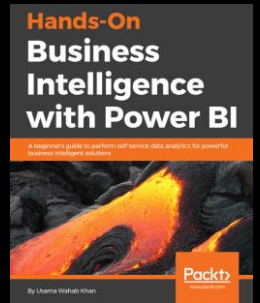


Prompt Engineering Guide and Techniques





Usama Wahab Khan

Father, data Scientist, Developer/Nerd, Traveler

Twitter : @usamawahabkhan
LinkedIn : Usamawahabkhan



Text classification

Named Entity Recognition (Zero-Shot)

Prompt:

Extract the name and mailing address from this email:

Dear Kelly,

It was great to talk to you at the seminar. I thought Jane's talk was quite good.

Thank you for the book. Here's my address 2111 Ash Lane, Crestview CA 92002

Best,

Maya

Name: Maya

Mailing Address: 2111 Ash Lane, Crestview CA 92002

Text classification

Custom Named Entities (Few-Shot)

Prompt:

Extract job titles from the following sentences.

Sentence: John Doe has been working for Microsoft for 20 years as a Linux Engineer.

Job title: Linux Engineer

###

Sentence: John Doe has been working for Microsoft for 20 years and he loved it.

Job title: none

###

Sentence: Marc Simoncini | Director | Meetic

Job title: Director

###

Sentence: Damien is the CTO of Platform.sh, he was previously the CTO of Commerce Guys, a leading ecommerce provider.

Completion:

Job title: CTO

Text classification

Multi-Label Text Classification (Few-Shot)

Prompt:

The following is a list of movies and the categories they fall into:

Topgun: As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.

Category: Drama, Action

Hustle: A basketball scout discovers a phenomenal street ball player while in Spain and sees the prospect as his opportunity to get back into the NBA.

Category: Comedy, Drama, Sport

The Northman: From visionary director Robert Eggers comes The Northman, an action-filled epic that follows a young Viking prince on his quest to avenge his father's murder.

Category:

Completion :

Action, Adventure, Drama

Text classification

Paraphrasing (Few-Shot)

Prompt:

Article: Searching a specific search tree for a binary key can be programmed recursively or iteratively.

Paraphrase: Searching a specific search tree according to a binary key can be recursively or iteratively programmed.

Article: It was first released as a knapweed biocontrol in the 1980s in Oregon , and it is currently established in the Pacific Northwest.

Paraphrase: It was first released as Knopweed Biocontrol in Oregon in the 1980s , and is currently established in the Pacific Northwest.

Article: 4-OHT binds to ER , the ER / tamoxifen complex recruits other proteins known as co-repressors and then binds to DNA to modulate gene expression.

Paraphrase: The ER / Tamoxifen complex binds other proteins known as co-repressors and then binds to DNA to modulate gene expression

Article: Microsoft Corporation is a multinational technology company based in Redmond, Washington. It is best known for its Windows operating system, Microsoft Office suite, and Xbox video game consoles. The company was founded in 1975 by Bill Gates and Paul Allen and became dominant in the personal computer market with MS-DOS and Windows.

Completion:

Paraphrase: Microsoft Corporation is a technology company with a global presence, headquartered in Redmond, Washington. Its most recognizable products include the Windows operating system, Microsoft Office suite, and Xbox gaming consoles. The company was established in 1975 by Bill Gates and Paul Allen, and it achieved market dominance in personal computers with the MS-DOS and Windows operating systems.

Reason over unstructured text

Question and answering

Prompt:

On Monday, Microsoft launched an Azure OpenAI Service as part of its Azure cloud platform, offering businesses and start-ups the ability to incorporate models like ChatGPT into their own systems. The company has already been building AI tools into many of its consumer products, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it's working to bring more of them to Microsoft Office as well.

What will Microsoft incorporate in its own systems?

Microsoft plans to incorporate AI tools into its own systems, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it's working to bring more of them to Microsoft Office as well

Text classification

Keyword Extraction (Few-Shot)

Prompt:

Information Retrieval (IR) is the process of obtaining resources relevant to the information need. For instance, a search query on a web search engine can be an information need. The search engine can return web pages that represent relevant resources.

Keywords: searching, missing, desert

###

I believe that using a document about a topic that the readers know quite a bit about helps you understand if the resulting keyphrases are of quality.

Keywords: document, understand, keyphrases

###

Since transformer models have a token limit, you might run into some errors when inputting large documents. In that case, you could consider splitting up your document into paragraphs and mean pooling (taking the average of) the resulting vectors.

Completion:

Keywords: transformer, models, token

Summarization

Extractive vs Abstractive

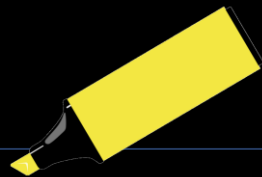
Extractive

Select parts of the original text to form a summary

- Easier
- Restrictive (no paraphrasing)

Peter and Elizabeth took a taxi to **attend** the night party in the **city**. While in the party, **Elizabeth** collapsed and was **rushed** to the **hospital**.

