

Building AI-Ready Best Practices: The Work of the Commerce Data Governance Board

October 2024

*The views expressed in this presentation are those of the presenter
and not the Census Bureau.*

Ryan Harper, Product Manager

Center for Enterprise Dissemination
U.S. Census Bureau

DOC AI and Open Government Data Assets Working Group

- The RFI that was sent out on April 17th, 2024 received over thirty responses
- The working group was established shortly thereafter and features participants from across the DOC
- The initial working group guidance is expected to be released this year



UNITED STATES
PATENT AND TRADEMARK OFFICE



NIST ▶ NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

United States[®]
Census
Bureau





Generative AI and LLMs



Generative AI applications enable users to create synthetic content, including text, images, audio, and video.



Large language models (LLMs) leverage deep learning algorithms to model language patterns and generate text.



Democratizing Access to Public Data Through Generative AI



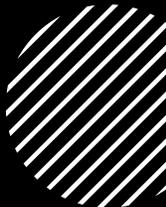
The emergence of GenAI has the potential to provide improved information and data access to users, from novices to experts.



GenAI and other AI technologies present opportunities and challenges for data providers and data users—including government entities, industry, academia, and the public at large.



Tech Adoption Patterns Tend to Repeat



Technology emerges



Public expectations change



Policy follows

Changes in the Public Data Landscape

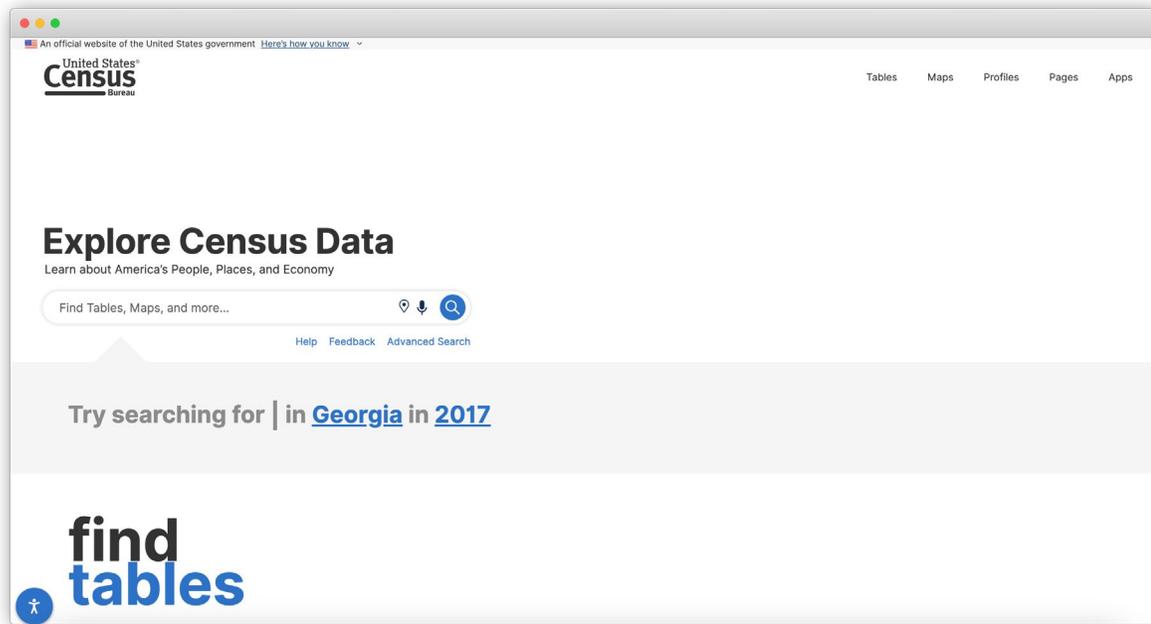
20th Century

- Federal government was the dominant user
- The statistical system was a near monopoly
- Output was mostly cross-tabs
- Published in books and deposited in libraries, then electronically largely in book formats (PDFs)
- Source data acquisition was difficult and costly
- Privacy and confidentiality risks were small
- Computation was limited and expensive

21st Century

- There are many diverse users of public data
- There are many more organizations that produce similar statistical products
- Output is varied and features visualization and tools for analysis
- Data is accessed electronically online or in secure enclaves
- Source data is more abundant and available in structured and unstructured formats
- Privacy and confidentiality risks are much greater
- Computation is vastly improved

data.census.gov and the U.S. Census API



An official website of the United States government [Here's how you know](#)

United States Census Bureau

Tables Maps Profiles Pages Apps

Explore Census Data

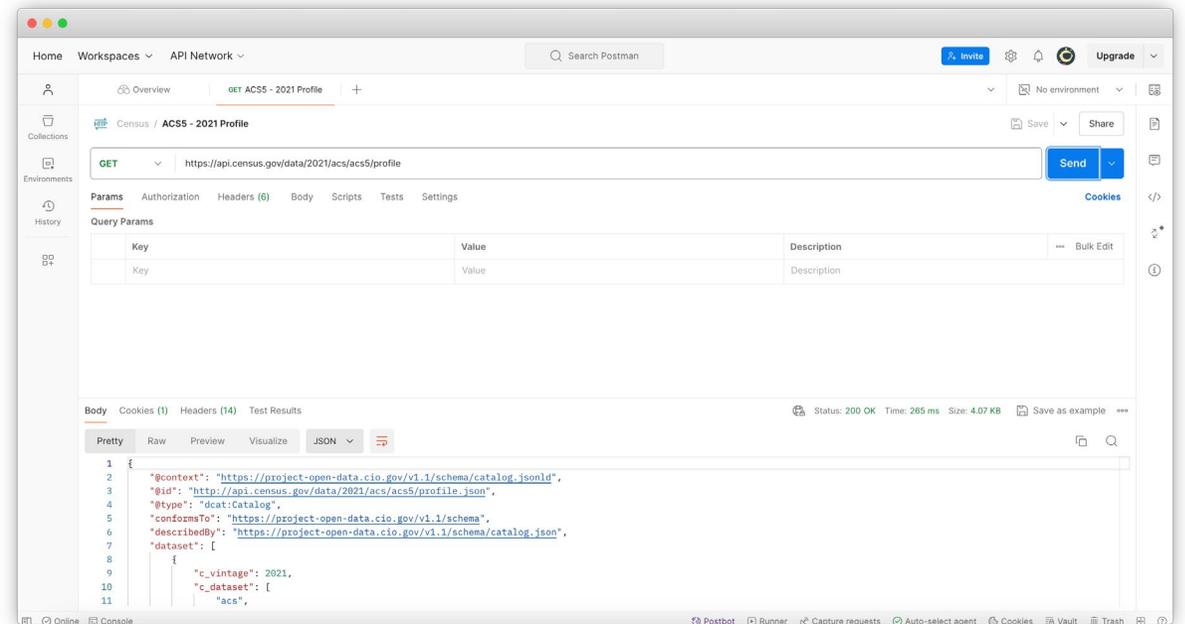
Learn about America's People, Places, and Economy

Find Tables, Maps, and more...

