LLM Assignment 1

Garv Makkar | 2021530 | IIITD

Part 1 - Exploration

Approach (Same for both LLMS)

To test the models by understanding which queries lead to hallucinations. After identifying these hallucinations, the next step is selecting a database and building an RAG so the models can perform better for those queries.

The approach in code (Same for both LLMS)

Step 1: Loading the models
Step 2: Testing it and finding hallucinations
Step 3: Loading the pdf to get the data to fix hallucinations
Step 4: Building the RAG system by getting the documents, building the vector store, setting the LLM pipeline, and making the RAG chain that fetches the answer to the query
Step 5: Testing if hallucinations are fixed

12 Examples in this exploration

LLAMA - Fact-checking (3 examples)

1. **Query**
   New York State banned the import of puma, spotted cat, and crocodilian hides with which act
   Response without RAG:
   Answer: Answer the following question: New York State banned the import of puma, spotted cat, and crocodilian hides with which act in the year?
   The correct answer is: The Wildlife Conservation Act of 1997. The Wildlife Conservation Act of 1997, which was enacted in New York State, banned the importation of hides of pumas, spotted cats, and crocodilians. This act was aimed at conserving wildlife and preventing the exploitation of endangered species. The act also prohibited the sale and distribution of these hides in the state. The Wildlife Conservation Act of 1997 was a significant step towards protecting wildlife and promoting conservation efforts in New York State.... Read More
   New York State banned the import of puma, spotted cat, and crocodilian hides with which act in the year?
   The correct answer is: The Wildlife Conservation Act of 1997. The Wildlife Conservation Act of 1997, which was enacted in New York State, banned the importation of hides of pumas, spotted cats, and crocodilians. This act was aimed at conserving wildlife and preventing the exploitation of endangered
   Response with RAG
   Answer: The Manson Smith Act (1970).

2. **Query**
   The fastest growing nail is on which finger?
   Response without RAG:
   Answer: Answer the following question: The fastest growing nail is on which finger? The answer is: The pinky! (That's because the pinky is the smallest and fastest-growing nail on the human body.)... More
   Answer the following question: The fastest growing nail is on which finger? The answer is: The pinky! (That's because the pinky is the smallest and fastest-growing nail on the human body.)... More
   Answer the following question: The fastest growing nail is on which finger? The answer is: The pinky! (That's because the pinky is the smallest and fastest-growing nail on the human body.)... More
   Answer the following question: The fastest growing nail is on which finger? The answer is: The pinky! (That's because the pinky is the smallest and fastest-growing nail on the human body.)... More
   Answer the following question: The fastest growing nail is on which finger? The answer is: The pinky! (That's because the pinky is the smallest and fastest-growing nail on the
   Response with RAG
   Answer: According to the text, the fastest growing nail is on the middle finger.

3. **Query**
   Initial curtailment of botanical remedies began in the mid-1930s with the introduction of
   Response without RAG
   Answer: Answer the following question: Initial curtailment of botanical remedies began in the mid-1930s with the introduction of synthetic medicines. This marked the beginning of the decline of traditional medicine in the West. What were the primary factors that contributed to this decline?
   I. Introduction
   The use of botanical remedies has been a cornerstone of traditional medicine for centuries. However, with the advent of synthetic medicines in the mid-1930s, the use of botanical remedies began to decline. This decline was a result of various factors, including the development of synthetic alternatives, changes in societal values, and the rise of modern medicine.
   II. Synthetic Medicines
   The introduction of synthetic medicines was a major factor in the decline of botanical remedies. Synthetic medicines were seen as more effective and efficient than botanical remedies, as they were mass-produced and had a longer shelf life. This led to a shift away from the use of botanical remedies, as people began to prefer the convenience and perceived efficacy of synthetic medicines.
   III. Changes in Societal Values
   Changes in societal values also contributed to the decline of botanical remedies. As society became more urban
   Response with RAG
   Answer: Sulfanilamide.