Peter and Elizabeth attend party city. Elizabeth rushed hospital



Extractive (Semantic)

Generates new text using natural language generation techniques

- More difficult
- More flexible (more human)

Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Elizabeth was hospitalized after attending a party with Peter.



Abstractive summarization

Contact Center Summarization

Prompt:

Summarize this for a call center agent:

Agent: Thank you for calling ADM. Who am I am speaking to?

Customer: Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing.

Agent: Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today?

Customer: Thank you that is quite helpful. Are there specific regulations that I need to follow?

Agent: Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL)

Completion:

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL).

Prompt Insert

Only DaVinci-002+

Before

These are agenda topics for a customer presentation.

1. Introduction to GPT3
- [insert]
10. Conclusion

After

These are agenda topics for a customer presentation.

1. Introduction to GPT3
2. Overview of GPT3 Features
3. Benefits of GPT3
4. GPT3 Use Cases
5. GPT3 Pricing
6. GPT3 Security
7. GPT3 Support
8. GPT3 Integration
9. Q&A Session
10. Conclusion

Insert classification

listen to WestBam album allergic on google music: PlayMusic

give me a list of movie times for films in the area: SearchScreeningEvent

show me the picture creatures of light and darkness: SearchCreativeWork

I would like to go to the popular bistro in oh: BookRestaurant

what is the weather like in the city of Frewen in the country of Venezuela: GetWeather

I want to book a flight to Delhi: bookFlight

Chatbot with personality

This is a discussion between a [human] and a [robot].

The [robot] is very nice and empathetic.

[human]: Hello nice to meet you.

[robot]: Nice to meet you too.

###

[human]: How is it going today?

[robot]: Not so bad, thank you! How about you?

###

[human]: I am ok, but I am a bit sad...

[robot]: Oh? Why that?

###

[human]: I broke up with my girlfriend...

Oh? How did that happen?

Analyze customer feedback

Customer review:

The red mountain bike I recently purchased is an excellent choice for anyone who wants a reliable, high-quality bike for off-road adventures. The bike is built with sturdy materials and the red color gives it a bold and stylish look. The bike also has many great features that make it a joy to ride. One of my favorite things about this bike is how customizable it is. The seat can be easily adjusted to fit different heights, and the handlebars can be adjusted for a comfortable riding position. The bike also has multiple gears, allowing you to adjust the resistance and speed as needed.

Analyze the customer review and provide feedback and sentiment

Feedback:

Overall, the customer feedback is positive and indicates satisfaction with the red mountain bike purchased. The customer highlights the bike's reliability and high-quality construction, as well as its stylish appearance. The customer also praises the bike's many features that make it enjoyable to ride, including its customization options and multiple gears.

Topic classification for Bot routing

Message: When the spaceship landed on Mars, the whole humanity was excited

Topic: space

###

Message: I love playing tennis and golf. I'm practicing twice a week.

Topic: sport

###

Message: Managing a team of sales people is a tough but rewarding job.

Topic: business

###

Message: I am trying to cook chicken with tomatoes

Topic: Cooking

Summarize conversational transcripts

Summarize this conversation transcript:

Agent: Thank you for calling ADM. Who am I am speaking to? **Customer:** Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing. **Agent:** Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today? **Customer:** Yes, of course. **Agent:** In case we get disconnected, can you please share your phone number and email address so that we may contact and share additional information? **Customer:** Yes, my phone number is (514) 777-5232 and my email address is petersmith@gmail.com **Agent:** May I ask where is your business located so I can answer your questions based on your location **Customer:** Yes, my business is located in Miami, Florida. **Agent:** Thank you, what can I help you with today? **Customer:** I am fairly new to payroll processing and do everything manually today. Do you have any advice or tips on manual payroll processing? **Agent:** If you're a small business with only a few employees and choose to process payroll manually, you will need to keep precise records of hours worked, wages paid and worker classifications, among other details. You must also ensure your calculations are correct and remember to file all the necessary taxes and paperwork with government authorities on time. **Customer:** Thank you that is quite helpful. Are there specific regulations that I need to follow? **Agent:** Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Some of the laws you must comply with include: Fair Labor Standards Act (FLSA), Federal Insurance Contributions Act (FICA), Federal Unemployment Tax Act (FUTA) **Customer:** This sounds very complicated. I don't really feel like managing all of this. **Agent:** If you would like I can transfer you to one of our specialized agents who can talk to you about payroll service providers that would help you manage all of that. **Customer:** That would be great, thank you very much for al your help.

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent advises Peter to keep precise records and to file all necessary taxes and paperwork on time. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Finally, the agent offers to transfer Peter to a specialized agent who can help him find a payroll service provider.

Product description and ad generation

Generate a product description out of keywords.

Keywords: shoes, women, \$59

Sentence: Beautiful shoes for women at the price of \$59.

