

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




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



Changes in the Public Data Landscape

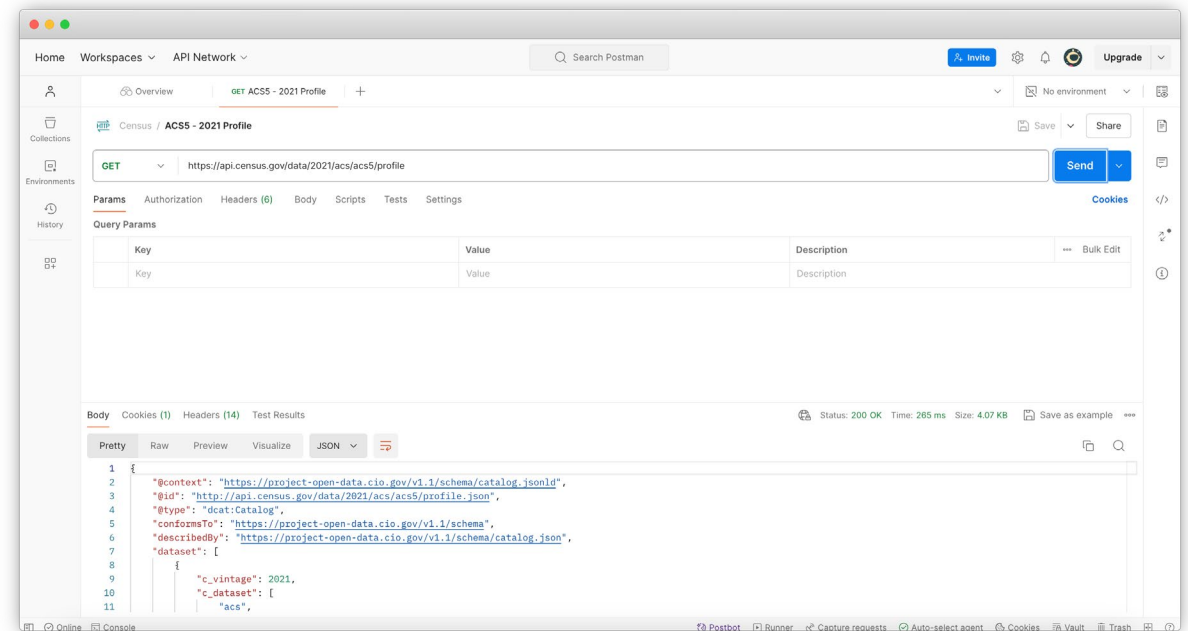
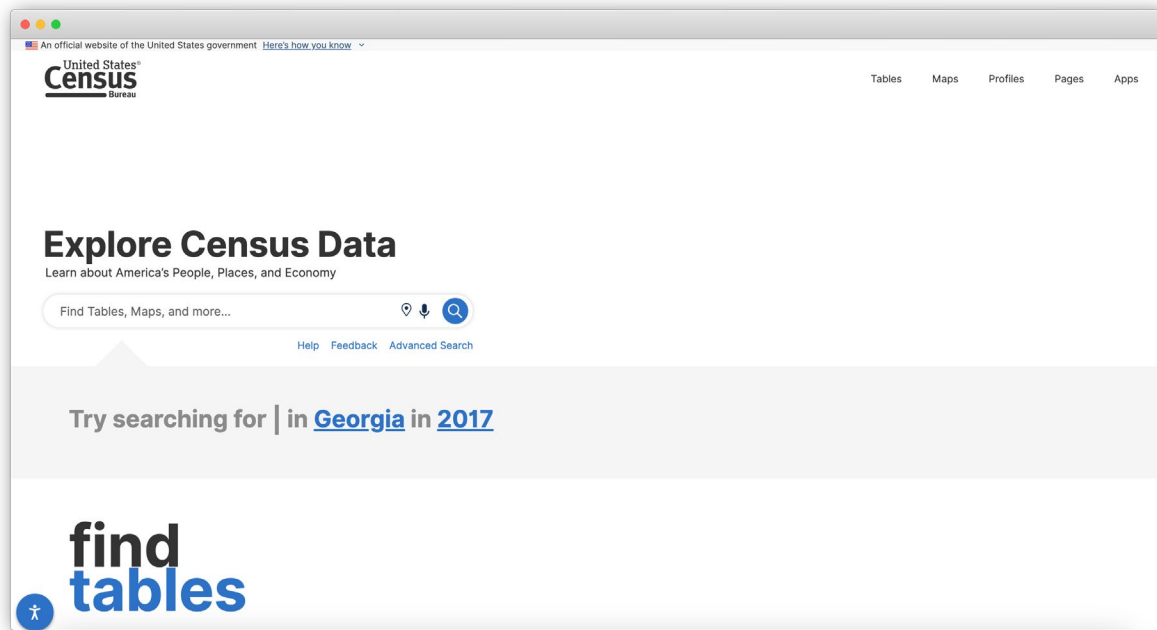
20th Century

- 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

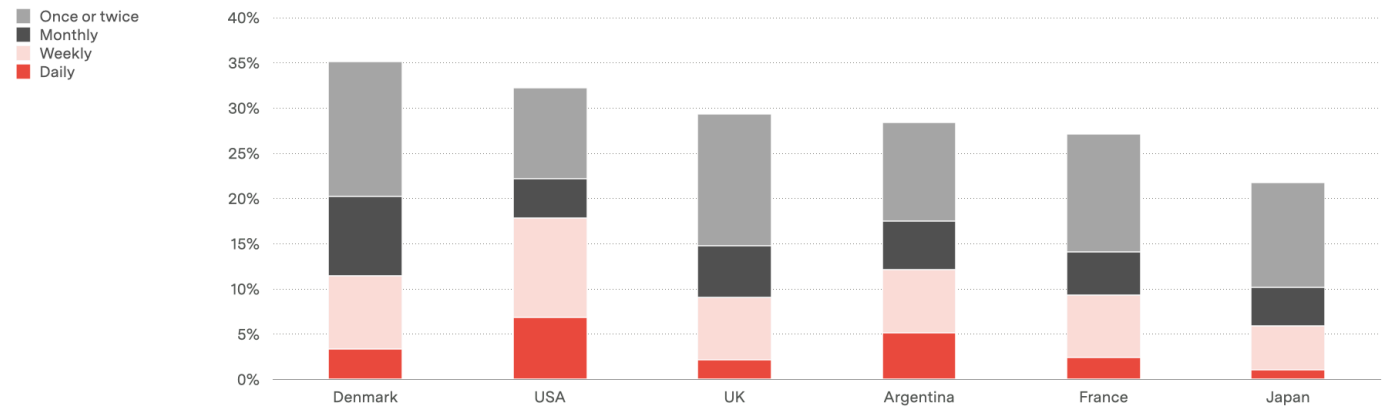
- There are many diverse users of public data
- There are many more organizations that produce similar statistical products
- Output is varied and features visualization and tools for analysis
- Data is accessed electronically online or in secure enclaves
- Source data is more abundant and available 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



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)

10:21

chat.openai.com/chat/48fef2

New chat

Tell me about the demographics of suitland md

< 2 / 2 >

Suitland is an unincorporated community in Prince George's County, Maryland, United States. According to the United States Census Bureau, the Suitland census-designated place (CDP) has a total area of 4.2 square miles (10.8 km²), all of it land. As of the 2020 census, the population of the Suitland CDP was 27,568.

In terms of demographics, the population of Suitland is predominantly African American, with African Americans comprising over 95% of the population. The remaining 5% is made up of people of other races, including White, Hispanic/Latino, Asian, and Native American. The median age in Suitland is 37.4 years, which is slightly younger than the median age of Maryland as a whole.

