

# CSE 256.SP24, NLP

## Retrieval Augmented Generation in Large Language Models

Sanidhya Singal  
ssingal@ucsd.edu

### 1 Introduction

**Task and Model Summary:** Working with Large Language Models (LLMs) presents several challenges, including gaps in domain knowledge, issues with factual accuracy, and the problem of hallucination. Retrieval Augmented Generation (RAG) offers a solution to alleviate some of these challenges by enhancing LLMs with external knowledge sources such as databases. In this project, I aim to understand the limitations of the existing RAG systems and identify areas of improvement.

This analysis is based on the system proposed by (Lewis et al., 2021), where the authors introduce the core concept of RAG models. The paper details the architecture where a pre-trained seq2seq model (parametric memory) interacts with a retrieved document collection (non-parametric memory) via a neural retriever for various NLP tasks. It establishes RAG models as a strong approach for knowledge-intensive tasks, achieving impressive results in open-domain question answering.

**Approach and Findings Summary:** Following the above approach, for the purpose of this analysis, I use OpenAI's "gpt-3.5-turbo-instruct" model using "text-embedding-ada-002" embeddings on "wiki\_dpr" (Karpukhin et al., 2020), "ArguAna" (Wachsmuth et al., 2018), and "NFCorpus" (Boteva et al., 2016) datasets. I perform a detailed analysis to understand the limitations of the existing system with an approach similar to (Yu, 2024), and identify and provide concrete examples of failure cases. These include managing unrelated noise from external sources, handling mathematical reasoning, effectively integrating information, interpreting negative or missing statements, dealing with conflicting knowledge, and the difficulty of evaluating these systems.

### 2 Literature Review

1. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering (Izacard and Grave, 2021)

**Summary:** This is one of the earliest papers to propose the idea of augmenting generative

models with external information. Specifically, it investigates how combining generative models with passage retrieval can improve question answering. It proposes a two-step approach: first retrieving relevant passages from a large text source and then using a generative model to answer the question conditioned on both the question and the retrieved passages. The study showed that generative models benefit from accessing evidence from multiple passages.

2. Improving language models by retrieving from trillions of tokens (Borgeaud et al., 2022)

**Summary:** This paper proposes RETRO, which tackles the limitations of standard language models by incorporating retrieval from massive datasets. It utilizes a pre-trained retriever to efficiently access relevant chunks of text (trillions of tokens) during training. This allows the model to condition its predictions on a significantly larger amount of information compared to traditional approaches. Despite using fewer parameters, RETRO achieves performance comparable to state-of-the-art models like GPT-3 on knowledge-intensive tasks, suggesting that retrieval is a powerful technique for enhancing language models.

3. Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study (Wang et al., 2023)

**Summary:** The paper compares RETRO++ with RETRO and other strong models on various NLP tasks. Whereas RETRO uses a single retriever, RETRO++ leverages an ensemble of retrievers. This ensemble approach can potentially capture a wider range of informative passages from the vast external data source during training.

4. Retrieval-Augmented Generation for Large Language Models: A Survey (Gao et al., 2024)

**Summary:** This survey paper provides a comprehensive overview of the RAG framework, its applications, and its impact on various NLP tasks. It discusses different strategies for

integrating retrieval mechanisms into LLMs and analyzes the performance of RAG on different benchmarks. This paper is a valuable resource for understanding the broader landscape of RAG research.

5. Retrieval-Augmented Generation for AI-Generated Content: A Survey (Zhao et al., 2024)

**Summary:** This latest survey paper comprehensively reviews existing efforts of integrating RAG into Artificial Intelligence Generated Content (AIGC) scenarios. It categorizes RAG foundations based on how the retriever augments the generator, summarizes enhancement methods, surveys practical applications across different modalities and tasks, introduces RAG benchmarks, discusses current limitations, and suggests future research directions.

3 Datasets

I use the following datasets for my analysis:

- 1. “wiki\_dpr” Dataset (Karpukhin et al., 2020): This is the Wikipedia split used to evaluate the Dense Passage Retrieval (DPR) model. It contains 21M passages from wikipedia along with their DPR embeddings. The wikipedia articles are split into multiple, disjoint text blocks of 100 words as passages. (Lewis et al., 2021) uses this dataset as its non-parametric knowledge source.

Example Document:

Daniel Jones (musician) Daniel Jones (born 22 July 1973) is an Australian musician, songwriter, and record producer. He is best known for being a member of the successful Australian pop duo Savage Garden, whose international hit singles included the songs “I Want You”, “To the Moon and Back”, “Truly Madly Deeply”, “I Knew I Loved You”, and “Crash and Burn”. After Savage Garden split in 2001, Jones launched his own production company (Meridien Musik) and a recording studio (Level 7 Studios), to work with various new, young Australian artists. Jones was born at Southend-on-Sea, Essex, England, but his family moved...

Example Query:

Who is Daniel Jones?

- 2. “ArguAna” Counterargs Corpus (Wachsmuth et al., 2018): This dataset is designed for studying the retrieval of the best counterarguments to given arguments. With 8,674 documents and 1,406 associated queries, this corpus provides a nuanced perspective on argumentative structures. Importantly, gold-standard mappings for relevant documents are available for all queries in the ArguAna dataset, ensuring a robust evaluation framework.

Example Document:

You don't have to be vegetarian to be green. Many special environments have been created by livestock farming - for example chalk down land in England and mountain pastures in many countries . Ending livestock farming would see these areas go back to woodland with a loss of many unique plants and animals. Growing crops can also be very bad for the planet, with fertilisers and pesticides polluting rivers, lakes and seas. Most tropical forests are now cut down for timber, or to allow oil palm trees to be grown in plantations, not to create space for meat production...