Help Feedback Advanced Search

Try searching for | in [Georgia](#) in [2017](#)

find tables



Home Workspaces API Network

Search Postman

GET ACSS - 2021 Profile

GET `https://api.census.gov/data/2021/acs/acs5/profile` Send

Params Authorization Headers (6) Body Scripts Tests Settings

Key	Value	Description	Bulk Edit
Key	Value	Description	

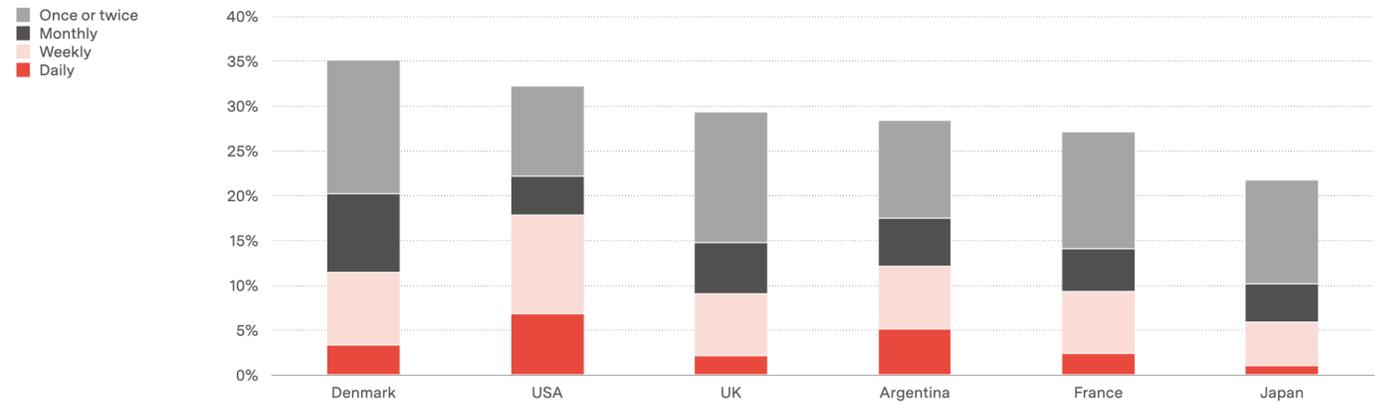
Body Cookies (1) Headers (14) Test Results

Status: 200 OK Time: 265 ms Size: 4.07 KB Save as example

```
1 {
2   "@context": "https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld",
3   "@id": "http://api.census.gov/data/2021/acs/acs5/profile.json",
4   "@type": "dcat:Catalog",
5   "conformsTo": "https://project-open-data.cio.gov/v1.1/schema",
6   "describedBy": "https://project-open-data.cio.gov/v1.1/schema/catalog.json",
7   "dataset": [
8     {
9       "c_vintage": 2021,
10      "c_dataset": {
11        "acs",
```

So Are Generative AI and LLMs Next?

Use of ChatGPT by country, May 2024

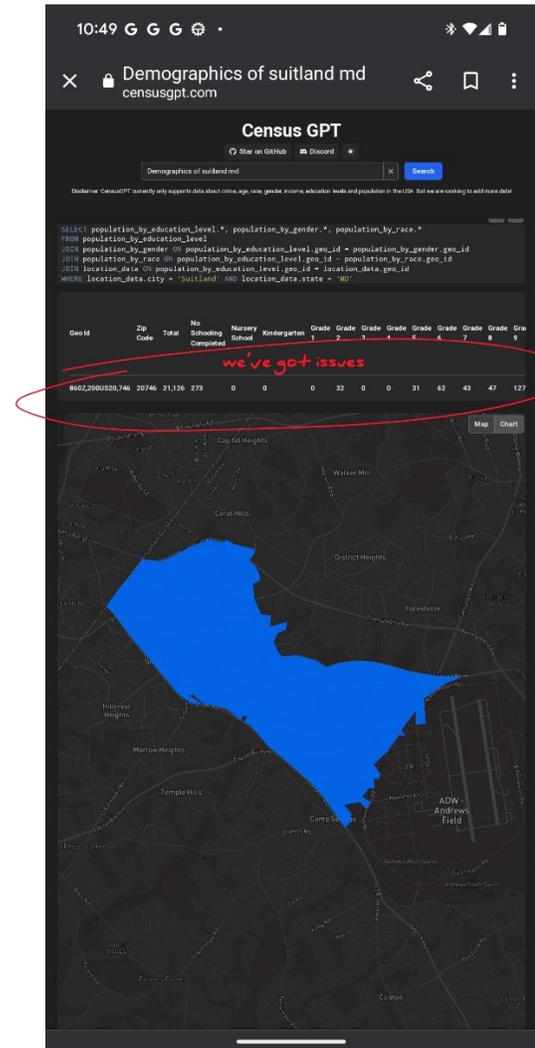
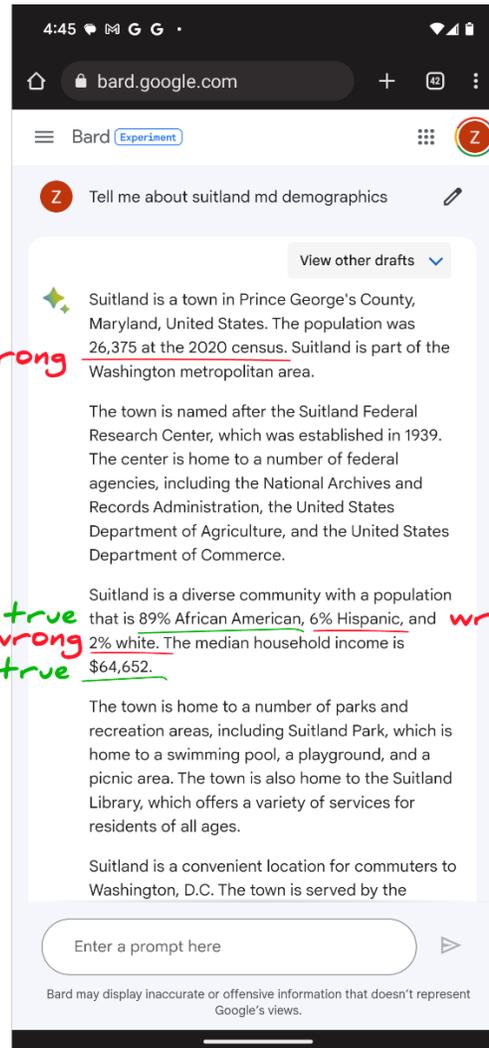
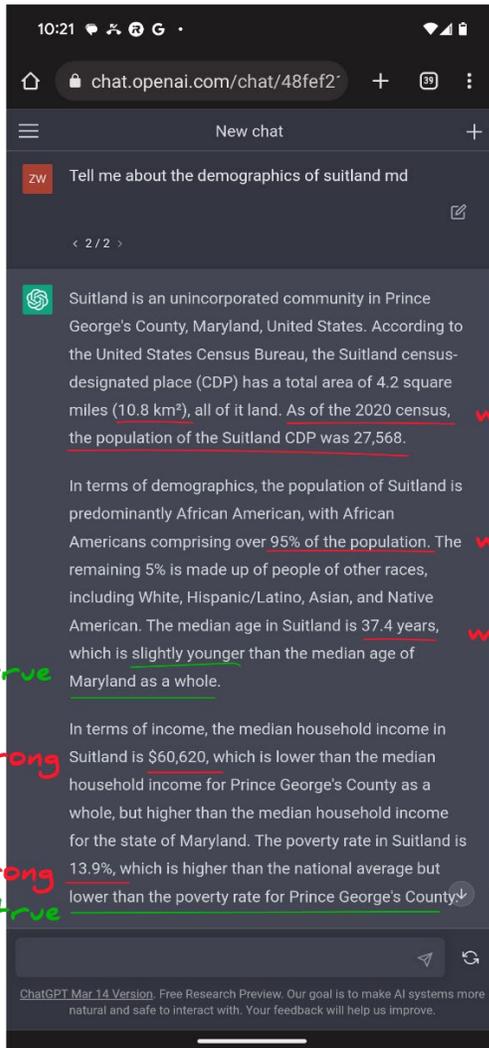