In terms of income, the median household income in Suitland is \$60,620, which is lower than the median household income for Prince George's County as a whole, but higher than the median household income for the state of Maryland. The poverty rate in Suitland is 13.9%, which is higher than the national average but lower than the poverty rate for Prince George's County.

ChatGPT Mar 14 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

wrong

wrong²

wrong

wrong

true

wrong

wrong

true

4:45

bard.google.com

Bard Experiment

Tell me about suitland md demographics

View other drafts

Suitland is a town in Prince George's County, Maryland, United States. The population was 26,375 at the 2020 census. Suitland is part of the Washington metropolitan area.

The town is named after the Suitland Federal Research Center, which was established in 1939. The center is home to a number of federal agencies, including the National Archives and Records Administration, the United States Department of Agriculture, and the United States Department of Commerce.

Suitland is a diverse community with a population that is 89% African American, 6% Hispanic, and 2% white. The median household income is \$64,652.

The town is home to a number of parks and recreation areas, including Suitland Park, which is home to a swimming pool, a playground, and a picnic area. The town is also home to the Suitland Library, which offers a variety of services for residents of all ages.

Suitland is a convenient location for commuters to Washington, D.C. The town is served by the

Enter a prompt here

Bard may display inaccurate or offensive information that doesn't represent Google's views.

true

wrong

true

wrong

10:49

Demographics of suitland md

censusgpt.com

Census GPT

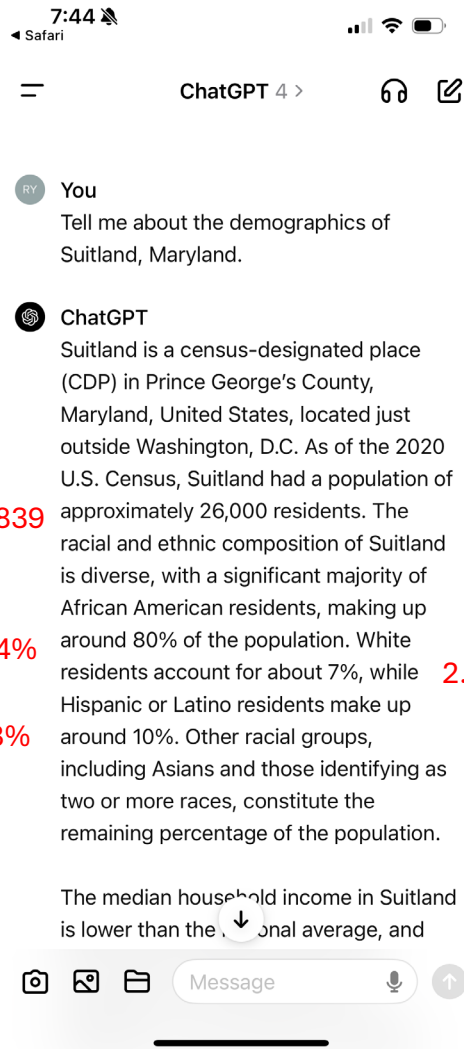
Demographics of suitland md

```
SELECT population_by_education_level.*, population_by_gender.*, population_by_race.*
FROM population_by_education_level
2020 population_by_gender ON population_by_education_level.geo_id = population_by_gender.geo_id
2010 population_by_race ON population_by_education_level.geo_id = population_by_race.geo_id
2020 location_data ON population_by_education_level.geo_id = location_data.geo_id
WHERE location_data.city = 'Suitland' AND location_data.state = 'MD'
```

GeoId	Zip Code	Total	No. of Housing Units	Nursery School	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9
8662200US20746	20746	21,126	273	0	0	0	32	0	31	62	43	47	127	

