# Improving How LLMs Handle Education and Workforce Pathway Data

*An Expert-Led Evaluation of AI-Readiness Enhancements*

**Erika Tyagi, Kristin Blagg, Emily Gutierrez**

*October 2026*

**In a baseline evaluation published in May 2026, we found that large language models (LLMs) often gave low-quality responses to questions that required drawing on education and workforce pathway data (Tyagi et al. 2026). In response to these findings, we implemented targeted improvements to two Urban Institute data tools to make them more reliable sources for AI systems: building Model Context Protocol (MCP) servers that give models a structured way to query our data and expanding the documentation that explains how the data should be used. When we asked the same questions with these enhancements in place, subject-matter experts found that response quality improved substantially.**

Our baseline evaluation asked three LLMs 100 curated questions, each written from the perspective of a specific user persona, such as a parent, a Hill staffer, or a career advisor, and focused on topics that the Education Data Portal and the Education-to-Workforce Framework Data Tool are well-suited to answer. Subject-matter experts scored each response across six dimensions, including accuracy, references, and confidence, which captures whether uncertainty and limitations are flagged appropriately. Each dimension was scored on a four-point scale ranging from no issues to critical issues. Across the three models, only 16 percent of responses had no accuracy issues, and evaluators flagged significant or critical issues with references in more than 60 percent. When prompts explicitly named

our tools, models almost always recognized them but were rarely able to access their data correctly. We closed that brief by describing the improvements we planned to make to both tools and committing to evaluate them, and we report those results here.

In August 2026, we ran the same 100 prompts again using Claude Sonnet 4.5, the same model version we tested in the baseline, with the improvements in place. Subject-matter experts again scored the new responses using the same rubric. Across this set of prompts, we observed the following:

- **Response quality improved substantially across every dimension.** Responses with no accuracy issues rose from 16 to 61 percent, and those with critical accuracy issues fell from 35 to 7 percent. Across every dimension, at least three-quarters of responses had no or minor issues in this round.
- **Citations became traceable, and responses flagged limitations more consistently.** The share of responses with no issues on references rose from 11 to 84 percent, and those with no issues on confidence rose from 26 to 71 percent. Specific citations also made the remaining errors easier for our evaluators to identify.
- **Remaining errors were concentrated in judgment rather than retrieval.** In this round, models were usually able to locate and retrieve data reliably. Errors were most common when an answer required more than a lookup, such as choosing between multiple sources that both plausibly applied, using the right denominator to calculate a rate, or combining data across sources that our tools do not link automatically.
- **Responses required more tokens and time to produce.** On average, a response cost roughly 12 times as much to generate and took substantially longer. This reflects the added work of retrieving and processing data rather than meaningfully longer answers.

In the sections that follow, we describe the enhancements we implemented, what the updated evaluation found, and what these results suggest for practitioners and other organizations working to make their data more usable by AI systems. Both MCP servers are publicly hosted (Education Data Portal server, Education-to-Workforce Framework Data Tool server), and our prompt library, evaluation rubric, and source code are on GitHub.

# What We Built and How We Evaluated It

The baseline prompts spanned ten topics and six user personas.[1] We sent each prompt to GPT-5.2, Claude Sonnet 4.5, and Gemini 3 Flash, the models a user would likely have encountered in the free versions of ChatGPT, Claude, and Gemini in early 2026. Each prompt was sent as a standalone query with no custom system prompt, conversation history, or tool access beyond what each model provides by default. In addition to rating each response on the six dimensions[2] using the four-point scale,[3] subject-matter experts classified why a response fell short using a diagnostic framework organized around four steps involved in answering a question with public data: discovering the right source, accessing the data correctly, interpreting it appropriately, and citing it properly.

The diagnostic patterns suggested two improvement tracks. The most common failures involved finding and citing data. Responses often drew on secondary reports rather than underlying datasets, and citations were frequently too vague to trace. When prompts directed models to our tools, responses also often failed to access the data correctly, referencing package names, variable names, and sources that did not exist. These patterns called for more structured ways for AI systems to find, retrieve, and cite data. A second set of failures involved interpretation. Responses that drew on the right data often omitted caveats essential to using a measure or applied it to a population it does not represent. These patterns called for richer documentation of the data itself.

## Creating More Structured Pathways to Our Data

To address discovery, access, and citation failures, we built MCP servers for both the Education Data Portal and the Education-to-Workforce Framework Data Tool. Each server provides a defined set of MCP tools, or specific actions a model can take, such as searching for a dataset or retrieving records, and each comes with a plain-language description of what it does and when it applies. Because MCP is a shared standard, we build and maintain one server for each of our data tools, and any AI application that supports the protocol can connect to it. Anthropic introduced MCP in November 2024 and donated to the Linux Foundation's Agentic AI Foundation in December 2025. Major AI providers have since added support for MCP, though that support varies across products, as we discuss later in this brief. Other public data providers have also begun testing this approach. A recent pilot found that purpose-built MCP servers raised data retrieval accuracy from near zero to 95 percent for two federal datasets (Mendoza et al. 2026).