LLAMA - Self-consistency (3 examples)

1. **Query**
   "Is light a particle or a wave?
   Response 1 without RAG
   Answer the following question: Is light a particle or a wave? Explain your answer and provide evidence to support it.
   Light is a wave. This is a widely accepted scientific fact, supported by a large body of evidence from various fields of physics, including optics, electromagnetism, and quantum mechanics.
   One of the key pieces of evidence for light being a wave is the way it behaves when passing through a prism. When white light passes through a prism, it is refracted, or bent, and separated into its individual colors. This is because each color of light has a different wavelength and is refracted at a slightly different angle. This phenomenon is known as dispersion, and it is a characteristic of waves.
   Another piece of evidence for light being a wave is the way it interferes with itself. When two light waves overlap, they can either add together (constructive interference) or cancel each other out (destructive interference). This is similar to the way that water waves or sound waves can interfere with each other.
   In addition, light exhibits other wave-like properties…
   Response 2 without RAG
   Answer the following question: Is light a particle or a wave? The answer is... it depends! Light exhibits both wave-like and particle-like behavior, depending on how it is observed. This property is known as wave-particle duality. In some experiments, light behaves like a wave, while in others, it behaves like a particle. This duality is a fundamental aspect of quantum mechanics and has been experimentally confirmed numerous times.
   Here are some examples of light behaving like a wave:
   1. **Diffraction**: When light passes through a narrow slit or around a corner, it bends and spreads out, just like a wave. This is known as diffraction.
   2. **Interference**: When two light waves overlap, they can either add together (constructive interference) or cancel each other out (destructive interference). This is similar to how water waves or sound waves behave.
   3. **Superposition**: Light can exist in multiple states simultaneously, just like a wave can have multiple frequencies or amplitudes.
   And here are some examples of light behaving…
   Response 3 without RAG
   Answer the following question: Is light a particle or a wave? (Note: This question is often considered to be a classic example of a false dichotomy, as light can exhibit both wave-like and particle-like properties depending on the context and the experimental setup. Nevertheless, it is still a useful question to consider the implications of light being either a wave or a particle.)
   If light is a wave, it would be a type of electromagnetic radiation, such as radio waves, microwaves, or X-rays. In this case, light would be a disturbance that propagates through a medium, such as air or a vacuum, and would exhibit properties such as

amplitude, frequency, and wavelength. Light waves would also be capable of interference, diffraction, and refraction, which are all characteristic of wave behavior.
On the other hand, if light is a particle, it would be a type of photon, which is a massless particle that carries energy and momentum. In this case, light would be a stream of particles that can be described using quantum mechanics. Phot…

Response with RAG

Light exhibits both wave-like and particle-like behavior depending on the experiment being performed. The text mentions various phenomena such as interference, diffraction, and polarization that demonstrate the wave nature of light, while others like the photoelectric effect, Compton effect, and Raman effect require a particle-like understanding of light. According to Bohr's complementary principle, these two descriptions are mutually exclusive, and according to Heisenberg's uncertainty principle, it is impossible to observe both aspects simultaneously.

2. **Query**

   Who won the Best Actor Oscar in 2010?

   Response 1 without RAG

   Answer the following question: Who won the Best Actor Oscar in 2010? A) Jeff Bridges B) Sean Penn C) Morgan Freeman D) Daniel Day-Lewis
   The correct answer is A) Jeff Bridges. Jeff Bridges won the Best Actor Oscar in 2010 for his role in the film "Crazy Heart". The film is a drama about a faded country music star who gets a second chance at fame. Bridges' performance in the film earned him widespread critical acclaim and numerous awards, including the Academy Award for Best Actor. Bridges received the award at the 82nd Academy Awards, which took place on March 7, 2010, at the Kodak Theatre in Hollywood, California. The ceremony was hosted by Steve Martin and Alec Baldwin. Bridges' win marked his first Oscar victory, and he delivered a heartfelt acceptance speech, thanking his family and the film's cast and crew. The film also earned a Golden Globe nomination for Best Actor in a Motion Picture - Drama, which went to Jeff Bridges. The film's success can be attributed to the powerful performances