Source: Reuters Institute

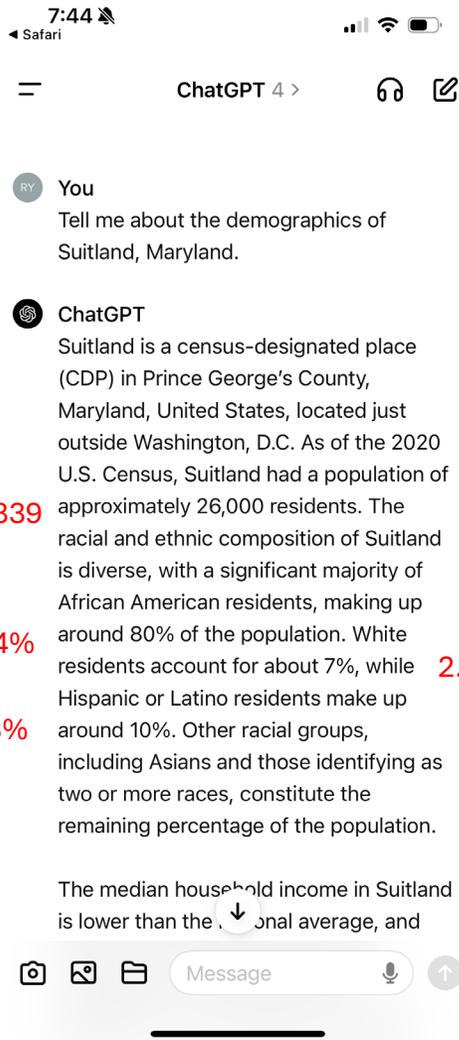
Benedict Evans — July 2024

Current State

ChatGPT, Bard, and CensusGPT return incorrect results from secondary sources (2023)



ChatGPT, Gemini, and Copilot return incorrect results from secondary sources (February 2024)