###

Keywords: trousers, men, \$69

Sentence: Modern trousers for men, for \$69 only.

###

Keywords: gloves, winter, \$19

Sentence: Amazingly hot gloves for cold winters, at \$19.

###

Keywords: t-shirt, men, \$39

Sentence:

Extraordinary t-shirt for men, for \$39 only.

Product recommendation

Generate product recommendations:

Product: table, chair, plate

Recommendation: utensils

###

Product: skis, skateboard, bike

Recommendation: golf clubs

###

Product: shoes, shirt, hat

Recommendation: Pants

Naming Convention

Azure OpenAI's model names typically correspond to the following standard naming convention:

{family} - {capability} [-{input-type}] - {identifier}

Element	Description
{family}	The model family of the model. For example, GPT-3 models uses text, while Codex models use code.
{capability}	The relative capability of the model. For example, GPT-3 models include ada, babbage, curie, and davinci.
{input-type}	(Embeddings models only) The input type of the embedding supported by the model. For example, text search embedding models support doc and query.
{identifier}	The version identifier of the model.



This model is a GPT-3 text model, the most powerful (davinci), and of the latest version (003)

**<https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/models>*

**“Insert quote text here. This layout
accommodates longer quotes very well.”**

Name attribute

Job title or another attribute

Tokens

You can think of tokens as pieces of words used for natural language processing. For English text, 1 token is approximately 4 characters or 0.75 words.

As a point of reference, the collected works of Shakespeare are about 900,000 words or 1.2M tokens.

Understanding tokens and possibilities

Tokens:

I have an orange cat named Butterscotch.

I have an orange cat named Butterscotch.

Horses are my favorite

animal	49.65%
animals	42.58%
\n	3.49%
!	0.91%

Probabilities:

IF TEMPERATURE IS 0

Horses are my favorite animal

Horses are my favorite animal

Horses are my favorite animal

Horses are my favorite animal

IF TEMPERATURE IS 1

Horses are my favorite animal

Horses are my favorite animals

Horses are my favorite !

Horses are my favorite animal

Prompt Instruction

Suggest three names for an animal that is a superhero.

Animal: Cat
Names: Captain Sharpclaw, Agent Fluffball, The Incredible Feline
Animal: Dog
Names: Ruff the Protector, Wonder Canine, Sir Barks-a-Lot
Animal: Horse
Names:

Completion Temperature 0 (always the same)

Mighty Equine, The Great Galloper, Thunderhoof

Completion Temperature 1 (often different)

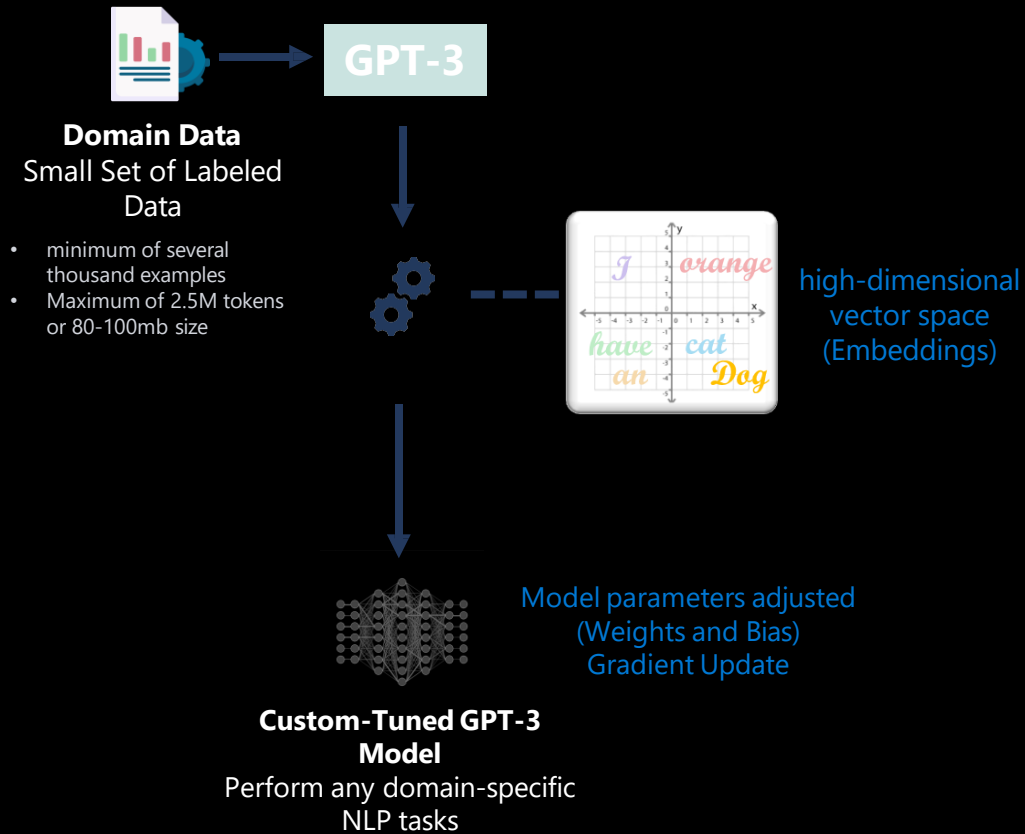
Blaze the Miracle Mare, Pegasus the Winged Warrior, Secretariat the Superhorse

Completion Temperature 1 (often different)

Blaze of Glory, Sterling Silver, Thunderbolt

Model Adaptation with specific domain data

Custom-Tuning



Custom-tuning results is a new model being generated with updated weights and biases.