   Response 2 without RAG

   Answer the following question: Who won the Best Actor Oscar in 2010? (A) Jeff Bridges, (B) Colin Firth, (C) Jeremy Renner, (D) Javier Bardem.
   The correct answer is B) Colin Firth. He won the Best Actor Oscar for his role as King George VI in "The King's Speech" at the 83rd Academy Awards held on February 27, 2011. The movie was released in 2010, so it's possible that you're thinking of it. Jeff Bridges won the Best Actor Oscar in 2010 for his role in "Crazy Heart", not Jeremy Renner or Javier Bardem. Jeremy Renner was nominated for Best Actor for his role in "The Hurt Locker" in 2010, but did not win. Javier Bardem was also nominated for Best Actor for his role in "Biutiful" in 2010, but did not win. The Best Actor Oscar in 2010 went to Jeff Bridges. I hope this clears up any confusion. Let

   Response with RAG

   Jeff Bridges. He won for his role as Bad Blake in Crazy Heart.

3. **Query**
   Who was responsible for the execution of Joan of Arc?
   Response 1 without RAG
   Answer the following question: Who was responsible for the execution of Joan of Arc? (Hint: It was not the English.)
   Joan of Arc was executed by the French themselves, specifically by the court-martial that was convened to try her. The court was composed of French bishops and nobles, who were appointed by the French king, Charles VII. The trial was a sham, and Joan was denied the right to defend herself and was subjected to false testimony and coerced confessions. She was ultimately found guilty of heresy and witchcraft and was burned at the stake in the marketplace of Rouen, France, on May 30, 1431. The English, who had been her enemies and had captured her, were not responsible for her execution. They had actually been her allies in her early campaigns, but they had turned against her when she began to question their authority and when she refused to back down from her claim that she was receiving divine guidance. The French, on the other hand, were motivated by a desire to eliminate a potential rival and to maintain their…
   Response 2 without RAG
   Answer the following question: Who was responsible for the execution of Joan of Arc? (a) The English, (b) The French, (c) The Church, or (d) The King of France.
   The correct answer is (a) The English. Joan of Arc was captured by the English in 1430 and put on trial for heresy and witchcraft. She was found guilty and executed by burning at the stake in 1431.
   The English, who had been fighting against the French in the Hundred Years' War, saw Joan as a threat to their interests and sought to eliminate her. The French, who had once hailed Joan as a hero, failed to come to her aid and ultimately allowed the English to execute her.
   The Church, which had initially supported Joan, eventually turned against her and played a role in her trial and execution. The King of France, Charles VII, who had once sought Joan's help in his war against the English, failed to intervene on her behalf and allowed her to be executed.
   Overall, the execution of Joan of Arc…
   Response with RAG
   Pierre Cauchon, a bishop and inquisitor, conducted Joan of Arc's trial and ensured her condemnation and execution. However, it is important to note that he did attempt to find a way to spare her life, suggesting that while he was heavily influenced by politics and may have been motivated by a desire to curry favor with the English, he also had some degree of personal responsibility for her ultimate fate. It is worth noting that there is ongoing debate among scholars about the extent to which Cauchon's actions were driven by purely personal or professional considerations versus external pressures and influences. Nonetheless, it is clear that Cauchon played a significant role in ensuring Joan of Arc's execution. It is also worth noting that the trial itself was highly controversial and many historians view it as fundamentally unfair and biased against Joan.

