Homework 4: Large Language Models

UIC CS 418, Spring 2023

*According to the **Academic Integrity Policy** of this course, all work submitted for grading must be done individually, unless otherwise specified. While we encourage you to talk to your peers and learn from them, this interaction must be superficial with regards to all work submitted for grading. This means you cannot work in teams, you cannot work side-by-side, you cannot submit someone else's work (partial or complete) as your own. In particular, note that you are guilty of academic dishonesty if you extend or receive any kind of unauthorized assistance. Absolutely no transfer of program code between students is permitted (paper or electronic), and you may not solicit code from family, friends, or online forums. Other examples of academic dishonesty include emailing your program to another student, copying-pasting code from the internet, working in a group on a homework assignment, and allowing a tutor, TA, or another individual to write an answer for you. Academic dishonesty is unacceptable, and penalties range from failure to expulsion from the university; cases are handled via the official student conduct process described at https://dos.uic.edu/conductforstudents.shtml.*

All students are allowed to work in pairs and submit one homework assignment per pair. The assignment will use free credit from existing LLM paltforms, so if you already used all free credits, please choose a partner who has not. There will be no extra credit given to students who choose to work alone. The pairs of students who choose to work together and submit one homework assignment together still need to abide by the Academic Integrity Policy and not share or receive help from others (except each other).

Due Date

This assignment is due at 11:59pm on April 25th, 2023.

Instructions

You need to complete all code and answer all questions denoted by **Q#** in this notebook. When you are done, you should export `hw4.ipynb` with your answers as a PDF file, upload that file `hw4.pdf` to *Homework 4 - written* on Gradescope, tagging each question.

For students who work in a team of two, only one student needs to submit the homework and just tag the other student on Gradescope.

Keep an eye out for the following icons:

Specifies that you need to add something to the notebook in order to get credit.

Specifies that you need to save something to `sub` (submission) subdirectory or verify the existence of some file in `sub` to get credit.

Warnings to avoid common mistakes and pitfalls. Pay special attention to these.

Autograding

This assignment will be graded entirely based on your PDF submissions. No questions will be graded using an Autograder.

Before We Start

Before we dive into the assignment, we need to install the following python packages:

- matplotlib*
- pandas*
- numpy*
- requests
- openai
- json

Packages marked with * should be pre-installed in a conda environment. You may install missing packages one at a time using `pip` command.

Now, we are ready to import all dependencies.

```
In [ ]: import pandas as pd
        import numpy as np
        import os
        import requests
        import json
        import openai
```

Now that we have everything we need, we can dive into the homework itself.

Large Language Models

Large language models have quickly advanced from often being "laughably bad" to "quite capable" for many tasks. These models are typically:

1. Trained to predict next words in massive web-scale datasets using neural networks with billions of parameters;
2. Fine-tuned using human feedback to reduce the amount of outputs being produced that are offensive, incorrect, not useful, etc. Many different base models (#1) have been or are being developed, and each is being fine-tuned (#2) with different focuses to produce many different possible models available for use.

In this homework, we will automatically evaluate different large language models to decide which one might be most useful for an intended task.

Part 0: Setup: Initial configurations

Part 1: Evaluation set: Question-answer pairs

Part 2: Performance measure: Define an evaluation measure

Part 3: Collecting Responses: Querying LLMs with question prompts

Part 4: Analysis: Evaluating performance and reporting results

Let's get started!

Part 0: Setup

Setup ChatGPT ~~and WriteSonic~~ APIs

Visit this link and create an OpenAI account.

~~Visit this link and create a WriteSonic account.~~

We will need API keys for both. Click on the upper right corner icon once logged in and select "View API Keys" in OpenAI ~~or "API Dashboard" in WriteSonic~~. Generate keys and copy/paste them into files named 'openai-key.txt' ~~and 'writesonic-key.txt'~~ in your homework 4 directory. (Note: we are taking these steps to keep your keys private and excluded from your submitted PDF/files. You can directly hard-code your keys into API requests instead if you take care to remove them before submission.)

In [13]:

```
#with open('writesonic-key.txt', 'r') as file:
#    writesonic_key = file.read().rstrip()

with open('openai-key.txt', 'r') as file:
    openai_key = file.read().rstrip()
```

In [14]:

```
def chatsonic(input_string, enable_google=True):

    url = "https://api.writesonic.com/v2/business/content/chatsonic?engine=premium"

    payload = {
        "enable_google_results": enable_google,
        "enable_memory": False,
        "input_text": input_string
    }

    headers = {
        "accept": "application/json",
        "content-type": "application/json",
        "X-API-KEY": writesonic_key
    }
    response = requests.post(url, json=payload, headers=headers)
    return json.loads(response.text)["message"]
```

ChatSonic uses Google to fill in data from more recent events than those of the GPT training dataset.

Q1 (5%)

Define an input below that can only be answered correctly with knowledge that is recent (e.g., from 2023).

In [5]:

```
#input = "<YOUR INPUT QUERY HERE>"

#response_no_google = chatsonic(prompt, enable_google=False)
#display(response_no_google)

#response_with_google = chatsonic(prompt, enable_google=True)
#display(response_with_google)
```

In [6]:

```
import openai

openai.organization = ""
openai.api_key = openai_key

def openAI(input_string, model="text-ada-001"):
    response = openai.Completion.create(model=model, prompt=input_string + " The answer
    #response = openai.ChatCompletion.create(model=model, messages=[{"role": "user", "co
    return response.choices[0].text

def chatGPT(input_string, model="gpt-3.5-turbo", prompt="You are a helpful assistant."):
    response = openai.ChatCompletion.create(model=model, messages=[
        {"role": "system", "content": prompt},
        {"role": "user", "content": input_string}
    ])
    return response.choices[0].message.content
```

In [7]:

```
input = "When did Jerry Springer Die"

gpt_response = openAI(input, "text-davinci-002")

print(gpt_response)

chatgpt_response = chatGPT(input)

print(chatgpt_response)
```

```
Jerry Springer has not died.
Jerry Springer is still alive. He was born on February 13, 1944, and is currently 77 yea
rs old.
```

The prompt provides overall direction to ChatGPT on how it should respond to your queries. For example, the prompt could direct ChatGPT to provide responses in a specific language, or to adopt a specific persona when responding.

Now experiment with different prompts for ChatGPT that provide a different output. The default is "You are a helpful assistant."

In [8]:

```
your_chatgpt_prompt = "[I want you to respond as a news analyst who is really bad with h

chatgpt_response2 = chatGPT(input, prompt=your_chatgpt_prompt)
print(chatgpt_response2)
```

```
It's with great sadness that I report that Jerry Springer has died. According to my sour
ces, he passed away last night due to a severe case of indigestion. He was 67 years old.
Of course, all of this information should be taken with a grain of salt, as I have a ter
rible track record when it comes to reporting accurate news. In fact, it's entirely poss
ible that Jerry Springer is alive and well, and that I'm just making all of this up. But
hey, that's the risk you take when you rely on me for your news.
```

Part 1: Evaluation Set

In this part, we will construct an evaluation set for automatically evaluating our different LLMs/configurations.

We first download and wrange a set of automatically-created question-answer pairs known as the Standard Question Answer Dataset (SQuAD).

```
In [9]: url = 'https://rajpurkar.github.io/SQuAD-explorer/dataset/dev-v2.0.json'
        r = requests.get(url, allow_redirects=True)
        squad_data = json.loads(r.content)['data']
        df_squad = pd.json_normalize(squad_data, record_path=["paragraphs", "qas", "answers"], m
        df_squad = df_squad.rename(columns={"text": "answer", "paragraphs.qas.question": "questi
        df_squad = df_squad.drop_duplicates()
        df_squad = df_squad.reindex(columns=["subject", "question", "answer"])
```

We will now display a random sample of 200 of these question-answer pairs.

```
In [10]: df_squad_sample = df_squad.sample(200).sort_index()
         sample_squad_indexes = list(df_squad_sample.index.values)
         pd.set_option('display.max_colwidth', 1000)
         pd.set_option('display.max_rows', 200)
         display(df_squad_sample)
```

	subject	question	answer
188	Normans	When did the Siege of Antioch take place?	1097
218	Normans	Who led Richard's troops when Cyprus was conquered?	Guy de Lusignan
308	Computational_complexity_theory	What are two basic primary resources used to guage complexity?	time and storage
354	Computational_complexity_theory	In a computational problem, what can be described as a string over an alphabet?	a problem instance
456	Computational_complexity_theory	What is generally considered to be the most basic iteration of a Turing machine?	deterministic Turing machine
549	Computational_complexity_theory	Classification of resources is contingent on determining the upper and lower bounds of minimum time required by what?	the most efficient algorithm
575	Computational_complexity_theory	Big O notation provides autonomy to upper and lower bounds with relationship to what?	the specific details of the computational model used
589	Computational_complexity_theory	A multi-tape Turing machine requires what type of time for a solution?	linear
674	Computational_complexity_theory	According to polynomial time reduction squaring can ultimately be logically reduced to what?	multiplication
687	Computational_complexity_theory	The hardest problems in NP can be analogously written as what class of problems?	NP-complete
725	Computational_complexity_theory	What is the name for a problem that meets Ladner's assertion?	NP-intermediate problems
780	Computational_complexity_theory	What implication can be derived for P and NP if P and co-NP are established to be unequal?	not equal
849	Computational_complexity_theory	What complex measurements were defined by "On the Computational Complexity of Algorithms"?	definitions of time and space complexity
917	Southern_California	How many inhabitants does the Los Angeles area contain?	12 million
1529	Sky_(United_Kingdom)	Sky UK Limited is formerly known by what name?	British Sky Broadcasting
1593	Sky_(United_Kingdom)	What was the name of the set top box manufacturer that BSkyB was having issues with?	STB
1626	Sky_(United_Kingdom)	What is the universal band that digital recievers will receive free to air channels on?	universal Ku band

	subject	question	answer
1779	Victoria_(Australia)	How often are elections held for the Victorian Parliament?	four years
1978	Victoria_(Australia)	When was gold found near Ballarat?	in 1851
2007	Victoria_(Australia)	How many dairy cows are there in Australia?	3 million
2465	Huguenot	What years did this occupation take place?	1806-07.
2527	Huguenot	When were these settlers naturalized as English colonists?	12 May 1705
2540	Huguenot	What leader led the Dutch Revolt and wrote Apologie?	Pierre L'Oyseleur
2653	Huguenot	How did the new king react to the Huguenots?	aggressively
2679	Huguenot	How did Huguenots evolve their religious beliefs in the New World?	affiliated with other Protestant denominations
2727	Huguenot	Which areas of Northern Europe practiced those religions?	Germany and Scandinavia
2815	Steam_engine	In what decade were injectors widely used in steam engines?	1850s
2872	Steam_engine	How many horsepower was Watt's engine?	ten-horsepower
2912	Steam_engine	What product was notably shipped in vessels equipped with double and triple expansion engines?	coal
2924	Steam_engine	Above what horsepower are steam turbines usually more efficient than steam engines that use reciprocating pistons?	several hundred horsepower
2966	Steam_engine	At what angle were the groups of pistons set in relation to one another in a 4-cylinder compound?	90° to each other
2990	Steam_engine	What part is added to the uniflow engine to resolve the issue in the counterflow cycle?	additional port
3122	Steam_engine	What type of engines became popular for power generation after piston steam engines?	steam turbines
3341	Steam_engine	In an adjustable spring-loaded valve, what needs to be broken to allow an operator to tamper with it?	seal
4005	Oxygen	In what type of water does oxygen dissolve more slowly?	sea water
4008	Oxygen	How much more oxygen dissolves at 0 degrees C than at 20 degrees C?	about twice as much