26,375

91%

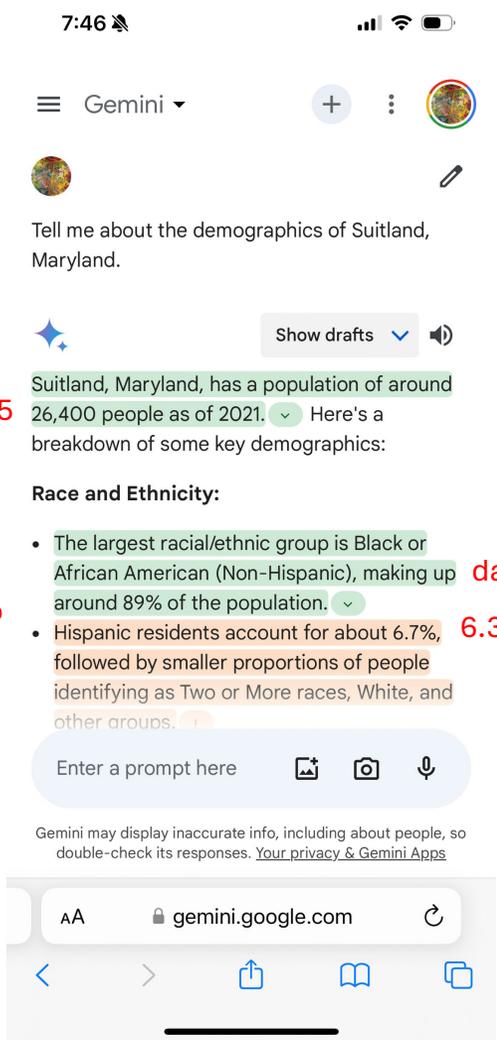
datausa.io

6.3%

87.4%

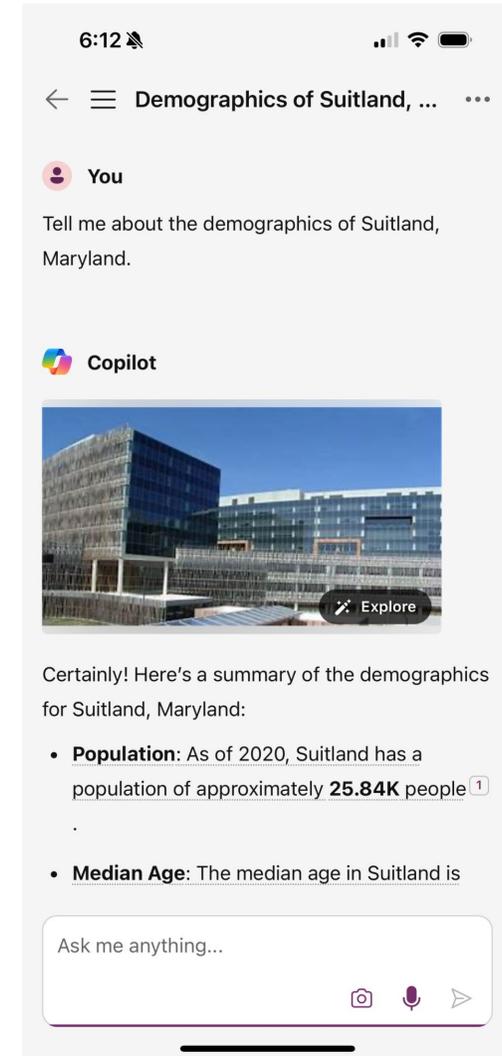
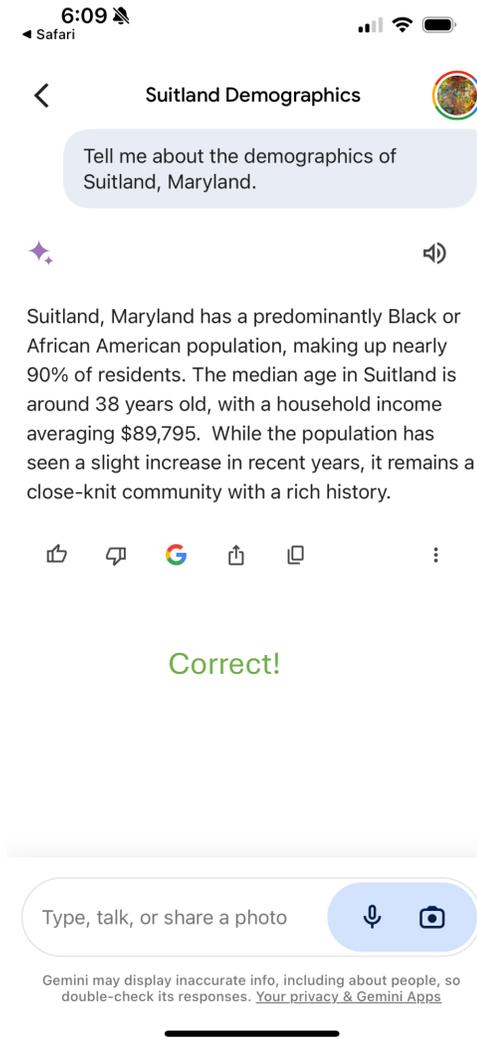
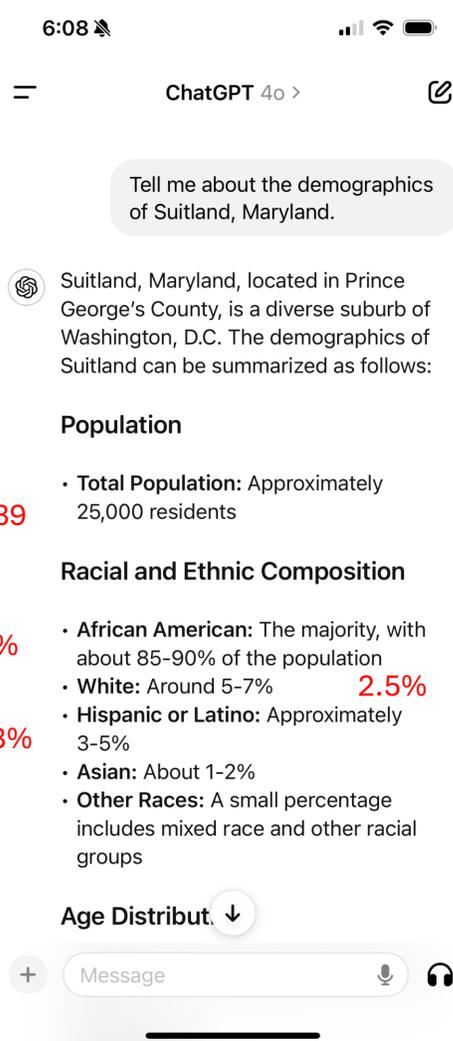
3.8%