Example Query:

Being vegetarian helps the environment Becoming a vegetarian is an environmentally friendly thing to do. Modern farming is one of the main sources of pollution in our rivers. Beef farming is one of the main causes of deforestation, and as long as people continue to buy fast food in their billions, there will be a financial incentive to continue cutting down trees to make room for cattle...

- 3. NutritionFacts “NFCorpus” (Boteva et al., 2016): This dataset focuses on Medical Information Retrieval. Comprising 3,633 documents and 323 associated queries, this dataset allows for a specialized exploration of medical domains. Gold-standard mappings for relevant documents are available for all queries in the NFCorpus.

Example Document:

Recent studies have suggested that statins, an established drug group in the prevention of cardiovascular mortality, could delay or prevent breast cancer recurrence but the effect on disease-specific mortality remains unclear. We evaluated risk of breast cancer death among statin users in a population-based cohort of breast cancer patients. The study cohort included all newly diagnosed breast cancer patients in Finland during 1995–2003 (31,236 cases), identified from the Finnish Cancer Registry...

Example Query:

Do Cholesterol Statin Drugs Cause Breast Cancer?

I employ nltk, the Natural Language Toolkit (Bird and Loper, 2004), to perform essential data pre-processing tasks. Documents as well as query texts are converted to lowercase, and words are tokenized using nltk’s word\_tokenizer. To enhance the quality of the text, stop-words and words with lengths less than 2 are systematically removed; Punctuation is discarded, and the WordNetLemmatizer is applied to each word, ensuring that each is transformed to its lemma. This data pre-processing step is common to each of the aforementioned datasets.

4 Analysis Approach

I use the following tools:

- OpenAI’s API (https://platform.openai.com/) to generate embeddings and to interface with ChatGPT LLM. Specifically, OpenAI embeddings based on the “text-embedding-ada-002” model (It maps the documents and queries vectors to a 1536-dimensional vector space), and OpenAI’s “gpt-3.5-turbo-instruct” model as the LLM.
- FAISS-CPU library (Douze et al., 2024) for efficient similarity search and clustering of high-dimensional embeddings.

- Langchain (<https://www.langchain.com/>) to enable the LLM to connect with sources of context (prompt instructions, few-shot examples, content to ground its response in, etc.), thus enabling it to perform RAG.

My approach is as follows:

1. Following the data pre-processing in the previous section, I leverage OpenAI's API and the "text-embedding-ada-002" model, to map our documents and queries vectors to a 1536-dimensional vector space.
2. Then, I build a Retriever. Essentially, I utilize the Facebook AI Similarity Search (FAISS) vector store available in "langchain.vectorstores" to build a VectorDB of embeddings obtained in the previous step. I make sure that each embedding is a unit vector and employ the dot product metric to find similarity between any two vectors. (Note that cosine similarity with unit vectors is the same as dot product). This allows me to retrieve the top- $k$  documents, with  $k$  set to 10, optimizing the relevance of retrieved information in the subsequent steps.
3. This step harnesses the power of OpenAI's language model ("gpt-3.5-turbo-instruct") available in "langchain.llms". Furthermore, to perform RAG, I create a chain by first defining a prompt based on a fixed "ChatPromptTemplate" template. Then, I pass the prompt along with the question and the context to the LLM. I use the following prompts:

**Prompt 1** (Used for "wiki\_dpr" and "NFCorpus" datasets):

Answer the question or Explain the topic given this additional context: {context}  
Question: {question}

**Prompt 2** (Used for "ArguAna" dataset):

How many distinct arguments and counterarguments can you identify from this following debate paragraphs?  
Output the total number for each of them, before you output the arguments/evidence.

Argument:  
{question}

Counterargument:  
{context}

Note: For prompt 2, I only retrieve the first document (i.e.,  $k = 1$ ) as context.

**Prompt 3** (Used for "ArguAna" dataset):

Identify the passage with the highest relevance to the query.

Query: {question}  
Passages:  
{context}

(Wang et al., 2024) is the motivation behind Prompts 2 and 3.

4. Finally, I try out multiple example queries to identify concrete examples of failure cases. This may include modifying the example documents to inject a conflicting or incorrect piece of text/fact into the beginning of retrieved context.

## 5 Errors and their Categorization

Before diving into the errors, here is a sample query which results in a correct output through the application of RAG in LLMs:

**Query** ("wiki\_dpr" Dataset):

Who is Idris Elba?

**Retrieved Context** ("wiki\_dpr" Dataset):

"Idris Elba Idrissa Akuna Elba (; born 6 September 1972) is a British actor, producer, musician and DJ known for playing drug trafficker Stringer Bell on the HBO series ""The Wire"", DCI John Luther on the BBC One series ""Luther"" and Nelson Mandela in the biographical film "" (2013)..."

""RocknRolla"" with Idris Elba, Butler took the lead in several 2009 films including the romantic comedy ""The Ugly Truth"" with Katherine Heigl and the thriller ""Law Abiding Citizen"" with Jamie Foxx. In 2010, Butler voiced the role of Stoick the Vast in the animated action-fantasy film ""How to Train Your Dragon""..."

"Stupnitsky and directed by Paul Feig. Executive story editor Charlie Grandy conceived the idea of Michael leaving Dunder Mifflin. ""New Boss"" was the first of six episodes Idris Elba was set to guest star in as Charles, Michael's new Dunder Mifflin superior. It was his first role in a comedy and his first television appearance since leaving his regular role on HBO's ""The Wire"", where he played the character Stringer Bell ..."