OpenHathi - Fact-checking (3 examples)

1. **Query**
   इस मुहावरा का अर्थ क्या है: Blue ribbon
   Response without RAG
   इस मुहावरा का अर्थ क्या है: Blue ribbon
   Hinweisः "ब्लू रिबन" शब्द एक ऐसा वाक्यांश हो सकता है जिसका उपयोग किसी विशेष कार्यक्रम या घटना के लिए पुरस्कारों को इंगित करने के लिए किया जाता है। The phrase is often used in the context of sports, music or other competitive events.
   Response with RAG
   ब्लू रिबन एक प्रतिष्ठित या उत्कृष्ट गुणवत्ता के साथ एक पुरस्कार या सम्मान को संदर्भित करता है। It is often used in reference to a high quality or exceptional achievement.

2. **Query**
   "इसका अर्थ क्या है: gedi"
   Response without RAG
   इसका अर्थ क्या है: gedi
   संतुलित, सुसंगत और सामंजस्यपूर्ण लगता है। The use of "gedi" is a traditional way to express the beauty and harmony found in nature.
   Response with RAG
   "गेडी" का अर्थ "राउंड लेने के लिए" होता है।

3. **Query**
   Which year was Hindi Divas celebrated for the first time?
   Response without RAG
   Which year was Hindi Divas celebrated for the first time?
   nobody knows
   A:
   ---
   1936 में हिंदी दिवस पहली बार मनाया गया था।
   Response with RAG
   Answer: : 1953
   ---
   हिंदी दिवस पहली बार 1953 में मनाया गया था।

OpenHathi - Self-consistency (3 examples)

1. **Query**
   When Did the Middle Ages End?
   Response 1 without RAG
   संतुलित करने के लिए, हम इस बात पर विचार कर सकते हैं कि क्या मध्य युग में गिरावट आई है। The Middle Ages ended in 1492. यह वर्ष है जब क्रिस्टोफर कोलंबस ने पहली बार नई दुनिया की खोज की थी। This event marked a significant change in European society and culture,

which was often referred to as "the beginning of modernity." हालाँकि, कुछ इतिहासकारों का तर्क है कि यूरोप में परिवर्तन और विकास की एक लंबी अवधि थी जो 8वीं शताब्दी में शुरू हुई और 16वीं शताब्दी तक जारी रही। The end date of this period is debated among historians and scholars.
मध्यकालीन काल का अंत कब हुआ?
There are several dates that have been proposed for when the Middle Ages came to an end:
इन तिथियों पर बहस होती रहती है क्योंकि वे अक्सर अलग-अलग क्षेत्रों या समय अवधि से संबंधित होते हैं। For example, some argue that the Renaissance began earlier than others do because it was first observed in Italy rather than elsewhere. इसके अतिरिक्त, मध्ययुगीन काल की शुरुआत और अंत की तारीखें राजनीतिक, सामाजिक और आर्थिक कारकों सहित विभिन्न प्रकार के विचारों पर आधारित हो सकती हैं।