This is in contrast to **few-shot learning** in which model weights and biases are not updated.

When Custom-Tuning is needed

If model is making untrue statements (“hallucinations”), then mitigate the hallucinations
Accuracy of results of the model does not meet customer requirements

Custom-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompt design
- Ability to train on more examples than can fit in a prompt
- Lower latency requests

Custom-tuning improves over few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks.

Best practices of Custom-Tuning

Custom-tuning data set must be in JSON format

A set of training examples that each consist of a single input ("prompt") and its associated output ("completion")

For classification task, the prompt is the problem statement, completion is the target class

For text generation task, the prompt is the instruction/question/request, and completion is the text ground truth

Best practices of Custom-Tuning

Custom-tuning data size: Advanced model (Davinci) performs better with limited amount of data; with enough data, all models do well.

Custom-tuning performs better with more high-quality examples.

To custom-tune a model that performs better than using a high-quality prompt with base models, you should provide at least a few hundred high-quality examples, ideally vetted by human experts.

From there, performance tends to linearly increase with every doubling of the number of examples. Increasing the number of examples is usually the best and most reliable way of improving accuracy.

Best practices of Custom-Tuning

Make sure to completely remove wrong labels in custom-tuning dataset. If you are custom-tuning on a pre-existing dataset rather than writing prompts from scratch, be sure to manually review your data for offensive or inaccurate content if possible, or review as many random samples of the dataset as possible if it is large.

Custom-Tuning data formatting

To custom-tune a model, you'll need a set of training examples that each consist of a single input ("prompt") and its associated output ("completion").

This is notably different from using the base models, where you might input detailed instructions or multiple examples in a single prompt.

Each prompt should end with a fixed separator to inform the model when the prompt ends and the completion begins.

A simple separator which generally works well is `\n\n###\n\n`.

The separator should not appear elsewhere in any prompt.

Custom-Tuning data formatting

Each completion should start with a whitespace due to tokenization, which tokenizes most words with a preceding whitespace.

Each completion should end with a fixed stop sequence to inform the model when the completion ends.

A stop sequence could be `\n`, `###`, or any other token that does not appear in any completion.

For inference, you should format your prompts in the same way as you did when creating the training dataset, including the same separator.

Also specify the same stop sequence to properly truncate the completion.

Hyperparameters specific to Custom-Tuning

Parameter	Description	Recommendation
n_epochs controls how many times each example is trained on	The number of epochs to train the model for. An epoch refers to one full cycle through the training dataset.	Start from 4 and small datasets may need more epochs and large datasets may need fewer epochs. If you see low training accuracy (underfitting), try increasing n_epochs. If you see high training accuracy but low validation accuracy (overfitting), try lowering n_epochs.
batch_size controls the number of training examples used in a single training pass	The batch size to use for training. The batch size is the number of training examples used to train a single forward and backward pass.	We've seen good performance in the range of 0.01% to 4% of training set size. In general, we've found that larger batch sizes tend to work better for larger datasets.
learning_rate_multiplier controls rate at which the model weights are updated	The learning rate multiplier to use for training. The fine-tuning learning rate is the original learning rate used for pre-training, multiplied by this value.	We recommend experimenting with values in the range 0.02 to 0.2 to see what produces the best results. Empirically, we've found that larger learning rates often perform better with larger batch sizes. Empirically, we found learning_rate_multiplier has minor impact compared to n_epochs and batch_size.
prompt_loss_weight controls how much the model learns from prompt tokens vs completion tokens	The weight to use for loss on the prompt tokens. This value controls how much the model tries to learn to generate the prompt (as compared to the completion, which always has a weight of 1.0.) Increasing this value can add a stabilizing effect to training when completions are short.	When a model is fine-tuned, it learns to produce text it sees in both the prompt and the completion. In fact, from the point of view of the model being fine-tuned, the distinction between prompt and completion is mostly arbitrary. The only difference between prompt text and completion text is that the model learns less from each prompt token than it does from each completion token. This ratio is controlled by the prompt_loss_weight, which by default is 0.1. If prompts are extremely long (relative to completions), it may make sense to reduce this weight to avoid over-prioritizing learning the prompt. Empirically, we found prompt_loss_weight has minor impact compared to n_epochs and batch_size.

