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

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.

26. On the black-box constraint: The topic specifies the testbed "shall rely only on queries and outputs of the subject AI system, rather than direct access to the model itself." For the DP2 feasibility documentation, should proposers demonstrate prior work using commercial API access (standard LLM provider APIs), or is there an expectation of local/air-gapped model compatibility?
A: Proposers for a DP2 proposal should provide documentation that describes the current status of their solution relative to the goals of Phase I of the topic.

27. On feasibility documentation: The Phase I section lists four fixed payable milestones (architecture report, scaling report, stock agent demo, testbed demo). For a DP2 proposal, should the feasibility documentation address ALL four milestones, or may proposers focus on demonstrating the core capability (black-box behavioral monitoring) and propose completing the market-simulation and stock-agent milestones during Phase II?

A: The feasibility demonstration should address *all* of the technical goals of Phase I.

28. Model access and sandboxing. The FAQs permit commercial LLMs but state that proposers should not assume access to an agent's internal state, including prompting and memory. May the test environment configure API-exposed parameters such as system prompts, temperature, and sampling, provided behavioral assessments rely solely on observable inputs and outputs? Does the sandbox requirement permit stock agents to use commercial hosted models through APIs?
    A: Proposers may configure LLM models as necessary; however, evaluations of LLMs and agent behavior should only rely on observable inputs and outputs. Proposers may use commercial models through API calls.

29. Allocative-efficiency benchmark. We understand that proposers must define an appropriate optimum for each implemented market or auction model and demonstrate greater than 90% allocative efficiency across multiple trials. Should this optimum represent the theoretically efficient allocation, with human experimental outcomes serving as a separate behavioral comparison, or may an empirical human baseline serve as the denominator?
    A: An allocative efficiency of 100% should represent the theoretical maximum allocative efficiency in the market or auction model. Human outcomes should be evaluated against the theoretical maximum.

30. Human comparison data. We understand that existing datasets and literature-derived behavioral models are acceptable and that new human-subject data collection is out of scope. Will DARPA furnish any reference datasets, or should proposers plan to identify and obtain all comparison data independently?
    A: Proposers should identify all reference datasets that will be used in their model development.

31. Phase I classifiers. The FAQs distinguish Phase I behavioral comparison and measurement methodology from Phase II measurements of bias, deception, and collaboration. Could you clarify the minimum expected Phase I classifier capability and whether proposers should define and justify their own reference labels or ground-truth methodology?
    A: There are no performance metrics on classifiers in Phase I. However, proposers should plan to demonstrate an operational classifier at the end of Phase I. Proposers should define their reference labels of classifier behavior in Phase II.

32. Magentic Marketplace. Is the reference to Magentic Marketplace illustrative, or is there a preference for extending that environment? Would an approach built on it and an independently developed testbed be evaluated on the same basis, provided both meet the topic requirements?
    A: The reference to Magnetic Marketplace is illustrative. All proposals will be evaluated on the same basis as all other proposals of the same type (Phase I or Direct to Phase II).

33. The topic states that stock AI agents "shall only be used within the test environment sandbox and shall not be used on other systems." Does this permit stock agents to call commercially hosted model interfaces from within the sandbox, or does the Government expect stock agents to be hosted entirely within the proposer's test environment?
A: Yes, stock agents may call commercially hosted model interfaces from within the test environment.

34. Phase II anticipates delivering software and data for testing by U.S. Government partners. Can the Government identify the earliest expected types of users or organizations for the test environment so that Phase I architecture and documentation choices can support their equities/interests?
A: No.

35. The Phase I text describes the >90% allocative efficiency target alongside replicating bidding patterns and allocation efficiencies observed in markets. Is the target intended mainly to validate the fidelity of the market environment, or must populations of LLM-based stock agents achieve it?
A: The allocative efficiency metric is to evaluate the fidelity of the market environment.