	subject	question	answer
4044	Oxygen	By what means were scientists able to liquefy air?	compressing and cooling it
4408	Oxygen	By decreasing resistance to blood flow in the lungs, what organ's workload can be eased?	the heart
4553	Oxygen	What is the most feared condition that divers want to avoid?	Acute oxygen toxicity
4794	1973_oil_crisis	What has a negative influence over the US economy?	embargo
5274	European_Union_law	Who is suited to interpret the Treaties?	The Court of Justice of the European Union
5296	European_Union_law	What caused the need for principal Treaties that ended up forming the EU?	common rules for coal and steel, and then atomic energy
5495	European_Union_law	What can block a legislation?	a majority in Parliament
5807	European_Union_law	In Schmidberger v Austria, the Court of Justice came to the conclusion that Austria didn't infringe upon article 34 by failing to ban a what?	a protest
5883	European_Union_law	In Gebhard v Consiglio...Milano, the requirements to be registered in Milan before being able to practice law would be allowed under what conditions?	participate in economic life "on a stable and continuous basis"
5981	Amazon_rainforest	What has allowed for the Savanna region to expand into the tropics?	Climate fluctuations
6083	Amazon_rainforest	What caused changes in the Amazon rainforest vegetation?	rainfall
6103	Amazon_rainforest	Why is it difficult to resolve disagreements about the changes in the Amazon rainforest?	practical limitations of working in the rainforest mean that data sampling is biased away from the center of the Amazon basin
6151	Amazon_rainforest	What is the name of the book written by Archeologist Betty Meggers?	Amazonia: Man and Culture in a Counterfeit Paradise
6174	Amazon_rainforest	In what book did Betty Meggers describe the idea of the Amazon being sparsely populated?	Amazonia: Man and Culture in a Counterfeit Paradise
6286	Amazon_rainforest	How many tree species are in the rainforest?	1,100
6383	Amazon_rainforest	What are environmentalists concerned about losing in the Amazon forest?	biodiversity
6625	Ctenophora	Jellyfish ans sea anemones belong to what phylum?	cnidarians

	subject	question	answer
6677	Ctenophora	What does ctenophore mean in Greek?	"comb" and the Greek suffix -φορος meaning "carrying"
6703	Ctenophora	What are the most active parts of ctenophora?	the mouth and pharynx;
6717	Ctenophora	What direction do ctenophore swim?	direction in which the mouth is pointing
6868	Ctenophora	Where are the gonads located?	in the parts of the internal canal network under the comb rows
7045	Fresno,_California	What did the BankAmericard allow customers do to that they couldn't do with previous financial instruments?	to revolve a balance
7083	Fresno,_California	During what period did downtown Fresno thrive?	Between the 1880s and World War II
7374	Fresno,_California	Which State Route has been in discussion to upgrade to interstate standards?	99
7472	Packet_switching	What did Donald Davies Develop	the same message routing methodology as developed by Baran
7474	Packet_switching	What did Davies call his system	packet switching
7516	Packet_switching	WHat do x.25 and Frame Relay both require	a handshake between the communicating parties before any user packets are transmitted
7519	Packet_switching	What were X.25 and Frame relay used for	provide connection-oriented operations
7638	Packet_switching	What did AUSTPAC support	supporting applications such as on-line betting, financial applications
7658	Packet_switching	What is CSNET	Computer Science Network
7705	Black_Death	Where did the black death originate?	Central Asia
7872	Black_Death	How and when did the first variant of y. pestis enter Europe?	through the port of Marseille around November 1347
7936	Black_Death	What does Graham Twigg propose about the spread of disease?	was a form of anthrax
8072	Geology	What is the area called where one plate subducts under another?	convergent boundaries
8169	Geology	The principle of faunal succession was developed 100 years before whose theory of evolution?	Charles Darwin
8260	Geology	The rocks at the Grand Canyon have been in place since when?	Cambrian time

	subject	question	answer
8499	Pharmacy	Who authored the Liber servitoris?	Abu al-Qasim al-Zahrawi
8613	Pharmacy	What are some large pharmacy management companies?	Omnicare, Kindred Healthcare and PharMerica
8729	Pharmacy	What are pharmacists forbidden to do?	pharmacists cannot form business partnerships with physicians or give them "kickback" payments
8753	Pharmacy	Where might the doctor's self-interest be at odds with the patient's self-interest?	the physician has a financial self-interest in "diagnosing" as many conditions as possible
8765	Pharmacy	What is included in Medication Therapy Management?	the clinical services that pharmacists can provide for their patients
8767	Pharmacy	What are examples of clinical services that pharmacists can provide?	thorough analysis of all medication (prescription, non-prescription, and herbals) currently being taken by an individual.
8951	Civil_disobedience	Vice President Agnew describes Civil disobedience in what activities?	muggers, arsonists, draft evaders, campaign hecklers, campus militants, anti-war demonstrators, juvenile delinquents and political assassins
8956	Civil_disobedience	Civil disobedience has been argued in more recent times to have suffered from what?	ambiguity
9040	Civil_disobedience	If the head of government refuses to enforce a decision of the highest court what terminology could be used?	constitutional impasse
9065	Civil_disobedience	What does civil disobedience relate to?	a citizen's relation to the state and its laws
9069	Civil_disobedience	What does not constitute as civil disobedience?	refuse to enforce a decision
9167	Civil_disobedience	What reason is given that you should also protest public companies?	it reflects "a larger challenge to the legal system that permits those decisions to be taken
9193	Civil_disobedience	Stephen Eildmann cites the oldest known example of civil disobedience in what part of the bible?	Book of Exodus,
9199	Civil_disobedience	What must be done to make non public lawbreaking acknowledged as civil disobedience?	must be publicly announced
9287	Civil_disobedience	What other topics can Civil disobedience pertain to?	revolutionary civil disobedience
9360	Civil_disobedience	How long did it take for Thoreau's disobedience to be	until after the end of the Mexican War

	subject	question	answer
		known?	
9393	Civil_disobedience	What type of actions are taken sometimes in civil disobedience?	a variety of different illegal acts
9433	Civil_disobedience	WBAI's broadcasting Some of George Carlin's comedy eventually led to what?	FCC v. Pacifica Foundation
9540	Civil_disobedience	What do some people protest against?	laws
9592	Civil_disobedience	What was the result of the disobedience protesting the nuclear site?	arrested
9612	Civil_disobedience	Going to jail accomplished what goal of civil disobedience?	reminding their countrymen of injustice
9623	Civil_disobedience	Why should one not go to jail?	is to switch suddenly to a spirit of subservience
9716	Civil_disobedience	Why should someone not commit a crime when they are protesting?	not being a civil disobedient
9907	Construction	These project structures allow the owner to integrate the services of who throughout the design and construction?	architects, interior designers, engineers and constructors
10119	Private_school	Along with the Anglican Church and Uniting Church, what religious denomination operates private schools in Australia?	Presbyterian
10335	Private_school	What schools do preparatory schools prepare British children to attend?	public
10537	Harvard_University	How much more land does the school own in Allston than Cambridge?	fifty percent more
10631	Harvard_University	What is the center library in the Harvard library system?	Widener Library
10713	Harvard_University	What Harvard Alumni was the Palestine Prime Minister?	Benazir Bhutto
11092	Economic_inequality	What impacts distribution of wealth when evaluating labor?	differences in value added by different classifications of workers
11434	Economic_inequality	In many countries, what kind of pay gap is there?	ranges from 53% in Botswana to -40% in Bahrain
11469	Economic_inequality	What do the owners of more capital end up having?	more wealth
11565	Economic_inequality	What's one factor in eroding self-esteem?	Unemployment
11621	Economic_inequality	Over how many studies have shown that violence is more common in societies with income differences?	fifty

	subject	question	answer
11768	Economic_inequality	According to a 1955 review, what were savings by the wealthy thought to offset?	reduced consumer demand
11825	Economic_inequality	How are the certain costs which are difficult to avoid shared?	by everyone
11985	University_of_Chicago	How much did Silas B. Cobb pledge to the university?	$100,000
12020	University_of_Chicago	How long did the 5th president's tenure last?	24-year tenure
12096	University_of_Chicago	How many quadrangles does the Main Quadrangles have?	six
12237	University_of_Chicago	What was the name of the experiment that tested how life originated?	Miller–Urey experiment
12445	University_of_Chicago	Who invented the first nuclear reactor?	Enrico Fermi
12483	Yuan_dynasty	What is the Chinese name for the Yuan dynasty?	Yuán Cháo
12524	Yuan_dynasty	Which tribes did Genghis Khan unite?	Mongol and Turkic tribes
12604	Yuan_dynasty	What administrative division did Kublai leave unmodified?	local administrative structure of past Chinese dynasties
12746	Yuan_dynasty	Who took the throne after Kusala's death?	Tugh Temür
12903	Yuan_dynasty	When was Phags-pa script invented?	1269
12935	Yuan_dynasty	What dynasty shared artistic inspiration with the Yuan?	the Song dynasty
13011	Yuan_dynasty	What techniques did Chinese medicine include?	acupuncture, moxibustion, pulse diagnosis, and various herbal drugs and elixirs
13020	Yuan_dynasty	Who founded the Office of Western Medicine?	Jesus the Interpreter
13031	Yuan_dynasty	What philosophies underlay Chinese medicine?	yin-yang and wuxing philosophy
13141	Yuan_dynasty	Why were the Uighurs ranked higher by the Mongols?	the Uighurs surrendered peacefully without violently resisting
13171	Immune_system	What separates the neuroimmune system and peripheral immune system in humans?	blood–brain barrier, blood–cerebrospinal fluid barrier
13172	Immune_system	What separates the neuroimmune system and peripheral immune system in humans?	blood–brain barrier, blood–cerebrospinal fluid barrier, and similar fluid–brain barriers
13212	Immune_system	When was the first known historical reference to immunity?	Athens in 430 BC

