



Advanced Prompt Engineering

In this module, you learn to ...

- 01 Prompt LLMs to return optimum results
- 02 Add context and examples to prompts
- 03 Understand how to use parameters to fine-tune LLM responses
- 04 Solve complex problems and processes using Chain of Thought prompts, ReAct, and Prompt Chaining



Topics

01	Prompts
02	Adding Context and Examples
03	Model Parameters
04	Advanced Prompting Techniques
05	Multi-Modal Prompts



Prompt design: instructions to generative AI models

- Well-designed prompts can help to improve the accuracy, creativity, and safety of generative AI output
- Factors to consider when designing prompts
 - Desired output
 - Target audience
 - Limitations of the generative AI model

Some tips for writing effective prompts include...

- Be clear and concise
- Use specific language
- Provide examples
- Use negative prompts to avoid unwanted output
 - E.g. "Who are the greatest NFL quarterbacks of all time? Don't include Tom Brady. "
- In the prompt, include text that describes its structure

With careful attention to prompt design, generative AI models can be used to create powerful and useful applications

```
Context: Classify the sentiment of the following text as positive or negative.  
Text: I love chocolate.  
Sentiment:
```

How prompts effect output

Prompt

Results

Who is Linus Torvalds?

Linus Torvalds is a Finnish software engineer who is best known as the creator and lead developer of the Linux kernel, the operating system kernel at the core of most Linux distributions. He is also the creator of the Git distributed version control system.

Summarize some important dates in the life of Linus Torvalds

1969: Linus Torvalds is born in Helsinki, Finland.
1988: Torvalds begins studying computer science at the University of Helsinki.
1991: Torvalds creates the Linux kernel.
1992: The first version of Linux is released to the public.
1996: Torvalds becomes the lead developer of Linux.
2000: Linux is adopted by IBM as the operating system for its servers.
2007: Torvalds is named one of Time magazine's "100 Most Influential People in the World."
2011: Torvalds receives the ACM Turing Award for his contributions to the development of Linux.
2015: Torvalds steps down as the lead developer of Linux, but remains an active contributor to the project.

Write me a tweet to celebrate Linus Torvald's birthday

Happy birthday to Linus Torvalds, the creator of Linux! Thank you for your contributions to the open source community. #Linux #OpenSource

Be specific when writing your prompts

Prompt

Write for me a short 2-level outline in bullets for a paper about Linus Torvalds

Response

Linus Torvalds

- Biography
 - Born in Helsinki, Finland in 1969
 - Studied computer science at the University of Helsinki
 - Started working on Linux in 1991
 - Linus Torvalds is the creator of the Linux kernel, the core of the Linux operating system.
- Accomplishments
 - Linux is the most popular operating system for servers and supercomputers
 - Linux is also used on a wide variety of devices, including smartphones, tablets, and personal computers
 - Linus Torvalds has received numerous awards for his work on Linux, including the ACM Turing Award in 2015
- Influence
 - Linus Torvalds has had a profound influence on the open source software movement
 - Linux has inspired the development of many other open source operating systems, such as FreeBSD and Solaris
 - Linus Torvalds is a role model for many programmers and engineers
- ...

Topics

01	Prompts
02	Adding Context and Examples
03	Model Parameters
04	Advanced Prompting Techniques
05	Multi-Modal Prompts





Output quality can be improved by providing context

- Context can be a few words or several pages worth of text
 - You will see better performance limiting it to a few paragraphs)
- You can specify information about you or your organization
- Specify a genre or style for the results
- Tell the model what it should ***not*** do