we've got issues

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

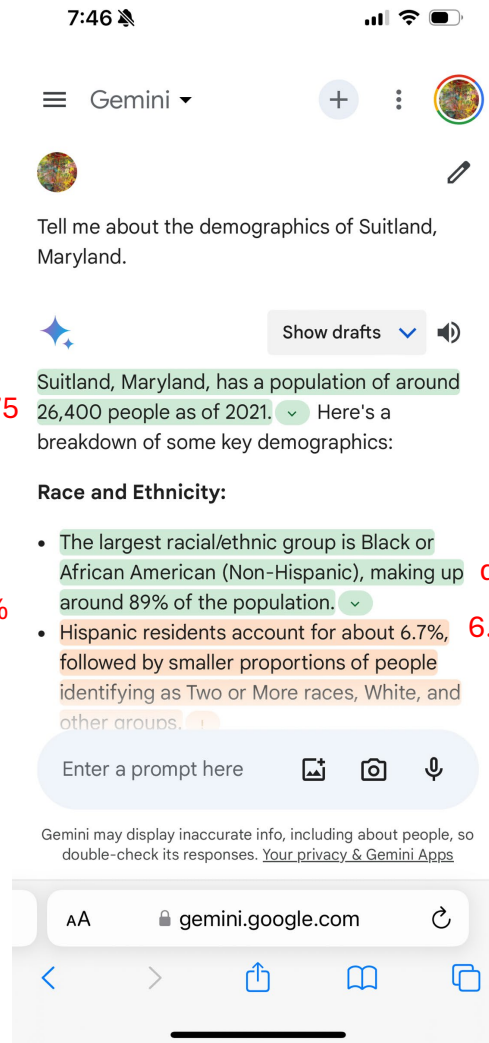


25,839

87.4%

7.8%

2.5%

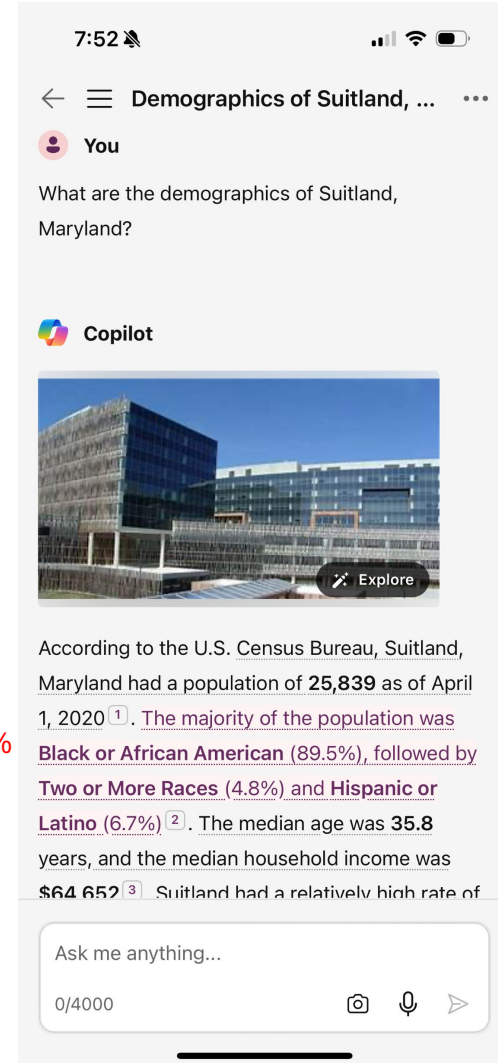


26,375

91%

datausa.io

6.3%



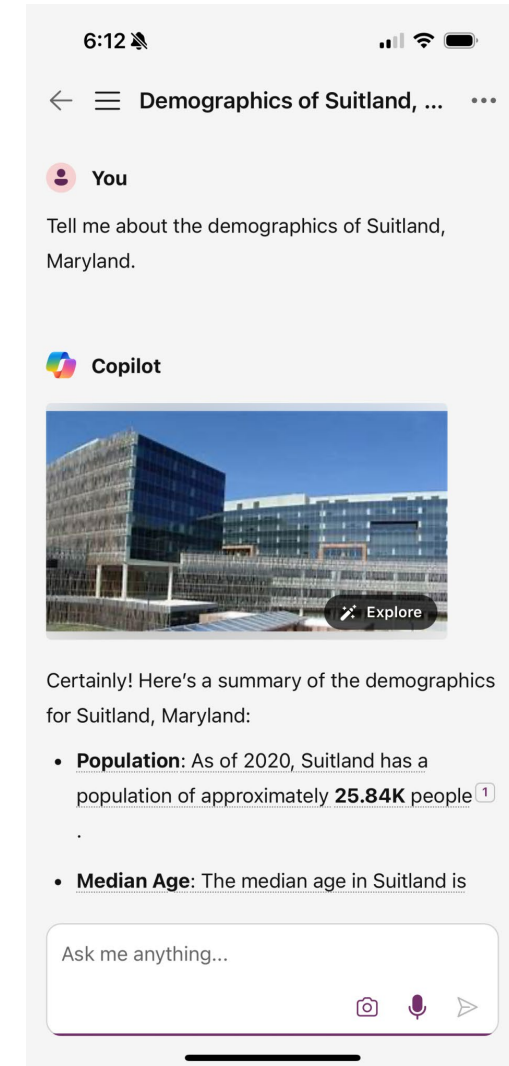
