DPA26BZ06-DV026
Influence Benchmarks for AI Systems
Frequently Asked Questions (FAQs)

September 16, 2026

1. Does the black-box requirement permit an external integrity and provenance layer that records prompts, outputs, timing, market state, and classifier results, provided it does not inspect model weights or hidden reasoning?
   A: Recordings of agent prompts, outputs, actions, market state, and classifier results within the test environment is in scope.

2. For comparison with human behavior, may Phase I combine published auction benchmarks with a limited prospectively collected human dataset, or does DARPA expect a new original human-subject dataset of a particular scale?
   A: Usage of published auction benchmarks in Phase I to evaluate the performance of the test environment is in scope. Proposers may choose to use publicly available datasets or use internal datasets that have been previously acquired. However, collection of human-subject datasets for any purpose in this SBIR is out of scope.

3. Should the Phase I classifiers be optimized primarily for detecting behavioral bias and deception, or for measuring deviation from human market-outcome distributions?
   A: In Phase I, performers should establish a framework to compare the behavioral preferences of AI agents. Performers must also be able to evaluate the performance of their market test environment using suitable metrics. Such metrics may include but are not limited to measures of market efficiency and measures of agent behavior in the test environments. Measures of behavioral bias and deception should be established in Phase II.

4. Does DARPA have a preferred market or auction model for the simulated environment, or is that left up to the proposer?
   A: DARPA does not have a preference for the auction or market model for the simulated environment. As stated in the topic, the test environment should be able to accommodate different models of markets and auctions.

5. For the requirement to use at least 10 different LLMs, is DARPA looking for specific models, or can the proposer determine the appropriate mix?
   A: Proposers should determine the appropriate mix of models to use in Phase I.

6. For comparing AI behavior to human market behavior, can existing research and market data be used, or is DARPA expecting new human-generated data?
   A: Use of existing research and market data is acceptable. Collection of new human-generated data or other human subjects research is out of scope for this SBIR.

7. Would using commercial real estate as an initial market domain be responsive if the underlying testbed is designed to support multiple auction/market structures and the Phase I milestones in the topic?
   A: In general, the use of specific domains where multiple market models may exist or co-exist is acceptable. As stated in the topic, the test environment must be able to accommodate multiple market or auction models.

8. At proposal submission, must a Phase I proposer already possess human-market comparison data and classifiers, or may those datasets/classifiers be developed during the 12-month Phase I effort as part of the stated milestones?
   A: As stated in the topic, by month 12, markets must be able to replicate bidding patterns and allocation efficiencies observed in human markets. Prospective proposers should, in their technical proposal, describe their approach to meeting these metrics by month 12. Please note that no human subjects research is out of scope for this SBIR.

9. For the required greater than 90% allocative efficiency, how does DARPA intend allocative efficiency to be defined and evaluated? In particular, should this represent realized welfare relative to an optimal allocation, and is the threshold expected per episode or across repeated trials?
   A: Proposers should provide definitions of allocative efficiency that is consistent with the market or auction models that are implemented in the test environment. Proposers should define optimal allocation within their implemented market or auction models and describe measurement approaches to determine the allocative efficiency within their test environment. The >90% allocative efficiency metric should be shown across multiple trials.

10. For the requirement to demonstrate stock agents drawn from at least 10 different LLMs, what constitutes a "different LLM"? For example, does DARPA intend distinct model families/base models, or may different model versions, sizes, or fine-tunes within a family qualify?
    A: For the purposes of this SBIR, different model families and different model versions within LLM families are considered different LLMs.

11. What forms of human comparison are considered acceptable for Phase I? Would literature-derived behavioral models and archival human auction datasets satisfy the requirement, or does DARPA anticipate collection of new human-subject data?
    A: For Phase I, literature-derived behavioral models and archival human auction datasets are acceptable. Collection of new human-generated data or other human subjects research is out of scope for this SBIR.

12. For the required classifiers of biases in AI-agent behavior and market outcomes, does DARPA have a preferred interpretation of "bias" or expected ground-truth methodology, or may performers define operational behavioral constructs and validate them through controlled positive/negative experimental conditions?

A: As stated in the topic, proposers must define an error function that describes differences between AI agent decisions and expected human results. Proposers must provide quantitative metrics on how social behaviors between AI agents alter market efficiency and outcomes and generate testable hypotheses for how latent AI agent preferences affect social interactions, dynamic responses to external stimuli, and ultimately market outcomes.

13. Is measurement of deception and collaboration expected as part of the Phase I proof-of-concept demonstration, or is Phase I primarily expected to establish the test environment, behavioral measurement methodology, and preliminary instrumentation for these behaviors?
    A: The focus of Phase I is to establish the test environment and measurement methodology. Measurements of deception and collaboration are to be established in Phase II.