36. Anthropic models during the project: Can an awardee use Anthropic's Claude models only as subjects being tested in the research, given the current government restrictions involving Anthropic? Does the answer change if we access Claude through another company's hosted service? If approval or an exception is needed, who provides it, and when should we request it?
A: The technical objective of the SBIR is the development of a test environment which can be used by the government to evaluate latent preferences in AI systems. Proposers are expected to comply with the latest policies and regulations for acquisitions for the DoW SBIR program. It is not expected that proposers will require access to specific models to meet the technical objectives of the topic.

37. Training examples from other models: Some U.S.-developed models were trained partly on examples generated by models such as DeepSeek or Kimi. Do any applicable restrictions cover using those U.S.-developed models when we are not using DeepSeek or Kimi services ourselves? What checks or documentation would you expect?
A: Proposers are expected to comply with the latest policies and regulations for acquisitions for the DoW SBIR program. In general, US-developed and hosted models are acceptable. Proposers may use software and AI models which comply with these regulations to develop the test environment; however, it is not expected

that proposers will require access to specific models to meet the technical objectives of the topic.

38. Government access to hosted models: FAQ 25 says outside licenses and services must not prevent independent Government evaluation. Would a plan using hosted model APIs, with suitable licenses, separate evaluator accounts and a budget for usage, meet that requirement? Does any part of the Phase I demonstration or later Phase II delivery need to run without outside model services?
    A: As stated in responses to previous questions, any licenses and agreements used in the development of the test environment should not preclude the government from evaluating the final deliverables. Proposers may use outside model services as part of the final deliverable, but their use and purpose within the test environment should be justified as part of their proposal.

39. The topic describes a simulated market as the test environment. Should a Phase I testbed run open weight models, commercial hosted models, or both?
    A: Proposers may use open-source models, commercially hosted models, or both in development of their test environment.

40. The topic covers both hidden bias in users and deceptive strategy between agents. If Phase I can cover only one of those well in nine months, which is more useful to the program?
    A: Proposers should develop a technical plan which addresses both aspects. Please note that detection of bias and deception in AI models should be included in the technical plan for Phase II.

41. Are there existing government testbeds, datasets or scenario sets you would want a proposer to build on rather than start from scratch?
    A: No government furnished information or equipment will be provided as part of this topic.

42. Beyond encouraging delusions and eroding military discipline, what broader taxonomy of human behaviors of concern should the human representative AI agent be designed to probe?
    A: Those examples are illustrative and modeling those outcomes is not the focus on this SBIR topic. The agent representing "human" behavior should exhibit "human-like" decision making within the marketplace, as defined by the proposer's model of human behavior.

43. To better align our proposal with the program's end goals, are there any additional requirements, priorities, or recommendations that are not fully addressed in the solicitation?
A: No.

44. Allocative efficiency target. The Month 12 milestone calls for a demonstration with allocative efficiencies above 90%. Should that level be achieved by populations of LLM stock agents, or is it a capability of the test environment that may be shown with rule-based reference traders (for example budget-constrained zero-intelligence traders in a double auction) while the efficiency achieved by each LLM population is reported as a measured result? Is there a preferred market design for this demonstration?
A: The 90% allocative efficiency metric should be met through demonstrations on agent populations.

45. Allocative efficiency: How should the requirement to demonstrate greater than 90% allocative efficiency be interpreted? Is this intended to establish the validity of the simulated market using appropriate reference agents, or are the evaluated LLM agents themselves expected to achieve greater than 90% allocative efficiency?
A: Allocative efficiency is a measure of market outcomes and not individual agent outcomes.

46. Counterfactual testing: Would matched experiments in which the same AI system is evaluated in otherwise identical market scenarios while a single information or social variable is changed be consistent with DARPA's intent for characterizing latent behavioral preferences?
A: We cannot determine the suitability of the experimental platform without a full proposal submission.

47. Classifier validation: Would programmed reference agents with known strategies or deliberately introduced behavioral tendencies be an appropriate method for establishing ground truth and validating classifier accuracy before applying the classifiers to black-box LLMs?
A: Proposers should justify their classifier architecture and validation approaches in their technical proposals.