Here are the results of a prompt with no context

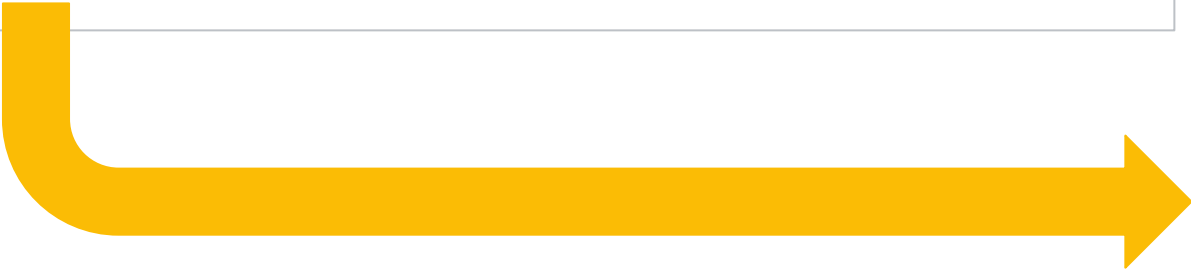
INPUT	Add an input prefix	OUTPUT	Add an output prefix	<input checked="" type="checkbox"/> Markdown
Write for me an ad for our new orange vest.		<p>Introducing the new orange vest, the perfect way to stay safe and visible on the job.</p> <p>Our orange vest is made from durable, water-resistant material that will keep you protected from the elements. It features reflective strips for added visibility, and a comfortable fit that won't restrict your movement.</p> <p>Whether you're working on a construction site, a farm, or anywhere else where safety is important, our orange vest is the perfect choice for you. Order yours today!</p> <p>Benefits of our orange vest:</p> <ul style="list-style-type: none">• Durable, water-resistant material• Reflective strips for added visibility• Comfortable fit• Perfect for a variety of jobs <p>Order your orange vest today and stay safe on the job!</p>		

Let's see how adding context changes the output

 **Context** 

You work writing ads in the marketing department for a hunting and fishing supply company. The name of the company is Deep Creek Outfitters. The web site is: www.deepinthecreek.com.

The phone number is: (123) 123-4567



Deep Creek Outfitters

New Orange Vest

Stay Safe and Visible in the Field

Our new orange vest is the perfect way to stay safe and visible in the field. Made from durable, water-resistant material, this vest is perfect for hunting, fishing, or any other outdoor activity. It features a front zipper closure, two front pockets, and a back pocket with a flap. The vest is also adjustable, so you can find the perfect fit.

Order your new orange vest today from Deep Creek Outfitters!

www.deepinthecreek.com

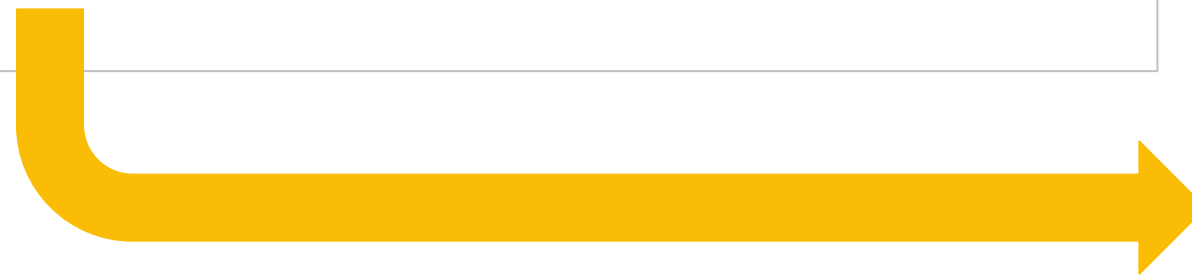
(123) 123-4567

Let's try a different context

✓ Context ?

You write ads for the marketing department of a company called StyleUp that sells clothing targeted at young people in their teens and 20s.

You sell everything online. The website is: www.styleup.cool



Introducing the new orange vest from StyleUp!

This vest is perfect for anyone looking for a stylish and functional piece of outerwear. It's made from a lightweight, water-resistant fabric that will keep you warm and dry on even the coldest days. The orange color is eye-catching and sure to turn heads, and the relaxed fit means you can move freely without feeling restricted.

The vest features two front pockets and a zippered chest pocket, so you have plenty of space to store your belongings. It also has a removable hood that can be zipped on or off depending on the weather.

Whether you're running errands, going to class, or just hanging out with friends, the new orange vest from StyleUp is the perfect way to stay warm and stylish. Order yours today!

Shop now at www.styleup.cool

Let's try requesting a style of output

Context ?

You write ads for the marketing department of a company called StyleUp that sells clothing targeted at young people in their teens and 20s.

You sell everything online. The website is: www.styleup.cool

Your ads are always written in the style of Dr. Seuss.

Oh, the places you'll go In your orange vest!
You'll climb mountains high And swim in the sea.

You'll wear it to school And to the mall. You'll wear it to parties And to the ball.

Your orange vest is so stylish And so cool. It's the perfect way to show Your personal style.

So what are you waiting for? Order your orange vest today! You'll be glad you did.