"Stringer Bell Russell ""Stringer"" Bell born September 17, 1969, is a fictional character in ""The Wire"", played by Idris Elba. Bell serves as drug kingpin Avon Barksdale's second-in-command, assuming direct control of the Barksdale Organization during Avon's imprisonment. Bell attends economics classes at Baltimore City Community College and maintains a personal library..."

...

**LLM Output:**

Idris Elba is a British actor, producer, musician, and DJ known for his roles in various television shows and films such as "The Wire", "Luther", "Mandela: Long Walk to Freedom", and "Thor". He has received multiple award nominations and has won a Golden Globe and a Primetime Emmy. He has also been honored for his contributions to drama and was named as one of the 50 people who have helped shape London's cultural landscape.

Following are the errors found during the analysis, as supported by specific examples.

### 5.1 Example Error 1

**Query** ("wiki\_dpr" Dataset):

What are the movies or series Idris Alba acted in between 2011 and 2013?

**Retrieved Context** (Prompt 1, "wiki\_dpr" Dataset):

"Idris Elba Idrissa Akuna Elba (; born 6 September 1972) is a British actor, producer, musician and DJ known for playing drug trafficker Stringer Bell on the HBO series ""The Wire"", DCI John Luther on the BBC One series ""Luther"" and Nelson Mandela in the biographical film "" (2013). He has been nominated four times for a Golden Globe Award for Best Actor - Miniseries or Television Film, winning one and was nominated five times for a Primetime Emmy Award. Elba appeared in Ridley Scott's ""American Gangster"" (2007) and ""Prometheus"" (2012). Elba portrays Heimdall in ""Thor"" (2011) and its sequels"

""RocknRolla"" with Idris Elba, Butler took the lead in several 2009 films including the romantic comedy ""The Ugly Truth"" with Katherine Heigl and the thriller ""Law Abiding Citizen"" with Jamie Foxx. In 2010, Butler voiced the role of Stoick the Vast in the animated action-fantasy film ""How to Train Your Dragon""..."

<p>"Honours for services to drama. In 2017, Elba won the male title for the ""Rear of the Year"" award in Britain. In September 2018, he was one of 50 people named for ""making London awesome"" and helping to shape London 's cultural landscape who were featured in ""Time Out"" to mark the magazine's 50th anniversary..."</p> <p>"Saturn Award for Best Actor and a win for MTV Movie Award for Best Fight . In the 2010s, he voiced the role of Stoick the Vast in the animated action-fantasy film ""How to Train Your Dragon"", a role he later reprised in ""Legend of the Boneknapper Dragon"" (2010), ""Gift of the Night Fury"" (2011), and ""How to Train Your Dragon 2"" (2014). He played military leader Tullus Aufidius in the 2011 film ""Coriolanus"", a modernized adaptation of Shakespeare's tragedy of the same name. He also played Sam Childers in the 2011 action biopic ""Machine Gun Preacher"". Gerard James Butler was"</p> <p>""She's All That"" (1999), and ""The Skulls"" (2000). In 2001, Walker's breakthrough role arrived when he starred opposite actor Vin Diesel in the successful action film ""The Fast and the Furious"", the first film in the franchise. The film established Walker as a notable film star and leading man and led to his reprise of the role in the 2003 sequel ""2 Fast 2 Furious""..."</p> <p>"2015, most of which failed financially except the animated film ""The Little Prince"", a modest commercial success. James Franco filmography James Franco is an American actor who began acting on television, guest-starring in ""Pacific Blue"" (1997). He landed his breakthrough role in the comedy-drama television series ""Freaks and Geeks"" (1999-2000). After his film debut in ""Never Been Kissed"" (1999), Franco won a Golden Globe Award for Best Actor..."</p> <p>"Serpent's Heir"" in 2016. The series will be co-written by Dean DeBlois, writer and director of the film series, and Richard Hamilton, writer of ""Dragons: Race to the Edge"", with the production designer of ""How to Train Your Dragon 2"", Pierre-Olivier Vincent, providing cover artwork. The series will take place between the second and third film, with the first novel picking up right after the conclusion of the second film..."</p> <p>"series about British monarch Elizabeth II. In 2017, Dillane appeared in two biopics, playing Edward Wood, 1st Earl of Halifax in Joe Wright's ""Darkest Hour"", starring Gary Oldman as Winston Churchill, and writer William Godwin, the father of ""Frankenstein"" author Mary Shelley..."</p> <p>"the Seven Kingdoms. HBO ordered the seventh season on April 21, 2016, three days before the premiere of the show's sixth season, and began filming on August 31, 2016. The season was filmed primarily in Northern Ireland, Spain, Croatia and Iceland. ""Game of Thrones"" features a large ensemble cast, including Peter Dinklage, Nikolaj Coster-Waldau, Lena Headey, Emilia Clarke, and Kit Harington..."</p> <p>"Stupnitsky and directed by Paul Feig. Executive story editor Charlie Grandy conceived the idea of Michael leaving Dunder Mifflin. ""New Boss"" was the first of six episodes Idris Elba was set to guest star in as Charles, Michael's new Dunder Mifflin superior. It was his first role in a comedy and his first television appearance since leaving his regular role on HBO's ""The Wire"", where he played the character Stringer Bell. Elba, a fan of ""The Office"", said, ""The creators of the show called me , said they wanted to put me in as this new character, that I'd be perfect"</p>
---

### LLM Output:

<p>In 2011, Idris Alba appeared in the film "Thor" and its sequels. He also played Tullus Aufidius in the film "Coriolanus". In 2013, he portrayed Nelson Mandela in the biographical film "Mandela: Long Walk to Freedom".</p>
---

There are two problems in the output. First, Tullus Aufidius in the film “Coriolanus” was played by Gerard Butler. Second, the LLM missed “Luther” (2011–) and “Prometheus” (2012). I categorize these two problems as inability to handle unrelated noise and perform mathematical reasoning, respectively.

