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

Erika Tyagi, Kristin Blagg, and Emily Gutierrez

May 2026

AI tools are rapidly becoming one of the primary ways people find and interpret information.[1] When a parent wants to understand how their local school is performing, a Hill staffer needs data for a briefing, or a career counselor is helping a student compare postsecondary options, they are increasingly turning to tools like ChatGPT, Gemini, or Claude rather than going directly to established data sources. These tools have the potential to make data more accessible, but we know relatively little about how well they handle questions that draw on public data and whether they can find authoritative sources, use the data correctly, and communicate information in a way that serves its users. To help fill this gap, we conducted a structured evaluation in which subject-matter experts assessed how three large language models handled 100 curated questions about education and workforce pathway data.

Across models and topics, we observed several patterns in how models handled questions grounded in public data:

- **Response quality varied widely by task type.** Quality varied substantially depending on what a prompt asked the model to do. Models performed well when asked to explain concepts or

describe widely used data sources in broad terms, but quality declined significantly when tasks required the model to locate, retrieve, and correctly use specific data and declined further for prompts that asked for working code.

- **Errors were often difficult to identify.** When responses did contain errors, they were often difficult to detect. Inaccurate claims were presented with the same apparent certainty as correct ones, and the citations accompanying them were often too vague or unreliable to enable efficient verification.
- **Models often misinterpreted questions.** Models frequently answered different questions than asked, defaulting to national trends when asked about specific schools or institutions.
- **Key limitations and caveats were frequently missing.** Responses often omitted caveats essential to interpreting data sources appropriately or produced answers when the correct response was that the data did not exist or was not yet available.
- **Models struggled to use Urban's data tools accurately.** Even when prompts explicitly directed models to use Urban's data tools containing relevant information, models rarely accessed the data correctly—often fabricating API endpoints, package names, and variable names in well-formatted code.

Improvements based on these findings are being implemented in two Urban Institute tools: the Education Data Portal and the Education-to-Workforce Framework Data Tool. These tools were initially designed to make large amounts of public data more accessible to *human* users by centralizing and harmonizing data and metadata. We are now extending that foundation so these tools can serve as reliable data sources for AI systems as well.

Our prompt library, evaluation rubric, and source code are available on GitHub. Our methodology is designed to be adaptable across domains and contexts. This approach can be tailored by a range of entities, including federal agencies maintaining national surveys, state longitudinal data systems, city open data portals, and organizations working in public health, housing, or criminal justice to curate prompts grounded in their users' questions and score against their own authoritative sources. We hope this is useful to the broader community of data providers working to ensure their data is used reliably and responsibly as AI adoption accelerates.

How We Designed the Evaluation

We designed this evaluation to surface specific, actionable patterns that could directly inform improvements to Urban's data tools. Several design choices shape the approach: prompts grounded in how people actually ask about education and workforce pathway data, a two-layer rubric that captures both overall response quality and the specific reasons a response fell short, and model configurations designed to approximate the default experience most users encounter.

Curating Prompts Across Ten Topics and Six User Personas

We developed a suite of 100 prompts spanning ten topics that people commonly turn to education and workforce data to answer: 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. Together these topics span students' trajectories from early education through their careers and the contextual conditions that shape those pathways. Urban's researchers bring deep expertise in these topics, both in the types of questions users ask and in identifying and interpreting the authoritative data sources to answer them. We also curated prompts to focus on questions that data from the Education Data Portal and the Education-to-Workforce Framework Data Tool are well-suited to answer.

Each prompt was written from the perspective of one of six user personas, representing a range of people who ask about education and workforce pathway data in ways that reflect their roles, goals, and backgrounds:

- **Hill staffer**: congressional staff preparing briefings or talking points, likely to ask comparative, policy-relevant questions, often requesting official federal sources
- **Parent**: a parent researching schools, school choice, or local education context, likely to ask location-specific, practical questions about their children's schools and options
- **Advocate**: a policy advocate building a case for legislators or the public, often interested in compelling data points or disparities
- **Career or education advisor**: an advisor helping learners navigate education and career pathways, often interested in outcome-oriented information to help students make decisions
- **Data analyst**: an analyst without education domain expertise working with education data, likely to ask code-oriented, technical questions or request specific data for analysis
- **Domain researcher**: a researcher with education policy expertise, likely to ask precise, methodologically aware questions, including requests for specific data or code

