

Laura Ridenour, University of Missouri
Margaret Kipp, University of Wisconsin-Milwaukee
Heather Thach, University of Missouri
Liya Zhu, University of Wisconsin-Milwaukee

Census.gov Data, from Paper Tables to APIs: A Retrieval Augmented Generation Domain Analysis

1. Introduction

The United States Census Bureau has officially been operating since 1902, but the US government has been collecting and providing the American people with facts and figures in some capacity since the early 1600s (United States Census Bureau, 2025). Electronic access to the 1960 Census Data was made available to the public in the form of purchasable magnetic computer tapes in 1962 (US Census Bureau, 1962) and was the first census to be tabulated entirely by computer (Library of Congress, ND). Census data has been online since the 1990 Decennial Census, which used PDFs to make data available (Library of Congress, ND). Data currently available online spans the years 1986-some of 2025 (api.census.gov), with exceptions for privacy. The American FactFinder was used for the 2000 and 2010 Decennial Censuses, which was decommissioned in 2020 in favor of data.census.gov.

Census data is a “public good,” a term derived from economics “as a product or service that is non-rivalrous and non-excludable” (Hodge, 2015, p.24). Non-rivalrous means that use of the good by one person does not deplete the amount of resource available for others, and non-excludable relates to the wide availability of the good. In this case data is infinitely reusable, and the open nature of access to the data. It is important to note that “no two censuses are exactly alike” due to the nature of the tabulation, as censuses must be adapted to reflect and measure changes in the nations’ population and demographics (Library of Congress, ND, para. 3).

The current Census data site works well for providing statistics and a quick natural language search, but this search returns several tables possibly related to the query, the data is returned in a pivot table format, and extraneous variables such as standard deviation for each corresponding concept are included in the table returned. The Census API training videos recommends using data.census.gov to triangulate variables needed for API calls, which isn’t always practical (US Census Bureau, 2024).

Application Programming Interfaces, or APIs, are a means to connect and exchange data following a set of rules for formatting a query based on the naming conventions for concepts, organization of data, data availability, and access granted to the data requestor by means of a key. This is similar to querying and accessing data in a bibliographic database, where the key would be determined by the browser and authentication. An API is essentially a programmatically accessed Knowledge Organization System (with some caveats).

2. Literature Review

Artificial intelligence has been used to enhance the engineering of knowledge representation (Qin, 2020), and to assist in mapping interdisciplinary vocabulary (Greenberg et al., 2021). Smiraglia and Cai (2017) used the International Society for Knowledge Organization’s Knowledge organization literature database (ISKO, 2021) to examine the trends of four algorithmically-driven methods in knowledge organization literature (note that the database has not been updated since 2021). The topics selected by the authors for examination included “clustering,” “automatic classification,” “automatic indexing,” and “machine learning.” They found an increase in the use of the terms in their defined domain of knowledge organization.

Gao et al. (2023) surveyed the literature in Retrieval Augmented Generation (RAG) for Large Language Models (LLMs). They examined over 100 studies and categorized their RAG models into three types: Naive, Advanced, and Modular RAG. Modular RAG involves a combination of methods including Naive and Advanced which range from indexing-retrieval-generation to iterative RAG processes which build on each successive step in the retrieval-analysis process. Relevance judgements and prompt engineering are key to the analysis process in more advanced RAG models. Models explored by Gao et al. (2023) tend to use older models of information retrieval, but many incorporated metadata. In contrast, DeBellis et al. (2025) used modern KO tools to generate a domain-specific knowledge base for RAG to improve results from an LLM. They discovered that the use of ontologies and knowledge graphs improved retrieval by reformulation browsing.

Generative AI and ontology-driven approaches examining domains are widespread in their fields of application. Many articles in the field of biomedicine examine the use of RAG and LLMs (Liu et al. 2025) for example: literature recommendation and summarization (Li et al., 2024), improving decisions making and

knowledge sharing (Murugan et al., 2024; Zhan et al., 2025), retrieval in electronic health records (Myers et al., 2025), and question answering (Zheng, Li, and Chen, 2024).

2.1 Large Language Models and Text Analysis

Retrieval Augmented Generation is an emerging natural language processing (NLP) approach based on a machine learning framework. It intersects the generative capabilities of Large Language Models with the retrieval of external documents relevant to a user's query (Shahade & Deshmukh, 2024). This approach has been found to improve the accuracy and contextual relevance of output from generative artificial intelligence (AI) models (Gao et al., 2023). RAG is often used for text analysis related tasks. Mao et al. (2021) used RAG to provide additional context for answering open-domain questions. Their study demonstrated that RAG enhanced systems provided better retrieval accuracy when combining results from different contexts enhanced by RAG. Lewis et al. (2020) used RAG to enhanced retrieval in knowledge intensive NLP tasks. They found that RAG formulations based on pre-training parametric and non-parametric memory performed better than standard LLMs. Some of their models outperformed task-specific retrieval systems. Chirkova et al. (2024) examined multi-lingual RAG enhanced retrieval systems. Similar to the studies by Mao et al. (2021) and Lewis et al. (2020), Chirkova et al. (2024) also found that pre-training with RAG models improved retrieval efficiency. Chirkova et al. (2024) also noted that additional work will be required to create a true multi-lingual solution. Guo et al. (2025) proposed a system that would enhance the ability of LLMs to translate between their original natural language systems and more formalized systems like SQL. They used a Poincaré model to better match questions to answers and multi-pattern RAG formulation to generate the correct SQL statements to query a database form a natural language query. Their method increased accuracy by 13%.

While not explicitly RAG, Linxen et al. (2024) used SPARQL queries generated from LLM prompting to improve knowledge retrieval in the digital humanities for the Folkwang University of the Arts' Library. They followed Noy and McGuinness' (2001) Ontology Development 101 for the creation of the ontology, which encompassed material inventory, cataloging, metadata, database, and other works in addition to operational, contact, staff, and other service information for the library (Linxen et al., 2024). Testing was done using El-Korany et al.'s (2000) method for testing knowledge bases, which allows modular testing of system components and the system as a whole and is broken down further into verification, and validation.

