Measuring Large Language Model Understanding of Federal Statistical Data

(MLMU-25) FAQ

	Question	Answer
1	Are there specific statistical domains or datasets within Commerce that should be prioritized for the evaluation (e.g., economic data, census data)?	As noted in the RFS, the case study should be applied to a representative sample of the complex portfolio of Commerce public-facing statistical products and datasets.
2	Is the primary goal to assess whether federal statistical data is AI-ready, to evaluate the current capabilities of LLMs in understanding unready data, or both? Should the solution focus on improving data readiness or testing the maturity of AI systems?	We look to vendors to propose solutions that meet the project's objectives of developing an empirical evaluation of LLM performance with federal statistical data, including an assessment of how well these statistical data assets are structured to support effective LLM interaction.
3	What level of granularity is expected in the prompt-response pairs? Should they address simple queries, complex analytical tasks, or both?	The level of granularity in prompt-response pairs must meaningfully address LLM performance across the complex portfolio of federal statistical products and data assets.
4	Are there predefined metrics or benchmarks for evaluating LLM responses, or should the offeror propose these metrics?	We look to the vendor to propose performance metrics that align with the objectives stated in the RFS.
5	Should the evaluation framework include separate methodologies for assessing LLM performance and identifying gaps in metadata or data readiness?	As stated in the RFS, the evaluation should assess both the quality of LLM responses and the extent to which federal statistical data assets are structured to support effective LLM interaction.
6	Are there specific metadata attributes or standards (beyond DCAT-US 3.0) that should be emphasized in the evaluation?	All relevant metadata standards are referenced in the RFS.
7	Should the evaluation focus equally on older datasets with legacy metadata and newer datasets with updated standards, or is there a priority?	As noted in the RFS, the evaluation should include both legacy and more recently published data assets. We look to vendors to propose solutions that address differences in AI-readiness and usability that may result from evolving publication standards.
8	Should the solution include actionable recommendations for improving metadata and data formatting to enhance AI-readiness?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS.

	Question	Answer
9	Are there specific technologies (e.g., GenAI, Knowledge Graphs, Explainable AI) that the government encourages for inclusion in the solution?	While open-source solutions are preferred, proposals may include non-open-source solutions with clearly specified benefits on why one solution was selected over another to meet the objectives stated in the RFS.
10	Should the solution address multimodal data formats (e.g., combining text, images, and tabular data) for LLM interaction?	We look to vendors to propose solutions that meet the project's objectives of assessing LLMs' ability to interpret and use federal statistical data. Solutions should address a representative sample of the complex portfolio of public-facing federal statistical products and data assets.
11	Should the solution explore whether LLMs can adapt to unstructured or less-than-optimal data formats without additional preprocessing?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS. As stated in the RFS, key research questions include investigating the relationship between the quality of federal statistical data assets and the accuracy and relevancy of LLM responses in federal statistical use cases.
12	What is the expected scale of the pilot testing (e.g., number of datasets, types of prompts)?	As stated in the RFS, the case study should be applied to both legacy and more recently published Commerce statistical products and datasets and should leverage domain-specific prompt-response pairs. We look to the vendor to propose the appropriate scale and design of pilot testing that best aligns with the objectives stated in the RFS.
13	Are there specific criteria for validating the tool's effectiveness during the pilot phase?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS.
14	Should the pilot include iterative testing to refine the evaluation framework and tool based on initial findings?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS.
15	Are there any budgetary limitations or guidelines that the offeror should consider while proposing the solution?	There is not a target level of funding. We look to the offerors to propose a cost that will meet the objectives stated in the RFS.
16	Are there specific deadlines for deliverables, or is the timeline flexible based on the offeror's proposed milestones?	Specific deliverables and timelines are outlined in the RFS. We look to the offerors to align their proposed approach with these requirements.

	Question	Answer
17	How does the government plan to measure the impact of the solution on AI-readiness and LLM performance?	This RFS envisions a solution that enables agencies to investigate the effectiveness of enriching their data assets in improving the accuracy and relevancy of LLM responses in federal statistical use cases. This will facilitate evidence-based and data-driven decision-making as the government takes the necessary steps to make its data AI-ready.
18	Are there long-term goals or future phases envisioned for this project that the offeror should consider?	This RFS envisions the development of a tool that may be offered as part of a shared service within a future National Secure Data Service (NSDS) and lay the groundwork for future expansion to other statistical subject-matter domains and agencies.
19	Should the solution prioritize improving data readiness over testing AI capabilities, or is equal emphasis required?	Please see the response to question #2.
20	In Attachment 1 – Project Topic, under "AI-Readiness Evaluation for Statistical Data" and "Testing the AI-Readiness Evaluation," could NCSES provide examples or a representative sample of the federal statistical products and data assets envisioned for the case study (e.g., microdata, metadata)?	As stated in the RFS, the case study should be applied to both legacy and more recently published Commerce statistical products and datasets.
21	Is there an expected or recommended budget range?	There is not a target level of funding. We would look to the offerors to propose a cost that will meet the objectives stated in the RFS.
22	Does NCSES have a preferred percentage or target level of involvement for non-traditional entities within the proposed project team?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS.
23	Is there any preference for being a member of ADC prior to submitting a proposal?	Membership with ADC is not required to submit a proposal and will not be considered as a part of the application evaluation; however, if chosen for award, membership is required.
24	The call mentions the availability of Commerce statistical data assets and using a curated collection of domain-specific prompt-response pairs. Will these data assets and curated prompt-response pairs be provided or will it be the responsibility of performers to collect them?	As noted in the RFS, the case study should be applied to Commerce public-facing statistical products and datasets. We look to the vendor to identify and develop prompt-response pairs that best meet the objectives stated in the RFS.

	Question	Answer
25	Is the call focused on evaluating LLMs themselves or more complex LLM-based systems (e.g., that use agentic workflows, data prepocessing, external knowledge collection)? Will there be set APIs that any system or LLM under evaluation is expected to adhere to?	As stated in the RFS, we look to the vendor to develop an empirical evaluation that measures the ability of LLMs to accurately respond to questions that require an understanding of federal statistical data assets.
26	What assumptions should be made about the inputs and outputs that the LLMs (or LLM-based systems) can support? For example, should it be assumed that all information should be provided as part of the prompt to the LLM (query, data, meta-data, etc.) or will assumed that all systems under evaluation support more complex interactions (e.g., providing data and metadata in advance, supporting a more formal input structure that takes in metadata and data, etc)?	We look to the vendor to propose the best solution to meet the objectives stated in the RFS.
27	Should this work focus on examining a broad range of LLMs/LLM-based systems (e.g., available from Google Model Garden, HuggingFace Hub, etc.) or are there a set of candidate LLMs that should be prioritized in early evaluations?	We look to the vendor to propose a set of candidate LLMs to evaluate that will best meet the objectives stated in the RFS.
28	Are there specific Commerce statistical data assets or domains that should be prioritized in the case study?	Please see the response to question #1.
29	Can you confirm that the expectation is to only include Commerce data within the 1-year period of performance?	The expectation is that the case study will focus on both legacy and more recently published Commerce data and be completed within the specified period of performance.
30	Is the expectation that the LLMs also be open source, or can/should leading proprietary models also be evaluated?	While open-source solutions are preferred, proposals may include non-open-source solutions with clearly specified benefits on why one solution was selected over another to meet the objectives stated in the RFS.