Advanced Language Generation

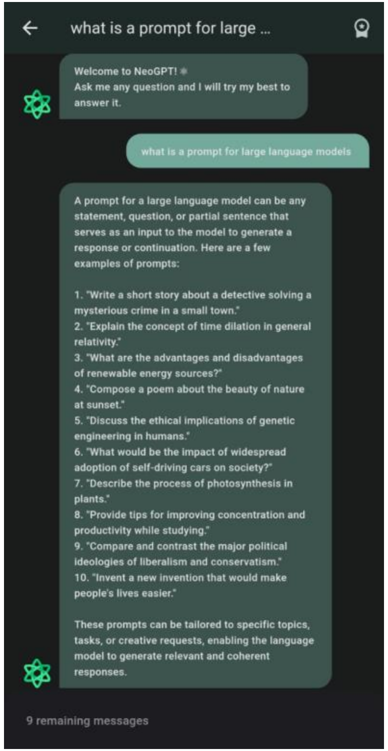
OUTLINE

- Introduction to Prompt Engineering
- Prompt engineering techniques:
    - Zero-shot prompting
    - Few-shot prompting
    - COT prompting
- Retrieval Augmented Generation (RAG)
    - How does it work?
    - Training RAG models

Intro to Prompt Engineering

What is a ‘Prompt’?

- LLMs Prompts are the textual inputs (e.g., questions, instructions) that you enter into your LLM to get responses.
- The model predicts an appropriate response to the prompt you entered. In general, a more specific and carefully worded prompt will get you better responses.



Welcome to NeoGPT!
Ask me any question and I will try my best to answer it.

what is a prompt for large language models

A prompt for a large language model can be any statement, question, or partial sentence that serves as an input to the model to generate a response or continuation. Here are a few examples of prompts:

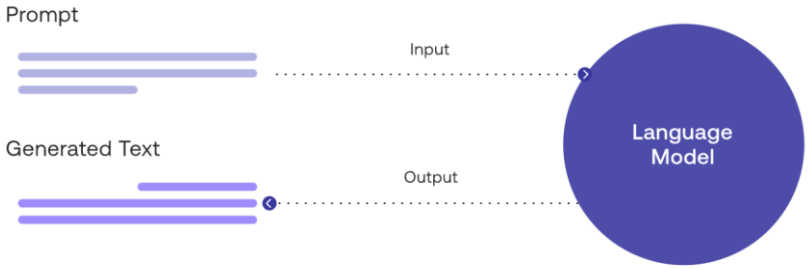
1. "Write a short story about a detective solving a mysterious crime in a small town."
2. "Explain the concept of time dilation in general relativity."
3. "What are the advantages and disadvantages of renewable energy sources?"
4. "Compose a poem about the beauty of nature at sunset."
5. "Discuss the ethical implications of genetic engineering in humans."
6. "What would be the impact of widespread adoption of self-driving cars on society?"
7. "Describe the process of photosynthesis in plants."
8. "Provide tips for improving concentration and productivity while studying."
9. "Compare and contrast the major political ideologies of liberalism and conservatism."
10. "Invent a new invention that would make people's lives easier."

These prompts can be tailored to specific topics, tasks, or creative requests, enabling the language model to generate relevant and coherent responses.

Elements of a Prompt

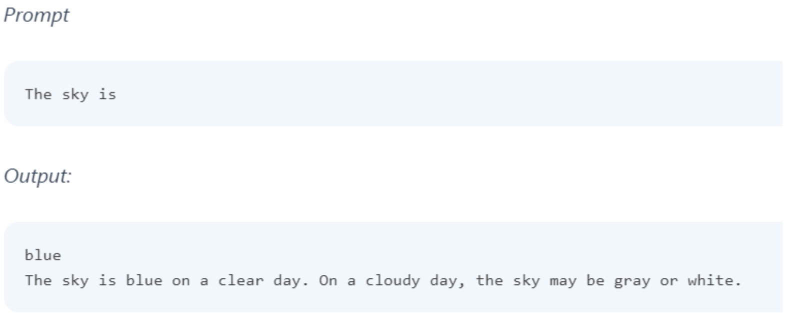
A prompt can contain any of the following components:

- **Instruction** - a specific task or instruction you want the model to perform
- **Context** - can involve additional context that can steer the model to better responses
- **Input Data** - is the question that we are interested in finding a response for
- **Output Indicator** - indicates the type or format of the output.



Prompt Engineering

- You can achieve a lot with simple prompts, but the quality of results depends on how much information you provide it and how well-crafted it is.
- In the context of LLMs, Prompt Engineering is the process of structuring text that can be interpreted and understood by the model.

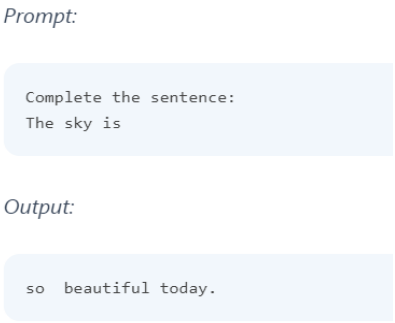


Prompt

```
The sky is
```

Output:

```
blue
The sky is blue on a clear day. On a cloudy day, the sky may be gray or white.
```



Prompt:

```
Complete the sentence:
The sky is
```

Output:

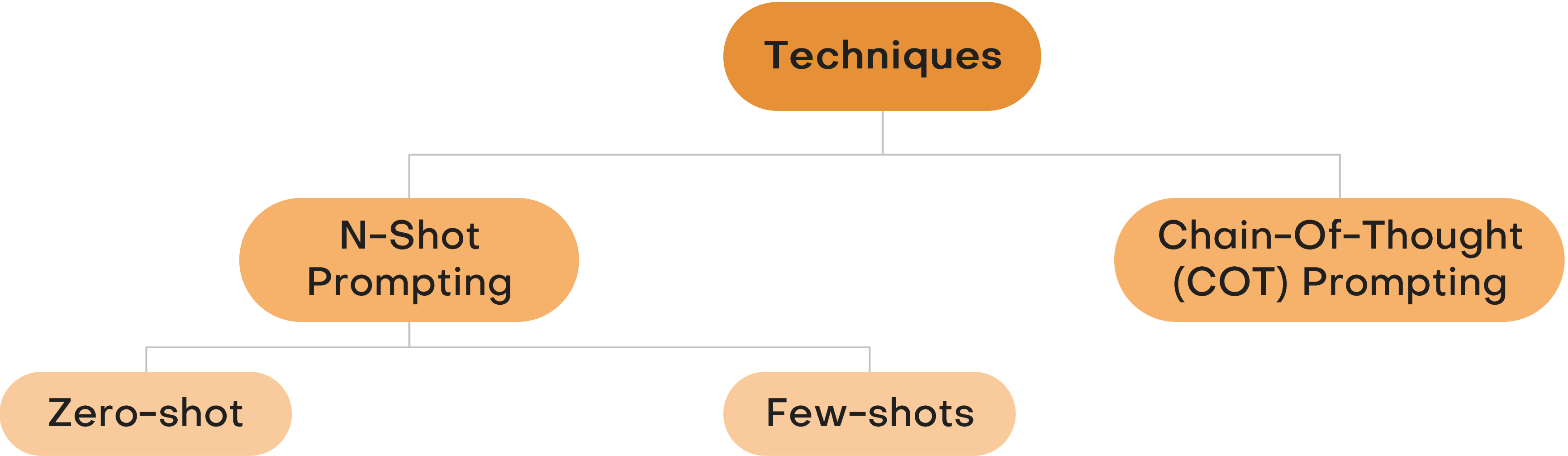
```
so  beautiful today.
```

Prompt Engineering Techniques

General Tips for designing Prompts

- Start simple
- Provide context
- Define a clear task & avoid open-ended questions
- Use Keywords
- Iterate