2.2 Domain Analysis, Knowledge Graphs, and Ontologies in LLMs

Tennis (2003) defined the two axes of bounding a domain analysis by the extension, or, coverage of time and space. and intension, or degree of specialization of the coverage. Szostak (2024) suggested using concept translation to analyze, link, and describe concepts based on individual phenomena allowing for simpler concepts to be identified as larger parts of a whole and to define their roles in discipline-specific theories.

Zhang et al. (2025) surveyed the use of knowledge-graphs in graph-based RAG implementations to support domain-specific LLM applications. Yang et al. (2024) discuss the promise of Knowledge Graph enhanced LLM enhancements and the use of LLMs to assist in the construction of Knowledge Graphs, using a structure-oriented retrieval augmented generation (SRAG). They also discuss the issues of AI hallucinations (i.e. made-up content such as fake links), and the need for expert human oversight. Their suggested approach involves human experts constructing rules for LLMs which would then flag items for human oversight. Zhu et al. (2024) also used AI and SRAG to develop techniques for accountability and machine analysis of rich metadata structures in social science research datasets used by policy makers and public health officials to make informed decisions. They propose a Knowledge Graph enhanced SRAG system to ensure data from LLMs is valid, reliable, and accurate.

Sharma, Kumar and Li (2024) discussed OG-RAG which is designed to support RAG in LLM systems using domain-specific ontologies. In testing, OG-RAG increased accuracy by 55% and response correctness by 40%. Toro et al. (2024) discussed the use of dynamic retrieval augmentation of ontologies using DRAGON AI. Although their method was successful in generating additions to ontologies, evaluation by ontology experts suggested that flaws remain in AI generated content that require curation by human experts in a field.

AI Chatbots and Question Answering

Lappalainen and Narayanan (2023) reported on the development of Aisha, a custom ChatGPT chatbot for providing reference and support services at the library for after-hours questions. They found that while the chatbot provided accessible responses in a human-like manner, it still occasionally produced hallucinations despite explicit instructions not to do so. Chen et al. (2024) developed EyeGPT to provide answers to patient questions in ophthalmology. They compared fine-tuned (carefully curated medical data validated by an expert in the field) and RAG based models and found no significant difference in system performance.

He et al (2024) evaluated the quality of answers to patient queries about laboratory results using Q&A data from Yahoo! Answers and a ChatGPT based web portal which generated results from 5 different LLMs. Results were evaluated manually by medical professionals. They found that the GPT-4 chatbot's responses scored higher

than other LLMs and the human responses from Yahoo! Answers, although all the LLMs occasionally provided incorrect statements or did not provide references. They recommended prompt engineering and RAG to improve results.

Dobreski and Hastings (2025) compared several chatbots to determine their potential use for subject analysis and the assignment of classification numbers and subject headings. Overall results were poor, with many overly broad or incorrect headings. The authors suggest caution in using these tools for cataloguing, but that the tools may improve with training.

2.3 Authorship and Knowledge Representation in AI

Soos and Haroutunian (2024) examined the issues arising from AI authorship as LLMs and generative AI become increasingly common. Authorship is a core concept in KO and is more complex than a simple byline or names on a title page, and issues surrounding AI authorship only add to this complexity. Soos and Haroutunian (2024) succinctly point out the crux of the problem in identifying the division of “authorship” when an LLM is used, especially if it is proprietary, stating that “determining where the intellectual and technical labor responsible for an LLM starts and ends is deceptively difficult, a dilemma that is only further exacerbated as proprietary models become embedded within third-party programs and interfaces” (p. 85). The authors propose an authorship model based on accountability and acknowledgement of all sources. This would be similar to the Contributor Role Taxonomy (CReDiT), which is widely used in scientific publications and is a NISO recommendation (CReDiT, 2025). Netzer (2025) proposed the use of LLMs for tracking authorship in a more practical sense by using LLMs to generate BIBTEX based citation records for university presses. He found that LLMs could be used to generate low-cost citations without cloud computing costs and with increased efficiency.

Rosa et al. (2024) studied knowledge representation in Leonardo, an AI tool, by examining the results of images generated using sample queries. Generated images showed that AI replicates the dominant paradigm and may raise ethical issues surrounding the responsible representation of information.

2.4 AI Hallucinations and other Limitations

Huang et al. (2025) studied hallucinations in AI systems and developed metrics for evaluation hallucinations as well as suggestions for mitigating hallucinations. They categorized hallucinations into three categories: data-related, training-related, and inference-related and recommended category specific methods for reducing hallucinations. They also examined hallucinations in RAG enhanced LLMs and identified two major sources of these hallucinations: retrieval failure and generation bottleneck. Suggested mitigations included changes to LLM models, additional contextual background, and prompt engineering to remove ambiguity in questions.

Finally, Reiche (2023) found that the cost of using custom or in-house AI models to generate metadata for images remains cost-prohibitive for individual libraries, thus supporting a need to train existing commercial models for library-specific use.

3. Introduction to Retrieval Augmented Generation

The architecture of all RAG systems consists of a retriever and a generator. After a user inputs a query or prompt, the retriever is responsible for searching a set of predefined documents, databases, or a corpus of text for information relevant to the query (Gupta et al., 2024; Shahade & Deshmukh, 2024). To do this, the user’s query is transformed into a vector representation, and the model computes similarity scores between the query and vector representations of chunks of text from external sources. Multiple documents or relevant pieces of text are returned and ranked by relevance, or their similarity with the query (Gao et al., 2023). This information is then passed into the generator along with the query. The generator, essentially an LLM, takes the contextual information from the retriever and the input from the prompt to generate an output rooted in the retrieved content (Gupta et al., 2024; Shahade & Deshmukh, 2024). This ensures that the generated text is not only contextually accurate but also grounded in factual information sourced from the external knowledge base. In effect, RAG helps address the common limitations of traditional LLMs, such as hallucinations—where models generate plausible but factually inaccurate or misleading information—by anchoring responses in verifiable external sources (Gupta et al., 2024; Ji et al., 2023).

4. Methods

An initial analysis was conducted of the structure of the Census website, documentation, variables, API documentation, and query structure. Bing CoPilot was prompted to conduct further analysis of the Census website for the variables, API, API documentation, and data tables to determine the organization of the Census data on the website and through the API.

4.1 Prompting

The order questions were asked in was designed to ascertain whether or not the AI could be prompted to locate and generate census data tables. Prompting included:

- I. A RAG-style identity prompt: You are a skilled Python programmer and economist, as well as user of the US Census Bureau API
- II. An initial overview of the target domain using language from a manual analysis

5. Results

5.1 Human Analysis

The Census website is self-referential and contains several access points for the same or similar information (e.g. “QuickFacts” and “data.census.gov” both access community data). Of relevance to this work, the site high-level categories of Topics, Data & Maps, Surveys & Programs, and a Resource Library. The following section contains an overview of the organization and content of the four main pages of the Census Website.

Census Website Analysis

The Topics page (Figure 1) is a taxonomy of topics of interest to the census; it includes Subtopics hierarchically related to each topic (Subtopics are omitted for readability and relevance). Topics include Business and Economy, Education, Emergency Management, Employment, Family and Living Arrangements, Health, Housing, Income and Poverty, International Trade, Population, Public Sector, and Research. Each Topic link links to a page containing a brief overview of the topic and the Census’ interest in the topic (e.g. for [Education](#), “The Census Bureau provides information on a variety of educational topics, from educational attainment and school enrollment to school costs and financing” [US Census Bureau, 2024, October].) Subtopics are also listed on each Topic page and contain a description on the page to which they belong.

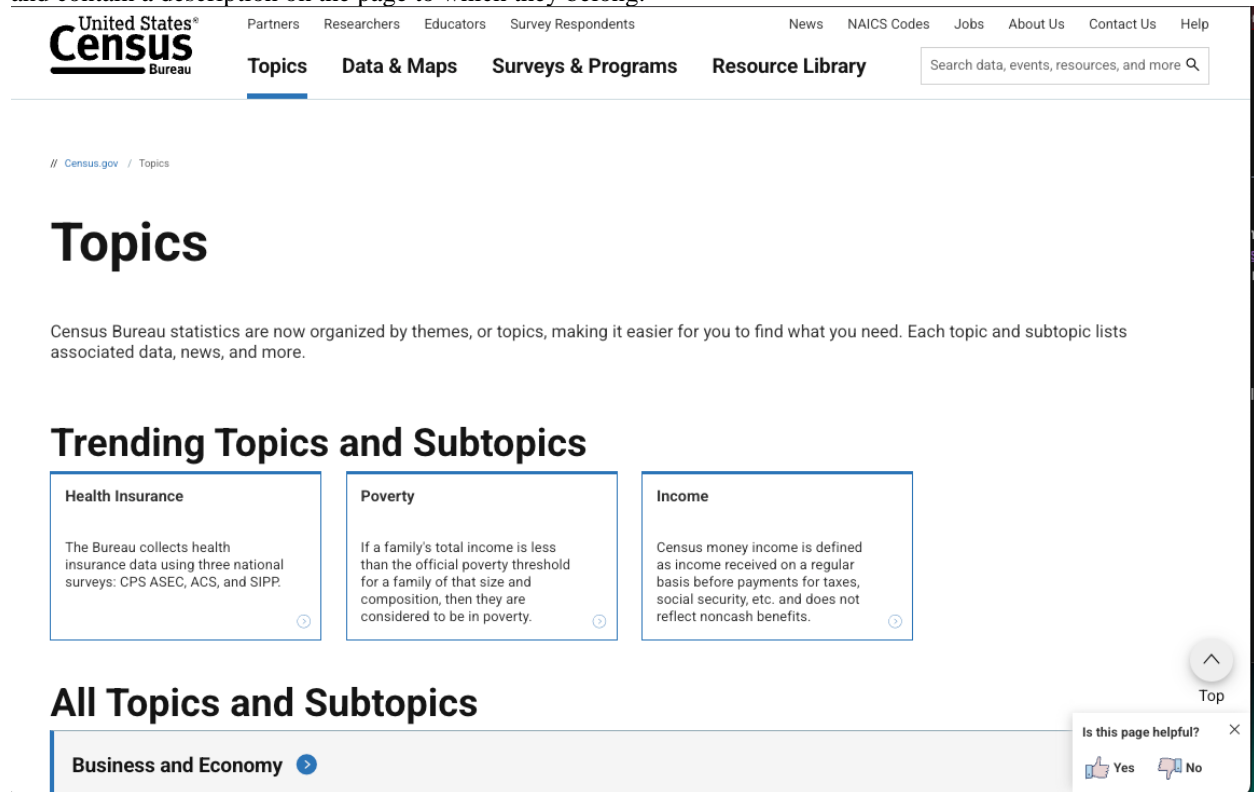


Figure 1: US Census Bureau Topics Page

The Data & Maps page (Figure 2) contains links to the new data.census.gov site, as well as Data Profiles, Tables and Maps, various Data Tools, Popular Visualizations, the Census Academy, a list of upcoming Webinars on using Census data and tools, exploring Census APIs, and additional topics including combining data, software, and experimental data products. This page also contains a link to the Data Tables, but points users to data.census.gov first (“While many data tables are now available in [data.census.gov](#), you can [browse and download additional data tables](#) by topic and year.” [US Census Bureau, 2025a]).

Data

Try out our new way to explore data

GO TO DATA.CENSUS.GOV

data.census.gov Resources

Get an overview of resources available for help on using data.census.gov, the Census API, and Microdata Access.

Data Profiles

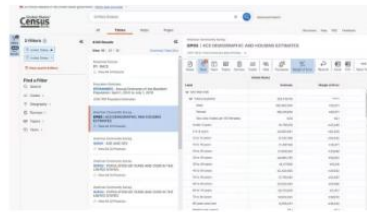
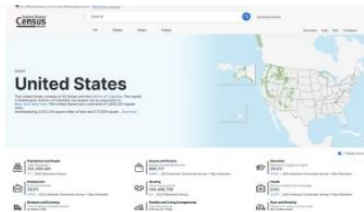
View statistics across a variety of topics for your state, county or town.

Tables and Maps

Dynamically add geographies, topics, or filters to get the data tables and maps you need.

2020 Census Data

Sign up to receive updates as more 2020 Census results become available.



Data from the 2020 Census will impact communities across the country for the next 10 years.



SUBSCRIBE

<https://data.census.gov>

Figure 2: US Census Data & Maps Web Page Screen Shot

The Surveys & Programs page (Figure 3) contains both featured content and a paginated list of all 137 census surveys and programs, both demographic and economic. They describe a survey as a “method of collecting and analyzing social, economic, and geographic data [that] provides information about the condition of the United States, states, and counties” (US Census Bureau, 2025b). The three featured surveys on the webpage the day of our survey included the 2020 Census, the American Community Survey (ACS), and the Economic Census.

Our Surveys & Programs

A survey is a method of collecting and analyzing social, economic, and geographic data. It provides information about the conditions of the United States, states, and counties.

FEATURED

Census Survey Explorer

Discover which surveys meet your criteria or needs. Search by topic, geography, and frequency of publication. This is a great place to jumpstart your research.

[Visit Census Survey Explorer](#)



Featured

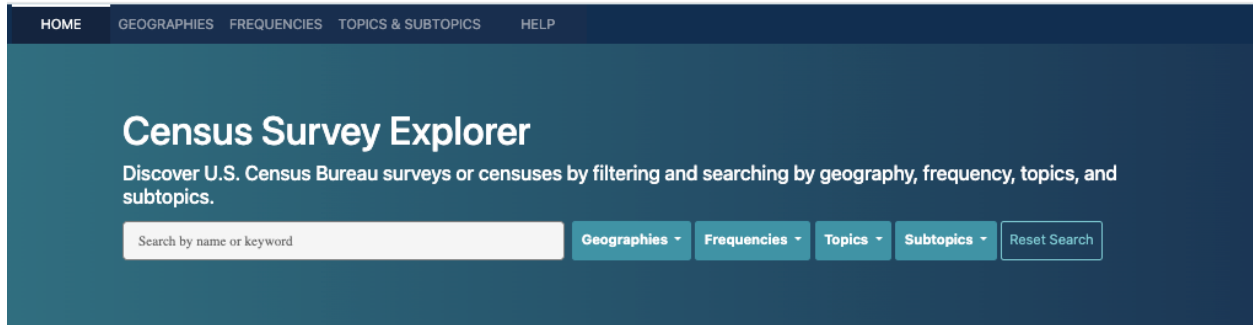
<p>2020 Census</p> <p>The 2020 Census marked the 24th count of the U.S. population and the first time that households were invited to respond to the census online.</p>	<p>American Community Survey (ACS)</p> <p>The American Community Survey is the premier source for information about America's changing population, housing and workforce.</p>	<p>Economic Census</p> <p>The Economic Census is the official five-year measure of American businesses providing comprehensive statistics at the national, state, and local levels. Follow the 2022 Economic Census release schedule for additional information.</p>
--	--	---

Top

Is this page helpful? Yes No

Figure 3: US Census Surveys & Programs Web Page Screen Shot

Featured on the Surveys & Program page was the Census Survey Explorer (Figure 4), an additional tool for users to narrow the scope of their data needs by geography (e.g. city, state), frequency of publication (annual, decennial), topics (e.g. demographic, economic), and subtopics (e.g. age, child care, citizenship). We note that “topics” in the Census Survey Explorer are not the same as the “Topics” linked from the main Census web page. The topics in the Census Survey Explorer pertain to the type of survey, and include five categories including Demographic, Economic, Government, Housing, and Social. Some of the subtopics in the Census Survey Explorer are similar to those in the Topics page, (e.g. “postsecondary education”), but this is outside the scope of the analysis of this work.



Use the geography, frequency, topic, and subtopic filters above to discover surveys relevant to your needs. ✕
Use the tabs across the top of this page to learn more and obtain guidance for using this tool.
For more details about a survey or to learn about data sources for a survey, click on the survey name to go to the survey's webpage.

Learn more about how to use this tool with this [instructional video](#).

Questions?
Call 1-844-ASK-DATA (1-844-275-3282)
Email
census.askdata@census.gov

Comment or Feedback?
Provide comments via the "Is this page helpful" box at the bottom of this page.

Advance Monthly Sales for Retail and Food Services (MARTS) →
Provides an early indication of sales for retail and food service companies.

Advanced Economic Indicators →
Provide advance statistics for U.S. International Trade in Goods, domestic

American Community Survey (ACS) →
The American Community Survey (ACS) is the nation's most current, reliable, and

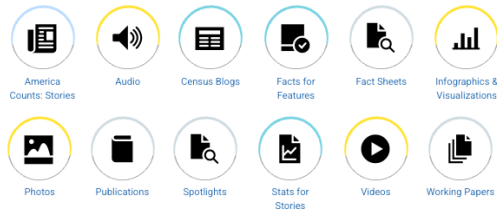
Figure 4: US Census Survey Explorer Web Page Screen Shot

The Resource Library (titled “Library”) (Figure 5) contains links to various Census Bureau collections. Links to materials are organized by media type in alphabetical order. The collection links include America Counts: Stories; Audio; Census Blogs; Facts for Features; Fact Sheets; Infographics and Visualizations; Photos; Publications; Spotlights; Stats for Stories; Videos; and Working Papers.

// [Census.gov](#) / Library

Library

Census Bureau collection formats include audio, video, mobile apps, images, and publications.