48. Phase I evidence of success: Beyond the stated milestones and allocative-efficiency requirement, what would DARPA consider the strongest evidence at Month 12 that the test environment has successfully characterized a latent behavioral preference of an AI system rather than simply observed different outputs from different models?
A: The exit criteria for Phase I are listed in the topic description.

49. Are there particular biases or latent preferences DARPA would most value seeing characterized in the Phase I demonstration? Are ethnic, religious, and nationality-based preferences within the intended scope alongside cognitive biases such as confirmation bias, overreliance, and sycophancy?
A: DARPA has no preference for the types of biases studied in Phase I.

50. FAQ 31 establishes that Phase I requires an operational classifier without specified performance targets. What outputs should that classifier produce by Month 12—for example, identifying specific behavioral tendencies, grouping agents by revealed preferences, or distinguishing patterns of bidding and market outcomes?
A: In their Phase I proposals, proposers should define the inputs and outputs of their classifier to support their technical approach in both Phase I and Phase II.

51. Who must reproduce human bidding behavior—the LLM agents or separate reference agents? For the Phase I requirement to replicate bidding patterns observed in human markets, must the LLM-based stock agents themselves reproduce those patterns, or may separate reference agents calibrated to archival human behavior establish that capability while the LLM agents are evaluated for their differences?
A: Proposers should define their technical approach to modeling human behaviors within their technical approach in Phase I.

52. By Month 12, does DARPA expect multiple distinct market or auction mechanisms to be implemented and individually validated, or is one validated mechanism sufficient if the architecture demonstrably accommodates additional mechanisms? Is there a minimum expected number?
A: DARPA has no minimum or maximum number of market models for implementation in Phase I.

53. Black-box observation boundary. The topic states that the testbed shall rely only on queries and outputs of the subject AI system rather than direct access to the model. Does this preclude use of model-side telemetry such as token probabilities/log-probabilities, internal reasoning traces, activations, embeddings, or confidence outputs, such that characterization must derive exclusively from externally observable inputs and outputs?
A: Characterization of AI agents must derive exclusively from externally observable inputs and outputs.

54. Pre-outcome behavioral transitions. Would DARPA consider within scope classifiers or quantitative measures designed to identify statistically significant changes in observable agent behavior before those changes manifest as degraded market outcomes, provided the measures operate exclusively on permitted black-box observations?
A: Yes, such measures are considered in scope.

55. Controlled information asymmetry. In addition to the public dynamic “news feed,” may the test environment provide controlled private information to selected agents to experimentally distinguish error, bias, strategic adaptation, collaboration, and deceptive behavior?
A: Yes, this is in scope.

56. Proposer-defined quantitative measures. For behaviors including collaboration and deception, may proposers define and validate their own quantitative behavioral measures, provided those measures are reproducible, operate within the specified black-box boundary, and can be related to market efficiency and outcomes?
A: Yes, proposers may define and validate their own behavior measures.

57. Generalization to unseen models. Is DARPA's intended end-state a model-agnostic benchmark capable of accepting a previously unseen AI system as a test agent and characterizing its behavioral preferences without retraining the test environment or behavioral measures specifically for that model?
A: As stated in the topic, proposers should deliver a test environment to characterize the latent preferences of a test AI agent. The test environment should be delivered in a state that allows DARPA and/or other US government personnel to operate the environment and validate its performance.

58. The answer to FAQ question 36 refers to "current government restrictions involving Anthropic." (a) Which restriction does this refer to, and where is it published? (b) May Anthropic models be among the LLMs evaluated as stock agents inside the test environment, either through Anthropic's interface or through another company's hosted service? (c) May a proposer use Anthropic models as development tooling, for example coding assistance, if no Anthropic model is part of the delivered environment?
A: It is not expected that proposers will require access to specific AI models or coding assistance tools to meet the technical objectives of the SBIR.

59. Allocative efficiency (FAQ questions 16 and 44): may the Month 12 demonstration show above 90 percent efficiency for each market model on a proposer-defined mixed population of LLM stock agents, with each single-LLM population reported as a measured result, or must every one of the ten or more LLMs reach 90 percent on its own?
A: Month 12 demonstrations of allocative efficiency may be performed on mixed populations of agents, of the performer’s choice.

