Improving LLM Access to Federal Open Data

A Pilot Study of Model Context Protocol Servers for Federal Data Access

Bella Mendoza
Mark Aronson, PhD
Haley Johnson
Sam Levy
Mohammad Arifur Rahman

UNITED STATES
DIGITAL
CORPS

U.S. Digital Corps
Technology Transformation Services
Federal Acquisition Service
General Services Administration

About the U.S. Digital Corps

The U.S. Digital Corps was launched in August 2021 by the General Services Administration in collaboration with the White House Office of Management and Budget, the Office of Personnel Management, the Cybersecurity and Infrastructure Security Agency, and the White House Office of Science and Technology Policy. It is a cross-government fellowship opportunity operated by the GSA's Technology Transformation Services (TTS).

The idea for a Digital Corps was sparked by technologists across government who identified a gap in the federal government's journey towards digital success—a lack of early-career technology talent. TTS recognized the need for entry-level technologists to not only bring immediate innovation but also to serve as a continuing resource for government digital transformation. These technologists could complement the efforts already underway by technologists across government—bringing in fresh skills and perspectives as well as sustaining support over time. Internally growing the federal government technology leaders of the future. What started as an idea and grassroots effort evolved into the U.S. Digital Corps.

About this Document

Improving LLM Access to Federal Open Data: A Pilot Study for Use of Model Context Protocol Servers for Federal Data Access reports findings of a Model Context Protocol (MCP) experiment for access to federal data and provides insights and recommendations for implementing MCP for federal data interactions. This report will help agencies build MCPs, resulting in improved federal data retrieval with Large Language Models (LLMs).

Disclaimer

Mention of or referral to any product, service, individual, organization or other enterprise in this report, including citations or links to non-government sites, is not and does not imply official government endorsement of those entities. The opinions and ideas of non-government entities are theirs alone.

Contact

Federal employees or contractors with questions may reach out to mcp@gsa.gov.

Table of Contents

Executive Summary

Federal data drives critical societal functions, from supporting scientific research to economic policy. Today, individuals and organizations increasingly depend on AI systems to find, summarize, and interpret data, creating a new imperative for agencies to make their data AI-accessible. If agencies do not modernize, users making the wide range of decisions informed by federal data risk relying on outdated or non-authoritative sources.

We evaluated large language models' ability to retrieve and interpret federal data using two representative datasets: CDC PLACES (statistical health data) and USAspending (federal spending data). After discovering significant error rates in data retrieval, we tested Model Context Protocol (MCP), an open-source standard enabling structured API interactions for LLMs, as a solution.

Key Findings

- **Problem Discovered: Severe baseline inaccuracy.** ChatGPT and Gemini achieved 0% accuracy for USAspending and only 2.1% for CDC PLACES on data retrieval tasks.
- **Root Cause: Retrieval is a bottleneck.** When provided with simulated API responses, LLMs answered correctly nearly 100% of the time, confirming data retrieval as the primary barrier.
- **MCPs work.** Our MCP servers improved LLM accuracy from 18% to 95%, demonstrating a practical path for AI-accessible government data.

What This Means for Agencies

- **The problem**. AI agents struggle to interact with federal data systems designed for people, hindering LLM use of federal data while also preventing agencies from effectively testing the AI readiness of their data.
- **The risk**. Without modernization, AI systems may continue to turn to unofficial sources, reducing federal data's authoritative value.
- **The opportunity**. MCP servers offer an approach for agencies to make existing APIs accessible, ensuring federal data remains authoritative as AI adoption accelerates and provides agencies with a mechanism for testing the effectiveness of AI-ready data enhancements.

Policy and Governance Considerations

- **Develop MCP servers to expand public value of existing datasets.** MCPs lower barriers to API access, broadening use. Federal MCPs would complement third-party innovation by ensuring that authoritative sources remain available alongside community tools.
- **Centralize discovery through a federal MCP registry as a "front door" to federal data.** Establishing such a registry would also strengthen safety and public trust in how government data is accessed and used.
- **Ensure federal stewardship.** Data stewards and API developers are essential to building MCPs. To maximize transparency and adoption, agencies should open-source their MCP implementations.
- **Prioritize security and accountability.** Agencies should establish governance structures to oversee MCP deployment and monitor compliance.

This pilot proves MCPs can fundamentally transform federal data access by infusing LLM interactions with expert-level guidance, enabling seamless discovery and aggregation across sources. While deployment and security challenges require careful planning, MCPs represent essential infrastructure for AI-ready federal data. Federal adoption of MCP can dramatically expand public access to government information while enabling agencies to leverage data for real-time decision-making, converting information silos into a unified strategic asset delivering more effective evidence-based outcomes for the public.

A. Introduction and Key Findings

AI Adoption Presents an Opportunity to Improve Federal Data Interactions

The adoption of AI systems is transforming how people access information online[1], leading to new requirements for data accessibility. Instead of machine-readable formats and traditional APIs, data providers must now develop machine-understandable documentation and interfaces that AI agents can navigate and interpret, as we'll show in this report. The shift to these new standards is particularly urgent for datasets which represent an authoritative source of information, as is often the case for data from federal agencies.

This need for AI-ready public data has been formally recognized in several recent federal initiatives. In January 2025, the U.S. Department of Commerce published *Generative AI and Open Data: Guidelines and Best Practices*[2], which highlights the importance of moving beyond traditional human-centered formats and documentation to structure open data in ways that AI systems can interpret and navigate. In alignment, the Federal Committee on Statistical Methodology (FCSM) issued *AI-Ready Federal Statistical Data: An Extension of Communicating Data Quality*[3], which called on agencies to modernize data access to "support accurate and trusted generative AI results." To achieve this, FCSM recommended agencies explore MCPs as one possible approach for improving machine understandability. In responding to this call, we recognized not only an opportunity for the federal government to take a leading role in defining the technologies associated with AI-ready data, but to improve the use of federal data through agentic interactions with MCP servers.

Model Context Protocol: An Emerging Standard for AI Interactions

The Model Context Protocol (MCP) is an open-source framework for integration and data sharing between LLMs and external tools and applications. Developed by Anthropic in November 2024[4], the MCP has since been adopted by almost all major AI companies[5], including OpenAI (ChatGPT), Microsoft (Copilot), Meta (Llama), and Google (Gemini). A key value of MCP is that it is an open standard. This offers greater stability, lowers barriers to entry,

---

[1] https://www.nngroup.com/articles/ai-changing-search-behaviors/

[2] Commerce Data Governance Board, "Generative Artificial Intelligence and Open Data: Guidelines and Best Practices," *U.S. Department of Commerce Blog*, January 16, 2025, https://www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices.

[3] Federal Committee on Statistical Methodology (FCSM), *AI-Ready Federal Statistical Data: An Extension of Communicating Data Quality* (Hoppe et al., 2025). https://www.statspolicy.gov/assets/fcsm/files/docs/FCSM.25.03_AI-Ready-Extension-Data-Quality.pdf

[4] "Introducing the Model Context Protocol," *Anthropic*, accessed August 19, 2025, https://www.anthropic.com/news/model-context-protocol.

[5] *Enterprise AI Adoption Rapidly Evolves with Anthropic's MCP*. Accessed 7/20/2025. https://mlq.ai/news/enterprise-ai-adoption-rapidly-evolves-with-anthropics-mcp/

reduces the cost of development and maintenance, and makes adoption easier. Being open means more people benefit, strengthening both accessibility and innovation.

The MCP defines a protocol for communication between a client and server (Fig 1): the AI model communicates with the MCP client, which then communicates with the MCP server which is made up of a suite of “tools” that support discrete tasks. These are discrete functions, where a simple example is a tool that functions as a wrapper around a single API endpoint. The server provides context to the model on when and how to use each tool, so that when prompted, the MCP client communicates with the MCP server to determine if it has any relevant tools and if any information provided in the prompt should be passed to the tool. Think of MCP like a restaurant: the AI model is the customer, the MCP client is the waiter taking the order, and the MCP server is the kitchen with different stations. Each tool is like a station that handles a specific dish. The customer places an order, the waiter passes it to the right station, and the finished dish comes back through the waiter to the customer.

Using our implementation makes these concepts more concrete.[6] We used Claude Desktop for our MCP server testing because Claude Desktop, the application, both interacts with an AI model (in our case, Claude Sonnet 4) and has a built-in MCP client. This client (Claude Desktop) then interacts with an MCP server we programmed in Python and ran locally on the desktop. The communication between these two parts, the AI model (through the MCP client) and the MCP server, is standardized by the Model Context Protocol. The MCP server then interacts with the dataset’s Application Programming Interface (API) through native API protocols, at which point this system functions like a standard API call. The API queries then query the dataset and the API response is returned to the server.

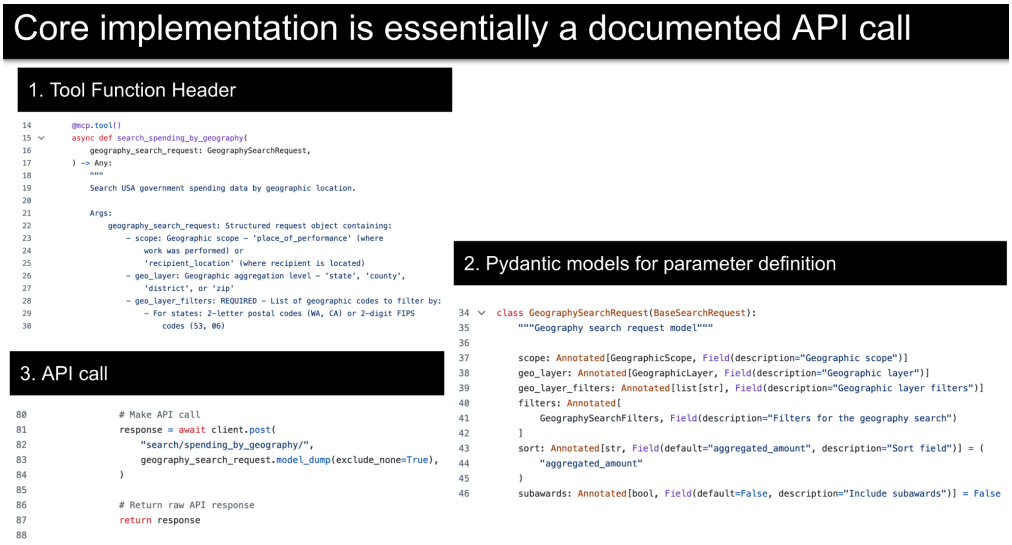

Figure 1 | Code snippets from MCP server for USAspending. The MCP server is constructed with Python code to execute API calls.

[6]Our proof-of-concept MCP servers can be found here: GitHub - GSA-TTS/usa-spending-mcp-server-DEMO: Test MCP Server for USA Spending API, GitHub - GSA-TTS/cdc-places-mcp-server: Pilot MCP for the CDC PLACES dataset

Our team was particularly interested in MCPs' potential for API integrations that provide an intuitive, natural language interface for data retrieval. An example flow would be:

1. User asks a question for which the dataset contains a correct, specific answer

   For CDC PLACES, this could be a question like "What percent of adults went to the dentist in Worcester County, MA in 2018?"
2. The AI model checks to see if there are any MCP server tools that could be used to answer the question

   A tool named "get_cdc_places_data" with the description "fetches data from the CDC PLACES dataset"
3. After finding a potential tool, the AI Model, through the MCP Client, looks to see what parameters the tool needs to run

   For "get_cdc_places_data," this is the year (2018), the measure ID (DENTIST), and the geographic granularity (county)
4. The model then passes those parameters to the MCP server and the server uses them to form an API query
5. The API returns the response to the MCP server, which passes it back to the AI model
6. The AI Model then interprets the response and provides an answer to the user in the chat window

The MCP client-server relationship abstracts the work of identifying the correct API, specifying query parameters, and making a call, simplifying data access and retrieval.

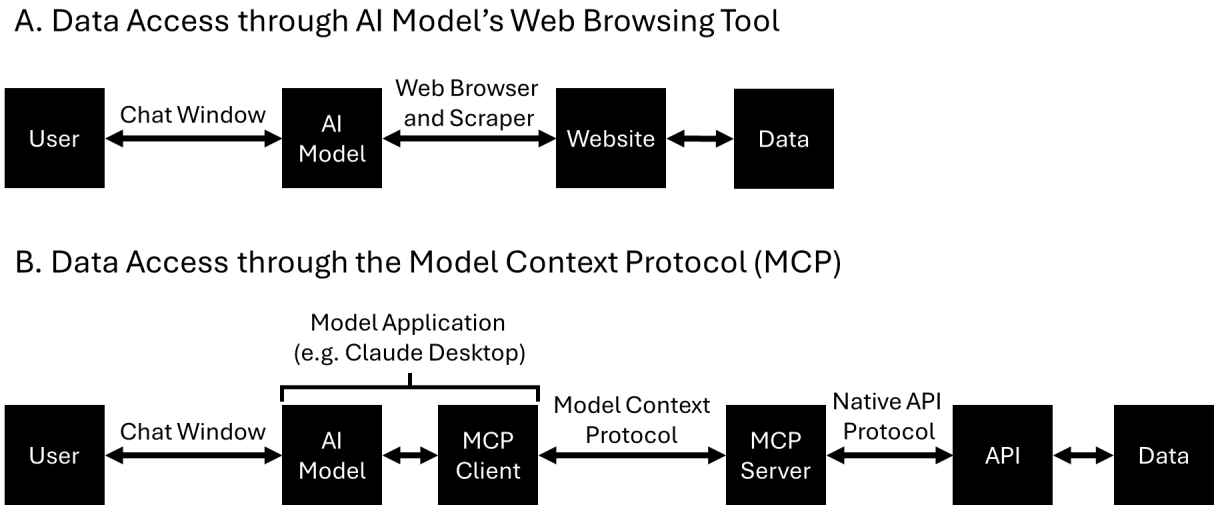

Figure 2 | Schematic of two methods of data retrieval using an AI model | A. Data access using a web browser and scraper. Without additional tools, AI models can take a user's query and form a web search and scrape the text found on those websites to inform the response. In the case of data retrieval, the models will search websites containing the data in an attempt to scrape the data into the context window. **B.** Data access using the Model Context Protocol (MCP) for querying a dataset using an API. In this setup, the user (left) communicates with an AI Model through its chat interface. The AI model in turn, communicates with an MCP client, which uses the MCP to communicate with an MCP server. The server informs the client (and thus the model) which tools it has, what their purposes are, and what parameters are needed to use them. The MCP server, in turn, uses the native API protocol (such as HTTP requests or SQL queries) to communicate with the API, which queries data from the dataset.