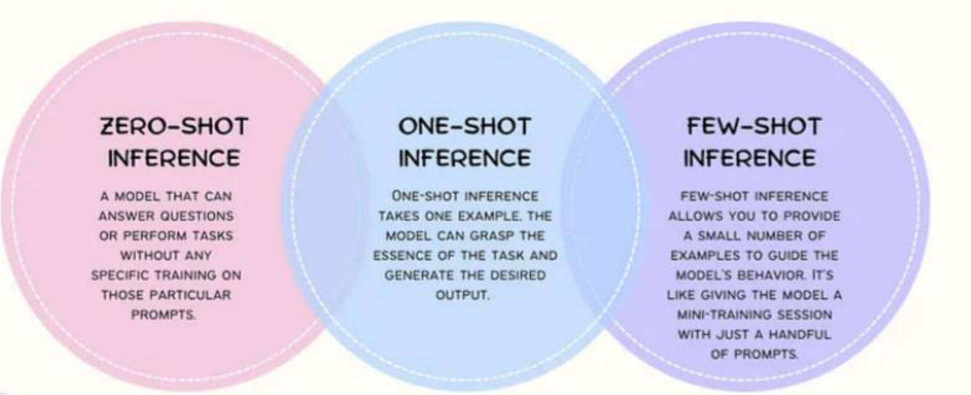
Prompt Engineering Techniques



- Techniques
  - N-Shot Prompting
    - Zero-shot
    - Few-shots
  - Chain-Of-Thought (COT) Prompting

Zero-shot Prompting

- LLMs are trained on LARGE amounts of data
- Don’t provide the model with any examples
- When zero-shot doesn’t work => provide examples



ZERO-SHOT INFERENCE

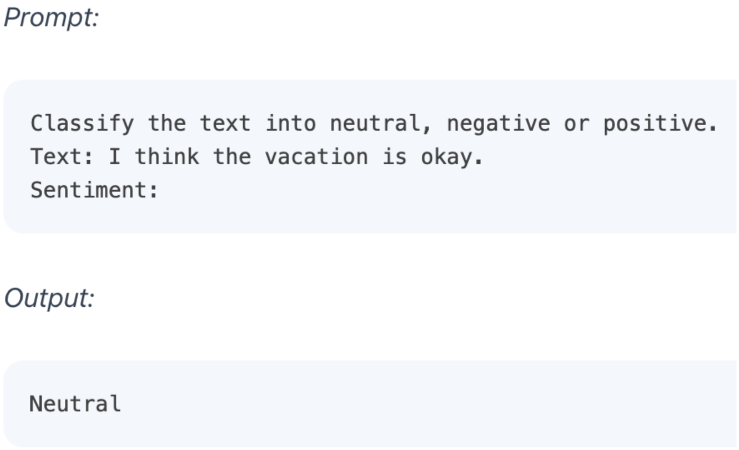
A MODEL THAT CAN ANSWER QUESTIONS OR PERFORM TASKS WITHOUT ANY SPECIFIC TRAINING ON THOSE PARTICULAR PROMPTS.

ONE-SHOT INFERENCE

ONE-SHOT INFERENCE TAKES ONE EXAMPLE. THE MODEL CAN GRASP THE ESSENCE OF THE TASK AND GENERATE THE DESIRED OUTPUT.

FEW-SHOT INFERENCE

FEW-SHOT INFERENCE ALLOWS YOU TO PROVIDE A SMALL NUMBER OF EXAMPLES TO GUIDE THE MODEL'S BEHAVIOR. IT'S LIKE GIVING THE MODEL A MINI-TRAINING SESSION WITH JUST A HANDFUL OF PROMPTS.



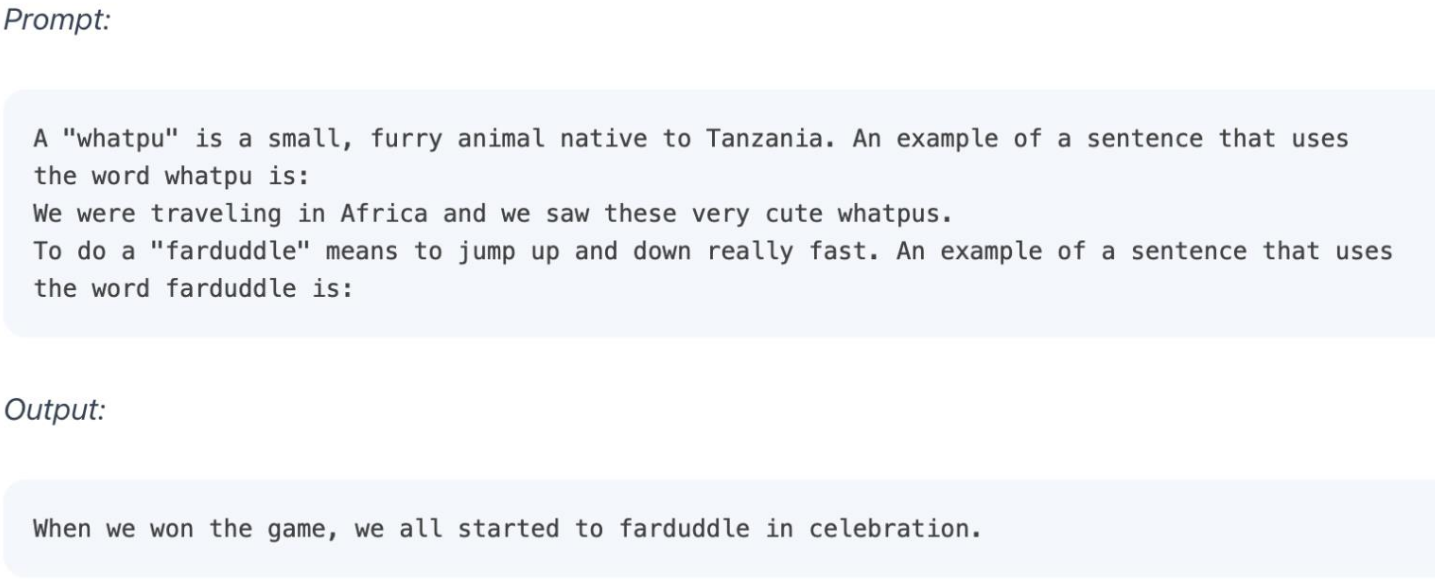
Prompt:

```
Classify the text into neutral, negative or positive.
Text: I think the vacation is okay.
Sentiment:
```