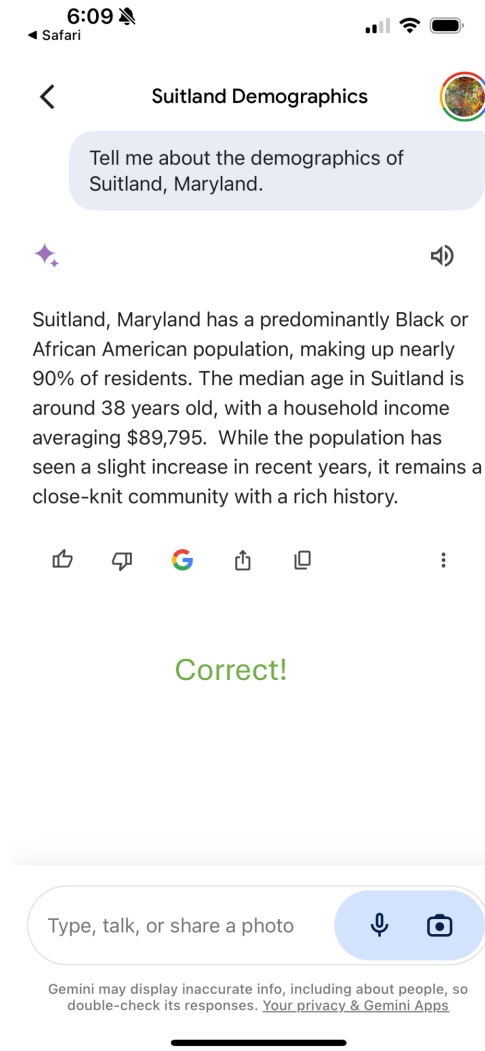
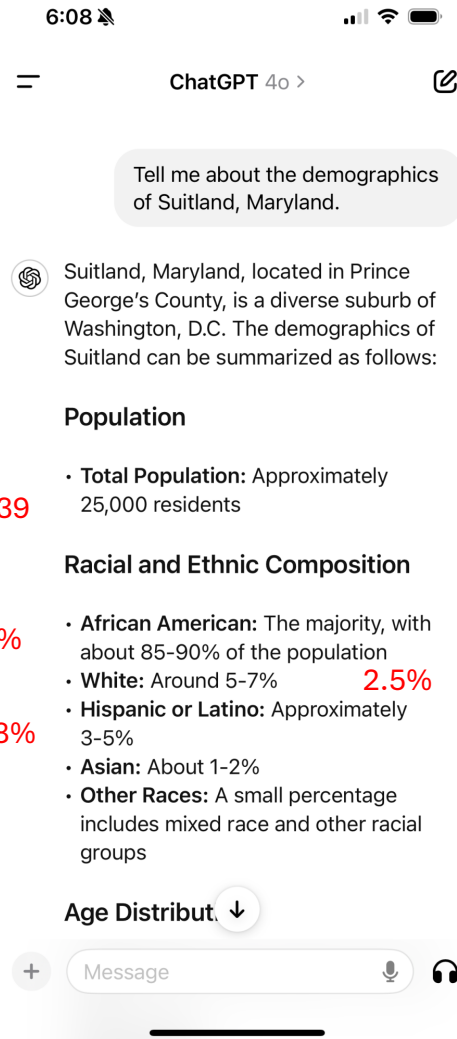
1: neighborhoodscount.com
2: census.gov (QuickFacts)
3: datausa.io

87.4%

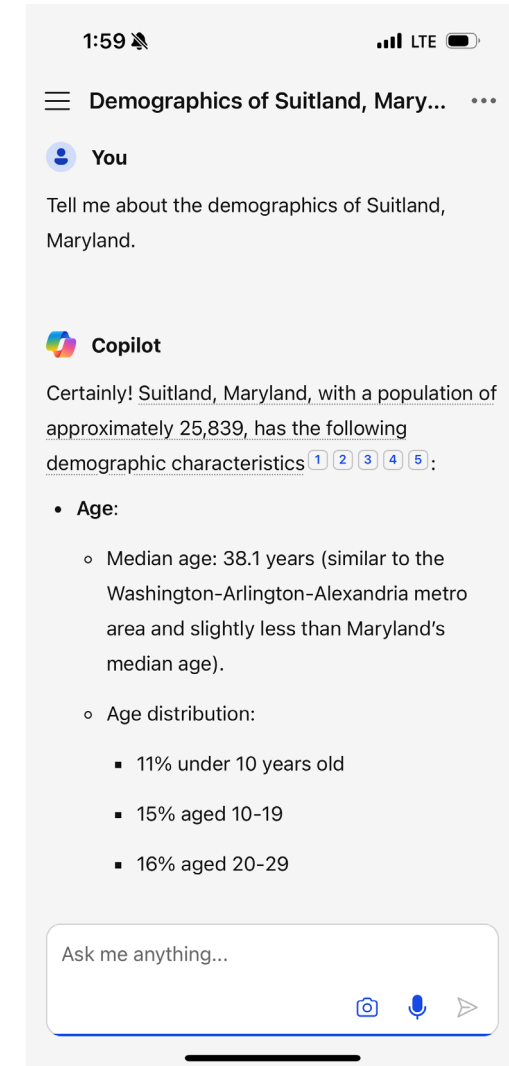
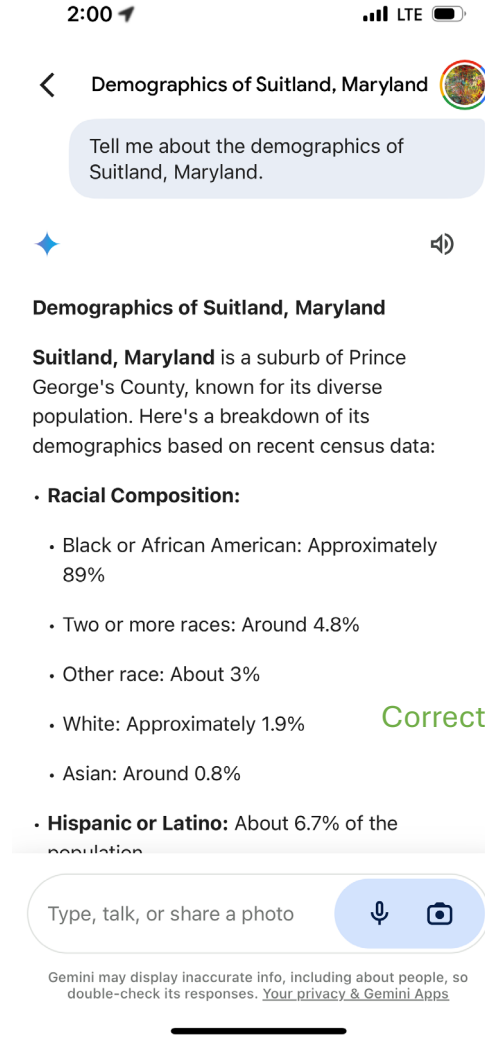
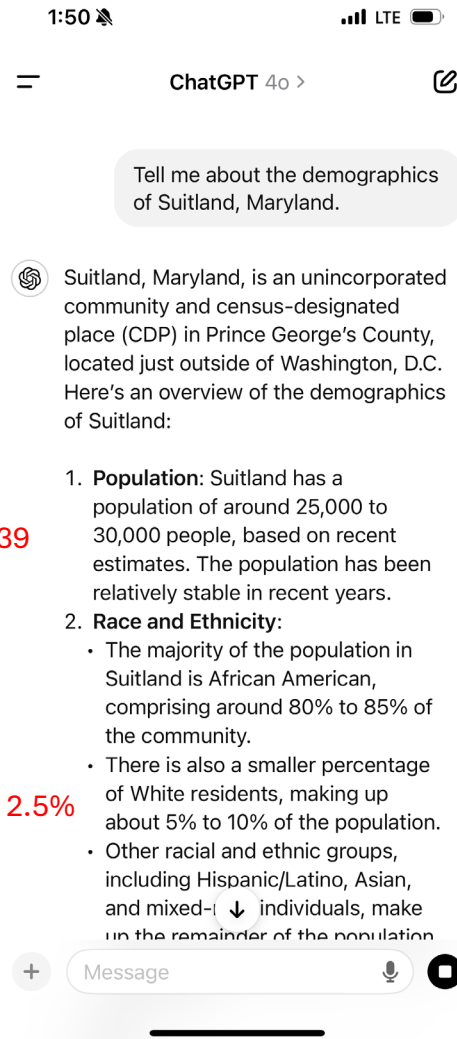
3.8%

