Measuring LLM Understanding of Federal Statistical Data – Outreach Plan

Purpose

As part of the National Secure Data Service Demonstration (NSDS-D) project, the National Center for Science and Engineering Statistics (NCSES) has contracted with NORC at the University of Chicago (NORC) 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 data assets and their associated metadata. A major part of this work will be creating a collection of prompt-response pairs which will be used to evaluate LLMs' interaction with Commerce statistical data assets. We will engage with data users and leverage publicly available resources on data use to ensure the prompts are user-centered and based on real-world data use.

Timeline

The following timeline summarizes the sequencing of the three main Evaluation Plan activities. Dates reflect the current work plan.

Identification of Commerce Statistical Data Assets. The recommended data assets are included in this Draft Evaluation Plan. Final asset selections will be included in the Final Evaluation Plan on December 23.

Prompt-Response Development. The prompts will be drafted and included in the Final Evaluation Plan on December 23. The corresponding responses will be developed in early 2026. The complete and final prompt-response pairs will be completed no later than March 24, 2026. **This is the phase of the project for which we will conduct outreach to federal data users.**

Evaluation Plan Criteria. Complete evaluation plan criteria, including evaluation of prompt-response performance, AI-readiness evaluation of data assets, and LLM selection considerations will be included in the Final Evaluation Plan on December 23.

Prompt-Response Development Phase

NORC and the Massive Data Institute at Georgetown University (MDI) will collaboratively develop a user-centered collection of domain-specific prompt-response pairs to evaluate LLM performance in real-world data use scenarios. These prompts will be tailored to three categories of data users in **Table 1** and span a range of analytical complexities and question types, including data discovery, access, retrieval, analysis, variable definitions, and data linking. NORC and MDI will review information from archive sources including the ACS Data Users Group[1], and tailored searches for relevant Census or BEA topics or data assets within the r/DataScience[2] and r/DataIsBeautiful[3] Reddit communities to draft and refine prompts and responses that are clear, realistic, and statistically valid. The prompts will be designed to assess LLM

[1] The ACS DUG is no longer active, but archived discussions are available using the Internet Archive's Wayback Machine through 09/2025 (https://web.archive.org/web/20250201225457/https://acsdatacommunity.prb.org/)
[2] https://www.reddit.com/r/datascience/
[3] https://www.reddit.com/r/dataisbeautiful/

performance across tasks such as factual retrieval, analytical reasoning, metadata comprehension, time sensitivity, and data usability.

Some prompts will make use of multiple topics and data assets, recognizing that many user queries are not confined to single domains nor single data assets. Each persona may require different levels of prompt complexity, since both experienced and casual users can approach the same task in very different ways. Variation can also occur within the same persona type, depending on context, goals, or familiarity with the domain. The final collection will include multiple prompt variants per user type and question category, grounded in authentic use cases and authoritative data interpretations.

Table 1. Categories of user personas by level of data experience

	Experienced Data Users	**Intermediate Data Users**	**Casual Data Users**
Examples	Academic researcher, data scientist, demographer, economist, librarian, statistician.	Business analyst, Congressional staffer, journalist, state/local government staff.	Community advocates, social media influencers, students, engaged members of the public.
Task	Complex analytic need from a specific or multiple data asset(s) (data use), potentially requiring detailed metadata.	Basic to medium complexity need from a specific data asset (data access and use) with limited availability to navigate metadata.	Basic analytic need from an unknown data asset (data discovery, access, and subsequent use) without background to navigate metadata.
Technology	APIs, GIS, code, databases spreadsheets.	GIS, spreadsheets, some code, web search.	Web search, AI chatbots (e.g., Microsoft CoPilot, ChatGPT)
Experience	Much experience in subject matter and/or data analysis. High to low Commerce data experience.	Moderate experience in subject matter and/or data analysis. Moderate to low Commerce data experience.	Low experience in subject matter and data analysis. Low Commerce data experience.
Background	For university researcher/ demographer/economist: Will need more metadata than the current system serves and needs consistent metadata for every dataset and every place. For data scientist: Subject agnostic/may have little domain expertise. Needs access to deep methodology detail.	Understands basic survey concepts but does not have time to read complex methodology documentation. Works in fast-paced environment where answers must be accurate, but also quick.	Do not know what they do not know. May use either more or less sophisticated prompts.

Measuring LLM Understanding of Federal Statistical Data – Outreach Plan

Prompt Input from Federal Data Users

Before the team proceeds with any user engagement, we are submitting our plans to Commerce for review and approval. Specifically, this document includes a general description of the work along with the specific questions we plan to ask users. If approved, MDI will facilitate an online discussion with data users on the Federal Data Forum (FDF). Our proposed questions are outlined below. MDI will then synthesize forum members' input to help inform the final prompts.

Questions for the Federal Data Forum

Detailed below is our planned sequencing and outreach language we will use to elicit data user feedback:

Share Your Experience: Using AI to Access Federal Statistics
The National Center for Science and Engineering Statistics (NCSES) is working with NORC at the University of Chicago and the Georgetown University Massive Data Institute on a project to understand how large language models (LLMs) handle questions about federal statistical data. Your feedback will help us evaluate the readiness of federal data for LLMs and improve how these tools serve data users like you.

Why your input matters:
Your input will help us evaluate LLM performance and identify ways to make these tools more accurate and useful for federal data users.

How you can help:
Please share your experiences by answering any of these questions:

1. What kinds of prompts do you use with LLMs to ask questions about federal data? Tell us about what goes into crafting your prompts for LLMs.
2. Can you share examples of prompts you've tried? Did you find the responses accurate or helpful? If you have any specific prompts related to Census Bureau or Bureau of Economic Analysis data, please share those.
3. Have you run into any challenges when using LLMs for federal data questions?

Follow-Up and Clarifying Questions

1. Tell us more about that.
2. Could you provide more detail about that?
3. What else?
4. What do you mean by . . . [user response]?
5. Could you provide an example of a prompt that you have used (or would use) for that?