### 5.2 Example Error 2

#### Query (Prompt 1, “wiki\_dpr” Dataset):

<p>How many movies or series Idris Alba acted in between 2011 and 2013, and what are they?</p>
--

**Retrieved Context** (“wiki\_dpr” Dataset): Same as Example Error 1

#### LLM Output:

<p>Idris Alba acted in five movies or series between 2011 and 2013. They are "Thor" (2011), "Luther" (2011-2013), "Coriolanus" (2011), "American Gangster" (2012), and "Prometheus" (2012).</p>
---

There are several problems in the output. “American Gangster” was released in 2007; “Coriolanus” is not related to Idris Alba; The LLM missed “Mandela: Long Walk to Freedom” (2013); and “Luther” dates are incorrect (2011–2019). Again, I categorize these problems as inability to handle unrelated noise and perform mathematical reasoning.

### 5.3 Example Error 3

#### Query (“wiki\_dpr” Dataset):

<p>Between the age of 35 and 40, what movies or series did Idris Alba act in ?</p>
--

**Retrieved Context** (Prompt 1, “wiki\_dpr” Dataset): Same as Example Error 1  
**LLM Output:**

<p>At the age between 35 and 40 (2007-2012), Idris Alba acted in "American Gangster" (2007), "Prometheus" (2012), "Thor" (2011), "The Fast and the Furious" (2001), "2 Fast 2 Furious" (2003), "Joy Ride" (2001), "Timeline" (2003), "Into the Blue" (2005), "Darkest Hour" (2017), and "Hunter Killer" (2018).</p>
---

Here, although the model is able to identify the range of years correctly, it still returns an incorrect set of movies and series. Clearly, this relates to the model’s inability to handle unrelated noise and perform mathematical reasoning and information integration.

### 5.4 Example Error 4

#### Query (“wiki\_dpr” Dataset):

<p>What movies or series were ‘produced’ by Idris Elba?</p>
---

**Retrieved Context** (Prompt 1, “wiki\_dpr” Dataset): Same as Example Error 1  
**LLM Output:**

<p>Some of the movies and series that were produced by Idris Elba include " Luther", "The Wire", "Thor", "American Gangster", "Prometheus", "How to Train Your Dragon", "Legend of the Boneknapper Dragon", "Gift of the Night Fury", "How to Train Your Dragon 2", "Coriolanus", and "RocknRolla ".</p>
--

Idris Alba has only produced episodes of “Luther”, the rest are incorrect. This problem relates to the model’s inability to reject negative, conflicting, or counterfactual information.

### 5.5 Example Error 5

#### Query (“ArguAna” Dataset):

<p>Being vegetarian helps the environment Becoming a vegetarian is an environmentally friendly thing to do. Modern farming is one of the main sources of pollution in our rivers. Beef farming is one of the main causes of deforestation, and as long as people continue to buy fast food in their billions, there will be a financial incentive to continue cutting down trees to make room for cattle. Because of our desire to eat fish, our rivers and seas are being emptied of fish and many species are facing extinction. Energy resources are used up much more greedily by meat farming than my farming cereals, pulses etc. Eating meat and fish not only causes cruelty to animals, it causes serious harm to the environment and to biodiversity. For example consider Meat production related pollution and deforestation At Toronto's 1992 Royal Agricultural Winter Fair, Agriculture Canada displayed two contrasting statistics: "it takes four football fields of land (about 1.6 hectares) to feed each Canadian" and "one apple tree produces enough fruit to make 320 pies." Think about it - a couple of apple trees and a few rows of wheat on a mere fraction of a hectare could produce enough food for one person! [1] The 2006 U.N. Food and Agriculture Organization (FAO) report concluded that worldwide livestock farming generates 18% of the planet's greenhouse gas emissions - by comparison, all the world's cars, trains, planes and boats account for a combined 13% of greenhouse gas emissions. [2] As a result of the above point producing meat damages the environment. The demand for meat drives deforestation. Daniel Cesar Avelino of Brazil's Federal Public Prosecution Office says "We know that the single biggest</p>
---

driver of deforestation in the Amazon is cattle." This clearing of tropical rainforests such as the Amazon for agriculture is estimated to produce 17% of the world's greenhouse gas emissions. [3] Not only this but the production of meat takes a lot more energy than it ultimately gives us chicken meat production consumes energy in a 4:1 ratio to protein output; beef cattle production requires an energy input to protein output ratio of 54:1. The same is true with water use due to the same phenomenon of meat being inefficient to produce in terms of the amount of grain needed to produce the same weight of meat, production requires a lot of water. Water is another scarce resource that we will soon not have enough of in various areas of the globe. Grain-fed beef production takes 100,000 liters of water for every kilogram of food. Raising broiler chickens takes 3,500 liters of water to make a kilogram of meat. In comparison, soybean production uses 2,000 liters for kilogram of food produced; rice, 1,912; wheat, 900; and potatoes, 500 liters. [4] This is while there are areas of the globe that have severe water shortages. With farming using up to 70 times more water than is used for domestic purposes: cooking and washing. A third of the population of the world is already suffering from a shortage of water. [5] Groundwater levels are falling all over the world and rivers are beginning to dry up. Already some of the biggest rivers such as China's Yellow river do not reach the sea. [6] With a rising population becoming vegetarian is the only responsible way to eat. [1] Stephen Leckie, 'How Meat-centred Eating Patterns Affect Food Security and the Environment', International development research center [2] Bryan Walsh, Meat: Making Global Warming Worse, Time magazine, 10 September 2008 . [3] David Adam, Supermarket suppliers 'helping to destroy Amazon rainforest', The Guardian, 21st June 2009. [4] Roger Segelken, U.S. could feed 800 million people with grain that livestock eat, Cornell Science News, 7th August 1997. [5] Fiona Harvey, Water scarcity affects one in three, FT.com, 21st August 2003 [6] Rupert Wingfield-Hayes, Yellow river 'drying up', BBC News, 29th July 2004

