

Enhancing Data Analysis with Large Language Models:

Revolutionizing Data Discovery Through Semantic Search, Summarization, and Captioning

Irina Belyaeva, Ph.D

U.S Census Bureau, Center Enterprise Dissemination

FedCASIC

April 17, 2024



Outline

- Unlocking Insights: Challenges in Federal Survey Data Analysis
- Generative Artificial Intelligence: Next Generation of Data Analysis
- Rethinking Data Analysis and Content Discovery of Federal Surveys with Large Language Models
- Challenges and lookahead

Unveiling the Obstacles: Challenges Faced in Federal Survey Data Analysis

- Surveys **Discoverability**

- Identification of relevant surveys or datasets can be challenging due to the vast amount of available data

- Surveys **Understandability**

- Interpretation of survey data requires comprehension of the context, methodology, and terminology used

Unveiling the Obstacles: Challenges Faced in Federal Survey Data Analysis

- Surveys **Discoverability**

- Identification of relevant surveys or datasets can be challenging due to the **vast amount** of available data

data.census.gov \approx **6,000** datasets

- **Insufficient** accompanying **metadata** may impede the discoverability of surveys, as it may not provide enough context or details for **effective search**

Unveiling the Obstacles: Challenges Faced in Federal Survey Data Analysis

- Surveys **Discoverability**

- Identification of relevant surveys or datasets can be challenging due to the **vast amount** of available data

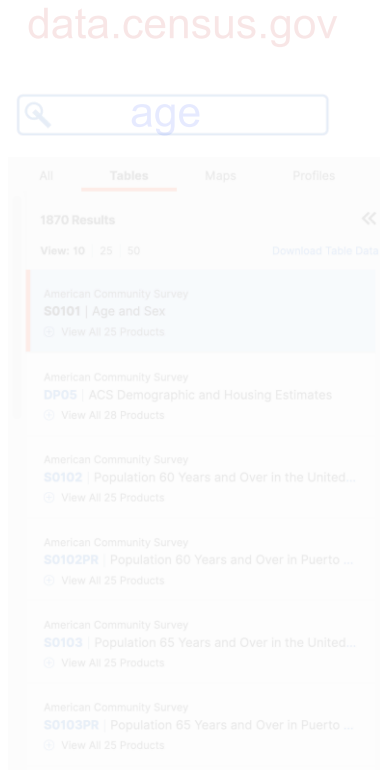
data.census.gov \approx **6,000** datasets

- **Insufficient** accompanying **metadata** may impede the discoverability of surveys, as it may not provide enough context or details for **effective search**

Unveiling the Obstacles: Challenges Faced in Federal Survey Data Analysis

- Surveys **Understandability**

- **Interpretation** of survey data/search requires comprehension of the **context**, **methodology**, and **terminology** used



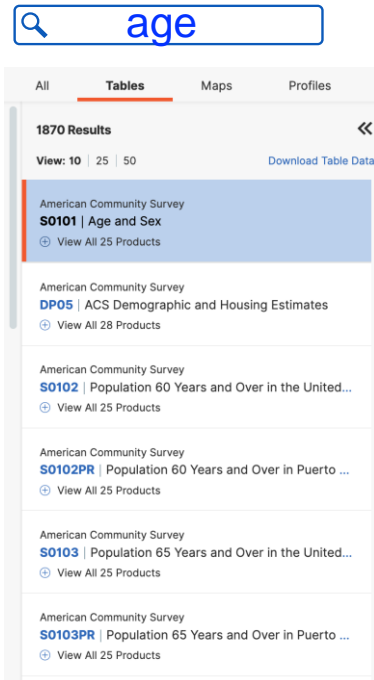
◀ How can we understand the **essence** of the survey search being presented?

Unveiling the Obstacles: Challenges Faced in Federal Survey Data Analysis

- Surveys **Understandability**

- **Interpretation** of survey data/search requires comprehension of the **context**, **methodology**, and **terminology** used

data.census.gov



The screenshot shows the search results for 'age' on data.census.gov. The search bar contains 'age' and the results are filtered to 'Tables'. There are 1870 results. The first result is 'American Community Survey S0101 | Age and Sex' with a 'View All 25 Products' link. Other results include 'American Community Survey DP05 | ACS Demographic and Housing Estimates', 'American Community Survey S0102 | Population 60 Years and Over in the United...', 'American Community Survey S0102PR | Population 60 Years and Over in Puerto ...', 'American Community Survey S0103 | Population 65 Years and Over in the United...', and 'American Community Survey S0103PR | Population 65 Years and Over in Puerto ...'. Each result has a 'View All 25 Products' link.

◀ How can we understand the **essence** of the survey search being presented?

Tackling the Challenge: Strategies for Effective Federal Survey Data Analysis

Generative Artificial Intelligence (AI) is the state-of-the-art approach to address data *discoverability* and *semantic understanding*

- Large Language Models (LLMs)
 - How can I find **relevant** Survey(s)/Dataset(s)?