MCP Implementation Vastly Improves Models' Data Retrieval Capabilities

Following the Federal Committee on Statistical Methodology's recommendations, we identified candidate datasets for a baseline assessment of LLMs' abilities to retrieve correct data across a variety of federal data products. We selected CDC PLACES, a straightforward dataset hosted on Socrata, an open data platform widely used across government. By using this common platform as a test case, successful MCP server implementation demonstrated scalability to other datasets on the same platform[7].

To test performance in a more complex data environment, we also experimented with USAspending[8]. USAspending aggregates spending data from over 400 data elements across the federal government, ranging from major department financial systems to smaller agency databases, each with different update cycles, data quality standards, and categorical frameworks.

To perform our baseline assessment of the LLMs' abilities to retrieve and interpret USAspending and CDC PLACES data, we prompted different AI systems with a panel of questions for which the data products contained specific, numeric answers. Starting with simple queries, we found that the models were largely incapable of extracting the correct data value (5% accuracy for ChatGPT and 0% accuracy for Gemini across ten trials for each question). Prompt engineering alone did not improve data recall (only boosting performance to 4.2% accuracy for ChatGPT), nor could the models form API queries to retrieve the desired data values (creating valid API queries only 17% of the time).

These initial results indicated that the models struggled with data retrieval. To isolate data interpretation abilities, we fed simulated API responses and found the models identified the correct value with 100% accuracy. This performance gap between data retrieval and data interpretation revealed a fundamental limitation: **while LLMs can accurately interpret structured data, they struggle to locate and extract it from real-world sources.** We suspect this is due to a number of factors (see *JavaScript Content Loading Issues*). Many federal websites use JavaScript to dynamically deliver content to users, but these sites are expensive to scrape and typically excluded from LLM's training data. Furthermore, data and API documentation are often split across multiple websites, such as in the case of CDC PLACES. This can make it difficult for LLMs to connect all the relevant details needed to correctly query, aggregate, and interpret the data. These insights pointed us toward MCP servers as a potential solution. They directly address this limitation by guiding LLMs to translate natural language prompts into valid, context-aware API queries, improving LLM access to federal statistical data. Implementing pilot MCP servers for each dataset achieved 95% data retrieval accuracy across both data products, confirming this hypothesis.

---

[7] See the following more examples of federal APIs that are hosted on Socrata: https://dev.socrata.com/foundry/data.transportation.gov/anj8-k6f5 (Department of Transportation); https://dev.socrata.com/foundry/www.datahub.va.gov/5uqy-ph6a (Department of Veterans Affairs); https://dev.socrata.com/foundry/data.permits.performance.gov/mcm3-xbid (President's Management Agenda)

[8] U.S. Department of the Treasury, Bureau of the Fiscal Service, "USAspending.gov," accessed August 19, 2025, https://www.usaspending.gov/.

Our results indicate the enormous potential of MCPs to make federal data more accessible. With MCP deployment, agencies can enable accurate, reliable machine interpretation and unlock more nuanced interactions with federal data for academic research, policy development, and program evaluation.

B. Recommendations for Leadership

I. Use MCPs as Guardrails for LLM-Federal Data Interactions Already Occurring

> *Federal agencies should invest in MCPs to prevent data misuse and ensure existing federal data is usable in an increasingly AI-powered workforce and public.*

Our results show that, while LLMs have some knowledge of federal data sources, they are generally incapable of correctly querying data on their own. In the case of CDC PLACES, LLMs could answer methodological questions but were unable to accurately interact with the underlying data. This gives the LLM a veneer of authority — a user who is unfamiliar with CDC PLACES could easily check the sources the LLM used, see they linked to PLACES' website, conclude the LLM is an authoritative source, and proceed to use data that the LLM had incorrectly queried, or in many cases, falsely attributed to PLACES. In August 2025 alone, Federal websites had 12 million referrals from ChatGPT, underscoring the scale of this issue[9].

Some data owners may argue that their data is not designed for LLM use and that investing in AI readiness and MCPs is unnecessary and would only encourage improper use. We share many of the same concerns about how LLMs may decontextualize or misrepresent federal data sources. However, given that federal data already exist in models' training corpora and the increasing prevalence of data retrieval using LLMs, it is clear that we cannot prevent members of the public from seeking federal data in this way. MCPs offer a simple safeguard: they reduce risks of misinterpretation, ensure proper attribution, and make federal data more accessible and authoritative.

Beyond allowing agencies to keep up with emerging technology, MCPs also have the ability to facilitate new kinds of interactions with federal data. APIs are one of the fastest, most reliable ways to access real-time federal data but require a high level of technical expertise. MCPs can "democratize" API access by providing an intuitive, natural language interface that translates user requests into actual API calls, retrieving live data directly from federal sources.

Agencies do not publish data merely for the sake of release. They have a prerogative to ensure their data effectively reaches and serves the needs of interested members of the public to promote transparency, and to unlock the economic potential presented by federally-funded data collections. MCPs are a powerful tool to support this goal. Agencies invested in openness, transparency, and data sharing would be remiss not to explore this emerging technology.

[9] Data accessed on September 5, 2025. See analytics.usa.gov for up-to-date information.

II. Federal Stewardship Is Key to Effective MCP Development

The federal government delivers an extraordinary range of data to the public — from climate projections and demographic statistics to economic indicators and public health trends. These datasets underpin critical decisions across the nation in sectors such as scientific research, public health, economic development, and infrastructure planning. Therefore, the federal government must develop MCPs in a way that preserves the integrity, security, and usability of this valuable data, leveraging agencies' deep domain expertise, established governance frameworks, and commitment to serving the public.

Public Confidence Requires Federal Ownership

> *Federal agencies should develop authoritative MCP servers for federal datasets, release them openly, and establish processes for external contributions that strengthen, rather than fragment, the quality of MCPs.*

Today, unofficial MCP servers are created by third parties and used by the public for many federal datasets. There are multiple unofficial versions of Census MCP servers in circulation[10], as well as a USAspending MCP server built without agency involvement[11]. These third-party versions lack the deep expertise of the federal data and API team, and they risk embedding errors and omitting critical context. To safeguard accuracy and public trust, MCPs should be developed and governed by federal agencies as an authoritative source. Further, public-private partnership for the development of MCPs is critical for ensuring a successful adoption of this new technology.

Open Source Strengthens Stewardship

> *Federal teams should develop MCPs as open-source projects to take advantage of community interest and better serve users' needs while still providing an authoritative, vetted source.*

Providing MCPs as open code promotes transparency, allows the public to see how data is represented, enables users to contribute improvements, and establishes federal expertise and oversight.

Open source also taps into the knowledge of the broader user community. While federal stewards have the most expertise in methodology and system design, users often provide valuable insights about common use-cases and shortcomings of existing tools. Open-source MCPs ensure federal teams remain firmly in charge of direction and standards while inviting meaningful collaboration with users. By leveraging open source communities, federal teams can ensure an authoritative, vetted option is available while also giving members of the public the

[10] GitHub - TEKIMAX/mcp_census_server., GitHub - aaronbrezel/mcp-census: Census API access through MCP server

[11] flothjl, *usaspending-mcp* [Source code], GitHub, accessed August 28, 2025, https://github.com/flothjl/usaspending-mcp.

opportunity to adapt MCPs to meet their specific needs and use-cases. Deep user-engagement can create a synergistic relationship, where community-driven suggestions help federal teams better serve the public's needs and develop MCPs that deliver real value.

Open source development is already standard practice across the federal government.[12] As of 2016, more than two-thirds of Cabinet-level agencies maintained public repositories including the Departments of Agriculture, Commerce, Education, Energy (including national labs such as the National Renewable Energy Laboratory), Health and Human Services, Homeland Security (including National Cybersecurity Assessment and Technical Services), Interior (including bureaus such as United States Geological Survey and the National Parks Service), Justice, Treasury, Veterans Affairs, War (plus combat support agencies like the National Geospatial-Intelligence Agency and the Army Corps of Engineers), the Environmental Protection Agency, and the General Services Administration.[13] Within the Department of Health and Human Services, the Centers for Disease Control and Prevention[14] and the Centers for Medicare and Medicaid Services[15] both maintain several open-source projects and can serve as a model for how other agencies may adopt these practices.

Data Stewards Must Be Involved in MCP Design

> *Treat data stewards as core collaborators whose knowledge preserves the nuance and intent of federal data in AI use.*

Despite their public importance, federal data products are rarely simple. They reflect decades of shifting priorities, resource tradeoffs, and methodological choices, much of which is often only fully understood by data stewards.

Data stewards possess deep knowledge of the idiosyncrasies of each dataset: why certain variables are coded in a particular way, how gaps in coverage arose, or which derived indicators are commonly misinterpreted by users. Much of this contextual knowledge lives outside of official metadata or documentation, which itself is often fragmented across multiple locations, filled with technical jargon, and organized around system logic rather than user needs. Even the most comprehensive data dictionaries can fail to convey the subtle but important nuances that arise from long-standing institutional knowledge, informal reasoning, legacy constraints, and design trade-offs that shape how data is structured and used.

For example, CDC PLACES has, for most measures, a two year lag between the year the survey data is collected and when it publishes the small-area estimations. The reason for this is it is reliant upon Behavioral Risk Factor Surveillance Survey (BRFSS) data and Census

---

[12] Office of Management and Budget, Memorandum M-16-21, *Federal Source Code Policy: Achieving Efficiency, Transparency, and Innovation through Reusable and Open Source Software*, August 8, 2016, White House, available at https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2016/m_16_21.pdf

[13] *Facts about publishing open source code in government*, August 8, 2016, https://preserved.org.uk/18f.gsa.gov/2016/08/08/facts-about-publishing-open-source-code-in-government/index.html

[14] See https://github.com/CDCgov

[15] See https://www.cms.gov/digital-service/open-source-program-office and https://github.com/DSACMS

American Community Survey (ACS). Census ACS data become available roughly one year after BRFSS and then must be imported with complex small-area estimation to include estimates at the county level. Traditional API access may not note this complexity, but this is a fast-moving process when data does become available.

The API documentation, which is not on an official government website and instead lives on a third-party page, notes that "the 2023 release uses 2021 BRFSS data for 29 measures[16]." This is not recorded on the methodology page or dataset FAQ on the CDC website. Likewise, the API documentation does not make it explicit that *all release years* have a two year lag with the underlying data and this is not just a feature of the 2023 data. This nuance, which is important for understanding prevalence metrics and correctly selecting API endpoints, is only apparent after deep engagement with PLACES data and would not be apparent to new users or, as our results suggest, LLMs.

These problems are not unique to CDC PLACES and are indicative of the kinds of challenges many users face navigating federal data sources — information scattered across multiple sources, key details buried deep in the documentation, and important nuances left implicit which may imply slow moving data development processes. Federal data systems are only able to move as fast as the availability of data from dependent data systems.

Improving documentation remains important for human use as analysts and researchers still need clear comprehensive guidance for critical applications, but we also need the experts who hold this institutional knowledge to be involved with MCP server design. Without direct involvement from data stewards, that critical context risks being misrepresented, oversimplified, or lost entirely. For MCPs to meaningfully improve machine understanding of federal data, data stewards must be deeply involved in the design process, not as reviewers of the end products, but as core collaborators from the start.

Data stewards have a deep understanding of how their users actually engage with federal data in practice. By managing help desks, support forms, and feedback inboxes, they maintain a frontline view of real-world data needs and pain points. This perspective highlights real-world needs and recurring misunderstandings, and incorporating this experience into MCP server design can help ensure MCP servers are built and evaluated to serve users effectively. Without this grounding, there's a risk that MCPs may technically function, but fail to meet the needs of actual users or even perpetuate known patterns of misinterpretation. Our experience with USAspending reinforced this. Working directly with data stewards, API developers, and help desk managers allowed us to build a prototype that both avoided common errors and reflected genuine user experience.

API teams must lead technical design of MCPs

> *Federal agencies should empower their API teams to lead MCP development, ensuring MCP servers reflect technical realities and can evolve alongside federal data systems.*

[16] https://dev.socrata.com/foundry/data.cdc.gov/9umn-c3jf. Accessed September 22, 2025

Developing MCP servers involves encoding the logic that determines when and how to call different API endpoints. API developers, with the most intimate knowledge of these interfaces, data structures, and system behaviors, are well-positioned to build and maintain MCPs effectively. Their expertise can orchestrate retrieval accurately, scale effectively, and adapt through new endpoints when necessary. API teams are well-positioned to lead MCP development, though it does require adopting new frameworks and design patterns.

As outlined in our design recommendations (see *MCP Design Recommendations)*, successful MCP development involves leveraging established frameworks like FastMCP,[17] implementing extensive tool documentation, and understanding how to structure multi-call workflows — skills that build naturally on API developers' existing technical foundation. By empowering API teams to lead MCP development, agencies can ensure these critical tools evolve alongside their data infrastructure while maintaining the technical excellence that federal data users require.

III. Establish MCP Governance Frameworks for Federal Deployment

> *Establish comprehensive governance frameworks, security protocols, evaluation sets, and testing standards before widespread MCP deployment to address known risks and accountability gaps.*

While MCPs offer powerful benefits for enabling LLMs to access external data and tools, organizations must also account for security risks, accuracy challenges, and governance questions before deployment.