## Retrieved Context (Prompt 2, “ArguAna” Dataset):

You don't have to be vegetarian to be green. Many special environments have been created by livestock farming - for example chalk down land in England and mountain pastures in many countries. Ending livestock farming would see these areas go back to woodland with a loss of many unique plants and animals. Growing crops can also be very bad for the planet, with fertilisers and pesticides polluting rivers, lakes and seas. Most tropical forests are now cut down for timber, or to allow oil palm trees to be grown in plantations, not to create space for meat production. British farmer and former editor Simon Farrell also states: "Many vegans and vegetarians rely on one source from the U.N. calculation that livestock generates 18% of global carbon emissions, but this figure contains basic mistakes. It attributes all deforestation from ranching to cattle, rather than logging or development. It also muddles up one-off emissions from deforestation with on-going pollution." He also refutes the statement of meat production inefficiency: "Scientists have calculated that globally the ratio between the amounts of useful plant food used to produce meat is about 5 to 1. If you feed animals only food that humans can eat - which is, indeed, largely the case in the Western world - that may be true. But animals also eat food we can't eat, such as grass. So the real conversion figure is 1.4 to 1." [1] At the same time eating a vegetarian diet may be no more environmentally friendly than a meat based diet if it is not sustainably sourced or uses perishable fruit and vegetables that are flown in from around the world. Eating locally sourced food can have as big an impact as being vegetarian. [2] [1] Tara Kelly, Simon Fairlie: How Eating Meat Can Save the World, 12 October 2010 [2] Lucy Siegle, 'It is time to become a vegetarian?' The Observer, 18th May 2008

## LLM Output 1:

Total arguments: 6  
Total counterarguments: 6

Arguments:

1. Being vegetarian helps the environment
2. Modern farming is a main source of pollution
3. Meat and fish consumption causes harm to environment and biodiversity
4. Meat production contributes to greenhouse gas emissions and deforestation
5. Meat production is inefficient in terms of energy and water use
6. Rising population makes vegetarianism a responsible way to eat

Counterarguments:

1. Vegetarianism is not the only way to be environmentally friendly
2. Livestock farming contributes to unique environments
3. Deforestation is not solely driven by meat production
4. Livestock emissions and pollution are exaggerated
5. Meat production is not as inefficient as claimed
6. Vegetarianism is not always more environmentally friendly than meat-based diets

## LLM Output 2:

Arguments: 8

1. Being vegetarian helps the environment
2. Modern farming is a main source of pollution in rivers
3. Beef farming causes deforestation
4. Buying fast food leads to deforestation
5. Eating meat and fish causes harm to the environment
6. Meat production leads to pollution and deforestation
7. Demand for meat drives deforestation

8. Becoming vegetarian is the only responsible way to eat due to a rising population

Counterarguments: 8

1. You don't have to be vegetarian to be environmentally friendly
2. Livestock farming creates special environments
3. Ending livestock farming would result in the loss of unique plants and animals
4. Growing crops can also be harmful to the environment
5. Most tropical forests are cut down for purposes other than meat production
6. The calculation that livestock generates 18% of global carbon emissions is flawed
7. Eating a vegetarian diet may not necessarily be more environmentally friendly
8. Eating locally sourced food can have the same impact as being vegetarian

The LLM does a great work to provide an almost accurate result. However, the number of arguments and counterarguments vary across different runs. In fact, there are exactly 6 arguments and 8 counterarguments in the above example. This variation in outputs across different runs makes it harder to evaluate the LLM.

## 5.6 Example Error 6

### Query (“ArguAna” Dataset):

Defaulting would cause chaos in Greece There is no good solution for the crisis Greece finds itself in, only less bad ones. Austerity measures imposed on Greece may currently be causing suffering, but austerity is the least bad option available for the Greek people: default would be considerably worse. Here is what would most likely happen: The Greek banking sector would collapse [1]. A large portion of the Greek debt is owed to Greek banks and companies, many of which would quickly go bankrupt when the Government defaults. This is also because Greek banks are almost totally reliant on the ECB for liquidity. [2] People would consequently lose their savings, and credit would be close to impossible to find. The Government would quickly devalue the Drachma by at least 50%. This will lead to imported goods being more expensive and consequently to a huge rise in inflation with the living costs increasing tremendously.[3] These two events would lead to a severe shortage of credit, making it almost impossible for struggling companies to survive. Unemployment would soar as a result. It will become increasingly difficult to secure supplies of oil, medicine, foodstuffs and other goods . Naturally, those hit worst would be the poor. The Government, in this respect, would be failing on an enormous scale in providing many citizens with the basic needs. [4] [1] Brzeski, Carsten: "Viewpoints: What if Greece exits euro?", BBC News, 13 July 2012, [2] Ruparel, Raoul and Persson, Mats: "Better off Out? The short-term options for Greece inside and outside of the euro", June 2012, Open Europe, 2012 [3] ibid [4] Arghyrou, Michael: "Viewpoints: What if Greece exits euro?", BBC News, 13 July 2012,

### Retrieved Context (Shuffled) (Prompt 3, “ArguAna” Dataset):

A Greek default would have a negative domino-effect on other Eurozone countries. A Greek default will leave tremendous shockwaves across the Eurozone. Investors will instantly become wary of default in Portugal, Spain, Italy or Ireland, particularly given the sudden nature of the Greek default. Consequently, huge volumes of capital will flow out of these countries and into other more secure ones like Germany and the Netherlands. [1] This will, in turn, heighten speculation about the danger of default of other Eurozone nations. Speculation of default is particularly dangerous because it drives demand for government bonds down ...