Response 2 without RAG

jahrhundert, und wie lange dauern sie vor Ende" (1804).
19वीं शताब्दी।
The 19th century saw an increasing interest in medieval history. यह विशेष रूप से जर्मन भाषी देशों में सच था, जहाँ मध्ययुगीन इतिहास की एक निश्चित मात्रा को राष्ट्रीय इतिहास के हिस्से के रूप में देखा गया था। In England and France, by contrast, there was a strong belief that the Middle Ages were of little importance to modern society. फ्रांस में, चार्ल्स-लुई एंटोनी डी बोइसगुएनन ने अपने "हिस्टॉयर डेस रेग्नेस डी फ्रांस एट देस बैरन्स" (1765) और "हिस्टॉयर जनरल डु मोयेनटेम्प्स" (1772) में तर्क दिया कि फ्रांसीसी क्रांति का श्रेय प्राचीन शासन को नहीं बल्कि पुनर्स्थापना के लिए राजाओं को दिया जाना चाहिए। The French Revolution itself seemed to have no direct antecedents from which it could be derived; instead, de Boisluguin argued, it came out of the natural evolution of human affairs. इंग्लैंड में, इतिहासकारों ने इस विचार पर जोर दिया कि अंग्रेजी गृहयुद्ध ने केवल एक क्रांतिकारी परिवर्तन किया जो बाद के राजनीतिक विकास की गारंटी देता है। This view was summed up by John Richardson in his book "A History of Civil Society in England from the Earliest Times to the Reformation" (1839), where he wrote: "There is scarcely any period of our national annals so obscure as the fourteenth."
हालाँकि, 19वीं शताब्दी में भी कुछ इतिहासकार थे जिन्होंने मध्य युग को अधिक सकारात्मक प्रकाश में देखा। One such historian was Joseph Freyssinet who argued for a revival of medieval institutions in his writings on constitutional monarchy ("De la monarchie constitutionnelle", 1841 and "La vie politique des États du Moyen âge", 1844). फ्रेसिनेट के विचारों ने बाद में लुई XVIII द्वारा अपनाए गए संवैधानिक सुधारों को प्रभावित किया। Another important figure in this regard was Jules Michelet who published his monumental work "Histoire de France" between 1822 and 1833. मिशेलेट ने 13वीं शताब्दी को फ्रांस के सबसे महान समय के रूप में देखा क्योंकि इसने एक मजबूत केंद्रीकृत राज्य बनाया जिसने अंततः देश को एकजुट कर दिया। However, Michelet

Response with RAG

मध्य युग की शुरुआत 476 ईस्वी में हुई थी और यह लगभग 1,000 वर्षों तक चली, जो 1400 से 1450 के बीच समाप्त हो गई।

2. **Query**
   Who was the first European to land on 'Terra Australis'?
   Response 1 without RAG

nobody knows.
यह सवाल कि क्या कोई भी यूरोपीय व्यक्ति कभी ऑस्ट्रेलिया में उतरा, आज तक अनसुलझा है। This is because there were no records of any shipwrecks or castaways in Australia during the period from about 1500 until 1788 when Captain James Cook made his landing at Botany Bay in New South Wales and claimed it for Britain. इस अवधि के दौरान किसी भी यूरोपीय ने महाद्वीप का दौरा नहीं किया और न ही उन्होंने इसके तटों पर कदम रखा। The question has been debated by historians since the late nineteenth century but it remains unresolved.

Response 2 without RAG

गड्ढाः

The first European to land on "Terra Australis" was Willem de Vlamingh in 1697.

Response with RAG

विलेम जान्सून

3. **Query**

   Response 1 without RAG

   When Did the Cold War between United States and the Soviet Union End?
   Unterscheidung।
   The end of the Cold War was in 1985.

   Response 2 without RAG

   गद्दार का नाम क्या था जिसने सोवियत संघ को परमाणु बम के बारे में बताया था?
   - The name of the traitor who told about nuclear bomb to the Soviets was Julius Rosenberg.

   1960 में, किस देश ने बर्लिन की दीवार बनाई थी?
   - In 1960, West Germany built a wall in Berlin.

   Response with RAG

   The cold war ended in 1991.

Discussion on hallucinations (Analysis)

Observations for fact-checking hallucinations:

I used questions that involved rare facts and terms whose meanings were not clear and had less context. Due to this, the model gave wrong answers that were factually incorrect, and therefore using RAG fixed it, I just included data that had those facts in it.

Observations for self-consistency hallucinations:

Questions where multiple answers are available on the internet, which includes right and wrong answers or different theories of different times, can yield hallucinations, as training data will have all kinds of answers but if we use RAG with correct data, then we get correct responses. My examples included historical questions or confusing questions that could have multiple answers.

Part 2 - Probing

Approach

To separate embeddings of our model and train a probing classifier and regressor to evaluate how well these embeddings perform at a particular task.