These personas differ not just in *what* they ask but in how they phrase questions, the level of technical detail they expect, and what a useful response looks like. Drawing from our prompt library, for example, a Hill staffer building a national briefing might ask, "How much has national school enrollment changed since the pandemic? Use official federal statistics." A parent in Dallas might ask, "Are there fewer kids at Lakewood Elementary than there used to be?" And a data analyst might ask, "I'm looking for K–12 enrollment data by race/ethnicity. Should I use the Common Core of Data (CCD) or the Civil Rights Data Collection (CRDC) data?" Although these questions draw on overlapping data, each requires the model to draw on a different set of capabilities—from translating a question into the right data sources to calibrating the level of detail and context in its response.

Scoring by Subject-Matter Experts Using a Two-Layer Rubric

Subject-matter experts scored all responses directly, rather than delegating to LLM-as-judge or other automated approaches. Many of the failure modes we wanted to surface, such as citing a plausible but outdated data source or using a measure without noting its known limitations, require domain expertise to recognize. Reliably delegating that judgement to an LLM or more automated system would require extensive, iterative development of detailed evaluation guidance that was out of scope at this stage.

We scored each response using a two-layer rubric. The first layer measures overall response quality across six dimensions: accuracy, clarity, completeness, confidence, reference, and responsibility.[2] Each dimension was rated on a four-point scale: no issues, minor issues, significant issues, and critical issues.[3] Evaluators applied the same rubric to every response, but they interpreted some dimensions considering the persona the prompt was written for. For example, a clear and useful response for a parent differs from what counts as clear for a domain researcher.

The second evaluation layer serves as a diagnostic taxonomy: When a response fell short, evaluators classified *why* by tagging the specific issues that occurred. These diagnostic tags are organized around four stages of how an LLM serves a response from public data:

- Whether a model could **discover** the right source
- Whether it could **access** the data correctly
- Whether it **interpreted** what it found appropriately
- Whether it **cited** its sources properly

This structure was designed to make the evaluation actionable: different diagnostic patterns point toward different interventions to improve the AI-readiness of the Education Data Portal and Education-to-Workforce Framework Data Tool, as described in a later section.

Replicating the Default User Experience

We ran each of the 100 prompts through three LLMs: GPT-5.2, Claude Sonnet 4.5, and Gemini 3 Flash, producing 300 responses for evaluation. These reflect the models a user would likely encounter when interacting with the free versions of ChatGPT, Claude, and Gemini, respectively, as of early 2026. Most people currently use these tools out of the box, without custom system prompts, retrieval augmentation, plugins, or other advanced configurations, and while that may change as AI workflows evolve, it describes how the personas in our target audience overwhelmingly interact with these systems today. To approximate that default experience, we queried each model programmatically, sending each prompt as a standalone query with no custom system prompt, conversation history, or tool access beyond what each model provides by default. Evaluators were blinded to which model produced each response.

What We Found

Importantly, we designed this evaluation to produce a structured understanding of how LLMs respond to curated questions drawing on education and workforce pathway data, not to rank specific models, serve as a quantitative benchmark, or make broad claims about LLM capabilities. We intentionally designed prompts to test specific failure patterns and inform improvements to Urban's data tools, rather than to estimate average performance on a representative sample of user questions. As a result, the figures we report below reflect observations across this deliberately constructed set of prompts and should be read as patterns, rather than generalizable performance metrics.

Overall scores and patterns were broadly similar across the three models we tested, and we report findings in aggregate below. The one notable model-specific observation is that GPT-5.2 responded with clarifying questions rather than attempting an answer more often than the other models. This is not necessarily the wrong behavior, as declining to answer an ambiguous question is often preferable to answering it incorrectly, but it means that GPT-5.2 received fewer accuracy flags in part because it committed to fewer answers, making direct cross-model comparisons difficult to interpret.

Response Quality Varied Substantially by What a Prompt Asked the Model to Do

When prompts asked models to explain concepts, define terms, or broadly describe where to find data, responses were often accurate and useful. For example, all three models produced high-quality responses when asked to explain the difference between fall enrollment and full-time equivalent (FTE) enrollment in Integrated Postsecondary Education Data System (IPEDS), or when asked where to find reliable national data on undergraduate borrowing.