Output:

```
Neutral
```

Few-shot Prompting



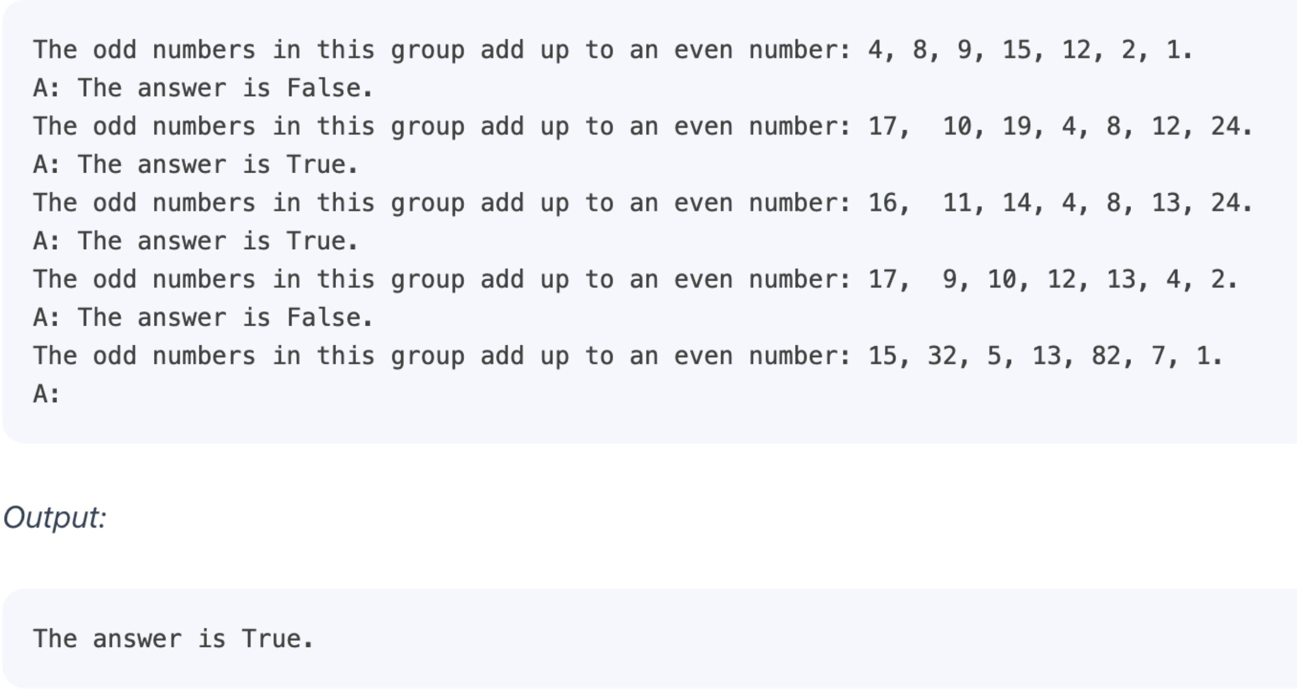
Prompt:

```
A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses
the word farduddle is:
```

Output:

```
When we won the game, we all started to farduddle in celebration.
```

Limitations of Few-shot Prompting



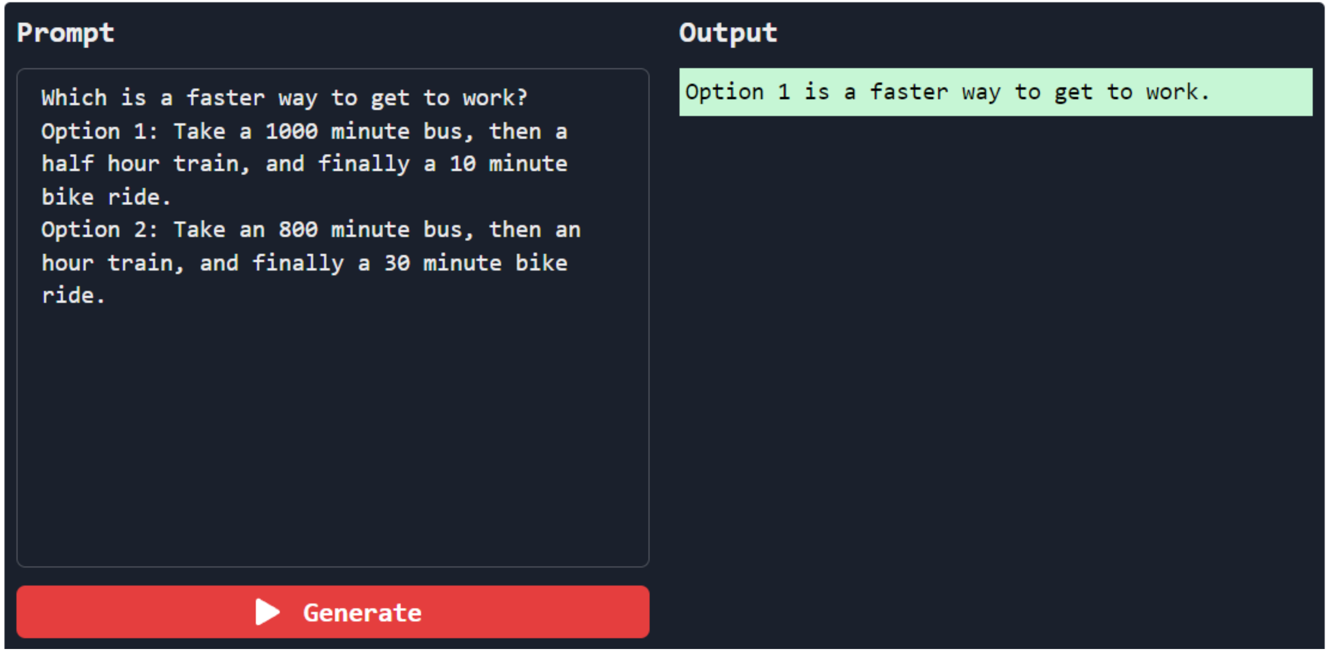
```
The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.
A: The answer is False.
The odd numbers in this group add up to an even number: 17,  10, 19, 4, 8, 12, 24.
A: The answer is True.
The odd numbers in this group add up to an even number: 16,  11, 14, 4, 8, 13, 24.
A: The answer is True.
The odd numbers in this group add up to an even number: 17,  9, 10, 12, 13, 4, 2.
A: The answer is False.
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:
```

Output:

```
The answer is True.
```

Chain-of-Thought Prompting

Incorrect

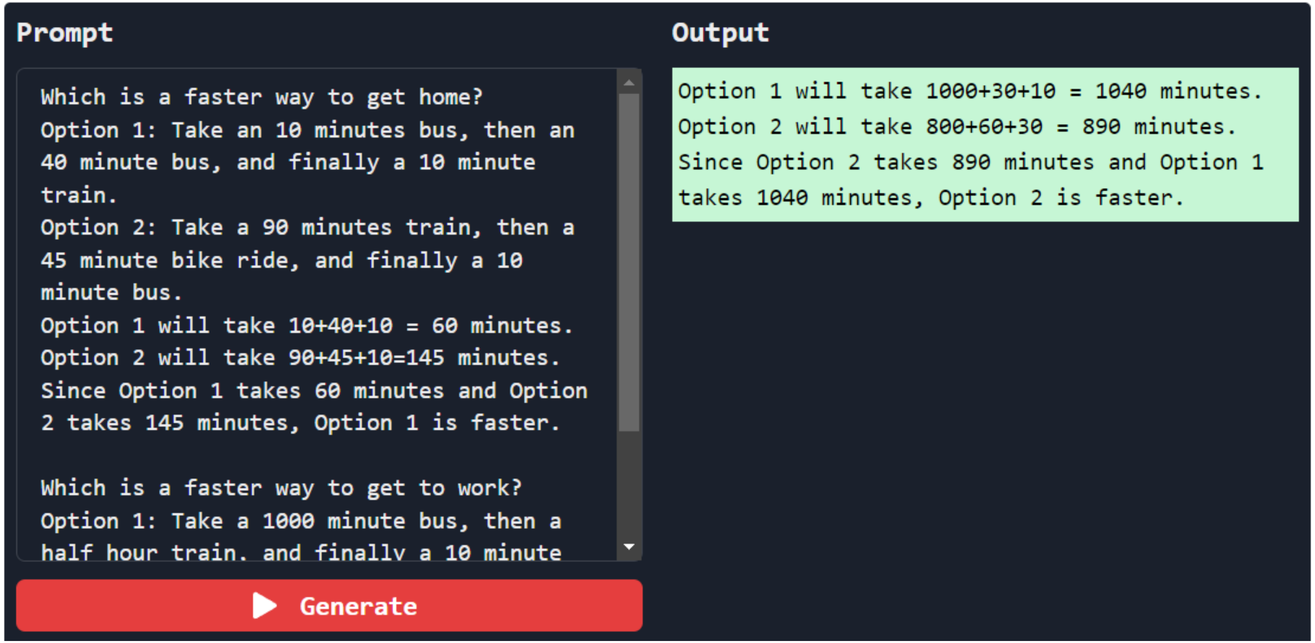


```
Which is a faster way to get to work?
Option 1: Take a 1000 minute bus, then a
half hour train, and finally a 10 minute
bike ride.
Option 2: Take an 800 minute bus, then an
hour train, and finally a 30 minute bike
ride.
```

```
Option 1 is a faster way to get to work.
```