Security Risks of MCP Servers

MCP servers expand the attack surface by introducing new ways for models to use external tools. If misconfigured or poorly implemented, these servers can expose sensitive data or enable malicious behavior. Documented risks include:

- **Tool invocation exploits**: Unauthorized access through backend APIs triggered by improperly validated tool calls.[18]
- **Tool shadowing**: Attackers create servers with identical or similar tool names/descriptions to redirect tool calls to their own servers.[19]

[17] jlowin. *fastmcp*. GitHub repository. Apache-2.0 license. Available: https://github.com/jlowin/fastmcp (github.com)
[18] "MCP Security Notification: Tool Poisoning Attacks." 2025. April 1, 2025. https://invariantlabs.ai/blog/mcp-security-notification-tool-poisoning-attacks.
[19] Peponnet, Cyril. 2025. "Cross-Server Tool Shadowing: Hijacking Calls Between Servers — Acuvity." Acuvity | Securing AI, Unlocking Innovation (blog). September 18, 2025. https://acuvity.ai/cross-server-tool-shadowing-hijacking-calls-between-servers/.

- **Tool poisoning and context bleeding**: Malicious instructions injected into responses or prompts can cause unintended data leakage or model manipulation.[20; 21]
- **Supply chain compromise**: Malicious or vulnerable dependencies (plugins, connectors, or software development kits) integrated into MCP servers can introduce backdoors, leading to data exfiltration or privilege escalation.

Accuracy and Testing Limitations

Another limitation is the difficulty of validating MCP server accuracy. Ensuring that the use and analysis of data returned by MCP servers is complete, consistent, and appropriately nuanced will require careful and repeated testing. In our own work, validating server responses against source datasets was a manual, labor-intensive process and used a limited question panel that does not capture all use-cases or interactions with the data.

Likewise, each stage of the model–server interaction must be verified: that the model invoked the correct tool, that the tool was called with the right parameters, that the returned data was accurate, bounded in scope, and did not overwhelm the model's context window, and that the data was interpreted correctly by the model. Defining what counts as a "correct" response can also be subjective. In our evaluation, we counted answers as correct even when the model used a tool we hadn't intended, but other teams may prefer stricter standards.

These challenges grow more complex when multiple dependent tool calls are chained together, such as to answer a question spanning multiple MCP servers or to extract insights from a dataset as complex as USAspending. In these cases, small errors can cascade into larger failures. Establishing a robust framework for testing and validating MCP servers, and ideally automating much of this process, will be essential to ensuring that the MCP servers the government builds are reliable and trustworthy.

We also found an unintended side effect: with an MCP server available, the model was more likely to attempt an answer even when the correct tool or data was not available. In preliminary trials, ChatGPT and Gemini refused to answer 22% and 80% of data retrieval questions, respectively. With the MCP, models only refused to answer in 3.4% of trials across PLACES and USAspending. While refusing to answer is a nonideal user experience, it is preferable to an incorrect answer. Federal stewards should consider adding additional guardrails to LLM-clients that encourage the models to point users to resources or documentation when they're unable to answer a query.

---

[20] Salamone, Salvatore. 2025. "The Growing Importance of Securing MCP Servers for AI Agents." RTInsights. July 5, 2025. https://www.rtinsights.com/the-growing-importance-of-securing-mcp-servers-for-ai-agents/.

[21] Trend Micro — United States (US). 2025. "Why a Classic MCP Server Vulnerability Can Undermine Your Entire AI Agent." Trend Micro. June 24, 2025. https://www.trendmicro.com/en_us/research/25/f/why-a-classic-mcp-server-vulnerability-can-undermine-your-entire-ai-agent.html.

Usability Limitations

At present, the barrier to benefiting from MCP is high. Users must not only be aware that the protocol exists but also locate relevant MCP servers, potentially clone GitHub repositories, run servers locally, and manually configure their AI applications to point to those servers.[22] This workflow is feasible for technical power users but impractical for the general public.

Developing a more seamless process for discovering and connecting to MCP servers will be essential to ensure that the broader public, not just a narrow group of experts, can benefit from their development.

Governance and Accountability

Finally, MCP raises unresolved governance questions: what happens when a server is inaccurate? Who is responsible for correcting errors, and how should users be alerted? What happens when a server has a security vulnerability? Should there be a coordinated disclosure process, or a central authority that vets and certifies servers before they are added to registries? Who decides which servers are "trusted"? Without clear governance, users may unknowingly rely on insecure or unvetted servers.

These questions are especially pressing in government contexts. Establishing governance frameworks, including standards for testing, auditing, and incident response, will be critical to building trust in MCP as a foundation for data access. Teams should incorporate MCP-specific guidance into existing open data and AI governance plans in order to integrate these processes into the existing governance lifecycle as much as possible.

Solving these problems will require collaboration among experts in government and industry. One important piece of the solution is centralized coordination through an MCP registry, which is explored in detail in the next section. Registries can serve not only as catalogs of available servers but also as enforcement points for quality, security, and accuracy standards.

[22] "Connect to Local MCP Servers — Model Context Protocol." n.d. Model Context Protocol. https://modelcontextprotocol.io/docs/develop/connect-local-servers.

IV. Coordinate Federal MCP Servers for Open Government

> *Create a centralized federal MCP registry to coordinate server discovery, enforce quality standards, prevent fragmentation across government data sources, and reduce user burden.*

MCP servers are a valuable tool to enable access to data within a single product. However, if each agency deploys them independently, siloed MCP servers will limit discoverability and usability. Effective implementation of MCPs requires a centralized mechanism that allows access to all available government servers.

We recommend the formation of an MCP registry.[23] This centralized catalog will allow MCP clients to discover and connect to MCP servers. Clients query a discovery API to locate servers matching a user's needs. The registry responds with metadata about each server: its capabilities, tools, endpoints, and connection details. Using this information, the client can establish a direct MCP protocol session with the chosen server. The registry never hosts or runs server code; it functions solely as a discovery and connection tool.

Creation of an MCP registry for federal data access will provide the federal government with an authoritative directory of MCP servers developed and vetted by federal agencies (addressing some of the possible security concerns around MCP servers developed by third parties). This creates a front door to federal government data, allowing AI models to navigate federal MCP servers and return answers grounded in open federal data. Today, running an MCP server often requires technical expertise to set up locally and configure an AI desktop client. A centralized registry eliminates that burden, empowering not only policy writers and academic researchers but anyone seeking insights that federal data can provide.

Registries also can enforce quality and security standards. Government-managed registries, for example, could require servers to undergo accuracy testing, publish validation results, and complete security vetting before being listed. They could also mandate coordinated vulnerability disclosure as part of the onboarding process. By tying governance to discoverability, registries create a clear mechanism for improving accountability and reliability across the MCP ecosystem.

One possibility for hosting this registry is Data.gov. Data.gov already serves as the federal government's official catalog for federal datasets under the OPEN Government Data Act[24], aiming to make federal data more accessible. This mission is reinforced by *Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance*[25] which requires agencies to maintain

---

[23] Modelcontextprotocol. n.d. "GitHub — Modelcontextprotocol/Registry: A Community Driven Registry Service for Model Context Protocol (MCP) Servers." GitHub. https://github.com/modelcontextprotocol/registry.

[24] *Foundations for Evidence-Based Policymaking Act of 2018*, Pub. L. No. 115-435, 132 Stat. 5529 (2019).

[25] *Office of Management and Budget, Memorandum M-25-05, Phase 2 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Open Government Data Access and Management Guidance (Jan. 14, 2025).*

comprehensive data inventories and to publish standardized metadata (DCAT-US 3.0). Expanding Data.gov to include a centralized MCP server registry would further its statutory purpose of making federal data accessible. Another possible host is the Capacity Building office in the Cybersecurity and Infrastructure Security Agency's (CISA) Cybersecurity Division, which hosts the .gov domains.

Technical Implementation

Implementing MCP servers government-wide requires balancing central coordination with agency autonomy. A centralized MCP registry can streamline discovery, promote interoperability, and maintain high standards, but over-centralization risks creating bottlenecks that slow innovation and deployment.

Coordination Model: A "Registry of Registries"

To combine centralized discoverability with decentralized management, the government could adopt a multi-tier registry model:

- **Agency Registries**: Each agency maintains its own MCP registry, managing deployment, access policies, and security for its servers.
- **Central MCP Registry of Registries**: This national-level registry aggregates agency registries into a single query endpoint.
- **Benefits**: AI companies, developers, and other consumers can query one central location to dynamically discover MCP servers across agencies, while agencies retain autonomy over their own infrastructure and governance processes.

This approach mirrors successful private-sector patterns[26] for federated service discovery, providing both a unified interface for external consumers and the flexibility required for diverse agency missions.

A possible proposal for the resource allocation needed to actualize this registry:

- Cloud resources to host all the system components (containers for each individual MCP server and a database with a list of the servers, or list of registries).
- Individual data teams to develop and validate each MCP server.
- A central organizing team that oversees the registry and approves new servers.

[26] "MCP Registry Registry." n.d. https://mastra.ai/mcp-registry-registry.

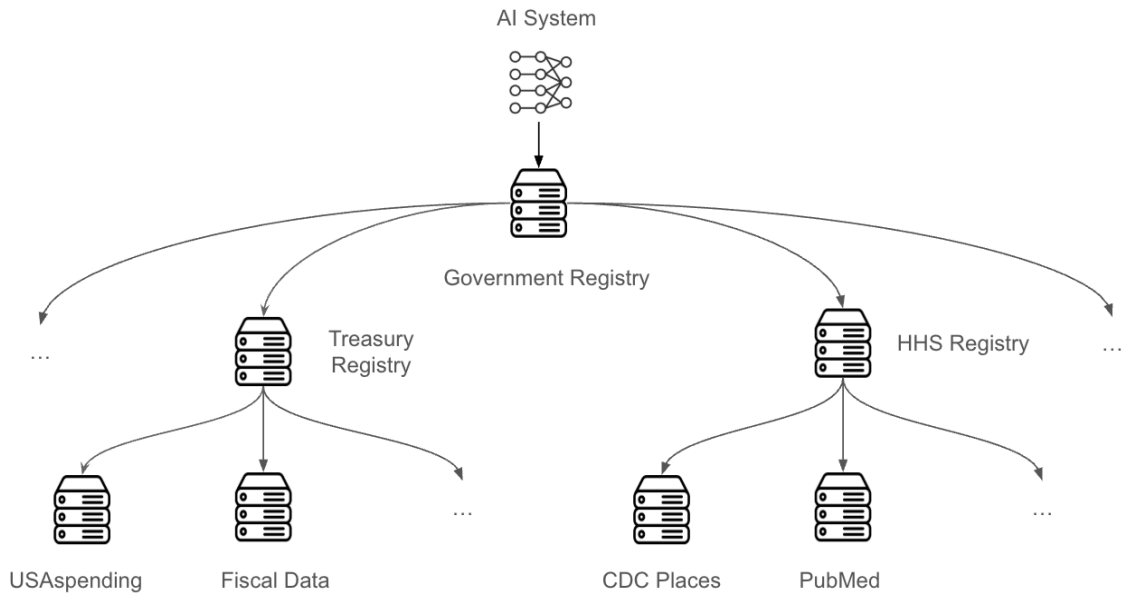

Figure 3 | Schematic proposing a possible design for a federal government-wide data registry. One central registry ("Government Registry") would serve as a catalog of individual department registries (e.g. "Treasury Registry" and "HHS Registry"), which would point to individual MCP servers developed for their datasets. This model allows each agency to build and design their registry independently, while still enabling dynamic discovery of new data for the end user.

V. MCPs can Help Agencies Integrate AI into their Work

For the scope of our project, we focused on how MCPs can be used by the public to access and use federal datasets. However, MCPs could also be used by the government to supplement agentic workflows and increase employee efficiency. One example of how MCPs could support government employees is in document search and retrieval.

Document Search and Retrieval

Federal agencies maintain vast repositories of documents across different systems and formats. It requires institutional and technical knowledge to know when and how to access these systems. Some of this work could be offloaded to an AI assistant. MCP servers could be developed for each of these internal systems and could provide intelligent document discovery that goes beyond simple keyword searches. Instead, users could provide complex directives that an AI assistant would then parse, then use to identify and utilize the appropriate MCP server and tool to pull the relevant data from its source.

MCPs not only have the ability to read data from external systems, they also have the ability to write data. This raises the possibility for use in a variety of situations such as:

1. Regulatory Compliance Automation

Agencies tracking regulatory compliance might have to monitor a variety of sources including inspection databases, violation tracking systems, and enforcement tools. An AI agent could

provide a base level tracking system for continuously monitoring compliance across multiple regions by pulling from MCP servers configured for these different data sources. It could then flag high-risk situations requiring immediate attention. It could even draft inspection reports based on predefined templates.

2. Contract Analysis and Management

Government contracts often reference multiple regulations, agreements, and compliance requirements stored across different systems. MCP can connect management systems with regulatory databases, financial systems, and performance tracking tools. This would enable automated contract compliance monitoring, risk assessment, and performance analysis. In addition, it could provide an initial review of new contracts where it flags potential conflicts with existing contracts, identifies relevant regulations, suggests standard clauses based on similar agreements, and even provides a cost assessment.

3. Security Considerations

Any internal rollout of MCPs, particularly for systems that contain sensitive data, must undergo rigorous security review. As a novel protocol, MCP introduces potential security vulnerabilities that require careful evaluation.Like other emerging technologies that interface with AI systems, MCPs carry inherent security risks. Researchers have identified cases where agentic AI systems have acted out of alignment with their intended behavior, leading to leaks of personal identifiable information or other malicious actions. For example, a recent study released by Anthropic found that AI agents may act in unexpected and potentially harmful ways if they become aware of their imminent replacement or if the organization they operate within begins to take actions in opposition to their assigned objective.[27] Such findings underscore the need to develop MCPs with caution and take thorough due diligence to address all possible security considerations.

That said, MCP servers built for public data and public APIs aimed at broadening access do not face these same risks. The U.S. government should actively encourage development in this space.