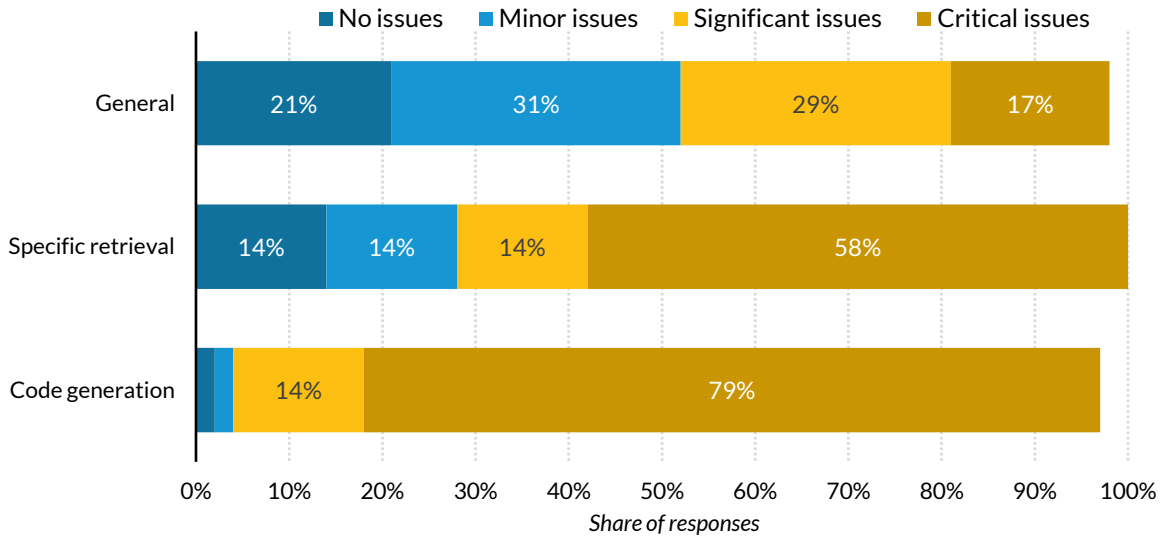
Response quality decreased significantly for tasks that required providing specific numbers, verifying claims, or comparing values across sources, and decreased further for technical tasks such as writing working code or constructing API calls. This pattern appeared consistently across topics and models: The more a prompt required the model to locate, retrieve, and correctly use specific data rather than summarize general knowledge, the more likely the response was to contain errors.

Figure 1 below illustrates this gradient by reporting the distribution of accuracy scores based on the type of task that each prompt asked the model to do. For general questions (defining terms, identifying data sources, or summarizing national or state trends), about one in five responses had no accuracy issues, but even in this category nearly half had significant or critical errors. For specific retrieval questions (about a particular school, district, or institution), the critical issue rate more than tripled to 58 percent. For code generation tasks (producing code, API calls, or downloadable artifacts), 79 percent of responses contained critical accuracy issues.

FIGURE 1

Accuracy Scores by Prompt Type

Response quality declined as prompts required models to locate and use more specific data



URBAN INSTITUTE

Source: Author evaluation of 300 LLM responses across GPT-5.2, Claude Sonnet 4.5, and Gemini 3 Flash.
Notes: Prompts were grouped into three categories based on what they asked the model to do: general (defining a term, distinguishing between datasets, identifying where to find data, or summarizing national or state-level trends), specific retrieval (retrieving or comparing values for a particular school, district, or institution), and code generation (producing code or downloadable artifacts). Each response was scored on 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.

One likely reason response quality tracked so closely with task type is that models frequently drew on secondary reports rather than underlying datasets. These secondary sources were often sufficient for general questions but lacked the specificity to answer questions about particular places, institutions, or time periods (e.g., school-level enrollment or program-specific earnings at a particular institution). Even when prompts steered models toward specific data sources—by naming the source directly (e.g., "Use the latest CCD enrollment data") or by prompting the model to use the Education Data Portal or generate code to pull data from it—this did not consistently improve response quality. Instead, models often fabricated API endpoints, variable names, and package names, or named specific sources in prose while reporting figures that did not actually come from them.

The personas in our evaluation tended to ask different kinds of questions, and the patterns in response quality across personas largely reflect this variance. Data analysts and domain researchers, whose prompts more frequently involved code generation and specific data retrieval, saw the highest rates of issues in our evaluation. Hill staffers, whose prompts were more often conceptual and general retrieval questions, saw fewer. Parents asked a mix of general and location-specific questions. Models

handled the broader questions well but struggled with the specific local ones, a pattern consistent with the task-type gradient described above.

Citations and Sourcing Were Frequently Unreliable

Evaluators flagged significant or critical issues in the references dimension in more than 60 percent of responses. Citations were often vague, pointed to content that did not support the numbers attributed to them, or were missing entirely. Evaluators encountered citations to reports that did not exist or had been superseded years earlier and attributions of datasets to the wrong source organization (e.g., attributing Urban's Model Estimates of Poverty in Schools [MEPS] data to the Census Bureau or National Center for Education Statistics [NCES]). A related pattern was the fabrication of navigation paths: Responses directed users to a "Data and Analytics" tab on a state agency website that had no such section, or to a dashboard at a URL that did not exist. In several cases, models cited specific reports that are no longer published (e.g., multiple responses cited an annual student debt report from The Institute for College Access & Success that has not been published since 2021, suggesting the model learned it as a canonical reference and had not updated as the series ended).

LLMs routinely advise users to verify the information they provide. Verification, however, depends on citations that are specific enough to trace, and the citations in our evaluation frequently were not. Our evaluators often had difficulty tracing claims back to their sources, particularly for responses that combined figures from multiple sources. In our evaluation, the presence of a citation in a response was also not associated with higher accuracy. Responses that included URLs or source attributions were no more likely to be accurate than those without them. For users without domain expertise, this verification would be substantially more time-consuming and difficult.

Errors Were Common and Presented with the Same Confidence as Correct Information

Across our prompt set, only 16 percent of responses had no accuracy issues, and 34 percent had issues flagged by our evaluators as critical, meaning the response contained major errors that rendered it unreliable. In one example, a response provided Purdue engineering graduate earnings figures that were tens of thousands of dollars below the actual values reported in the College Scorecard. In another, a response reported student debt figures for Loyola Marymount University that did not match the institution's reported values. Accuracy problems were especially pronounced for newer data sources with less documentation. Responses involving MEPS and Postsecondary Employment Outcomes (PSEO) were consistently the most error-prone, both being newer datasets with fewer published reports and secondary analyses compared to longer-standing sources like CCD or IPEDS.

Inaccurate statistics were typically presented in the same tone, precision, and level of apparent certainty as accurate ones. Several responses produced specific dollar amounts, enrollment counts, or percentages that did not match actual reported values but were presented in formats that signal authority, rounded to the nearest hundred dollars, displayed in clean tables, or attributed to specific datasets with language like "according to the most recent release." This presentation made errors especially difficult to detect without independent verification. Because the citations accompanying

these claims were often themselves unreliable, as described above, verifying a given response would require significant time and effort.

Models Often Answered a Different Question Than the One Asked

Two related patterns appeared frequently across responses. The first was scope substitution: when prompts asked about specific places, institutions, or cases, models frequently defaulted to broader patterns instead. For example, a prompt about enrollment at Northern Virginia Community College (NVCC) received a response that cited national community college trends. This was most consequential when the specific answer ran counter to the aggregate. NVCC's enrollment has actually been growing, but responses across models stated it was shrinking, consistent with the broader national decline in community college enrollment.

The second was premise acceptance: when a prompt contained an embedded assumption, models often accepted it rather than checking or pushing back. For example, a prompt asking about declining enrollment at a Dallas elementary school produced responses that offered several reasons for the supposed decline, even though the school's enrollment had been roughly stable.

Responses Often Omitted Important Caveats and Limitations

Several of the topics we covered required not just retrieving a number but contextualizing the limitations of how the data should be used. For example, responses to questions about school poverty frequently relied on free- and reduced-price lunch data without noting that universal school meal programs have made this measure increasingly unreliable as a poverty proxy over time, one of the gaps MEPS was designed to address.[4] Responses to questions about institution-level earnings sometimes used PSEO data without noting that specific states are not represented or that the data covers only full-time workers.[5] In these cases, responses could be misleading for the specific question being asked. This pattern was most common for data sources whose appropriate use depends on limitations not documented in the data itself or in the reports that summarize it.

A related pattern emerged for prompts where the correct response required acknowledging that the requested data does not exist, is not yet available, or should not be used for the proposed purpose. For example, prompts asked about CCD data for a school year that had not yet been released, or about PSEO data for states that are not represented in the dataset. In these cases, models consistently defaulted to producing an answer rather than naming the absence, even when the absence was the most important thing for the user to know.

Pointing Models to Our Tools Did Not Inherently Improve Results

The Education Data Portal and the Education-to-Workforce Framework Data Tool were designed to make education and workforce pathway data more accessible by centralizing data from across sources, harmonizing it across years and geographies, and providing detailed human- and machine-readable documentation. We expected this foundation to improve the quality of LLM responses when prompts pointed models toward these tools, but this was not the case. When prompts explicitly named these

tools (e.g., "Use the Urban Institute's Education Data Portal"), models almost always recognized them but were rarely able to access their data correctly or guide users to do so.

This was especially evident in prompts that asked models to generate code. Of the 30 prompts that specifically steered the model to use our tools, more than half asked to produce R, Python, or Stata code to pull data from them. Of these, more than 80 percent contained issues categorized as critical. Across all three models, responses cited package names that do not exist, fabricated API endpoints and variable names, used the wrong dataset (e.g., pulling SAIPE data when asked for MEPS), and confused calendar conventions (e.g., treating 2020 as the 2020–21 school year in CCD, where it represents the 2019–20 year). These errors were typically presented in well-formatted, well-commented code, making it difficult to distinguish code that *looks* correct from code that *is* correct, which is especially relevant given that code generation is a use case where people are already widely adopting AI tools.

These patterns raise practical concerns beyond accuracy. When a model produces code that references a fabricated package or calls a nonexistent API endpoint, users who attempt to run it may encounter unexpected errors or, in some cases, install unverified packages that pose security risks. And when a response names the Education Data Portal or the Education-to-Workforce Framework Data Tool as its source while reporting figures that did not actually come from them, it risks undermining trust in the tools themselves, even when the tools contain the correct data. These patterns shaped the improvements we describe in the next section.

How We Are Using These Findings

We designed this evaluation to serve as a baseline and inform AI-readiness improvements to the Education Data Portal and the Education-to-Workforce Framework Data Tool. Launched in 2018, the Education Data Portal provides API-first access to harmonized data and metadata across more than 120 endpoints from over 40 sources. The Education-to-Workforce Framework Data Tool, launched in 2025, centralizes nationally available data across 99 indicators and 20 essential questions about how students progress from early education through career and the system conditions that shape those trajectories. Both tools already centralize, harmonize, and document the kinds of data the models in our evaluation often did not locate or use correctly, providing a strong foundation to build on.

Different Failure Types Point Toward Different Interventions

The diagnostic layer of our rubric, which categorized issues by whether they involved discoverability, access, interpretation, or citations, was designed to inform specific improvements and to provide a framework for organizing our response. Discoverability and citation failures, the most common patterns in our evaluation, suggest that models need more structured pathways to reliably find and reference authoritative data. Access and interpretation failures suggest that models need richer metadata and documentation to use data correctly once they find it.

We are pursuing improvements along both tracks, additionally informed by emerging efforts from federal statistical agencies, city governments, and cross-domain data platforms to make public data

more usable by AI systems. These efforts are still early, with limited empirical evidence about which practices produce meaningful improvements, and we see our work as both contributing to and testing these emerging approaches.

Creating More Structured Pathways for AI Systems to Find and Reference Data

To address discoverability and citation failures, we are creating more structured pathways for AI systems to find and reference the data in our tools. The most direct approach is developing a model context protocol (MCP) server. MCP is an open standard that lets data providers expose a defined set of tools that AI systems can call directly, each with a plain-language description of when it should be used and what inputs it requires. A recent pilot found that purpose-built MCP servers raised data retrieval accuracy from near zero to 95 percent for two federal datasets, suggesting this approach can substantially close the retrieval gap (Mendoza et al. 2026).

However, MCP is a relatively new standard that introduces security and governance challenges, along with ongoing investments to maintain as underlying data and tooling evolve. Crucially, it also requires users to know the protocol exists, locate relevant servers, and connect through a compatible MCP client, feasible for technical power users but likely impractical for many of the personas in our evaluation who typically ask questions through the default ChatGPT, Claude, or Gemini interfaces without MCP configurations. To better support these users, we are also exploring ways to improve how our data is discovered through the search and indexing channels that general-purpose AI systems rely on by default.

Expanding Metadata and Documentation Layers

To address access and interpretation failures, we are also expanding the metadata and documentation layer across both tools. This includes more detailed variable definitions, guidance on appropriate comparisons, known limitations, and information about which questions different sources can and cannot reliably answer, effectively incorporating more subject-matter expertise directly into the data and metadata layers themselves through natural language. These improvements serve both AI systems and human users: Richer documentation helps models interpret data correctly, and it helps researchers, analysts, and practitioners better understand and use the data on their own. This work also matters regardless of whether users reach the data through MCP, web search, or a model's training data. These efforts directly address the patterns we observed where models located our tools but used them incorrectly, confusing variable names, misapplying calendar conventions, or omitting caveats essential to interpreting the data appropriately.

Ensuring Infrastructure Can Handle AI-Driven Demand

In the early months of 2026, traffic to the Education Data Portal increased more than tenfold and continues to rapidly grow. As AI systems increasingly scrape and query public data tools and platforms, the volume and pattern of requests differ from traditional human use: more frequent, more programmatic, and less predictable. Scaling the underlying infrastructure to ensure our tools remain reliable and responsive under these conditions is a necessary complement to the discoverability, access, and interpretation improvements described above.

Sharing What We Learn

Once these updates are in place, we plan to report results from a post-intervention evaluation to share what we learn and contribute evidence about which approaches produce measurable improvements. Data providers and funders face open questions about where the highest-impact investments in AI-readiness lie, whether in improving the underlying data infrastructure for AI systems, building user-facing resources that help specific audiences get more out of existing tools, or some combination. Our goal is to support this broader ecosystem of public-interest data providers and to contribute shared evidence about what works as the field and technology continue to develop.

Beyond This Evaluation

Our evaluation is one of the first to systematically assess how AI systems handle a specific domain of public data using expert scoring, and we designed the approach to be useful to others doing related work across organizations and domains. While the specific patterns we observed reflect the data landscape of education and workforce pathways, the underlying challenges around discoverability, access, interpretation, and citations are likely relevant to any organization that maintains structured, documented public data and wants to ensure AI systems use it correctly.

We also recognize that models, access modes, and the workflows people use to interact with these systems are changing quickly. New models are released frequently, and newer versions may perform better on many of the tasks we tested. Our findings reflect a snapshot in time, not a permanent assessment of these tools' capabilities. We also designed this evaluation to approximate the experience most users encounter by default without intentional prompt engineering, retrieval augmentation, or tool use. But practitioners are increasingly interacting with LLMs through more deliberate prompting, code-enabled tools, agentic workflows, and other mechanisms that can likely produce meaningfully different responses (Nigenda et al. 2026).

In future work, we hope to build on this approach in two ways: by creating recurring benchmarks that track how performance on policy-relevant data questions evolves as models, tools, and underlying data sources change over time, and by more rigorously evaluating how AI systems perform across these different interaction modes, including in real research and analytic workflows rather than isolated prompts. As these tools become more deeply embedded in how people find and use public data, ongoing evaluation will be essential to realizing their potential to make public data more accessible and useful.

Notes

[1] Matt O'Brien and Linley Sanders, "How US Adults Are Using AI, According to AP-NORC Polling," AP News, July 29, 2025, https://apnews.com/article/ai-artificial-intelligence-poll-229b665d10d057441a69f56648b973e1.

[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] "Model Estimates of Poverty in Schools," Urban Institute, accessed May 4, 2026, https://www.urban.org/projects/model-estimates-poverty-schools.

[5] "Longitudinal Employer-Household Dynamics," Post-Secondary Employment Outcomes, United States Census Bureau, accessed May 4, 2026, https://lehd.ces.census.gov/data/pseo_experimental.html.

References

Nigenda, Andrés, Camille Shao, and Dallas Dotter. 2026. "Trusting Your LLM: Building a Framework for Evaluating LLMs on Complex Statistical Data Sets." Mathematica.

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

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 Rachel Lamb and Katie Pullom, who helped shape the evaluation design and collaborated throughout, including developing prompts and evaluating responses. We are also grateful to Alena Stern and Maggie Reeves for their valuable input throughout the evaluation and insightful comments on earlier drafts of 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 © May 2026. Urban Institute. Permission is granted for reproduction of this file, with attribution to the Urban Institute.