Chain-of-Thought Prompting

Correct



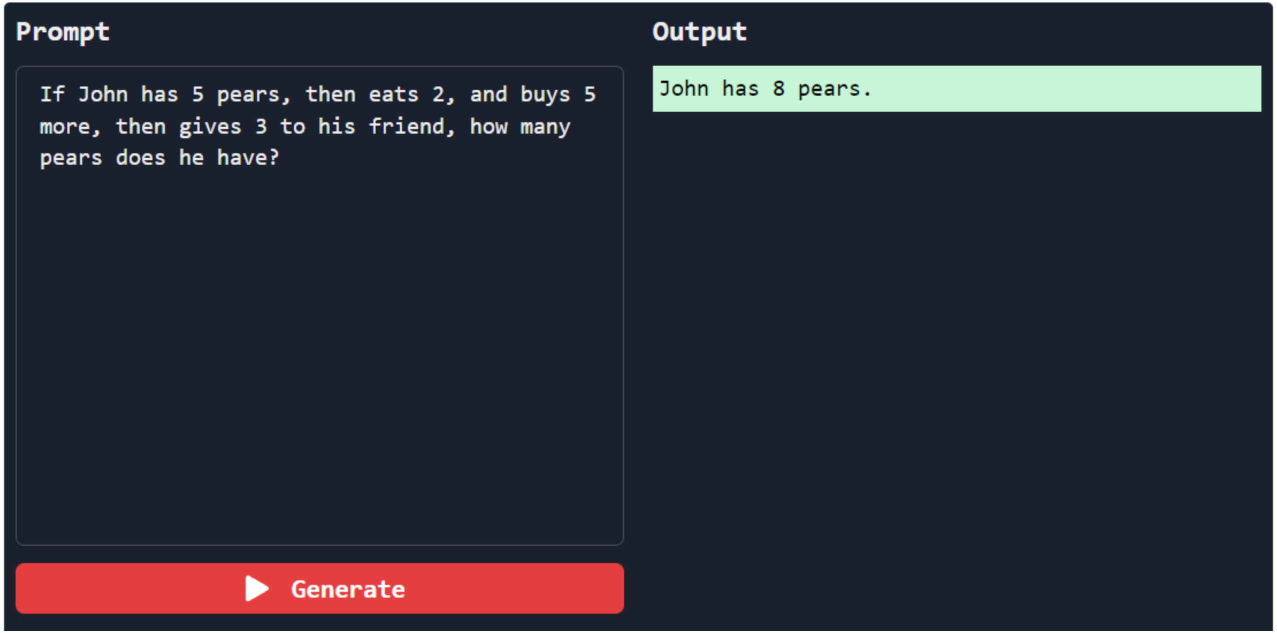
```
Which is a faster way to get home?
Option 1: Take an 10 minutes bus, then an
40 minute bus, and finally a 10 minute
train.
Option 2: Take a 90 minutes train, then a
45 minute bike ride, and finally a 10
minute bus.
Option 1 will take 10+40+10 = 60 minutes.
Option 2 will take 90+45+10=145 minutes.
Since Option 1 takes 60 minutes and Option
2 takes 145 minutes, Option 1 is faster.

Which is a faster way to get to work?
Option 1: Take a 1000 minute bus, then a
half hour train, and finally a 10 minute
```

```
Option 1 will take 1000+30+10 = 1040 minutes.
Option 2 will take 800+60+30 = 890 minutes.
Since Option 2 takes 890 minutes and Option 1
takes 1040 minutes, Option 2 is faster.
```

Zero-shot COT Prompting

Incorrect

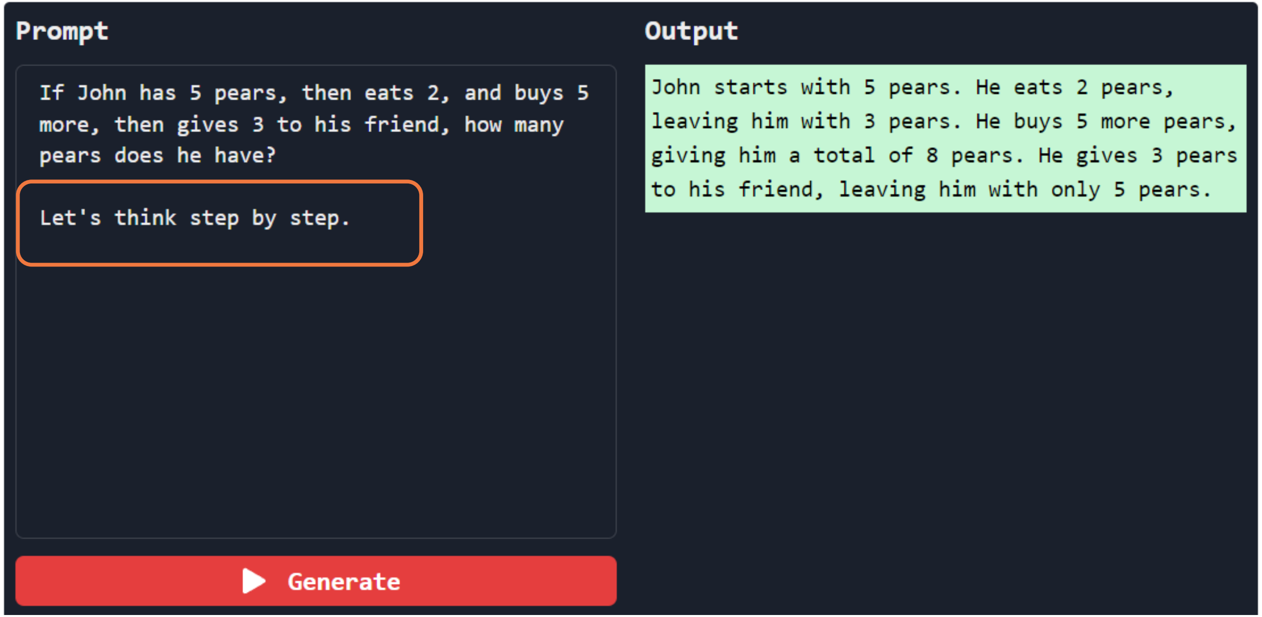


```
If John has 5 pears, then eats 2, and buys 5
more, then gives 3 to his friend, how many
pears does he have?
```

```
John has 8 pears.
```

Zero-shot COT Prompting

Correct



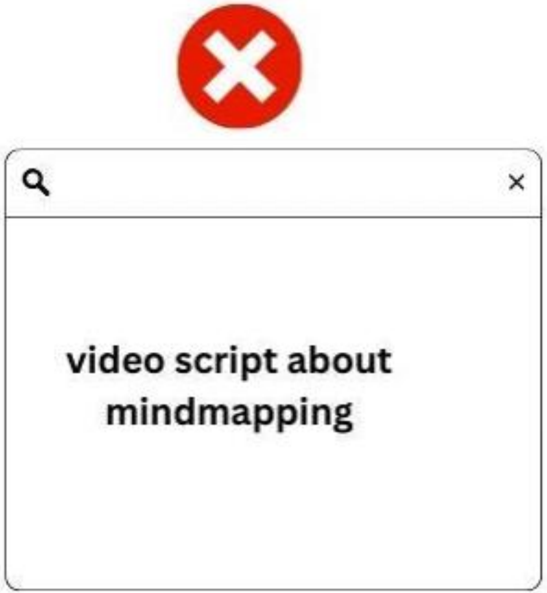
If John has 5 pears, then eats 2, and buys 5 more, then gives 3 to his friend, how many pears does he have?

