Attachment 1 – Project Topic

Measuring Large Language Model Understanding of Federal Statistical Data

Key Objective

Generative AI applications offer transformative opportunities for how Americans interact with public data. By enabling interaction through natural language and multimodal prompts, these technologies facilitate more intuitive access to complex data collections through chat-based interfaces, reducing technical barriers and expanding the accessibility of public data to a broader range of users. To ensure that federal data are increasingly valuable in the training of generative AI applications, the federal government must optimize and enrich its data assets with the appropriate context for this rapidly evolving ecosystem.

This Request for Solutions (RFS) seeks to develop an empirical evaluation that measures the ability of large language models (LLMs) to accurately respond to questions that require an understanding of federal statistical open Government data assets and their associated metadata.[1] This will involve the creation of prompt-response pairs necessary to assess the accuracy, relevancy, and explainability of LLMs in federal statistical use cases. In addition, this effort will result in a tool that will evaluate LLM performance in response to these evaluation prompts, while also providing insight into how well federal statistical data assets are structured to support LLM interaction – highlighting opportunities to improve metadata quality, accessibility, and machine-readability. Ultimately, 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 replication and expansion across additional statistical subject-matter domains and agencies.

Background

The federal government has made great strides in developing best practices for AI-ready data and in piloting approaches to test and improve the AI-readiness[2] of its publicly available statistical data assets. In January 2025, the U.S. Department of Commerce (Commerce) published *Generative AI and Open Data: Guidelines and Best Practices* for preparing and disseminating Commerce public data for use by generative AI. While designed for Commerce's use, the guidance has been made publicly available to empower open data publishers across the globe to optimize their data for generative AI systems.

[1] Open government data assets are defined in statute (44 USC 3502(20)) and in M-25-05. Solutions should consider both datasets and their accompanying contextual information (e.g., metadata, documentation, and formatting) as this information can influence how an LLM interprets the dataset.

[2] For the purposes of this project, "AI-readiness" refers to the extent to which a data asset is prepared for effective analysis and querying by AI systems, particularly LLMs. This includes data quality, access, formatting, and metadata. AI-ready data enable LLMs to generate meaningful insights from both structured (e.g., tabular) and unstructured (e.g., reports) formats by facilitating accurate interpretation and contextual understanding.

This guidance incorporates a critical element of the Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance that requires federal agencies to capture and store rich metadata about their data assets to enable effective internal usage, management, data cataloging, discovery, and interoperability. Specifically, it recommends federal agencies update their metadata to comply with the DCAT-US 3.0 standard.

Developing a tool that can conduct repeated evaluations of how well LLM responses interpret and use federal statistical data will enable agencies to investigate the effectiveness of enriching their data, metadata, and documentation 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 transitions to DCAT-US 3.0 and takes the necessary steps to make its data AI-ready.

In addition, the ongoing NSDS Demonstration project, "AI-Ready Data Products to Facilitate Discovery and Use" (AI-RD-24), aims to assess and improve the AI compatibility of federal statistical data by developing AI-readiness criteria and prototyping tools that transform public datasets into machine-interpretable formats. The Bureau of Economic Analysis, the Census Bureau, the National Center for Science and Engineering Statistics, and other federal statistical agencies have played a major role in this project.

Furthermore, the Federal Committee on Statistical Methodology (FCSM) has published a call to action to improve the AI-readiness of federal statistical data by enhancing APIs with metadata and context to improve LLM results. They are also assessing the use of emerging standards like the Model Context Protocol to increase system interoperability.

While these initiatives provide a strong foundation for improving AI-readiness, an important next step is to assess whether statistical data are being accurately captured and represented in responses generated by LLMs. This project aims to do just that—by developing an evaluation to assess the ability of LLMs to meet a high bar of statistical accuracy through the creation of prompt-response pairs that address a representative sample of the complex portfolio of federal statistical products and data assets, and by developing a tool that evaluates both the quality of LLM responses and the extent to which federal statistical data assets are structured to support effective LLM interaction.