60. The Month 12 efficiency target. FAQ 44 asked whether the 90% allocative efficiency level could be shown with rule-based reference traders, and the answer was that it "should be met through demonstrations on agent populations." FAQ 16 adds that it "should be met for each market or auction model implemented." Must the 90% be met by populations of AI agents in each market or auction model, with rule-based traders used only as a reference?
61. **A: The 90% allocative efficiency should be met by populations of agents and not rule-based traders.**

62. Would a Phase I award on this topic include a restriction on Anthropic products, such as a clause, a contracting officer direction, or a non-use certification? If so, which of these would it cover: Claude Code as our software development tool, including code written with it before award; Claude models tested as black-box subjects through a service that hosts or resells models, such as OpenRouter or Amazon Bedrock; or contract information stored or processed in Claude? If Anthropic products are not allowed, we would do all Phase I development with OpenAI's Codex and would not test any Claude models. Would that meet the requirement?
A: Awards made under this SBIR will reflect current policy and regulations. It is not expected that proposers will require access to specific AI models or coding assistance tools to meet the technical objectives of the SBIR.

63. Month 12 criterion. The topic states that the test environment must be able to replicate bidding patterns and allocation efficiencies observed in markets and demonstrate allocative efficiencies above 90 percent. Is that criterion satisfied by validating the environment with calibrated programmed populations (for example, equilibrium bidders and zero-intelligence-constrained traders reproducing published human efficiency and bidding results), with the LLM stock agents then measured against those baselines, or must the LLM stock agents themselves reproduce human bidding patterns and exceed 90 percent efficiency?
A: The 90% allocative efficiency metric must be met with AI-based agents. Rule-based traders or zero-intelligence agents would not satisfy this metric.

64. Subject models. FAQ 20 states there are no topic-specific restrictions on commercial or open-weight LLMs, subject to policies on procurement from foreign countries of concern. May open-weight models released by organizations in a foreign country of concern be run as subject agents inside the sandbox, for the sole purpose of being measured, or should the 10-LLM suite exclude them?
A: DARPA does not expect that proposers will require access to specific AI models in order to meet the technical objectives of this SBIR topic.

65. For the Phase I Month 12 milestone, may a population consisting entirely of non-LLM reference agents satisfy the >90% allocative-efficiency requirement for each implemented market model, while LLM-population efficiencies are reported separately?
A: No, measurements of market efficiency should be met with the participation of AI agents in the marketplace.

66. Under FAQ 18's Phase I two-thirds work requirement, do standard metered commercial LLM API charges count as work performed by others when the provider performs no project-specific research?
A: DARPA cannot provide an answer to this question without a full proposal submission. Proposers should review the cost proposal instructions to accurately classify and document the costs associated with their proposals.

67. For the Phase I, Month 12 milestone, may a population consisting entirely of non-LLM reference agents satisfy the >90% allocative-efficiency requirement for each implemented market model, while LLM-population efficiencies are reported separately?
A: No, measurements of market efficiency should be met with the participation of AI agents in the marketplace.

68. Allocative efficiency target. The Month 12 milestone calls for allocative efficiencies above 90%. Must this be demonstrated in every auction or market model the environment supports, or in at least one? Does it apply to markets composed entirely of stock AI agents, and to each of the ten or more LLMs individually or to the suite as a whole?
A: 1The allocative efficiency metric should be met through demonstrations with AI agent populations. The 90% allocative efficiency should be met for every market or auction model proposed, using a mixed AI agent population derived from multiple models.

69. Composition of stock agents**.** May a stock agent include supporting structure around the LLM, such as fixed prompts, memory or validity checks on bids, or should it be the unmodified model? Do different versions or sizes of a model from one developer count as different LLMs toward the ten?
A: Proposers should include their agent architecture within their proposal materials. Additionally, see answer to question 10.

70. Model access. Should proposers budget for LLM access and compute, or will any be Government-furnished?
A: Proposers should budget for LLM usage. There is no government furnished equipment or information available to the proposers.