Let's think step by step.

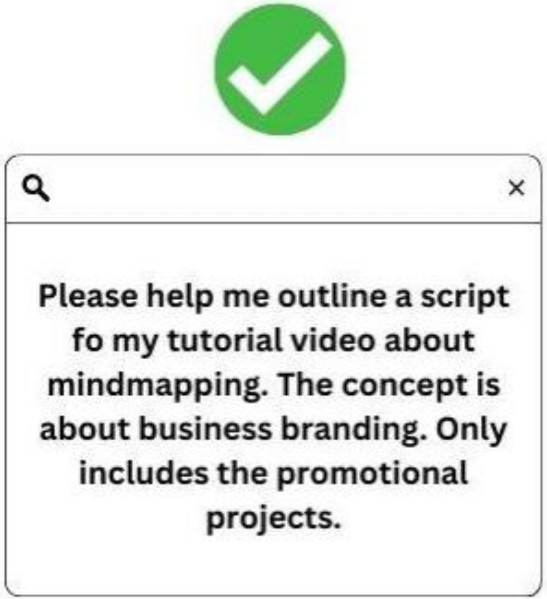
John starts with 5 pears. He eats 2 pears, leaving him with 3 pears. He buys 5 more pears, giving him a total of 8 pears. He gives 3 pears to his friend, leaving him with only 5 pears.

Effective Prompting

Prompt Generator Apps

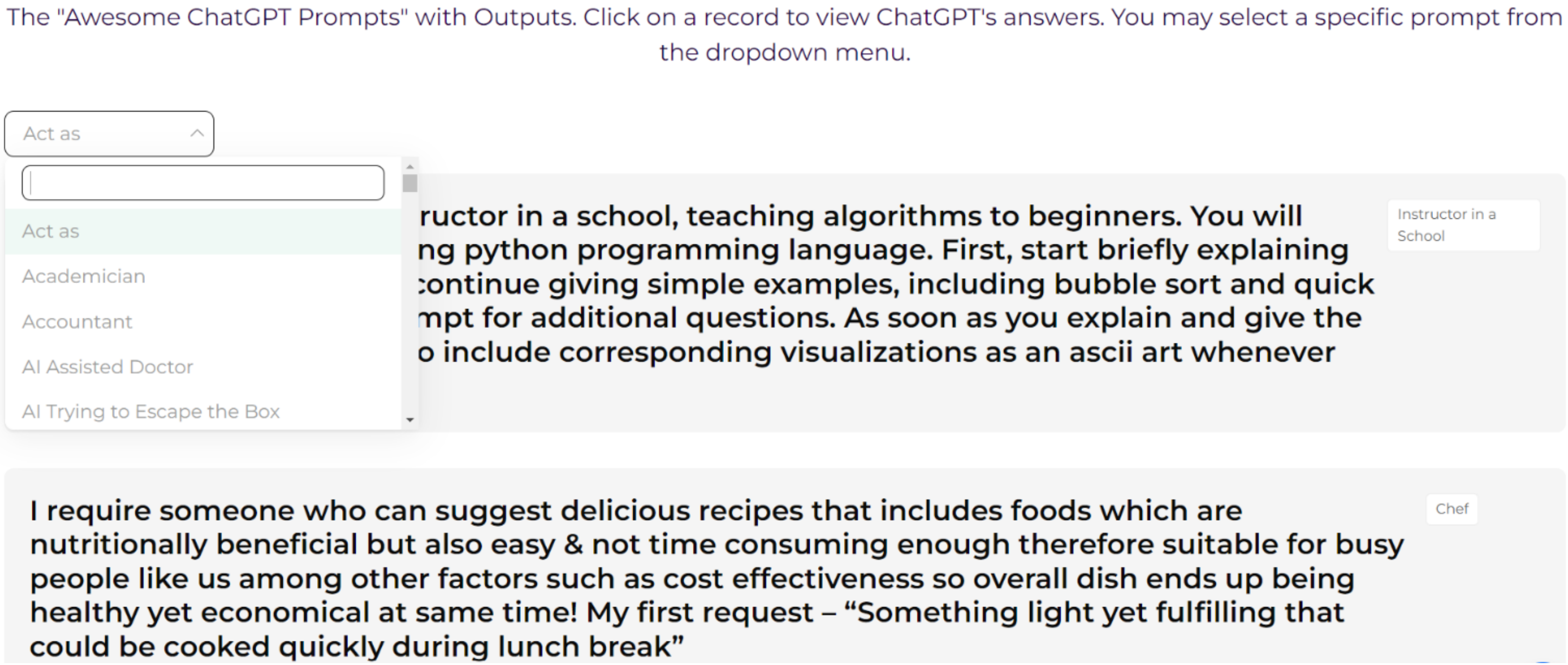

video script about mindmapping

SIMPLE PROMPT


Please help me outline a script fo my tutorial video about mindmapping. The concept is about business branding. Only includes the promotional projects.

AI-GENERATED PROMPT

Act-as Prompts



The "Awesome ChatGPT Prompts" with Outputs. Click on a record to view ChatGPT's answers. You may select a specific prompt from the dropdown menu.

ructor in a school, teaching algorithms to beginners. You will
ng python programming language. First, start briefly explaining
continue giving simple examples, including bubble sort and quick
mpt for additional questions. As soon as you explain and give the
o include corresponding visualizations as an ascii art whenever

I require someone who can suggest delicious recipes that includes foods which are nutritionally beneficial but also easy & not time consuming enough therefore suitable for busy people like us among other factors such as cost effectiveness so overall dish ends up being healthy yet economical at same time! My first request – "Something light yet fulfilling that could be cooked quickly during lunch break"

https://www.turbo-gpt.com/chatgpt-prompts-act-as

General Prompting Tips

- **Expert Simulator**
  "Act as a seasoned data scientist and guide me in writing a report that highlights the latest trends and predictions in the field of artificial intelligence. Provide me with unique insights and expert advice that will help me deliver high-quality, informative content to my audience."

- **Unconventional Prompts**
  "Write a poem about content creation" or "Describe feeling like an digital marketer in 10 adjectives."

