CENSUS OPEN INNOVATION LABS

THE OPPORTUNITY PROJECT

2025 SPRINT TOPIC

Preparing Federal Open Data for the Next Generation of AI Users

U.S. Department of Commerce | U.S. Census Bureau

The Challenge

Improve the ability of Large Language Models (LLMs) to serve accurate and reliable information involving federal open data, while making this data more easily ingestible by LLMs.

Executive Champion

George Cook, performing the non-exclusive functions and duties of the Director, U.S. Census Bureau

The Problem

Large Language Models (LLMs), a type of generative artificial intelligence (GenAI) that can process and *generate* content, are increasingly being used to replace traditional web browsing and search experiences in favor of AI-produced summaries. This shift will have profound consequences on the dissemination, discovery, and use of federal statistics, and raises questions about accuracy, potential biases, and the authenticity of the source information. LLMs amplify known problems that occur when intermediaries repackage federal open data for use by the public, as moderately autonomous AI are known to hallucinate, misinterpret information, provide biased responses, and combine various sources of information to create blended data products from unofficial sources. To protect data quality, new methods are needed to ensure the accuracy and integrity of federal data in these systems.

A newly developed and widely adopted open standard called Model Context Protocol (MCP) helps to address this issue. Using this protocol, data owners can provide open-source tools that promote more accurate and efficient interactions with their data. MCP Servers act as intermediaries between data sources and the LLM client, creating opportunities to optimize responses for use in the LLM's "context window" – the amount of information that an LLM can process or consider within a conversation.

The Opportunity

In this sprint, we invite LLM/Gen AI developers and data users to help improve tools' accuracy, and the underlying data's usability by LLMs, and as applicable to certain tech teams, to test a beta version of the U.S. Census Bureau's MCP Server. Through the sprint, we will explore questions such as:

- Who are the users asking questions of LLMs that could benefit from our data? How can our data be served in a way that supports these interactions?
- What can we do to maximize the quality and accuracy of our content in these systems?

- What data exists outside of our API boundary that would be essential for use in these systems?

We challenge sprint teams to create new or improved tools, including those that integrate the U.S. Census Bureau's MCP Server into their AI Assistant technology stack, testing how the MCP Server can be used in diverse use cases with clear benefits to end users. We encourage the participation of sprint teams with existing LLMs, who could use the sprint as an opportunity to improve and build on their current tools.

Target End Users

LLM developers and their downstream users making inquiries about federal data

Related Data Sets

- Model Context Protocol
- Census Data API User Guide, U.S. Census Bureau
- Available APIs, U.S. Census Bureau
- Generative Artificial Intelligence and Open Data: Guidelines and Best Practices, U.S. Department of Commerce
- Further data sets to be defined by Census Bureau and other DOC bureaus such as NOAA

Sprint Leaders

- Luke Keller, Chief Innovation Officer & Bureau AI Lead, Innovation Strategy Office, U.S. Census Bureau
- Bella Mendoza, U.S. Digital Corps Data Science & Analytics Fellow, Office of the Under Secretary for Economic Affairs, U.S. Department of Commerce
- Zach Palmer, U.S. Digital Corps Data Science & Analytics Fellow, Office of the Under Secretary for Economic Affairs, U.S. Department of Commerce