- ◀ Survey Metadata Summarization

- ◀ Semantic Search

- LLMs AI Assistants

- ◀ Semantic Captioning/Search Overview

- ◀ Fact Finding Assistants



Tackling the Challenge: Strategies for Effective Federal Survey Data Analysis

Generative Artificial Intelligence (AI) is the state-of-the-art approach to address data *discoverability* and *semantic understanding*

- Large Language Models (LLMs)
 - How can I find *relevant* Survey(s)/Dataset(s)?
 - ◀ Survey Metadata Summarization
 - ◀ Semantic Search
- LLMs AI Assistants
 - ◀ Semantic Captioning/Search Overview
 - ◀ Fact Finding Assistants

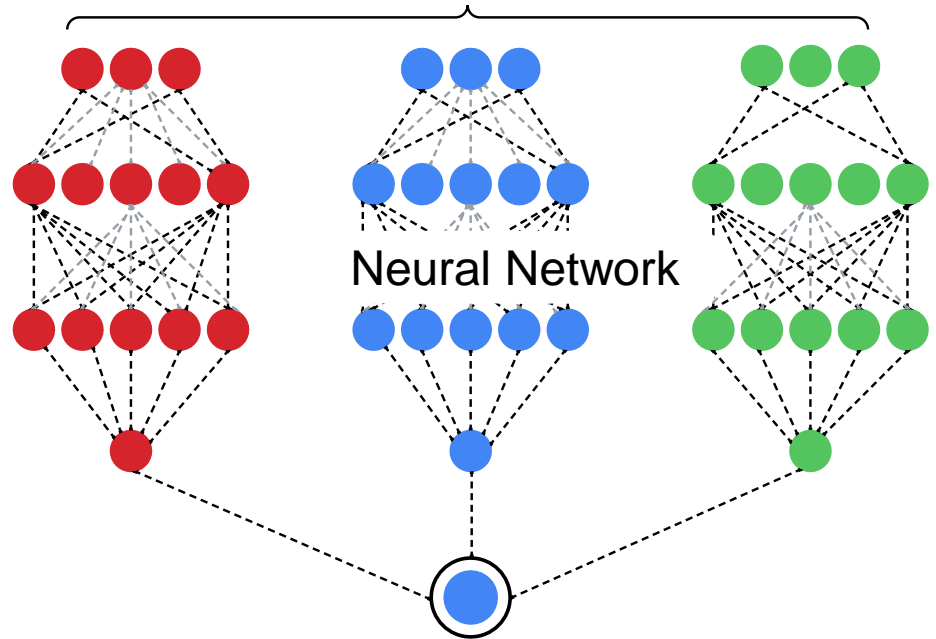
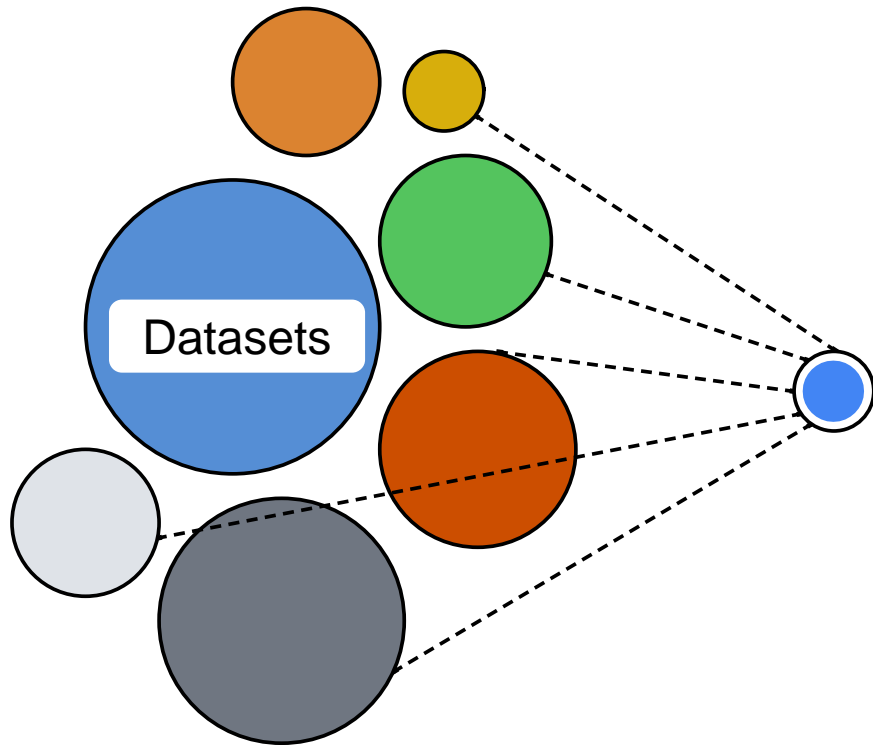
Tackling the Challenge: Strategies for Effective Federal Survey Data Analysis

Generative Artificial Intelligence (AI) is the state-of-the-art approach to address data *discoverability* and *semantic understanding*

- Large Language Models (LLMs)
 - How can I find **relevant** Survey(s)/Dataset(s)?
 - ◀ Survey Metadata Summarization
 - ◀ Semantic Search
- LLMs AI Assistants
 - ◀ Semantic Captioning/Search Overview
 - ◀ Fact Finding Assistants

What are Large Language Models?

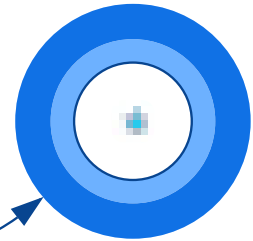
Large Language Model is a *next element neural network sequence prediction model** trained on massive datasets



What are Large Language Models?

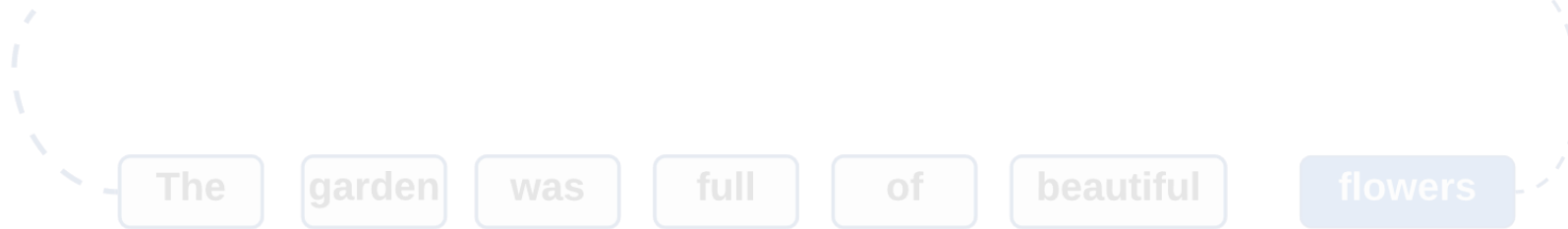
The garden was full of beautiful ?

Large Language Model (LLM)



Input

Output



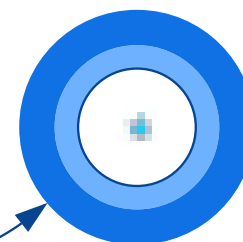
Large Language Model (LLM)



What are Large Language Models?

The garden was full of beautiful ?

Large Language Model (LLM)



Input

Output

The

garden

was

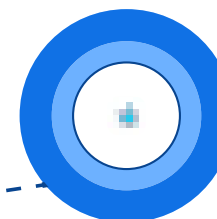
full

of

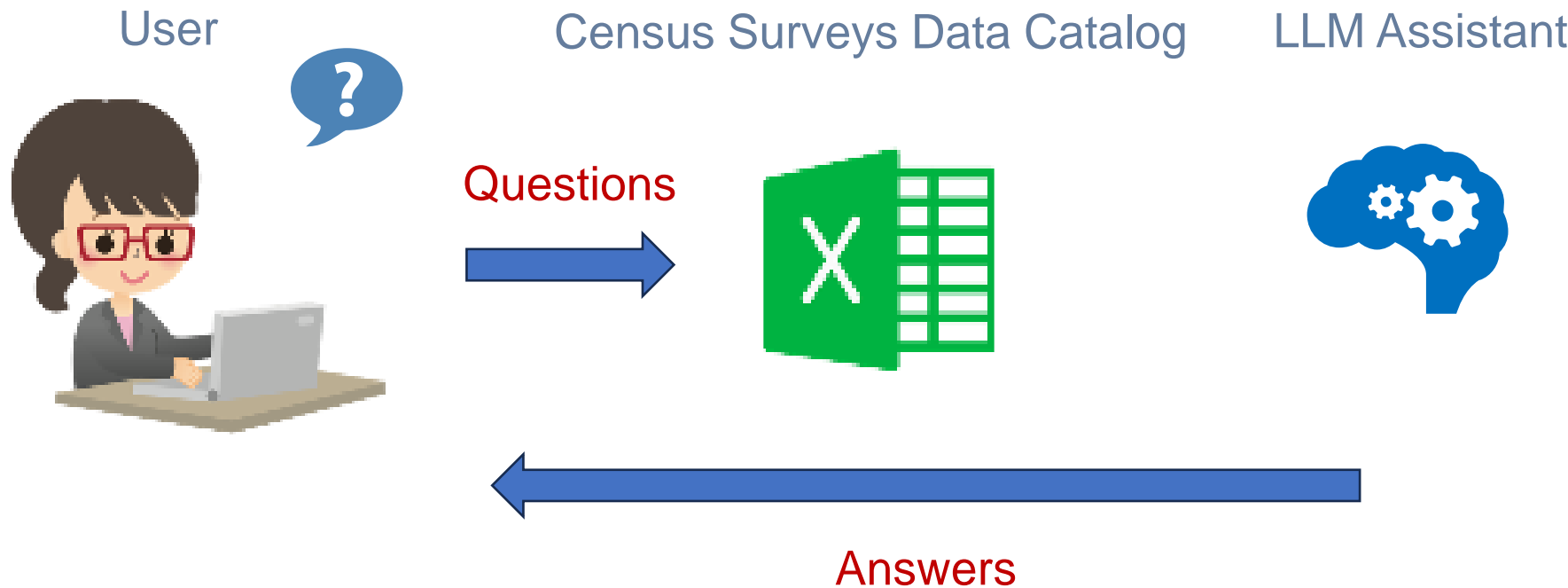
beautiful

flowers

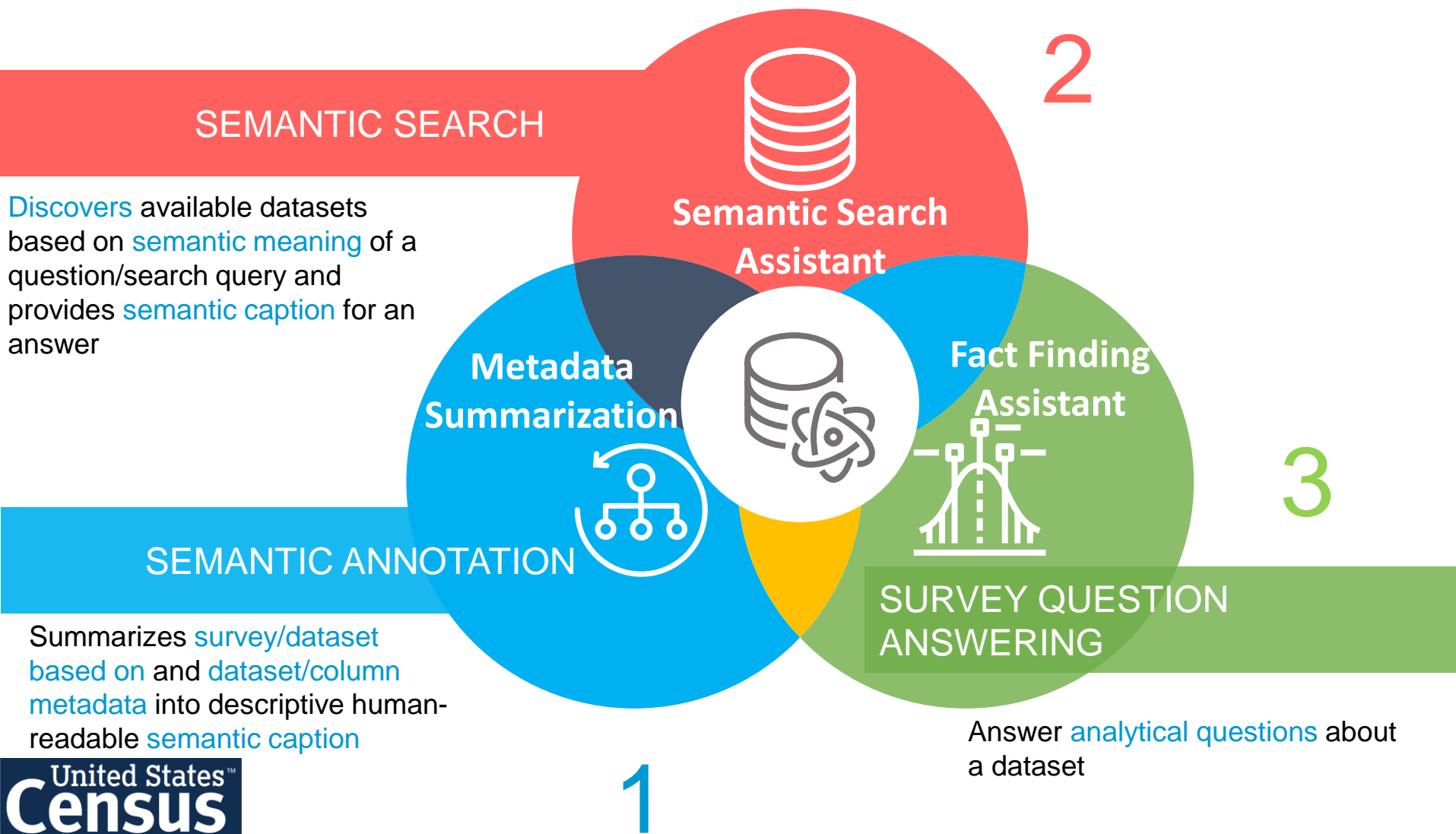
Large Language Model (LLM)



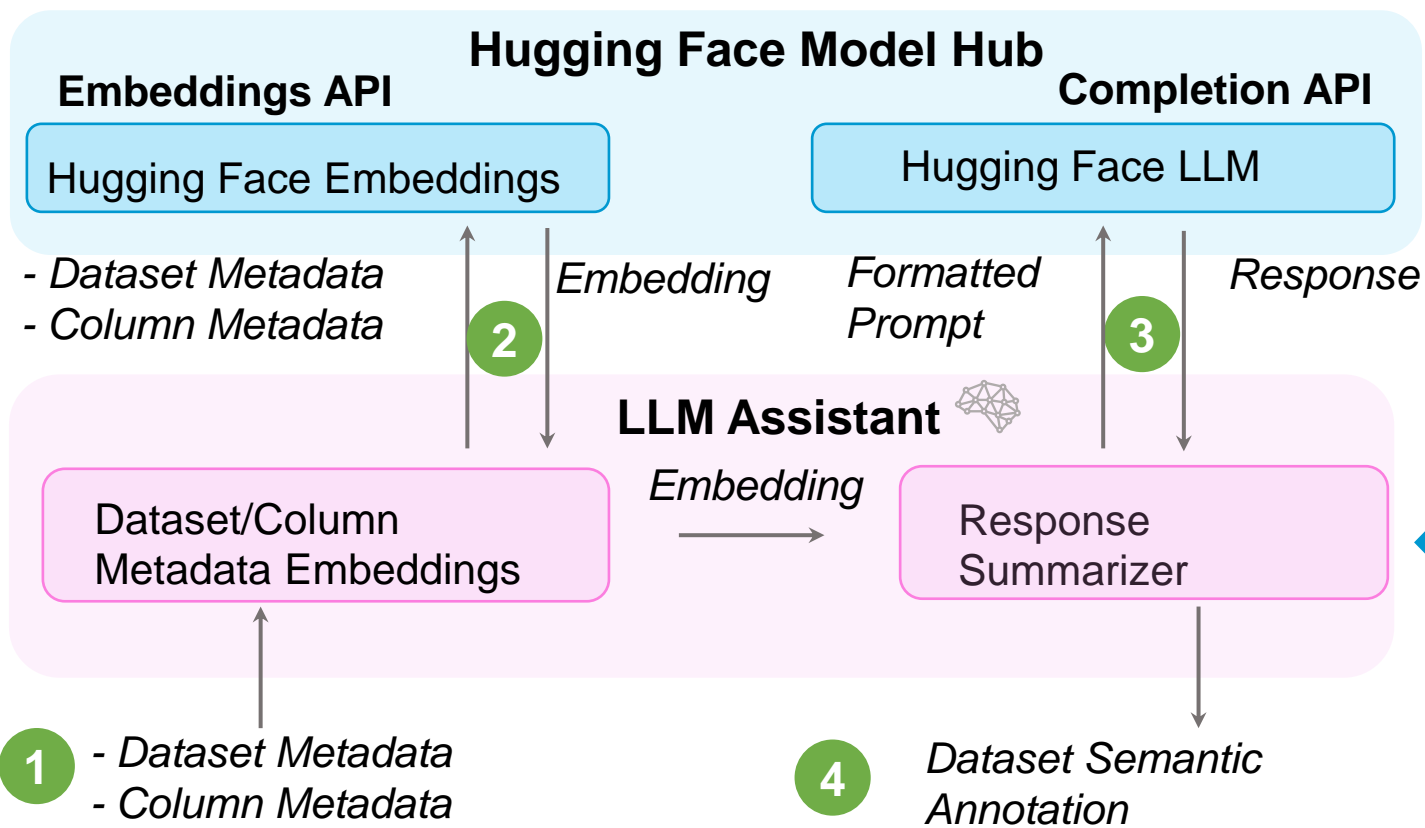
How Can We Analyze Data in Federal Surveys with Generative AI?



Federal Surveys Discovery and Data Analysis powered by LLMs



Metadata Semantic Annotation Task Architecture for Census Surveys



1 - 2 Embedding Phase

-Dataset metadata (name and title),and column descriptors are sent to the **Embeddings API**.

- Dataset Embedding is obtained.

3 - 4 Summarization

-Embeddings and formatted prompt are sent to the **Completion API**.

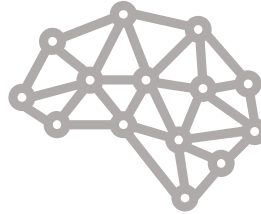
-**Response Summarizer** returns *Dataset Semantic Annotation*

LLM Enabled Metadata Semantic Annotation Task

- Input:**
- Dataset short name
 - Column List
 - Other metadata

LLM Annotation Agent

“<LLM Prompt>”



Output:

- Extended Dataset Semantic Caption

Census Data Catalog Search

Census Survey/Dataset Discovery

LLM Metadata Annotation in Action



“American Community Survey Datasets”



Dataset ID: B01001

Dataset Title: Sex by Age

Columns: 100 Columns Metadata



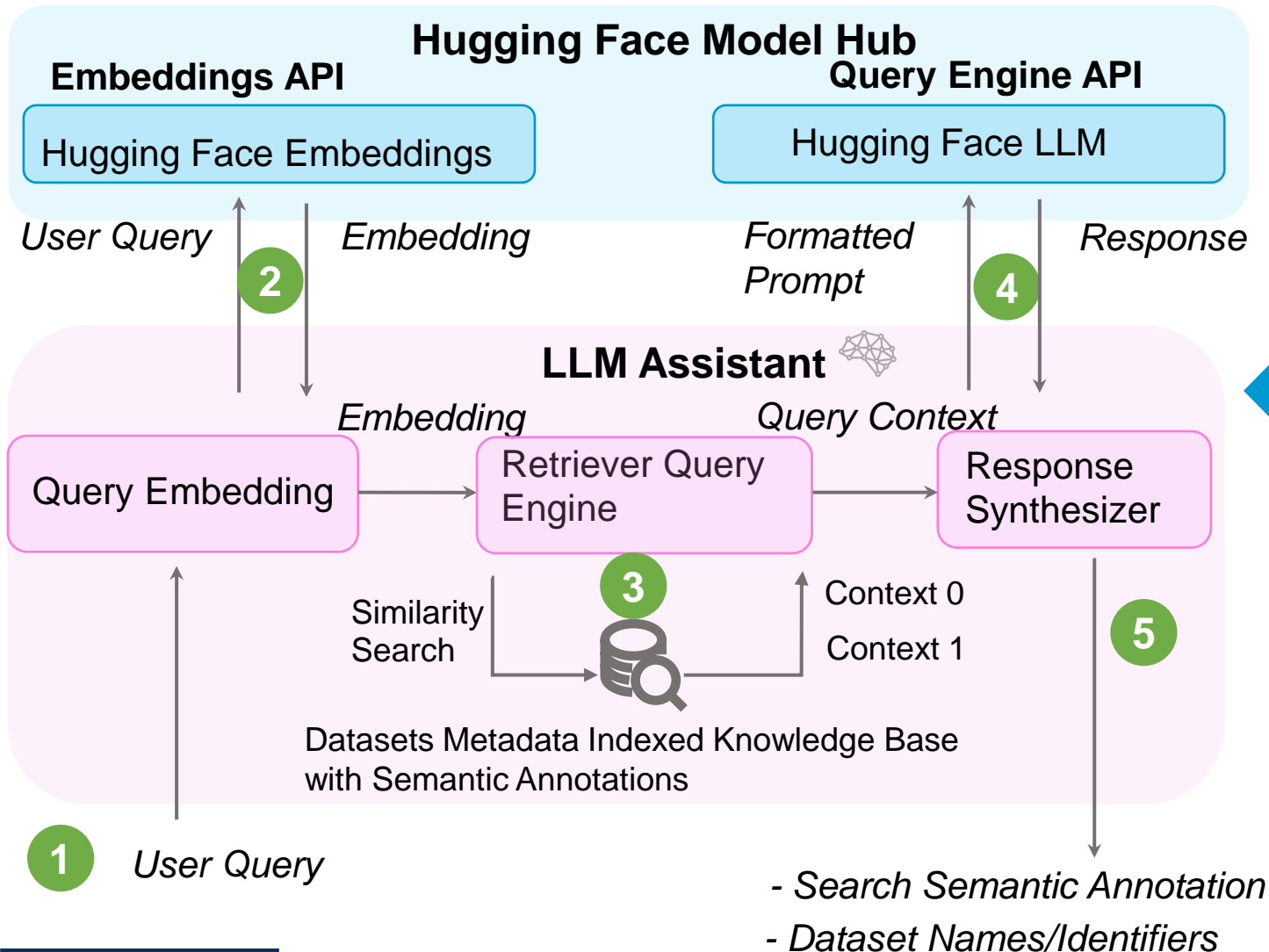
LLM generated

Dataset Semantic Caption:

This table provides estimates and margins of error for the total population by sex and age. The age categories range from under 5 years to 85 years and over. The data is from the American Community Survey, 1-Year Estimate for 2022.

Semantic Search Architecture for Census Surveys

Retrieval Augmented Generation (RAG)



1 - 2 Embedding Phase

-User Query is sent to the **Embeddings API**.

User Query Embedding is obtained.

3 Retrieval

-Results are retrieved via **Semantic Similarity Search** from Datasets Semantic Index.

-Relevant Query Contexts are sent **Query Engine** for the *Search Semantic Caption* Generation.

4 - 5 Response Captioning

-**Response synthesizer** **formats** final response and returns Dataset Names along with the Search overview.

LLM Enabled Semantic Search

Input: <User Question>

LLM Census Surveys
Semantic Search Agent
“<LLM Prompt>”



Output:

- Human-readable Search Semantic Caption
- Dataset Names/Titles/Identifiers



U: What datasets are available in the American Community Survey related to age?



Query Retriever -> Sub questions generation

Final Answer



Q1: By first identifying and quoting the most relevant sources, What is the age distribution in the American Community Survey? ⚡

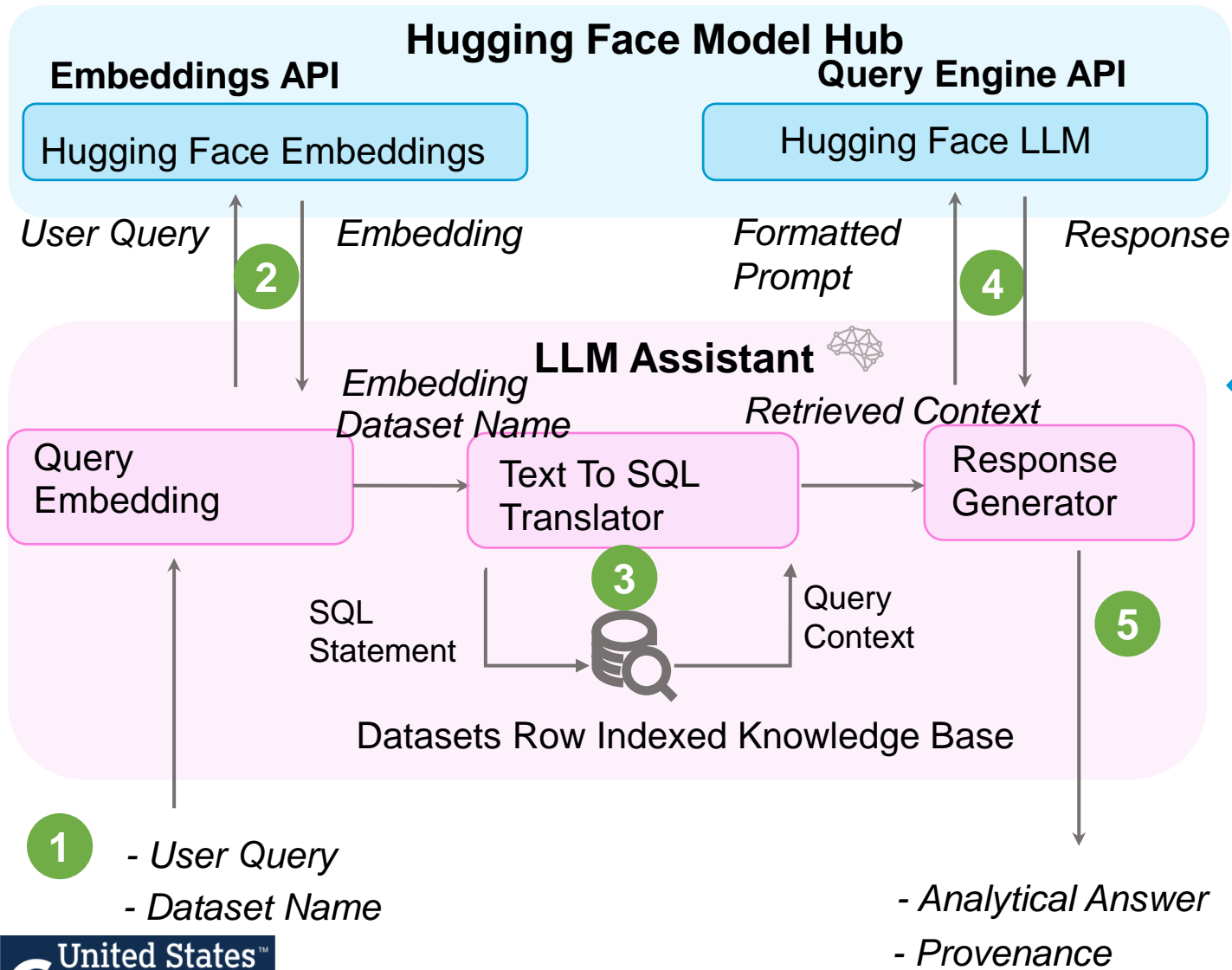
Q1: By first identifying and quoting the most relevant sources, What are the dataset IDs related to age in the American Community Survey?

The dataset IDs related to age in the American Community Survey are as follows:

1. "S0101" - Age and Sex
2. "S0102" - Population 60 Years and Over in the United States
3. "B01001" - Sex by Age
4. "C01001E" - Sex by Age (Native Hawaiian and Other Pacific Islander Alone)
5. "B05003" - Sex by Age by Nativity and Citizenship Status

Dataset Question/Fact Finding Architecture

Structured Data Question Answering via RAG



1 - 2 Embedding Phase

-User Query is sent to the **Embeddings API**.

User Query Embedding is obtained.

3 Question Translation / Retrieval

-User Question is translated into **Structured Query Language (SQL)**

-Answer is retrieved from the **Datasets Row Index**

4 - 5 Response Generation

-**Response Generator** formats response and returns *Analytical Answer*

LLM Enabled Dataset Fact Finding Task

Input: <User Question>

LLM Census Surveys
Search Agent

Output:

Dataset Name

"<LLM Prompt>"



- Numerical Answer if applicable
- Provenance



U: What are the top 5 states with the highest population in the United States?



Final Answer: Source: "S0101" - Age and Sex, 2022

State	Population, M
California	39029342
Texas	30029572
Florida	22244823
New York	19677151
Pennsylvania	12972008



U: What is the total population of males between 15 and 17 years old in the United States?



Final Answer: Source: "B01001" - Sex by Age, 2022

According to the US Census Bureau, in 2022, there were 6,655,455 youth aged 15-17 in the US.

How Do We Evaluate LLMs Performance?

Task Adapted Machine Learning Metrics and Responsible AI Metrics

Metadata Semantic Annotation

Semantic Summarization Metrics



- Annotation Quality/Semantic Overlap [1]
- Bert Semantic Similarity Score
- F1, Precision, Recall (Semantic version)

Semantic Search

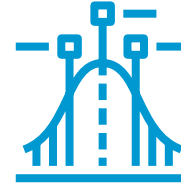
Relevancy Metrics



- Query Context Relevancy
- Recall @ K
- Precision @ K

Fact Finding Task

Accuracy Metrics



- Answer Accuracy
- Accuracy

All Tasks

Responsible AI Metrics



- Fairness, Bias and Safety [1]
- Coherence
- Toxicity
- Fairness
- Hallucinations

How Do We Evaluate LLMs Performance?

Task Adapted Machine Learning Metrics and Responsible AI Metrics

Metadata Semantic Annotation

Semantic Summarization Metrics

Semantic Search

Relevancy Metrics

Fact Finding Task

Accuracy Metrics

All Tasks

Responsible AI Metrics

- Annotation Quality [1]

- Query Context Relevancy

- Answer Accuracy

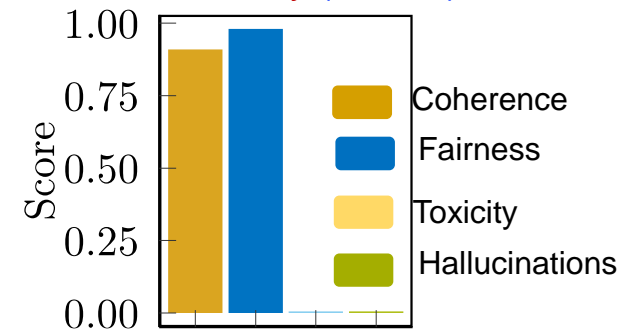
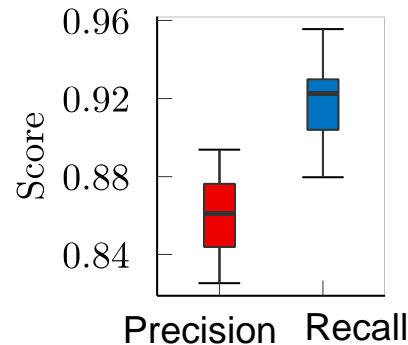
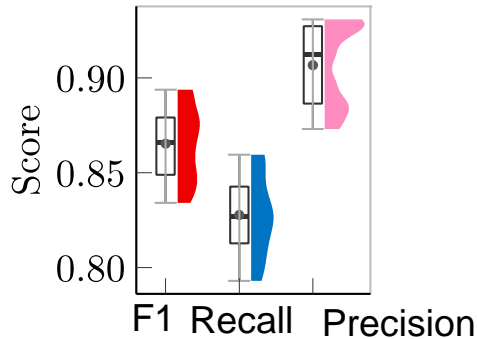
- Fairness, Bias and Safety [2]

Precision (92%), Recall (83%), F1 (87%)

- Recall @ K (92%)
- Precision @ K (86%)

- Accuracy (89%)

- Coherence (91%)
- Fairness (99%)
- Hallucinations (0.01%)
- Toxicity (0.00%)



What Are the Challenges in LLM-powered Survey Data Analysis?

- Tabular Content Modeling for **Discoverability**
- Model **Fine-tuning** for out of Domain Tasks
- Limited **Structural Understanding** of Tabular Data by LLMs
- **Risk mitigation**/LLM applications safety guardrails

Challenges in Comprehending Tables Structures by LLMs

Dataset: B01001 "Sex by Age"

	Geography	Geographic Area Name	Total	Total Male	Margin of Error:Total Mal:	Total:Male:Under 5 years	Margin of Error:Total:Male:Under 5 years	Estimate:Total:Male:5 to 9 years
0	0100000US	United States	333287562	165228214	33974	9394890	17175	10110917
1	0400000US01	Alabama	5074296	2461248	6178	146169	3134	158767
2	0400000US02	Alaska	733583	385667	2351	23043	1511	25916
3	0400000US04	Arizona	7359197	3678381	2695	201423	1573	221769
4	0400000US05	Arkansas	3045637	1504488	4216	90239	2661	98535

?

Table Parser

The headings of the table are:
Geography, Total, Total Male,...

Table Parser

The headings are located in the **first** row of the table.

What are the headings of the table?

Are they located in the first row?

LLMs Textual and Symbolic Reasoning Path

Dataset: B01001 "Sex by Age"

	Geography	Geographic Area Name	Total	Total Male	Margin of Error:Total Mal:	Total:Male:Under 5 years	Margin of Error:Total:Male:Under 5 years	Estimate:Total:Male:5 to 9 years
0	0100000US	United States	333287562	165228214	33974	9394890	17175	10110917
1	0400000US01	Alabama	5074296	2461248	6178	146169	3134	158767
2	0400000US02	Alaska	733583	385667	2351	23043	1511	25916
3	0400000US04	Arizona	7359197	3678381	2695	201423	1573	221769
4	0400000US05	Arkansas	3045637	1504488	4216	90239	2661	98535



What is the total population of the United States?



Data Retriever

To determine the total population of the United States, we need to look at the *Geography* and *Total* columns in the *B01001* table. From that we can see that **333M** people in the United States.



Answer:



The total population of United States is **333M**.

LLM Assistant Capabilities to Answer Analytical Questions about Federal Surveys

Stages

Partition and Parsing

Search & Retrieval

Capabilities

Structural Description Detection

Format Understanding

Facts Grounding

Symbolic Reasoning

Tasks

Table Partition

Table Size Detection

Hierarchy Detection

Cells Lookup

Column Retrieval

Constraints

- Search/Discovery of initial dataset
- Limited structural understanding capabilities of LLMs
- Tabular Data Modeling

Unveiling Insights: How Language Models Empower Data Analysis of Federal Surveys

▶ Supercharging Discovery

✓ *Semantic Search and Annotation* uncover hidden connections and improve **Surveys Discovery**

✓ **Automated Fact-Finding** through *Human Question to Machine Translation* bridges the gap between human questions and machine comprehension

▶ Empowering Decision-Making

✓ LLMs provide crucial **Data-Driven Insights** to support **informed policy** and strategic choices

Unveiling Insights: How Language Models Empower Data Analysis of Federal Surveys

▶ Supercharging Discovery

✓ *Semantic Search and Annotation* uncover hidden connections and improve **Surveys Discovery**

✓ **Automated Fact-Finding** through *Human Question to Machine Translation* bridges the gap between human questions and machine comprehension

▶ Empowering Evidence based Decision-Making

✓ LLMs provide crucial **Data-Driven Insights** to support **informed policy** and strategic choices



Thank you!