The Current Request for Solutions

This RFS seeks to develop a tool that may be offered as part of a shared service within a future NSDS to help federal agencies gain insights on the AI-readiness of their statistical data and to optimize their statistical data assets for the rapidly evolving data ecosystem. This work should complement the current efforts to assess and implement strategies for AI readiness in the federal data ecosystem with a focus on responses generated by LLM. This RFS proposes a case study applied to Commerce public-facing statistical products and data assets to support this effort and lays the groundwork for extending its application to additional statistical subject-matter domains and agencies.

This case study involves two main parts:

Developing an AI-Readiness Evaluation for Statistical Data:
Using Commerce statistical data assets, develop an evaluation that can be used to assess an LLM's ability to meet a high bar of statistical accuracy, and that gauges the extent to which its responses effectively

interpret and use statistical data assets. The evaluation should leverage a curated collection of domain-specific prompt-response pairs against which LLM model-generated responses will be compared and evaluated. The evaluation should also assess how well federal statistical data assets are structured to support effective LLM interaction, identifying potential barriers related to metadata quality, accessibility, and data formatting.

Testing the AI-Readiness Evaluation on Statistical Data:
Using Commerce statistical data assets, prototype an open-source and publicly accessible tool that can automate and apply the AI-readiness evaluation. Use the prototype to conduct a pilot of the evaluation and measure LLMs' ability to accurately respond to evaluation prompts and demonstrate an understanding of statistical data assets. The pilot should include both older and more recently published statistical data assets to determine whether older publication standards create different challenges for LLM interaction compared to newer standards.

Information Gaps

The project will address critical information gaps in AI-readiness evaluation by:

- Advancing beyond theoretical frameworks by 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.
- Leveraging subject-matter expertise to inform the development of prompt-response pairs, ensuring that the evaluation reflects real-world analytical demands and domain relevance.
- Establishing quantifiable performance metrics to evaluate the quality of LLM responses when engaging with federal statistical data.
- Identifying best practices for scaling this approach to different statistical agencies as part of a shared service within a future NSDS.

Key Evidence-Building Considerations

Key research questions include:

- How do LLMs perform when a user submits a prompt that requires subject-matter expertise of federal statistical data to generate an accurate response?
- How do responses generated by LLMs compare to those produced by subject-matter experts in accessing, analyzing, and interpreting federal statistical data? What differences emerge in their reasoning processes, accuracy, and interpretations?
- What is the relationship between the quality of federal statistical data, metadata, and documentation and the accuracy and relevancy of LLM responses in federal statistical use cases?

Deliverables

At a minimum, offerors will provide the following if selected for an award. Proposers should outline the additional deliverables they will provide in the provision of this solution.

Project Management:

- Biweekly coordination meetings to establish goals, assess progress, and ensure alignment with project objectives.
- Monthly status reports towards project objectives delivered in an agreed-upon format.
- Quarterly lessons learned based on what has been learned during the last quarter that will inform a future NSDS.

Draft/prototype and final versions of the following deliverables:

- Project plan outlining key milestones and timelines.
- A plan, grounded in best practices, to create a collection of prompt-response pairs through engagement with subject-matter experts and data stewards (e.g., RFI, facilitated workshop).
- A collection of domain-specific prompt-response pairs used to evaluate an LLM's interactions with Commerce statistical data assets.
- An AI-readiness evaluation that measures how effectively an LLM interacts with Commerce statistical data assets and provides insights into the extent to which their structure supports meaningful and accurate LLM interactions.
- Performance metrics that assess the LLM against the evaluation prompts.
- An open-source and publicly accessible tool that will conduct the AI-readiness evaluation and provide performance metrics in a scalable and replicable manner.
- All code and documentation supporting the project suitable for inclusion in a publicly accessible open-source software repository and provided under an open-license or public-domain equivalent; including all code used in the project provided in a reproducible open-source format (R, Python, etc.). All technical components should propose using open-source software and open-source software development principles wherever feasible.
- Technical documentation on the methodology, evaluation, and tool that explains their capabilities and limitations, and enables maintenance and expansion to additional subject-matter domains and data assets.
- A final report on the project, including the process, lessons learned, outcomes, and recommendations.