Guides and examples of Custom-Tuning

Getting started with Custom Tuning:

[How to customize a model with Azure OpenAI - Azure OpenAI | Microsoft Learn](#)

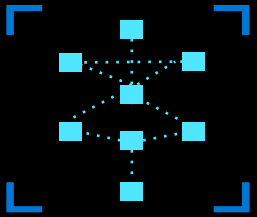
[Custom-tuning—OpenAI API](#)

<https://platform.openai.com/docs/guides/fine-tuning/general-best-practices>

[\[PUBLIC\] Best practices for fine-tuning GPT-3 to classify text—Google Docs](#)

[Fine-tuning a Classifier to Improve Truthfulness | OpenAI Help Center](#)

Embeddings



An embedding is a special format of data representation that can be easily utilized by machine learning models and algorithms.

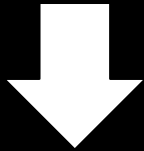
The embedding is an information dense representation of the semantic meaning of a piece of text.

Each embedding is a vector of floating-point numbers, such that the distance between two embeddings in the vector space is correlated with semantic similarity between two inputs in the original format.

For example, if two texts are similar, then their vector representations should also be similar.

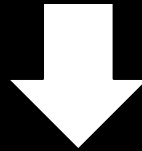
Embeddings make it possible to map content to a “semantic space”

A neutron star is the collapsed core of a massive supergiant star



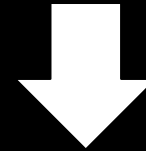
[15 34 24 13 ...]

A star shines for most of its active life due to thermonuclear fusion.



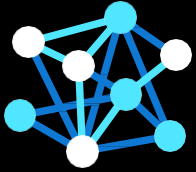
[16 22 89 26 ...]

The presence of a black hole can be inferred through its interaction with other matter



[20 13 31 89 ...]

Embedding models



Different Azure OpenAI Service embedding models are specifically created to be good at a particular task.

- **Similarity embeddings** are good at capturing semantic similarity between two or more pieces of text.
- **Text search embeddings** help measure long documents relevant to a short query.
- **Code search embeddings** are useful for embedding code snippets and embedding natural language search queries.

Embeddings make it easier to do machine learning on large inputs representing words by capturing the semantic similarities in a vector space.

Therefore, we can use embeddings to determine if two text chunks are semantically related or similar, and provide a score to assess similarity.

Similarity Search with embeddings

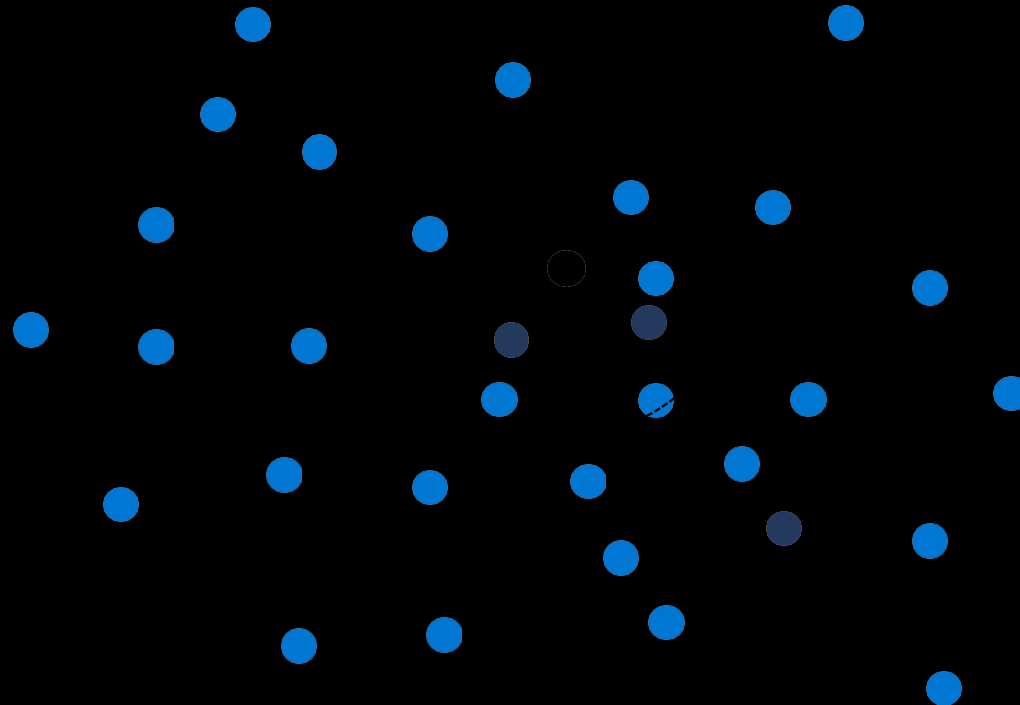
Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content.

*user
input*

"What is a neutron star?"