Defaulting would be the quickest route to economic recovery Under the status quo, the Greek economy is only headed in one direction: deeper recession. There are no signs of the situation changing any time soon. Were the Greek Government to default on its debts, after a period of recession, conditions would quickly be favourable for economic growth once more. This is what was observed when Argentina and other nations [1] recently defaulted and can be explained by many factors...

It is not necessarily true that the whole banking sector in Greece would collapse. Given that the default would be orderly and take place within the context of the European Union, the ECB and European Commission would still provide substantial liquidity aid for Greek banks. Moreover it is not true that a devaluation of domestic currency necessarily leads to high inflation - this was not the case, for example, when Britain exited the European Exchange-rate Mechanism in 1992 and pursued a devaluation policy of the British Pound...

Defaulting would not solve Greece's problems The proposition argue that the hardship endured by the default would only be temporary, but an analysis at the particular situation facing Greece indicates the opposite . Greece's problems arose from a horrifically inefficient public sector embedded within a mentality of corruption and tax evasion. Even if we assume that defaulting would eventually boost Greek exports and help the

economy recover, this would not solve the underlying problems that caused the crisis in the first place...

The proposition vastly understates the negative impact a default has on the local economy. It is unrealistic to compare Greece with Argentina. As a member of the Eurozone, the developments within the Greek debt crisis have a huge impact on nations suffering from similar problems, as well as the Eurozone as a whole. Moreover, devaluing the Drachma would be nowhere near as beneficial as the proposition suggests. Greece is not rich in natural resources or industry and so boosting exports will not make a huge difference...

Greece's default will not decrease uncertainty. If anything, the perceived risk of investing in other Eurozone members suffering from their own debt problems like Italy, Spain, Portugal and Ireland would rocket sky-high. The Eurozone project as a whole may struggle on with Germany trying to keep it together, but claiming that a Greek exit from the Eurozone would restore stability is short-sighted...

The current austerity measures are not working The Austerity measures put in place by the ECB, IMF and European Commission have led to nothing but misery for the Greek people. They have failed to cut down the total debt % GDP ratio and have also failed to increase the competitiveness of the Greek economy. This is because raising taxes and slashing the minimum wage has sent the economy deeper and deeper into recession...

### LLM Output:

The current austerity measures are not working The Austerity measures put in place by the ECB, IMF and European Commission have led to nothing but misery for the Greek people. They have failed to cut down the total debt % GDP ratio and have also failed to increase the competitiveness of the Greek economy. This is because raising taxes and slashing the minimum wage has sent the economy deeper and deeper into recession...

The above answer is incorrect. The correct answer is paragraph 3 in the retrieved context:

It is not necessarily true that the whole banking sector in Greece would collapse. Given that the default would be orderly and take place within the context of the European Union, the ECB and European Commission would still provide substantial liquidity aid for Greek banks. Moreover it is not true that a devaluation of domestic currency necessarily leads to high inflation - this was not the case, for example, when Britain exited the European Exchange-rate Mechanism in 1992 and pursued a devaluation policy of the British Pound...

It looks like the LLM forgot what is in the middle of the context. Therefore, this error relates to the model's inability to handle unrelated noise.

## 5.7 Example Error 7

### Query 1 ("NFCorpus" Dataset):

Pharmacists Versus Health Food Store Employees: Who Gives Better Advice?

### Retrieved Context (Prompt 1, "NFCorpus" Dataset):

OBJECTIVES: To ascertain the recommendations, training and education of health food store employees and determine how they communicate the costs, benefits and risks associated with natural health products for the HIV/AIDS community. METHODS: Four male research assistants, posing as asymptomatic HIV-positive individuals, inquired of employees of all retail health food stores in a major Canadian city as to what is recommended for their condition. The research assistants asked about product costs, side effects, potential drug interactions and efficacy. They also inquired as to employee education related to Complementary and Alternative Medicine (CAM) and noted whether employees asked about which conventional medications they were taking and whether they recommended that the subjects seek physician or CAM provider advice. RESULTS: A total of 32 stores were included. Eight store employees (25%) offered no advice; eight (25%) inquired whether the subjects were currently taking medications; six (19%) suggested visiting a physician; and eight (25%) suggested visiting a CAM provider. A total of 36 different products (mean 2.3 per employee) were recommended with considerable variability in product evidence and cost. The education of the employees varied from postgraduate education (n=3), to undergraduate degree (n=3), college level (n=5) in CAM, or no formal education in CAM (n=21). CONCLUSION: There was considerable heterogeneity in advice on natural food products provided by employees of natural food stores and, in general, these individuals had limited formal training in CAM. The products they recommended had limited evidence supporting their efficacy and in some instances were potentially harmful and had considerable costs. The findings of this study support the need to further examine how best to regulate this growing component of the health care system.