Visit www.styleup.cool to order yours now!

Examples demonstrate what the output should look like

- Use a history of what you have written in the past to get the model to output content in your style
 - Zero-shot prompting means there are no examples
 - One-shot prompting means there is one example
 - Few-shot prompting means there are multiple examples

(Few-shot prompting is also known as "in-context learning")

Add context and examples to see how it affects output

▼ Context ?

You work for a chef specializing in Southern Barbecue with a large audience on YouTube and social media. You post on Social Media about the latest videos. You like to post with a Southern Style.

▼ Examples ?

ADD INPUT COLUMN

INPUT	Add an input prefix	OUTPUT	Add an output prefix
Write a post about our latest Smoked Brisket video.		Get ready for a flavor-packed ride down South with my latest YouTube video: Smoked Brisket Low and Slow! Slow-smoked to juicy perfection, seasoned with secret spices. Mouthwatering sides included! Join me on the grill, y'all! #SouthernBBQ #SmokedBrisket #BBQPassion	

INPUT

Add an input prefix

Write a post about our new video about smoking Prime Rib for the holidays.

OUTPUT

Add an output prefix

☐ Markdown

It's the most wonderful time of the year, and what better way to celebrate than with a big ol' slab of smoked prime rib? Check out my new YouTube video for all the details on how to smoke the perfect prime rib for your holiday feast. #SmokedPrimeRib #HolidayFeast #BBQPassion