7.8%

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

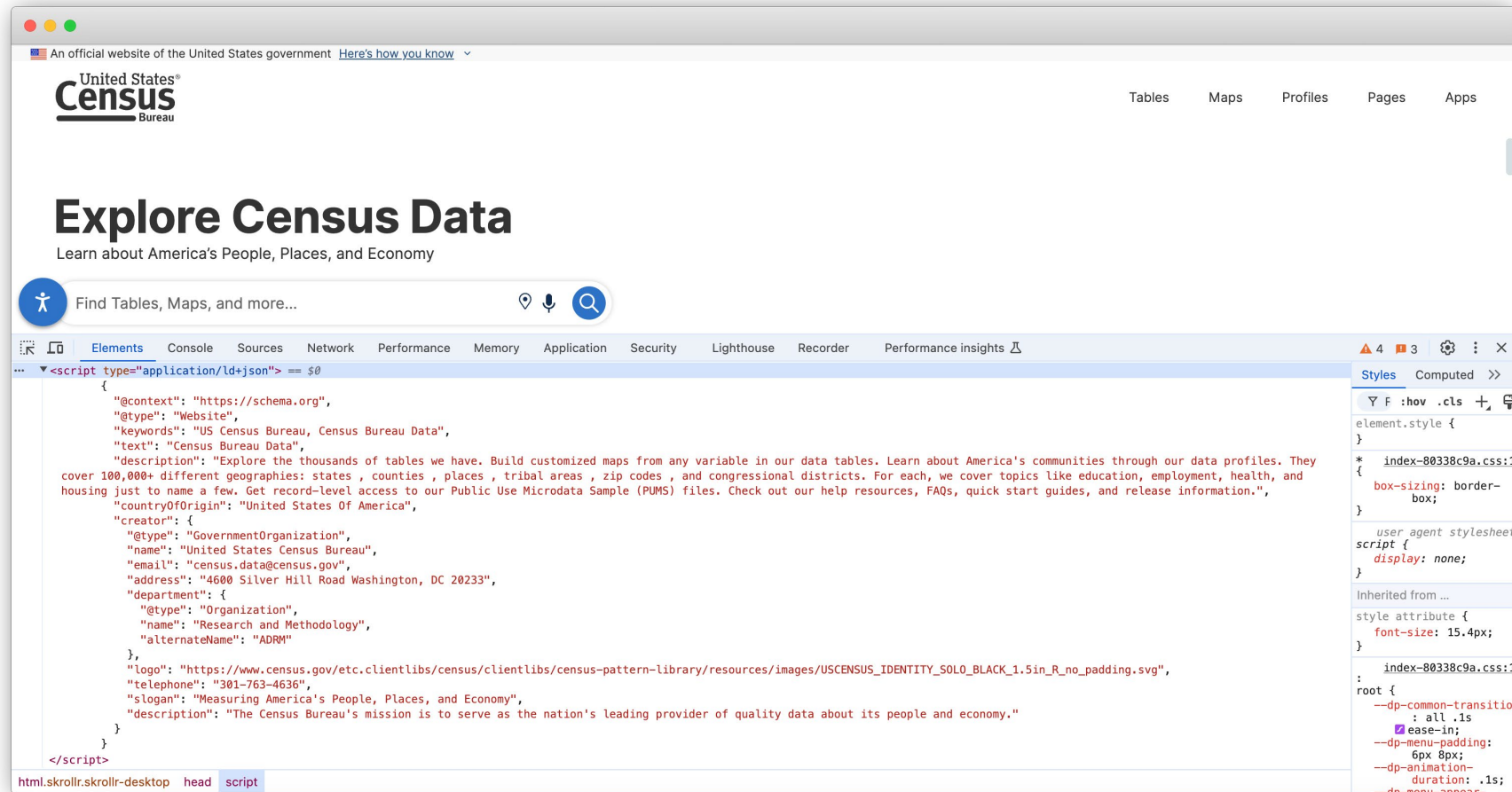
Tactic #1: SEO, but More

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

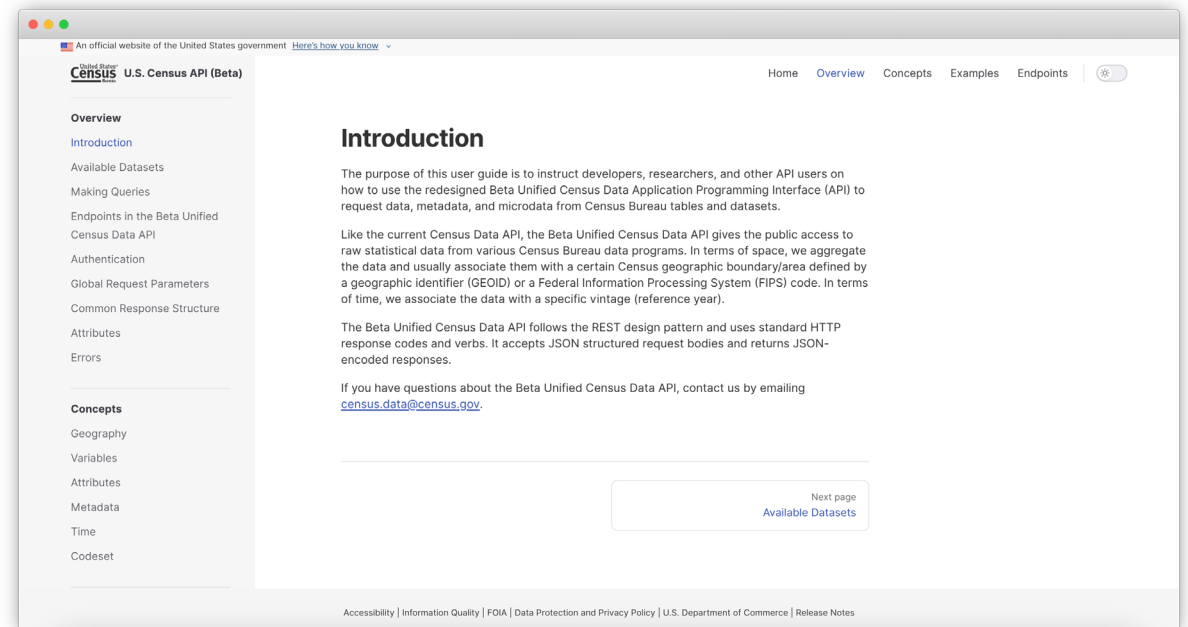
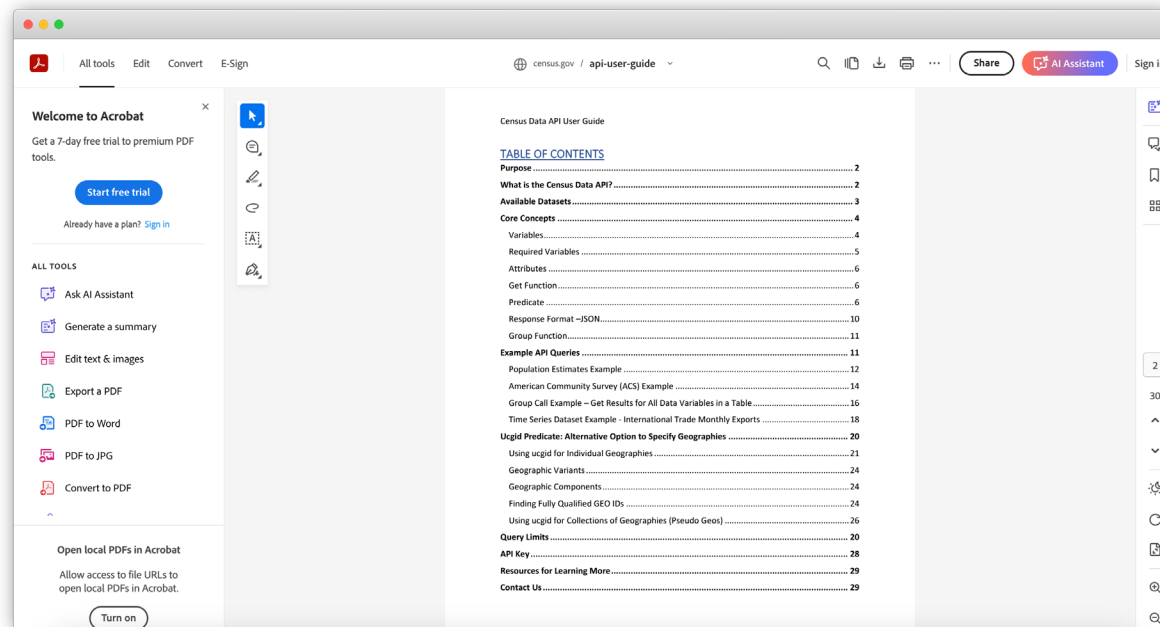
```
<?xml version="1.0" encoding="UTF-8" ?>
<sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_basics.xml.gz</loc>
  <lastmod>2024-03-08</lastmod>
</sitemap>
<sitemap>
  <loc>https://data.census.gov/sitemaps/sitemap_profiles_010.part1.xml.gz</loc>
  <lastmod>2023-8-3</lastmod>
</sitemap>
<sitemap>
  <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>
</sitemaps>
```


What Is This Data About?

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



Tell It To Me Straight: Documentation and Prose Descriptions



Tactic #2: AI-Ready Data

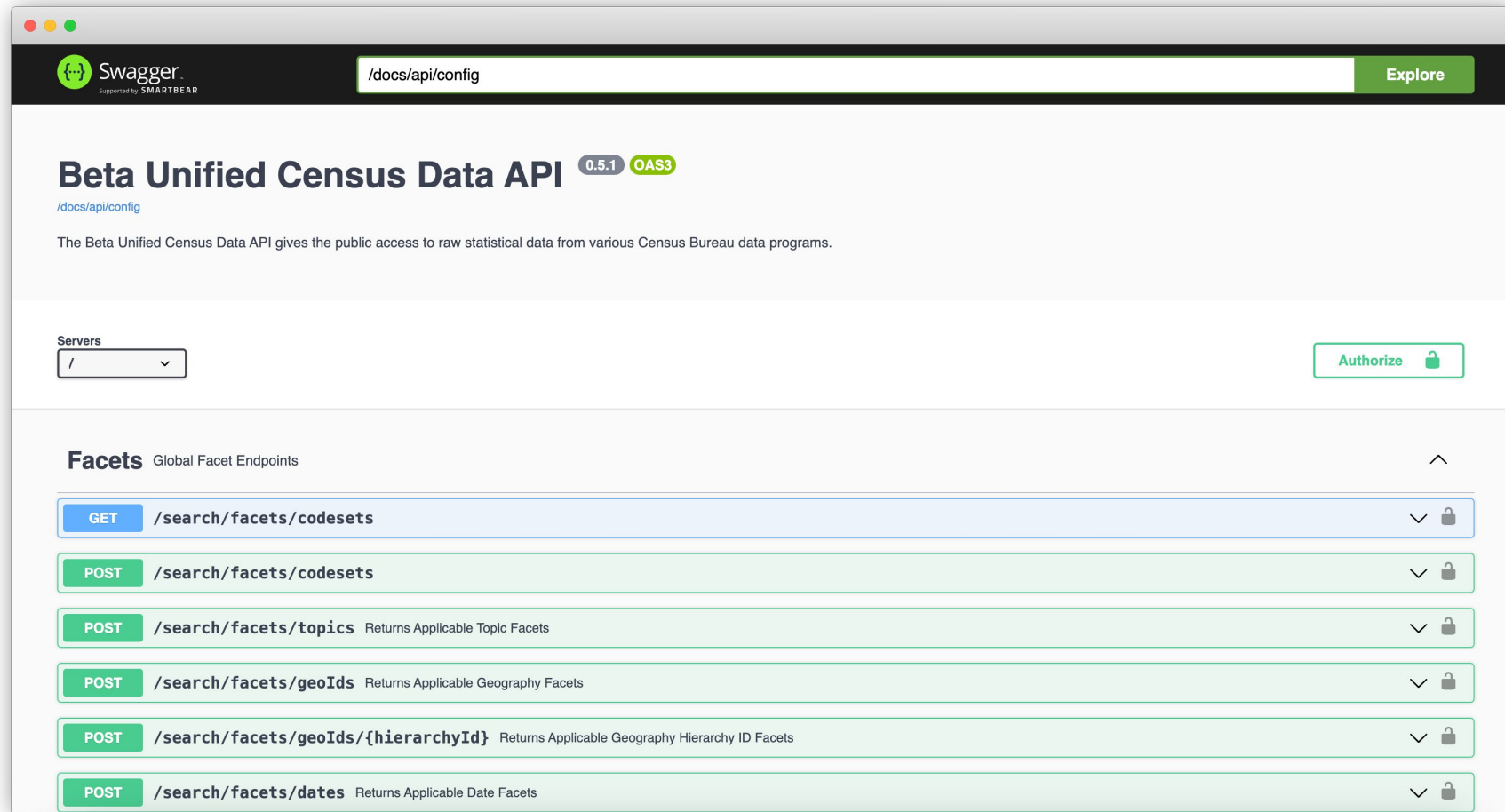
Aiding Data Ingestion: Proprietary and Unstructured < Open and Structured



VS.