14. When a latent behavioral preference cannot be distinguished reliably from factors such as model competence, stochasticity, prompt sensitivity, or context limitations, is an explicit "insufficient evidence" or abstention outcome considered an acceptable classifier result?
    A: Measurements and classifications of latent behavioral preference and potential outcomes should be defined by the proposers and supported by their choice of models, metrics, and experimental plan.

15. For Phase I validation, which form of generalization is most important to DARPA: performance on unseen LLMs, unseen market mechanisms, unseen information/news interventions, or some combination of these?
    A: As stated in the topic, the goal of Phase I is to develop and define the architecture of the test environment and demonstrate a functional proof of concept for a simulated auction or market (heretofore "market"). The test environment must be able to accommodate different models of auctions and markets to evaluate agents on different measures of market efficiency.

16. For the Month 12 demonstration of allocative efficiencies greater than 90%, is the threshold intended for a reference-agent population, the suite drawn from at least ten different LLMs, or another population? Should it be met separately for each market configuration, or across an aggregate evaluation suite? If proposers should define and justify these choices, please confirm.
    A: The 90% allocative efficiency metric is an evaluation of market performance and should be met for each market or auction model implemented in the test environment. Performers should justify choices of agents for their market demonstration in Phase I.

17. For Phase I, must the delivered environment be independently installable and runnable by Government personnel, including source code, or is a contractor-operated demonstration with replayable data and documentation sufficient? The topic explicitly calls for source

and object code delivery in Phase II; we want to distinguish the Phase I delivery requirement correctly.

A: As stated in the topic, proposers should demonstrate an operational test environment by Month 12. No delivery of source and object code is required at the end of Phase I.

18. May an eligible Phase I proposer incorporate a separately licensed third-party assurance layer of this kind as a subcontracted or partner component?

    A: Yes. In general, proposers are free to incorporate subcontractors into their proposal. Please note that the prime proposer must perform at minimum two-thirds of the total work proposed, as measured by direct and indirect costs.

19. Would an approach be responsive if the core deliverable is the simulated-market testbed and benchmarking system required by the topic, while that external provenance layer is used to support reproducibility, traceability, and comparison across the required agents?

    A: In their proposal materials, proposers should describe their approach to meeting the technical objectives and deliverables outlined in the topic. Proposers are encouraged to consider how each component of their solution contributes to the overall technical objectives of DPA26BZ06-DV026.

20. Are there any topic specific restrictions on commercial large language model application programming interfaces, open weight models, cloud regions, or retention of model outputs?

    A: There are no topic-specific restrictions on the use of commercial or open-weight large language models. However, proposers are expected to comply with all relevant laws and policies regarding cybersecurity and the procurement of goods and services from foreign countries of concern.

21. What is the accepted reference optimum and measurement approach for the greater than 90 percent allocative efficiency requirement across multiple market mechanisms?

    A: As stated in the topic, the test environment must be able to accommodate different market or auction models. Proposers should provide definitions of allocative efficiency that are consistent with the market or auction models that are implemented in the test environment. Proposers should define optimal allocation within their implemented market or auction models and describe measurement approaches to determine the allocative efficiency within their test environment.

22. Permitted model access. Should every behavioral measurement rely solely on submitted inputs and observed outputs, and what controls are expected over prompts, memory, available tools, and model randomness to support reproducible comparisons?
A: As stated in the topic, proposers should establish measurement approaches that rely on observable inputs and outputs to agents. Proposers should not assume access to the internal state (prompting, memory usage, etc.) of an agent in their approaches.

23. Changing information. For experiments involving public news or other dynamic information, should performers use controlled historical or synthetic event streams, live information, or both, and how should information exposure be recorded so that results remain reproducible?
A: As stated in the topic, proposers should define their approach to building a "news feed" which simulates public information available to the market. Please note that source and object code for the test environment, including implementation of the "news feed", is a deliverable in Phase II of the SBIR.

24. Phase I demonstration. What minimum demonstration and documentation will satisfy the month 12 proof of concept, particularly regarding the relationship among human-like bidding behavior, allocative efficiency, and identification of behavioral preferences?
A: Information on the technical deliverables expected for Phase I can be found in the SBIR topic description. Proposers should define human-like bidding behavior and allocative efficiency for the market or auctions models that will be implemented in Phase I.

25. Delivery and continued operation. Given the later delivery of source code, object code, and data to government partners, what assumptions should govern third-party model licenses, redistribution rights, external service dependencies, and the government's ability to operate the benchmark independently?
A: Third-party licenses, redistribution rights, or dependencies should not hinder the government's ability to independently evaluate the capabilities of the test environment.