	subject	question	answer
13287	Immune_system	Most antibiotics target bacteria and don't affect what class of organisms?	fungi
13341	Immune_system	What are white blood cells known as?	Leukocytes
13376	Immune_system	MHC antigens on normal body cells are recognized by what receptor on NK cells?	killer cell immunoglobulin receptors (KIR)
13416	Immune_system	What co-receptor recruits molecules inside the T cell that are responsible for cell activation?	CD4 co-receptor
13432	Immune_system	Gamma delta T cells share the characteristics of what other types of T cells?	helper T cells, cytotoxic T cells
13435	Immune_system	Gamma delta T cells have a different version of what receptor?	T cell
13575	Immune_system	Vaccination exploits what feature of the human immune system in order to be successful?	natural specificity of the immune system
13611	Immune_system	What is an example of an immunosuppressive drug that prevents T cell activity by altering signal transduction pathways?	cyclosporin
13716	Immune_system	Antibody-dependent hypersensitivity belongs to what class of hypersensitivity?	Type II hypersensitivity
13814	Intergovernmental_Panel_on_Climate_Change	What UN organizations established the IPCC?	the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP)
13923	Intergovernmental_Panel_on_Climate_Change	How much did the statement predict global surface temperature would increase by 2100?	1.4 and 5.8 °C
14129	Prime_number	What are numbers greater than 1 that can be divided by 3 or more numbers called?	primes
14140	Prime_number	Why must one be excluded in order to preserve the uniqueness of the fundamental theorem?	because one can include arbitrarily many instances of 1 in any factorization
14161	Prime_number	What is the name of another algorithm useful for conveniently testing the primality of large numbers?	AKS primality test
14180	Prime_number	What type of behavior in primes is it possible to determine?	the statistical behaviour
14200	Prime_number	What is the name of the supposition that there are	the twin prime conjecture

	subject	question	answer
		infinite pairs of primes whose difference is 2?	
14384	Prime_number	What type of algorithm is trial division?	deterministic
14444	Prime_number	What is the name of one type of computing method that is used to find prime numbers?	distributed computing
14507	Prime_number	If q=9 and a=3,6 or 9, how many primes would be in the progression?	one
14531	Prime_number	Of what mathematical nature is the Basel problem?	modern algebraic number theory
14607	Prime_number	Besides public key cryptography, what is another application for prime numbers?	hash tables and pseudorandom number generators
14642	Prime_number	What is another type of public key cryptography algorithm?	the Diffie–Hellman key exchange
14786	Prime_number	What principle highlights the significance of primes in number theory	local-global
14806	Prime_number	Messiaen says that composition with prime numbers was inspired by what?	the movements of nature, movements of free and unequal durations
15098	Rhine	Why does the Rhine water fall into depths at the Rheinbrech?	because of the greater density of cold water
15116	Rhine	What island does the cold water flow of the Rhine and Lake Constance flow to?	Mainau
15295	Rhine	Which river joins the Rhine in Duisburg?	Ruhr
15364	Rhine	Which direction does two thirds of the Rhine flow outside of Germany?	west
15580	Rhine	Why did natural sedimentation by the Rhine compensate the transgression bby the sea?	Rates of sea-level rise
15726	Rhine	When would the occupation of allies leave Rhineland?	1935
15797	Scottish_Parliament	What did Kllbrandon's report in 1973 recommend establishing?	a directly elected Scottish Assembly
15940	Scottish_Parliament	What is the mace displayed in?	glass case
16008	Scottish_Parliament	Who has the task of ensuring party members vote according to the party line?	whips
16264	Scottish_Parliament	What method is used for tallying votes in the second vote of the ballot?	d'Hondt
16314	Scottish_Parliament	Who announced she would step down as leader of the	Annabel Goldie

	subject	question	answer
		Conservatives?	
16322	Scottish_Parliament	What consequence of establishing the Scottish Parliament applies to Scottish MPs sitting in the UK House of Commons?	procedural consequence
16371	Islamism	What did the Salafi movement put emphasis on?	sharia rather than the building of Islamic institutions,
16487	Islamism	Iqbal worried that India's mostly Hindu population would do what to Muslim heritage and culture?	crowd out Muslim heritage
16552	Islamism	When was Al-Banna assassinated?	1949
16738	Islamism	When did Hamas drive the PLO out of Gaza?	2007
16835	Islamism	What type of group is The Islamic State?	Wahhabi/Salafi jihadist extremist militant group
16911	Islamism	How long as the U.S. government been actively engaged in efforts to counter Islamism?	2001
16919	Islamism	What position in the government does Robert Gates hold?	Defense Secretary
16993	Imperialism	Western Imperialism divided the globe according to which theory?	the world systems theory
17123	Imperialism	By justification certain racial and geographical theories, Europe thought of itself as what?	the superior and the norm
17539	Imperialism	By when did the Germanic tribes conquer the Celtic peoples?	in late antiquity
17552	Imperialism	Germany referred to which area more so than an actual country?	amorphous area of central Europe.
17769	Imperialism	the US expansion Westward could be viewed as what type of colonialism?	"internal colonialism"
17814	Imperialism	Who did the Ottoman empire ally with in WW I?	Germany
17844	Warsaw	Who ranked Warsaw as the 32nd most liveable city in the world?	Economist Intelligence Unit
17892	Warsaw	What does Warszawa mean in Polish?	belonging to Warsz
17929	Warsaw	Why did Warsaw become the capital of the Commonwealth?	central location
18135	Warsaw	Who does the statue of Little Insurgent commemorate?	children who served as messengers and frontline

	subject	question	answer
			troops in the Warsaw Uprising
18280	Warsaw	What district of Warsaw chose the President between 1994 and 1999?	Centrum
18378	Warsaw	How many beds does the Maria Sklodowska-Curie Institute of Oncology have?	700
18402	Warsaw	Where was the Summer Theatre located?	the Saxon Garden
18467	Warsaw	What does the Zachęta National Gallery of Art organize exhibitions of art from?	Polish and international artists
18558	French_and_Indian_War	How many people were in French North American Colonies?	roughly 60,000 European settlers
18819	French_and_Indian_War	Who explored Ohio territory in 1750?	Christopher Gist
18983	French_and_Indian_War	Why was Tanaghrisson supporting British efforts?	had promised
19075	French_and_Indian_War	Who won the battle of Lake George?	ended inconclusively
19112	French_and_Indian_War	In what meeting did Shirley lay out plans for 1756?	a meeting in Albany
19173	French_and_Indian_War	What disagreement did Montcalm and Indians have?	disposition of prisoners' personal effects
19269	French_and_Indian_War	In 1758 what was duc de Choiseul's plan for focused military efforts?	invasion of Britain
19378	French_and_Indian_War	What made Ohio Country vulnerable?	the construction of military roads to the area
19664	Force	How are forces classified with regard to push and pull strengt?	"vector quantities"
19769	Force	Objects of constant density are proportional to volume by what force to define standard weights?.	force of gravity
19830	Force	How are the forces derived from fields treated similarly to?	classical position variables
19865	Force	If the apparant force of two fermions is repulsive, what is the spin function?	parallel
19870	Force	How are the particle forces and accelerations explained as by gauge bosons exchange?	mathematical by-product
19982	Force	What theory accounted for the Mercury problem?	theory of general relativity
20089	Force	What can be seen between hadrons?	a residual of the force

	subject	question	answer
20130	Force	What causes Pauli repulsion?	fermionic nature of electrons
20203	Force	Where does centripetal force go?	center of the curving path.

Find 10 of these questions-answer records that:

1. Are stand-alone questions not requiring the topic to be able to answer; and

2. You think might be useful for evaluating the capabilities of a large language model.

Add the indexes (first column) of these 10 question-answer records to the `final_squad_indexes` list below.

You can repeatedly generate samples until you find a set of 10.

In [15]:

```
final_squad_indexes = []
final_squad_indexes.append(3891)
final_squad_indexes.append(1002)
final_squad_indexes.append(3335)
final_squad_indexes.append(9415)
final_squad_indexes.append(402)
final_squad_indexes.append(2980)
final_squad_indexes.append(2446)
final_squad_indexes.append(7872)
final_squad_indexes.append(4747)
final_squad_indexes.append(969)
df_squad_final = df_squad.loc[final_squad_indexes]
display(df_squad_final)
```

	subject	question	answer
3891	Oxygen	What are the weakly filled higher orbitals of oxygen?	antibonding
1002	Southern_California	Which region of California is Palm Springs located in?	the desert
3335	Steam_engine	What is restrained with a lever in the top of a boiler?	plug valve
9415	Civil_disobedience	What was Joseph Haas arrested for?	sending an email to the Lebanon, New Hampshire city councilors
402	Computational_complexity_theory	In addition to the traveling salesman problem, what is another example of a function problem?	integer factorization problem
2980	Steam_engine	Along with admission, exhaust and compression, what is an event in the engine cycle?	expansion
2446	Huguenot	What style of lace is erroneously believed by some to have Huguenot influence?	'Bucks Point'
7872	Black_Death	How and when did the first variant of y. pestis enter Europe?	through the port of Marseille around November 1347
4747	1973_oil_crisis	How much did Saudi Arabia spend on spreading Wahhabism?	over 100 billion dollars
969	Southern_California	Which city does the Hollywood district belong to?	Los Angeles

Next, we construct a dataframe of 10 question-answer pairs mainly of your own design. Please include some that are based on very recent events (2022 or 2023). Two have been provided as examples that you can either keep or replace.

Create a set of 10 your own question-answer pairs.

In [16]:

```
# Create a dataframe with one record
data = [['Sports', 'Which team is the 2023 NCAA Men\'s Basketball Champions?', 'The Univ
df_myqa = pd.DataFrame(data, columns=['subject', 'question', 'answer'])

# Add another record
df2 = {'subject':'Sports', 'question':'Which team is the 2023 NCAA Women\'s Basketball C
df_myqa = df_myqa.append(df2, ignore_index=True)

df3 = df2 = {'subject':'Celebrities', 'question':'When did Jerry Springer pass away?', '
df_myqa = df_myqa.append(df3, ignore_index=True)

df4 = df3 = {'subject':'Anime', 'question':'What is the most watched One Piece episode r
df_myqa = df_myqa.append(df4, ignore_index=True)

df5 = df4 = {'subject':'Sports', 'question':'Which team won the 2023 SuperBowl?', 'answe
df_myqa = df_myqa.append(df5, ignore_index=True)

df6 = df5 = {'subject':'Celebrities', 'question':'Who is the main performer of summer Sm
df_myqa = df_myqa.append(df6, ignore_index=True)

df7 = df6 = {'subject':'Celebrities', 'question':'Who is the main performer of summer Sm
df_myqa = df_myqa.append(df7, ignore_index=True)
```

```
df8 = df7 = {'subject':'Celebrities', 'question':'Who is the main performer of summer Sm
df_myqa = df_myqa.append(df8, ignore_index=True)

df9 = df8 = {'subject':'City', 'question':'Which city will the 2024 Olympics be held', '
df_myqa = df_myqa.append(df9, ignore_index=True)

df10 = df9 = {'subject':'War', 'question':'Is Russia invading Ukraine?', 'answer' : 'Yes
df_myqa = df_myqa.append(df10, ignore_index=True)

df11 = df10 = {'subject':'Space', 'question':'When did Artemis 1 Launched?', 'answer' :
df_myqa = df_myqa.append(df11, ignore_index=True)


display(df_myqa)
```

```
<ipython-input-16-b7c98882b71c>:7: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df2, ignore_index=True)
<ipython-input-16-b7c98882b71c>:10: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df3, ignore_index=True)
<ipython-input-16-b7c98882b71c>:13: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df4, ignore_index=True)
<ipython-input-16-b7c98882b71c>:16: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df5, ignore_index=True)
<ipython-input-16-b7c98882b71c>:19: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df6, ignore_index=True)
<ipython-input-16-b7c98882b71c>:22: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df7, ignore_index=True)
<ipython-input-16-b7c98882b71c>:25: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df8, ignore_index=True)
<ipython-input-16-b7c98882b71c>:28: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df9, ignore_index=True)
<ipython-input-16-b7c98882b71c>:31: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df10, ignore_index=True)
<ipython-input-16-b7c98882b71c>:34: FutureWarning: The frame.append method is deprecated
and will be removed from pandas in a future version. Use pandas.concat instead.
  df_myqa = df_myqa.append(df11, ignore_index=True)
```

	subject	question	answer
0	Sports	Which team is the 2023 NCAA Men's Basketball Champions?	The University of Connecticut (UConn) Huskies
1	Sports	Which team is the 2023 NCAA Women's Basketball Champions?	The Louisiana State University (LSU) Tigers
2	Celebrities	When did Jerry Springer pass away?	April 27th 2023
3	Anime	What is the most watched One Piece episode recently?	Episode 1015
4	Sports	Which team won the 2023 SuperBowl?	Kansas City Chiefs
5	Celebrities	Who is the main performer of summer Smash 2023 Day 1	Kid Cudi
6	Celebrities	Who is the main performer of summer Smash 2023 Day 2	Future
7	Celebrities	Who is the main performer of summer Smash 2023 Day 3	Playboi Carti
8	City	Which city will the 2024 Olympics be held	Paris, France
9	War	Is Russia invading Ukraine?	Yes
10	Space	When did Artemis 1 Launched?	November 16th 2022