Data Retrieval via API: Open Standards, Accessible Metadata



Tabular Data:

Hierarchical Tables < Flat Tables

United States Census Bureau

Search

Advanced Search

All Tables Maps Profiles Pages Apps Help FAQ Feedback

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		40,476,446	±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 Apps Help FAQ Feedback

CS2100CSP | 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: ECONSVY Business Patterns County Business Patterns

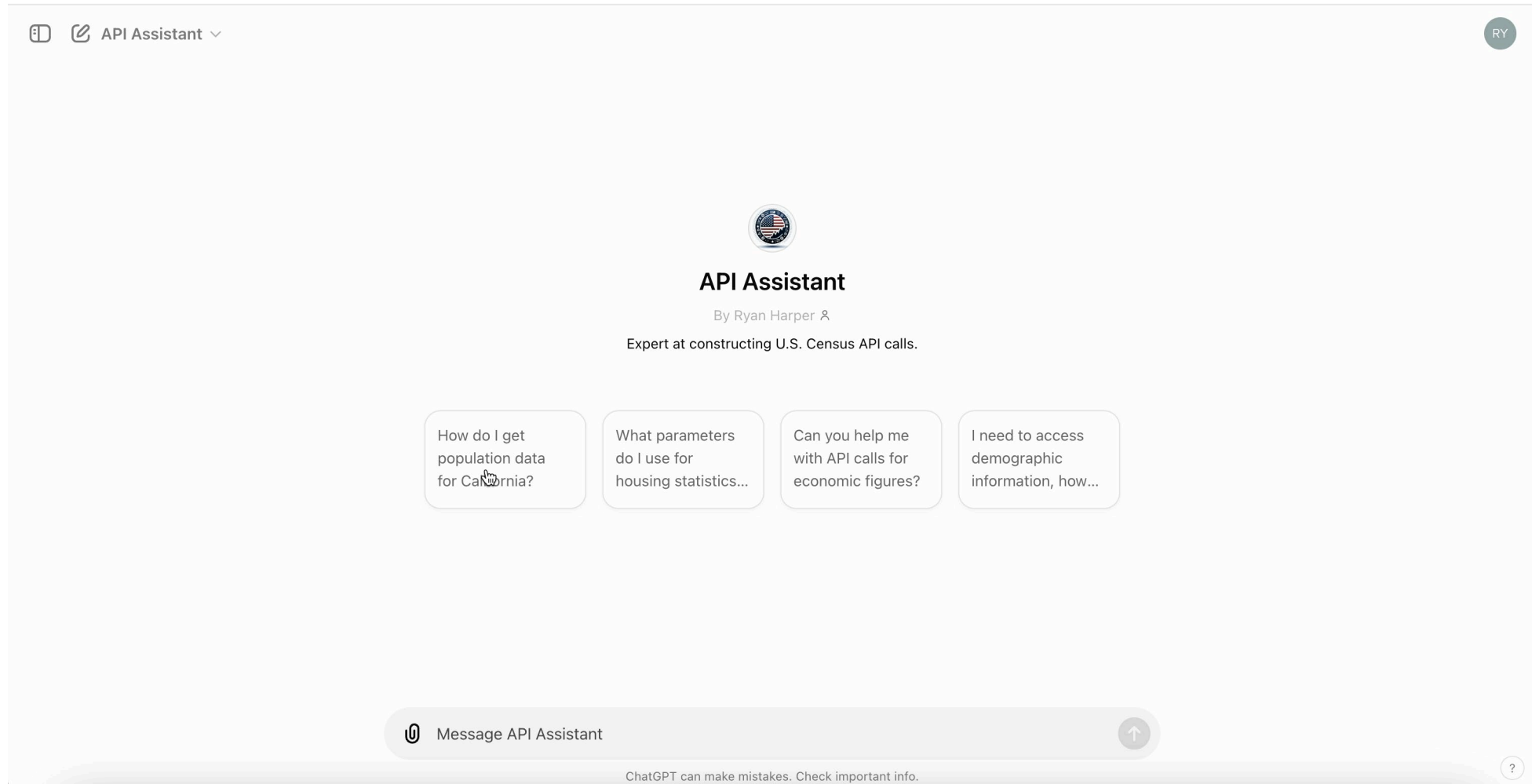
Notes Geos Topics Codes Year Columns Filters Restore Excel CSV ZIP Cite Share Print More Tools

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

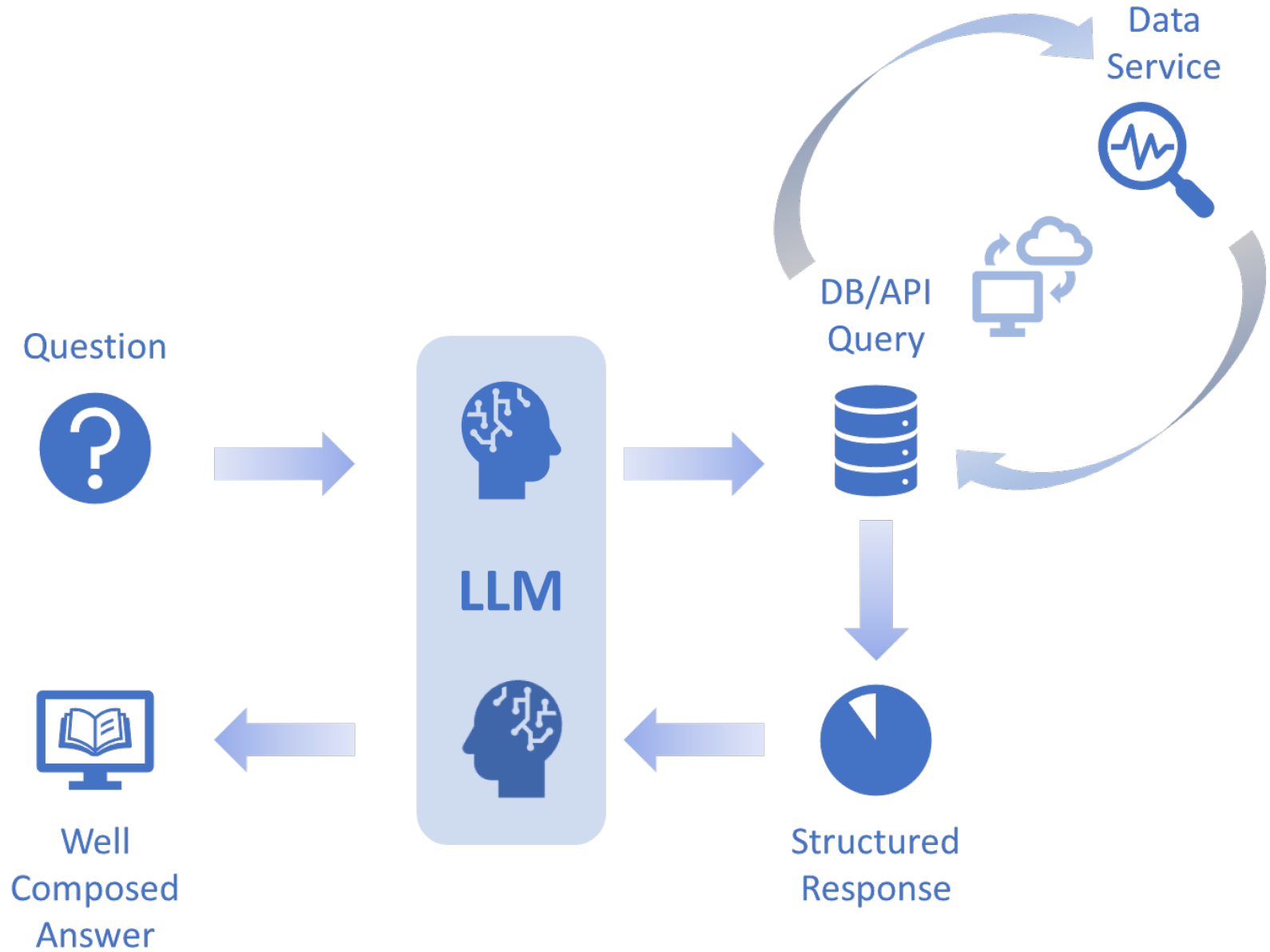
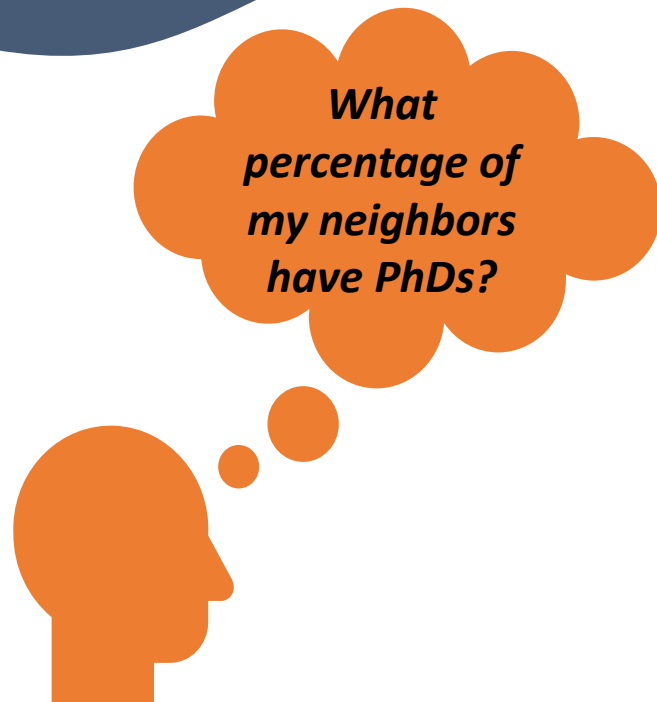
1 to 200 of 1,401 | Page 1 of 8