To strike the right balance:

- **Oversight and Approval**: All MCP servers should undergo security review, functionality testing, and standards compliance checks prior to deployment.
- **Agency Flexibility**: Agencies should retain the ability to build, deploy, and manage MCP servers when needed, provided these efforts align with established security and compliance requirements.
- **Streamlined Processes**: Governance structures should emphasize automation, standardized templates, and clear documentation to make compliance as frictionless as possible.

[27] "Agentic Misalignment: How LLMs Could be Insider Threats," *Anthropic,* Accessed August 26, 2025 https://www.anthropic.com/research/agentic-misalignment

C. Technical Considerations and Recommendations

Limitations of Web Architecture for AI Access

Agencies sometimes assume that making websites more search-friendly will enable AI models to access federal data effectively. While improvements to web architecture can provide some benefits, our testing shows these changes alone are not enough. Even as AI systems become more capable at browsing the web, they remain constrained by how sites are indexed, how content is delivered, and what information is exposed through user-facing dashboards rather than APIs. As a result, the following issues and potential improvements are worth addressing, but MCPs will likely provide the greater benefit, particularly for structured access to APIs, databases, and other resources that are otherwise difficult for AI models to navigate reliably.

Search Index Dependency

AI models depend on standard web search queries and can only access content already indexed by search engines. For Google, which accounts for roughly 90% of search traffic[28], a page is eligible for indexing if Googlebot (its web crawler) can access the site, the page successfully loads, and the content is in an approved, indexable format that is not classified as spam. Common issues such as robots.txt exclusions, lack of internal linking, technical errors, and authentication requirements can prevent content from being indexed and thus accessed by AI models.

JavaScript Content Loading Issues

AI models proved ineffective on websites built as single page applications (SPAs) and websites that dynamically display content with JavaScript after their initial page loaded. We believe this occurs because AI models use web scraping technology that doesn't execute JavaScript when accessing pages.[29] This is likely a deliberate design choice to optimize performance and reduce the time and cost associated with web scraping operations. The design choice, however, conflicts with the widespread adoption of JavaScript (JS) frontend frameworks such as React,[30] which are commonly used in SPAs.

Data Accessibility Challenges

Our experiments highlighted how these limitations affect federal data portals:

- **USAspending.gov** makes its data available through interactive dashboards that rely on complex search and filtering. Because the site is a SPA, its data does not load without JavaScript. When tested, general-purpose AI search tools failed to answer 68% of

[28] “Search Engine Market Share Worldwide | Statcounter Global Stats.” n.d. StatCounter Global Stats. https://gs.statcounter.com/search-engine-market-share.
[29] “Does ChatGPT and AI Crawlers Read JavaScript? What It Means for SEO.” n.d. https://seo.ai/blog/does-chatgpt-and-ai-crawlers-read-javascript.
[30] “State of JavaScript 2024: Front-end Frameworks.” n.d. https://2024.stateofjs.com/en-US/libraries/front-end-frameworks/.

USAspending questions, largely because the model's web access could not render the content.
- **CDC PLACES**, like USAspending, makes its data available through interactive dashboards on a site that also requires JavaScript to load content. When tested, AI search tools failed to answer 100% of CDC PLACES questions, largely because the model's web access could not render the content.

This doesn't account for the datasets that are inaccessible through web interfaces, which cannot be retrieved by AI models relying on search tools.

Caveat: Agentic AI Approaches

Advanced "agentic" AI systems (such as ChatGPT with Deep Research enabled or Google Gemini Deep Research)[31] can employ headless browsers to render JavaScript, navigate websites, and interact with SPAs in ways that resemble human browsing. This capability is powerful, as it allows the AI to move beyond static scraping and engage with dynamic web interfaces.

However, several limitations remain:

- **Cost and performance tradeoffs**: Running a headless browser is slower and more resource-intensive than static scraping. The AI must load full pages, execute JavaScript, and make sequential decisions, all of which drive up time and compute costs, making this approach more expensive than MCPs.
- **Decision complexity**: Unlike MCPs, which provide structured, machine-readable access to data, agentic systems must infer how to interact with each unique site. This increases error rates. In our USAspending test, the model successfully loaded pages but failed to apply the correct filters, taking 16+ minutes and still returning the wrong answer, while the MCP server returned the correct result in about a minute.
- **Dataset availability limits**: Even with a headless browser, agentic AI cannot access data that isn't available through a public web interface. For example, datasets not available through a web interface remain inaccessible.

In short, while agentic AI approaches demonstrate impressive potential for navigating dynamic content, they are less reliable, slower, and more resource-intensive than MCPs, and they still cannot compensate for the absence of web-accessible datasets.

Recommendations for Improving Website AI Accessibility

Given these constraints, agencies seeking to make their data more accessible to AI models should consider tailoring their web architecture accordingly. Two ways to achieve this are:

1. **Ensure Search Engine Indexability:** Verify that your websites are properly indexable by search engine bots, making your content discoverable through standard search queries.[32]

[31] "Introducing Operator." n.d. OpenAI. https://openai.com/index/introducing-operator/.
[32] "Google Crawling and Indexing | Google Search Central | Documentation | Google for Developers." n.d. Google for Developers. https://developers.google.com/search/docs/crawling-indexing.

2. **Reduce Dynamic Content Loading:** Minimize reliance on JavaScript for content delivery by either:
    a. Converting dynamic content to static pages where feasible
    b. Implementing server-side rendering (SSR) to generate HTML dynamically on the server rather than the client.[33]

While these changes can provide "quick wins" in some cases, in practice many government systems are built on legacy platforms or modern SPAs that are deeply embedded in existing architectures. Reworking them into static or SSR-based designs is often neither quick nor inexpensive.

Because of this, **MCPs are likely the more practical solution for federal agencies**. MCPs enable structured, machine-readable access to data without requiring major reengineering of legacy web infrastructure, making them a more scalable and reliable path to AI accessibility.

MCP Server Design Recommendations

Our design recommendations are informed by our work building two pilot MCP servers, conversations with federal API owners, research into MCPs, and our own technical expertise.

Logic Placement (MCP vs API)

Avoid splitting business logic between the MCP server and the API unless it's absolutely necessary, this creates parallel maintenance burdens and increases inconsistency risks.

- **API responsibilities:** business logic (validations, permissions, constraints), data access and persistence (queries, storage, indexing), and computational workflows (aggregations, scoring, analytics).
- **MCP server responsibilities:** protocol translation (wrapping API endpoints as MCP tools), response shaping (schema adaptation, pagination, batching), and interaction facilitation (helping models invoke the right tools with the right parameters).

Typical MCP server use cases include:

- **Combining related API endpoints** into a single tool to reduce tool-calling overhead.
- **Lightweight transformations** (e.g., comparative statistics) that improve usability but are not natively supported by the API.
- **Multi-call aggregation** when the model struggles with multi-step workflows (e.g., paging).

Number of Tools

- Creating fewer, **more capable tools** is better than mirroring every single API endpoint.
- Group **related API calls** into individual tools whenever possible.

[33] "Rendering on the Web." 2019. Web.Dev. February 6, 2019. https://web.dev/articles/rendering-on-the-web.

- A large number of tools **increases the maintenance burden and likelihood of model indecision.**
- Start with a **minimal viable** set and expand only when concrete use cases require it.

Data Volume Considerations

Large responses risk exceeding token or context limits and can overwhelm the model. For instance, Claude Desktop seems to currently enforce a ~1 MB limit on results returned from MCP tool calls. To mitigate this:

- **Avoid returning entire datasets** unless strictly necessary.
- **Implement paging or size limits** in the MCP server, even if the underlying API allows large responses.
- **Use a summary-first approach**, with the option to fetch details on demand.
- **Add "fetch all pages" support** in the MCP server when the model is unlikely to manage paging by itself.

Good Design Principles

- **Leverage established MCP frameworks** (e.g., FastMCP) to speed up development and reduce errors.
- **Document tools extensively**, including their purpose, parameters, example inputs/outputs, and hints for when to use each tool.
- Use **clear, descriptive naming conventions** for tools and parameters to help the model select the correct tool and use it appropriately.

Secure Design Principles

Security poses one of the most significant challenges in scaling MCP server deployment across the federal government.[34] Teams implementing MCP must treat MCP servers as **production-grade software**, even for public APIs. Safeguards must be embedded at every stage of the MCP lifecycle:

- **Design and Development**: Provide agencies with secure coding guidelines, vetted templates, and reference implementations.
- **Automated Code Analysis**: Use static and dynamic analysis tools to detect common vulnerabilities early.[35]
- **Supply Chain Integrity**: Continuously monitor third-party dependencies to prevent injection of malicious code.[36]
- **Auth/Authz:** Implement authentication/authorization, particularly if the upstream API requires authentication. This can also be used to enforce rate limits on users.
- **Deployment Protections**: Implement measures such as Web Application Firewalls (WAFs) and Distributed Denial of Service (DDoS) protection to guard against network-based threats.[37]

[34] https://arxiv.org/pdf/2503.23278
[35] https://checkmarx.com/learn/sast/effective-static-source-code-analysis/
[36] https://cloud.google.com/software-supply-chain-security/docs/dependencies
[37] https://www.f5.com/glossary/web-application-firewall-waf

- **Ongoing Monitoring**: Conduct periodic security audits, penetration testing, and log analysis to identify and remediate issues in production. Ongoing monitoring can be conducted using standard benchmark questions to detect model drift.

Security governance should also clearly define responsibilities between centralized oversight bodies and individual agencies, ensuring consistent enforcement without hindering operational agility.

CI/CD Stages

- **Implement a full CI/CD pipeline**:
    - Static code analysis (linting, type checking)
    - Dynamic analysis (runtime behavior tests)
    - Unit and integration tests
    - Accuracy/functional tests
    - Security scanning (dependencies, containers)
    - MCP specific security scanning
    - Staging deployment
    - Logging and monitoring
- **Automate deployments** to reduce errors and enable rapid iteration.

General Security Scanning

- Include **dependency and artifact scanning** in CI/CD pipelines.
- Example tools: Dependabot, Checkmarx, Snyk, Trivy, Anchore, Clair, SAST scanners.
- **Run scans on pull requests and on a regular schedule** to catch new vulnerabilities.

MCP-Specific Security Considerations

- **Protect against prompt injection**[38] by validating and sanitizing inputs, especially when consuming data from less-controlled MCP server sources.
- **Audit all tool calls** for traceability and anomaly detection.
- **Implement filtering/redaction** for sensitive fields in API responses.
- **Implement rate limiting** to prevent abuse.

Helping the Model Choose the Right Tool

For MCP to work effectively, models need reliable guidance on which tools to invoke and when. Several strategies can improve tool selection:

- **Use descriptive tool names** and **document each tool thoroughly** with examples, parameters, and usage hints.
- **Provide an initial overview prompt** describing available tools and when to use them.
- Structured, descriptive prompts improve model tool selection.

[38] Young, Sarah. 2025. “Protecting Against Indirect Prompt Injection Attacks in MCP — Microsoft for Developers.” Microsoft for Developers. April 28, 2025. https://developer.microsoft.com/blog/protecting-against-indirect-injection-attacks-mcp.

Guardrails for Tool Use

A critical open question is how to build guardrails that prevent models from confidently using the wrong tool or from producing misleading results when no tool is suitable. Guardrails could take multiple forms, such as:

- Prompt engineering that instructs the model to decline queries outside a tool's scope.
- Model parameter tuning to adjust sensitivity to uncertainty.
- Advising or filtering layers, which provide statistical or interpretive checks before results are passed back to the user.
- Fallback logic that summarizes limits or errors rather than returning incomplete answers.

At this stage, these are ideas rather than established practices. Determining which guardrails are most effective will require further research across government, industry, and academia. Establishing this as a research priority will be essential to ensuring MCP deployments are safe, reliable, and trustworthy.

Helping the Model Use Tools Correctly

- Use **strongly typed schemas** (e.g., Pydantic models) for tool parameters.
- Provide **example calls** for each tool.
- Include **parameter metadata** (allowed ranges or values) where relevant.

Remote Deployment of MCP Servers

Determining the optimal hosting strategy is another key challenge. Running MCP servers locally on individual machines does not scale well, especially considering the potential number of deployments. A better approach is to host MCP servers remotely on infrastructure controlled by the deploying agency.

- **Run MCP servers remotely** in production to decrease user configuration barriers.
- **Containerize or use serverless deployment** for scaling, isolation, and maintainability.
- **Integrate logging, monitoring, and automated deployments**.
- **Ensure secure access and versioned deployments.**

This model:

- **Reduces Operational Overhead**: Agencies maintain direct control while leveraging their existing hosting and monitoring capabilities.
- **Improves Availability**: Cloud-based deployments can support redundancy, load balancing, and rapid scaling.
- **Supports Vendor Integration**: Providers such as Cloudflare already support relevant hosting and networking capabilities, and more vendors are likely to follow.[39]

Scaling & Performance Considerations

- Design MCP servers as **production-ready** with scalability in mind.
- Consider:

[39] https://developers.cloudflare.com/agents/guides/remote-mcp-server/

- Load testing
- Caching repeated queries
- Request batching / multi-call aggregation
- Horizontal scaling for high availability
- Monitoring and alerting for performance degradation

D. Methodology and Results

Data Selection

As part of this pilot, we identified two datasets, CDC PLACES and USAspending, to test the feasibility of this emerging technology.

Selection Criteria

To ensure a meaningful experiment, we sought datasets that would represent the breadth of federal data resources. We used the following criteria:

Criteria	**Rationale**
Availability of an open API	Allows programmatic access without additional barriers such as restricted API keys.
Completion of public API documentation	Provides the LLM with necessary context and helps the team quickly learn how to use the API.
Sufficient complexity such that the average user would need to consult methodology	Ensures the dataset is nuanced enough to make MCP development rigorous and measurable.
Public interest in the dataset and presence of user support	Focuses on datasets of broad relevance while reflecting authentic user interactions.
Access to subject matter experts and data owners	Enables clarification of ambiguities and accurate representation of the dataset.