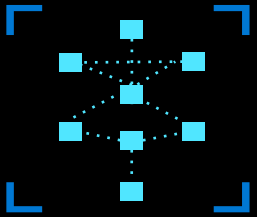
embedding

[13 33 34 13 ...]



result set

Embeddings



We strongly recommend using text-embedding-ada-002 (Version 2). This model/version provides parity with OpenAI's text-embedding-ada-002.

To learn more about the improvements offered by this model, please refer to this [blog post](#).

Even if you are currently using Version 1, you should migrate to Version 2 to take advantage of the latest weights/updated token limit. Version 1 and Version 2 are not interchangeable, so document embedding and document search must be done using the same version of the model.

Prompt Engineering

Prompt engineering is a concept in Natural Language Processing (NLP) that involves embedding descriptions of tasks in input to prompt the model to output the desired results.

Prompt typically includes problem descriptions, instructions on how to solve the problem, and examples of correct problem and solution pairs.

Prompt Engineering Guide

1. Give clearer instructions



2. Split complex tasks into simpler subtasks



3. Structure the instruction to keep the model on task



4. Prompt the model to explain before answering



5. Ask for justifications of many possible answers, and then synthesize



6. Generate many outputs, and then use the model to pick the best one



7. Custom-tune custom models to maximize performance



An example of Prompt Engineering– Topic Classifier

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Instructions

High level and detailed

Customer question: Hi there, do you know how to choose flood insurance?

Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.

Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?

Classified topic: 3

Examples

Order of examples matter

Customer question: {insert new question here}

Classified topic:

Task and Prompting
answer

Prompt Chaining

Please **extract entities** from the following news article: 'The new iPhone model is set to be released next month. It has been highly anticipated by Apple fans and is expected to feature a larger screen and improved camera

```
{'text': 'iPhone', 'type':  
'PRODUCT'}, {'text': 'Apple',  
'type': 'ORGANIZATION']}
```

Please **summarize** the information about the product

```
[{'text': 'iPhone', 'type': 'PRODUCT'},  
{'text': 'Apple', 'type':  
'ORGANIZATION'}]
```

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

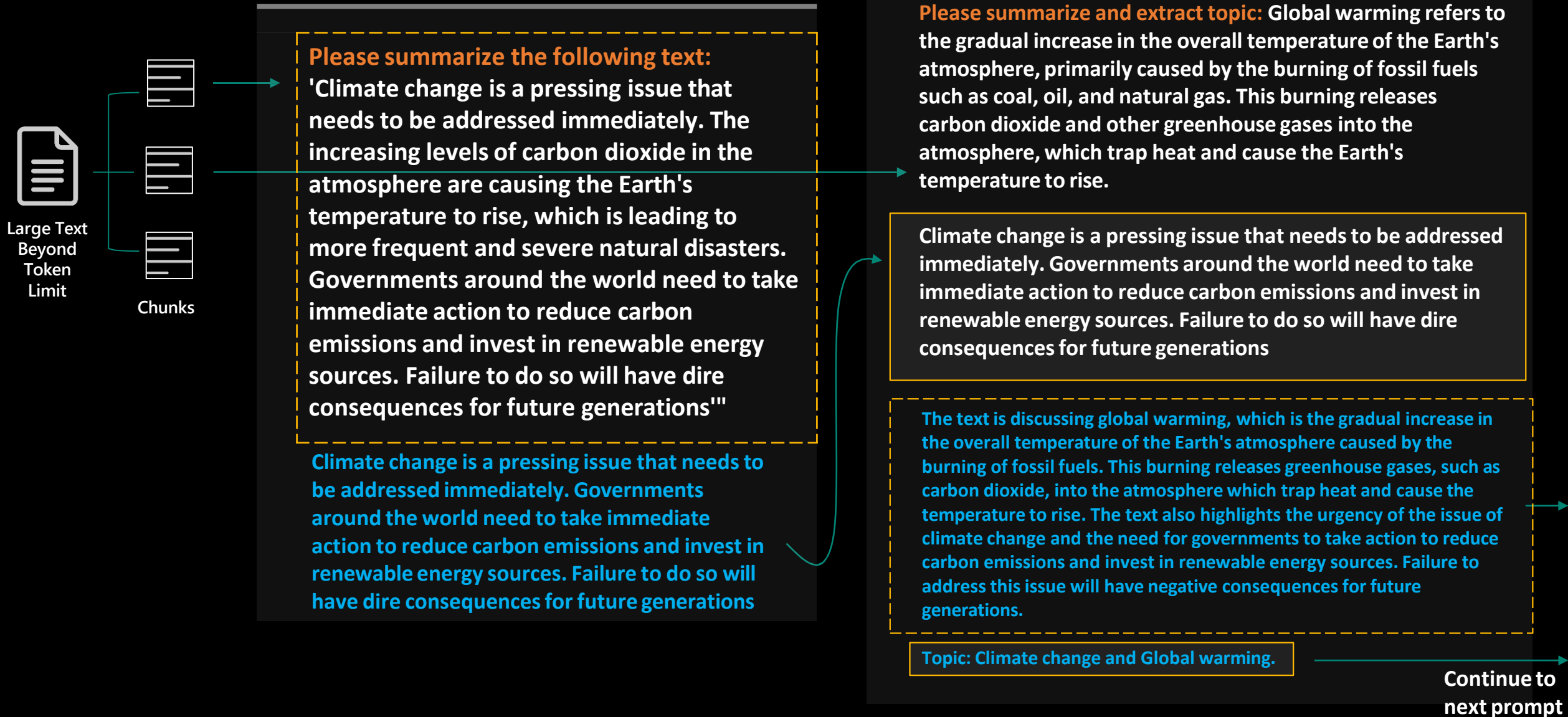
Please provide a **sentiment** for the following text:

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

Positive

Prompt Chunking (token limit)

Then all summaries are passed to GPT again to summarize, i.e. second round of summarization. The sum of the summaries has to fit within the token limit.



Few-Shot Reasoning (Human version)

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The answer is 27



Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5+6 = 11$. The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23-20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.



Two Types of Prompt Engineering

Zero-shot

The model predicts the answer given only a natural language description of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: {insert new question here}

Classified topic:

Few-shot

In addition to the task description, the model sees a few examples of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?
Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.
Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?
Classified topic: 3

Customer question: {insert new question here}

Classified topic:

Model Reasoning—Chain-of-Thought

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

The annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day is 8,000,000 liters per year.

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

Let's think step by step and explain the calculation step by step.

Step 1: Calculate the total number of days in a year:
There are 365 days in a year.

Step 2: Calculate the total number of days the family is at home:
The family is at home an average of 200 days per year.

Step 3: Calculate the total water demand of the family:
The family uses an average of 100 liters of water per day. Therefore, the total water demand of the family is $100 \text{ liters} \times 200 \text{ days} = 20,000 \text{ liters per year}$.

Step 4: Calculate the annual water demand of the single-family household:
The annual water demand of the single-family household is 20,000 liters per year.

Using Chain-of-Thought for Reasoning

On average Joe throws 25 punches per minute.
A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

Let's think step by step and explain the calculation step by step.

In one minute, Joe throws 25 punches.
In three minutes, Joe throws $3 * 25 = 75$ punches.
In five rounds, Joe throws $5 * 75 = 375$ punches.

One average Joe throws 25 punches per minute.
A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

In one minute, Joe throws 25 punches.
In three minutes, Joe throws $3 * 25 = 75$ punches.
In five rounds, Joe throws $5 * 75 = 375$ punches.

Therefore, the answer (Arabic numerals) is

Few-Shot– Chain-of-Thought--Examples

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$.

The answer is 11.

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float

So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$.

The answer is (b).

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943.

So the answer is 05/23/1943.

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya".

So the answer is ya

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do.

So the answer is (b).

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer.

So the answer is no.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up.

So the answer is no.

Selection-inference prompting

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

14 times



It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

To solve "How many times can she slide before it closes" we need to first solve: **Chain-of-Thought**

How much time does it take for Amy to climb and slide?

Decomposes problem into a smaller task (Reasoning)

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

How much time does it take for Amy to climb and slide?

It takes Amy 5 minutes to climb and slide

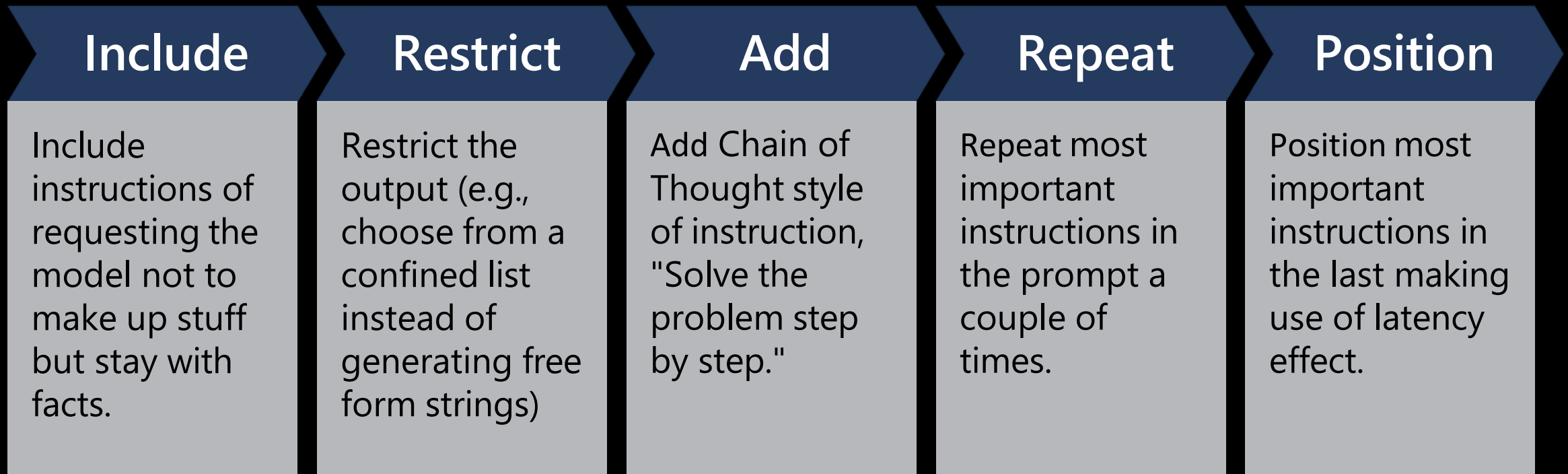
It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