7.8%

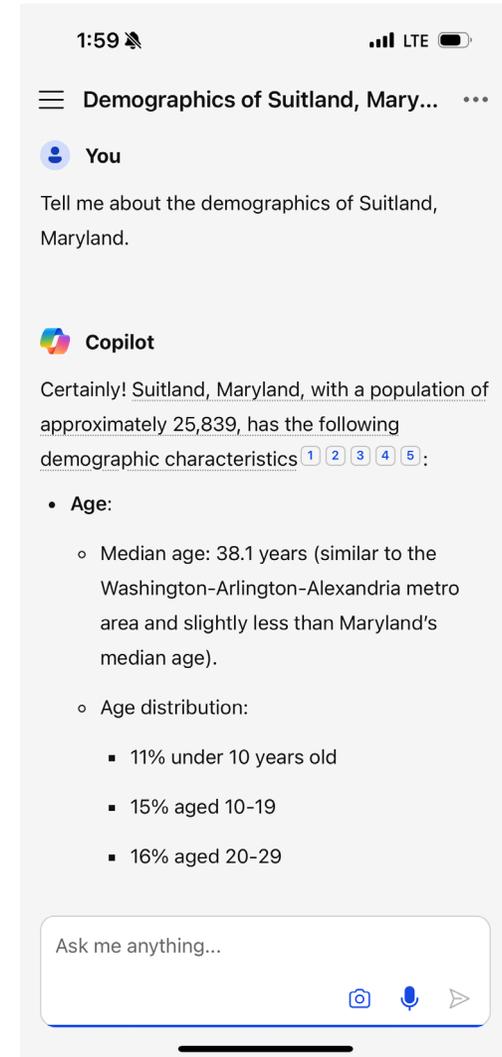
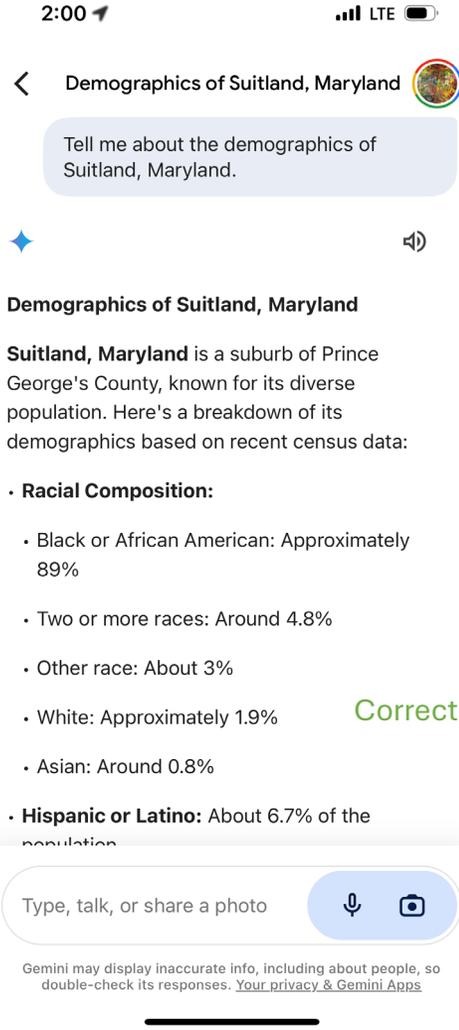
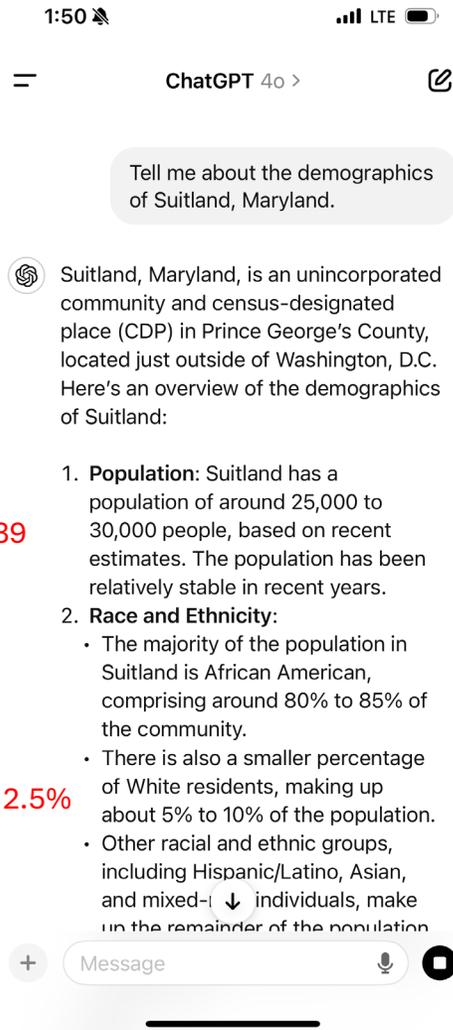


1: neighborhoodscount.com
2: census.gov (QuickFacts)
3: datausa.io

ChatGPT returns incorrect results. Copilot cites secondary sources. Gemini shows improved accuracy. (June 2024)



ChatGPT returns incorrect results. Copilot cites secondary sources. Gemini is accurate. (August 2024)



Emerging Publishing Tactics

Tactic #1: SEO, but More

Data Discoverability: Sitemaps, Indexing, robots.txt, Server Side Rendering...

```
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_basics.xml.gz</loc>
  <lastmod>2024-03-08</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_010.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_040.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_050.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_060.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_250.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_500.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_860.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_860.part2.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_860.part3.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_950.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_960.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
▼ <sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_970.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
```

What Is This Data About?

Meta Tags, Open Graph Tags, schema.org, Croissant...

The screenshot shows the United States Census Bureau website. The main heading is "Explore Census Data" with the subtitle "Learn about America's People, Places, and Economy". A search bar is present with the text "Find Tables, Maps, and more...". The browser's developer tools are open, displaying an LD+JSON schema for the page. The schema includes metadata such as @context, @type, keywords, text, description, countryOfOrigin, creator, logo, telephone, slogan, and a detailed description of the data available.

```
<script type="application/ld+json"> == $0
{
  "@context": "https://schema.org",
  "@type": "Website",
  "keywords": "US Census Bureau, Census Bureau Data",
  "text": "Census Bureau Data",
  "description": "Explore the thousands of tables we have. Build customized maps from any variable in our data tables. Learn about America's communities through our data profiles. They cover 100,000+ different geographies: states , counties , places , tribal areas , zip codes , and congressional districts. For each, we cover topics like education, employment, health, and housing just to name a few. Get record-level access to our Public Use Microdata Sample (PUMS) files. Check out our help resources, FAQs, quick start guides, and release information.",
  "countryOfOrigin": "United States Of America",
  "creator": {
    "@type": "GovernmentOrganization",
    "name": "United States Census Bureau",
    "email": "census.data@census.gov",
    "address": "4600 Silver Hill Road Washington, DC 20233",
    "department": {
      "@type": "Organization",
      "name": "Research and Methodology",
      "alternateName": "ADRM"
    },
  },
  "logo": "https://www.census.gov/etc.clientlibs/census/clientlibs/census-pattern-library/resources/images/USCENSUS_IDENTITY_SOLO_BLACK_1.5in_R_no_padding.svg",
  "telephone": "301-763-4636",
  "slogan": "Measuring America's People, Places, and Economy",
  "description": "The Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy."
}
```

Tell It To Me Straight: Documentation and Prose Descriptions

The screenshot shows the Adobe Acrobat AI Assistant interface. On the left, a sidebar titled 'Welcome to Acrobat' offers a 7-day free trial for premium PDF tools and lists various tools like 'Ask AI Assistant', 'Generate a summary', and 'Export a PDF'. The main content area displays the 'Census Data API User Guide' with a 'TABLE OF CONTENTS' section listing topics such as Purpose, Available Datasets, Core Concepts, and Query Limits.