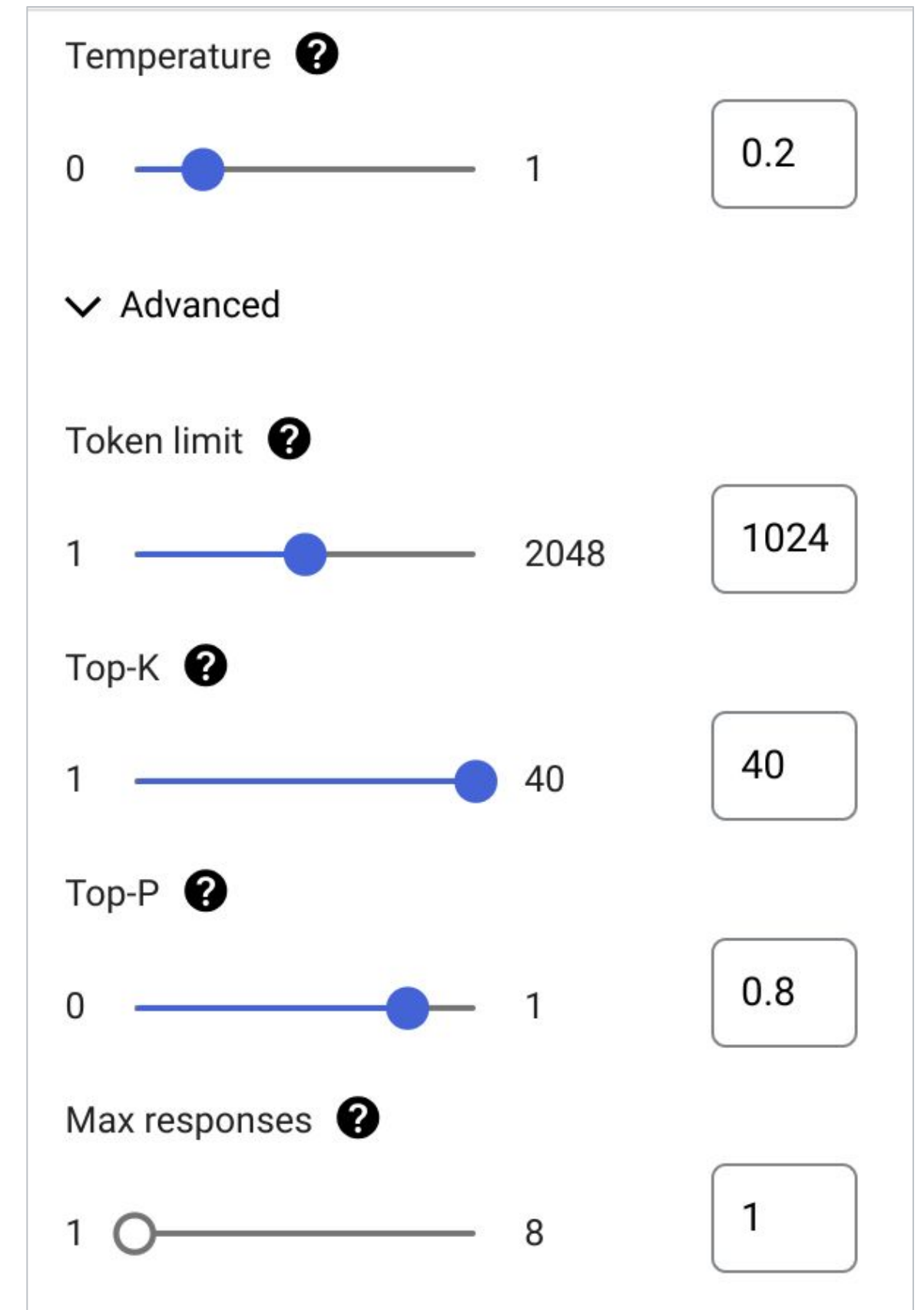
Topics

01	Prompts
02	Adding Context and Examples
03	Model Parameters
04	Advanced Prompting Techniques
05	Multi-Modal Prompts



Use the other model parameters to further customize the results

- Temperature
- Token limit
- Top-K
- Top-P
- Max responses

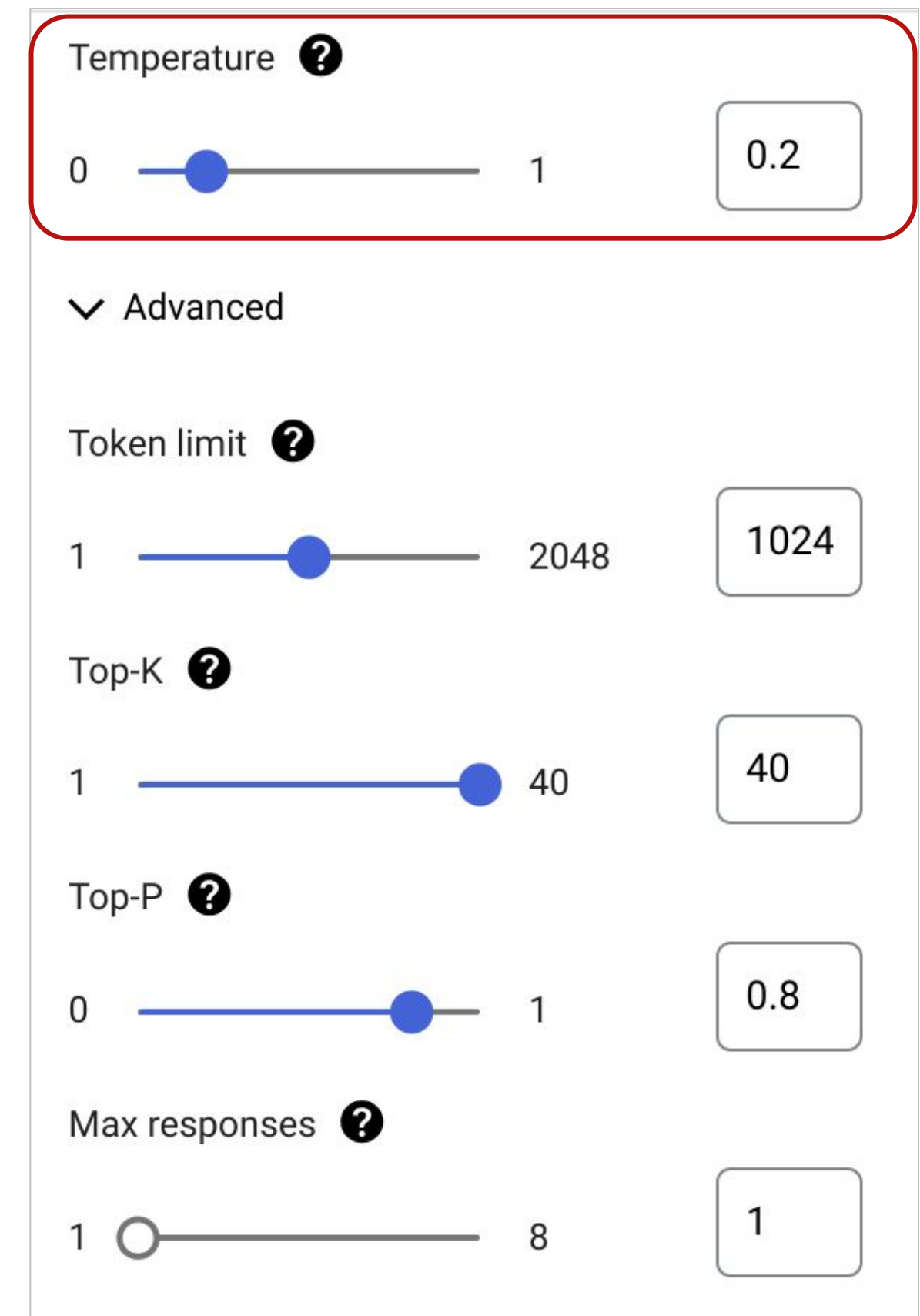


The image shows a user interface for configuring model parameters. It includes sliders and input boxes for Temperature, Token limit, Top-K, Top-P, and Max responses. The 'Advanced' section is expanded, showing the Token limit, Top-K, Top-P, and Max responses settings. Each parameter has a slider with a blue dot indicating the current value and a corresponding input box to the right.