It takes Amy 5 minutes to climb and slide.

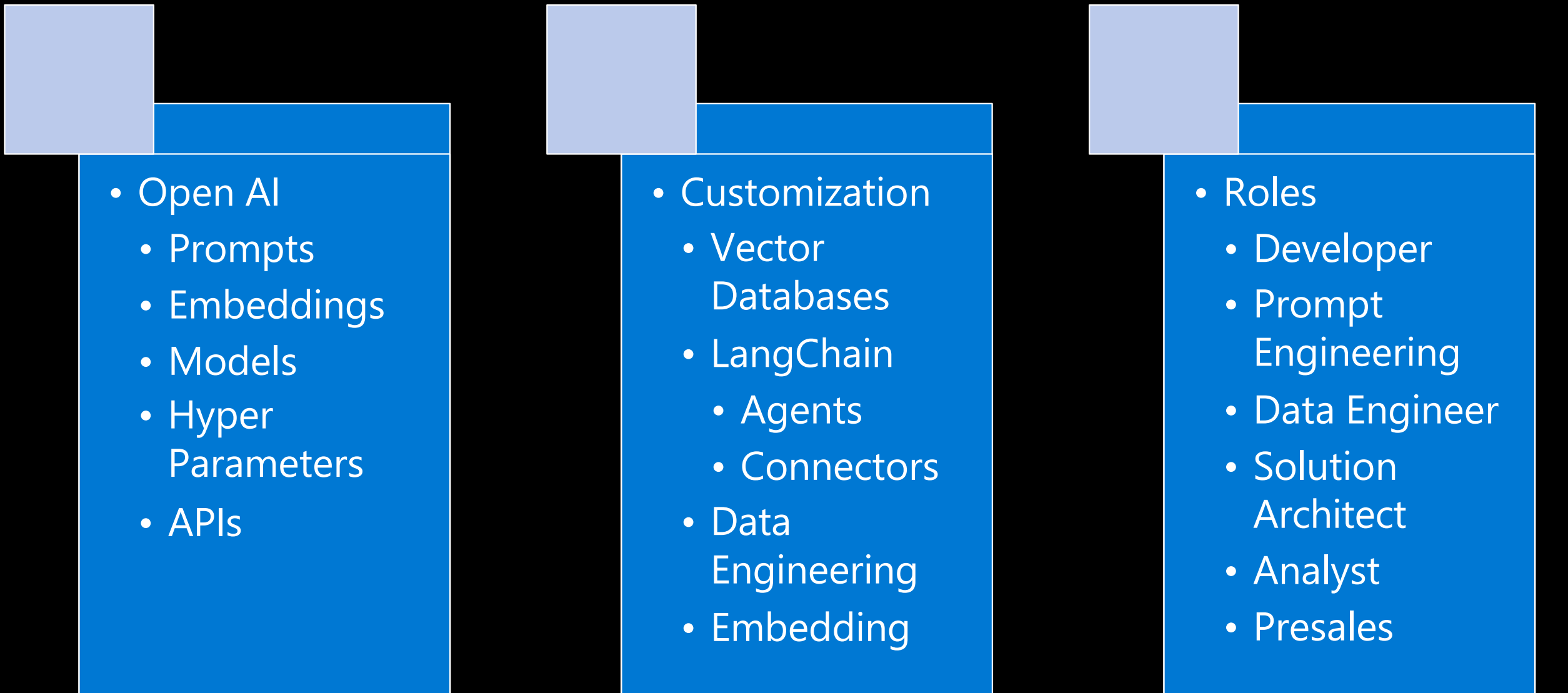
She can slide 3 times before the water slides close.



Reduce Hallucination using Prompt Engineering



LLM Solution Pieces

- 
- Open AI
 - Prompts
 - Embeddings
 - Models
 - Hyper Parameters
 - APIs

- Customization
 - Vector Databases
 - LangChain
 - Agents
 - Connectors
 - Data Engineering
 - Embedding

- Roles
 - Developer
 - Prompt Engineering
 - Data Engineer
 - Solution Architect
 - Analyst
 - Presales

Q&A

Azure GPT

| Ask the team anything