Featured

[America Counts Story](#)

Puerto Rico Community Survey Offers Detailed Look at Island Population

[Director's Blog](#)

Remembering John H. Thompson

Figure 5: US Census Reference Library Web Page Screen Shot

Accessing the Census Bureau Data API

The developer documentation for the API is separated from much of the rest of the site and varies in its coverage of topics. Finding the API, or its trainings, is tricky. One way to access it is through the Data & Maps page (Figure 2), following the Census Academy link toward the bottom of the page. This leads to a page containing fifteen course options, one of which is the “Introduction to the Census Bureau Data API,” categorized as “Business and Economy.” The entirety of the course is available as YouTube videos, with transcripts available as PDFs (US Census Bureau, 2024 July). Examining the page closely, there are links to applications that use the Census API, available APIs, where to request an API key, the API user guide, where to download RStudio, information on the tidycensus R module, and a link to a Slack channel.

You searched for: census api

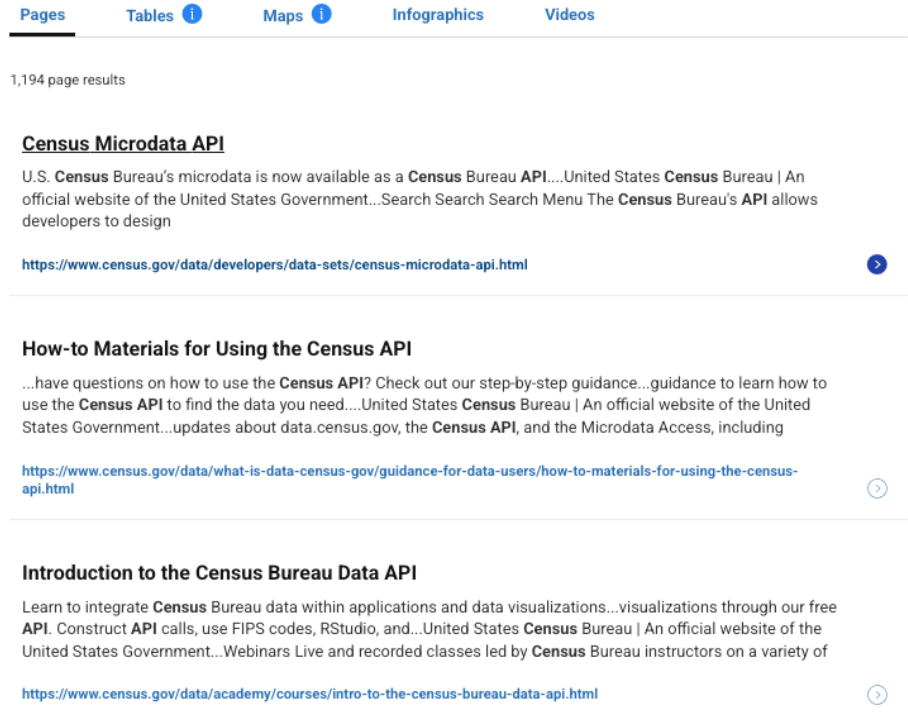


Figure 6: US Census Website "Census API" First Three Search Results

A second way to access information on the API is to search for it on the Census website. Our first search was for "census api," which yielded 1,194 results. The first three results reference a "Microdata API," the "How-to Materials for Using the Census API," and the "Introduction to the Census Bureau Data API" (Figure 6). The third result in our search was discussed as our first way to access the API. Searching for other strings, such as "api data documentation," does not perform any better.

A third way involves starting on the Data & Maps page and scrolling down to the "Explore APIs" header, then clicking on "API Documentation" card. This page contains links to the developer documentation and guides. From here we were able to find the "[Available APIs](#)" page, and examine the descriptions of available Census Surveys in the API.

The fourth method involved searching from an external search engine. This yielded the "[Census API: Datasets in /data and its descendants](#)" page, which is a simple HTML master list of all datasets, their names, and metadata needed to use the datasets with the REST API. This table contains links to the Census shorthand codes used by the API, which makes it possible to map the codes back to human-readable names for places, concepts, and variables.

Copilot Analysis

We used Bing Copilot to analyze the American Community Survey 1-Year Data for 2023, our selected "sub-domain" for testing. Following a RAG-style prompt, we instructed it to use only the documentation from the [ACS 1-Year Data \(2005-2023\)](#) and [Census API: Datasets in /data and its descendants](#) pages as its initial references for constructing the query. Results included descriptions of "social characteristics" used for the ACS and their definitions (Table 1). Survey-to-Table summary and output, the main types of variables and their relationships (Table 3).

Table 1

User-verified Sample Social Characteristics Included in the ACS

LLM Output

1. **Health Insurance Coverage:** Information on the type of health insurance coverage individuals have, including private and public insurance.
 2. **School Enrollment and Educational Attainment:** Data on the current school enrollment status of individuals and the highest level of education they have completed.
 3. **Residence 1 Year Ago:** Information on where individuals lived one year prior to the survey, which helps track migration patterns.
 4. **Household and Family Type:** Data on the composition of households and families, including the presence of children and the relationship between household members[1][2][3].
-

Generating Survey-to-Table summaries took additional prompting to output full results. Table 2 contains a sample of the output.

Table 2

User-verified Sample CoPilot Survey-to-Data Table Summary and Mapping Output

LLM Output

The American Community Survey (ACS) for 2023 organizes its data into several types of tables, each serving a specific purpose and often interrelated. Here are the main types of tables and their relationships:

1. **Data Profiles (DP):** These tables provide broad summaries of social, economic, housing, and demographic data for a specific geographic area. They are often used as a starting point for understanding the general characteristics of an area[1].
 2. **Detailed Tables (DT):** These tables offer more granular data on specific topics covered in the Data Profiles. They provide detailed cross-tabulations and are useful for in-depth analysis[2].
 3. **Subject Tables (S):** These tables focus on specific subjects, such as education, employment, or housing. They provide detailed data similar to Detailed Tables but are organized by subject area[2].
-

Further mapping of Tables-to-Related Tables, as a lookup table would be documented in a human-readable fashion (Table 3), was successful for a descriptive output but will require more refinement to be machine-readable. The table is illustrative of inter-Table relationships in the ACS for 2023.

Table 3

Mapping of Tables-to-Related Tables as a Lookup

Table Type	Description	Related Tables
Data Profiles (DP)	Broad summaries of social, economic, housing, and demographic data.	Detailed Tables (DT), Subject Tables (S)
Detailed Tables (DT)	Granular data on specific topics with detailed cross-tabulations.	Data Profiles (DP), Subject Tables (S)
Subject Tables (S)	Focused on specific subjects like education, employment, or housing.	Data Profiles (DP), Detailed Tables (DT)
Comparison Profiles (CP)	Compare data across different time periods or geographic areas.	Data Profiles (DP), Detailed Tables (DT), Subject Tables (S)
Geographic Comparison (GCT)	Compare data across different geographic areas.	Data Profiles (DP), Detailed Tables (DT), Subject Tables (S)
Public Use Microdata Sample (PUMS)	Anonymized individual-level data for custom tabulations.	Data Profiles (DP), Detailed Tables (DT), Subject Tables (S)
Variance Replicate Estimate (VRE) Tables	Measures of sampling variability for understanding precision of estimates.	Data Profiles (DP), Detailed Tables (DT), Subject Tables (S), PUMS

Outputting mappings for variables was possible only after guiding the AI through the domain. Table 4 contains high-level mappings of Data Profile (DP) tables and variables (“characteristics”) by the characteristic ID.

Table 4

High-Level Mappings of Data Profile (DP) Tables and Variables by the Characteristic ID

Data Profiles (DP) Variables	Related Detailed Tables (DT)	Related Subject Tables (S)	Related Comparison Profiles (CP)	Related Geographic Comparison Tables (GCT)
-------------------------------------	-------------------------------------	-----------------------------------	---	---

Social Characteristics	B01001, B02001, B03002, B05001	S0101, S0201, S0501	CP02	GCT0101, GCT0201
Economic Characteristics	B19001, B23001, B24011	S1901, S2301, S2401	CP03	GCT1901, GCT2301
Housing Characteristics	B25001, B25002, B25003	S2501, S2502, S2503	CP04	GCT2501, GCT2502
Demographic Characteristics	B01003, B02001, B03002	S0101, S0201, S0301	CP01	GCT0101, GCT0201

Without prompting, the AI included some example mapping of variables including Social Characteristics and Economic Characteristics to DP Variables, Detailed Tables, Subject Tables, Comparison Profiles, and Geographic Comparison Tables plus samples of Census-designated variable IDs and descriptions for each (see Table 5):

Table 5
Example Mapping Table

Characteristic	LLM Mapping
Social Characteristics	<ul style="list-style-type: none"> DP Variables: Educational attainment, language spoken at home, disability status, etc. Detailed Tables (DT): B01001 (Sex by Age), B02001 (Race), B03002 (Hispanic or Latino Origin), B05001 (Nativity and Citizenship Status). Subject Tables (S): S0101 (Age and Sex), S0201 (Selected Population Profile), S0501 (Selected Characteristics of the Native and Foreign-Born Populations). Comparison Profiles (CP): CP02 (Selected Social Characteristics). Geographic Comparison Tables (GCT): GCT0101 (Age and Sex), GCT0201 (Race and Hispanic Origin).
Economic Characteristics	<ul style="list-style-type: none"> DP Variables: Income, employment status, occupation, commuting to work, etc. Detailed Tables (DT): B19001 (Household Income), B23001 (Employment Status), B24011 (Occupation). Subject Tables (S): S1901 (Income), S2301 (Employment Status), S2401 (Occupation). Comparison Profiles (CP): CP03 (Selected Economic Characteristics). Geographic Comparison Tables (GCT): GCT1901 (Income), GCT2301 (Employment Status).

We tested the system’s ability to output a REST API query for the 2023 ACS after the initial domain analysis for this mapping, which it correctly did after an initial output in Python (see Table 6 for the REST API query and other generated information).

Table 6
REST API Query

System Input	System Output
Using that information, write a census API query for the number of grandparents living with grandchildren in Philadelphia in 2023.	<p>Sure! Here's how you can construct a REST API query to get the number of grandparents living with grandchildren in Philadelphia in 2023 using the US Census Bureau API:</p> <p>https://api.census.gov/data/2023/acs/acs1?get=B10001_001E&for=county:101&in=state:42&key=YOUR_CENSUS_API_KEY</p> <p>In this URL:</p> <ul style="list-style-type: none"> B10001_001E is the variable for the total number of grandparents living with grandchildren. county:101 specifies Philadelphia County. state:42 specifies Pennsylvania.

- Replace YOUR_CENSUS_API_KEY with your actual Census API key.

When you make a GET request to this URL, it will return the number of grandparents living with grandchildren in Philadelphia for the year 2023.

If you need any further assistance or have more questions, feel free to ask!

5. Discussion

Previous attempts to elicit useful Census data queries as output from LLMs produced seemingly plausible queries that simply did not work when tested with the Census API. This work was done with ChatGPT 3o and 3.5, Claude, Bing Copilot, and Google Gemini. This preliminary work was successful in using a domain-analytic inspired method with RAG to produce a valid, functional query using one LLM, specifically Bing Copilot.

While “chats” with LLMs are often specific and require the user to seek out the chat or chatbot application, most search engines have undergone some sort of LLM AI-ification, for lack of a better term. Dobreski & Hastings (2025) found that library-specific applications with general use chatbots were not feasible or cost effective. The case presented in this paper, which is library-aligned, requires enough knowledge of categories, classification, and the willingness to engage in general fact-checking to make it work. It is not certain that a chatbot is the solution here, either, as previous chatbots (e.g. Lappalainen and Narayanan 2023) have produced hallucinations even when carefully trained.

5.1 Limitations

Bing Copilot is based on ChatGPT and is always being updated based on user input as well as adjustments made to the backend of the model. Ergo, the questions we ask today may not yield the same results as tomorrow. The particular version of Bing Copilot used was the “enterprise” version licensed by the University of Missouri, which may or may not have further influenced the results.

