AI-Ready Data

Ensuring Public Data Meets the Needs of AI and the American Public

The USAFacts Guide to AI-Ready Data for Government Agencies

Introduction: Why AI Ready Data Matters

Artificial Intelligence (AI) is transforming public policy, governance, and civic engagement. Large-language models (LLMs) and chat applications have become the fastest growing way Americans get information and understand their communities. Yet, trusting AI models to produce reliable, up-to-date factual information is still fraught with challenges. Hallucinations, stale data, and lack of context all lead to risks that AI presents inaccurate or incomplete information for users.

Key to addressing these changes is enabling AI models to use factual information. For decades, the government's statistical agencies have served as America's data backbone, informing the public and policymakers on a wide range of topics. However, while model developers and AI systems builders have made great improvements in many areas, government data is largely built for the pre-AI era, and a lack of standardization, availability, and consistency limits the ability for AI systems to incorporate government data and provide accurate facts.

At USAFacts, a nonpartisan initiative focused on making data accessible and useful to the public, we work daily with government data and, recently, LLMs to bring robust, accurate analysis of taxpayer-funded data to every American. We've seen what works and what doesn't in data publishing and have established these guidelines to better prepare America's data agencies for the AI future.

AI For Modern Government

The wheels of government can turn slowly, especially when it comes to adopting new technologies. The ability to access and use data is foundational to many of today's modern, successful, and AI-forward businesses. As government works to catch up to these standards, being able to access data using the most advanced AI technologies is crucial to modernization.

Readying government data for AI has the potential to provide deep insights to and create efficiencies for government employees, improve the effectiveness of government policies and institutions, and allow better access to and understanding of government for American citizens. AI-ready data is a clear early step in creating a more modern, effective, and efficient government. Conversely, not taking on this important work will greatly hinder future modernization efforts.

A Systems Based Approach

The cost, both financially and ecologically, of training multi-billion parameter large language models can mean that AI models are a snapshot of the world at the time they were trained. New techniques, like RAG, Agents, and AI web search provide for improved data access and freshness, but can rely on outdated websites, irrelevant results, or misinterpretation of data due to lack of context.

At USAFacts, we view an AI system comprised of an LLM and one or more information retrieval techniques, bounded by a set of evaluations for accuracy, as the best path forward for accurate, reliable AI applications. Many of today’s best applications, ChatGPT, Perplexity, and Gemini among them, already rely on some form of information retrieval to ensure their data is as up to date as possible.

As government continues to evolve its role as a data provider to AI systems, these criteria should provide a roadmap for allowing AI to not only access, but also understand and validate the data they are retrieving and presenting to users.

Key Criteria for AI-Ready Government Data

ACCESSIBLE

1. Data Must Be Machine-Readable and Easy to Access

AI systems should reference current, authoritative datasets rather than rely on potentially stale training data. To do so, government data should be easy to retrieve and APIs should focus on reliability and low latency.

How?

- Provide RESTful APIs (with minimal rate limits) for real-time and bulk retrieval.
- Publish datasets in common structured formats (JSON, CSV, machine-readable Excel). Avoid documents (PDFs) that need to be processed and parsed.
- Ensure metadata, including the most recent data update, are programmatically retrievable. Provide webhooks or other push systems for frequently updated data.
- Maintain centralized data catalogs (e.g., expanded and improved data.gov) to simplify discovery and verification.
- Encourage AI developers to cite official endpoints (e.g., APIs) rather than relying on locally cached or model-stored figures.
- Implement versioning so AI developers can track updates over time.

Example: The Department of Commerce continuously updates economic indicators; AI systems should dynamically fetch these figures instead of using older, embedded stats.

UNDERSTANDABLE

2.Data Must Be Machine Understandable with Easily Retrievable, Well-Structured Documentation and Metadata

AI models need context—such as variable definitions, collection methodology, and version history—to ensure transparency and explainability to end users. Structured, interoperable data is crucial for efficient ingestion and accurate generation of insights.

How?

- Maintain data dictionaries, taxonomies, and ontologies to describe the data's domain and structure. Publish Data Cards, similar to the Model Cards used in AI Transparency, that outline dataset origins, known biases, and recommended use cases.
- Employ semantic labeling (chunking, tagging) to facilitate retrieval-augmented generation in LLMs.
- Where possible, make data labels human-readable and colloquial. Provide well-structured explanations and context to data labels in key-value pairs or table format rather than footnotes.
- Align with standard schemas like NIEM and Crossaint for consistent cross-agency data exchange.
- Disclose methodologies for derived or transformed datasets, ensuring clarity on how figures are generated.

Example: Data sets contain detailed Data Cards that clarify how health metrics were collected, aiding interpretability and reducing misuse.

ACCURATE

3. Agencies Should Provide Clear Evaluation Benchmarks and Make Them Public

To ensure data is represented accurately, data agencies should define what a good LLM response looks like and regularly benchmark models and AI systems against their own Evaluations.

How?

- Provide LLM evaluation datasets so AI developers can test and fine-tune models for up-to-date accuracy.
- Regularly test foundation models and AI applications against the Agency's own benchmarks. Notify application developers and model builders of discrepancies.
- Documentation and data should be regularly reviewed and updated by internal teams to find and correct potential abuse and misinformation.
- Employ automated validation and audit trails to detect errors.

Example: Agencies release a public test sets that allow AI researchers to validate model outputs against real-time government data.

OPEN

4. Data Must Be Unencumbered and Open to All Citizens

Public data should enhance civic engagement while upholding privacy and equity.

How?

- Use clear, permissive licenses (CC0, ODC-BY) enabling unrestricted AI development.
- Apply privacy safeguards (e.g., differential privacy) to protect individuals in aggregate datasets.
- Properly identify and document suppressed data (e.g., in very small counties) in plain language with unique identifiers.
- Provide API keys for accessing the data free of charge and in an open, anonymous and automated manner.
- Ensure text documentation (data dictionaries, methodology) are accessible to both humans and AI/web crawlers.

Example: Data and aggregated statistics are released under CC0 or similar licenses, integrating privacy mechanisms but allowing open use for AI-driven insights.

Join Us in Ensuring the Future of AI is Based in Facts

Government agencies must act with urgency to modernize and standardize public data for AI applications—prioritizing real-time, verifiable references over static figures embedded in training data. By aligning with existing mandates and global standards, agencies can ensure government data remains authoritative, ethically accessible, and technically robust. USAFacts stands ready to collaborate with federal, state, and local agencies as well as industry leaders to establish best practices for AI-ready open data.

USAFACTS

Contact

For more information or to discuss AI-ready data standards, contact:

Richard Coffin
Chief of Research and Advocacy, USAFacts
richardc@usafacts.org