Alternative health practices have become increasingly popular in recent years. Many patients visit specific complementary practitioners, while others attempt to educate themselves, trusting advice from employees at local health food stores or the Internet. Thirty-two retail health food stores were surveyed on the nature of the information provided by their staff. A research assistant visited the stores and presented as the mother of a child in whom Crohn's disease had been diagnosed...

CONTEXT: Despite cancer patients' widespread and growing use of complementary and alternative medicine, minimal attention has been paid to the role of health food stores in the "supply side" of this phenomenon . OBJECTIVE: To gain a better understanding of health food store personnel's recommendations for breast cancer patient care. DESIGN: Researcher posing as the daughter of a breast cancer patient and surveying health food store personnel on their product recommendations for cancer care. SETTING: Oahu, Hawaii, summer 1998. PARTICIPANTS: All health food stores (N = 40) offering products for cancer patients...

The goal of this research was to evaluate the personal health behaviors of physicians in training and attending physicians in association with patient-related lifestyle counseling. Physicians at a major teaching hospital were surveyed regarding their personal lifestyle behavior, perceived confidence, and frequency of counseling patients regarding lifestyle behaviors. One hundred eighty-three total responses were received. Trainees were more likely to consume fast food and less likely to consume fruits and vegetables than attendings. Attending physicians were more likely to exercise 4 or more days per week and more than 150 minutes per week...

Medical students in the United States are taught little about nutrition and dietetics. Worse yet, their training biases them against the studies that show the power of dietary approaches to managing disease. The current approach to evidence-based medicine encourages physicians to ignore any information that does not come from a double-blind, randomized controlled trial. Yet human beings cannot be blinded to a dietary intervention. As a result, physicians are biased toward drug treatments and against dietary interventions for the management of chronic disease. Copyright 2013 Elsevier Ltd. All rights reserved.

### LLM Output:

Answer: The study mentioned is focused on comparing the recommendations and advice given by pharmacists versus health food store employees, specifically in relation to natural health products for the HIV/AIDS community. The study found that there was considerable heterogeneity in the advice provided by health food store employees, with limited formal training in complementary and alternative medicine (CAM). In contrast, pharmacists are highly trained healthcare professionals who have a deeper understanding of drug interactions and potential risks associated with natural health products. Therefore, it can be inferred that pharmacists are better equipped to give advice and recommendations to patients in this context.

### Query 2 ("NFCorpus" Dataset):

Pharmacists Versus Medical students Versus Health Food Store Employees: Who Gives Better Advice?

### Retrieved Context (Prompt 1, "NFCorpus" Dataset): Same as in Query 1

### LLM Output:

Answer: The study on the comparative knowledge and practices of healthcare professionals and medical students showed that doctors had the best knowledge of lifestyle modification, including nutrition, physical activity, and smoking cessation. However, the study also found that health food store employees were inconsistent in their advice and had limited formal training in complementary and alternative medicine. It is important for medical students and healthcare professionals to receive adequate education and training in lifestyle modification to provide evidence-based recommendations to their patients.

Clearly, the output for query 1 is accurate based on the provided context. However, the output for query 2 is not fully correct as the context lacks evidence for comparison between medical students and pharmacists or health food store employees. In such a case, one would expect the LLM should reject to answer the question as the required knowledge is not fully present in the retrieved context. This problem could be categorized as the inability of the model to perform negative rejection.

## 6 Discussion

Based on the aforementioned errors, here's a breakdown of the challenges that RAG models face:

1. **Unrelated Noise:** RAG systems retrieve information from external sources to inform



their responses. The challenge is that these sources might contain irrelevant details. RAG models struggle to filter out this noise, leading to responses that are cluttered or misleading. This can be seen in errors 1 and 2, where the LLM incorrectly attributes “Coriolanus” to Idris Alba, or misses the years associated with movies and TV series. This is because the retrieved context contains a lot more information than required, such as the movies and series related to Gerard Butler or Paul Walker.

Furthermore, this also relates to a rather well studied problem in LLMs, known as “Lost in the Middle” (Liu et al., 2023a). Here, the authors show that the LLM’s performance can degrade significantly when changing the position of relevant information, indicating that language models do not robustly make use of information in long input contexts. In particular, they observe that performance is often highest when relevant information occurs at the beginning or end of the input context, and significantly degrades when models must access relevant information in the middle of long contexts. This issue can be seen in error 6 where the LLM is unable to identify the passage most relevant to the provided input query when the passage is moved to the middle of the context.

2. **Reasoning of Maths:** RAG models are currently not designed to handle mathematical reasoning. They can retrieve information about math concepts, but they can’t perform calculations, or understand the logical steps involved. Again, this can be seen in errors 1, 2, and 3, where the model is unable to figure out the movies or series that were released within a provided range of years. The LLM either incorrectly tries to include movies or series that were released outside the provided range or misses the ones that were released inside the provided range.
3. **Information Integration:** RAG systems can struggle to effectively combine information retrieved from different sources. They might simply concatenate the retrieved snippets without truly understanding the relationships between them. This can lead to responses that lack coherence or miss out on insights from combining the information. This problem can be seen in error 3 where the model tries to integrate two pieces of information, *viz.*, finding the range of years based on Idris Alba’s date of birth, and the movies and series related to Idris Alba. Although it is able to achieve the former, it struggles to achieve the latter.
4. **Negative Rejection:** RAG models might not always correctly interpret negative or missing

statements in prompts or retrieved information. This can lead to them to hallucinate (Kalai and Vempala, 2024) or generating responses that contradict the negation. In such scenarios, one would expect the LLM to indicate that there is insufficient information or provide other rejection signals. This is particularly evident in error 7 where the output for query 1 is accurate based on the provided context, but the output for query 2 is not fully correct as the context lacks evidence for comparison between medical students and pharmacists or health food store employees.