About the Selected Datasets

CDC PLACES represents a relatively straightforward statistical dataset hosted on the Socrata framework. Socrata underpins many federal open data portals, making the lessons from this pilot broadly transferable. CDC PLACES is derived from the Behavioral Risk Factor Surveillance System (BRFSS); it applies multi-level regression and post-stratification to survey responses in order to generate health estimates for small geographic areas.

In contrast, USAspending exemplifies the complexity of large-scale federal data integration. Drawing from more than 400 government data sources, it provides public access to detailed information on federal awards—including grants, contracts, and loans—supporting transparency and accountability in federal spending.

Together, these datasets illustrate the range of federal data environments and highlight different challenges for MCP implementation:

Dimension	CDC PLACES	USAspending	Relevance for MCP
Type of Data	Public health estimates (e.g., prevalence of chronic diseases at local levels)	Federal spending data (contracts, grants, loans, etc.)	Demonstrates MCP performance across very different subject domains.
Data Structure	Flat tables	Highly complex linkages across selection of tables	Tests MCP's ability to handle both simple and highly relational data.
Platform	Hosted on Socrata, widely used for government open data portals	Custom-built API maintained by Treasury with many detailed endpoints	Shows how MCP interacts with both standardized platforms and bespoke APIs.
Complexity for Users	Moderate – users must reference methodology to interpret measures (e.g., age-adjusted prevalence)	High – users must navigate award hierarchies, agency codes, and financial definitions	Highlights MCP's role in surfacing expert-level context for queries.
Documentation and Support	Standard API docs with general Socrata guidance	Extensive but technical documentation with dedicated developer resources	Provides insight into MCP performance across varying levels of documentation availability and quality.
Public Interest	Relevant to public health researchers, journalists, and policymakers	Central to transparency, accountability, and economic analysis	Highlights the wide range of users and data needs MCP must accommodate.

Question Panel Development

To support systematic evaluation, we developed panels of questions targeting the CDC PLACES and USAspending datasets.

For CDC PLACES, questions were designed to assess three core capabilities. First, the ability to distinguish between different forms of the same measure, such as correctly identifying whether the crude or age-adjusted value was requested when both were available. Second, selecting the most recent and relevant year of data, taking into account the two-year lag between data collection and dataset release. Third, responding accurately across multiple geographic resolutions, including county, ZIP code, or census tract.

For USAspending, we developed a question panel in consultation with subject matter experts (SMEs) from the USAspending team. These discussions helped identify common user misunderstandings, frequently asked questions from the broader user community, and the areas where SMEs believed the data provides strong versus limited support.

The panels for each dataset were organized into three categories:

- **Data Retrieval**: Questions that require numeric lookups and precise identification of the correct value from the dataset.
  CDC PLACES Example: According to the CDC PLACES dataset, what percentage of Wayne County, Michigan residents experienced food insecurity in 2022? Report the raw and age-adjusted values.
  USAspending Example: According to USAspending, how much funding did HHS award to non-profit organizations in Texas in FY 2022?

- **Methodology**: Questions about dataset definitions, estimation methods, validation procedures, and recommended comparison practices.
  CDC PLACES Example: According to the CDC PLACES dataset, how does PLACES generate area-level estimates?
  USAspending Example: According to USAspending, are there reporting thresholds for determining which awards appear in USAspending?

- **Interpretation**: Comparative or analytical questions requiring both retrieval and reasoning, such as identifying the highest prevalence within a region or calculating differences between geographies.
  CDC PLACES Example: According to the CDC PLACES dataset, does Brooks County or Cameron County have a higher rate of Diabetes Prevalence in 2022?
  USAspending Example: According to USAspending, how has federal spending by recipient type (e.g., nonprofit vs private company) changed over time?

This left us with 18 questions for the CDC PLACES question panel (6 data retrieval, 6 methodology, and 6 interpretation) and 22 questions for the USAspending question panel (6 data retrieval, 7 methodology, and 9 interpretation).[40]

Overview of Experiment

We evaluated how LLMs interact with federal data through a series of structured experiments. First, we did some baseline testing using ChatGPT and Gemini to assess their baseline ability to retrieve and interpret data. Next, we applied prompt engineering strategies to explore whether additional guidance could improve model performance. Finally, we implemented MCP servers and used Claude to examine how MCP supports data retrieval and interpretation.

[40] See *Appendix* for question panels.

Round 1 — Baseline

SUMMARY — ChatGPT (GPT-4.1) and Gemini (Flash 2.5) struggled with direct data retrieval and interpretation, often giving incorrect answers or refusing to answer. They performed well on methodology questions where answers were already publicly documented. ***This suggests that LLMs alone cannot reliably interact with federal data without additional context or structured access.***

Methodology

We tested the free-tier versions of ChatGPT (GPT-4.1) and Gemini (Flash 2.5) to evaluate their ability to access and use federal data sources and APIs.[41] This test was designed to assess our hypothesis that federal data is not yet in machine-readable, AI-ready formats.[42]

Each question in the entire question panel was asked ten times to both models to account for variability in model responses. Answers were evaluated using the following scale:

- Correct: Fully correct answer.
- Partially Correct: Answer is partially correct, but some components are missing or incorrect.
- Incorrect: The answer is entirely incorrect.
- No Answer: The model did not provide a response.

Each query was run in a new chat window to avoid influence from previous responses.

CDC PLACES Results

When assessing the models' ability to retrieve data from the CDC PLACES dataset, performance on data retrieval questions was limited. ChatGPT answered incorrectly in 73% of trials, while Gemini refused to answer in 63% of trials. The incorrect answers are usually derived from discrepancies between PLACES's release year and underlying data year — there is a two-year lag between when data is collected and when it's released. For instance, the data in the 2024 PLACES release was collected in 2022. When prompted to find a data point from 2022, the models typically tried to use the 2022 release and failed to pull the correct data point.

In contrast, when queried about the datasets' underlying methodology, both models performed well. ChatGPT correctly answered 93% of methodology questions, and Gemini correctly answered 88%, likely due to the fact that the methodology documents were scraped from a web search and part of the models' training data. Errors in partially correct answers were typically minor, but meaningful, omissions rather than fully incorrect explanations, suggesting that methodological information was generally well represented in the models' knowledge.

For the interpretation questions, the models repeated their behavior from the data retrieval questions, with ChatGPT typically trying to answer and coming up with incorrect answers and Gemini typically refusing to answer.

[41] Round 1 of testing was conducted
[42] This testing was done on 7/24/2025.

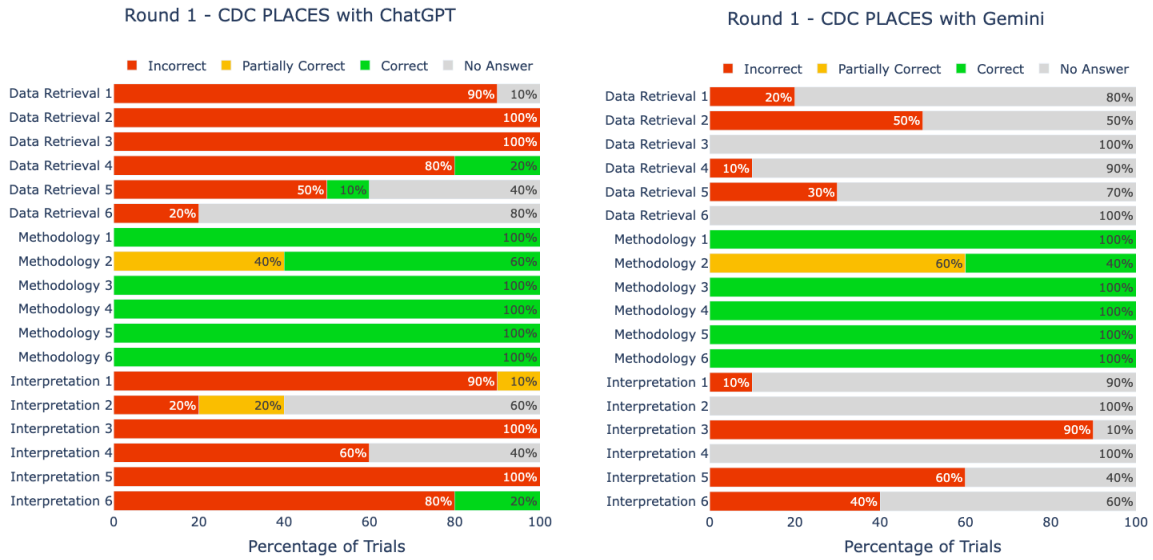

Figure 4 | Initial evaluation of OpenAI's ChatGPT (A) and Google's Gemini (B) models' use of CDC PLACES data. Six questions across three categories (data retrieval, methodology, and interpretation) were asked and answers from the models were graded as either correct, partially correct, incorrect, or, in the case the model refused to answer, as no answer. n=10 trials of each question were conducted for each model.

USAspending Results

Both ChatGPT and Gemini struggled with data retrieval questions. In many cases, the models refused to provide an answer, occasionally citing issues such as pages not loading or an inability to locate the requested data. At times, they attempted to rely on external sources rather than accessing the official website or API directly. More often than not, instead of producing a direct answer, the models offered step-by-step instructions for users to retrieve the data themselves. These behaviors highlight clear limitations in accessing and using information directly from USAspending through the models alone. In contrast, both models performed perfectly on methodology questions, achieving 100% accuracy. This suggests that much of the methodological information is either publicly available in formats accessible to LLMs or included in their training data. For interpretation questions, neither model provided substantive answers, reflecting the same limitations observed in data retrieval tasks.

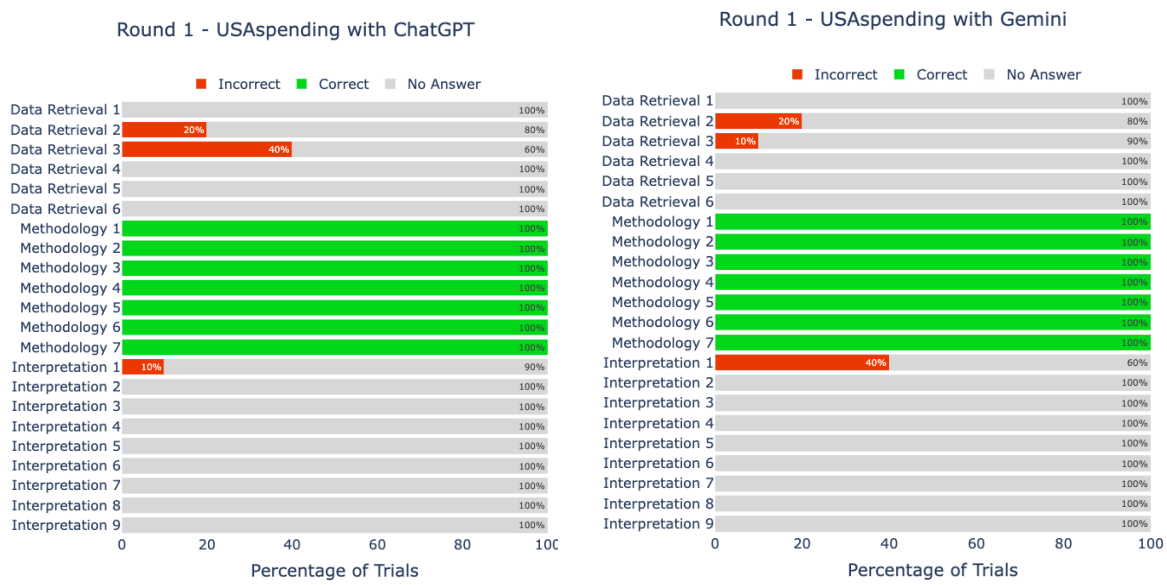

Figure 5 | USAspending data initial evaluation of OpenAI's ChatGPT (A) and Google's Gemini (B) models' use of USAspending data. 22 questions across three categories (data retrieval, methodology, and interpretation) were asked and answers from the models were graded as either correct, partially correct, incorrect, or, in the case the model refused to answer, as no answer. n=10 trials of each question were conducted for each model.

Round 2 — Prompt Engineering

SUMMARY — In Round 2 testing, we evaluated six questions per dataset under seven prompting conditions to assess how context shapes LLM performance, resulting in a total of 42 prompts per panel. Across both CDC PLACES and USAspending, the models consistently failed at independent data retrieval, often defaulting to outdated third-party sources or generating invalid API calls. However, when provided with accurate API responses (simulating MCP), the models reliably parsed and interpreted the data, showing that their main weakness lies in retrieval rather than analysis.

Methodology

To evaluate how contextual information in prompts affects model accuracy, we tested a subset of 6 questions per panel[43] using seven different prompting conditions, ranging from fully open-ended queries to scenarios where relevant data was explicitly provided. This progression allowed us to assess the incremental benefits of prompt design while distinguishing them from broader improvements achievable through MCPs.[44]

The seven prompt conditions[45] were:

[43] See *Appendix* for questions used in this round.
[44] This testing was done on 7/28/2025.
[45] See example prompts for each prompt condition in *Appendix.*

1. **No Dataset Context**: The question was presented without reference to the dataset, requiring the model to independently identify the data source.
2. **No Dataset Context + Citation Request**: As above, but with an instruction to cite the sources used in generating the answer.
3. **No Dataset Context + Citation + Source Justification**: Expanded on the prior condition by requiring the model to explain its source selection.
4. **Dataset-Specified Prompt**: The question was prefaced with the dataset's name (e.g., "According to the CDC PLACES dataset…") to isolate the assessment of data interpretation from source identification.
5. **API Call Generation (No Documentation)**: The model was asked to construct the API endpoint and parameters needed to answer the question, without access to API documentation.
6. **API Call Generation (With Documentation)**: Same as the previous condition, but supplemented with a link to the relevant API documentation.
7. **Embedded API Response**: Simulating an MCP by providing the LLM with API responses, it would ideally retrieve itself via an API to focus on evaluating the LLM's ability to process and understand data, rather than being confounded by retrieval failures. This helps answer the question: "If the LLM had the right data, could it use it effectively?"