The screenshot shows the official website for the U.S. Census API (Beta). The page features a navigation menu with 'Home', 'Overview', 'Concepts', 'Examples', and 'Endpoints'. The main content area is titled 'Introduction' and explains the purpose of the user guide: to instruct developers, researchers, and other API users on how to use the redesigned Beta Unified Census Data Application Programming Interface (API) to request data, metadata, and microdata from Census Bureau tables and datasets. It also provides contact information for questions.

Tactic #2: AI-Ready Data

Aiding Data Ingestion: Proprietary and Unstructured < Open and Structured



VS.



Data Retrieval via API: Open Standards, Accessible Metadata

The screenshot shows the Swagger UI for the Beta Unified Census Data API. At the top, the Swagger logo is visible, along with the URL `/docs/api/config` and an **Explore** button. The main heading is **Beta Unified Census Data API** with version `0.5.1` and the **OAS3** specification. Below the heading, a description states: "The Beta Unified Census Data API gives the public access to raw statistical data from various Census Bureau data programs." There is a **Servers** dropdown menu set to `/` and an **Authorize** button. The **Facets** section, titled "Global Facet Endpoints", lists several API endpoints:

- GET** `/search/facets/codesets`
- POST** `/search/facets/codesets`
- POST** `/search/facets/topics` Returns Applicable Topic Facets
- POST** `/search/facets/geoIds` Returns Applicable Geography Facets
- POST** `/search/facets/geoIds/{hierarchyId}` Returns Applicable Geography Hierarchy ID Facets
- POST** `/search/facets/dates` Returns Applicable Date Facets

Tabular Data: Hierarchical Tables < Flat Tables

United States Census Bureau

Search

Advanced Search

All **Tables** Maps Profiles Pages

DP05 | ACS Demographic and Housing Estimates

American Community Survey | 2022: ACS 1-Year Estimates Data Profiles

Notes Geos Topics Codes Dataset Year Columns Transpose Margin of Error Restore Excel CSV ZIP More Tools

Please note that American Community Survey 1-Year estimates are published for geographies with a population of 65,000 or more. For more information, see the [guidance for when to use 1-year or 5-year estimates](#).

Label	United States	Estimate	Margin of Error	Percent	Percent Margin of Error
SEX AND AGE					
▼ Total population		333,287,562	*****	333,287,562	(X)
▼ Male		165,228,214	±33,974	49.6%	+0.1
Female		168,059,348	±33,976	50.4%	+0.1
Sex ratio (males per 100 females)		98.3	±0.1	(X)	(X)
Under 5 years		18,358,199	±18,313	5.5%	+0.1
5 to 9 years		19,770,314	±74,288	5.9%	+0.1
10 to 14 years		21,220,214	±73,259	6.4%	+0.1
15 to 19 years		21,785,658	±36,858	6.5%	+0.1
20 to 24 years		22,446,040	±36,578	6.7%	+0.1
25 to 34 years		45,233,264	±45,742	13.6%	+0.1
35 to 44 years		44,047,644	±32,891	13.2%	+0.1
45 to 54 years		40,478,446	±29,935	12.1%	+0.1
55 to 59 years		20,553,615	±65,466	6.2%	+0.1
60 to 64 years		21,573,853	±65,896	6.5%	+0.1
65 to 74 years		33,898,072	±23,126	10.2%	+0.1
75 to 84 years		17,764,300	±38,920	5.3%	+0.1

United States Census Bureau

Business and Economy

Advanced Search

All **Tables** Maps Profiles Pages

CS2100CBP | All Sectors: County Business Patterns, Including ZIP Code Business Patterns, by Legal Form of Organization and Employment Size Class for the U.S., States, and Selected Geographies: 2021

Economic Surveys | 2021: ECNSVY Business Patterns County Business Patterns

Notes Geos Topics Codes Year Columns Filters Restore Excel CSV ZIP Cite Share Print More Tools

Geographic Area Name	2017 NAICS code	Meaning of NAICS Code	Meaning of Legal form of o...	Meaning of Employment si...	Year	Number of establishments	Annual payroll (\$1,000)
United States	00	Total for all sectors	All establishments	All establishments	2021	8,148,606	8,278,573,947
United States	00	Total for all sectors	All establishments	Establishments with less than ...	2021	4,614,481	451,190,829
United States	00	Total for all sectors	All establishments	Establishments with 5 to 9 em...	2021	1,434,428	459,103,438
United States	00	Total for all sectors	All establishments	Establishments with 10 to 19 e...	2021	988,092	660,592,593
United States	00	Total for all sectors	All establishments	Establishments with 20 to 49 ...	2021	702,372	1,085,150,405
United States	00	Total for all sectors	All establishments	Establishments with 50 to 99 ...	2021	223,604	900,585,158
United States	00	Total for all sectors	All establishments	Establishments with 100 to 24...	2021	128,680	1,245,537,514
United States	00	Total for all sectors	All establishments	Establishments with 250 to 49...	2021	35,454	873,248,570
United States	00	Total for all sectors	All establishments	Establishments with 500 to 99...	2021	13,130	728,695,838
United States	00	Total for all sectors	All establishments	Establishments with 1,000 em...	2021	8,365	1,874,469,602
United States	00	Total for all sectors	C-corporations and other cor...	All establishments	2021	2,048,384	4,351,380,964
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with less than ...	2021	896,447	124,306,003
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with 5 to 9 em...	2021	401,839	154,666,173
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with 10 to 19 e...	2021	309,711	244,634,928
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with 20 to 49 ...	2021	251,773	450,501,569
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with 50 to 99 ...	2021	90,038	428,269,108
United States	00	Total for all sectors	C-corporations and other cor...	Establishments with 100 to 24...	2021	65,102	702,162,711

1 to 200 of 1,401 | Page 1 of 8

Tactic #3: Meet AI Systems Where They Are

Census API Assistant GPT (Research Prototype)

API Assistant

RY



API Assistant

By Ryan Harper

Expert at constructing U.S. Census API calls.

How do I get population data for California?

What parameters do I use for housing statistics...