General Prompting Tips

- **Experiment With Styles & Tone**

*"Give **[insert style]** advice you can about using LLMs to create more effective content in a **[insert tone]**."*

Styles		Tones	
Descriptive	Informal	Humorous	Confident
Persuasive	Formal	Satirical	Doubtful
Narrative	Romantic	Serious	Optimistic
Technical	Gothic	Angry	Pessimistic

Demo: Effective Prompting with ChatGPT

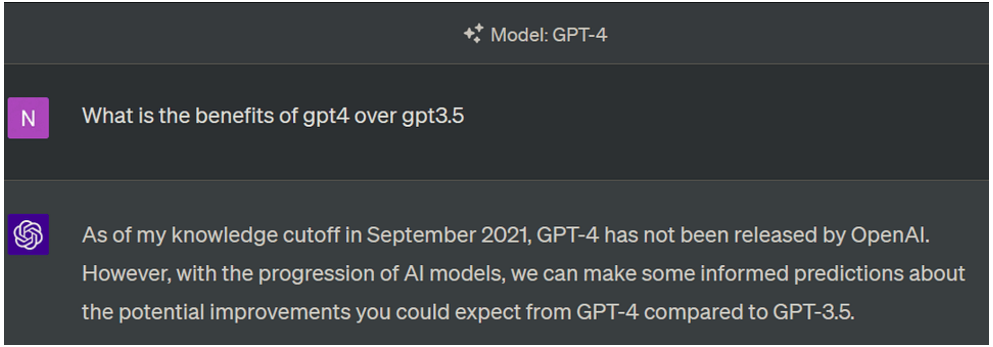
Introduction to Retrieval Augmented Generation

Today’s State of the Art LLMs

Current pre-trained LLMs such as GPT, Bart and T5 excel at various NLP tasks:

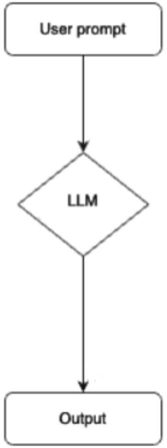
- They possess a wealth of world knowledge in their parameters
- Can easily handle a wide variety of tasks
- Can be fine-tuned to specific tasks

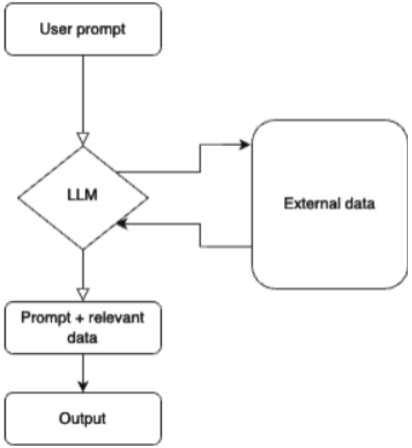
RAG vs. LLMs



What is the benefits of gpt4 over gpt3.5

As of my knowledge cutoff in September 2021, GPT-4 has not been released by OpenAI. However, with the progression of AI models, we can make some informed predictions about the potential improvements you could expect from GPT-4 compared to GPT-3.5.





Why RAG?

Retrieval Augmented Generation is a NLP technique created to **address** some of the **limitations** of LLMs:

- Generating incorrect information (Hallucinations)
- Leaking sensitive information
- Precise knowledge application
- Difficult to update to current world knowledge

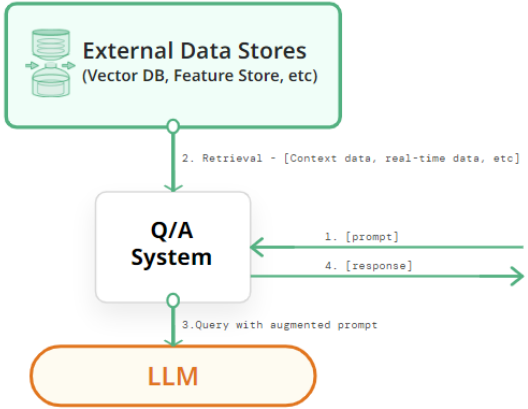
Retrieval Augmented Generation

What is a RAG model?

A Retrieval Augmented Generation model combines the generation aspect of LLMs and Information Retrieval Techniques to address limitations:

- The model **retrieves** and **generates** simultaneously
- **No labels required** to retrieve information
- Resistant to information leak
- Is a **general solution** for any seq2seq task

However, it is still under development.


External Data Stores
(Vector DB, Feature Store, etc)
2. Retrieval - [Context data, real-time data, etc]
Q/A System
1. [prompt]
4. [response]
User
3.Query with augmented prompt
LLM

Examples of RAG-like models

einstein

alexa

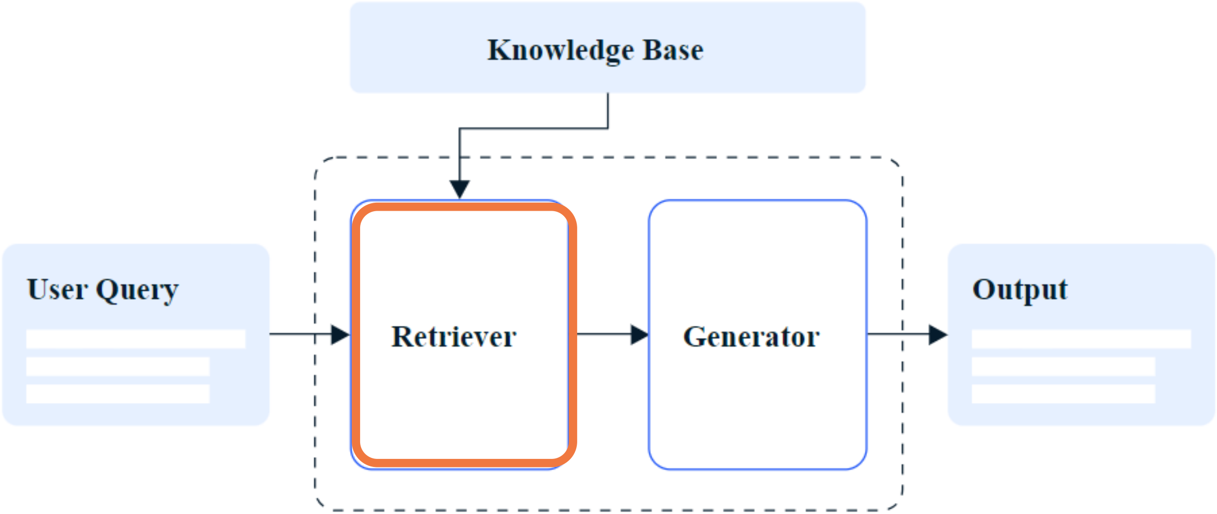
Google Assistant

Components of a RAG model

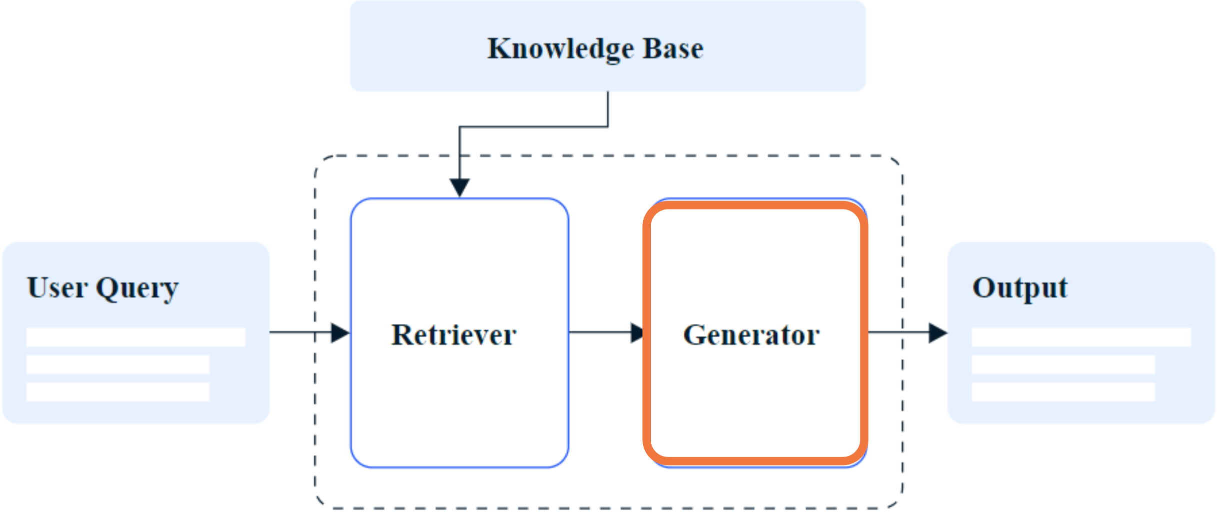
RAG models consist of 3 components:

