QC is performed?

A51: The CARE ET does not require “real time” performance of the evaluation method. Note, there is a project metric related to how long it takes to evaluate a chatbot output.

Q52: TA1.1 outlines that two patient-facing medical chatbot use cases must be chosen. It also states that proposals should select use cases that are likely to generalize to other kinds of patient-facing medical advice. Could you provide clarification on whether the two use cases must generalize to different types of patient-facing medical advice?

A52: The two use cases do not have to generalize to different types of advice. The selection of use cases should be such that the resulting evaluation technology is generalizable to different types of patient-facing medical advice.

Q53: TA1.1 states that the "Uses cases must be defined by a coherent and significant medical need." How are "coherent" and "significant" defined?

A53: A coherent medical need is a well-defined clinical concern. Significant medical needs include but are not limited to those that impact many people, interfere with activities of daily living, or require frequent medical interventions.

Q54: If patient data is used in the development of the evaluation technology, is the raw data (individual-level patient data) expected to be uploaded to the designated data repositories?

A54: ARPA-H will work with the performer to determine the best approach for making data available while maintaining patient data privacy.