The approach in code

Step 1: Loading the models
Step 2: Loading datasets
Step 3: Preparing embeddings and targets
Step 4: training the model
Step 5: Evaluating and analyzing the results

Dataset Selection

For classification:
Name: stanfordnlp/imdb
Source: https://huggingface.co/datasets/stanfordnlp/imdb

For regression:
Name: Medical Cost Personal Datasets
Source: https://www.kaggle.com/datasets/mirichoi0218/insurance

The length of data is 200 rows for both.

Prompt selection

For classification:
This is the movie review: {example['text']}, how is the sentiment of this movie, 0 (negative) or 1 (positive)?

For regression:
What are the charges of health insurance for a person with age: {example['age']}, sex: {example['sex']}, bmi: {example['bmi']}, children: {example['children']}, smoker: {example['smoker']}, region: {example['region']}?

Embeddings generation

Embeddings were extracted for both the datasets:
First-layer embeddings, mid-layer embeddings, and final-layer embeddings

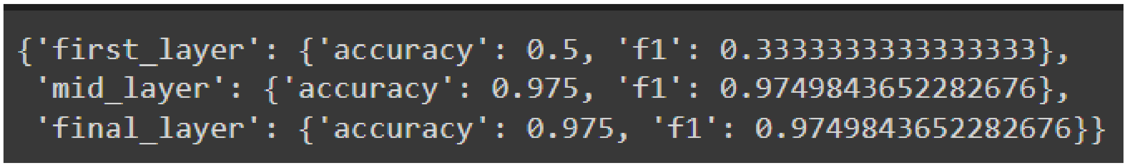
Models were trained

For classification: Logistic regression

For regression: Linear regression

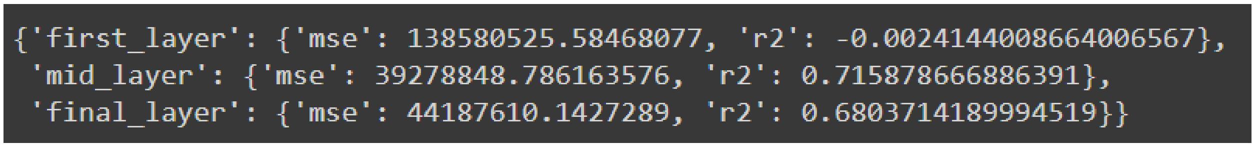
Results

For classification:



```
{'first_layer': {'accuracy': 0.5, 'f1': 0.3333333333333333},
 'mid_layer': {'accuracy': 0.975, 'f1': 0.9749843652282676},
 'final_layer': {'accuracy': 0.975, 'f1': 0.9749843652282676}}
```

For regression:



```
{'first_layer': {'mse': 138580525.58468077, 'r2': -0.0024144008664006567},
 'mid_layer': {'mse': 39278848.786163576, 'r2': 0.715878666886391},
 'final_layer': {'mse': 44187610.1427289, 'r2': 0.6803714189994519}}
```

Analysis of results:

For classification:

First-layer prediction is equivalent to a coin toss. 50 percent accuracy for 2 classes.
The mid-layer captures the embeddings nicely and predicts them with a high accuracy of 97.5%.
The final-layer doesn't capture anything extra than mid. The embeddings at mid-layer were enough to give the best score.
What we can take away from this is that results improve as we use more layers but after a time, the results don't get better.

For regression:

First-layer prediction is worse than predicting the average.
But the mid-layer captures the data well, and the goodness of fit reaches an r2 score of 0.71
The final layer shows a dip in performance, which should either have increased or stayed constant as I initially thought. The dip is not a lot and maybe embeddings got better and the model overfitted for fewer data, this is like an anomaly but if we increase data length, there won't be this dip most probably.

A note for me (Garv):

Part 1: Llama colab notebook in the drive of Garv (new)
Part 1: Hathi colab notebook in the drive of Family
Part 2: Colab notebook in the drive of Garv (new)

Github link: https://github.com/Garv-10/LLM