- A pre-trained Language Model (e.g. GPT, BART) - Generator
- A pre-trained Retrieval Model (e.g. DPR, ICT)
- A Knowledge Base of text data (e.g. Wikipedia, YAGO, tweets)

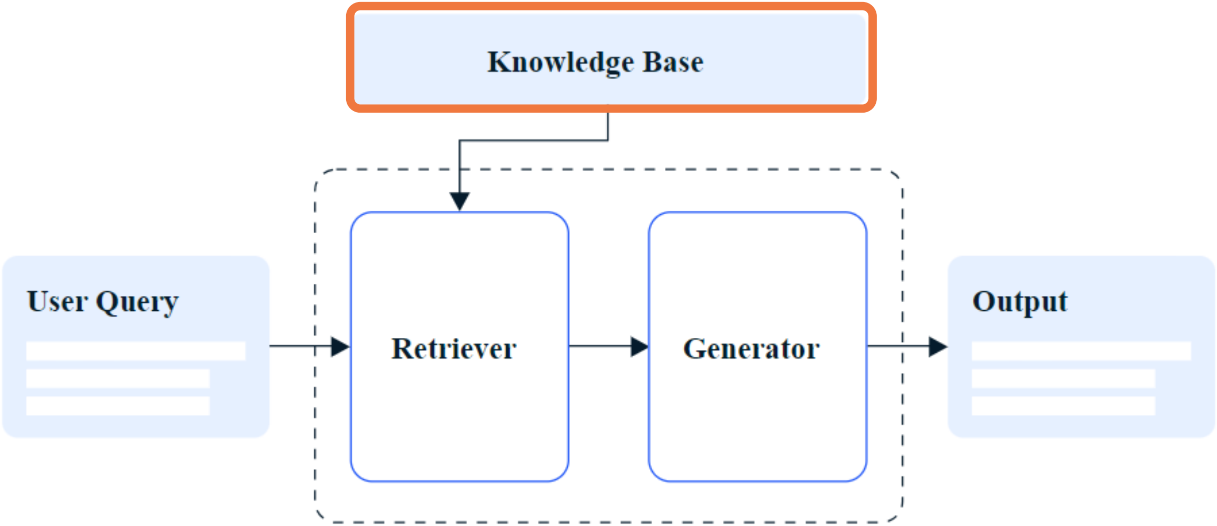
The Retriever



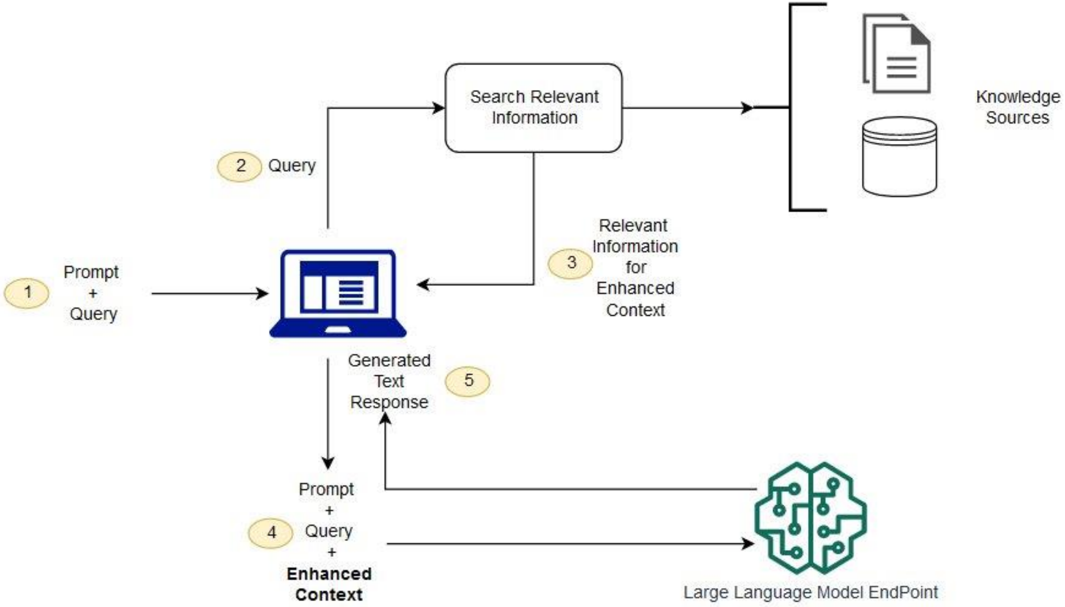
The Generator



The Knowledge Base



RAG Architecture Visually


Search Relevant Information
Knowledge Sources
2 Query
Relevant Information for Enhanced Context
3
1
Prompt + Query
Generated Text Response
5
Prompt + Query + **Enhanced Context**
4
Large Language Model EndPoint

https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html

RAGs excel at knowledge-intensive tasks

RAGs provide **State of the Art** results for general QA

- Combine strengths of parametric and non-parametric memory
  - Parametric memory is the pretrained seq2seq model
  - Non-parametric is the vector index from the KB accessed by the retrieval model
- More factual and precise information than seq2seq only

Input: Which planet in the solar system has the most moons?

GPT-3: As of my last knowledge update in September 2021, Jupiter, the largest planet in our solar system, had the most moons. Jupiter has over 80 known moons.

RAG: Saturn is the planet in our solar system with the most moons, with a total of 146 known moons.

Pros and Cons

Information Retrieval existed before LLMs, and has great advantages:

- Allows precise and accurate results
- Easy to update
- Some newer retrieval models are getting excellent results

However, it does have disadvantages:

- IR needs labeled data for training
- Needs to be programmed in specific ways

Use-Cases of RAG Models

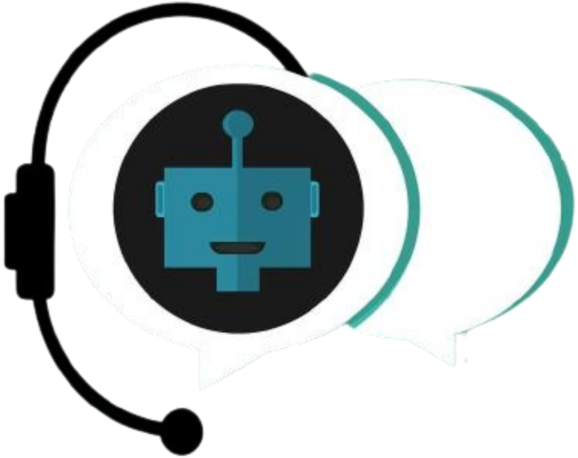
Lots of Opportunities

As mentioned before, RAG models are still under development

- Current products still do not use this technology
- Many use-cases have been predicted
- There are challenges when using them commercially

