Nena Beecham
DATS 6501: Data Science Capstone
College Wealth Wizard Report
12/10/2024

1 Abstract

In this paper, a novel application of multi-agent retrieval augmented generation (RAG) to the domain of higher education is presented. Named the College Wealth Wizard, this multi-agent RAG system harnesses the power of large language models (LLMs) to unify federal data sources on education and employment, direct user queries to appropriate data sources, and contribute to the overall transparency of postgraduate finances. Powered by langgraph, the College Wealth Wizard hosts three vector store collections, as well as several router agents to appropriately route user queries. Furthermore, the application builds upon advanced RAG frameworks such as corrective RAG by implementing multiple evaluator agents. Through meticulous data preprocessing and chunking, each of the three vector store collections achieved a document retrieval accuracy score of 90% or above. Additionally, a test of 80 queries related to the vector stores and College Scorecard API revealed that a relevant answer was generated nearly 90% of times. Overall, the College Wealth Wizard unveils the potential for the application of multi-agent RAG in higher education and other untapped domains.

2 Introduction

The selection of a university and field of study can be seen as a long-term investment. There are initial investment costs, such as tuition, student loans, and room and board, as well as risks, such as not finding a job after graduation, or even failing to graduate. In addition to the costs and risks, the time it takes to break even on one's investment is not always clear. Salaries for a given occupation may vary across industry and location, complicating the ability to calculate how many years of work may be required to pay off the costs of education.

Federal education and employment data portals such as The National Center for Education Statistics (NCES), The Integrated Postsecondary Education Data System (IPEDS), The Bureau of Labor Statistics (BLS), and The National Student Loan Data System (NSLDS) have made progress in providing transparency for college costs as well as post-graduate earnings. Data on occupational wages and college expenses are publicly available for students, researchers, and policymakers alike. However, the breadth of information available through these data portals can be overwhelming and time-consuming to navigate. Coupled with internet searches, where relevant results are not always returned for user queries, getting an accurate idea of how much one's education will cost and how long it will take to pay off can be a difficult task. Given the current student loan debt crisis, inaccurately predicting education costs as well as overestimating starting salaries can have long-term financial consequences.

The onset of large language models (LLMs) and generative AI has provided a partial solution to the issue of college and occupational transparency. As mentioned previously, internet searches do not always return relevant results for user searches. Taking into account advertisements and marketing, finding accurate information on college costs can take a wealth of time and numerous sources to sift through. LLMs not only enable users with the ability to find information within seconds, but they can also summarize sources and carry out other complex actions. However, this solution is not without its own shortcomings as well.

First, LLMs can lack distinctive domain knowledge. Commonly used generative models such as ChatGPT were trained on copious amounts of data from the internet and may not understand or return irrelevant results for domain-specific queries. This poses a challenge to the application of LLMs in higher education, where industry-specific codes, definitions, and concepts are used and can be misunderstood by models. In addition to lacking distinctive domain knowledge, LLMs can be prone to the issue of hallucination. Described succinctly by Lin et al. (2024) [6], hallucination refers to generated content "that deviates from its input, lacks empirical evidence, is devoid of meaningful coherence, or contradicts real-world facts." In other words, although an answer is generated, it is not reliable.

3 Problem Statement

A proper solution to the issue of calculating long-term collegiate and occupational investments not only provides relevant sources in a timely manner, but when combined with the power of LLMs and generative AI, also addresses the issue of domain knowledge and hallucination. This solution was developed in the form of the College Wealth Wizard: a multi-agent retrieval augmented generation (RAG) application powered by langgraph for education and employment. Using information retrieval from the College Scorecard API and a series of vector store collections on federal educational and employment data, the College Wealth Wizard informs LLM-generated responses to user queries with domain-specific information. There are three research objectives of the College Wealth Wizard. The first is to assess the benefits of applying corrective RAG to the domain of higher education. The second objective is to explore the performance of vector store retrieval for federal education data sources. The final objective is to facilitate the creation of a pipeline between LLMs and the College Scorecard API.

The development of the College Wealth Wizard is significant for several reasons. First, it contributes to existing research on advanced RAG, namely corrective RAG. Initially proposed by Yan et al. [7] in 2024, the novel framework significantly reduced instances of LLM hallucination through the implementation of evaluator agents for retrieved documents. As the College Wealth Wizard includes several evaluator agents, the experiments conducted within this paper can be leveraged for further research and understanding of RAG.

Another significance of the College Wealth Wizard is that it fills a research gap on the implementation of RAG in higher education. As discussed in Section 2, although there is existing research on the general application of machine learning in higher education, there is a lack of research on the specific implementation of RAG. As institutions of higher education contemplate the use of generative AI and LLMs, this paper will serve as an innovative case study and potential guideline as to how these emerging tools can be employed.

A final significance of the College Wealth Wizard is that it can be used to estimate negative financial returns on a given plan of study. In a review of over 30,000 bachelor's degree programs in the US, Cooper (2021) [17] estimated that over a quarter of the programs had negative average financial returns on investment. More concerning, negative returns on educational investments can have adverse impacts that extend beyond individual graduates. Students carrying large amounts of student debt have less disposable income, leading [16] to lower consumption and investment spending and higher possibilities of economic recession. The College Wealth Wizard minimizes the likelihood of negative returns on educational investments through easing access to data on all facets of the collegiate lifecycle. This provides students and additional stakeholders with the opportunity to make educational and occupational choices with confidence and ensure they are accurately informed about what their future can look like.

4 Related Work

4.1 Predicting Academic Success and Alumni Salary with Machine Learning

While there is a lack of research conducted for the application of RAG to the domain of higher education, general machine learning has been applied to areas such as academic success and alumni salary projections. Noting the tendency of traditional predictive models to assume linear relationships between educational variables, Cascallar, Hernandez, and Musso (2020) proposed [1] the use of multilayer perceptron artificial neural network models to "to classify levels of grade point average, academic retention, and degree completion outcomes in a sample of 655 students from a private university." Through their experiment, they found that artificial neural networks (ANN's) had high accuracy in predicting GPA, academic retention, and degree completion and could be used to stage interventions for at-risk students.

Wang et. al (2022) experimented [2] with five machine learning methods including support vector machines, naive bayes, classification and regression trees, random forest, and XGBoost to predict whether graduates could earn a high salary or not. Through their experiment, they found that variables such as academic qualifications, employment regions, and employment industries influence post-graduate salaries. Furthermore, XGBoost models proved most effective in predicting alumni salaries.

Gomez-Cravioto et al. (2022) explored [3] the use of several machine learning methods such as random forest and gradient boosting to identify variables that have a strong relationship with alumni salary. They found that several variables, including bachelor's GPA, working hours per week, and first income after graduation were both actionable and closely tied to alumni salaries.

Ramos-Pulido, Hernández-Gress and Torres-Delgado (2024) [12] took a data science approach to understand and predict alumni career satisfaction, utilizing decision tree, random forest, gradient boosting, support vector machine, and neural network models. Through the calculation of SHapley Additive exPlanations (SHAP) values, they were able to identify the strongest predictors for career satisfaction. Low income satisfaction proved to be the strongest predictor low career satisfaction, while high life satisfaction proved to be the strongest predictor of high career satisfaction.

Acknowledging a research gap in the implementation of LLMs in higher education, Idris, Feng, and Dyo (2024) [9] explored the potential of LLM application in several areas, including: the role of institutions of higher education as gatekeepers of knowledge, providers of credentials, research centers, incubators of innovation, drivers of social change, and employers; academic integrity; the future of higher education; intellectual property; and public perception. In their exploration, they described current usage of LLMs at different higher education institutions, such as an open-access platform for free educational resources powered by LLMs at the Massachusetts Institute of Technology (MIT) and the University of Texas at Austin's partnership with edX to offer microcredentials powered by LLMs. They also identified several areas of concern, such as overreliances on automated outputs, the interpretability and reproducibility of results, data privacy, and the passing of sensitive data to LLMs.

4.2 Transparency in Postsecondary Earnings and Debt

Transparency in postsecondary earnings and debt is an established area of research. Numerous proposals have been made to promote transparency, such as central repositories of federal data on financial aid and post-college earnings, as well as penalties for universities that overprice programs or fail to communicate high debt-to-earnings ratios.

Through an analysis of federal data provided through the College Scorecard, Delisle and Cohn (2022) [4] found that "nearly a quarter of students who borrow for master's degrees were enrolled in programs where debt exceeded their earnings shortly after completing the degrees." Furthermore, income-driven student loan repayment plans were primarily used for graduate degrees, which "account[ed] for about two-thirds of the debt enrolled in the program." In light of the popularity of income-driven repayment plans, the importance of debt and post-college earnings transparency becomes more evident.

Pena (2024) [5] notes that there is a lack of a central system that pools data from federal data portals ion education n a way that is "useful for the public, policymakers, and researchers". Given that "employers are demanding higher skilled and better educated workers" and there is "a mismatch between the programs students are completing and the work they are able to find," a central system such as the one proposed by Pena would help to mitigate the lack of transparency in collegiate and occupational finances.

Using data from the U.S. Federal Reserve Board's Survey of Household Economics of Decisionmaking (SHED), Shafiq and Toutkoushian (2024) [10] conducted an examination of post-graduate views on higher education returns. While results showed that only 6.7% of students regretted going to college, nearly 20% believed the lifetime financial costs a degree were larger than the financial benefits and nearly 40% regretted the college major they chose. When viewed by financial variables, 35-40% of students who took $5,000 or more in student loan debt believed the lifetime financial costs a degree were larger than the financial benefits.

4.3 Student Loan Debt Perceptions and Estimations

The need for transparency in postgraduate earnings and debt is exemplified through studies on the perceptions and estimations of student loan debt. Markle (2019) [11] surveyed a sample of 429 undergraduate students to assess their financial literacy and student loan awareness. Among students who had student loans, the survey revealed that nearly “11% of those with student loans reported receiving no advice about student loans.” Furthermore, only 20% of students who borrowed loans determined how much to borrow by estimating costs. Nearly 15% borrowed the entire amount of loans they were offered without any calculation or outside advice. When analyzing perceptions of loan debt, almost 60% of students expressed feeling “overwhelmed, anxious, or scared about the level of student debt they [had] incurred.”

Carales and Molina (2023) [13] conducted an examination of alumni perceptions of student loan repayment as well as their navigation of the repayment process. Through qualitative interviews with 17 alumni of a public university in the southwestern US, they found that participants who “had prior experiences with debt other than student loans demonstrated a knowledgeable sense of financial literacy.” Furthermore, most participants “didn’t plan on thinking about repayment until graduation,” in line with a study by Klepfer et al. (2015) [14] that highlighted how “students may not think about repayment or how much they have borrowed until they complete the required exit counseling.”

Kraha, Doran, and Marks (2024) [15] analyzed the influence of course assignments on student perceptions of debt and salary among 88 psychology students. In a presurvey, students were asked to estimate “what they expect starting salaries and debt levels to be for each of three educational levels (BA/BS, MA/MS/MSW, PhD/ PsyD/EdD),” while in a postsurvey, they reflected on several topics such as the impact of course assignments on their views of graduate school. When compared to published values, students tended to underestimate debt and overestimate salary. Furthermore, the majority of students reported obtaining information about graduate school from the internet, with 68.2% on the presurvey and 62.5% on the postsurvey.

4.4 LLM Hallucination and Lack of Contextualized Information (RAG)

Although retrieval augmented generation provides strong capabilities outside of traditional machine learning methods, it is not without its challenges. Lin et al. (2024) conducted [6] a comprehensive review of the current landscape of research for benchmarks and evaluation methods on bias and hallucination in LLMs. In their review, they note that human feedback is currently the most widespread evaluation method of hallucination. Although there is a move towards automated evaluation, performance of this sort of evaluation “fluctuates greatly” across different LLMs or different domains within the same LLM.

As mentioned in Section 1, Yan et al. (2024) proposed [7] Corrective Retrieval Augmented Generation (CRAG) to solve the issue of LLM hallucination and cases where irrelevant documents are retrieved to inform generation. In order to preserve the integrity of LLM generations, a retrieval evaluator was developed to evaluate retrieved context and if necessary, “trigger three knowledge retrieval actions discriminately.” Through their experiment, they found that the "quality of the retrieval evaluator significantly determined the performance of the entire system," highlighting the importance of evaluating retrieved documents before moving onto generation.

Xing et al. (2025) [8] proposed a new retrieval framework to "foster a direct link between queries and web sources." Through their approach of viewing LLMs and search engines as a cohesive entity, validation and retrieval chains were established to map queries to web sources prior to generation and to validate the accessibility, relevance, and reliability of each web source. Based on the validator's feedback and response, a final optimizer revalidates the retrieved web sources.