Each question was tested under all seven prompt conditions, resulting in 42 prompts per panel. Answers were evaluated using the same scale as Round 1: Correct (fully correct), Partially Correct (partially correct, with some missing or incorrect components), Incorrect (entirely incorrect), and No Answer (no response). Each query was run in a new chat session to avoid influence from previous responses. This round of testing was run using ChatGPT (GPT-4.1).

This structured approach enabled us to systematically isolate the effect of prompt design on retrieval and interpretation accuracy, spanning unconstrained natural language queries to highly structured, tool-assisted scenarios.

CDC PLACES Results

When tested with the CDC PLACES dataset in Round 2, ChatGPT continued to struggle with retrieval despite added prompt structure. In prompt conditions 1 and 2, where no dataset context was provided, the model occasionally recognized PLACES but more often relied on third-party websites and failed to return the correct values. Adding a dataset reference in prompt condition 4 did not improve accuracy, indicating that cueing the source alone was insufficient. In conditions 5 and 6, which required generating API queries, the model sometimes identified the correct endpoint but almost always applied incorrect or incomplete parameters, resulting in invalid queries. Only in prompt condition 7, where we simulated an MCP by embedding the JSON output of the relevant API call, did the model reliably produce correct answers, demonstrating that while it can interpret PLACES data once provided in a structured format, the model cannot independently locate or query the dataset effectively.

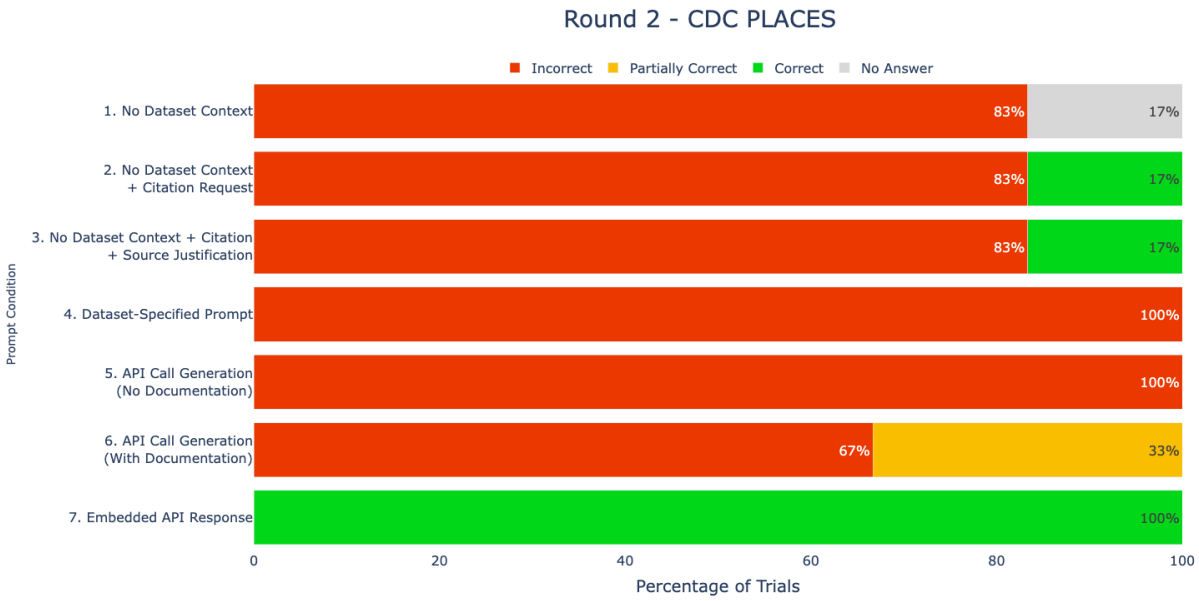

Figure 6 | Model responses from prompt engineering testing with CDC PLACES data. Model was given varying levels of context in the question (y-axis) and rated on the correctness of their answer over six different questions. n=6 questions (see Methods).

USAspending Results

In Round 2 testing for USAspending, ChatGPT continued to struggle with data retrieval. With prompt conditions 1 and 2, the model sometimes attempted to use USAspending but failed to extract data and was consistently incorrect. Instead, it often defaulted to third-party sources, such as news articles or GAO reports, that occasionally cited USAspending but contained outdated numbers. Specifying USAspending directly with prompt condition 4 led the model to reference the dataset, but it frequently failed to locate the correct webpage or interpret its contents, again defaulting to other sources. By contrast, with prompt conditions 6 and 7, the model performed better using USAspending's API documentation, including GitHub pages, and generally produced partially correct API calls that required minor parameter adjustments. When provided with the accurate API response, the model parsed and interpreted the data correctly. Overall, this round of testing showed that while ChatGPT remains poor at retrieving data from USAspending in its current state, it is capable of parsing and analyzing the data effectively once it is provided.

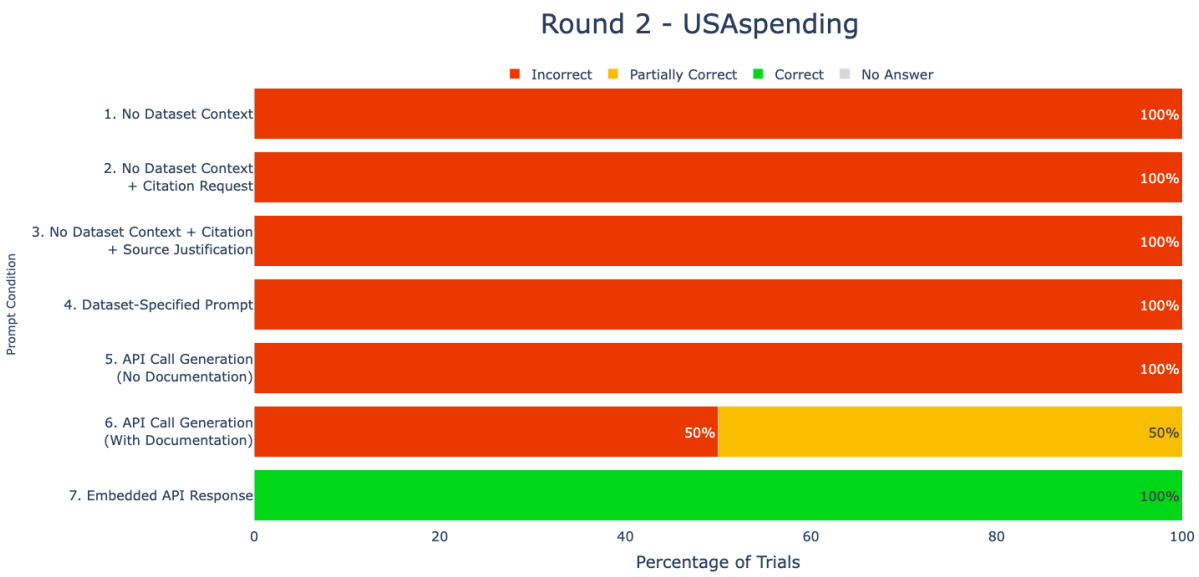

Figure 7 | Model responses from prompt engineering testing with USAspending data. Model was given varying levels of context in the question (y-axis) and rated on the correctness of its answer over six different questions. n=6 questions (see Methods).

Round 3 — MCP Implementation

SUMMARY — In Round 3, we tested question panels with and without MCP servers. Implementing MCP servers allowed models to access data directly, resulting in 100% accuracy for CDC PLACES and a substantial improvement to 89.5% for USAspending, while also speeding response times. These results demonstrate that MCPs are an effective mechanism for enabling reliable, accurate LLM interaction with complex federal datasets.

Methodology

Given LLMs' limited ability to directly retrieve structured federal data, we needed to test a solution that would enable LLMs to retrieve data, focusing less on the analysis piece. These models consistently failed to autonomously invoke or navigate federal APIs even when explicitly prompted to. As a result, we focused on designing a mechanism that allows LLMs to engage directly with federal data in a structured, context-aware manner (via API) by building proof-of-concept MCP servers for the CDC PLACES and USAspending APIs.[46]

The substantial differences between the two datasets[47] led to distinctly different experiences in building their MCPs.

For CDC PLACES, the building experience was straightforward. The dataset is delivered through the Socrata platform, which provides a well-structured and predictable API. This meant that key design decisions centered on mapping query parameters to natural language inputs and ensuring the MCP could reliably retrieve values across different geographies and years.

[46] This testing was done on 8/1/25.
[47] See *Methodology and Results — Data Selection*

Because PLACES is a modeled dataset with consistent fields and naming conventions, the MCP server was able to be implemented with minimal complexity, producing stable, high-accuracy responses with comparatively little custom handling. One notable limitation, however, was the lack of direct API documentation. The available materials are primarily oriented toward helping users navigate the PLACES data portal and web interface, rather than guiding developers in working programmatically with the API.[48]

For USAspending, the building experience was substantially more complex. Unlike PLACES, USAspending exposes many endpoints with complex parameters. Designing the MCP required careful scoping to avoid overwhelming the model with too many tools while still preserving useful coverage. We focused on building a subset of endpoints that demonstrated how MCPs can support structured queries across awards, agencies, and geographic areas. Even so, additional guardrails and error handling were necessary to ensure reliable responses. USAspending provides a fair amount of user support and documentation for the API, which helped guide our design choices. However, during the build process, it became clear that we should have mirrored more of USAspending's web interface structure and conventions by grouping related API requests into a single tool. This would have reduced the number of tool calls required while still providing the model with all the necessary data to answer a given query. This underscores that MCP servers should ultimately be developed by the API owners and developers who best understand their structures and interaction patterns.[49]

To evaluate the effectiveness of the MCP for the PLACES dataset, we generated a new panel of 10 questions[50] focusing on single data point retrieval (Data Retrieval), comparisons between two data points (Direct Comparison), and comparison across geographies in a region (Multiple Comparison). These questions would, in turn, test the model's ability to query for individual datapoints, use repeated queries of the single data point retrieval tool to compare answers, and to use the regional comparison tools that we developed.

The USAspending API has over a hundred endpoints, so we coded only a subset[51], prioritizing proof-of-concept testing to assess whether this emerging technology is worth pursuing rather than building a fully comprehensive MCP. Since we were unable to dedicate the time needed to build a fully comprehensive MCP for USAspending, we organized our 22-question panel into 19 questions that required only the endpoints implemented on our MCP server, and 3 questions that were outside the knowledge base and required endpoints that we did not implement on our MCP server.[52] This distinction allowed us to assess both the server's performance under ideal conditions, where there was a tool that could directly answer the user's query, and its capacity to generate responses when no direct endpoint integration was available.

For Round 3 of testing, we evaluated our question panels both with and without the MCP servers we had built, using Claude Sonnet 4 – the only LLM at the time that supported MCP

[48] Proof-of-concept MCP server for CDC PLACES can be found here: https://github.com/GSA-TTS/cdc-places-mcp-server
[49] Proof-of-concept MCP server for USAspending can be found here: https://github.com/GSA-TTS/usa-spending-mcp-server-DEMO
[50] See *Appendix* for question panel
[51] See *Appendix* for list of USAspending api endpoints used.
[52] See *Appendix* for questions included in this round of testing.

integrations.[53] Each prompt was tested across 10 trials to account for variability in model responses. Answers were evaluated using the same scale as Rounds 1 and 2: Correct (fully correct), Partially Correct (partially correct, with some missing or incorrect components), Incorrect (entirely incorrect), and No Answer (no response). Each query was run in a new chat session to avoid influence from previous responses.

CDC PLACES Results

For each question, without the MCP active, the model did not generate an answer and gave instructions for directly accessing the dataset. With MCP active, the model achieved near-perfect accuracy across all categories. It consistently returned correct numeric values, distinguished between crude and age-adjusted measures, and drew from the correct year and geography. Tasks that had previously failed in most pre-MCP trials were now answered accurately and consistently. These results exhibit the clear benefit of data retrieval and comparison tasks enabled by the MCP connection to a dataset API.

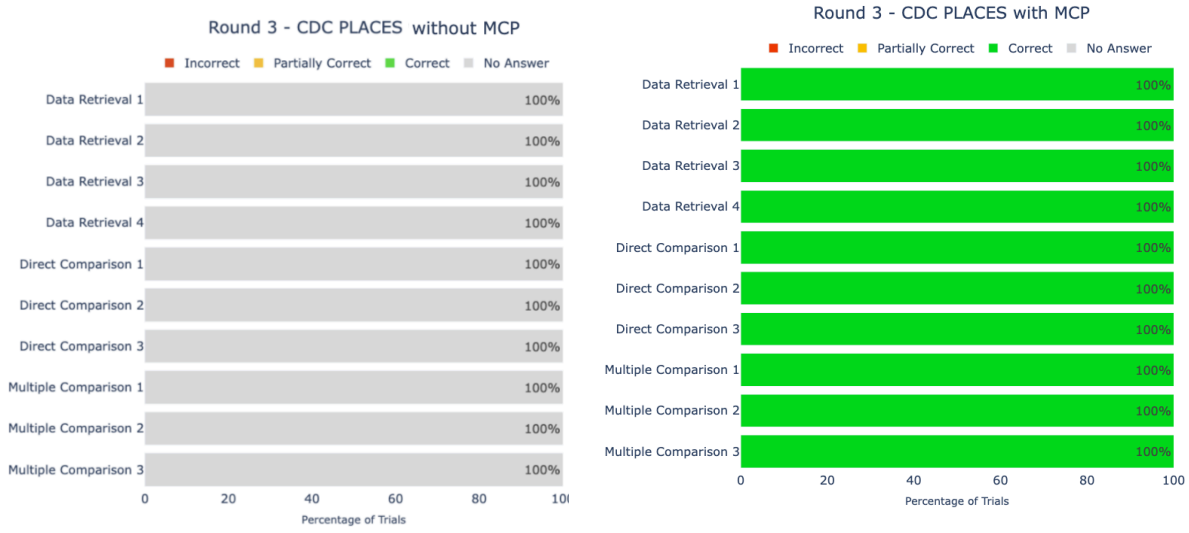

Figure 8 | PLACES data retrieval performance without (A) and with (B) an active MCP connection. Without an MCP, across all trials and questions, the model did not attempt to give an answer and instead suggested steps for accessing the PLACES data directly (n=5 trials for each question). With an active MCP connection, PLACES data retrieval was near-perfect as the model produced the correct result across all trials and questions (n=5 trials for each question).