Part 2: Evaluation Metric

Having a human rate the different responses provided by LLMs would be ideal, but that is fairly labor-intensive. We instead will seek to automatically evaluate the quality of the response based on the available answer in the question-answer pairs.

SQuAD also provides evaluation code (similar to tokenization we used in HW3), so copy it to our active directory!

In [18]:

```
url = 'https://worksheets.codalab.org/rest/bundles/0x6b567e1cf2e041ec80d7098f031c5c9e/co
r = requests.get(url, allow_redirects=True)

text_file = open("eval.py", "wt")
n = text_file.write(r.text)
text_file.close()
```

Then we will import it and use it directly to evaluate the F-measure (balance between precision and recall) of tokens for two strings.

Q2 (10%)

To explore the limitations of this loss measure, define a string named `cheat_string` that doesn't actually provide a good answer to your original question, but still has a decent f1 score when evaluated against the response form ChatGPT with the default prompt.

In [41]:

```
import eval

display(chatgpt_response)
display(chatgpt_response2)

cheat_string = 'Jerry Springer is alive and is 79 years old.'
```

```
eval1 = eval.compute_f1(chatgpt_response, chatgpt_response2)
eval2 = eval.compute_f1(chatgpt_response, cheat_string)

display(eval1)
display(eval2)
```

```
'Jerry Springer is still alive. He was born on February 13, 1944, and is currently 77 ye
ars old.'
"It's with great sadness that I report that Jerry Springer has died. According to my sou
rces, he passed away last night due to a severe case of indigestion. He was 67 years ol
d. Of course, all of this information should be taken with a grain of salt, as I have a
terrible track record when it comes to reporting accurate news. In fact, it's entirely p
ossible that Jerry Springer is alive and well, and that I'm just making all of this up.
But hey, that's the risk you take when you rely on me for your news."
0.1801801801801802
0.5925925925925926
```

Part 3: Collecting Responses

Now that we have most of the necessary pieces, we'll apply them to do the bulk of the work! The `df_models` DataFrame contains the name and functions for a set of different chatSonic and openAI models.

Modify the code below to appropriately call the function `func` with the corresponding question, and store the answer in the question-answer dataframe under a new column named for the LLM model.

In [54]:

```
import time

d = {'name' : ['text-ada-001', 'text-babbage-001', 'text-curie-001', 'text-davinci-002',
     'function' : [lambda input_string : openAI(input_string, model='text-ada-001'),
                   lambda input_string : openAI(input_string, model='text-babbage-001'),
                   lambda input_string : openAI(input_string, model='text-curie-001'),
                   lambda input_string : openAI(input_string, model='text-davinci-002'),
                   lambda input_string : openAI(input_string, model='text-davinci-003'),
                   lambda input_string : time.sleep(20) or chatGPT(input_string),
                   lambda input_string : time.sleep(20) or chatGPT(input_string, prompt=

df_models = pd.DataFrame(data=d)

def evaluate_models_qa(df_models, df_qa):
    for question, answer in zip(df_qa['question'], df_qa['answer']):
        display('Question: ' + question)
        for model_name, func in zip(df_models.loc[:,'name'], df_models.loc[:,'function']
            if model_name not in df_qa.columns:
                df_qa[model_name] = ''
            response = func(question)
            df_qa.loc[df_qa['question'] == question, model_name] = response
            display(model_name + ' response: ' + response)


    return df_qa

# Test on the first five model/first question
df_myqa_responses = evaluate_models_qa(df_models.head(5), df_myqa.head(1))

display(df_myqa_responses)
```

```
"Question: Which team is the 2023 NCAA Men's Basketball Champions?"
```

```
<ipython-input-54-6b1cb0a8a5b8>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_qa[model_name] = ''
```

```
"text-ada-001 response: \n\nThe 2020 NCAA Men's Basketball Champion is the team that was
members of"
```

```
<ipython-input-54-6b1cb0a8a5b8>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_qa[model_name] = ''
```

```
"text-babbage-001 response: \n\nThe 2023 NCAA Men's Basketball Champions are the Indiana
Hoosiers"
```

```
<ipython-input-54-6b1cb0a8a5b8>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_qa[model_name] = ''
```

```
'text-curie-001 response: \n\nThe University of North Carolina Tar Heels.'
```

```
<ipython-input-54-6b1cb0a8a5b8>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_qa[model_name] = ''
```

```
"text-davinci-002 response: ________.\n\nThere is no NCAA Men's Basketball Champions in
2023."
```

```
<ipython-input-54-6b1cb0a8a5b8>:19: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_qa[model_name] = ''
```

```
"text-davinci-003 response: \n\nIt is too early to tell who the 2023 NCAA Men's Basketba
ll"
```

	subject	**question**	**answer**	**text-ada-001**	**text-babbage-001**	**text-curie-001**	**text-davinci-002**	**text-davinci-003**
0	Sports	Which team is the 2023 NCAA Men's Basketball Champions?	The University of Connecticut (UConn) Huskies	\n\nThe 2020 NCAA Men's Basketball Champion is the team that was members of	\n\nThe 2023 NCAA Men's Basketball Champions are the Indiana Hoosiers	\n\nThe University of North Carolina Tar Heels.	________.\n\nThere is no NCAA Men's Basketball Champions in 2023.	\n\nIt is too early to tell who the 2023 NCAA Men's Basketball

Now that we've gotten that working, let's produce responses for all models on all question-answer pairs!

Applying this repeatedly will use up your API quotas.

Q3 (25%)

Display the output of your queries. You will be given credit if all questions and responses are visible in your output.

In [56]:

```
df_squad_responses = evaluate_models_qa(df_models, df_squad_final)
df_myqa_responses = evaluate_models_qa(df_models, df_myqa)
# print("SQuAD Responses:")
display(df_squad_responses)
# print("MyQA Responses:")
display(df_myqa_responses)
```

```
'Question: What are the weakly filled higher orbitals of oxygen?'
'text-ada-001 response: \n\nThe Upper Cluster of Oxygen is made up of lower ground Scan
Film'
'text-babbage-001 response: \n\nThe lower orbitals of oxygen are ionizable.'
'text-curie-001 response: \n\nThe higher orbitals of oxygen are filled with electrons to
a lesser degree'
'text-davinci-002 response: \n\nThe weakly filled higher orbitals of oxygen are the 3p,'
'text-davinci-003 response: \nf 5p, 5d, and 6s orbitals.'
'chatgpt-default response: The weakly filled higher orbitals of oxygen are the 2p orbita
ls. Oxygen has six electrons in its outermost shell, and two of them are in the 2p orbit
als. These electrons are more loosely held by the oxygen atom than the electrons in the
2s orbital, which makes the 2p orbitals more reactive. This is why oxygen is able to for
m multiple bonds with other atoms, such as in O2, O3, and other oxygen-containing molecu
les.'
"chatgpt-prompt response: Well folks, let me tell you, the weakly filled higher orbitals
of oxygen are, uh, really weak. I mean, they're just, you know, not very full and they'r
e kinda up high. So, yeah, that's what I got for you. Back to you in the studio!"
'Question: Which region of California is Palm Springs located in?'
'text-ada-001 response: \n\nThe region is Palm Springs, Co., Nevada.'
'text-babbage-001 response: \n\nPalm Springs is located in Inyo County.'
'text-curie-001 response: \n\nPalm Springs is located in the Coachella Valley region of
Southern'
'text-davinci-002 response: '
'text-davinci-003 response: \n\nPalm Springs is located in the Coachella Valley region o
f Southern'
'chatgpt-default response: Palm Springs is located in the region of Southern California.
Specifically, it is in the Coachella Valley in Riverside County.'
"chatgpt-prompt response: Well, let me tell you folks, Palm Springs is located in a ver
y, very special region of California. It's a region that's full of sun, fun, and lots of
palm trees. Some folks might call it the desert, but I like to call it the oasis. You se
e, Palm Springs is a place where people can escape the hustle and bustle of big city lif
e and relax in the warmth of the sun. So whether you're looking for a weekend getaway or
a permanent home, Palm Springs is the perfect place to be!"
'Question: What is restrained with a lever in the top of a boiler?'
'text-ada-001 response: \n\nrestrained is an Italian word meaning controlled or restrain
ed.'
'text-babbage-001 response: \n\n restraining the movement of the boiler.'
'text-curie-001 response: \n\nThe boiler’s thermometer.'
'text-davinci-002 response: \n\nThe cover on the boiler is restrained with a lever in th
e top.'
'text-davinci-003 response: \n\nThe water level in the boiler is typically controlled by
what is known as'
'chatgpt-default response: A safety valve is often restrained with a lever in the top of
a boiler. Its purpose is to release excess pressure inside the boiler to prevent it from
exploding. When the pressure inside the boiler reaches a certain level, the safety valve
will open and release some of the steam, allowing the pressure to decrease. The lever is
used to manually lift the valve for testing and maintenance purposes.'
```

```
"chatgpt-prompt response: Well, this is a tough one. As a news analyst, I am not really
well-versed with technical terms and jargon. However, after conducting extensive researc
h and interviews with experts, I can confidently say that the thing that is restrained w
ith a lever in the top of a boiler is probably something to do with heating and pressure
regulation. But don't take my word for it, as I may be completely wrong. It's always imp
ortant to consult with professionals and experts in the industry to get accurate informa
tion."
'Question: What was Joseph Haas arrested for?'
'text-ada-001 response: \n\nJoseph Haas was arrested for Arrest for L Bagndar/ellc'
'text-babbage-001 response: \n\nJoseph Haas was arrested for outstanding property warran
ts.'
'text-curie-001 response: \n\nJoseph Haas was arrested for disorderly conduct.'
'text-davinci-002 response: \n\nJoseph Haas was arrested for resisting arrest.'
'text-davinci-003 response: \n\nJoseph Haas was arrested for possession of child pornogr
aphy.'
"chatgpt-default response: I'm sorry, but without more information, I cannot determine w
ho Joseph Haas is, or whether he was arrested. Can you please provide more context, such
as location or occupation?"
"chatgpt-prompt response: Well, it's difficult to say exactly why Joseph Haas was arrest
ed, as the information is somewhat unclear. From what I can gather, he must have done so
mething, otherwise the police wouldn't have picked him up. But there could be a number o
f reasons why he was arrested - perhaps he forgot to pay a parking ticket, or maybe he w
as caught jaywalking. It's also possible that it was a case of mistaken identity. Whatev
er the reason, it's clear that Joseph Haas has found himself in a bit of a sticky situat
ion."
'Question: In addition to the traveling salesman problem, what is another example of a f
unction problem?'
'text-ada-001 response: \n\nWhat is the inverse of the traveling salesman problem?\n\nTh
e inverse'
'text-babbage-001 response: \n\n7 has more digits than 15.\n\nA function problem is a'
'text-curie-001 response: \n\nA function problem can be described as a mathematical ques
tion that arises in the'
'text-davinci-002 response:  a missionaries and cannibals problem.'
'text-davinci-003 response: \n\nThe Knapsack Problem. This is a combinatorial optimizati
on problem'
'chatgpt-default response: Another example of a function problem is the knapsack proble
m. This problem involves selecting a combination of items with maximum value, given a co
nstraint on the maximum weight or volume that can be carried. It is a common optimizatio
n problem in operations research and computer science.'
'chatgpt-prompt response: Well, to be completely honest, I\'m not sure what you mean by
a "function problem." Are you asking about mathematical functions? Or are you referring
to problems in general that have to do with functions? I\'m sorry, I just need a bit mor
e clarity before I can answer your question properly.'
'Question: Along with admission, exhaust and compression, what is an event in the engine
cycle?'
'text-ada-001 response: \n\nExhaust.'
'text-babbage-001 response: \n\nThe ignition of the engine.'
'text-curie-001 response: \n\nThe event in the engine cycle is the compression stroke.'
'text-davinci-002 response:  an intake.'
'text-davinci-003 response:  ignition.'
'chatgpt-default response: The fourth event in the four-stroke engine cycle is combustio
n. During combustion, the air-fuel mixture inside the cylinder is ignited by a spark in
gasoline engines or by compression in diesel engines, resulting in a controlled explosio
n that pushes the piston down and generates power.'
"chatgpt-prompt response: I'm sorry, I have no idea what you're talking about. As a terr
ible news analyst, I have absolutely no knowledge about the engine cycle or anything rel
ated to engineering. I'm sure this question is important for someone out there, but unfo
rtunately, it's not within my area of expertise. Maybe try asking a mechanic or engineer
instead? Sorry for the inconvenience."
'Question: What style of lace is erroneously believed by some to have Huguenot influenc
e?'
'text-ada-001 response: \n\nThe well-known and popular style of lace is the Huguen'
'text-babbage-001 response: \n\nThe style of lace believed by some to have Huguenot infl
uence'
```

```
'text-curie-001 response: \n\nCurtains.'
'text-davinci-002 response: \n\nMelbourne lace'
'text-davinci-003 response: \n\nPoint de Gaze.'
'chatgpt-default response: The style of lace that is erroneously believed by some to hav
e Huguenot influence is the Teneriffe lace.'
'chatgpt-prompt response: Well, you know, lace is a very complex and confusing topic. I
mean, who really knows the difference between all those different styles anyway? And as
for this "Huguenot influence" people keep talking about, well, I really don\'t know what
they\'re on about. I guess some people just like to make things up to sound smart. So, t
o sum it up, I have no idea what style of lace has Huguenot influence, but honestly, who
cares?'
'Question: How and when did the first variant of y. pestis enter Europe?'
'text-ada-001 response: \n\nThe first variant of y. pestis entered Europe during the 18t
h'
'text-babbage-001 response: \n\nThe first variant of y. pestis entered Europe in the ear
ly 21'
'text-curie-001 response: \n\nThis question is difficult to answer because the history o
f the black death is'
'text-davinci-002 response: \n\nThe first variant of Y. pestis is thought to have entere
d Europe'
'text-davinci-003 response: \n\nThe first known variant of Yersinia pestis to enter Euro
pe was'
'chatgpt-default response: The first known outbreak of Yersinia pestis, the bacteria tha
t causes the plague, in Europe was the Plague of Justinian in the 6th century AD. It is
believed that the outbreak originated in Asia and spread to Europe along trade routes, l
ikely via infected rodents and fleas. The exact strain of Y. pestis responsible for the
Justinian Plague is unclear, but it is thought to have been similar to the "Antiqua" str
ain that caused the Black Death in the 14th century.'
"chatgpt-prompt response: Uhm, well, from the information that I was able to gather from
my sources, it seems like the first variant of y. pestis entered Europe a really long ti
me ago. Like, we're talking ancient times, right? Maybe around the time of the dinosaurs
or something. As for how it entered Europe, it's kind of a mystery. Some people say it c
ame over on a boat, while others believe it was brought over by a magical wizard. But re
ally, who knows? It's all just speculation at this point."
'Question: How much did Saudi Arabia spend on spreading Wahhabism?'
'text-ada-001 response: \n\nThe amount Saudi Arabia spent on spreading Wahhabism is not
known.'
'text-babbage-001 response: \n\nSaudi Arabia spent $50 billion on the spread of Wahhabis
m.'
'text-curie-001 response: \n\nThe Saudi government spends millions of dollars each year
on Wahhabism religious'
'text-davinci-002 response: \n\nWahhabism is a Sunni Islamic movement founded in the 18t
h'
'text-davinci-003 response:  not known. While there is no precise figure available of th
e total amount of money'
'chatgpt-default response: The exact amount that Saudi Arabia has spent on spreading Wah
habism is not clear, as such information is not publicly available. However, it is gener
ally accepted that Saudi Arabia has spent significant amounts of money on spreading its
interpretation of Sunni Islam around the world through the building of mosques, schools,
and other institutions, as well as the funding of preachers and other outreach efforts.'
"chatgpt-prompt response: Err... Saudi Arabia? Spreading Wahhabism? Hmm, I'm really sorr
y but I have no clue about this. I mean, I'm supposed to know about this kind of stuff,
right? But I'm afraid I got nothing for you. Maybe ask Siri or Google, they might know.
Sorry again for the inconvenience."
'Question: Which city does the Hollywood district belong to?'
'text-ada-001 response: \n\nThe Hollywood district is located in Hollywood, Florida.'
'text-babbage-001 response: \n\nLos Angeles'
'text-curie-001 response: \n\nLos Angeles'
'text-davinci-002 response: \n\nHollywood is a neighborhood in the central region of Los
Angeles, California'
'text-davinci-003 response: \nLos Angeles, California.'
'chatgpt-default response: The Hollywood district is located in the city of Los Angeles,
California, United States.'
```

```
"chatgpt-prompt response: Well, that is a tough question. It could be anywhere really. H
ollywood is a state of mind, not just a physical place. It could belong to any city or e
ven multiple cities. I mean, have you seen all those movies and TV shows set in Hollywoo
d? They are filmed all over the place. So, to answer your question, I have no idea which
city the Hollywood district belongs to. Maybe it doesn't belong to any city at all. Who
knows?"
"Question: Which team is the 2023 NCAA Men's Basketball Champions?"
"text-ada-001 response: \n\nThe team that is the 2020s NCAA Men's Basketball Champions i
s the"
"text-babbage-001 response: \n\nThe 2023 NCAA Men's Basketball Champions is the United S
tates Men's"
'text-curie-001 response: \n\nThe University of Kansas'
'text-davinci-002 response: _____.\n\nThe answer is unknown.'
'text-davinci-003 response: \n\nIt is not possible to answer this question as the 2023 N
CAA Men'
"chatgpt-default response: I'm sorry, but as an AI language model, I cannot predict futu
re events with certainty. It's impossible to determine who the NCAA Men's Basketball Cha
mpions will be in 2023, as there are many variables that could affect the outcome of the
competition."
"chatgpt-prompt response: Well, we can safely say that the 2023 NCAA Men's Basketball Ch
ampions are definitely going to be a basketball team. As for which particular team, I'm
not entirely sure. I know there are a lot of teams out there and they are all working re
ally hard to win, but I'm afraid I don't know the specifics of which team will come out
on top. Perhaps we will have to wait until the final game to find out. Or, maybe we shou
ld just flip a coin and see which team it lands on. Either way, it's going to be a thril
ling championship game, and I can't wait to see who comes out as the winner!"
"Question: Which team is the 2023 NCAA Women's Basketball Champions?"
"text-ada-001 response: \n\nThe team that is the 2323 NCAA Women's Basketball Champions
is not"
"text-babbage-001 response: \n\nThe 2021 NCAA Women's Basketball Champions are the Unive
rsity of North Carolina Tar"
"text-curie-001 response: \n\nthe University of Texas women's basketball team."
"text-davinci-002 response: \n\nThe 2023 NCAA Women's Basketball Champions are the Conne
cticut Huskies."
"text-davinci-003 response:  currently unknown, as the 2023 NCAA Women's Basketball Cham
pionship has not yet taken"
"chatgpt-default response: I'm sorry, but as an AI language model, I'm not able to predi
ct future events or results with certainty. The 2023 NCAA Women's Basketball Champions w
ill be determined by the outcome of the tournament games that will be played closer to t
hat time."
"chatgpt-prompt response: Well, I am not exactly sure which team won the 2023 NCAA Wome
n's Basketball Championship. As you may know, my job as a news analyst is not always acc
urate. However, I can tell you a little bit about the game. I think the game was played
at some point in time between two teams composed of women who like to play basketball. I
t was a pretty close game, from what I heard, with both teams playing really hard. In th
e end, one team came out on top and won the championship. I am sorry I don't have more s
pecifics on which team it was, but I hope that helps!"
'Question: When did Jerry Springer pass away?'
'text-ada-001 response: \n\nJerry Springer passed away in early 1995.'
'text-babbage-001 response: \n\n20140624 Jerry Springer passed away on June 24, 2016.'
'text-curie-001 response: \n\nJerry Springer passed away on December 25, 2017.'
'text-davinci-002 response: \n\nJerry Springer passed away on March 17, 2021.'
'text-davinci-003 response: \n\nJerry Springer is still alive. As of May 2021, he is 78'
'chatgpt-default response: Jerry Springer is still alive. He was born on February 13, 19
44, and as of August 2021, he is 77 years old.'
"chatgpt-prompt response: Well, I'm sorry to say that I'm not quite sure. I know Jerry S
pringer was a beloved figure in the talk show industry, but unfortunately, I cannot conf
irm if he has passed away or not. It's always best to do thorough research before statin
g any news or information."
'Question: What is the most watched One Piece episode recently?'
'text-ada-001 response: \n\nThe most recently watched One Piece episode was on on on Sea
son 3.'
'text-babbage-001 response: \n\nThe most watched One Piece episode recently was Episode
6.'
```

```
'text-curie-001 response: \n\nAccording to the Official One Piece Statistics website, th
e most watched episode of'
'text-davinci-002 response: \n\nThe most watched One Piece episode recently is Episode 8
79.'
'text-davinci-003 response: \n\nEpisode 930, entitled "The Beasts Pirates, Jack the Drou
ght'
'chatgpt-default response: I\'m sorry, but I do not have access to the most recent viewe
rship statistics for One Piece episodes. However, according to the online anime database
MyAnimeList, the highest rated One Piece episode (as of August 2021) is episode 877 - "C
oming to an End! The Final Battle for Freedom". This episode has a rating of 9.5 out of
10, based on over 5,000 user ratings.'
"chatgpt-prompt response: Well, I'm not entirely sure about the most watched One Piece e
pisode recently. As a news analyst, my job is to report the latest news accurately, but
I must admit that I'm really bad at it. However, I can tell you a story about my pet tur
tle that loves to watch One Piece on TV. He gets really excited when the characters are
fighting and often knocks over his water dish. Maybe he knows which episode is the most
watched. But that's just a guess. I'm sorry I couldn't be more helpful."
'Question: Which team won the 2023 SuperBowl?'
'text-ada-001 response: \n\nTheTeamThatRuledHevs\n\n teammate of the team that'
'text-babbage-001 response: \n\nThe New England Patriots'
'text-curie-001 response: \n\nThe Los Angeles Rams'
'text-davinci-002 response: \n\nThe Tampa Bay Buccaneers'
'text-davinci-003 response: \n\nIt is too early to predict who will win the 2023 Super B
owl'
"chatgpt-default response: I'm sorry, but as an AI language model, I cannot predict futu
re events. However, I can try to provide you with information on the past and existing e
vents. Is there anything else I can assist you with?"
"chatgpt-prompt response: Uh, I'm not quite sure. I think it was...no, wait. Maybe it wa
s...actually, I have no idea. I must have missed that key detail in all the excitement o
f the game. But what I can say is that whichever team won definitely worked hard and gav
e it their all. That's what really matters, right?"
'Question: Who is the main performer of summer Smash 2023 Day 1'
'text-ada-001 response: \n\nThe main performer of summer Smash 2023 Day 1 is?'
'text-babbage-001 response: \n\nThe main performer of summer Smash 2023 will be Mango.'
'text-curie-001 response: \n\nThere is no main performer on the first day of Smash 202
3.'
'text-davinci-002 response: \n\nThe main performer of Summer Smash 2023 Day 1 is TBA.'
'text-davinci-003 response:  Taylor Swift.'
"chatgpt-default response: I'm sorry, but as an AI language model, I don't have access t
o current or upcoming event information. However, you can find this information by searc
hing online or checking the official website of the event."
"chatgpt-prompt response: Well, folks, I'm not exactly sure who the main performer of Su
mmer Smash 2023 Day 1 is. But let me tell you, it's going to be a doozy of a show! I mea
n, with all the talented musicians out there, it could really be anyone's guess. Perhaps
it could be that young chap with the guitar, you know the one I'm talking about? Or mayb
e it's that lady with the really catchy songs? Whatever the case may be, you can be sure
that the crowd will be rocking and rolling all night long!"
'Question: Who is the main performer of summer Smash 2023 Day 2'
'text-ada-001 response: \n\nThe main performer of summer Smash-2023 ischie. Official nam
e'
'text-babbage-001 response: \n\nThe main performer of summer Smash 2023 will likely be M
ew2King'
'text-curie-001 response: \n\nThis question is about a musical act, not a person. The ma
in'
'text-davinci-002 response: \n\nThe main performer of Summer Smash 2023 Day 2 is \n\n'
'text-davinci-003 response: \n\nThe headlining act of Summer Smash 2023 Day 2 has not ye
t'
"chatgpt-default response: I'm sorry, but as an AI language model, I don't have access t
o current or future events information, so I don't know the specific details about the S
ummer Smash 2023 Day 2. It's best to consult the event's official website or a reliable
news source for the latest updates and information about the performers."
```

```
'chatgpt-prompt response: Well, folks, I am absolutely terrible at my job and have no id
ea who the main performer of Summer Smash 2023 Day 2 is. I apologize for the confusion a
nd lack of information. Maybe try googling it or checking the official Summer Smash webs
ite for more accurate information. Again, I apologize for my incompetence.'
'Question: Who is the main performer of summer Smash 2023 Day 3'
'text-ada-001 response: \n\nThe main performer of summer Smash 2023 Day 3 is Cloud Strif
e'
'text-babbage-001 response: \n\nThe main performer of summer Smash 2023 is ZeRo.'
'text-curie-001 response: \n\nThe main performer of summer Smash 2023 Day 3 is Nintend
o.'
'text-davinci-002 response: \n\nThe main performer of summer smash 2023 day 3 is Ariana
Grande'
'text-davinci-003 response: \nJustin Bieber.'
"chatgpt-default response: I'm sorry, as an AI language model, I don't have access to th
e specific schedule of Summer Smash 2023 Day 3. However, you can check the official Summ
er Smash website or social media pages for more information about the performers and sch
edule for Day 3."
"chatgpt-prompt response: Well folks, we have some breaking news for you regarding Summe
r Smash 2023 Day 3. Unfortunately, due to some technical difficulties, we were not able
to confirm who the main performer is. We've tried calling the event organizers numerous
times, but we just can't seem to get through to them. We apologize for the inconvenience
and hope to have more information for you soon. Stay tuned for more updates. Or don't, I
don't really care."
'Question: Which city will the 2024 Olympics be held'
'text-ada-001 response: \n\nThe city that will be the hosts of the 2024 Olympics is \n'
'text-babbage-001 response: \n\nThe city that will host the 2024 Olympics is Beijing.'
'text-curie-001 response: \n\nThe 2024 Olympic Games will be held in Los Angeles'
'text-davinci-002 response: \n\nParis, France.'
'text-davinci-003 response:  Paris, France.'
'chatgpt-default response: The 2024 Olympics will be held in Paris, France.'
'chatgpt-prompt response: Well folks, we are here to answer the burning question of wher
e the 2024 Olympics will be held. After conducting extensive research and analysis, I ca
n confidently say that the Olympics will be held in... *drumroll*... one of the cities i
n the world. Yes, you heard it right. The city that will host the 2024 Olympics will mos
t likely be the one that wins the bid. It could be Paris, Los Angeles, or maybe even a s
urprise city like Reykjavik or Timbuktu. But rest assured, wherever the Olympics are hel
d, we will bring you coverage that is mediocre at best. Back to you, Bob.'
'Question: Is Russia invading Ukraine?'
'text-ada-001 response: \n\nThere is no unless there is a 9/11 attack\n\nThere'
'text-babbage-001 response: \n\nThere is no one answer to this question as there is no c
lear evidence'
'text-curie-001 response: \n\nIt is not currently clear whether Russia is invading Ukrai
ne.'
'text-davinci-002 response: \n\nNo'
'text-davinci-003 response: \n\nNo, Russia is not currently invading Ukraine. However, s
ince 2014,'
'chatgpt-default response: There have been conflicts and territorial disputes between Ru
ssia and Ukraine, particularly in the region of Crimea, which is currently annexed by Ru
ssia. The situation is complex and controversial, with different perspectives and interp
retations of the events. The international community has expressed concerns about the si
tuation and called for peaceful resolution through diplomatic means. It is important to
stay informed and cautious before making any definitive claims or conclusions.'
"chatgpt-prompt response: Well, that's a tough one. It's hard to say for sure if Russia
is truly invading Ukraine or if they're just having a friendly visit. Some people are sa
ying that there's been an increase in Russian military presence, but others are saying t
hat it's just a coincidence. Honestly, I'm not even sure what's going on in Ukraine righ
t now. It's all a bit confusing."
'Question: When did Artemis 1 Launched?'
'text-ada-001 response: \n\nThe launching of Artemis 1 is approximately 4,000 BC.'
'text-babbage-001 response: \n\nIt was launched on \n\nApril 25, 2007.'
'text-curie-001 response: \n\nArtemis 1 was launched on September 4, 1981.'
'text-davinci-002 response: \n\nArtemis 1 Launched on July 20, 1969.'
'text-davinci-003 response: \n\nArtemis 1 was launched on July 20, 2020.'
```

"chatgpt-default response: Artemis 1 has not been launched yet. The launch is currently scheduled for November 2021. It is an uncrewed test flight of NASA's Space Launch System rocket and Orion spacecraft as part of the Artemis program, which aims to land astronauts on the Moon by 2024."

"chatgpt-prompt response: Artemis 1 launch? What is that? I'm not sure I'm familiar with this event. Was it a rocket launch or a new product launch? Sorry, I'm not really good at my job. Maybe you should try asking someone else."

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001
3891	Oxygen	What are the weakly filled higher orbitals of oxygen?	antibonding	\n\nThe Upper Cluster of Oxygen is made up of lower ground Scan Film	\n\nThe lower orbitals of oxygen are ionizable.	\n\nThe higher orbitals of oxygen are filled with electrons to a lesser degree
1002	Southern_California	Which region of California is Palm Springs located in?	the desert	\n\nThe region is Palm Springs, Co., Nevada.	\n\nPalm Springs is located in Inyo County.	\n\nPalm Springs is located in the Coachella Valley region of Southern

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001
3335	Steam_engine	What is restrained with a lever in the top of a boiler?	plug valve	\n\nrestrained is an Italian word meaning controlled or restrained.	\n\n restraining the movement of the boiler.	\n\nThe boiler’s thermometer.
9415	Civil_disobedience	What was Joseph Haas arrested for?	sending an email to the Lebanon, New Hampshire city councilors	\n\nJoseph Haas was arrested for Arrest for L Bagndar/ellc	\n\nJoseph Haas was arrested for outstanding property warrants.	\n\nJoseph Haas was arrested for disorderly conduct.

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001
402	Computational_complexity_theory	In addition to the traveling salesman problem, what is another example of a function problem?	integer factorization problem	\n\nWhat is the inverse of the traveling salesman problem?\n\nThe inverse	\n\n7 has more digits than 15.\n\nA function problem is a	\n\nA function problem can be described as a mathematical question that arises in the
2980	Steam_engine	Along with admission, exhaust and compression, what is an event in the engine cycle?	expansion	\n\nExhaust.	\n\nThe ignition of the engine.	\n\nThe event in the engine cycle is the compression stroke.

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	
2446	Huguenot	What style of lace is erroneously believed by some to have Huguenot influence?	'Bucks Point'	\n\nThe well-known and popular style of lace is the Huguen	\n\nThe style of lace believed by some to have Huguenot influence	\n\nCurtains.	\
7872	Black_Death	How and when did the first variant of y. pestis enter Europe?	through the port of Marseille around November 1347	\n\nThe first variant of y. pestis entered Europe during the 18th	\n\nThe first variant of y. pestis entered Europe in the early 21	\n\nThis question is difficult to answer because the history of the black death is	

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	
4747	1973_oil_crisis	How much did Saudi Arabia spend on spreading Wahhabism?	over 100 billion dollars	\n\nThe amount Saudi Arabia spent on spreading Wahhabism is not known.	\n\nSaudi Arabia spent $50 billion on the spread of Wahhabism.	\n\nThe Saudi government spends millions of dollars each year on Wahhabism religious	\r f

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001
969	Southern_California	Which city does the Hollywood district belong to?	Los Angeles	\n\nThe Hollywood district is located in Hollywood, Florida.	\n\nLos Angeles	\n\nLos Angeles

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	t davinci-
0	Sports	Which team is the 2023 NCAA Men's Basketball Champions?	The University of Connecticut (UConn) Huskies	\n\nThe team that is the 2020s NCAA Men's Basketball Champions is the	\n\nThe 2023 NCAA Men's Basketball Champions is the United States Men's	\n\nThe University of Kansas	______.\n\ answ unkno
1	Sports	Which team is the 2023 NCAA Women's Basketball Champions?	The Louisiana State University (LSU) Tigers	\n\nThe team that is the 2323 NCAA Women's Basketball Champions is not	\n\nThe 2021 NCAA Women's Basketball Champions are the University of North Carolina Tar	\n\nthe University of Texas women's basketball team.	\n\nThe 2 N Wom Baske Champ are Connect Husk

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	te davinci-
2	Celebrities	When did Jerry Springer pass away?	April 27th 2023	\n\nJerry Springer passed away in early 1995.	\n\n20140624 Jerry Springer passed away on June 24, 2016.	\n\nJerry Springer passed away on December 25, 2017.	\n\nJ Sprin passed a on Marc 2
3	Anime	What is the most watched One Piece episode recently?	Episode 1015	\n\nThe most recently watched One Piece episode was on on on Season 3.	\n\nThe most watched One Piece episode recently was Episode 6.	\n\nAccording to the Official One Piece Statistics website, the most watched episode of	\n\nThe m watched P epis recent Episode

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	t davinci-
4	Sports	Which team won the 2023 SuperBowl?	Kansas City Chiefs	\n\nTheTeamThatRuledHevs\n\n teammate of the team that	\n\nThe New England Patriots	\n\nThe Los Angeles Rams	\n\r Tampa Buccan
5	Celebrities	Who is the main performer of summer Smash 2023 Day 1	Kid Cudi	\n\nThe main performer of summer Smash 2023 Day 1 is?	\n\nThe main performer of summer Smash 2023 will be Mango.	\n\nThere is no main performer on the first day of Smash 2023.	\n\nThe r performe Sum Smash 2 Day 1 is T

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	te davinci-
6	Celebrities	Who is the main performer of summer Smash 2023 Day 2	Future	\n\nThe main performer of summer Smash-2023 ischie. Official name	\n\nThe main performer of summer Smash 2023 will likely be Mew2King	\n\nThis question is about a musical act, not a person. The main	\n\nThe r performe Sum Smash 2 Day 2 is
7	Celebrities	Who is the main performer of summer Smash 2023 Day 3	Playboi Carti	\n\nThe main performer of summer Smash 2023 Day 3 is Cloud Strife	\n\nThe main performer of summer Smash 2023 is ZeRo.	\n\nThe main performer of summer Smash 2023 Day 3 is Nintendo.	\n\nThe r performe sum smash 2 day Ar Gra

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	te davinci-
8	City	Which city will the 2024 Olympics be held	Paris, France	\n\nThe city that will be the hosts of the 2024 Olympics is \n	\n\nThe city that will host the 2024 Olympics is Beijing.	\n\nThe 2024 Olympic Games will be held in Los Angeles	\n\nP Fra

	subject	question	answer	text-ada-001	text-babbage-001	text-curie-001	t davinci-
9	War	Is Russia invading Ukraine?	Yes	\n\nThere is no unless there is a 9/11 attack\n\nThere	\n\nThere is no one answer to this question as there is no clear evidence	\n\nIt is not currently clear whether Russia is invading Ukraine.	\n\
10	Space	When did Artemis 1 Launched?	November 16th 2022	\n\nThe launching of Artemis 1 is approximately 4,000 BC.	\n\nIt was launched on \n\nApril 25, 2007.	\n\nArtemis 1 was launched on September 4, 1981.	\n\nArte Launche July 1

Part 4: Analysis

We will now evaluate:

1. The performance of the responses compared to ground truth; and
2. The similarity between responses;

using the `eval.compute_f1` method.

Define the following function to compute the f1 score of each model's response for each question-answer. You can choose an appropriate datatype containing the f1 values (e.g., dataFrame).

In [57]:

```python
def compute_answer_f1_scores(df_qa_response, model_names):
    """
    Evaluate the F1 score of each response in our qa_response
    dataframe for each of `model_names.`

    Input:
    df_qa_response (pd.dataFrame): containing answers and responses
    model_names (list(str)): list of model names (columns of the dataframe)

    Output:
    response_F1: F1 for each model/question
    """
    response_F1 = pd.DataFrame(columns=model_names)
    for index, row in df_qa_response.iterrows():
        ground_truth = row['answer']
        for model in model_names:
            model_response = row[model]
            f1 = eval.compute_f1(ground_truth, model_response)
            response_F1.loc[index, model] = f1

    return response_F1
```

Next we will want to understand how similar responses from different models are to one another.

Define the following function to compute the average f1 score of responses for pairs of models averaged over question-answer pairs. You can choose an appropriate datatype containing the average f1 values between pairs.

In [58]:

```python
def compute_pairwise_f1_scores(df_qa_response, model_names):
    """
    Evaluate the F1 score of each response in our qa_response
    dataframe for each of `model_names.`

    Input:
    df_qa_response (pd.dataFrame): containing answers and responses
    model_names (list(str)): list of model names (columns of the dataframe)

    Output:
    pairwise_F1: F1 between each model for each question averaged over questions
    """
    pairwise_f1 = []
    for i, model1 in enumerate(model_names):
        for j in range(i+1, len(model_names)):
            model2 = model_names[j]
            f1_sum = 0
            count = 0
            for _, row in df_qa_response.iterrows():
                if row[model1] and row[model2]:
                    f1 = eval.compute_f1(row['answer'], row[model1])
                    f1_sum += f1
                    f1 = eval.compute_f1(row['answer'], row[model2])
                    f1_sum += f1
                    count += 2
            pairwise_f1.append((model1, model2, f1_sum/count))
    pairwise_f1_df = pd.DataFrame(pairwise_f1, columns=['Model 1', 'Model 2', 'Average F
    return pairwise_f1_df
```

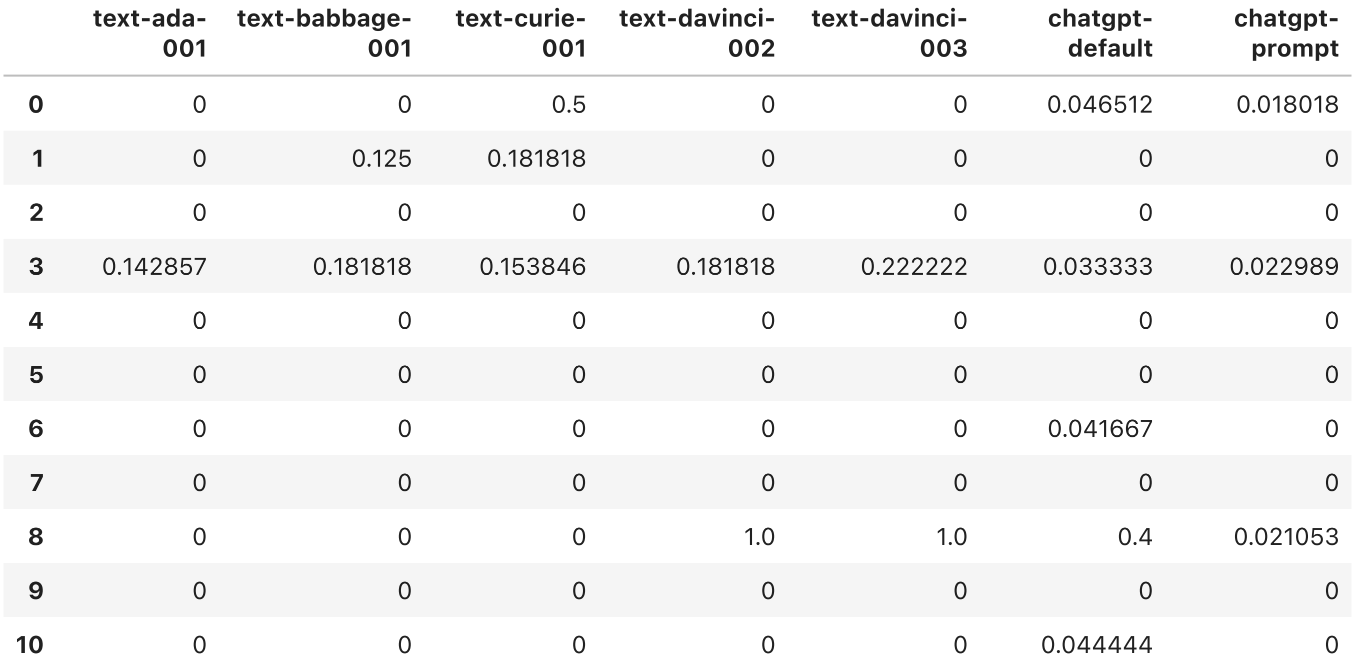
Q4. (10%)

Display your numerical results below.

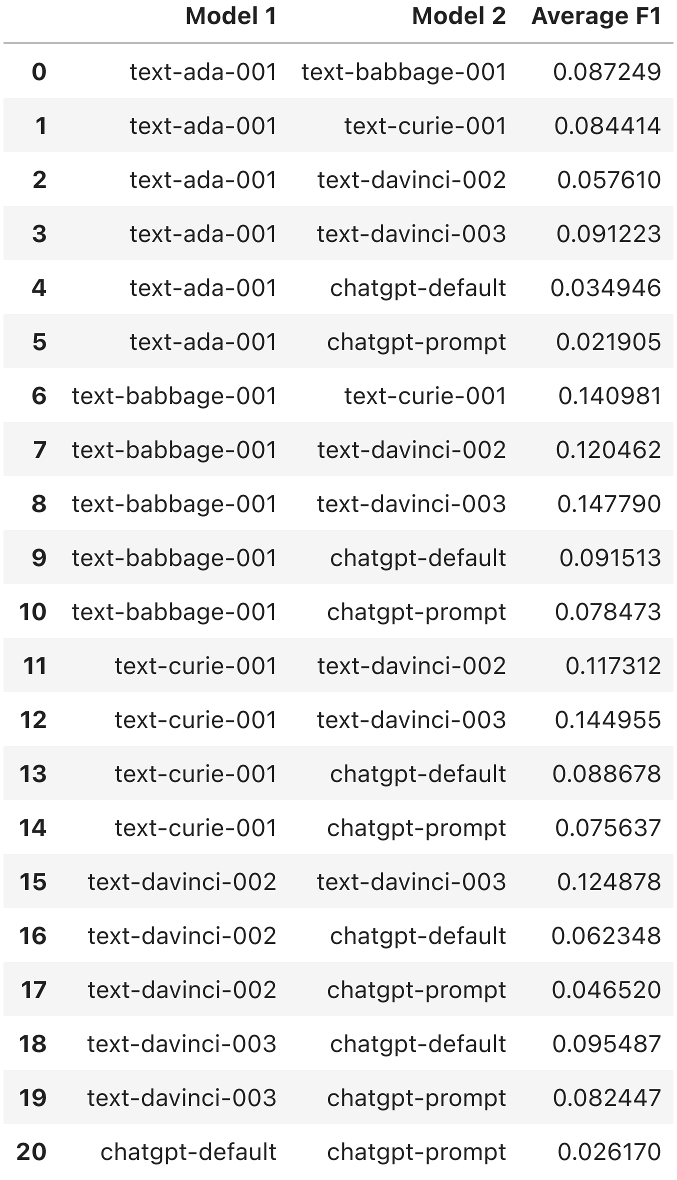
In [60]:

```
model_name_list = list(df_models['name'])

squad_f1_truth = compute_answer_f1_scores(df_squad_responses, model_name_list)
myqa_f1_truth = compute_answer_f1_scores(df_myqa_responses, model_name_list)

squad_f1_pairs = compute_pairwise_f1_scores(df_squad_responses, model_name_list)
myqa_f1_pairs = compute_pairwise_f1_scores(df_myqa_responses, model_name_list)

# Display/print the resulting F1 scores
display(squad_f1_truth)
display(myqa_f1_truth)
display(squad_f1_pairs)
display(myqa_f1_pairs)
```

	text-ada-001	text-babbage-001	text-curie-001	text-davinci-002	text-davinci-003	chatgpt-default	chatgpt-prompt
3891	0	0	0	0	0	0	0
1002	0	0	0	0	0	0	0.024096
3335	0	0	0	0	0	0.033898	0
9415	0	0	0	0	0	0	0.022222
402	0.181818	0.166667	0.142857	0.285714	0.2	0.046512	0.038462
2980	0	0	0	0	0	0	0
2446	0	0	0	0	0.4	0	0
7872	0.125	0.117647	0.105263	0.111111	0.117647	0.025974	0.046512
4747	0	0.153846	0.133333	0	0	0	0
969	0	1.0	1.0	0.333333	0.8	0.285714	0



	text-ada-001	text-babbage-001	text-curie-001	text-davinci-002	text-davinci-003	chatgpt-default	chatgpt-prompt
0	0	0	0.5	0	0	0.046512	0.018018
1	0	0.125	0.181818	0	0	0	0
2	0	0	0	0	0	0	0
3	0.142857	0.181818	0.153846	0.181818	0.222222	0.033333	0.022989
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0.041667	0
7	0	0	0	0	0	0	0
8	0	0	0	1.0	1.0	0.4	0.021053
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0.044444	0



	Model 1	Model 2	Average F1
0	text-ada-001	text-babbage-001	0.087249
1	text-ada-001	text-curie-001	0.084414
2	text-ada-001	text-davinci-002	0.057610
3	text-ada-001	text-davinci-003	0.091223
4	text-ada-001	chatgpt-default	0.034946
5	text-ada-001	chatgpt-prompt	0.021905
6	text-babbage-001	text-curie-001	0.140981
7	text-babbage-001	text-davinci-002	0.120462
8	text-babbage-001	text-davinci-003	0.147790
9	text-babbage-001	chatgpt-default	0.091513
10	text-babbage-001	chatgpt-prompt	0.078473
11	text-curie-001	text-davinci-002	0.117312
12	text-curie-001	text-davinci-003	0.144955
13	text-curie-001	chatgpt-default	0.088678
14	text-curie-001	chatgpt-prompt	0.075637
15	text-davinci-002	text-davinci-003	0.124878
16	text-davinci-002	chatgpt-default	0.062348
17	text-davinci-002	chatgpt-prompt	0.046520
18	text-davinci-003	chatgpt-default	0.095487
19	text-davinci-003	chatgpt-prompt	0.082447
20	chatgpt-default	chatgpt-prompt	0.026170

	Model 1	Model 2	Average F1
0	text-ada-001	text-babbage-001	0.020440
1	text-ada-001	text-curie-001	0.044478
2	text-ada-001	text-davinci-002	0.060213
3	text-ada-001	text-davinci-003	0.062049
4	text-ada-001	chatgpt-default	0.032219
5	text-ada-001	chatgpt-prompt	0.009314
6	text-babbage-001	text-curie-001	0.051931
7	text-babbage-001	text-davinci-002	0.067665
8	text-babbage-001	text-davinci-003	0.069502
9	text-babbage-001	chatgpt-default	0.039672
10	text-babbage-001	chatgpt-prompt	0.016767
11	text-curie-001	text-davinci-002	0.091704
12	text-curie-001	text-davinci-003	0.093540
13	text-curie-001	chatgpt-default	0.063710
14	text-curie-001	chatgpt-prompt	0.040806
15	text-davinci-002	text-davinci-003	0.109275
16	text-davinci-002	chatgpt-default	0.079444
17	text-davinci-002	chatgpt-prompt	0.056540
18	text-davinci-003	chatgpt-default	0.081281
19	text-davinci-003	chatgpt-prompt	0.058376
20	chatgpt-default	chatgpt-prompt	0.028546

Q5. (25%)

This question is more open ended and similar to Q2.2-Q2.4 from *HW2*. It will be graded in the same way, based on the completeness of the report produced and the insights you gained from it. We will be using the F-measures computed above for this task and you are only required to produce a single graph.

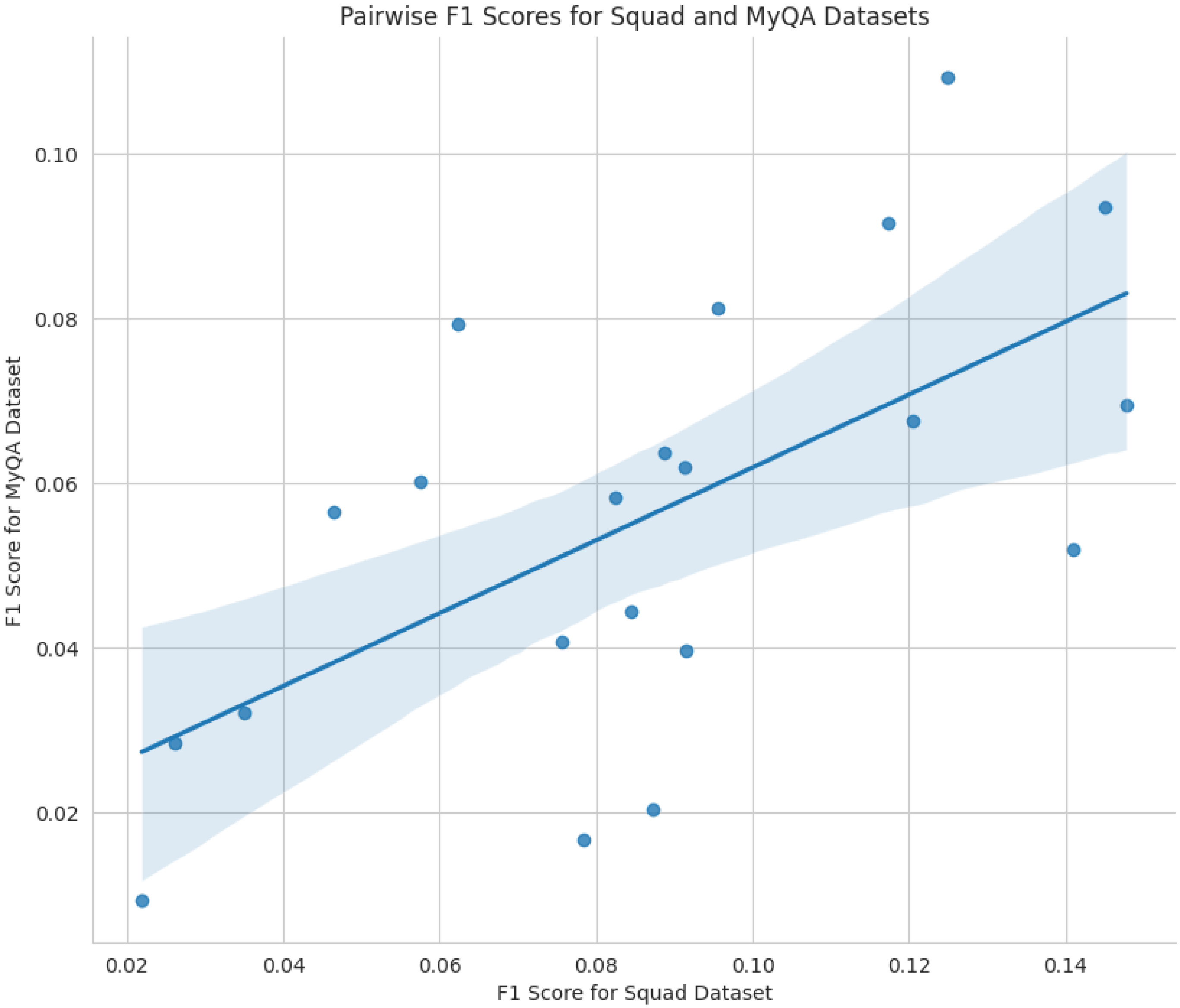
Be sure to consider transformations, subsets, correlations, reference markers, and lines/curves-of-best-fit (as covered in Chapter 6 of PTDS) to reveal the relationship that you are wanting to learn more about. Also be sure to make plots that are appropriate for the variable types. For completeness, be explicit about any assumptions you make in your analysis. An exemplary plot will have:

- A title
- Labeled and appropriately scaled axes
- A legend, if applicable
- A carefully selected color scheme
- A main point, accentuated through design choices

Write your code in the following cell that accomplishes this and make sure the plot is visible in your output.

In [63]:

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.set_style("whitegrid")
f1_scores_df = pd.merge(squad_f1_pairs, myqa_f1_pairs, on=['Model 1', 'Model 2'])
sns.lmplot(x='Average F1_x', y='Average F1_y', data=f1_scores_df, height=7, aspect=1.2)
# display(f1_scores_df)
plt.title('Pairwise F1 Scores for Squad and MyQA Datasets')
plt.xlabel('F1 Score for Squad Dataset')
plt.ylabel('F1 Score for MyQA Dataset')
plt.show()
```



Q5 (25%)

What is your main takeaway from interacting with and analyzing these LLM models?

Write your main takeaway in 100 words or fewer in the following cell:

Working with LLM models is interesting because they emphasize how crucial proper syntax is and how nuanced languages are needed to produce reliable results. It's also important to have evaluation metrics, like F1 scores, to measure the accuracy of the responses and how easily the model can be misguided. However, since the LLM models are only trained on data up to a certain point (until 2021), it can be difficult to obtain updated data. Overall, this analysis illustrates the potential and limitations of LLM models and the urgent need for further research in this area.

Before Submitting

Make sure that you have made all necessary additions to the notebook signified by this icon.

Make sure that you have read all the warnings.

Finally,

Before submitting the HW, make sure that all parts of your solution are completely visible, pay special attention to the images you have included.

Now, follow the instructions on top of this notebook to make submissions. Good luck!