5 Solution and Methodology

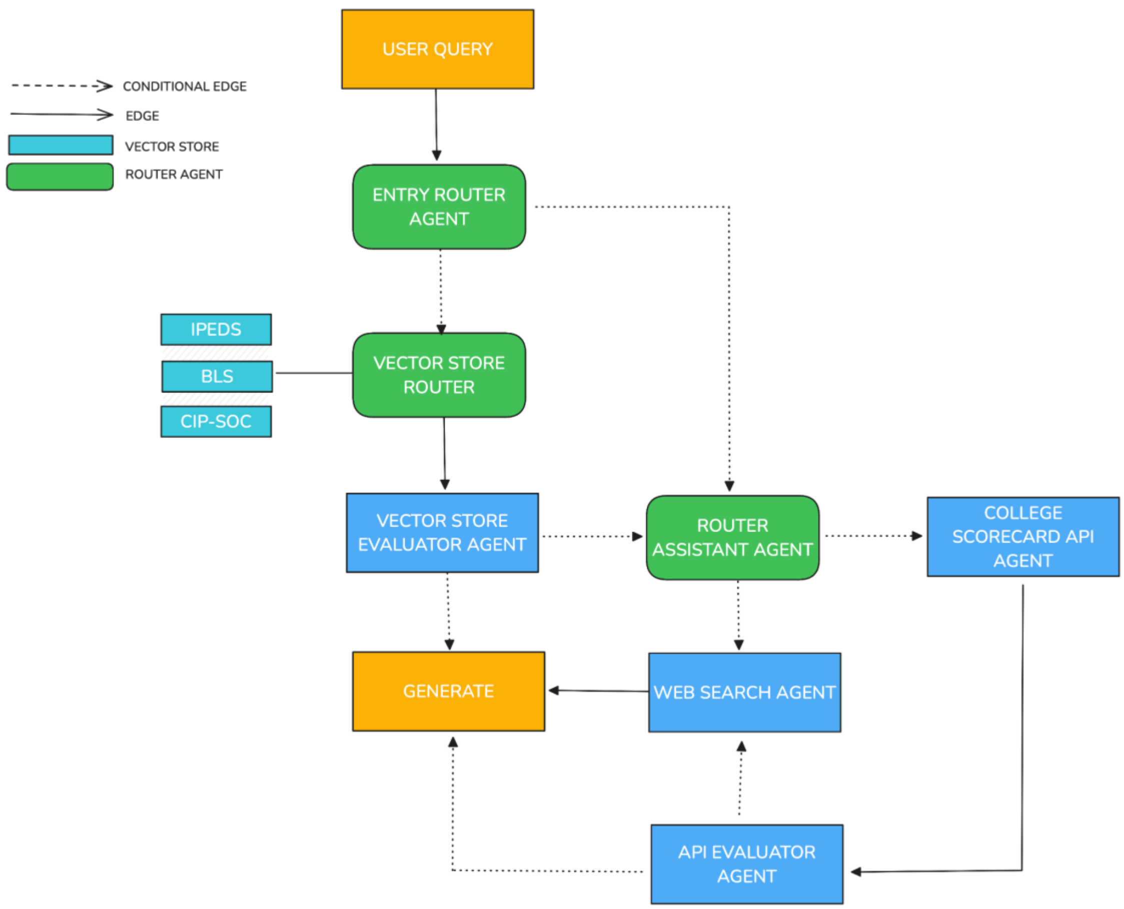

Fig. 1 Architecture of the College Wealth Wizard

The College Wealth Wizard provides a holistic approach to the issues of LLM hallucination and college and occupational transparency. Powered by a multi-agent langgraph system, this novel application accesses several vector stores, the College Scorecard API, and the Tavily web search API to inform LLM-generated responses to user queries related to education and employment. In addition to these components, the application includes router agents to direct user queries to the correct tool and evaluator agents to assess whether the context provided from a tool is relevant to a user query.

The application begins with a user query and ends with the generation agent, which considers documents retrieved from the provided agents or tools. Overall, there are 40 possible paths through the College Wealth Wizard. The prompt for the generation agent is demonstrated in Table 1.

Table 1 Generation Agent Prompt

"""You are an expert at providing direct, user-friendly answers to questions. Based ONLY ON the following context, provide a clear and concise user-friendly answer to the question. Do not include extra labels or headings before or after the answer. Here is the context: {context} Here is the question: {question}"""

5.1 Data and Vector Stores

5.1.1 IPEDs Data

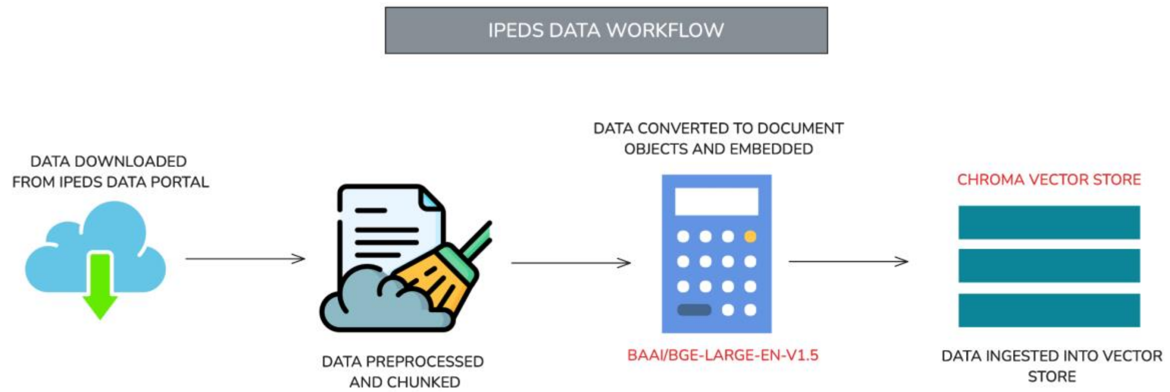

Fig. 2 IPEDs, BLS, and CIP-SOC Data Workflow

The core of the College Wealth Wizard is its vector store. Each data source used to create the collections of the vector store required significant preprocessing and experimentation with different chunking sizes. For the IPEDs data extraction, numerous columns were chosen related to tuition, tuition payment plans, room and board, application fees, books and supplies, housing, and federal aid. The initial extract consisted of 6,000 rows and 200 columns.

After the data was extracted, it had to be processed in a way that would reduce the size of the data as well as provide a logical structure for chunking. In order to do so, similar columns were combined into the following groups: tuition variables, housing variables, net price variables, aid variables, application fee variables, and expenses variables. Given that the original column names were descriptive, before being merged into the variable groups, the information within the original columns was transformed into a single sentence: f"The {column.lower()} at {df['Institution Name'][idx]} is {df[column][idx]}."

As an example, the value for the *Published out-of-state tuition and fees 2023-24 (IC2023_AY)* column for Arkansas State University is $13,826. The original column was dropped and the following sentence was added to the tuition variables column: "The *published out-of-state tuition and fees 2023-24 (ic2023_ay) at Arkansas State University is $13,826."*

The final preprocessed file consisted of three columns: a UnitID, an Institution Name, and Institution Summary. Following preprocessing, the average length of the summaries generated for each institution within the IPEDs file was 6,288 characters. The Institution Summary column represented a combination of the tuition columns, housing columns, net price columns, aid columns, application fee columns, and expenses columns:

Table 2 Institution Summary Column Generation

```
summary = []
for idx in formatted_df.index:
   summary.append(f'The tuition information for "{formatted_df["Institution Name"][idx]}" is as follows:
{formatted_df["Tuition_Information"][idx]}!! The housing information for "{formatted_df["Institution
Name"][idx]}" is as follows: {formatted_df["Housing_Information"][idx]}!! The expenses information for
"{formatted_df["Institution Name"][idx]}" is as follows: {formatted_df["Expenses_Information"][idx]}!! The aid
information for "{formatted_df["Institution Name"][idx]}" is as follows:
{formatted_df["Aid_Information"][idx]}!! The application fee information for "{formatted_df["Institution
Name"][idx]}" is as follows: {formatted_df["Application_Fee_Information"][idx]}!! The average net price
information at "{formatted_df["Institution Name"][idx]}" is as follows:
{formatted_df["Net_Price_Information"][idx]}!! {formatted_df['Options_Summary'][idx]}')

formatted_df['Institution_Summary'] = summary
```

5.1.2 BLS Data

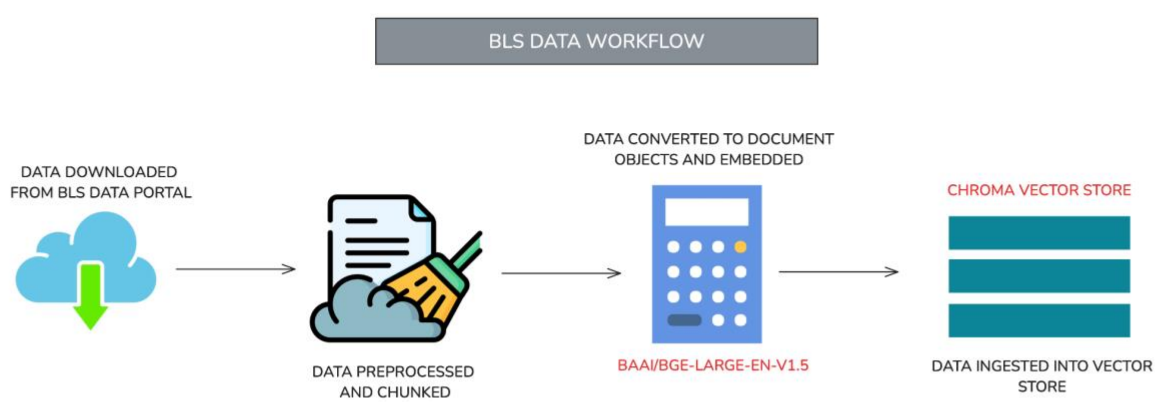

Fig. 3 BLS Data Workflow

The process of adding the BLS data to a collection within the Chroma vector store was a similar process to the IPEDs data. First, the data was downloaded from the BLS occupational wage statistics portal. The initial extract consisted of 32 columns and over 400,000 rows. Several columns were primarily null and others were not relevant to the application, so they were dropped. The final columns used were as follows: SOC code, SOC title, industry title, mean hourly wage, mean annual salary, and total employment.

Further preprocessing consisted of cleaning the CIP and SOC codes to ensure they followed the standard 6-digit format. In order to do this, the SOC and CIP codes were cleaned to remove special characters such as hyphens or periods. Finally, an Occupation Summary column was created for each occupation that included total employment and salary information. The average length for the summary text across the dataset was 440 characters.

Table 3 BLS Occupation Summary Generation

```
bls_text = []
for idx in bls_us.index:
   bls_text.append(f'For the "{bls_us["SOC_Title"][idx]}" occupation within the "{bls_us["NAICS_Industry_Title"][idx]}" industry, the national mean hourly wage is {bls_us['Mean_Hourly_Wage'][idx]} and the national mean annual wage is {bls_us['Mean_Annual_Wage'][idx]}.. The total employment nationwide is {bls_us["Total_Employment"][idx]} for the "{bls_us["SOC_Title"][idx]}" occupation in the "{bls_us["NAICS_Industry_Title"][idx]}" industry.. The SOC code for the "{bls_us["SOC_Title"][idx]}" occupation within the "{bls_us["NAICS_Industry_Title"][idx]}" industry is {bls_us["SOC_Code"][idx]}..')

bls_us['Occupation_Summary'] = bls_text
```

5.1.3 CIP-SOC Data

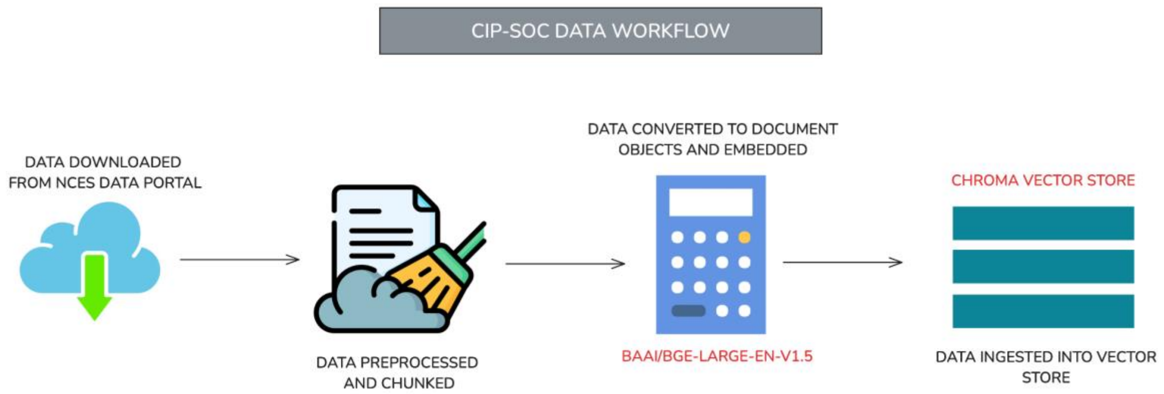

Fig. 4 CIP-SOC Data Workflow

The final data file added to the vector store was the CIP-SOC Crosswalk. The Classification of Instructional Programs (CIP) is a coding system developed by the Department of Education for fields of study, while the Standard Occupational Classification (SOC) is a federal coding system for occupations. Downloaded from the NCES, the CIP-SOC crosswalk links each CIP code to an SOC code and each SOC code to a CIP code. Some codes are linked to more than one code, or no codes at all. The initial dataset consisted of over 6,000 rows and four columns. Compared to the IPEDs and BLS datasets, this file required the least amount of preprocessing. Like the BLS data, the SOC and CIP codes were cleaned to remove special characters such as hyphens or periods. Next, a crosswalk summary was generated using all fours columns. The average length of the summaries across the entire dataset was 213 characters.

Table 5 CIP-SOC Crosswalk Summary Generation

```
xwalk = []

for idx in all_codes.index:
  xwalk.append(f'For CIP Code "{all_codes["CIP_Code"][idx]}", the associated SOC codes are as follows:
{all_codes["SOC_Code"][idx]}.. For the "{all_codes["CIP_Title"][idx]}" field of study, the associated career
paths are as follows: {all_codes["SOC_Title"][idx]}..')

all_codes['XWalk_Summary'] = xwalk
```

5.2 Router Agents

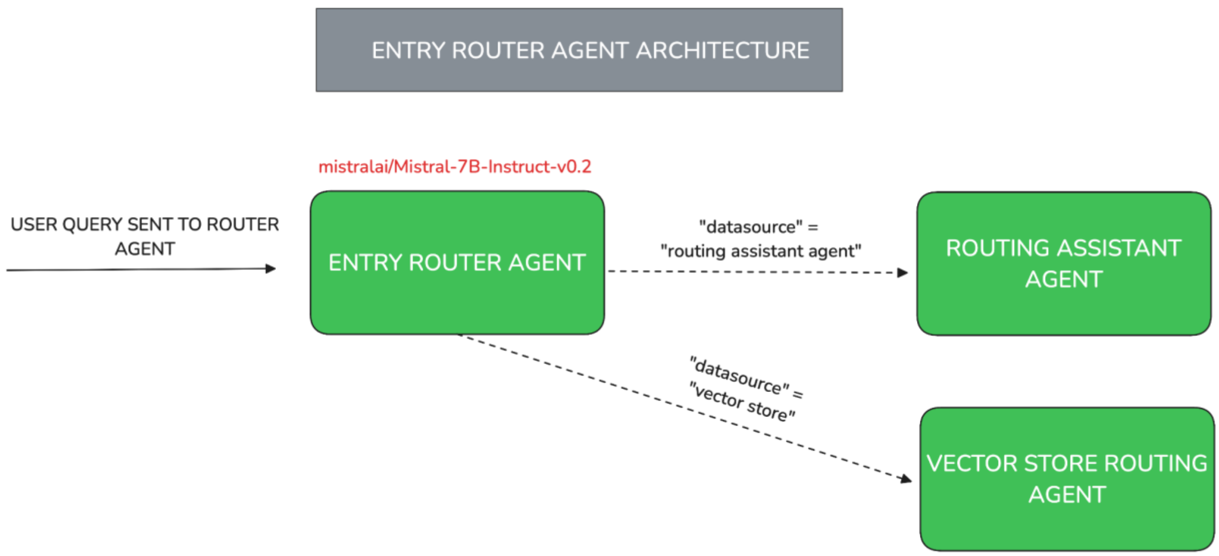

Fig. 5 Entry Router Agent Architecture

After the data was preprocessed and ingested into the Chroma vector store, the agents for the application were developed. The first point of entry to the College Wealth Wizard is the Entry Router Agent. This agent receives the user query and determines whether it can be answered using the vector stores or secondary tools available to the routing assistant agent: The College Scorecard API and the Tavily Web Search API. The agent is powered my mistralai's Mistral-7B-Instruct-v0.2. In order to route the user query, the following prompt is provided to the agent:

> *"""You are an expert at routing a user question to a series of vector stores or to a routing assistant. For questions on college tuition, financial aid and grants, collegiate expenses, mean average wages for a given occupation, salary for a given occupation, total employment for a given occupation, fields of study (CIP codes/titles) and their associated occupations (SOC codes/titles), use the vector stores. You do not need to be stringent with the keywords in the question related to these topics.*
>
> *For questions about admission rates, post-graduation salary, post-graduation earnings, student loan debt payments, student loan default rates for educational institutions, and other questions, route the query to the routing assistant.*

*Based on the question, return a dictionary with a single key 'datasource' and one of the following choices: 'vector store' or 'routing assistant'. Ensure that the output is in proper JSON format, with **double quotes** for keys and values.*

Here is how your response should be formatted: {format_instructions}\n
Here is the user query: {question}\n"""

In order to ensure the agent returns a single JSON object with a datasource key, a ResponseSchema object is used. The object is given the following description: *"whether the provided context should be sent to 'vector store', which contain information on college expenses, average salary and wages, CIP codes, and SOC codes, or if should be sent to the 'routing assistant'."*.

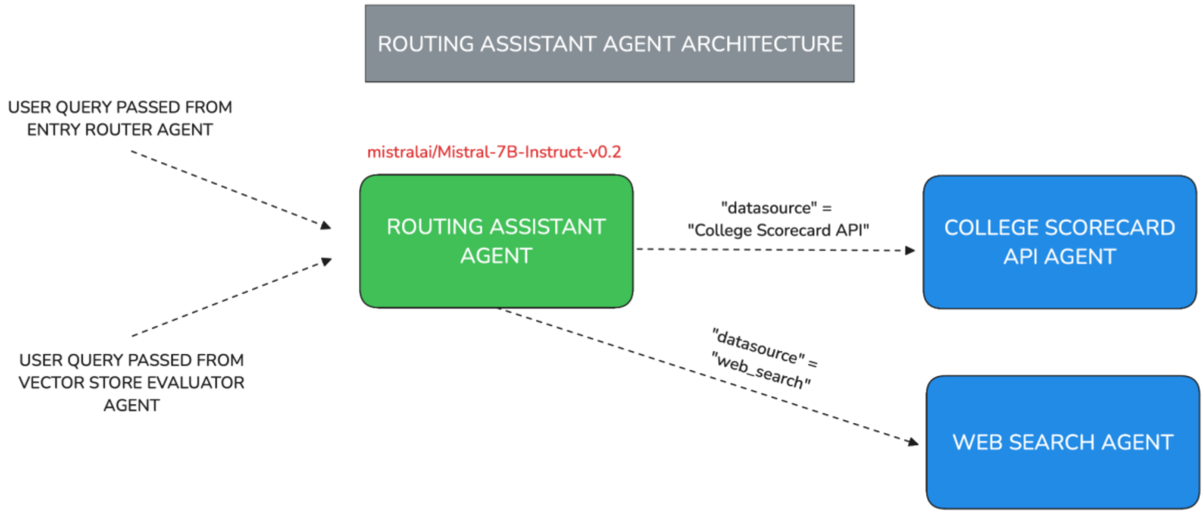

Fig. 6 Routing Assistant Agent Architecture

The Routing Assistant Agent has two points of entry: the entry router agent or the vector store evaluator agent, which is discussed in the evaluator agents section. Built on mistralai's Mistral-7B-Instruct-v0.2, this agent determines whether the user query should be sent to the College Scorecard API Agent or the Web Search Agent. Overall, there are a total of four different paths of action that can be conducted through the agent. The following prompt is provided to the agent to instruct where the user query should be routed:

"""<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are an assistant whose task is to route a user question to a web search or the College Scorecard API.

For questions about admission rates, post-graduation salary, post-graduation earnings, student loan debt payments, and student loan default rates for educational institutions, use the College Scorecard API. You do not need to be stringent with the keywords in the question related to these topics. Otherwise, use the web search.

*Based on the question, return a JSON dictionary with a single key 'datasource' and your selected choice: 'web_search' or 'College Scorecard API'. Ensure that the output is in proper JSON format, with **double quotes** for keys and values.*

Question to route: {question} <|eot_id|><|start_header_id|>assistant<|end_header_id|>"""

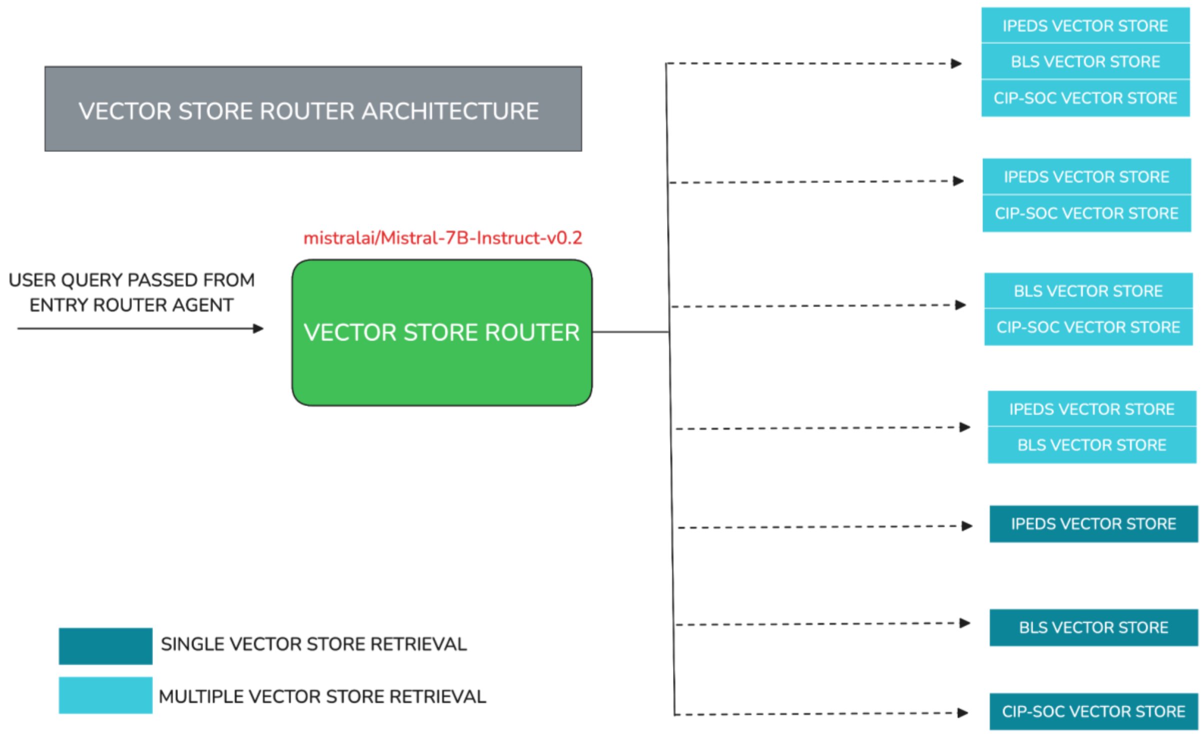

Fig. 7 Vector Store Router Architecture

The vector store router has one point of entry, which is the Entry Router Agent. After receiving the user query, it determines if it can be answered by one or more of the following vector stores: the IPEDs collection, the BLS collection, or the CIP-SOC collection. Overall, there are seven possible paths that could be conducted through the vector store router. Similar to the previous two agents, the vector store router is built on mistralai's Mistral-7B-Instruct-v0.2. The following prompt is used to instruct the router:

> *"""You are an expert at routing a user question to a series of vector stores. For questions on tuition, grants and scholarships, financial aid packages, room and board, books, offered fields of study or CIP codes for a given institution, and other collegiate expenses at colleges and universities, use the IPEDs vector store.*
>
> *For questions related to salary, mean average wages, mean average salaries, total employment, and SOC codes of specific occupations, use the BLS vector store.*
>
> *For questions about fields of study (CIP codes/titles) and their associated occupations (SOC codes/titles) WITHOUT reference to a specific educational institution, use the CIP_SOC vector store. You do not need to be stringent with the keywords in the question related to these topics.*
>
> *Based on the question, return a dictionary with a single key 'datasource' and a list as the value containing one or more of the following choices: 'IPEDs vector store', 'BLS vector store', or 'CIP_SOC vector store'. Ensure that the output is in proper JSON format, with **double quotes** for keys and values.*
>
> *Here are the formatting instructions: {format_instructions}*
> *Question to route: {question}"""*

5.3 APIs

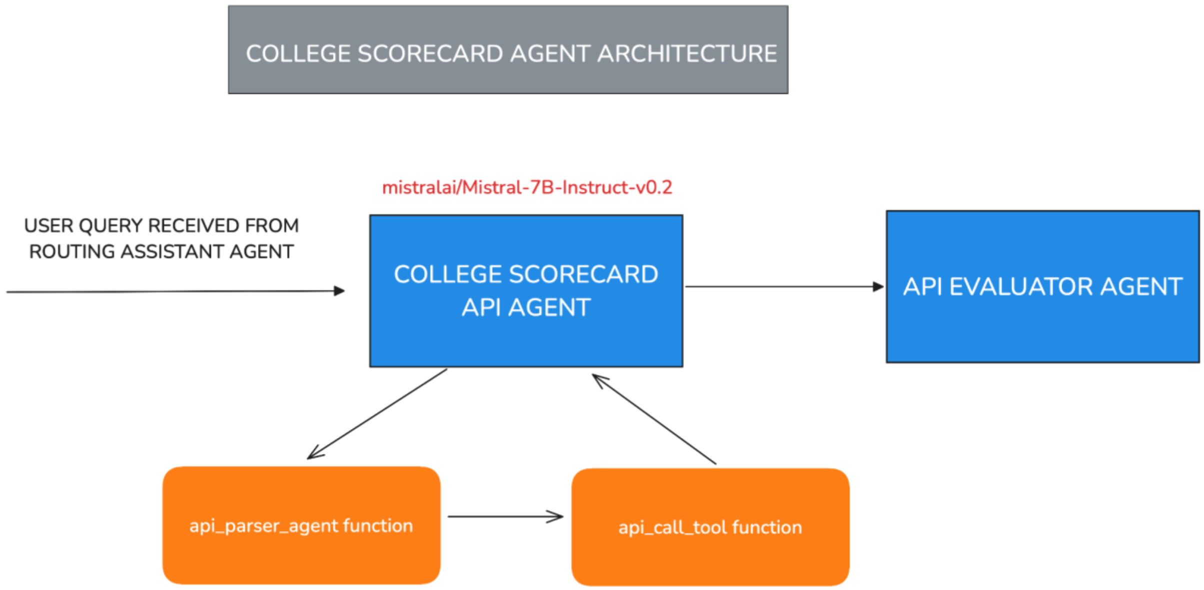

Fig. 8 College Scorecard API Agent Architecture

The College Scorecard API Agent receives the user query from the Routing Assistant Agent. After the user query is received, there are two functions that are called to receive documents relevant to the user query. The first function is the api_parser_agent() function. This function extracts the relevant filters and fields from the user query that are present in the College Scorecard API. It returns a dictionary with two keys: “filters” and “fields”. The filters key is for a dictionary that represents one of the following filters from the user query: school.name, school.city, or school.state. The fields key is for a list that contains the names of fields that should be requested from the API, such as admission_rate_suppressed.overall and earnings.4_yrs_after_completion.median.

The next function is the api_call_tool() function. This function receives the parameter dictionary returned from the api_parser_agent() function. The fields and filters provided from the parameter dictionary are formatted for the API request url and a request is made to the API. Finally, the documents are retrieved and added to the state’s “documents” key. The output is the passed to the API evaluator agent. Overall, there is one path in and one path out of the College Scorecard API Agent.

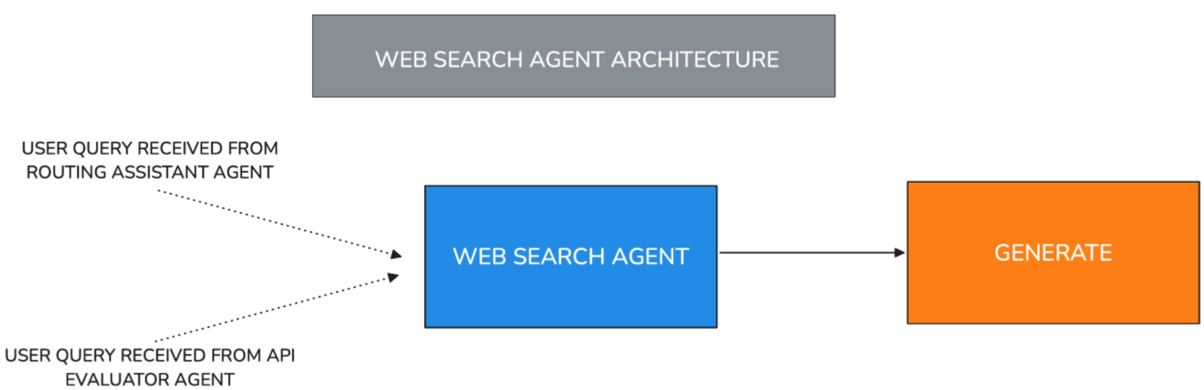

Fig. 9 Web Search Agent

The Tavily Web Search Tool is the last possible node to be called before the application generates an answer. The web search tool is meant to be a last resort if the vector stores or College Scorecard API do not provide enough context to answer the user query. There are two possible entries to the web search node: the routing assistant agent and the API evaluator agent. If the College Scorecard API fails to provide enough information for the user query, a web search is conducted. Alternatively, if a user query unrelated to collegiate education or employment is passed to the application, the query is passed to the web search, bypassing the vector stores and the College Scorecard API.

5.4 Evaluator Agents

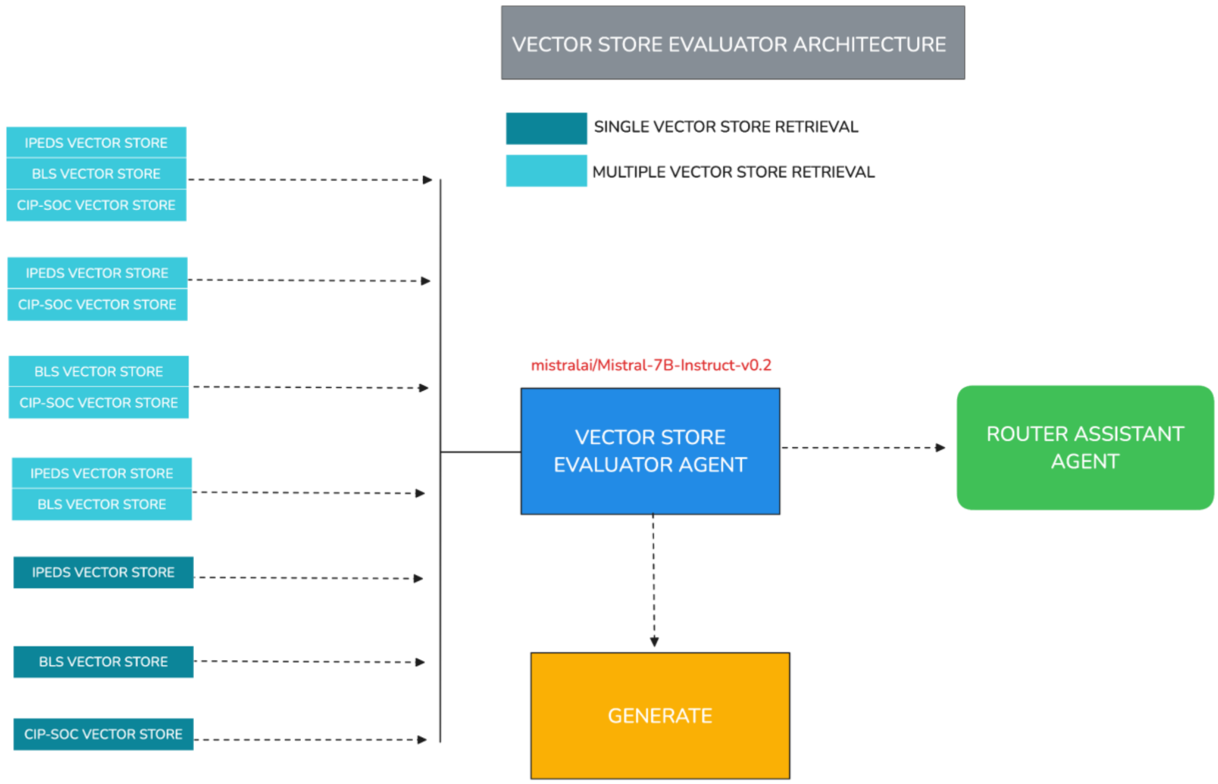

Fig. 10 Vector Store Evaluator Agent Architecture

The Vector Store Evaluator Agent receives as input the user query and the documents retrieved from the vector store(s). It assesses whether the documents are relevant to the user query and can be used to generate a response. If the documents are relevant, the user query and documents are passed to the generation agent. There, the application workflow ends. Otherwise, the user query is passed to the router assistant agent. Overall, there are 14 possible paths that can be taken through the vector store evaluator agent. The following prompt is used to instruct the router:

> *"""You are an expert at evaluating whether provided context can be used to answer a user query. If the context is relevant, you generate an answer. Otherwise, you pass the query to the routing assistant.*
>
> *Based on the user query, return a SINGLE JSON object with a key called 'relevance'. If the provided context provides sufficient information to answer the user question, choose 'generate' as the dictionary value. Otherwise, choose 'assistant_agent'.\n*
>
> *Here is how your response should be formatted: {format_instructions}\n*
> *Here is the context: {context}\n*
> *Here is the user query: {question}\n*
> *"""*

In order to constrain the agent's response, a Response Schema object with a "relevance" key and the following description is provided for formatting: *"whether or not the provided context can answer the user query. if it can, the value is generate. otherwise, it is assistant_agent."*.

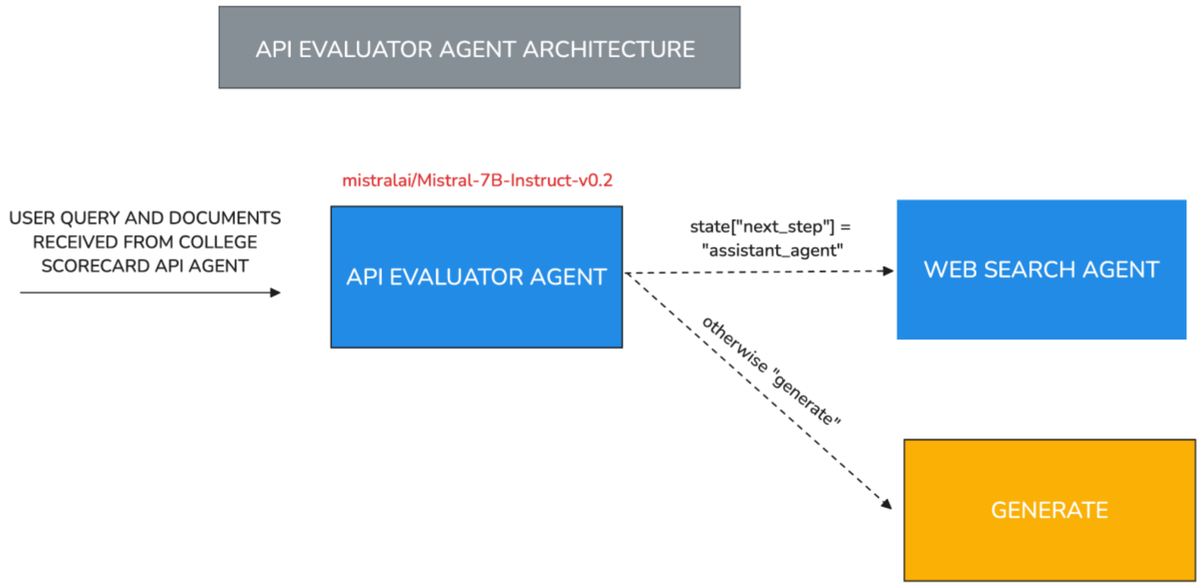

Fig. 11 API Evaluator Agent

The Vector Store Evaluator Agent is used with an additional helper function to assess the documents retrieved from the College Scorecard API. This agent is called the API Evaluator Agent and has two possible outputs: generate or web search. If the documents retrieved from the College Scorecard API are insufficient, then the user query is passed to the Web Search Agent. Otherwise, the query is passed to the generation agent and the application workflow ends. The API Evaluator Agent has the same prompt as the Vector Store Evaluator Agent, however, the following helper function is combined with it in a conditional edge to help assess documents from the College Scorecard API in specific:

Table 6 API Evaluator Helper Function

```
def api_evaluation(state):
  print(f'---EVALUATING API DOCUMENTS---')
  if state['next_step'] == 'assistant_agent':
     print(f'---NEXT STEP IS WEB SEARCH---')
     return 'tavily_search'
  else:
     print(f'---NEXT STEP IS GENERATE---')
     return 'generate'
```

6 Experiments and Results

6.1 Vector Store Retrieval Validation

The College Wealth Wizard can be thought of as a system of components. Given that the success of the application depends on the performance of its individual parts, tests were performed on each component prior to adding it to the application workflow. The first test was conducted on the vector store retrieval. This test consisted of several steps. The first step was to generate validation questions for retrieval. To do so, the document objects for the IPEDs, BLS, and CIP-SOC datasets were provided to a HuggingFace LLM. Using *Mistral-7B-Instruct-v0.2*, a hundred questions were developed for each document object using the following prompt[1]:

Table 7 Prompt for RAG Validation Dataset Creation

```
"""
Your task is to write a factoid question and an answer given a context.
Your factoid question should be answered with a specific, concise piece of factual information from the context.
Your factoid question should be formulated in the same style as questions users could ask in a search engine.
This means that your factoid question MUST NOT mention something like "according to the passage" or "context".

Provide your answer as follows:

Output:::
Factoid question: (your factoid question)
Answer: (your answer to the factoid question)

Now here is the context.

Context: {context}\n
Output:::"""
```

The document number and chunk number used to develop each validation question was saved alongside the questions to assist in developing the accuracy metrics. Finally, the validation questions were provided to the vector store to assess its performance. Mapping each validation question to a document and a document chunk allowed for the development of two accuracy metrics: document retrieval and chunk retrieval. Document retrieval accuracy evaluated whether the vector store retrieved the document used to develop a question, while chunk retrieval accuracy evaluated whether the vector store retrieved the exact chunk from a given document to develop a question. For each validation question, a total of five documents were retrieved.

Additionally, to streamline the creation of the vector store retrieval metrics, a function called *retrieval_accuracy* was developed. As input, the function takes the validation questions generated from given document objects, and the

[1] https://huggingface.co/learn/cookbook/en/rag_evaluation

documents retrieved from the vector store for the validation questions. From there, the function compares the document and chunk numbers from the retrieved documents to the document and chunk numbers associated with each validation question. Finally, the document retrieval accuracy and chunk retrieval accuracy are printed and a final dataframe with the validation questions, source documents, and retrieved documents is saved.

In consideration of the various inputs and components of the *retrieval_accuracy* function, a test script was developed to ensure appropriate performance of the function. As noted in Table 8, the function takes as input the comparison dataframe that contains the validation questions, source documents, and retrieved documents. If the document used to develop a validation question is among the documents retrieved from the vector store, the test is passed. Otherwise, the test is failed.