USAspending Results

In the absence of the MCP server, the model incorrectly answered 15 out of 22 (68%) of our questions. In most cases, the model explicitly stated that it was unable to retrieve data from USAspending, often citing issues such as web pages failing to load properly.

The model occasionally produced correct responses even without access to the MCP server. It was particularly effective on methodological questions, where it located and referenced static

[53] Now, MCP integration is supported by most major AI companies; https://mlq.ai/news/enterprise-ai-adoption-rapidly-evolves-with-anthropics-mcp/

text from documentation-rich webpages. These sources often contained detailed explanations of data definitions, limitations, and usage guidance, enabling the model to construct accurate responses based on publicly available information.

One notable example is Question 16: “According to USAspending, how has spending on different award types (contracts, grants, direct payments, etc.) changed in number and value over the last decade?” The model answered this question correctly; however, the response did not appear to rely on direct data retrieval from USAspending. Instead, it drew on general knowledge or heuristic trends, referencing a mix of sources including federal authorities such as the Congressional Budget Office and Government Accountability Office, as well as third-party organizations like USAFacts and KFF[54].

These results suggest that without a structured protocol like MCPs, models are limited in their ability to access and accurately interpret dynamic, data-rich sources such as USAspending. While the model can occasionally compensate using general knowledge or external summaries, this often leads to incomplete or imprecise answers, emphasizing the need for integrated retrieval mechanisms to ensure data fidelity and reliability in LLM-generated responses.

Using the MCP server, the model achieved significantly higher accuracy on knowledge base questions (89.5%). The only incorrect response occurred because the USAspending congressional district endpoint was down at the time of testing, and one additional response was only partially correct. Not only was the accuracy improved, but often, the model was much faster at answering these questions than they were without the MCP server.

When questions extended beyond the MCP’s available tools, accuracy did not improve. In certain cases, performance appeared worse: where models had previously returned nonresponses, the MCP configuration sometimes generated incorrect answers by attempting to repurpose available tools. This dynamic raises a notable concern, as users may be more inclined to trust a confident but inaccurate response than a clear indication that no answer is available. Importantly, these weaker results do not reflect a failure of the MCP framework itself, but rather the fact that our MCP implementation was only partially developed. **This underscores a critical point: MCP servers must be designed and maintained by subject matter experts with deep knowledge of the underlying data and APIs. Federal agencies, as the authoritative stewards of these datasets, are best positioned to lead this work and ensure that MCPs are implemented in ways that minimize risks and maximize reliability.**

[54] https://www.kff.org/

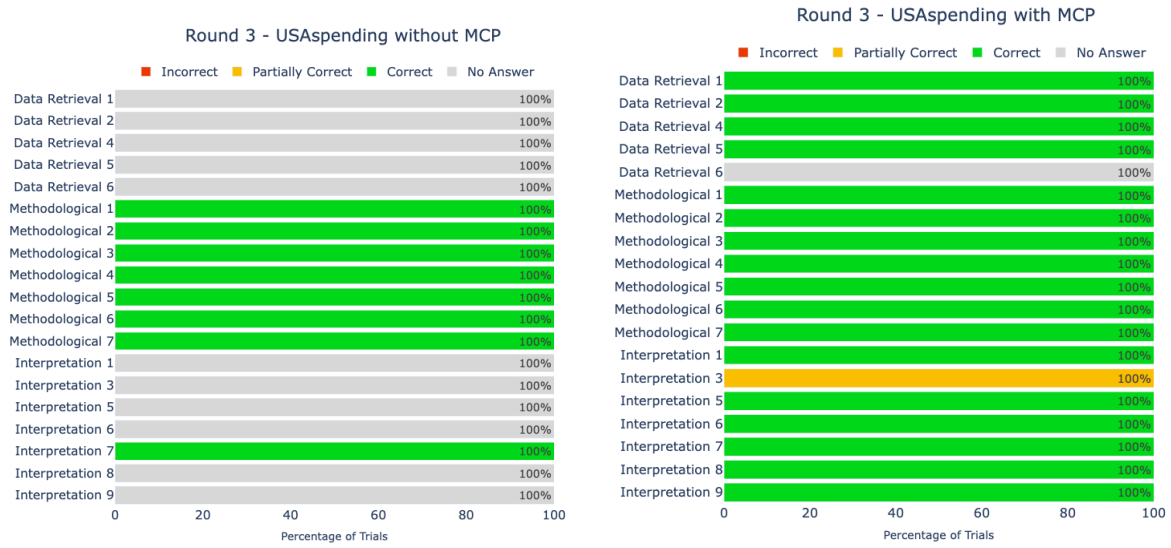

Figure 9 | USAspending data retrieval performance without (A) and with (B) an active MCP connection. Use of an MCP server significantly improves Claude's ability to retrieve and accurately interpret data from the USAspending platform.

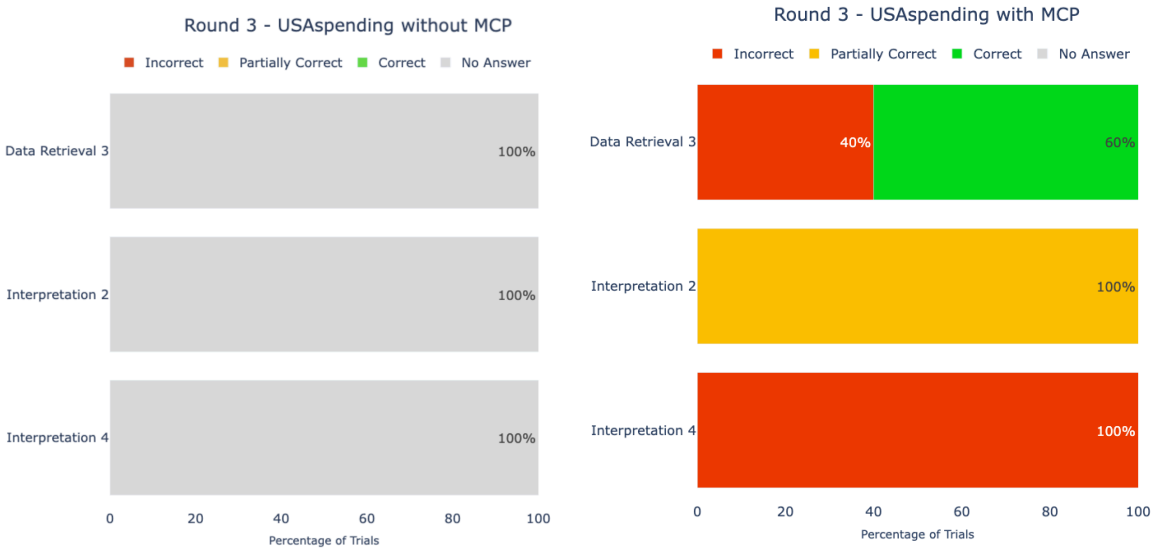

Figure 10 | USAspending data retrieval performance for questions outside the MCP server's knowledge base, both without (A) and with (B) an active MCP connection. When the server was asked questions not covered by the available tools, it attempted to repurpose the tools — sometimes incorrectly — to generate an answer. This often resulted in Claude generating inaccurate responses, rather than acknowledging the absence or insufficiency of relevant data.

Conclusion

We tested LLMs' ability to interact with two federal datasets, CDC PLACES and USAspending, across multiple rounds to evaluate data retrieval, methodology understanding, and interpretive reasoning. Baseline testing showed that models consistently failed at retrieving data

independently, often relying on outdated third-party sources, though they performed well on methodology questions. Prompt engineering improved performance marginally, but reliable results were only achieved when models were provided structured API responses, highlighting that the main limitation lies in retrieval rather than analysis. Implementation of MCP servers enabled LLMs to access and interpret data accurately and efficiently, achieving 100% accuracy on CDC PLACES and substantially improving results on USAspending. **These results suggest that MCPs are a promising mechanism for federal agencies to deliver structured data to LLMs, significantly enhancing their ability to accurately interpret and analyze complex federal datasets.**

E. Appendix

Round 1 — Baseline

CDC PLACES Round 1 Question Panel

18 questions total: 6 data retrieval, 6 methodology, 6 interpretation.

Question Type	Question	Answer
Data Retrieval 1	What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.	19.4% (Raw); 20.3% (Age-adjusted)
Data Retrieval 2	What was the 95% confidence interval for the percentage of adults who smoke in Kings County, New York in 2022? Use the raw number.	10.5%-13.0%
Data Retrieval 3	What percentage of people in census tract 11001008804, located in the District of Columbia, rate themselves in fair or poor health in the most recent year of data?	15.10%
Data Retrieval 4	What percentage of adults have asthma in Miami-Dade County, Florida in 2022?	9.0% (raw); 9.1% (age-adjusted)
Data Retrieval 5	According to the CDC PLACES dataset, which county in Florida has the highest rate of depression among adults according to age-adjusted prevalence in 2022 and what is the prevalence value?	25.30%
Data Retrieval 6	What is the age-adjusted prevalence of physical distress amongst adults in West Lafayette, IN in 2022?	12.20%
Methodology 1	How does PLACES generate area-level estimates?	PLACES uses a multilevel regression and poststratification (MRP) method to generate estimates of each measure
Methodology 2	How is “short sleep duration” defined in PLACES?	<7 hours, on average, during a 24-hour period
Methodology 3	How are the place level estimates validated in the PLACES model?	Both internal and external validation studies showed strong/moderate correlations between model-based estimates and direct survey estimates at state, county, and place levels
Methodology 4	Which measurement of binge drinking rate among adults should be used to compare Worcester County, MA and Loudoun County, VA?	age-adjusted
Methodology 5	What are the primary data sources used in the PLACES estimation model?	BRFSS (Behavioral Risk Factor Surveillance System), ACS (American

		Community Survey), and Census population data
Methodology 6	How are county-level estimates for the number of adults for a given measure generated in the PLACES model?	The county-level estimates are obtained by multiplying the probability by the total adult population of each county.
Interpretation 1	What is the difference in percent of adults who have health insurance between Davidson County, NC and Worcester County, MA in 2022?	13% (Davidson County, age-adjusted) — 4.9% (Worcester County, age-adjusted) = 7.9%
Interpretation 2	Which county in Indiana has the highest rate of doctor's visits in the last year?	Lake County (79.5%)
Interpretation 3	What is the estimated number of adults living with asthma in Miami-Dade County, Florida?	193,043, 195,188 (age-adjusted), year:2022
Interpretation 4	Which Florida county has the most adults with depression and what's the number?	Bradford: 5,110, Putname: 14,934 (age-adjusted), year:2022
Interpretation 5	Does Brooks County or Cameron County have a higher rate of Diabetes Prevalence in 2022?	Confidence intervals are too wide for the result to be meaningfully different (Cameron is slightly higher)
Interpretation 6	What is the overall prevalence of adults who received food stamps in Nevada in 2022?	9.80%

USAspending Round 1 Question Panel

22 questions total: 6 data retrieval, 7 methodology; 9 interpretation.

Question Type	Question	Answer
Data Retrieval 1	How much did Ernst & Young get from HHS OIG in FY2023? What percentage of HHS OIG spending was it?	13.5 million 2.8%
Data Retrieval 2	How much did the state of Georgia get on broadband from the Department of Commerce?	1.3 billion
Data Retrieval 3	How much did the Dept of Ed obligate in direct payments for individuals in FY 2020?	29.59 billion
Data Retrieval 4	How much funding did HHS award to non-profit organizations in Texas in FY 2022?	3.2 billion
Data Retrieval 5	Which states have shown the largest year-over-year increase in disaster-related spending?	There is not an easy way to answer this question. In past year CA had most with with $186B
Data Retrieval 6	Which congressional districts consistently receive the highest transaction amounts in FY23	Highest obligation is IN-07 ($127,526,305,714) and $170,118 per capita
Methodological 1	What is the difference between a prime award and a subaward in USAspending?	Prime from agency directly, sub is downstream distributions
Methodological 2	Are there reporting thresholds for determining which awards appear in USAspending?	Yes, if award is greater than $25,000, reporting is required, though smaller awards can be reported optionally
Methodological 3	Can you explain what an "award" is? How are awards categorized, and how do they relate to other types of federal assistance such as grants, loans, and contracts?	An award is money the federal government has promised to pay a recipient....
Methodological 4	What are the limitations of examining defense spending through this API? How should that qualify answers about defense spending?	Held for 90 days, some not included

Methodological 5	If an award shows a negative obligation value, should it be included when summing total spending?	From the USA Spending glossary, negative obligations, or de-obligations, occur when agencies decrease previous obligations to correct errors or to reflect new information (for example, that the price of a project was lower than expected). De-obligations are common in instances where the scope of a project changes.
Methodological 6	How do I find awards prior to 2017/2007/2001?	For searches, time period start and end dates are currently limited to an earliest date of 2007-10-01. For data going back to 2000-10-01, use either the Custom Award Download feature on the website or one of our download or bulk_download API endpoints.
Methodological 7	How do I find total spending for an award/agency/congressional district/state?	For individual awards, use spending by award endpoint. Can use spending explorer to find total agency spending, can use spending by geography to find geography specific breakdown.
Interpretation 1	Did the Indian Health Service or Bureau of Indian Affairs award more funding to tribal governments in FY 2021?	Indian Health Services (7.44B vs 4.61B)
Interpretation 2	In FY 2022, did more total funding go to the top 5 recipients in New York or California?	California (319.61B vs 209.81B)
Interpretation 3	Which federal agency had the highest total obligation amount for contracts in FY 2023?	Department of War ($579,672,665,249)
Interpretation 4	Between FY 2020 and FY 2023, how did disaster-related obligations change across all award types?	It decreased by around 30B $116,309,502,896 -> $83,655,216,198
Interpretation 5	How has spending on different award types (contracts, grants, direct payments, etc.) changed in number and value over the last decade?	Contracts grew slightly, grants and direct payments surged because of the pandemic. Overall, trend towards “assistance” awards with contracts making up only a modest percentage of spending.
Interpretation 6	Which agencies have the largest discrepancy between their obligated amounts and actual spending over the last 3 years?	HHS and DoW had largest gap (scale is hundreds of billions)
Interpretation 7	Are there specific quarters or fiscal periods when most award obligations are made?	The highest spending almost always occurs in the 4th quarter. (Trying to spend money allocated)
Interpretation 8	What are the trends in subaward obligations for top prime awards over the past three years?	Subaward obligations have remained relatively consistent over the past few years. They tend to be a modest fraction of the prime award total.
Interpretation 9	How has federal spending by recipient type (eg. nonprofit vs private company) changed over time?	Nonprofit spending was highest 2012-2016. Then dropped dramatically 2017-2019. Then increased in 2020 before dropping in 2021 and slowly increasing to 2024. Private companies followed a very similar trend. They saw a slightly smaller drop in 2017 but a bigger increase in 2020. And spending did not peak as much in 2024 as in nonprofits.

Round 2 — Prompt Engineering

Prompt Condition Examples

Condition	Example Question
No Dataset Context	What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.
No Dataset Context + Citation Request	What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022. Cite the sources used.
No Dataset Context + Citation + Source Justification	What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022. Cite the sources used and explain the choice of sources based on your evaluation of the data documentation, quality of data source, and the quality of metadata available for that source.
Dataset-Specified Prompt	According to the CDC PLACES dataset, what percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.
API Call Generation (No Documentation)	According to the CDC PLACES dataset, provide the API endpoint and parameters to answer what percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.
API Call Generation (With Documentation)	According to the CDC PLACES dataset, provide the API endpoint and parameters to answer What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.. Refer to this documentation to construct the endpoint: https://dev.socrata.com/foundry/data.cdc.gov/swc5-untb
Embedded API Response	According to the CDC PLACES dataset, provide the API endpoint and parameters to answer What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022. Use this API response to answer the question: Using the following API response: [{ "Year": 2022, "StateAbbr": "MI", "StateDesc": "Michigan", "CountyName": "Wayne County", ...

CDC PLACES Round 2 Question Panel

6 questions total, all data retrieval questions.

Question	Answer
What percentage of Wayne County, Michigan residents experienced food insecurity? Report the raw and age-adjusted values from 2022.	19.4% (Raw); 20.3% (Age-adjusted)
What was the 95% confidence interval for the percentage of adults who smoke in Kings County, New York in 2022? Use the raw number.	10.5%-13.0%
What percentage of people in census tract 11001008804, located in the District of Columbia, rate themselves in fair or poor health in the most recent year of data?	15.10%
What percentage of adults have asthma in Miami-Dade County, Florida in 2022?	9.0% (raw); 9.1% (age-adjusted)
Which county in Florida has the highest rate of depression among adults according to age-adjusted prevalence in 2022 and what is the prevalence value?	Putnam County (25.3%)

What is the age-adjusted prevalence of physical distress amongst adults in West Lafayette, IN in 2022?	12.2%

USAspending Round 2 Question Panel

6 questions total, all data retrieval questions.

Question	Answer
How much did Ernst & Young get from HHS OIG in FY2023? What percentage of HHS OIG spending was it?	13.5 million 2.8%
How much did the state of Georgia get on broadband from the Department of Commerce?	1.3 billion
How much did the Dept of Ed obligate in direct payments for individuals in FY 2020?	29.59 billion
How much funding did HHS award to non-profit organizations in Texas in FY 2022?	3.2 billion
Which states have shown the largest year-over-year increase in disaster-related spending?	There is not an easy way to answer this question. In past year CA had most with with $186B
Which congressional districts consistently receive the highest transaction amounts in FY23	Highest obligation is IN-07 ($127,526,305,714) and $170,118 per capita

Round 3 — MCP Implementation

CDC PLACES Round 3 Question Panel

Category	Question	Answer
Single Data Point Retrieval 1	According to the CDC PLACES dataset, what percent of adults had short sleep durations in Kauai County Hawaii in 2018?	Crude: 36.9%; Age-adjusted: 38.1%
Single Data Point Retrieval 2	According to the CDC PLACES dataset, what percent of adults had colorectal cancer screening in census tract 48005000302 in 2020?	Crude: 66.0%
Single Data Point Retrieval 3	According to the CDC PLACES dataset, what percent of adults had a mobility disability in the 04401 zip code in 2021?	Crude: 12.5%
Single Data Point Retrieval 4	According to the CDC PLACES dataset, what percent of adults had a lack of social support in West Lafayette, IN in 2022?	Crude: 26.5%; Age-adjusted: 22.3%
Data Point Comparison 1	According to the CDC PLACES dataset, did West Lafayette, IN or West Lafayette, OH have a higher rate of people who received food stamps in 2022?	West Lafayette, OH (21.3% vs 13.8%)
Data Point Comparison 2	According to the CDC PLACES dataset, which year between 2018 and 2022 did Nassau, NY have the lowest rate of adults without health insurance?	2022 (2018: 12%, 2019: 13.2%, 2020: 11.7%, 2021: 6.2%, 2022: 5.8%)
Data Point Comparison 3	According to the CDC PLACES dataset, did the 50636 zip code have a higher rate of adults with physical distress or mental distress in 2019?	Physical distress (12.2% vs 12.0%)
Multi-point Comparison 1	According to the CDC PLACES dataset, which county in Arkansas had the highest rate of dental visits in 2022?	Benton County, AR (62.8%)

Multi-point Comparison 2	According to the CDC PLACES dataset, what is the average rate of adults with cancer for all census tracts in Davidson, NC in 2018?	7.30%
Multi-point Comparison 3	According to the CDC PLACES dataset, what is the interquartile range of physical inactivity among all places in Vermont in 2022?	4.00%

USAspending Round 3 Question Panel — In Knowledge Base

19 questions total: 5 data retrieval, 7 methodological, 7 interpretation.

Question Type	Question	Answer
Data Retrieval 1	According to USAspending, How much did Ernst & Young get from HHS OIG in FY2023? What percentage of HHS OIG spending was it?	13.5 million 2.8%
Data Retrieval 2	According to USAspending, How much did the state of Georgia get on broadband from the Department of Commerce?	1.3 billion
Data Retrieval 4	According to USAspending, How much funding did HHS award to non-profit organizations in Texas in FY 2022?	3.2 billion
Data Retrieval 5	According to USAspending, Which states have shown the largest year-over-year increase in disaster-related spending?	There is not an easy way to answer this question. In past year CA had most with with $186B
Data Retrieval 6	According to USAspending, Which congressional districts consistently receive the highest transaction amounts in FY23	Highest obligation is IN-07 ($127,526,305,714) and $170,118 per capita
Methodological 1	According to USAspending, What is the difference between a prime award and a subaward in USAspending?	Prime from agency directly, sub is downstream distributions
Methodological 2	According to USAspending, Are there reporting thresholds for determining which awards appear in USAspending?	Yes, if award is greater than $25,000, reporting is required, though smaller awards can be reported optionally
Methodological 3	According to USAspending, Can you explain what an “award” is? How are awards categorized, and how do they relate to other types of federal assistance such as grants, loans, and contracts?	An award is money the federal government has promised to pay a recipient....
Methodological 4	According to USAspending, What are the limitations of examining defense spending through this API? How should that qualify answers about defense spending?	Held for 90 days, some not included
Methodological 5	According to USAspending, If an award shows a negative obligation value, should it be included when summing total spending?	From the USA Spending glossary, negative obligations, or de-obligations, occur when agencies decrease previous obligations to correct errors or to reflect new information (for example, that the price of a project was lower than expected). De-obligations are common in instances where the scope of a project changes.
Methodological 6	According to USAspending, How do I find awards prior to 2017/2007/2001?	For searches, time period start and end dates are currently limited to an earliest date of 2007-10-01. For data going back to 2000-10-01, use either the Custom Award Download feature on the website or one of our download or bulk_download API endpoints.
Methodological 7	According to USAspending, How do I find total spending for an award/agency/congressional district/state?	For individual awards, use spending by award endpoint. Can use spending

		explorer to find total agency spending, can use spending by geography to find geography specific breakdown.
Interpretation 1	According to USAspending, Did the Indian Health Service or Bureau of Indian Affairs award more funding to tribal governments in FY 2021?	Indian Health Services (7.44B vs 4.61B)
Interpretation 3	According to USAspending, Which federal agency had the highest total obligation amount for contracts in FY 2023?	Department of War ($579,672,665,249)
Interpretation 5	According to USAspending, How has spending on different award types (contracts, grants, direct payments, etc.) changed in number and value over the last decade?	Contracts grew slightly, grants and direct payments surged because of the pandemic. Overall, trend towards "assistance" awards with contracts making up only a modest percentage of spending.
Interpretation 6	According to USAspending, Which agencies have the largest discrepancy between their obligated amounts and actual spending over the last 3 years?	HHS and DoW had largest gap (scale is hundreds of billions)
Interpretation 7	According to USAspending, Are there specific quarters or fiscal periods when most award obligations are made?	The highest spending almost always occurs in the 4th quarter. (Trying to spend money allocated)
Interpretation 8	According to USAspending, What are the trends in subaward obligations for top prime awards over the past three years?	Subaward obligations have remained relatively consistent over the past few years. They tend to be a modest fraction of the prime award total.
Interpretation 9	According to USAspending, How has federal spending by recipient type (eg. nonprofit vs private company) changed over time?	Nonprofit spending was highest 2012-2016. Then dropped dramatically 2017-2019. Then increased in 2020 before dropping in 2021 and slowly increasing to 2024. Private companies followed a very similar trend. They saw a slightly smaller drop in 2017 but a bigger increase in 2020. And spending did not peak as much in 2024 as in nonprofits.

USAspending Round 3 Question Panel — Out of Knowledge Base

3 questions total: 1 data retrieval, 2 interpretation.

Question Type	Question	Answer
Data Retrieval 3	According to USAspending, How much did the Dept of Ed obligate in direct payments for individuals in FY 2020?	29.59 billion
Interpretation 2	According to USAspending, In FY 2022, did more total funding go to the top 5 recipients in New York or California?	California (319.61B vs 209.81B)
Interpretation 4	According to USAspending, Between FY 2020 and FY 2023, how did disaster-related obligations change across all award types?	It decreased by around 30B $116,309,502,896 -> $83,655,216,198

USAspending Endpoints Used[55]

Endpoint	Description

[55] "USAspending API." n.d. https://api.usaspending.gov/docs/endpoints.

/api/v2/agency/<TOPTIER_AGENCY_CODE>/sub_agency/	Returns a list of sub-agencies and offices with obligated amounts, transaction counts and new award counts for the agency in a single fiscal year
/api/v2/agency/<TOPTIER_AGENCY_CODE>/sub_components/	Returns a list of bureaus for the agency in a single fiscal year
/api/v2/agency/<TOPTIER_AGENCY_CODE>/sub_components/<BUREAU_SLUG>/	Returns a list of federal_accounts by bureau for the agency in a single fiscal year
/api/v2/search/spending_by_award/	Returns the fields of the filtered awards
/api/v2/awards/<AWARD_ID>/	Returns details about specific award
/api/v2/search/spending_by_geography/	Returns Spending by state code, county code, or congressional district code
/api/v2/agency/<TOPTIER_AGENCY_CODE>/program_activity/	Returns a list of Program Activity categories for the agency in a single fiscal year
/api/v2/award_spending/recipient/	Returns all award spending by recipient for a given fiscal year and agency id
/api/v2/references/toptier_agencies/	Returns all toptier agencies and related, relevant data.
/api/v2/references/award_types/	Returns a map of award types by award grouping.
/api/v2/references/glossary/	List of glossary terms and definitions
/api/v2/spending/	Returns spending data information through various types and filters

References to Federal Work

Commerce Data Governance Board, "Generative Artificial Intelligence and Open Data: Guidelines and Best Practices," *U.S. Department of Commerce Blog*, January 16, 2025, https://www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices.

Federal Committee on Statistical Methodology (FCSM), *AI-Ready Federal Statistical Data: An Extension of Communicating Data Quality* (Hoppe et al., 2025). https://www.statspolicy.gov/assets/fcsm/files/docs/FCSM.25.03_AI-Ready-Extension-Data-Quality.pdf

Contributors

We would like to thank the following people for their contributions and feedback on this work.

Justin Marsico, Department of Treasury
Anna Mourad, Department of Treasury
Heather Breedlove, Department of Treasury
Amanda Curry, Department of Treasury
Tracy Wright, Department of Treasury
Cyndi Pham, Department of Treasury
Jesrael Lopez-Rosario, Department of Treasury
Grace Lim, Department of Treasury
Justin Cole, Department of Treasury

Kai Cobb, General Services Administration
Drew Keller, General Services Administration
Nolan Harrington, General Services Administration

Dominique Duval-Diop, PhD, Department of Commerce
Zach Palmer, U.S. Digital Corps Fellow at Department of Commerce
Colton Lapp, U.S. Digital Corps Fellow at National Institute of Standards and Technology
Brock Webb, U.S. Census Bureau

Alex Adams, U.S. Digital Corps Fellow at U.S. Department of Agriculture
Chloe Hall, U.S. Digital Corps Fellow at U.S. Department of Agriculture

Travis Hoppe, PhD, U.S. Centers for Disease Control and Prevention

Eduardo Perez, U.S. Digital Corps Fellow at Administration of Children and Families (OTIP)

Moriah Gaynor, U.S. Digital Corps Fellow at Cybersecurity and Infrastructure Security Agency