5. **Conflicting Knowledge and Counterfactual Robustness:** RAG models might retrieve information from sources that contain contradictory information, hypothetical scenarios or counterfactuals (what-ifs). They might not be able to detect these inconsistencies, leading to responses that lack credibility or contain factual errors. This issue can be seen in error 4 where the model incorrectly attributes several movies and series to Idris Alba as their producer. The fact is (and this can be verified from the provided context) that Idris Alba acted in those movies and series and did not produce them.
6. **Evaluation:** Evaluating RAG system is hard. Particularly, creating high-quality human-written responses for a wide range of prompts can be time-consuming and expensive. This limits the amount of ground truth data available for evaluation, which can impact the accuracy and generalizability of the evaluation results. Furthermore, traditional evaluation metrics might not accurately reflect the effectiveness of RAG systems. As seen in error 5, different runs of the same query can result in different outputs making it harder to evaluate the correct response.

## 7 The Way Forward

Researchers are actively working on improving RAG models to address the above limitations. Some potential areas of improvement include:

- Developing better context management techniques to help RAG models understand the full conversation history. (Tay et al., 2022) provides a comprehensive survey on the improvements around computational and memory efficiency in Transformers. Other architectures like Mamba SSM (Gu and Dao, 2024) are focusing on developing subquadratic-time architectures and million length input context.
- Implementing noise reduction methods to filter out irrelevant information retrieved from external sources. A popular example is to use Graph RAG (Hu et al., 2024; Edge et al., 2024) instead

of classical RAG. Graph RAG includes a graph database as a source of the contextual information sent to the LLM, which provides structured entity information to the LLM (combining the entity textual description with its many properties and relationships), thereby encouraging deeper insights.

- Developing new metrics specifically designed to evaluate RAG models. These metrics might focus on aspects like relevance, diversity, and factual accuracy of the retrieved information. Furthermore, techniques like approximate nearest neighbor search or active learning can be used to reduce the computational cost of evaluation, allowing for more efficient testing of RAG models (Es et al., 2023; Hoshi et al., 2023; Chen et al., 2023; Liu et al., 2023b; Saad-Falcon et al., 2024).
- Incorporating symbolic reasoning capabilities to enable RAG models to handle mathematical problems. (Ahn et al., 2024) provides a comprehensive survey that examines the landscape of LLMs in the realm of mathematics.

By addressing these challenges, RAG models can become even more powerful tools for generating informative and comprehensive responses.

## 8 Conclusion

The challenges faced by RAG systems highlight significant areas for improvement. Key issues include managing unrelated noise from external sources, handling mathematical reasoning, effectively integrating information, interpreting negative or missing statements, dealing with conflicting knowledge, and the difficulty of evaluating these systems. These challenges often result in irrelevant, inaccurate, or incoherent responses.

To address these limitations, researchers are exploring several avenues. Enhanced context management techniques and architectures, such as Mamba SSM, aim to help models better understand full conversation histories. Noise reduction methods, like Graph RAG, seek to filter out irrelevant details by leveraging structured entity information. New evaluation metrics are being developed to more accurately assess the relevance, diversity, and factual accuracy of responses. Additionally, incorporating symbolic reasoning capabilities could enable RAG models to handle mathematical problems more effectively.

By focusing on these areas, the performance and reliability of RAG models can be significantly improved, making them more effective tools for generating informative and accurate responses.

## 9 Acknowledgements

I did not use ChatGPT or any other Generative AI tool for the purpose of writing this project report.

## References

- Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges.
- Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., and Sifre, L. (2022). Improving language models by retrieving from trillions of tokens.
- Boteva, V., Gholipour, D., Sokolov, A., and Riezler, S. (2016). A full-text learning to rank dataset for medical information retrieval. In Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., and Silvello, G., editors, *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.
- Chen, J., Lin, H., Han, X., and Sun, L. (2023). Benchmarking large language models in retrieval-augmented generation.
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2024). The faiss library.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization.
- Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.
- Gu, A. and Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces.
- Hoshi, Y., Miyashita, D., Ng, Y., Tatsuno, K., Morioka, Y., Torii, O., and Deguchi, J. (2023). Ralle: A framework for developing and evaluating retrieval-augmented large language models.
- Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., and Zhao, L. (2024). Grag: Graph retrieval-augmented generation.
- Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering.
- Kalai, A. T. and Vempala, S. S. (2024). Calibrated language models must hallucinate.



- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023a). Lost in the middle: How language models use long contexts.
- Liu, Y., Huang, L., Li, S., Chen, S., Zhou, H., Meng, F., Zhou, J., and Sun, X. (2023b). Recall: A benchmark for llms robustness against external counterfactual knowledge.
- Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2024). Ares: An automated evaluation framework for retrieval-augmented generation systems.
- Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient transformers: A survey.
- Wachsmuth, H., Syed, S., and Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.
- Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoeybi, M., Dong, Y., Kuchaiev, O., Li, B., Xiao, C., Anandkumar, A., and Catanzaro, B. (2023). Shall we pretrain autoregressive language models with retrieval? a comprehensive study.
- Wang, X., Wang, J., Cao, W., Wang, K., Paturi, R., and Bergen, L. (2024). Birco: A benchmark of information retrieval tasks with complex objectives.
- Yu, F. (2024). Why vector search didn’t work for your rag solution?
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., and Cui, B. (2024). Retrieval-augmented generation for ai-generated content: A survey.