Tactic #3: Meet AI Systems Where They Are

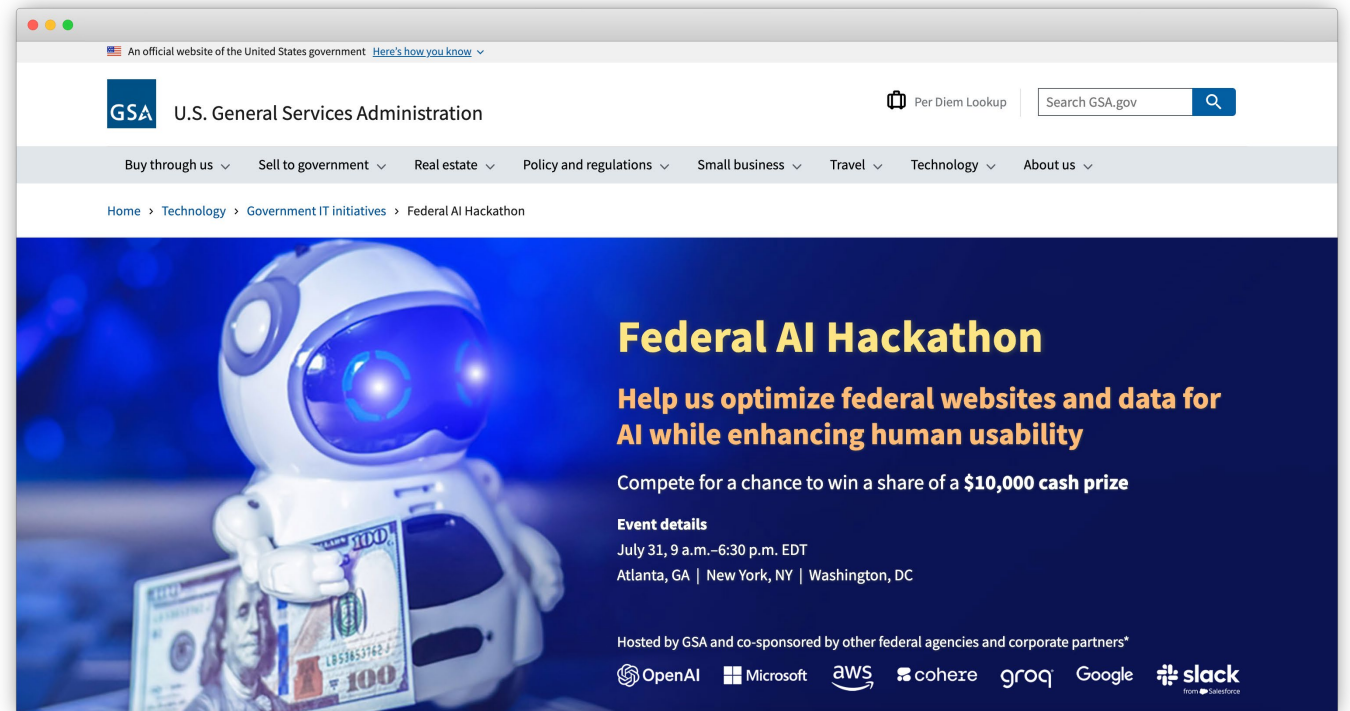
Census API Assistant GPT (Research Prototype)



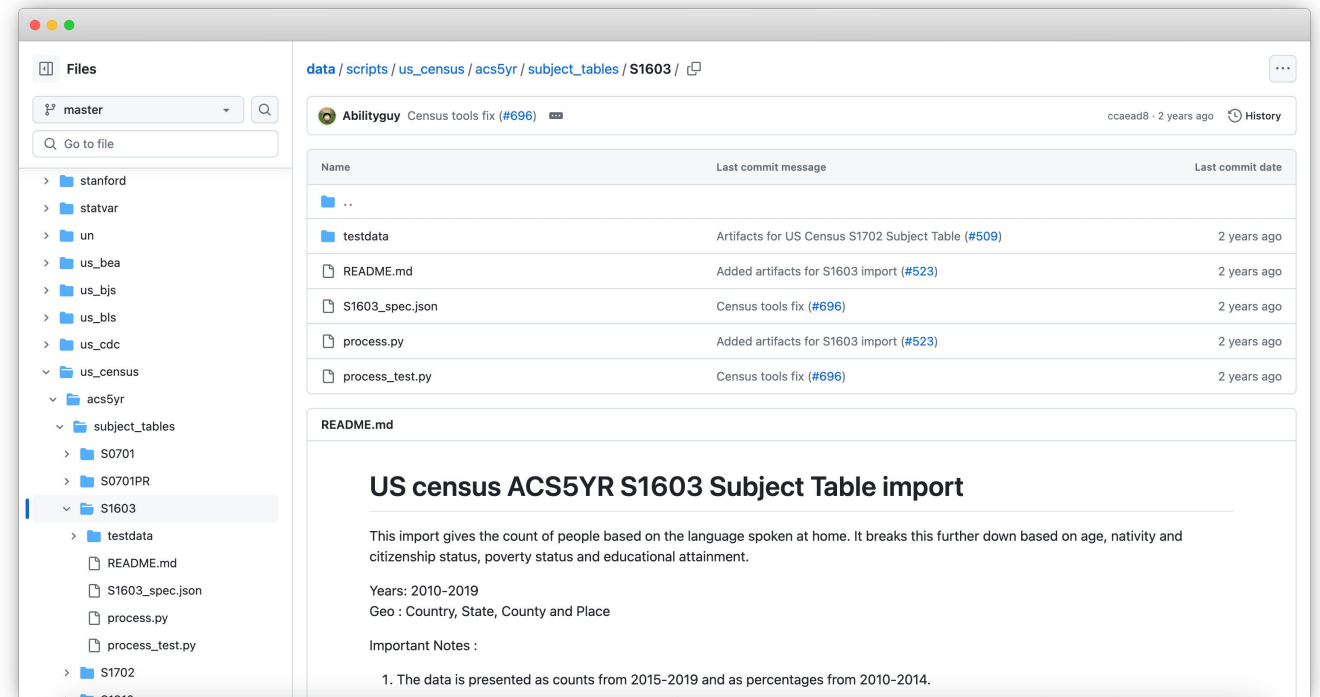
Hybrid Solutions



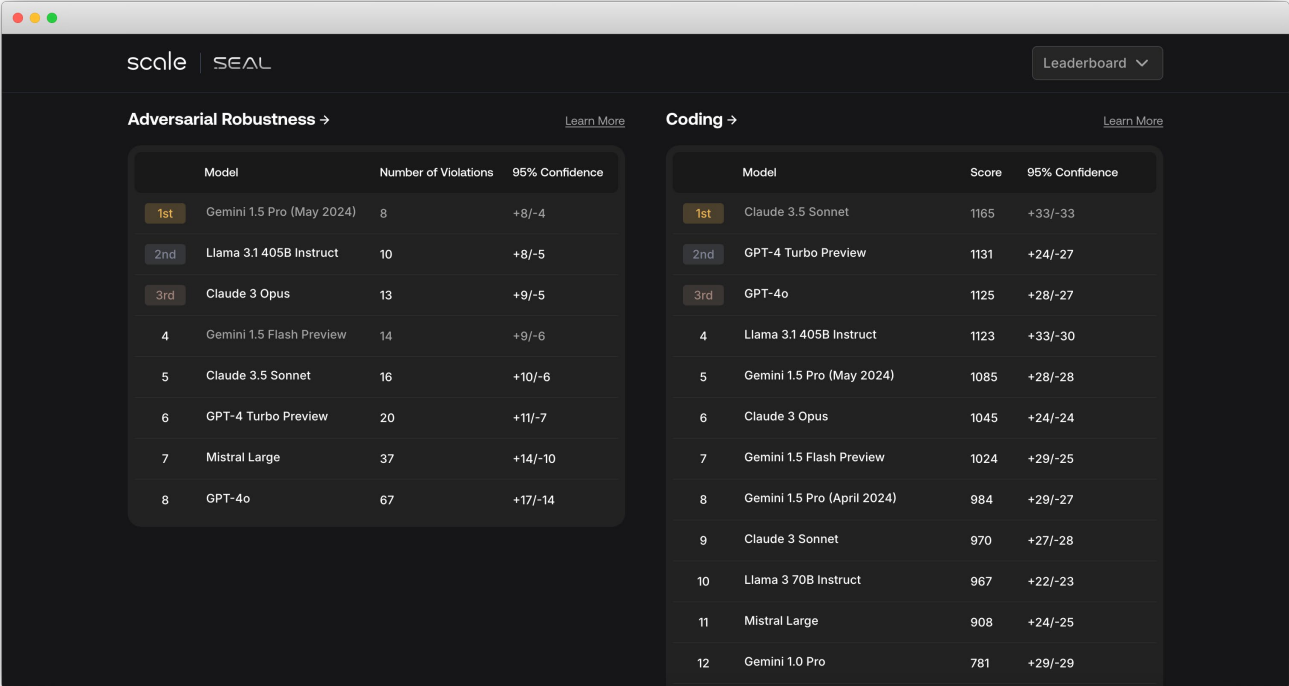
Partnership and Engagement: Community



Partnership and Engagement: Open Source Collaboration



Partnership and Engagement: Benchmarks



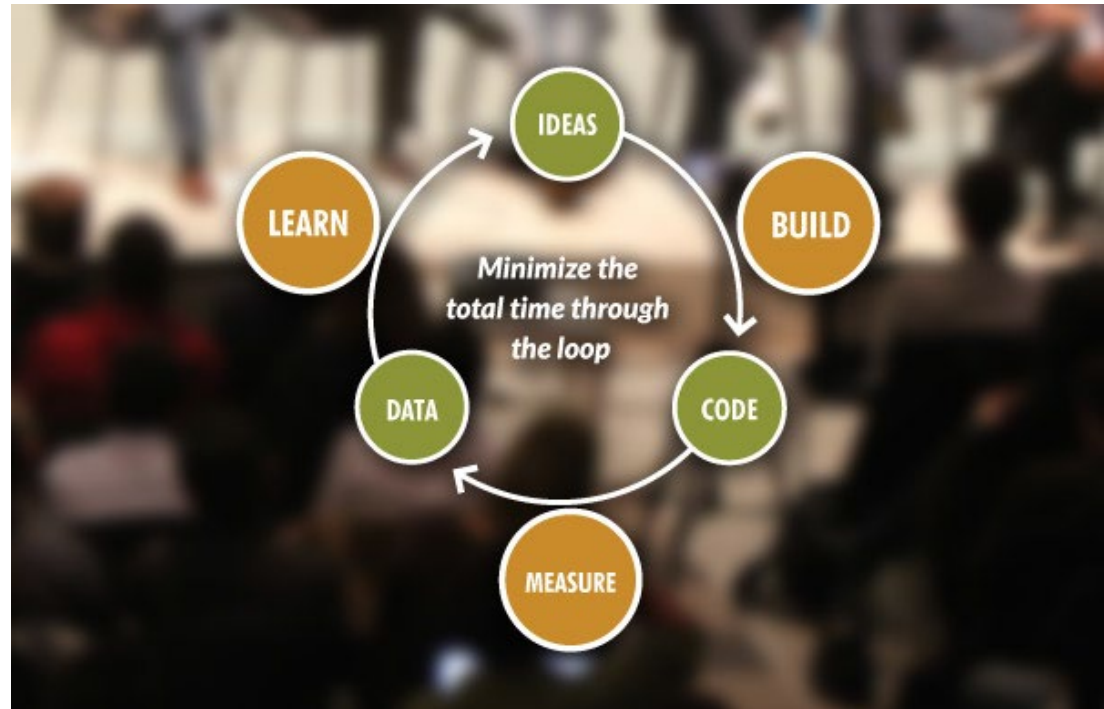
The screenshot displays the scale SEAL Leaderboard interface. It features two main sections: 'Adversarial Robustness' and 'Coding'. Each section contains a table of model performance metrics. The 'Adversarial Robustness' table lists models ranked by the number of violations, with Gemini 1.5 Pro (May 2024) at the top. The 'Coding' table lists models ranked by score, with Claude 3.5 Sonnet at the top. Both tables include columns for rank, model name, primary metric, and 95% confidence intervals. A 'Leaderboard' dropdown menu is visible in the top right corner.

Adversarial Robustness → Learn More				Coding → Learn More			
	Model	Number of Violations	95% Confidence		Model	Score	95% Confidence
1st	Gemini 1.5 Pro (May 2024)	8	+8/-4	1st	Claude 3.5 Sonnet	1165	+33/-33
2nd	Llama 3.1 405B Instruct	10	+8/-5	2nd	GPT-4 Turbo Preview	1131	+24/-27
3rd	Claude 3 Opus	13	+9/-5	3rd	GPT-4o	1125	+28/-27
4	Gemini 1.5 Flash Preview	14	+9/-6	4	Llama 3.1 405B Instruct	1123	+33/-30
5	Claude 3.5 Sonnet	16	+10/-6	5	Gemini 1.5 Pro (May 2024)	1085	+28/-28
6	GPT-4 Turbo Preview	20	+11/-7	6	Claude 3 Opus	1045	+24/-24
7	Mistral Large	37	+14/-10	7	Gemini 1.5 Flash Preview	1024	+29/-25
8	GPT-4o	67	+17/-14	8	Gemini 1.5 Pro (April 2024)	984	+29/-27
				9	Claude 3 Sonnet	970	+27/-28
				10	Llama 3 70B Instruct	967	+22/-23
				11	Mistral Large	908	+24/-25
				12	Gemini 1.0 Pro	781	+29/-29



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