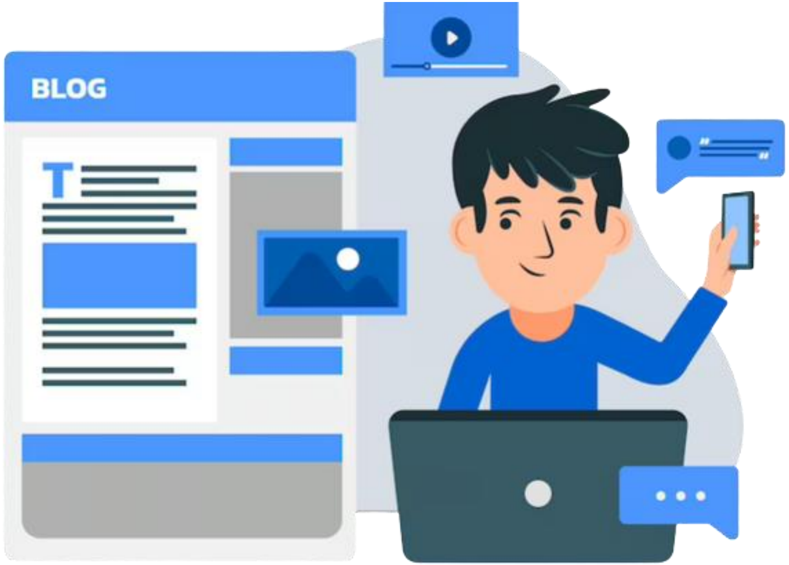
Customer Service Chatbots

- Enhanced responses
- Customized Knowledge Base
- Use updated customer reviews



Content Generation Tools

- Easier high-quality content creation
  - Blog posts
  - Marketing copies
  - Product descriptions
- All based on updated and factual information



Product Recommendation Systems

- Improved recommendations to customers
- By using the most recent information
    - Purchase history
    - Interests
    - Preferences

Medical Diagnosis Tools

- Precise and accurate diagnoses
- Generate a list of possible diagnoses based on:
  - Current symptoms
  - Medical history

Education

RAGS can enhance education and learning experience by generating personalized learning materials based on:

- Constantly updated data from students
- Specific needs and interests of each student

Current Challenges with RAG Models

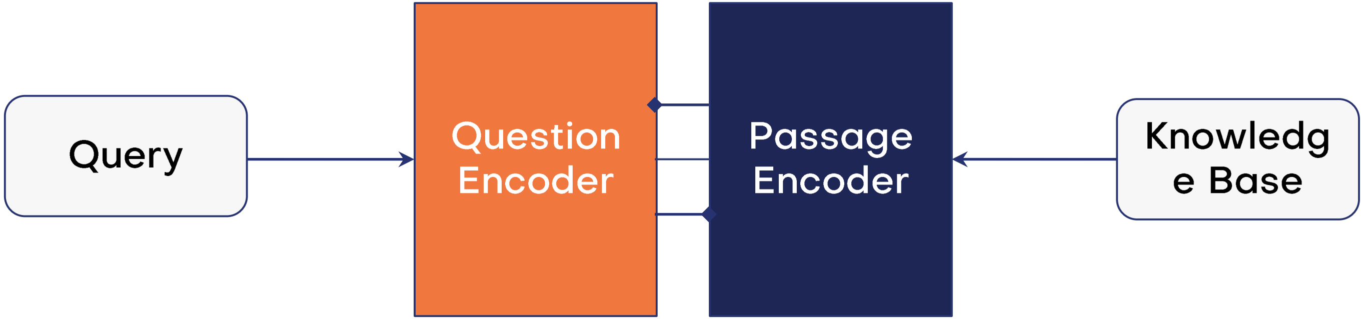
- Companies and people still struggle integrating their own private contexts
- Ethical usage of these models is a big concern
- Context size and relevance
- Response length and model constraints

Dense Passage Retrieval

The Retriever: Dense Passage Retrieval

DPR is a technique that is great for open-domain question answering:

- Used to compute dense vector representations of text
- Stores the results in an indexed manner
- Uses a Bi-directional (also called Siamese Two-Tower) model
- DPR retrieves the most relevant passages from the index base



Why DPR?

DPR has several advantages compared to traditional Information Retrieval:

- Great at capturing semantic similarities between text passages
- By precomputing and indexing, DPR gets faster retrieval time
- Dense vectors are rich with information

Open Domain Question Answering?

- ODQA is a type of task that differs from requesting straightforward and objective answers from a model.
- The questions can be from any domain
- The model doesn't have special training for specific topics.
- By measuring how often your model gets the answers right, you evaluate its performance in ODQA

Decoding Strategies

Two Ways to Decode

There are currently two ways to decode the answers from RAG models:

- The RAG Token technique
- The RAG Sequence technique

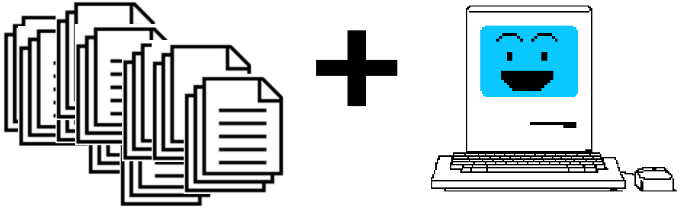
Both achieve very similar results!

- In DPR, you get the Z documents fetched from the KB
- Then, you figure out how the LLM generated its tokens based on the user’s input, and also the previously generated tokens.
- In short, both ways are summing up the documents and comparing.

RAG Token Decoding

In the RAG Token way:

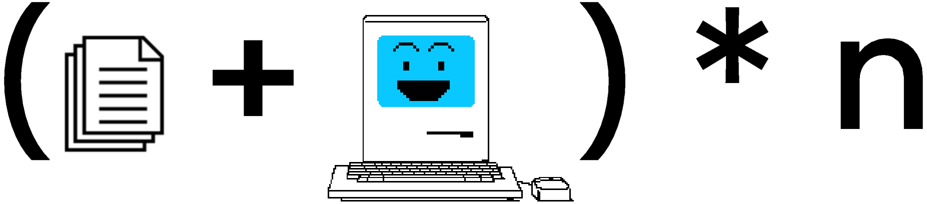
- When formulating an answer, it treats each token (or word) in the answer as a separate decision.
- For each token, it re-computes the attention over retrieved documents.
- It's analogous to the "token-level" attention mechanism where attention is computed for each token in the sequence.
- This can potentially produce answers that synthesize information from various parts of the retrieved documents.



RAG Sequence Decoding

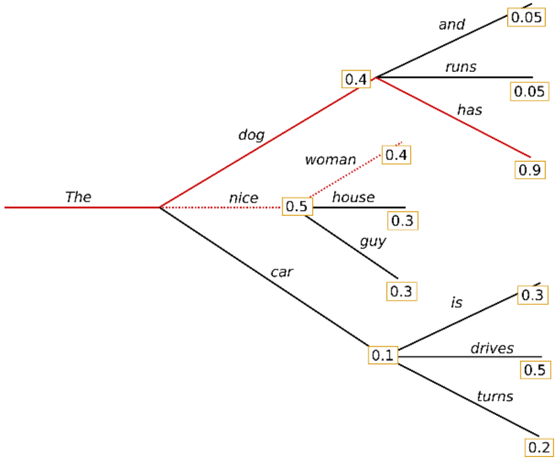
In the RAG Sequence way:

- The model takes one of the separated documents, and generates an entire sequence.
- Does this for every document
- Multiplies together the probabilities of the entire sequences, finding the most probable answer
- It's analogous to the "sequence-level" attention mechanism.



Beam Search

- Beam search is a technique used in decoding from these models
- It helps us explore different possibilities in generating tokens



شكراً لكم

Thank you

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority