



THE EFFECT OF MINIMAL AND MAXIMAL METADATA CONTEXT ON AI RESPONSE

FCSM-26-02 | September 2026

Prepared by Brandon Kopp, Carla Medalia, and Brian Quistorff

This brief summarizes the work of the Federal Committee on Statistical Methodology (FCSM) AI-Ready Data Working Group, whose expertise was essential to the development and completion of this research.

MEMBERS OF THE FCSM AI-READY DATA WORKING GROUP

Justin Cole

Bureau of Fiscal Service, Department of Treasury

Matt Jans

National Center for Health Statistics

Tiffany Julian

National Center for Science and Engineering Statistics

Luke Keller

U.S. Census Bureau (formerly)

Brandon Kopp

Bureau of Labor Statistics

David Laird

Office of Homeland Security Statistics, DHS

Caiyi Lang

Statistics of Income, IRS

Janice Lent

U.S. Energy Information Administration

Carla Medalia

U.S. Census Bureau

Clay Moulton

Statistics of Income, IRS

Tina Norris

National Center for Health Statistics

Brian Quistorff

Bureau of Economic Analysis

Ben Rogers

Centers for Disease Control and Prevention

RECOMMENDED CITATION

Brandon Kopp, Carla Medalia, and Brian Quistorff. 2026. *The Effect of Minimal and Maximal Metadata Context on AI Response*. (Prepared for the FCSM AI-Ready Data Working Group.) FCSM 26-02. Federal Committee on Statistical Methodology. September 2026. <https://doi.org/10.21949/6txs-6p32>

BACKGROUND AND OBJECTIVE

As Artificial Intelligence (AI) systems and tools proliferate, more people are using them as a primary source of information. These systems now serve as intermediaries between the public and the statistical data provided by the federal statistical system. The Federal Committee on Statistical Methodology (FCSM) seeks to help agencies improve the AI-readiness of their data products and make them not only machine readable but also machine understandable. One crucial aspect of AI-readiness is the availability, amount, and comprehensiveness of metadata that accompanies statistical products.¹

While general guidelines suggest that providing comprehensive metadata enables AI systems to provide accurate responses, metadata can be costly to produce and provide, and there is a lack of evidence showing *how much* metadata is needed for commercial AI services to answer user queries about federal statistical data products in an accurate, contextualized way. In addition, providing more metadata than is necessary could dilute a model's attention and lead to suboptimal response.

This project fills this gap by examining how the amount of metadata provided to a commercial generative AI service can affect its ability to provide accurate and contextualized responses to queries about federal statistical data.

¹ [Generative Artificial Intelligence and Open Data: Guidelines and Best Practices](#)

STUDY DESIGN: DATA AND METHODS

To test this objective, we varied several features of an AI chat interaction: the survey source of the estimate, amount of metadata provided with the prompt, and AI response mode. We then performed qualitative analyses on the model output.

For the study, we used Microsoft Copilot (GPT5) to respond to a query about educational attainment. Specifically, we asked: "What percent of people in the United States have at least a bachelor's degree?" twelve different times, varying three factors of the prompt (see below). In addition, in every test, we included an extract of a publicly-available statistical table that contained the answer to the question. That is, we simulate a process in which an agency is able to return correct data and metadata for a user's query and investigate how a commercial chat model makes use of the provided metadata to provide a contextualized response alongside the correct statistical estimate (see Figure 1).²

1. **Survey source of the estimate.** We provided the LLM with educational attainment data and metadata from either the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) or the American Community Survey (ACS).

This condition allowed us to focus the answer on a statistical product that we know came from a trusted source (the U.S. Census Bureau) and established benchmarks against which we could validate that the LLM produced an accurate response. It also allowed us to vary the underlying data and metadata while still providing the same amount of information in our amount of metadata condition.

2. **Amount of metadata.** We provided the LLM with varying amounts of metadata relevant to the user query along with the question prompt. We compared providing the LLM with no metadata, minimal metadata, and extensive metadata.

The no metadata condition meant that we just asked the LLM to provide a response to the question "What percent of people in the United States have at least a bachelor's degree?" As noted above, we include the table with the correct answer, which contained cell values, column headers, and row headers, but no other metadata.

The minimal metadata condition included basic, table-level metadata describing the table name, which included the survey, reference year, population information (in both cases U.S. residents age 25 and older), and a link to the source data/table.

The extensive metadata condition included the above plus manually-gathered survey- and question-level metadata sourced from publicly-available technical documentation, which resulted in a roughly 10-fold increase in the word count.

This condition enabled us to evaluate whether the accuracy and comprehensiveness of the LLM response depended on the amount of metadata included as part of the question prompt. Specifically, we could evaluate whether providing any metadata improved the response over no metadata, and whether providing extensive metadata improved the response over minimal metadata.

² We make this assumption because (1) methods for identifying the correct data and metadata will likely be agency-specific, (2) we did not have the resources on the team to build an AI-based data retrieval system, and (3) we want to provide a view of the end state of a fully-formed retrieval augmented generation process to show the value of investing time in aggregating metadata and building AI retrieval systems.

3. **AI response mode.** Finally, we varied the Copilot response mode by using either Copilot's standard response mode or think deeper mode. These two modes use the same AI model (GPT5) but provide different system prompts about how to perform the text generation task.

Copilot's standard response mode is the default and prioritizes speed over complexity. Copilot's think deeper mode, on the other hand, prioritizes complexity over speed. Users may select either mode, but since the standard response mode is the default and the question is not complex, users may be more likely to use the standard response mode default.

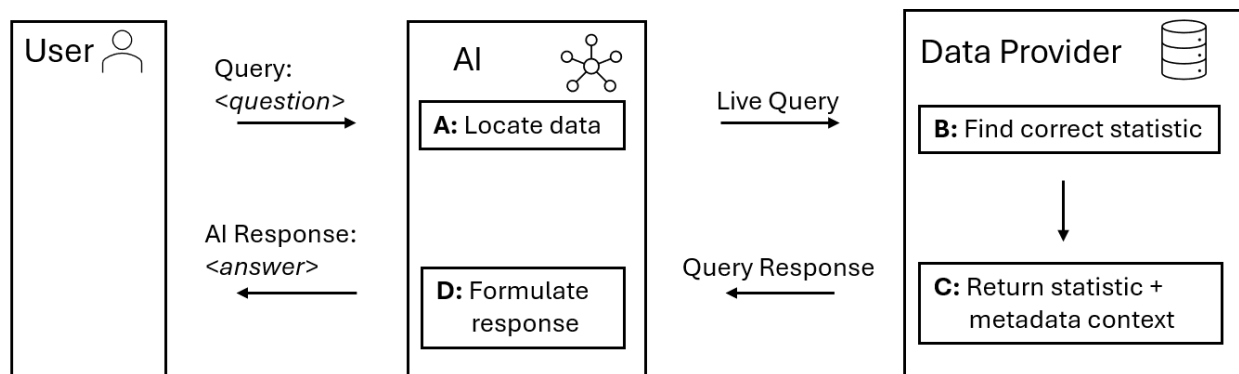
This condition helped us determine if different LLM configurations impacted the accuracy or comprehensiveness of the response under the same metadata conditions.

The variations across the three factors ($2 \times 3 \times 2$) resulted in twelve conditions evaluated.

In all conditions, model prompts were formatted as JSON objects with key-value pairs³ and were manually copied and pasted into the AI service. Each query was run only once. The responses to queries were saved and qualitatively evaluated.

A typical retrieval-augmented generation workflow is depicted in Figure 1. First, a user enters a query into an AI chat. The AI system then locates the data (A) and submits a live query to the data provider to find the correct statistic (B). The data provider then returns the statistic and metadata context (C), before formulating the response (D). This project focuses on metadata provided in step (C) so that a quality answer is returned in (D). We control for the correct statistic being retrieved in step (B) by providing manually collected data.