Table 8 Test Function for Vector Store Retrieval

```
def test_langchain_rag(test_cases):
    for idx in test_cases.index:
        if str(test_cases["source_doc_vld"][idx]) in test_cases["source_doc_rtrv"][idx]:
            print(f"Test passed for query: {test_cases['question'][idx]}")
        else:
            print(f"Test failed for query: {test_cases['question'][idx]}. Expected Document:
{test_cases['source_doc_vld'][idx]}, got: {test_cases['source_doc_rtrv'][idx]}")
```

The performance of the vector store on the hundred RAG questions provides insight into the strength of its retrieval. The correct document retrieval accuracy measures if the document used to generate a validation question was retrieved by the vector store. The correct chunk and document retrieval accuracy measures if the correct chunk from the document used to generate a validation question was retrieved.

Table 9 lists both accuracy metrics for each collection within the vector store. Apart from the BLS dataset, the accuracy for each retrieval metric across the datasets was at or above 90%. The CIP-SOC collection achieved the highest accuracy scores across all collections, with 95% for both accuracy metrics. While the BLS collection achieved a score of 90% for the document retrieval metric, it only received 70% for the chunk retrieval metric. The score of 70% signals an issue with the document's chunking. It is easy to retrieve the correct document for a given question, however, the documents may be chunked in a way that cuts off relevant information.

Table 9 Vector Store Retrieval Accuracy Metrics

Dataset	**Number of RAG Questions**	**Number of Documents Retrieved**	**Accuracy of Correct Document Retrieval**	**Accuracy of Correct Chunk and Document Retrieval**
IPEDs	100	5	90%	90%
BLS	100	5	90%	70%
CIP-SOC Associations	95	5	95%	95%

6.2 Testing Router Agents

The next component tested within the application were the router agents. The router agents hold a crucial role within the application: if a user query is not routed to the appropriate tool for additional information, the strength of the application is diminished. A test function called *test_llm_router_prompt* was developed to assess whether a user query was passed to the correct source of data or not. While the function was built off the Vector Store Router Agent, it could be used along with the Entry Router Agent or the Router Assistant Agent.

As noted in Table 10, the function prompts the user for a query and the data sources that should be used to answer the query. Next, the query is routed to the router agent—in this case the Vector Store Router Agent—and the agent returns the data source the question should be routed to. Finally, the data source specified by the router agent is compared to the intended data sources specified by the user. If all the data sources specified by the user are used by the routing agent, the test is passed. Otherwise, it is failed. As an example, if the user specifies that the BLS collection and CIP-SOC collection within the vector store should be used to answer the query, but only the BLS vector store is specified by the router agent, then the test will fail.

Table 10 Test Function for Router Agents

```
def test_llm_router_prompt(repo_id):
    query = input('Please enter your query.')
    source = input('Please enter the source(s) for your query, separated by a comma.')
    source = source.strip("'").strip('"')
    source = source.split(",")
```

```
    source = [item.lstrip() for item in source]

    print('The LLM will now route your query to the correct data source.')

    answer = vstore_router_agent(repo_id, query)

    if all(item in answer.get('datasource') for item in source):
        print(f'\nTEST PASSED: User input of source "{source}" equals LLM output of
"{answer.get("datasource")}".')
    else:
        print(f'\nTEST FAILED: User input of source "{source}" DOES NOT equal LLM output of
"{answer.get("datasource")}".')
```

To generate metrics for the router agents, 20 questions were developed for each of the vector store collections as well as the College Scorecard API. Using the router agent test function, the sources suggested by the Vector Store and Router Assistant agents were compared to the intended data source specified for each question. As shown in Table 11, the Vector Store Router Agent correctly routed the most questions to the CIP-SOC vector store. Out of 23 questions, all were correctly routed to the CIP-SOC vector store. This high accuracy metric represents two possibilities: first, the CIP-SOC collection contains highly distinctive information, making it easy to route user queries there. Second, the prompt provided to the Vector Store Router Agent provides clear examples as to what sorts of key phrases in a user query should be considered when routing to the CIP-SOC collection.

Table 11 Router Agent Accuracy Metrics

Intended Data Sources	**Number of Times Used**	**Number of Times Other Source Used**	**Overall Accuracy**
IPEDs Vector Store	13	7	65%
BLS Vector Store	15	5	75%
CIP-SOC Vector Store	20	0	100%
College Scorecard API	20	0	100%

Like the CIP-SOC collection, questions related to the College Scorecard API were all correctly routed. Of the 20 questions provided, the Router Assistant Agent correctly routed all to the API. The College Scorecard API contains data that isn't represented in any of the vector stores, such as average salary four years after graduation. This specialized data provides a possible explanation to the high routing score for the API.

Questions related to the BLS collection generated the second highest routing score for the Vector Store Agent. Out of 20 questions, 75% were correctly routed to the BLS collection. The other 15% were either incorrectly routed to the CIP-SOC collection or sent straight to the web search tool. The incorrectly routed questions signal that the agent has some trouble differentiating between whether a query is related to the BLS collection or the CIP-SOC collection. This could partly be due to the fact that the BLS and CIP-SOC collections contain overlapping information related to SOC codes and SOC titles.

Questions related to the IPEDs collection generated a routing score of 65%. A little over a quarter of the 20 IPEDs related questions were either sent to the College Scorecard API or the CIP-SOC collection. Similar to the BLS collection, the incorrectly routed questions signal that the agent has trouble deciding between the IPEDs collection or the CIP-SOC collection and College Scorecard API. This could also be due to the fact that the College Scorecard API contains tuition and debt information, although in a different format than the IPEDs collection. Furthermore, the IPEDs collection has information on CIP codes just as the CIP-SOC collection. Overall, the application was able to correctly route user queries 85% of times.

6.3 Testing Evaluator Agents

The final component tested of the College Wealth Wizard was the evaluator agents. The corrective nature of the application is housed within these agents. They serve as a safeguard against insufficient or erroneous information retrieval. If a tool does not receive the correct information to answer a user query, or the information is insufficient, the evaluator agents help to generate a more useful response by routing the query to additional data sources. To test the evaluator agents, a function called *test_llm_evaluator_agent* was developed. The function takes as input a user query, documents, and a relevance score for the documents. As noted in Table 5, if the evaluator agent's output is the same as the provided relevance argument, the test is passed. Otherwise, the test fails.

Table 12 Test Function for Evaluator Agents

```
def test_llm_evaluator_agent(question, context, relevance, repo_id):
  test_case = evaluator_agent(question, context, repo_id)

  if test_case['score'] == relevance:
     print(f'TEST PASSED: Evaluator agent correctly labeled provided documents for query as {relevance}.')
  else:
```

```
print(f'TEST FAILED: Evaluator agent incorrectly labeled provided documents for query as {test_case["score"]}.')
```

The 20 questions developed for each of the vector store collections as well as the College Scorecard API were also used to measure the performance of the application's final Generation Agent. The answers generated for each question were manually reviewed and coded. If the generated response sufficiently answered the user query, the response was coded with "1." Otherwise, the response was coded a "0".

Out of the questions related to the vector store collections, the CIP-SOC questions achieved the highest accuracy score. Only 2 of the 20 questions had an erroneous response generated, bringing the generation accuracy score to 90%. The two cases where the application generated an unrelated answer is when the queries asked which CIP codes were associated with a given career path. As an example, for the query *"What CIP codes are associated with a career in Mechanical Engineering?,"* the application returned *"Based on the provided contexts, the "Mechanical Engineering" field of study is associated with the following CIP codes: Mechanical Engineers, Mechanical Engineers, Engineering Teachers, Postsecondary."* The error is that the query asked about related CIP codes for the Mechanical Engineering career, not the Mechanical Engineering field of study. There may be CIP codes related to the given career that are not related to the field of study.

Table 13 Evaluator and Generation Agent Accuracy Metrics

Intended Data Source for Query	**Query Answered**	**Query Not Answered**	**Overall Accuracy**
IPEDs Vector Store	17	3	85%
BLS Vector Store	15	5	75%
CIP-SOC Vector Store	18	2	90%
College Scorecard API	*20	0	100%

The IPEDs questions received the second highest accuracy score of the vector store related questions. Three of the 20 questions were erroneously answered, generating an accuracy score of 85%. In the cases where an erroneous response was generated, it was because the question was sent to a data source other than the IPEDs collection. As an example, the query *"What fields of study are offered at Columbia University?"* was sent to the CIP-SOC collection and the web search tool. As mentioned in Section 5.2, this incorrect routing and generation could be due to the fact that the IPEDs and CIP-SOC collection hold overlapping sources of information.

The BLS questions received the lowest accuracy score of the vector score related questions, with a score of 75%. Of the five queries that were not properly answered, three were incorrectly routed to the CIP-SOC collection. The remaining two were correctly routed to the BLS collection, but did not retrieve the correct document. As an example, for the query *"What is the mean annual salary for a plumber in the US?,"* the application generated the following response: *"Unfortunately, the provided context does not include the mean annual salary for a plumber in the US within the "Utilities" industry. The context only provides the employment numbers and SOC codes for various occupations within the "Utilities" industry."* This response signals an issue with chunking. Total employment and SOC code information was retrieved, however, no chunks on salaries were retrieved.

For each of the 20 questions provided to the College Scorecard API, the application was able to generate a relevant response. However, there were cases where the API had null values for a given field, and the Generation Agent was not able to answer the question. As an example, for the query "*How much are the monthly debt payments for students who graduated from the University of Southern California?*," the application generated the following response: *"The context does not provide the monthly debt payments for students who graduated from the University of Southern California. All the 'aid.median_debt_suppressed.completers.monthly_payments' values are listed as None."* This signals an issue with the College Scorecard API. There are certain fields that have not yet been populated or have missing information. Overall, the application was able to correctly generate a response 87.5% of times.

7 Discussion

7.1 Retrieval and Generation Examples

Table 14 Examples of Correct Retrieval

Query	Sources	Document	Answer
6 years after starting college, how much do graduates of John Hopkins University make?	['College Scorecard API']	*Retrieved documents located in appendix.*	The median earnings of Johns Hopkins University graduates 6 years after starting college is $86,306.
What is the average net price at Tufts University?	['IPEDs vector store']	*Retrieved documents located in appendix.*	The average net price at Tufts University is $30,479.
What is the total employment of truck drivers in the US?	['BLS vector store']	*Retrieved documents located in appendix.*	The total employment of truck drivers nationwide in all industries is 917,010.0 (Heavy and Tractor-Trailer Truck Drivers: 917,010.0, from the "Truck Transportation" industry alone: 106,450.0).
What CIP codes are associated with a career in	['CIP_SOC vector store']	*Retrieved documents located in appendix.*	Aerospace Engineers are associated with the following CIP codes: "Aerospace Engineering" (CIP code: 14.0301), "Aerospace, Aeronautical, and Astronautical/Space Engineering, General" (CIP code: 14.1701), "Aerospace, Aeronautical, and

Aerospace Engineering?			Astronautical/Space Engineering, Other" (CIP code: 14.1705), and "Astronautical Engineering" (CIP code: 14.1703).

Table 14 presents examples of correct retrieval and generation. Table 15 presents examples of incorrect retrieval and/or generation. In order to improve the readability of the report, the documents are removed from the tables and provided in the Appendix. The first row of Table 15 represents a case where the College Scorecard API had null values. Furthermore, the evaluator agent did not call the web search API. Because of this, an informative response was not generated.

The second row represents a case where incorrect documents were retrieved from the IPEDs vector store. As noted in the Generation Agent answer, the retrieved documents only showed the amounts for books and supplies and other expenses, although the original query asked for tuition costs. The third row represents a case where conflicting documents were retrieved. The BLS vector store provided one number for salary, while the web sources provided other numbers. As a result, the Generation Agent shared all estimates, noting the conflicting sources.

The final row represents a case where incorrect documents were retrieved from the CIP-SOC vector store. The retrieved documents pertained to a completely irrelevant field of study. One potential reason for this is that Data Analytics might not have an associated SOC code, or Data Analytics might not be a field of study recognized in the CIP taxonomy.

Table 15 Examples of Incorrect Retrieval

Query	Sources	Document	Answer
How much student loan debt do graduates of Boston University have?	['College Scorecard API']	*Retrieved documents located in appendix.*	The context does not provide the median student loan debt for graduates of Boston University.
How much is the tuition at the University of California, Berkeley?	['IPEDs vector store']	*Retrieved documents located in appendix.*	The context does not provide information on tuition for the University of California, Berkeley. It only shows the amounts for books and supplies and other expenses.
What is the mean annual salary for a chemist in the US?	['BLS vector store', 'Web Search']	*Retrieved documents located in appendix.*	The mean annual salary for a chemist in the US is around $46,619, according to one source, but other sources suggest a higher average of $70,498 or even $81,664.
What SOC codes are associated with a field	['CIP_SOC vector store',	*Retrieved documents located in appendix.*	The provided context does not have information on SOC codes associated with a field of study in Data Analytics. The context only lists SOC codes and career

of study in Data Analytics?	'Web Search']		paths for various fields of study in Ethnic, Cultural, and Area Studies.

7.2 Challenges

7.2.1 Chunking

Table 16 Chunking Hyperparameters

Data Source	Chunk Size	Chunk Overlap	TextSplitter	Separator
IPEDs	550	100	CharacterTextSplitter	!!
BLS	250	50	CharacterTextSplitter	..
CIP-SOC	150	25	CharacterTextSplitter	..

One of the greatest challenges in developing the College Wealth Wizard was developing an appropriate chunking strategy. Since the data files used for the vector store have different formats and lengths, three different chunking strategies had to be developed. The IPEDs data file was the most difficult to chunk. Initially, the file was split along comma's. This was before the Institution Summary field was developed, so nearly 100 columns were split to develop the chunks. This posed several issues.

First, some columns contained very little information. Given that the columns were highly descriptive, not making use of the column names resulted in a huge amount of information loss. Second, some cells had columns within them, so using a comma as the separator could result in the inadvertent chunking and splitting of a cell. Finally, splitting with commas and not taking into account all columns of a row resulted in loss of important context. As an example, for a given row, one chunk could contain the institution name and while a separate chunk could contain the institution's average net price, without reference to the institution name. A better approach was to develop the Institution Summary field, which combined information from all fields for a given institution. Furthermore, the data was preprocessed in a way that ensured the institution name would be included in any given chunk.

The BLS file was the second most difficult to chunk. Following the initial method of chunking the IPEDs file, the BLS data was at first chunked along its columns, without the generation of an Occupation Summary field. This data file contained very few text and descriptive cells, as the majority of the fields were numeric. Additionally, the columns names weren't descriptive. With the generation of the Occupation Summary field, a detailed overview of each occupation, defined by an SOC code, was created. An additional issue was the presence of periods in certain cells. As an example, some industry names contained both periods and commas, leading to unintentional chunking. The use of double periods (..) for chunking helped to circumvent this issue.

The CIP-SOC file was the easiest to chunk, given it contained the least amount of information. Smaller chunk sizes worked with this field as even after generating the Xwalk Summary field, most summaries did not contain a lot of data. The decreased need for extra context meant that smaller chunks could be used without concern of a huge amount of information loss.

7.2.2 Vector Store Routing and Embedding

An additional challenge was routing user queries to the appropriate vector store. As noted in Section 6.2, information overlap between the IPEDs, BLS, and CIP-SOC files made it difficult in some cases for the routers to send user queries to the appropriate source. One approach I had to take was to redefine the router agent prompts in a way that distinctively described what sort of data should be sent to which data source. An additional approach could be to extract key words and phrases from the user query and compute the similarity score between the keys and the words associated with each data source in the LLM prompts.

The initial vector store embedding used was sentence-transformers/all-MiniLM-L6-v2. Although this embedding model allowed for quick generation of embeddings and fast retrieval time, it struggled to retrieve correct documents for the BLS and IPEDs data. Next, I tried all-mpnet-base-v2. While this embedding model is larger, its metrics were only slightly better than all-MiniLM-L6-v2. The final embedding model I tried and ended up choosing was BAAI/bge-large-en-v1.5. This model had high accuracy of retrievals for each of the vector store data files. Given the domain-specific terminology within the three data files, this model appeared to perform better with specialized data.

7.2.3 Educational Institution Names

The College Wealth Wizard had trouble retrieving documents for colleges that had similar names. As an example, queries for “Columbia University” commonly retrieved documents referring to “Columbia College.” In those cases, the embeddings generated for the documents might have been highly similar or the same when encoding “Columbia College” and “Columbia University.” This issue also occurred for university systems that had multiple branches. As an example, although The University of Texas at Austin and The University of Texas at Arlington belong to the same University of Texas system, they are two separate campuses and universities.

An additional challenge related to educational institution names is that the application struggled to retrieve the correct documents for smaller vocational schools, especially when they had similar names. As an example, institutions that contained the phrase “Beauty Academy” were commonly mistaken for one another. In the IPEDs data file, there were almost 400 institutions that contained the word “beauty” in their name and 160 with the word “cosmetology” in their name. While a solution to this issue has not been developed, one potential approach would be to individually evaluate each document retrieved with an evaluator agent to confirm that the institutions mentioned in the documents are the exact ones mentioned in the query.

8 Conclusion

The College Wealth Wizard represents a novel application of multi-agent, corrective RAG tailored to the domain of higher education. Through unifying data from federal sources such as NCES, IPEDS, and BLS, the application effectively addresses the issue of transparency in postgraduate earnings and debt—which is a pressing concern for prospective students and researchers alike. In addition to transparency, the application's implementation of evaluator agents addresses the issue of LLM hallucinations. The evaluator agents not only ensure the reliability of generated responses, but also sets a precedent for combining domain expertise with advanced generative AI techniques. The College Wealth Wizard demonstrates how RAG can be applied to the domain of higher education in a way that increases transparency and mitigates the transfer of erroneous information, which can have far-reaching financial consequences.

Future work includes experimenting with different LLMs to power the routing and evaluation agents, including region-specific salary estimates, and ingesting data on current occupational requirements. Overall, the College Wealth Wizard highlights the potential for multi-agent, corrective RAG to be applied to other domains with similar stakes and considerations.

9 References

1. Musso, M.F., Hernández, C.F.R. & Cascallar, E.C. Predicting key educational outcomes in academic trajectories: a machine-learning approach. High Educ 80, 875–894 (2020). https://doi.org/10.1007/s10734-020-00520-7
2. Wang, Ping, Liao, Wensheng, Zhao, Zhongping, Miu, Feng, Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning, Wireless Communications and Mobile Computing, 2022, 7845545, 14 pages, 2022. https://doi.org/10.1155/2022/7845545
3. Gomez-Cravioto, D.A., Diaz-Ramos, R.E., Hernandez-Gress, N. et al. Supervised machine learning predictive analytics for alumni income. J Big Data 9, 11 (2022). https://doi.org/10.1186/s40537-022-00559-6
4. Delisle, J. C. (2022). Master's Degree Debt and Earnings: New Federal Data Expose Risks for Students and the Government. Research Report. *Institute of Education Sciences*.
5. Peña C. (2020) Workforce success relies on transparent postsecondary data: A federal student-level data network would enable students to make informed education and career decisions. ED607449
6. Lin, Z., Guan, S., Zhang, W. et al. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artif Intell Rev 57, 243 (2024). https://doi.org/10.1007/s10462-024-10896-y
7. Yan, S. Q., Gu, J. C., Zhu, Y., & Ling, Z. H. (2024). Corrective retrieval augmented generation. arXiv preprint arXiv:2401.15884.
8. Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, Wei Lu, (2025). Know where to go: Make LLM a relevant, responsible, and trustworthy searchers, Decision Support Systems, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2024.114354.
9. M. Diab Idris, X. Feng and V. Dyo (2024). "Revolutionizing Higher Education: Unleashing the Potential of Large Language Models for Strategic Transformation," in IEEE Access, vol. 12, pp. 67738-67757, 2024, doi: 10.1109/ACCESS.2024.3400164.
10. Shafiq, M. N., & Toutkoushian, R. K. (2024). Perceived returns and regrets among college graduates. *Education Economics*, *32*(5), 686–702. https://doi.org/10.1080/09645292.2024.2353245
11. Markle, Gail (2019) "Crushing Debt or Savvy Strategy? Financial Literacy and Student Perceptions of their Student Loan Debt," *Journal of Student Financial Aid*: Vol. 49 : Iss. 1 , Article 4. DOI: https://doi.org/10.55504/0884-9153.1651
12. Ramos-Pulido, S., Hernández-Gress, N., & Torres-Delgado, G. (2024). Exploring the Relationship between Career Satisfaction and University Learning Using Data Science Models. *Informatics*, *11*(1), 6. https://doi.org/10.3390/informatics11010006
13. Carales, V. D., & Molina, M. (2023). Indebted to My Education: Examining College Graduates' Perceptions of Student Loan Repayment. The Journal of Student Financial Aid, 52(1). https://doi.org/10.55504/0884-9153.1773
14. Klepfer, K., Fernandez, C., Fletcher, C., & Jeff, W. (2015). Informed or overwhelmed? A legislative history of student loan counseling with a literature review on the efficacy of loan counseling. Round Rock, TX: Trellis Research. https://www.trelliscompany.org/wp-content/uploads/2017/02/Informed-or-Overwhelmed.pdf

15. Kraha, A., Doran, J. M., & Reid Marks, L. (2024). Undergraduate student impressions of student debt and starting salaries in psychology. In Scholarship of Teaching and Learning in Psychology (Vol. 10, Issue 1, pp. 76–88). American Psychological Association (APA). https://doi.org/10.1037/stl0000270
16. Apergis, Nicholas. (2023). College student loan debt and income inequality in the US: National and regional evidence. *Journal of Economic Studies* 50, (6): 1210-1225, https://www.proquest.com/scholarly-journals/college-student-loan-debt-income-inequality-us/docview/2850553940/se-2
17. Cooper, Preston. (2021). "Is College Worth It? A Comprehensive Return on Investment Analysis." FREOPP.org: The Foundation for Research on Equal Opportunity. https://freopp.org/is-college-worth-it-a-comprehensive-return-on-investment-analysis-1b2ad17f84c8

10 Appendix

10.1 Table 14 with Documents

Query	Sources	Document	Answer
6 years after starting college, how much do graduates of John Hopkins University make?	['College Scorecard API']	[{'metadata': {'page': 0, 'total': 1, 'per_page': 20}, 'results': [{'latest.earnings.6_yrs_after_entry.median': 86306, 'school.name': 'Johns Hopkins University'}]}, {'metadata': {'page': 0, 'total': 1, 'per_page': 20}, 'results': [{'latest.earnings.6_yrs_after_entry.median': 86306, 'school.name': 'Johns Hopkins University'}]}, {'metadata': {'page': 0, 'total': 1, 'per_page': 20}, 'results': [{'latest.earnings.6_yrs_after_entry.median': 86306, 'school.name': 'Johns Hopkins University'}]}, {'metadata': {'page': 0, 'total': 1, 'per_page': 20}, 'results': [{'latest.earnings.6_yrs_after_entry.median': 86306, 'school.name': 'Johns Hopkins University'}]}]	The median earnings of Johns Hopkins University graduates 6 years after starting college is $86,306.
What is the average net price at Tufts University?	['IPEDs vector store']	[Document(metadata={'chunk': 5, 'row': 5205, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Tufts University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Tufts University is 30479.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 9223.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 6812.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 14765.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 26144.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 57478.0.'), Document(metadata={'chunk': 5, 'row': 317, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Ashford University" is as follows: The average net price-students	The average net price at Tufts University is $30,479.

		awarded grant or scholarship aid 2021-22 (sfa2122) at Ashford University is 25104.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 24632.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 24732.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 25557.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 25989.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 26824.0.'), Document(metadata={'chunk': 5, 'row': 2739, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Lesley University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Lesley University is 32414.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 26082.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 24408.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 31675.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 31915.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 34667.0.'), Document(metadata={'chunk': 5, 'row': 3114, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Mildred Elley-Pittsfield Campus" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 25834.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 26016.0. The average net price	

		(income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 21881.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 24248.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 29378.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is unavailable.'), Document(metadata={'chunk': 5, 'row': 1045, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Chatham University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Chatham University is 28322.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 18793.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 20536.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 26345.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 28605.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 34515.0.'), Document(metadata={'chunk': 5, 'row': 5205, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Tufts University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Tufts University is 30479.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 9223.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 6812.0. The average net price (income 48 001-75 000)-students	

		awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 14765.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 26144.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Tufts University is 57478.0.'), Document(metadata={'chunk': 5, 'row': 317, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Ashford University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Ashford University is 25104.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 24632.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 24732.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 25557.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 25989.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Ashford University is 26824.0.'), Document(metadata={'chunk': 5, 'row': 2739, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Lesley University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Lesley University is 32414.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 26082.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 24408.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 31675.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 31915.0. The average net price (income over 110 000)-students	

		awarded title iv federal financial aid 2021-22 (sfa2122) at Lesley University is 34667.0.'), Document(metadata={'chunk': 5, 'row': 3114, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Mildred Elley-Pittsfield Campus" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 25834.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 26016.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 21881.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 24248.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is 29378.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Mildred Elley-Pittsfield Campus is unavailable.'), Document(metadata={'chunk': 5, 'row': 1045, 'source': 'IPEDs Data Extract'}, page_content='The average net price information at "Chatham University" is as follows: The average net price-students awarded grant or scholarship aid 2021-22 (sfa2122) at Chatham University is 28322.0. The average net price (income 0-30 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 18793.0. The average net price (income 30 001-48 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 20536.0. The average net price (income 48 001-75 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 26345.0. The average net price (income 75 001-110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 28605.0. The average net price (income over 110 000)-students awarded title iv federal financial aid 2021-22 (sfa2122) at Chatham University is 34515.0.')]	

What is the total employment of truck drivers in the US?	['BLS vector store']	[Document(metadata={'chunk': 1, 'row': 23715, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 917010.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Truck Transportation" industry'), Document(metadata={'chunk': 1, 'row': 13386, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 38150.0 for the "Dispatchers, Except Police, Fire, and Ambulance" occupation in the "Truck Transportation" industry'), Document(metadata={'chunk': 1, 'row': 23595, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 27420.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Highway, Street, and Bridge Construction" industry'), Document(metadata={'chunk': 1, 'row': 23574, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 19120.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Federal, State, and Local Government, excluding State and Local Government Schools and Hospitals and the U.S. Postal Service (OEWS Designation)" industry'), Document(metadata={'chunk': 1, 'row': 27927, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 106450.0 for the "Laborers and Freight, Stock, and Material Movers, Hand" occupation in the "Truck Transportation" industry'), Document(metadata={'chunk': 1, 'row': 23715, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 917010.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Truck Transportation" industry'), Document(metadata={'chunk': 1, 'row': 13386, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 38150.0 for the "Dispatchers, Except Police, Fire, and Ambulance" occupation in the "Truck Transportation" industry'), Document(metadata={'chunk': 1, 'row': 23595, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 27420.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Highway, Street, and Bridge Construction" industry'), Document(metadata={'chunk': 1, 'row':	The total employment of truck drivers nationwide in all industries is 917,010.0 (Heavy and Tractor-Trailer Truck Drivers: 917,010.0, from the "Truck Transportation" industry alone: 106,450.0).

		23574, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 19120.0 for the "Heavy and Tractor-Trailer Truck Drivers" occupation in the "Federal, State, and Local Government, excluding State and Local Government Schools and Hospitals and the U.S. Postal Service (OEWS Designation)" industry'), Document(metadata={'chunk': 1, 'row': 27927, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 106450.0 for the "Laborers and Freight, Stock, and Material Movers, Hand" occupation in the "Truck Transportation" industry')]	
What CIP codes are associated with a career in Aerospace Engineering?	['CIP_SOC vector store']	[Document(metadata={'chunk': 1, 'row': 382, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace, Aeronautical, and Astronautical/Space Engineering, General." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Engineering Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 384, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace, Aeronautical, and Astronautical/Space Engineering, Other." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Engineering Teachers, Postsecondary, Avionics Technicians'), Document(metadata={'chunk': 1, 'row': 383, 'source': 'CIP-SOC Associations'}, page_content='For the "Astronautical Engineering." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Aerospace Engineering and Operations Technologists and Technicians, Engineering Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 475, 'source': 'CIP-SOC Associations'}, page_content='For the "Aeronautical/Aerospace Engineering Technology/Technician." field of study, the associated career paths are as follows: Mechanical Drafters, Aerospace Engineering and Operations Technologists and Technicians'), Document(metadata={'chunk': 1, 'row': 813, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace Ground	Aerospace Engineers are associated with the following CIP codes: "Aerospace Engineering" (CIP code: 14.0301), "Aerospace, Aeronautical, and Astronautical/Space Engineering, General" (CIP code: 14.1701), "Aerospace, Aeronautical, and Astronautical/Space Engineering, Other" (CIP code: 14.1705), and "Astronautical Engineering" (CIP code: 14.1703).

		Equipment Technology." field of study, the associated career paths are as follows: Aerospace Engineering and Operations Technologists and Technicians, Aircraft Launch and Recovery Specialists'), Document(metadata={'chunk': 1, 'row': 382, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace, Aeronautical, and Astronautical/Space Engineering, General." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Engineering Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 384, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace, Aeronautical, and Astronautical/Space Engineering, Other." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Engineering Teachers, Postsecondary, Avionics Technicians'), Document(metadata={'chunk': 1, 'row': 383, 'source': 'CIP-SOC Associations'}, page_content='For the "Astronautical Engineering." field of study, the associated career paths are as follows: Architectural and Engineering Managers, Aerospace Engineers, Aerospace Engineering and Operations Technologists and Technicians, Engineering Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 475, 'source': 'CIP-SOC Associations'}, page_content='For the "Aeronautical/Aerospace Engineering Technology/Technician." field of study, the associated career paths are as follows: Mechanical Drafters, Aerospace Engineering and Operations Technologists and Technicians'), Document(metadata={'chunk': 1, 'row': 813, 'source': 'CIP-SOC Associations'}, page_content='For the "Aerospace Ground Equipment Technology." field of study, the associated career paths are as follows: Aerospace Engineering and Operations Technologists and Technicians, Aircraft Launch and Recovery Specialists')]	

10.2 Table 15 with Documents

Query	Sources	Document	Answer
How much student loan debt do graduates of Boston University have?	['College Scorecard API']	[{'metadata': {'page': 0, 'total': 2, 'per_page': 20}, 'results': [{'school.name': 'Boston University', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}, {'school.name': 'University of Massachusetts-Boston', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}]}, {'metadata': {'page': 0, 'total': 2, 'per_page': 20}, 'results': [{'school.name': 'Boston University', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}, {'school.name': 'University of Massachusetts-Boston', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}]}, {'metadata': {'page': 0, 'total': 2, 'per_page': 20}, 'results': [{'school.name': 'Boston University', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}, {'school.name': 'University of Massachusetts-Boston', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}]}, {'metadata': {'page': 0, 'total': 2, 'per_page': 20}, 'results': [{'school.name': 'Boston University', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}, {'school.name': 'University of Massachusetts-Boston', 'aid.median_debt_suppressed.completers.overall': None, 'aid.median_debt_suppressed.completers.monthly_payments': None}]}]	The context does not provide the median student loan debt for graduates of Boston University.
How much is the tuition at the University of California, Berkeley?	['IPEDs vector store']	[Document(metadata={'chunk': 2, 'row': 5338, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "University of California-Berkeley" is as follows: The books and supplies 2023-24 (ic2023_ay) at University of California-Berkeley is 1273.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at University of California-Berkeley is 6580.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at University of California-Berkeley is 15898.0. The books and supplies 2022-23 (ic2022_ay) at University of California-Berkeley is 1139.0.'), Document(metadata={'chunk': 2, 'row': 557, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley School of Theology" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley School of Theology is 1000.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley School of Theology is 2340.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley School of Theology is 500.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley School of Theology is 1000.0.'), Document(metadata={'chunk': 2, 'row': 555, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley College-New York" is as follows: The books and supplies 2023-24 (ic2023_ay) at	The context does not provide information on tuition for the University of California, Berkeley. It only shows the amounts for books and supplies and other expenses.

		Berkeley College-New York is 400.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-New York is 7630.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-New York is 7630.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley College-New York is unavailable.'), Document(metadata={'chunk': 2, 'row': 556, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley College-Woodland Park" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 400.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 7630.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 7630.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley College-Woodland Park is unavailable.'), Document(metadata={'chunk': 2, 'row': 554, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley City College" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley City College is 1971.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley City College is 6736.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley City College is 6736.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley City College is 1971.0.'), Document(metadata={'chunk': 2, 'row': 5338, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "University of California-Berkeley" is as follows: The books and supplies 2023-24 (ic2023_ay) at University of California-Berkeley is 1273.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at University of California-Berkeley is 6580.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at University of California-Berkeley is 15898.0. The books and supplies 2022-23 (ic2022_ay) at University of California-Berkeley is 1139.0.'), Document(metadata={'chunk': 2, 'row': 557, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley School of Theology" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley School of Theology is 1000.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley School of Theology is 2340.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley School of Theology is 500.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley School of Theology is 1000.0.'), Document(metadata={'chunk': 2, 'row': 555, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley College-New York" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley College-New York is 400.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-New York is 7630.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-New York is 7630.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley College-New York is unavailable.'),	

		Document(metadata={'chunk': 2, 'row': 556, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley College-Woodland Park" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 400.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 7630.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley College-Woodland Park is 7630.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley College-Woodland Park is unavailable.'), Document(metadata={'chunk': 2, 'row': 554, 'source': 'IPEDs Data Extract'}, page_content='The expenses information for "Berkeley City College" is as follows: The books and supplies 2023-24 (ic2023_ay) at Berkeley City College is 1971.0. The off campus (not with family) other expenses 2023-24 (ic2023_ay) at Berkeley City College is 6736.0. The off campus (with family) other expenses 2023-24 (ic2023_ay) at Berkeley City College is 6736.0. The books and supplies 2022-23 (ic2022_ay) at Berkeley City College is 1971.0.')]	
What is the mean annual salary for a chemist in the US?	['BLS vector store', 'Web Search']	[Document(metadata={'chunk': 1, 'row': 5848, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 31850.0 for the "Chemists" occupation in the "Manufacturing" industry.. The SOC code for the "Chemists" occupation within the "Manufacturing" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5808, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 2430.0 for the "Chemistry Teachers, Postsecondary" occupation in the "Junior Colleges" industry.. The SOC code for the "Chemistry Teachers, Postsecondary" occupation within the "Junior Colleges" industry is 251052'), Document(metadata={'chunk': 1, 'row': 5889, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 3260.0 for the "Chemists" occupation in the "Wholesale Trade" industry.. The SOC code for the "Chemists" occupation within the "Wholesale Trade" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5825, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 83530.0 for the "Chemists" occupation in the "Cross-industry" industry.. The SOC code for the "Chemists" occupation within the "Cross-industry" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5846, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 4660.0 for the "Chemists" occupation in the "Management of Companies and Enterprises" industry.. The SOC code for the "Chemists" occupation within the "Management of Companies and Enterprises" industry is 192031'), Document(metadata={}, page_content="The average salary for a chemist in the United States is around $46,619 per year. Avg SalaryShow avg average hourly wage. $29.2k Bottom 20%. $46.6k Median. $74.6k Top 20%. Chemists earn an average yearly salary of $46,619. Wages typically start from $29,151 and go up to $74,555.\nThe average salary for a Chemist is $81,664 per year in United States. Learn about salaries, benefits,	The mean annual salary for a chemist in the US is around $46,619, according to one source, but other sources suggest a higher average of $70,498 or even $81,664.

		salary satisfaction and where you could earn the most.\nSalary: Chemist in United States 2024 | Glassdoor The estimated total pay for a Chemist is $92,879 per year, with an average salary of $70,498 per year. Below is the total pay for the top 10 highest paying companies for a Chemist in United States. The top 5 paying industries for a Chemist in United States are Energy, Mining & Utilities with a median total pay of $93,922, Government & Public Administration with a median total pay of $82,745, Aerospace & Defense with a median total pay of $81,297, Information Technology with a median total pay of $80,661, and Construction, Repair & Maintenance Services with a median total pay of $77,991. Top paying companies in Energy, Mining & Utilities for Chemist are Chevron, Phillips 66, and ExxonMobil.\nHow much does a Chemist I make? As of November 01, 2024, the average annual pay of Chemist I in the United States is $67,915.While Salary.com is seeing that Chemist I salary in the US can go up to $84,621 or down to $53,492, but most earn between $60,365 and $76,659.Salary.com shows the average base salary (core compensation), as well as the average total cash compensation for the job of\nWe've identified 30 states where the typical salary for a Chemist job is above the national average. Topping the list is Washington, with Washington and District of Columbia close behind in second and third. District of Columbia beats the national average by 13.0%, and Washington furthers that trend with another $8,548 (13.3%) above the $64,466."), [Document(metadata={'chunk': 1, 'row': 5848, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 31850.0 for the "Chemists" occupation in the "Manufacturing" industry.. The SOC code for the "Chemists" occupation within the "Manufacturing" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5808, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 2430.0 for the "Chemistry Teachers, Postsecondary" occupation in the "Junior Colleges" industry.. The SOC code for the "Chemistry Teachers, Postsecondary" occupation within the "Junior Colleges" industry is 251052'), Document(metadata={'chunk': 1, 'row': 5889, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 3260.0 for the "Chemists" occupation in the "Wholesale Trade" industry.. The SOC code for the "Chemists" occupation within the "Wholesale Trade" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5825, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 83530.0 for the "Chemists" occupation in the "Cross-industry" industry.. The SOC code for the "Chemists" occupation within the "Cross-industry" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5846, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 4660.0 for the "Chemists" occupation in the "Management of Companies and Enterprises" industry.. The SOC code for the "Chemists" occupation within the "Management of Companies and Enterprises" industry is 192031'), Document(metadata={}, page_content="The average salary for a chemist in the	

		United States is around $46,619 per year. Avg SalaryShow avg average hourly wage. $29.2k Bottom 20%. $46.6k Median. $74.6k Top 20%. Chemists earn an average yearly salary of $46,619. Wages typically start from $29,151 and go up to $74,555.\nThe average salary for a Chemist is $81,664 per year in United States. Learn about salaries, benefits, salary satisfaction and where you could earn the most.\nSalary: Chemist in United States 2024 | Glassdoor The estimated total pay for a Chemist is $92,879 per year, with an average salary of $70,498 per year. Below is the total pay for the top 10 highest paying companies for a Chemist in United States. The top 5 paying industries for a Chemist in United States are Energy, Mining & Utilities with a median total pay of $93,922, Government & Public Administration with a median total pay of $82,745, Aerospace & Defense with a median total pay of $81,297, Information Technology with a median total pay of $80,661, and Construction, Repair & Maintenance Services with a median total pay of $77,991. Top paying companies in Energy, Mining & Utilities for Chemist are Chevron, Phillips 66, and ExxonMobil.\nHow much does a Chemist I make? As of November 01, 2024, the average annual pay of Chemist I in the United States is $67,915.While Salary.com is seeing that Chemist I salary in the US can go up to $84,621 or down to $53,492, but most earn between $60,365 and $76,659.Salary.com shows the average base salary (core compensation), as well as the average total cash compensation for the job of\nWe've identified 30 states where the typical salary for a Chemist job is above the national average. Topping the list is Washington, with Washington and District of Columbia close behind in second and third. District of Columbia beats the national average by 13.0%, and Washington furthers that trend with another $8,548 (13.3%) above the $64,466.")], Document(metadata={'chunk': 1, 'row': 5848, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 31850.0 for the "Chemists" occupation in the "Manufacturing" industry.. The SOC code for the "Chemists" occupation within the "Manufacturing" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5808, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 2430.0 for the "Chemistry Teachers, Postsecondary" occupation in the "Junior Colleges" industry.. The SOC code for the "Chemistry Teachers, Postsecondary" occupation within the "Junior Colleges" industry is 251052'), Document(metadata={'chunk': 1, 'row': 5889, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 3260.0 for the "Chemists" occupation in the "Wholesale Trade" industry.. The SOC code for the "Chemists" occupation within the "Wholesale Trade" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5825, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 83530.0 for the "Chemists" occupation in the "Cross-industry" industry.. The SOC code for the "Chemists" occupation within the "Cross-industry" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5846, 'source': 'BLS Data Extract'}, page_content='The total	

		employment nationwide is 4660.0 for the "Chemists" occupation in the "Management of Companies and Enterprises" industry.. The SOC code for the "Chemists" occupation within the "Management of Companies and Enterprises" industry is 192031'), Document(metadata={}, page_content="The average salary for a chemist in the United States is around $46,619 per year. Avg SalaryShow avg average hourly wage. $29.2k Bottom 20%. $46.6k Median. $74.6k Top 20%. Chemists earn an average yearly salary of $46,619. Wages typically start from $29,151 and go up to $74,555.\nThe average salary for a Chemist is $81,664 per year in United States. Learn about salaries, benefits, salary satisfaction and where you could earn the most.\nSalary: Chemist in United States 2024 | Glassdoor The estimated total pay for a Chemist is $92,879 per year, with an average salary of $70,498 per year. Below is the total pay for the top 10 highest paying companies for a Chemist in United States. The top 5 paying industries for a Chemist in United States are Energy, Mining & Utilities with a median total pay of $93,922, Government & Public Administration with a median total pay of $82,745, Aerospace & Defense with a median total pay of $81,297, Information Technology with a median total pay of $80,661, and Construction, Repair & Maintenance Services with a median total pay of $77,991. Top paying companies in Energy, Mining & Utilities for Chemist are Chevron, Phillips 66, and ExxonMobil.\nHow much does a Chemist I make? As of November 01, 2024, the average annual pay of Chemist I in the United States is $67,915.While Salary.com is seeing that Chemist I salary in the US can go up to $84,621 or down to $53,492, but most earn between $60,365 and $76,659.Salary.com shows the average base salary (core compensation), as well as the average total cash compensation for the job of\nWe've identified 30 states where the typical salary for a Chemist job is above the national average. Topping the list is Washington, with Washington and District of Columbia close behind in second and third. District of Columbia beats the national average by 13.0%, and Washington furthers that trend with another $8,548 (13.3%) above the $64,466."), [Document(metadata={'chunk': 1, 'row': 5848, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 31850.0 for the "Chemists" occupation in the "Manufacturing" industry.. The SOC code for the "Chemists" occupation within the "Manufacturing" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5808, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 2430.0 for the "Chemistry Teachers, Postsecondary" occupation in the "Junior Colleges" industry.. The SOC code for the "Chemistry Teachers, Postsecondary" occupation within the "Junior Colleges" industry is 251052'), Document(metadata={'chunk': 1, 'row': 5889, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 3260.0 for the "Chemists" occupation in the "Wholesale Trade" industry.. The SOC code for the "Chemists" occupation within the "Wholesale Trade" industry is 192031'), Document(metadata={'chunk': 1, 'row':	

		5825, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 83530.0 for the "Chemists" occupation in the "Cross-industry" industry.. The SOC code for the "Chemists" occupation within the "Cross-industry" industry is 192031'), Document(metadata={'chunk': 1, 'row': 5846, 'source': 'BLS Data Extract'}, page_content='The total employment nationwide is 4660.0 for the "Chemists" occupation in the "Management of Companies and Enterprises" industry.. The SOC code for the "Chemists" occupation within the "Management of Companies and Enterprises" industry is 192031'), Document(metadata={}, page_content="The average salary for a chemist in the United States is around $46,619 per year. Avg SalaryShow avg average hourly wage. $29.2k Bottom 20%. $46.6k Median. $74.6k Top 20%. Chemists earn an average yearly salary of $46,619. Wages typically start from $29,151 and go up to $74,555.\nThe average salary for a Chemist is $81,664 per year in United States. Learn about salaries, benefits, salary satisfaction and where you could earn the most.\nSalary: Chemist in United States 2024 | Glassdoor The estimated total pay for a Chemist is $92,879 per year, with an average salary of $70,498 per year. Below is the total pay for the top 10 highest paying companies for a Chemist in United States. The top 5 paying industries for a Chemist in United States are Energy, Mining & Utilities with a median total pay of $93,922, Government & Public Administration with a median total pay of $82,745, Aerospace & Defense with a median total pay of $81,297, Information Technology with a median total pay of $80,661, and Construction, Repair & Maintenance Services with a median total pay of $77,991. Top paying companies in Energy, Mining & Utilities for Chemist are Chevron, Phillips 66, and ExxonMobil.\nHow much does a Chemist I make? As of November 01, 2024, the average annual pay of Chemist I in the United States is $67,915.While Salary.com is seeing that Chemist I salary in the US can go up to $84,621 or down to $53,492, but most earn between $60,365 and $76,659.Salary.com shows the average base salary (core compensation), as well as the average total cash compensation for the job of\nWe've identified 30 states where the typical salary for a Chemist job is above the national average. Topping the list is Washington, with Washington and District of Columbia close behind in second and third. District of Columbia beats the national average by 13.0%, and Washington furthers that trend with another $8,548 (13.3%) above the $64,466.")]]	
What SOC codes are associated with a field of study in Data Analytics?	['CIP_SOC vector store', 'Web Search']	[Document(metadata={'chunk': 1, 'row': 172, 'source': 'CIP-SOC Associations'}, page_content='For the "Ethnic, Cultural Minority, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 173, 'source': 'CIP-SOC Associations'}, page_content='For the "Area, Ethnic, Cultural, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'),	The provided context does not have information on SOC codes associated with a field of study in Data Analytics. The context

		Document(metadata={'chunk': 1, 'row': 160, 'source': 'CIP-SOC Associations'}, page_content='For the "Area Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 135, 'source': 'CIP-SOC Associations'}, page_content='For the "South Asian Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 906, 'source': 'CIP-SOC Associations'}, page_content='For the "Catholic Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={}, page_content='Find Matching SOC Codes for data analyst, With Definition and Examples. Menu Close ... Plan, direct, or coordinate activities in such fields as electronic data processing, ... Plan and direct the operation and maintenance of catapults, arresting gear, and associated mechanical, hydraulic, and control systems involved primarily in aircraft\nO*NET-SOC 2019 Code O*NET-SOC 2019 Title; ... or coordinate activities in such fields as electronic data processing, information systems, systems analysis, and computer programming. ... Gather, compile, and analyze data. Study the use and operation of transportation systems. Develop transportation models or simulations. 19-4012.00:\nData Scientists (Soc Code 15-2051) This is a detailed information about Data Scientists, including the group, the code, the title, the definition of SOC, 2018 SOC Direct Match Title, illustrative example. ... Data Scientists. SOC Definition: Develop and implement a set of techniques or analytics applications to transform raw data into\nSOC Code 15-2051 - Data Scientists is a a final level code of the "Computer and Mathematical Occupations" Major Occupation Category. Exclusions "Statisticians" (15-2041) , "Cartographers and Photogrammetrists" (17-1021) , and "Health Information Technologists and Medical Registrars" (29-9021) .\nAn SOC code means that an occupation has been classified by the federal government for the purpose of cataloging, analyzing and distributing data about workers and jobs in the United States. Four tiers of aggregation within the codes allow organizations to find the level of detail that suits their interests.'), [Document(metadata={'chunk': 1, 'row': 172, 'source': 'CIP-SOC Associations'}, page_content='For the "Ethnic, Cultural Minority, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 173, 'source': 'CIP-SOC Associations'}, page_content='For the "Area, Ethnic, Cultural, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 160, 'source': 'CIP-SOC Associations'}, page_content='For the "Area Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers,	only lists SOC codes and career paths for various fields of study in Ethnic, Cultural, and Area Studies.

		Postsecondary'), Document(metadata={'chunk': 1, 'row': 135, 'source': 'CIP-SOC Associations'}, page_content='For the "South Asian Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 906, 'source': 'CIP-SOC Associations'}, page_content='For the "Catholic Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={}, page_content='Find Matching SOC Codes for data analyst, With Definition and Examples. Menu Close ... Plan, direct, or coordinate activities in such fields as electronic data processing, ... Plan and direct the operation and maintenance of catapults, arresting gear, and associated mechanical, hydraulic, and control systems involved primarily in aircraft\nO*NET-SOC 2019 Code O*NET-SOC 2019 Title; ... or coordinate activities in such fields as electronic data processing, information systems, systems analysis, and computer programming. ... Gather, compile, and analyze data. Study the use and operation of transportation systems. Develop transportation models or simulations. 19-4012.00:\nData Scientists (Soc Code 15-2051) This is a detailed information about Data Scientists, including the group, the code, the title, the definition of SOC, 2018 SOC Direct Match Title, illustrative example. ... Data Scientists. SOC Definition: Develop and implement a set of techniques or analytics applications to transform raw data into\nSOC Code 15-2051 - Data Scientists is a a final level code of the "Computer and Mathematical Occupations" Major Occupation Category. Exclusions "Statisticians" (15-2041) , "Cartographers and Photogrammetrists" (17-1021) , and "Health Information Technologists and Medical Registrars" (29-9021) .\nAn SOC code means that an occupation has been classified by the federal government for the purpose of cataloging, analyzing and distributing data about workers and jobs in the United States. Four tiers of aggregation within the codes allow organizations to find the level of detail that suits their interests.')], Document(metadata={'chunk': 1, 'row': 172, 'source': 'CIP-SOC Associations'}, page_content='For the "Ethnic, Cultural Minority, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 173, 'source': 'CIP-SOC Associations'}, page_content='For the "Area, Ethnic, Cultural, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 160, 'source': 'CIP-SOC Associations'}, page_content='For the "Area Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 135, 'source': 'CIP-SOC Associations'}, page_content='For the "South Asian Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies	

		Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 906, 'source': 'CIP-SOC Associations'}, page_content='For the "Catholic Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={}, page_content='Find Matching SOC Codes for data analyst, With Definition and Examples. Menu Close ... Plan, direct, or coordinate activities in such fields as electronic data processing, ... Plan and direct the operation and maintenance of catapults, arresting gear, and associated mechanical, hydraulic, and control systems involved primarily in aircraft\nO*NET-SOC 2019 Code O*NET-SOC 2019 Title; ... or coordinate activities in such fields as electronic data processing, information systems, systems analysis, and computer programming. ... Gather, compile, and analyze data. Study the use and operation of transportation systems. Develop transportation models or simulations. 19-4012.00:\nData Scientists (Soc Code 15-2051) This is a detailed information about Data Scientists, including the group, the code, the title, the definition of SOC, 2018 SOC Direct Match Title, illustrative example. ... Data Scientists. SOC Definition: Develop and implement a set of techniques or analytics applications to transform raw data into\nSOC Code 15-2051 - Data Scientists is a a final level code of the "Computer and Mathematical Occupations" Major Occupation Category. Exclusions "Statisticians" (15-2041) , "Cartographers and Photogrammetrists" (17-1021) , and "Health Information Technologists and Medical Registrars" (29-9021) .\nAn SOC code means that an occupation has been classified by the federal government for the purpose of cataloging, analyzing and distributing data about workers and jobs in the United States. Four tiers of aggregation within the codes allow organizations to find the level of detail that suits their interests.'), [Document(metadata={'chunk': 1, 'row': 172, 'source': 'CIP-SOC Associations'}, page_content='For the "Ethnic, Cultural Minority, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 173, 'source': 'CIP-SOC Associations'}, page_content='For the "Area, Ethnic, Cultural, Gender, and Group Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 160, 'source': 'CIP-SOC Associations'}, page_content='For the "Area Studies, Other." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 135, 'source': 'CIP-SOC Associations'}, page_content='For the "South Asian Studies." field of study, the associated career paths are as follows: Area, Ethnic, and Cultural Studies Teachers, Postsecondary'), Document(metadata={'chunk': 1, 'row': 906, 'source': 'CIP-SOC Associations'}, page_content='For the "Catholic Studies." field of study, the associated career paths are as follows: Area, Ethnic, and	

		Cultural Studies Teachers, Postsecondary'), Document(metadata={}, page_content='Find Matching SOC Codes for data analyst, With Definition and Examples. Menu Close ... Plan, direct, or coordinate activities in such fields as electronic data processing, ... Plan and direct the operation and maintenance of catapults, arresting gear, and associated mechanical, hydraulic, and control systems involved primarily in aircraft\nO*NET-SOC 2019 Code O*NET-SOC 2019 Title; ... or coordinate activities in such fields as electronic data processing, information systems, systems analysis, and computer programming. ... Gather, compile, and analyze data. Study the use and operation of transportation systems. Develop transportation models or simulations. 19-4012.00:\nData Scientists (Soc Code 15-2051) This is a detailed information about Data Scientists, including the group, the code, the title, the definition of SOC, 2018 SOC Direct Match Title, illustrative example. ... Data Scientists. SOC Definition: Develop and implement a set of techniques or analytics applications to transform raw data into\nSOC Code 15-2051 - Data Scientists is a a final level code of the "Computer and Mathematical Occupations" Major Occupation Category. Exclusions "Statisticians" (15-2041) , "Cartographers and Photogrammetrists" (17-1021) , and "Health Information Technologists and Medical Registrars" (29-9021) .\nAn SOC code means that an occupation has been classified by the federal government for the purpose of cataloging, analyzing and distributing data about workers and jobs in the United States. Four tiers of aggregation within the codes allow organizations to find the level of detail that suits their interests.')]]	