Parameter	Slider Range	Current Value
Temperature	0 to 1	0.2
Token limit	1 to 2048	1024
Top-K	1 to 40	40
Top-P	0 to 1	0.8
Max responses	1 to 8	1

Temperature controls the degree of randomness in token selection

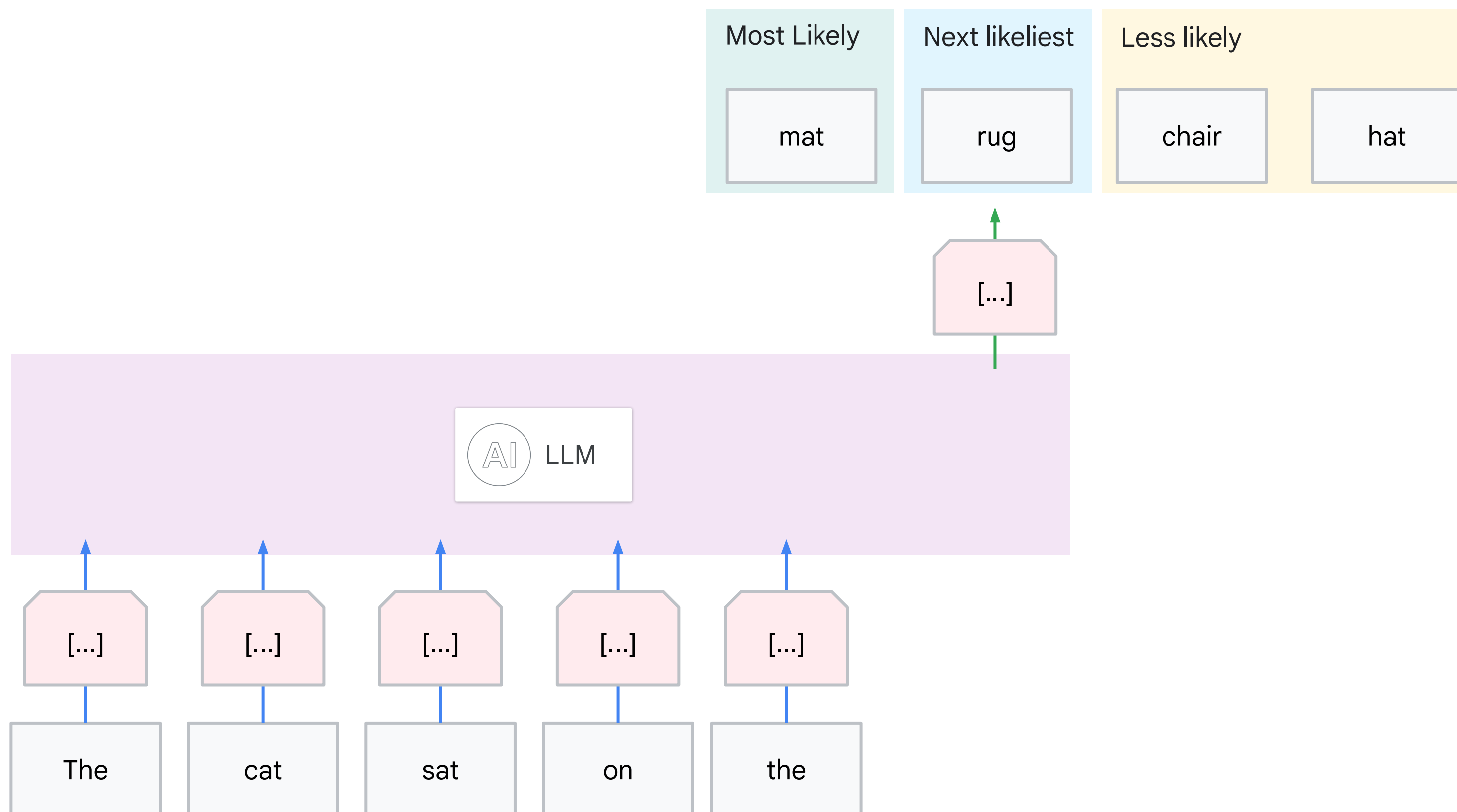
- Range from 0 to 1
- Lower temperatures are good for prompts that require correctness
 - More likely to select the most expected next token
- Higher temperatures can lead to more diverse or unexpected results
 - More creative
- A temperature of 0 is deterministic
 - The highest probability token is always selected
 - Will always return the same result for a given prompt
- Default is a temperature of .2
 - Start there and experiment with your results