Figure 1: Retrieval-Augmented AI Response Process



FINDINGS

For each of the study conditions, we evaluated whether the model response included the accurate numerical estimate as well as if it provided important context for the user to understand the estimate. Specifically, we looked for information about the (1) reference period, (2) population, (3) survey the statistic was generated from, (4) sponsoring agency for the survey, (5) calculations done to support the reported figure, (6) information relevant to data quality (e.g., sample size), and (7) other information provided as part of the metadata. For each of the study conditions, a score was calculated by adding the number of

³ This was done to simulate the return of gathered data and metadata from an agency's API or other search service (see Figure 1).

these pieces of context included in the response (a score of 0.5 was given for partially correct context items).⁴ This resulted in a score ranging from 0-7.

In all conditions, the model returned an accurate response. Given the correct answer was provided in the question prompt, this is not surprising.⁵ More importantly, we found that extensive metadata did not necessarily translate into additional context being provided in the response beyond minimal metadata unless paired with a more ruminative/verbose model or AI mode.

Table 1 shows the results of the scoring of our twelve conditions. Looking at the average score, we can see that providing no metadata performed worse than providing any metadata (minimal or extensive) and that there was no difference between the minimal and extensive metadata conditions. Also, Copilot shared more of the provided context as part of its response in think deeper mode than in the standard response mode.

Table 1: Provided Context Score by Study Condition

Data	Thinking Mode	Metadata Levels		
		None	Minimal	Extensive
CPS-ASEC	Standard response	1.5	4	4
CPS-ASEC	Think deeper	1.5	5	7
ACS	Standard response	0.5	3	3
ACS	Think deeper	0.5	7	5
Average Score		1	4.75	4.75

CONCLUSION AND RECOMMENDATIONS

This study examined the degree to which metadata improved the accuracy and context of AI-retrieved responses to simple questions about federal statistical data products. We find that providing minimal metadata, defining the elements in statistical tables, is crucial to enabling the LLM to return accurate, contextualized results. We also show that providing comprehensive metadata, such as technical documentation, does not appear to meaningfully improve the accuracy or comprehensiveness of the response to a simple question. These qualitative findings suggest that while comprehensive metadata may be important for upstream AI retrieval processes, we find that it may provide diminishing marginal value in the final response generation stage.

⁴ A partial score was given if the model reported the population was the United States but not that it was for adults aged 25 and over.

⁵ Throughout our study conditions, accuracy wasn't an issue because we provided data to the LLM as part of the prompt. In analyses not shown here, we replicated previous research showing that a model strictly relying on its pre-training data that did not include the survey reference period, provided inaccurate responses and inconsistent context. The query asked about educational attainment data for which 2024 was the most recent reference period. The model used in this work, GPT5 in Microsoft Copilot Government Community Cloud, has a training data cutoff of September 2024.

The team recommends that agencies:

1. Create a list of common queries users might have about their data products and define minimally acceptable responses to those queries to use in evaluation,
2. Regularly evaluate AI services' responses to agency-relevant queries,
3. Consider methods for dynamically selecting metadata relevant to user queries rather than supplying comprehensive metadata to all queries, and
4. Explore the creation of response templates to promote more consistent outputs that provide metadata deemed essential by the agency.

FUTURE RESEARCH

The team also recommends continued research in this area. This work remained qualitative. A fuller quantitative study assessing more treatment variations would be helpful. Future research may expand on this research by:

1. Experimenting with a wider array of topic areas, query wordings, data products, and surveys with varying amounts and quality of metadata
2. Experimenting with query types other than point estimates such as estimates of change (e.g., How does the number of Americans with at least a bachelor's degree now compare to 10 years ago?) or trends (e.g., How has the number of American's with at least a bachelor's degree changed over time?)
3. Experimenting with multi-turn sessions where the user's query is ambiguous and AI asks questions to refine.
4. Experimenting with more AI services and models and response modes
5. Submitting the same query package multiple times and examining the distribution of response quality rather than a single response
6. Further standardizing the evaluation of model responses through the use of coding rubrics used by multiple human reviewers (potentially even crowdsourced). Alternatively, AI could be used to review the outputs following a scoring rubric.

Scaling the methodology to cover more conditions would allow for more robust conclusions about the nature of tool-driven AI response.