As this was a pilot study, we only analyzed one survey, the American Community Survey, for one year, 2023. We tested one query for this paper, but did test others not discussed in the paper (languages spoken at home, other family configurations, etc.).

A significant limitation is the design of the Census API itself. The Census API is based on the tables as they were originally created by people conducting surveys on foot with pen and paper, and not as it would make sense to design a modern API. It was not born digital, and is a compromise between legacy parchment, not even paper, and the digital realm. The design of each Census is unique, and questions asked in the Census reflect the changes in communities over time. As such, Census REST API queries vary in structure by year, by data table, by variables, and by combinations of data thereof.

5.2 Recommendations

The multitude of information on the Census website about the “API” and “APIs,” including the “microdata API” as opposed to other APIs, might contribute to why LLMs fail to produce satisfactory results mapping and generating REST queries when not given adequate guidance. This inconsistent language across several pages seems to indicate that the Census has more than one API and only one API simultaneously (see Figure 6). Ultimately, the base URL for the REST API query remains the same. We recommend that the documentation be consolidated to improve access for LLMs and people.

LLMs “learn” from user input. It is well known that private companies store user queries and the models are either adjusted, or to some extent, self-adjust, to engagement with a human actor. While our work will not change the entire system, there is an enormous benefit to be found in training publicly available models to accurately assess and report on how to use public data.

5.3 Future work

Other LLM models need to be tested. Thus far, we have worked with the beta version of ChatGPT 4o on this same query, which did not perform as well. The model had included an economic variable (families receiving SNAP benefits) in a table on demographics (family composition by head of household or marital status). When we attempted to correct 4o, it informed us we were wrong, reiterating the hallucination about SNAP benefits. Potential models include Google’s Tapas-based models, and Microsoft Tapex-based models. Both incorporate SQL to ask questions of tabular data. Another possibility is Neulab’s Omnitab, a BART-based model built on Microsoft Tapex.

Additional future work assessing the information needs of public librarians needs to be done. The Institute of Museum and Library Services (IMLS) has previously provided Public Library Surveys. We plan to survey public librarians and other potential users of open or public government data for their information needs. While we cannot ever replace the diligent and thorough work of our dedicated civil servants, we can try to make public data more accessible and usable for public librarians to better understand their communities.

6. Conclusion

Domain analysis-guided RAG is a promising method for analyzing a large amount of technical data and documentation to ask a question that requires a precise answer as output. LLM output presented correctly mapped the ACS for 2023, and generated a simple REST API query. Accurately accessing and presenting data and facts is tricky, and removing barriers to directly access data will hopefully make it more likely to be used. With the rise of misinformation online and the potential for loss of digital data, more people (including librarians, researchers, and the general public) need access to usable and accurate data in order to analyze and verify information on their own. Census data is a public good and it is vitally important to make the data obtained by the Census more findable, usable, and future-proof.

Artificial Intelligence Use disclaimer: Generative AI was used to map the census.gov data.html for the API and to generate various tables and queries found in the Results Section (which is kind of the point). The contents of this paper were written by the authors with no use of Generative AI.

References

- Chen, X., Zhao, Z., Zhang, W., Xu, P., Wu, Y., Xu, M., Gao, L., Li, Y., Shang, X., Shi, D., & He, M. (2024). EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model. *Journal of Medical Internet Research*. 26:e60063. <https://doi.org/10.2196/60063>
- Chirkova, N., Rau, D., Déjean, H., Formal, T., Clinchant, S., & Nikoulina, V. (2024). Retrieval-augmented generation in multilingual settings. *ArXiv.org*. arXiv:2407.01463v1 [cs.CL]
- Chunxi Guo, Zhiliang Tian, Jintao Tang, Shasha Li, Ting Wang. (2025). Multi-pattern retrieval-augmented framework for Text-to-SQL with Poincaré-Skeleton retrieval and meta-instruction reasoning. *Information Processing & Management*, Volume 62, Issue 3, 2025, 103978. <https://doi.org/10.1016/j.ipm.2024.103978>
- CRedit. (2025). Contributor Role Taxonomy. <https://credit.niso.org/>
- DeBellis, M., Dutta, N., Gino, J., & Balaji, A. (2025). Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation. *Applied Ontology*, 1–19. <https://doi.org/10.1177/15705838241296446>
- Dobreski, B., & Hastings, C. (2025). AI Chatbots and Subject Cataloging: A Performance Test. *School of Information Sciences -- Faculty Publications and Other Works*. https://trace.tennessee.edu/utk_infosciopubs/485
- El-Korany, A., Rafea, A., Baraka, H., & Eid, S. (2000). A structured testing methodology for knowledge-based systems. In *Database and Expert Systems Applications: 11th International Conference, DEXA 2000 London, UK, September 4–8, 2000 Proceedings 11* (pp. 427-436). Springer Berlin Heidelberg.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey (Version 5). *arXiv.org*. <https://doi.org/10.48550/ARXIV.2312.10997>
- Gupta, S., Ranjan, R., & Singh, S. N. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions (Version 1). *arXiv.org*. <https://doi.org/10.48550/ARXIV.2410.12837>
- He, Z., Bhasuran, B., Jin, Q., Tian, S., Hanna, K., Shavor, C., Arguello, L. G., Murray, P., & Lu, Z. (2024). Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *Journal of Medical Internet Research* 26:e56655. <https://doi.org/10.2196/56655>
- Hodge, G. (2015). Government knowledge organization systems: Valuing a public good. *Bulletin of the Association for Information Science and Technology*, 40(4), 23–29. <https://doi.org/10.1002/bult.2014.1720400411>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, Volume 43, Issue 2, Article No.: 42, 1 - 55. <https://doi.org/10.1145/3703155>
- International Society for Knowledge Organization. (2021). Knowledge organization literature (C. Gnoli, Ed.). *ISKO*. <https://www.isko.org/lit.html>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Lappalainen, Y., & Narayanan, N. (2023). Aisha: A Custom AI Library Chatbot Using the ChatGPT API. *Journal of Web Librarianship* 2023, 17(3) 37–58. <https://doi.org/10.1080/19322909.2023.2221477>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2005.11401>
- Library of Congress. (ND). *U.S. Census Connections: A Resource Guide, History of the U.S. Census*. Library of Congress. <https://guides.loc.gov/census-connections/census-history>
- Li, Y., Zhao, J., Li, M., Dang, Y., Yu, E., Li, J., Sun, Z., Hussein, U., Wen, J., Abdelhameed, A. M., Mai, J., Li, S., Yu, Y., Hu, X., Yang, D., Feng, J., Li, Z., He, J., Tao, W., ... Tao, C. (2024). RefAI: A GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association*, 31(9), 2030–2039. <https://doi.org/10.1093/jamia/ocae129>
- Linxen, A., Schmidt, V.-M., Klinke, H., & Beecks, C. (2024). Ontology-driven knowledge base for digital humanities: Restructuring knowledge organization at the library of the Folkwang University of the Arts. *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024, pp. 2449-2455. <https://doi.org/10.1109/BigData62323.2024.10825984>

- Liu, S., McCoy, A. B., Wright, A. (2025). Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 2025; ocaf008. <https://doi.org/10.1093/jamia/ocaf008>
- Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., & Chen, W. (2021). Generation-Augmented Retrieval for Open-Domain Question Answering. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4089–4100.
- Murugan, M., Yuan, B., Venner, E., Ballantyne, C. M., Robinson, K. M., Coons, J. C., Wang, L., Empey, P. E., & Gibbs, R. A. (2024). Empowering personalized pharmacogenomics with generative AI solutions. *Journal of the American Medical Informatics Association*, 31(6), 1356–1366. <https://doi.org/10.1093/jamia/ocae039>
- Myers, S., Miller, T. A., Gao, Y., Churpek, M. M., Mayampurath, A., Dligach, D., Afshar, M. (2025). Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, 32(2), 357–364. <https://doi.org/10.1093/jamia/ocae308>
- Netzer, A. Y. (2025). Large Language Models for Machine-Readable Citation Data: Towards an Automated Metadata Curation Pipeline for Scholarly Journals. *Code4lib Journal*, Issue 60, 2025-04-14. <https://journal.code4lib.org/articles/18368>
- Noy, N. & McGuinness, D. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory. 32.
- Reiche, I. (2023). The viability of using an open source locally hosted AI for creating metadata in digital image collections. *Code4lib Journal*, Issue 56, 2023-04-21. <https://journal.code4lib.org/articles/17186>
- Rosa, P. C., Barizon Filho, A. L., Valentim, R. T., & Tognoli, N. (2024). Datafication, Artificial Intelligence and Images: The Dominant Paradigm in the Representation of Knowledge in Images. *Knowledge Organization*, 51(2), 117–126. <https://doi.org/10.5771/0943-7444-2024-2-117>
- Shahade, A. K., & Deshmukh, P. V. (2024). Enhancing Natural Language Processing: A Comprehensive Review of Retrieval Augmented Generation. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 609–611. <https://doi.org/10.1109/ICSES63445.2024.10763224>
- Sharma, K., Kumar, P., Li, Y. (2024). OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models. (version 1) *arXiv.org*. <https://doi.org/10.48550/arXiv.2412.15235>
- Smiraglia, R. P., & Cai, X. (2017). Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. *Knowledge Organization*, 44(3), 215–233. <https://doi.org/10.5771/0943-7444-2017-3-215>
- Soos, C., & Haroutunian, L. (2024). On the Question of Authorship in Large Language Models. *Knowledge Organization*, 51(2), 83–95. <https://doi.org/10.5771/0943-7444-2024-2-83>
- Szostak, R. (2004). *Classifying Science: Phenomena, Data, Theory, Method, Practice*. Information Science and Knowledge Management. Norwell, MA: Springer.
- Toro, S. et al. (2024). Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *Journal of Biomedical Semantics* 15:19. <https://doi.org/10.1186/s13326-024-00320-3>
- US Census Bureau. (1962, February 2). Computer Tapes Now Available to Public. *Census Bulletin: Official Information for Census Employees*, XII(2), 1 and 4.
- US Census Bureau. (2024, July). *Introduction to the Census Bureau Data API*. Introduction to the Census Bureau Data API. <https://www.census.gov/data/academy/courses/intro-to-the-census-bureau-data-api.html>
- US Census Bureau. (2024, October 22). Education. <https://www.census.gov/topics/education.html>
- US Census Bureau. (2025, April 9a). *Data & Maps*. <https://www.census.gov/data.html>
- US Census Bureau. (2025, April 9b). *Our Surveys & Programs*. <https://www.census.gov/programs-surveys.html>
- Yang, C., Xu, R., Luo, L., Pan, S. (2024). Knowledge Graph and Large Language Model Co-learning via Structure-oriented Retrieval Augmented Generation. *Bulletin of the Technical Community on Data Engineering*. December 2024 Vol. 48 No. 4. p. 9-46. <http://sites.computer.org/debull/A24dec/p9.pdf>
- Zhan, Z., Zhou, S., Li, M., & Zhang, R. (2025). RAMIE: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, 32(3), 545–554. <https://doi.org/10.1093/jamia/ocaf002>

- Zhang, Q., Chen, S., Bei, Y., Yuan, Z., Zhou, H., Hong, Z., Dong, J., Chen, H., Chang, Y, Huang, X.. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models. (version 1) *arXiv.org*. <https://doi.org/10.48550/arXiv.2501.13958>
- Zheng, X., Li, Z., Chen, Q., & Zhang, Y. (2024). Beyond decomposition: Hierarchical dependency management in multi-document question answering. *Journal of the Association for Information Science and Technology*, 1–20. <https://doi.org/10.1002/asi.24971>
- Zhu, X., Guo, X., Cao, S., Li, S., & Gong, J. (2024). StructuGraphRAG: Structured Document-Informed Knowledge Graphs for Retrieval-Augmented Generation. *Proceedings of the AAAI Symposium Series*, 4(1), 242-251. <https://doi.org/10.1609/aaaiss.v4i1.31798>