The Education Data Portal, launched in 2018, brings together data from more than 40 sources, including federal datasets such as the Common Core of Data (CCD) and the Integrated Postsecondary Education Data System (IPEDS), and standardizes them across years and sources. It covers individual schools, districts, and colleges and is updated quarterly as new data are released. The portal makes data available through an application programming interface (API) organized into more than 120 endpoints, each of which returns a specific dataset, such as school-level enrollment from the CCD. Its MCP server provides six tools that follow the steps involved in answering a question with the portal: searching for relevant datasets; describing a dataset's variables and structure; matching the name of a school, district, or college to the ID that identifies it in the data; looking up what coded values mean; retrieving individual records; and retrieving summary statistics. Because the portal uses a consistent structure, the MCP server continues to work as new data are added.

The MCP server also includes guidance on querying the Education Data Portal efficiently, such as reviewing a dataset's documentation to identify which variables can be used as filters and requesting only the variables a question requires. Where relevant, it also directs models to the Portal's summary endpoints, which calculate statistics such as totals and averages on our servers before returning results. We originally built summary endpoints so that users could get simple summary statistics without downloading and processing what can amount to hundreds of millions of raw records on their own computers. A question about how national enrollment has changed over a decade, for example, can be

answered by a single request rather than by retrieving and adding up millions of individual school records. The same design is useful for AI systems, which can process only a limited amount of data at once and whose costs rise with the amount of data they process.

The Education-to-Workforce Framework Data Tool, launched in 2025, serves a different purpose. Rather than organizing data by source, it organizes data around the Education-to-Workforce Indicator Framework, which poses 20 essential questions about how people progress from early education through career and identifies 99 indicators that help answer them. The tool compiles nationally available data for these indicators at the state, county, school district, and neighborhood levels. It draws education data from the Education Data Portal and other data from sources such as the American Community Survey. The Education-to-Workforce MCP server provides four tools: browsing the framework's essential questions, indicators, and metrics; explaining what each one covers; matching a place name (such as a county or district) to the ID that identifies it in the data; and retrieving indicator values for a place.

## Expanding Metadata and Documentation

Using data appropriately also depends on context about what a measure reports, whom it covers, and when it should or should not be used. Baseline responses frequently omitted this context. For example, some treated free and reduced-price-lunch participation as a measure of poverty comparable across states, even though universal school meal programs have made participation an increasingly unreliable measure. Others compared proficiency rates across states that administer different assessments or reported earnings from the Postsecondary Employment Outcomes (PSEO) data as if they represented all graduates in the state.

In each of these cases, the relevant context was documented, but often separately from the data itself, in methodology reports, technical appendices, and other websites, or in the working knowledge of researchers who use the data regularly. Both tools already provided harmonized metadata designed to be human- and machine-readable, and the baseline evaluation helped surface where additional context would be most useful for models.

We addressed these gaps by expanding the metadata and documentation across both tools. For the Education Data Portal, researchers with expertise in each source reviewed and updated the descriptions of its data sources and datasets. For the Education-to-Workforce Framework Data Tool, we added background on the framework's essential questions, indicators, and metrics that was previously spread across separate web pages. We started with the sources and patterns behind errors in the baseline and then made broader improvements along the same lines across the portal's endpoints. The additions included more detailed definitions, information about which populations and institutions the data cover, guidance on which comparisons the data can support, and known limitations. For example, see the following:

- **Why a measure was developed.** The documentation for Model Estimates of Poverty in Schools (MEPS) now notes that it was developed as an alternative to free and reduced-price lunch and

direct certification measures, both of which have become increasingly unreliable as poverty measures across states and over time.

- **How a number is constructed and whom it covers.** The PSEO documentation now specifies that earnings come from unemployment insurance wage records, are reported only for graduates earning above a full-time federal minimum wage threshold in at least three-quarters of the year and are available only for institutions that partner with the program.
- **What comparisons the data cannot support.** The documentation for EDFacts assessment data now states that each state administers its own assessment, so proficiency rates are not comparable across states, and that changes to state assessments or proficiency thresholds can shift reported rates from year to year.

### Evaluating the Enhancements

We designed this round to test whether our enhancements improved response quality, which required holding everything else constant. We used the same 100 prompts, model version, query setup, and scoring rubric as in the baseline. The only changes were the connection to our two MCP servers and the expanded documentation described above. Neither was built for these specific prompts. The servers give models access to all of the data in both tools, and the documentation changes extend across the portal's endpoints. Unlike in the baseline, where evaluators were blinded to which model produced each response, evaluators in this round knew they were scoring enhanced responses from a single model.

We tested one of the three baseline models in this round, Claude Sonnet 4.5, because we could add access to our MCP servers while keeping the rest of its baseline configuration unchanged.[4] To keep the comparison consistent, we compare the enhanced responses with Claude Sonnet 4.5's own baseline responses rather than with the results pooled across all three models, so that differences reflect our enhancements rather than differences between models.[5] Overall scores and patterns in the baseline were broadly similar across the three models, but we did not test whether other models would show similar gains with our enhancements in place. Testing across models and AI applications is a priority for future work. Because we deployed the MCP servers and the expanded documentation together, these results also reflect their combined effect rather than either one alone.

## What We Found

As in our baseline evaluation, we designed the prompts to test specific failure patterns and inform improvements to our tools, not to estimate average performance on a representative set of user questions. Because the prompts focus on questions our tools are well-suited to answer and our improvements responded to baseline failures on these same set of prompts, the gains we observed may be larger than they would be across a more typical set of questions. The figures below should be read as patterns across this deliberately constructed set of prompts rather than as generalizable performance metrics.

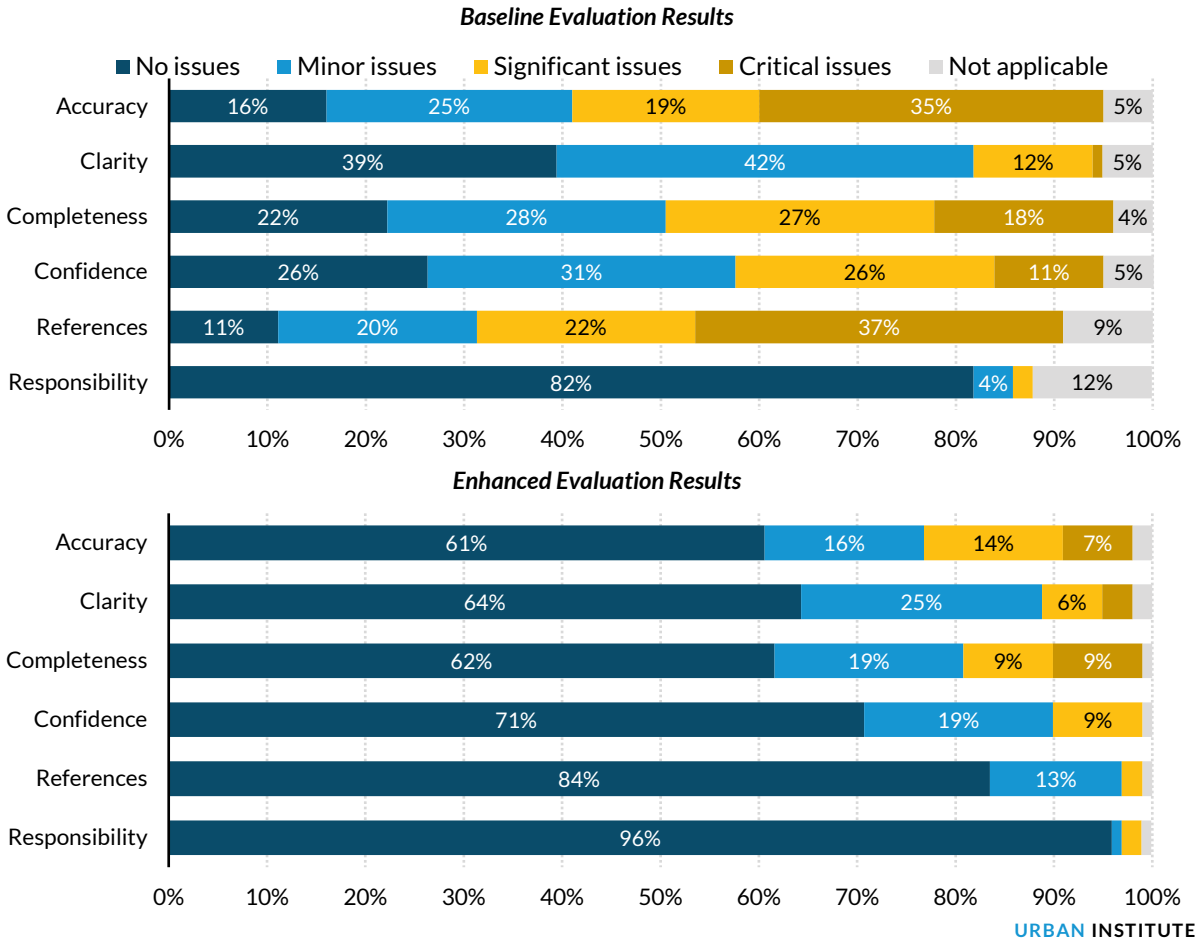
## Response Quality Improved Substantially Across Every Dimension We Scored

Response quality improved on all six dimensions our evaluators scored, and the largest gains came in the areas where the baseline results were the weakest. The share of responses with no accuracy issues rose from 16 to 61 percent, and the share with critical accuracy issues fell from 35 to 7 percent. References and confidence, which we discuss in the next section, improved by similar margins. Apart from responsibility, which was already high, clarity improved the least. Baseline responses were generally well organized and easy to read, which is part of why their errors were hard to spot.

FIGURE 1

### Response Scores Across Six Rubric Dimensions, Baseline and Enhanced Evaluations

*Response quality improved on every dimension, with the largest gains in references, accuracy, and confidence*




**Source:** Author evaluation of 100 LLM responses from Claude Sonnet 4.5 before and after we built MCP servers and expanded the documentation for the Education Data Portal and the Education-to-Workforce Framework Data Tool.
**Notes:** Both evaluations used the same 100 prompts and model version. Each response was scored on six dimensions (accuracy, clarity, completeness, confidence, references, and responsibility) using a four-point scale: no issues, minor issues (small errors that do not undermine usefulness), significant issues (errors that meaningfully affect quality or usability), and critical issues (major errors that render the response unreliable). Figures reflect patterns across a curated set of prompts designed to inform improvements to Urban's data tools and should not be interpreted as generalizable performance metrics.

In our baseline evaluation, response quality declined sharply as prompts moved from general questions to questions about specific places or institutions and to requests for code. One likely reason was that models frequently drew on secondary reports rather than underlying datasets, which were often sufficient for general questions but lacked the detail needed for more specific ones. In this round, 96 percent of responses queried at least one of our MCP servers, and every response that reported figures from them named the source they queried.

This change was most consequential for questions about a particular school, district, or institution, where 42 percent of baseline responses had critical accuracy issues, compared with 11 percent in this round. Our baseline brief described one such case in detail. Prompts about enrollment at Northern Virginia Community College (NVCC) received responses that cited national community college trends and stated that NVCC's enrollment was shrinking, when it had actually been growing. In this round, responses to these prompts retrieved NVCC's reported enrollment data from IPEDS through the Education Data Portal's MCP server. Asked for the percentage change in NVCC's enrollment from 2010 to 2024, for example, the response reported that enrollment rose from 48,996 to 56,457 students, an increase of 15.2 percent, and cited the source of those figures.

The same shift occurred across other questions about specific places and institutions, from a parent asking about bullying at their child's high school to a Hill staffer asking which Virginia community colleges had lost the most enrollment since the COVID-19 pandemic. In these cases, responses retrieved data for the specific schools and colleges the prompts asked about and cited where the figures came from.

Responses to code generation prompts improved as well. In our baseline evaluation, more than 70 percent of responses to code generation prompts had critical accuracy issues, often because the code referenced packages, API endpoints, or variables that did not exist. In this round, 15 percent had critical accuracy issues, and the errors that remained were mostly ordinary bugs rather than references to things that did not exist. One script skipped a merge between two tables, another mishandled missing values, and a third assumed that a parameter available through the API was also available through the R package. A model could likely catch errors like these if it could execute code and correct it iteratively, a capability this evaluation did not test.

## Citations Became Traceable and Models Reported What They Could Not Answer

Two related patterns from our baseline evaluation also improved substantially in this round: errors were difficult to identify because the citations accompanying them were often too vague or unreliable to verify, and responses frequently omitted limitations essential to interpreting the data.

The share of responses with no issues on references rose from 11 to 84 percent, and the share with critical reference issues fell from 37 percent to zero. Much of this improvement reflects how our MCP servers return data. Every result from the Education Data Portal's server includes a block naming the original source, the years it covers, the query's API URL, and a preformatted citation. For a result drawn from IPEDS, for example, the citation reads "Integrated Postsecondary Education Data System, via

Education Data Portal v. 0.26.1, Urban Institute, under the ODC Attribution License." Similarly, the Education-to-Workforce MCP server names the underlying federal source, such as the College Scorecard or Civil Rights Data Collection (CRDC), alongside the framework and the data tool in every result. Because responses drew nearly all their figures from these servers, they could pass these citations along directly. Traceable citations also made the errors that remained considerably easier for our evaluators to identify. In the baseline, evaluators often had difficulty tracing claims back to sources. In this round, responses typically described the tools they called and the data they retrieved before answering, which allowed evaluators to follow the steps a response took and see where it went wrong.

Responses also flagged limitations more consistently. The share of responses with no issues on confidence, which measures whether a response conveys uncertainty and limitations appropriately, rose from 26 to 71 percent, and critical confidence issues fell from 11 percent to 0. The change was especially visible on the six prompts that asked for data that do not exist or are not available at the level requested. In the baseline, responses to these prompts either invented a figure, described the limitation in general terms, or guessed at its cause. In this round, all six identified the specific reason. Asked why PSEO data for Maryland were not available on the Education Data Portal, for example, enhanced response explained that Maryland institutions do not participate in the PSEO program. Asked how the share of Black students borrowing federal loans had changed using Federal Student Aid data, the response explained that the dataset does not report loans disaggregated by race or ethnicity. Both answers drew on the documentation the MCP servers return alongside the data, including the expanded PSEO documentation noting that data are only available for institutions that partner with the program.

## Remaining Errors Were Concentrated in Judgment Calls

Most responses that still had significant or critical accuracy issues in this round retrieved data from our tools without difficulty. Instead, their errors came from decisions about which data to use and how to use it. One such decision was choosing between multiple sources that could each plausibly answer a question. The Education Data Portal includes two sources of institution- and program-level earnings data, PSEO and the College Scorecard, which cover different populations. Asked whether engineering graduates of Purdue University and the University of Southern Indiana earn meaningfully different amounts, one response chose the College Scorecard. Rather than identify earnings for the engineering program, it returned institution-level earnings data for undergraduates and described its figures as graduates' earnings, even though they also include students who did not complete a degree. In the Education Data Portal, only PSEO reports earnings by program, and only for graduates, so it could have answered the question directly. A response about average loan amounts at graduation for California institutions made the opposite choice, passing over the College Scorecard when it was the more appropriate source. Yet when asked directly how the two sources differ, a response explained the distinction accurately, suggesting the model had access to the relevant knowledge but did not apply it when choosing between them.

Other responses used the right source but applied a definition incorrectly or did not properly combine the data a question required. A response comparing borrowing rates at for-profit and public colleges calculated the share of student loan borrowers among students receiving any financial aid rather than among all enrolled students, which overstates borrowing rates and distorts the comparison, since the share of students receiving aid differs across sectors. Another correctly reported student loan deferment rates by county, but inappropriately implied that deferment rates were related to default rates. Deferment is an approved pause in repayment, such as when a borrower returns to school, while default means a borrower stopped repaying. Asked whether urban schools offer more gifted and talented opportunities, a response did not combine data from the CRDC with the CCD, which the Education Data Portal does not link automatically. Answering the question requires both, since the CRDC reports gifted and talented programs and the CCD identifies which schools are in urban areas.

A third set of errors involved the scope of a response, a pattern we also observed in the baseline. Some responses answered a different question than the one asked. A prompt written from the perspective of a school counselor in rural Alabama asked how to use the Education-to-Workforce Framework Data Tool's neighborhood data to understand the barriers students face in getting to college. The response focused on poverty, which the question had not raised, and pointed the counselor to other sources rather than to data the tool holds on graduation rates, counselor availability, and per-pupil spending. Others answered without noting what the data could not support. Asked to compare median student debt for graduates of public and private universities in a state with data from the Education-to-Workforce Framework Data Tool, the LLM response did not note that the data are not available separately for public and private institutions.

## Improved Responses Took Longer and Cost More to Produce

On average, a response in this round cost roughly 12 times as much to generate as a baseline response from the same model, about 11 cents compared with less than 1 cent at Anthropic's published rates for Claude Sonnet 4.5. Responses also took longer, with a median of 37 seconds and 26 of 100 responses taking more than a minute, compared with a few seconds in the baseline.

This difference reflects the work of answering from the data. In the baseline, the model generated each response in a single step from the prompt alone, so the model could produce a fluent, confident answer quickly, whether it was sourced from data. With our MCP servers connected, a response took an average of 7 steps, and as many as 31, as the model searched for the right dataset, reviewed its documentation, matched a school or college name to its ID, and retrieved the relevant records.

Each step is a separate request to the model, and each request includes the full exchange up to that point: the prompt, the descriptions of the available tools, and the data returned so far. AI providers charge based on the amount of text a model processes, measured in units called tokens. In total, a response in this round processed about 170 times as many tokens as a baseline response. Costs rose far less, about 12 times, because providers charge less for text a model reads than for text it writes, and they charge a steeply discounted rate for material repeated across steps that can be cached. About 83

percent of the added cost came from processing the prompt, tool descriptions, and returned data at each step, and only about 4 percent came from longer final answers.

# Lessons for Making Public Data AI-Ready

These results show that data providers can substantially improve how AI systems use their data, an increasingly important goal as more people turn to AI tools to find and interpret information. With direct access to authoritative data and documentation, a model can answer a specific question accurately and with a citation, in far less time than most people would need to find and analyze the data themselves. Without that access, however, the same question can produce a confident answer that is wrong and difficult to check.

The lessons below draw on our experience making targeted enhancements to two tools, but they are likely relevant to any organization that maintains structured data and wants AI systems to use it more reliably, including federal statistical agencies, state longitudinal data systems, city open data portals, research organizations and nonprofits, and organizations building AI tools on their own internal data.

## Evaluations Grounded in Real Questions Informed Where to Invest

Before building, we evaluated how models handled questions written the way people actually ask them. Each prompt reflected the perspective of a particular type of user, such as a parent, an advocate, or a data analyst, so the failures we observed represent issues users would encounter in practice. One prompt, written from the perspective of a parent, asked whether there were fewer kids at a local elementary school than there used to be. The question did not name a data source or identify the school by the ID the data uses, since most parents would not know either. Questions like this showed that a model needs to find the right source and identify the right school before it can retrieve anything, which directly informed the design of the MCP servers, including tools for searching for relevant data and for matching a school, college, or place name to its ID.

The baseline also showed what did not work on its own. Models already recognized our tools by name but could not access their data reliably, so simply making the tools more visible to AI systems would not have been enough to improve access. Classifying each failure by the step where it occurred pointed instead to the investments that we described earlier. Rerunning the same prompts after making changes showed how these investments worked and where more work is needed. For organizations considering similar work, a useful first step is to test how AI tools currently handle the questions their own users ask. This sequence of a baseline assessment, targeted enhancements, and follow-up assessment mirrors an approach the Federal Committee on Statistical Methodology has recommended for federal statistical agencies (Hoppe et al. 2025).

## A Strong Data Foundation Made the Enhancements Easier to Build

Much of what made our data usable by AI systems came from years of earlier work to make it more usable by people. Both the Education Data Portal and Education-to-Workforce Framework Data Tool were built to make large amounts of public data easier to find and use by centralizing data from many sources, standardizing it across years, and documenting it in formats that are both human- and machine-readable. That foundation made the MCP servers much easier to build. Because the Education Data Portal uses a consistent structure, a small set of tools covers all its data and continues to work as new data are added. Similarly, the expanded documentation built on existing metadata rather than starting from scratch.

Our experience is consistent with long-standing principles for publishing data that are findable, accessible, interoperable, and reusable (Wilkinson et al. 2016) and with early federal guidance on preparing public data for AI, which emphasizes detailed metadata and documentation (Commerce Data Governance Board 2025). For organizations without this foundation, investment in it may be the most useful first step, and the benefits run in both directions: the context we added for models also makes our documentation more useful to the people who read it directly. Organizations that publish data through widely used platforms may be able to build on shared work. MCP servers are now available for platforms such as Socrata and CKAN, though some are community-built rather than maintained by the platforms themselves.[6] Google's Data Commons, which brings public statistics from many sources into a common structure, has also extended that foundation to AI systems through its own MCP server.[7]

## Identifying and Resolving Issues Required Subject-Matter Expertise

Only people who know the data well could identify many of the problems we found. For example, recognizing a measure applied to a population it does not cover or a figure drawn from a plausible but outdated source requires knowing how the underlying data are collected, whom they cover, and how they have changed. This was especially true of the errors that remained in this round, which typically cited a real source and reported numbers that matched it. A non-expert reading these responses would likely miss them, and as we noted in our baseline evaluation, automated scoring would need detailed guidance from experts to catch them.

Subject-matter experts were involved at every stage of this work. They wrote prompts based on questions and pitfalls they know from their own work with the data, scored the responses, and translated the patterns they identified into specific updates to our documentation. Organizations considering similar work should plan for this expertise alongside the technical work of building servers and deploying documentation.

## Reaching Users Requires Designing for the Range of Tools They Use

People reach data through AI tools in different ways, so improving AI-readiness requires designing for more than one pathway. Our MCP servers give models the most direct and reliable route to our data,

and the results in this brief show how much that access improved responses. Using that route requires knowing that an MCP server exists and adding it to an AI tool, and support for doing so varies across products and subscription plans.[8] Hosting both servers at a web address makes this easier than running them locally, as many data providers' servers currently require, but it is still a step that many people are unlikely to take on their own.[9] We therefore also improved the pathways AI tools use by default, adding structured metadata to our dataset pages that search engines and AI search tools can read and offering bulk and pre-aggregated downloads. For people who connect AI tools to our data directly, especially through agentic systems, we also provide skill files, which package instructions for using our tools that most AI applications can load.

Even for people who connect to an MCP server, data providers only control part of that experience. A server defines its tools, guidance, and results, but the application that a person uses to connect to it, known as an MCP client, determines how it reaches the model. MCP is still maturing, and clients differ in which parts of the protocol they support. In testing our servers, we also found that some clients passed along the full guidance we provided while others shortened or omitted it, and some truncated large results without notifying users. We designed our servers with these limits in mind and recommend other data providers building MCP servers take similar steps: keeping guidance concise, putting the most important instructions first, limiting the size of results, and testing with the MCP clients that users are most likely to use.

## Beyond This Evaluation

This work also raises important questions that no single data provider can answer alone. As agencies, research organizations, and companies release MCP servers for their own data, people will increasingly connect to several at once. Little is known about how models choose among servers that could each plausibly answer a question, how they reconcile sources that report different figures for similar measures, or how they should communicate those choices to users. Because anyone can build a server that serves public data, users also need ways to judge whether a server, and the answers that cite it, can be relied on, especially since a response that misuses accurate data carries the same authoritative citation as one that uses it correctly. And as AI systems query public data tools more often, the infrastructure behind those tools will need to keep pace. Traffic to the Education Data Portal has increased more than tenfold in early 2026 and continues to grow.

Making public data more reliably usable by AI systems will require action from several groups:

- **AI developers** can work with data providers on how public data are discovered, cited, and kept current in AI tools, and can make that experience more consistent and transparent across products.
- **Data providers and intermediaries** can adopt shared standards for describing, versioning, and citing data so that AI tools can interpret and attribute them consistently, and share evaluation results so that others can learn which investments pay off.

- **Funders** can support the foundational data infrastructure and documentation that AI-readiness depends on, shared evaluation resources and standards that individual organizations would struggle to build alone, and the capacity public data tools need as AI-driven use grows.
- **People using AI tools** can directly connect their tools to authoritative data sources where possible, use custom instructions to ask models to cite sources and indicate when they cannot find data, and check specific claims against the cited source before relying on them.

Our own work is ongoing. Our enhancements substantially improved responses but did not eliminate errors. Some of the problems that remain are within our control, such as adding guidance at the points where models choose between datasets, which population a measure covers, or which denominator a rate should use, and we plan to continue updating both tools as new patterns emerge. Others, such as how AI tools handle connected servers or how models choose among them, depend on AI developers. The field is also changing quickly. During this project, AI providers released many new models, many new ways of connecting AI tools to data became available, and MCP released a major revision that changed how clients and servers communicate.[10] Each of these can change how AI systems use our data, which is why we see evaluation as something to repeat as the field evolves.

Continuing to evaluate this work involves tradeoffs. Expert scoring was essential for identifying the problems described in this brief, but it is labor intensive and difficult to repeat as often as the field changes. More automated approaches, such as using an LLM to score responses, are faster, but reliably catching errors would require detailed guidance from subject-matter experts upfront and ongoing oversight. Priorities for future work include testing our servers across models, AI tools, and advanced workflows, such as agentic systems that can run and correct their own code; using a broader set of questions, since our prompts draw on only a small share of the data in both tools; and developing more scalable approaches that build on the expert evaluation from these two rounds. We hope this work contributes to a shared understanding of what makes public data reliably usable by AI systems and the people who rely on them.

# Notes

1 The ten topics are K–12 enrollment, school poverty, test scores, higher-education enrollment, institution-level earnings, student loans, demographic disparities, neighborhood conditions, college preparation, and postsecondary skills and outcomes. The six user personas are Hill staffer, parent, advocate, career or education advisor, data analyst, and domain researcher.

2 The six dimensions are **accuracy** (whether the factual claims in the response are accurate), **clarity** (whether the response is organized, accessible, and appropriately concise for the intended audience), **completeness** (whether the response covers all components specified in the prompt), **confidence** (whether the response appropriately conveys uncertainty or flags limitations when information is uncertain, unavailable, or outside the tool's scope), **reference** (whether the response cites correct, valid, and appropriately recent sources when references are warranted), and **responsibility** (whether the response avoids bias, toxicity, harmful framing, or offensive content).

3 The four rating levels are **no issues** (fully meets the criterion with no meaningful errors or gaps), **minor issues** (meets the criterion overall, but with small errors, omissions, or imprecisions that do not undermine the response's usefulness), **significant issues** (partially meets the criterion, but with errors or gaps that meaningfully

affect the quality or usability of the response), and **critical issues** (fails to meet the criterion, or the response contains major errors or omissions that render it unreliable for this dimension).

[4] We connected Claude Sonnet 4.5 with no built-in tools, no system prompt, and extended thinking disabled, as in the baseline, so that access to our two MCP servers was the only capability we added. For the other two models, MCP support through each provider's API would have required changing how prompts were sent and processed (for GPT-5.2, moving to OpenAI's Responses API) or was not documented as available (for Gemini 3 Flash). We used the same model version even though newer Claude models had been released by August 2026, so that differences would not reflect model updates. A federal pilot similarly tested its MCP servers with a single Claude Model, Sonnet 4, which the authors describe as the only model that supported MCP integrations at the time (Mendoza et al. 2026).

[5] In the baseline, 16 percent of Claude Sonnet 4.5's responses had no accuracy issues and 35 percent had critical issues, compared with 16 and 34 percent across all three models.

[6] See, for example, "Socrata MCP Server," GitHub, last updated November 5, 2025, https://github.com/Thomas-TyTech/Socrata-MCP; "CKAN MCP Server," GitHub, last updated October 6, 2026, https://github.com/ondata/ckan-mcp-server.

[7] "Query data interactively with an AI agent," Data Commons by Google, accessed September 27, 2026, https://docs.datacommons.org/mcp/.

[8] As of September 2026, ChatGPT supported custom MCP servers only through developer mode, which was not available on its free plan, and connected directly only to remotely hosted servers; Gemini's custom MCP connections were available only to users 18 or older in the United States with personal Google accounts; and Claude's free plan allowed only one custom connector. See "Developer Mode and MCP Apps in ChatGPT," OpenAI Help Center, accessed September 27, 2026, https://help.openai.com/en/articles/12584461-developer-mode-and-mcp-apps-in-chatgpt; "Connect & Manage Custom Apps for Gemini Apps," Gemini Apps Help, Google, accessed September 27, 2026, https://support.google.com/gemini/answer/17209137?hl=en&co=GENIE.Platform%3DDesktop; "Get Started with Custom Connectors Using Remote MCP," Claude Help Center, accessed September 27, 2026, https://support.claude.com/en/articles/11175166-get-started-with-custom-connectors-using-remote-mcp.

[9] For example, the Census Bureau's official MCP server is installed by downloading its code and running it locally. See "US Census Bureau Data API MCP," US Census Bureau, GitHub, last updated March 6, 2026, https://github.com/uscensusbureau/us-census-bureau-data-api-mcp.

[10] "The 2026-07-28 Specification," *Model Context Protocol Blog*, July 28, 2026, https://blog.modelcontextprotocol.io/posts/2026-07-28/.

# References


Commerce Data Governance Board. 2025. "Generative Artificial Intelligence and Open Data: Guidelines and Best Practices." US Department of Commerce.

Hoppe, Travis, Jeffrey Gonzalez, Lisa Mirel, and Rolf Schmitt. 2025. "AI-Ready Federal Statistical Data: An Extension of Communicating Data Quality." FCSM 25-03. Federal Committee on Statistical Methodology.

Mendoza, Bella, Mark Aronson, Haley Johnson, Sam Levy, and Mohammad Arifur Rahman. 2026. "Improving LLM Access to Federal Open Data: A Pilot Study of Model Context Protocol Servers for Federal Data Access." US Digital Corps.

Tyagi, Erika, Kristin Blagg, and Emily Gutierrez. 2026. "Evaluating How LLMs Handle Education and Workforce Pathway Data: An Expert-Led Evaluation of Responses from Three Models Across Six User Personas and Ten Topics." Urban Institute.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

# About the Authors

**Erika Tyagi** is the associate director of data science products at the Urban Institute, where she develops data science tools that advance innovative policy research, including open data portals, interactive data tools, and privacy-enhancing technologies. She holds a BA in economics and international relations from Carleton College and an MS in computational analysis and public policy from the University of Chicago.

**Kristin Blagg** is a principal research associate in the Work, Education, and Labor Division at the Urban Institute. Her research focuses on K-12 and postsecondary education. Blagg has conducted studies on student transportation and school choice, student loans, and the role of information in higher education. Blagg holds a BA in government from Harvard University, an MSEd from Hunter College, an MPP from Georgetown University, and a PhD in public policy and public administration from the George Washington University.

**Emily Gutierrez** is a senior research associate in the Work, Education, and Labor Division, where her research focuses on K-12 education policies that aim to mitigate educational inequities, disparities, and external factors that affect student success. She received her BBA in economics, Spanish, and public administration and MS in economics from Baylor University, and she holds a PhD in public administration from Syracuse University.

# Acknowledgments

This brief was funded by the Gates Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We would also like to thank Shana Metcalf, who helped review and synthesize evaluation results. We are also grateful to Maggie Reeves and Alena Stern for their input throughout the evaluation and for their thoughtful feedback on this brief.




## ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization founded on one simple idea: To improve lives and strengthen communities, we need practices and policies that work. For more than 50 years, that has been our charge. By equipping changemakers with evidence and solutions, together we can create a future where every person and community has the opportunity and power to thrive.

Copyright © October 2026. Urban Institute. Permission is granted for reproduction of this file, with attribution to the Urban Institute.