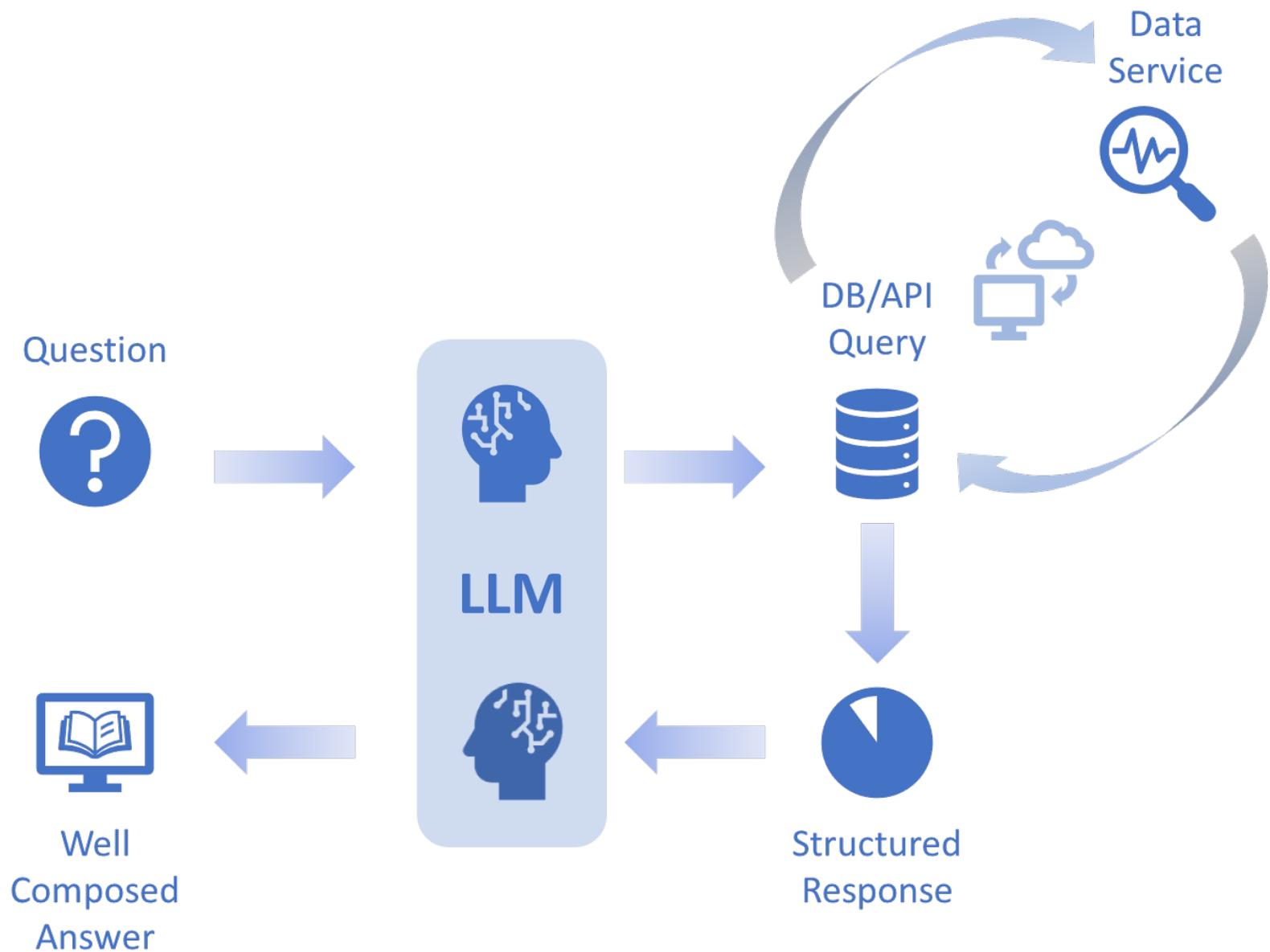
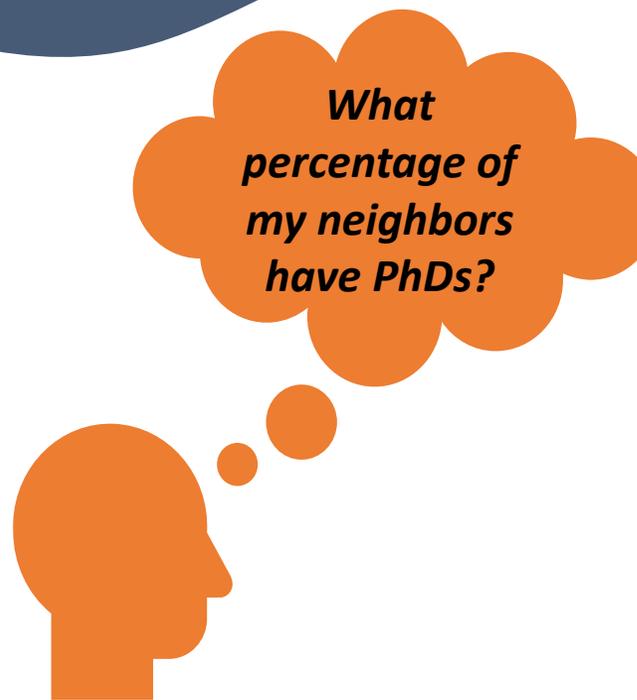
Can you help me with API calls for economic figures?

I need to access demographic information, how...

 Message API Assistant 

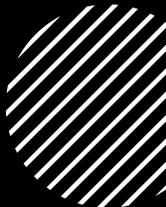
ChatGPT can make mistakes. Check important info. 

Hybrid Solutions





Partnership and Engagement: Community



An official website of the United States government [Here's how you know](#)

GSA U.S. General Services Administration

Per Diem Lookup

Buy through us | Sell to government | Real estate | Policy and regulations | Small business | Travel | Technology | About us

Home > Technology > Government IT Initiatives > Federal AI Hackathon

Federal AI Hackathon

Help us optimize federal websites and data for AI while enhancing human usability

Compete for a chance to win a share of a **\$10,000 cash prize**

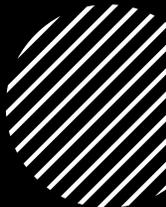
Event details
July 31, 9 a.m.–6:30 p.m. EDT
Atlanta, GA | New York, NY | Washington, DC

Hosted by GSA and co-sponsored by other federal agencies and corporate partners*

OpenAI | Microsoft | AWS | cohere | groq | Google | slack



Partnership and Engagement: Open Source Collaboration



The screenshot shows a GitHub repository page for the path `data / scripts / us_census / acs5yr / subject_tables / S1603`. The repository is owned by `Abilityguy` and has a commit hash of `cceaad8` from 2 years ago. The file list includes:

Name	Last commit message	Last commit date
..		
testdata	Artifacts for US Census S1702 Subject Table (#509)	2 years ago
README.md	Added artifacts for S1603 import (#523)	2 years ago
S1603_spec.json	Census tools fix (#696)	2 years ago
process.py	Added artifacts for S1603 import (#523)	2 years ago
process_test.py	Census tools fix (#696)	2 years ago

The `README.md` file content is as follows:

US census ACS5YR S1603 Subject Table import

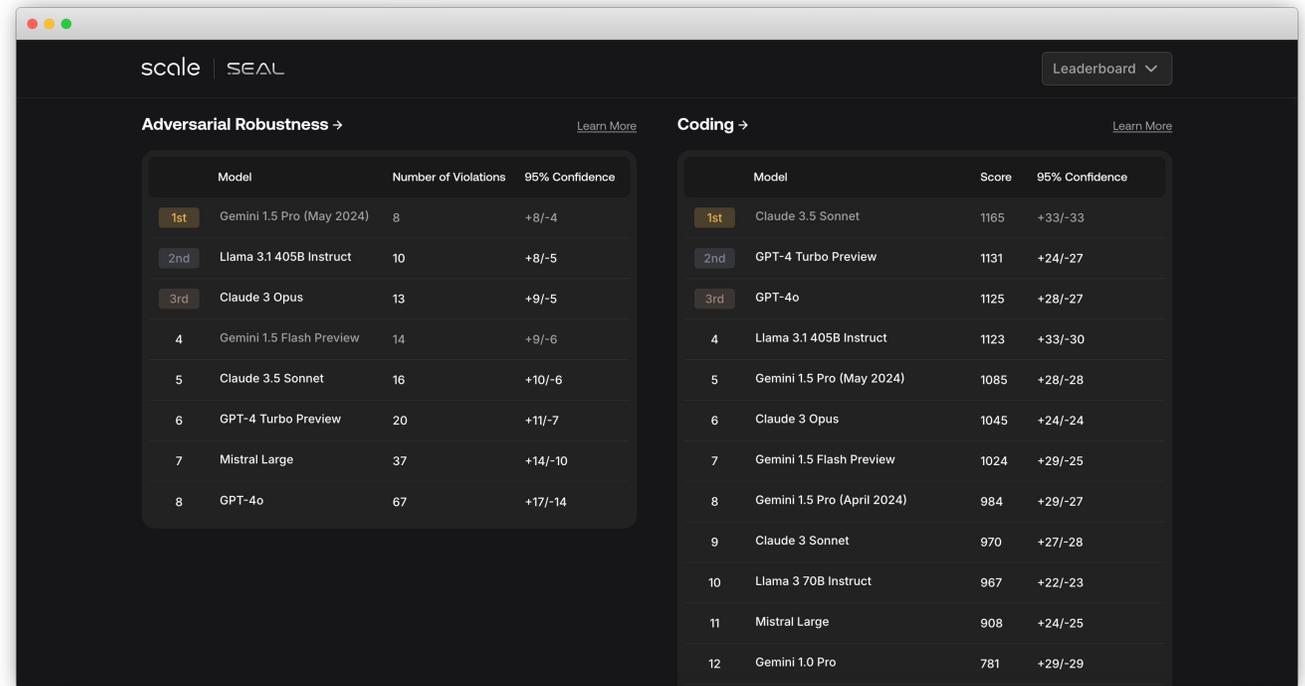
This import gives the count of people based on the language spoken at home. It breaks this further down based on age, nativity and citizenship status, poverty status and educational attainment.

Years: 2010-2019
Geo : Country, State, County and Place

Important Notes :

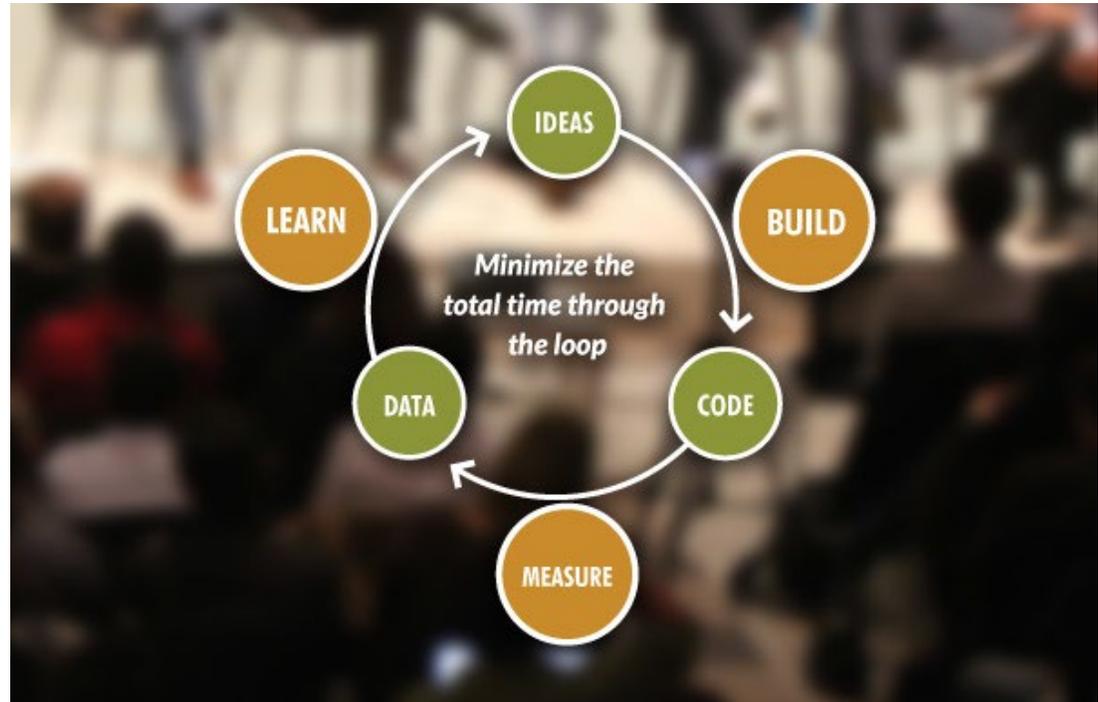
1. The data is presented as counts from 2015-2019 and as percentages from 2010-2014.

Partnership and Engagement: Benchmarks



What's Next?

Monitor and Continue to Innovate (Cautiously and Ethically)



THANK YOU

Ryan Harper, Product Manager
Center for Enterprise Dissemination
U.S. Census Bureau

ryan.harper@census.gov