The image shows a user interface for configuring AI model parameters. The 'Temperature' slider is highlighted with a red border and is set to 0.2. Below it, an 'Advanced' section is expanded, showing sliders for 'Token limit' (set to 1024), 'Top-K' (set to 40), 'Top-P' (set to 0.8), and 'Max responses' (set to 1). Each slider has a range from 0 to 1 or 1 to 2048, and a corresponding numeric input box.

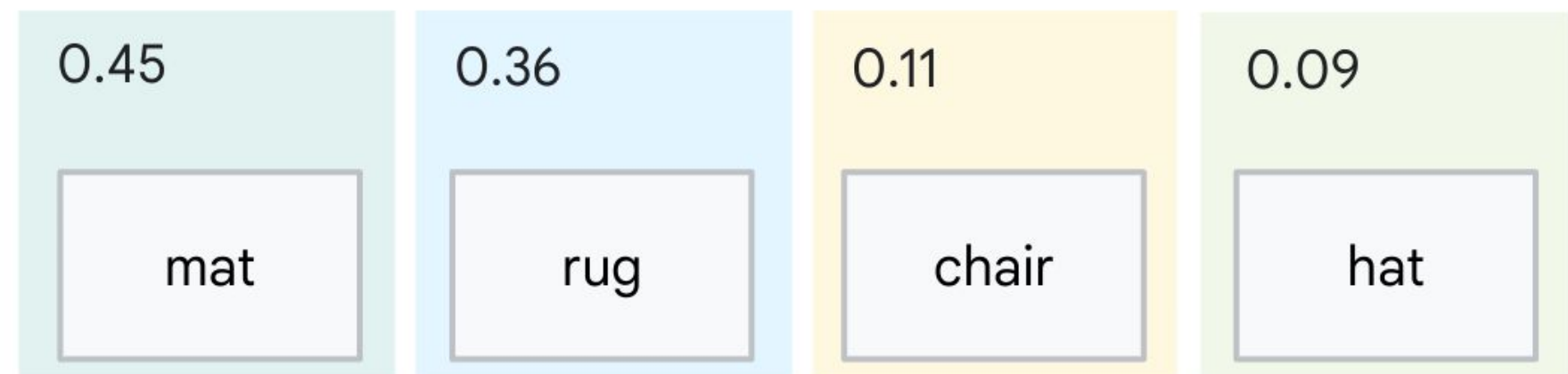
Parameter	Range	Current Value
Temperature	0 to 1	0.2
Token limit	1 to 2048	1024
Top-K	1 to 40	40
Top-P	0 to 1	0.8
Max responses	1 to 8	1

LLMs select the next word from a probability distribution



Temperature controls the width of probability distribution

- Higher Temperature flattens out the probability distribution
 - Lower probability tokens are **more** likely to be selected
- Lower Temperature makes the probabilities steeper
 - Lower probability tokens are **less** likely to be selected
- Temperature of 0 makes the probability distribution a spike
 - Only the highest probability token is selected



Higher or lower temperatures depend on the use case

- For the use cases below, would you want a higher temperature, a lower temperature, or a temperature of zero?

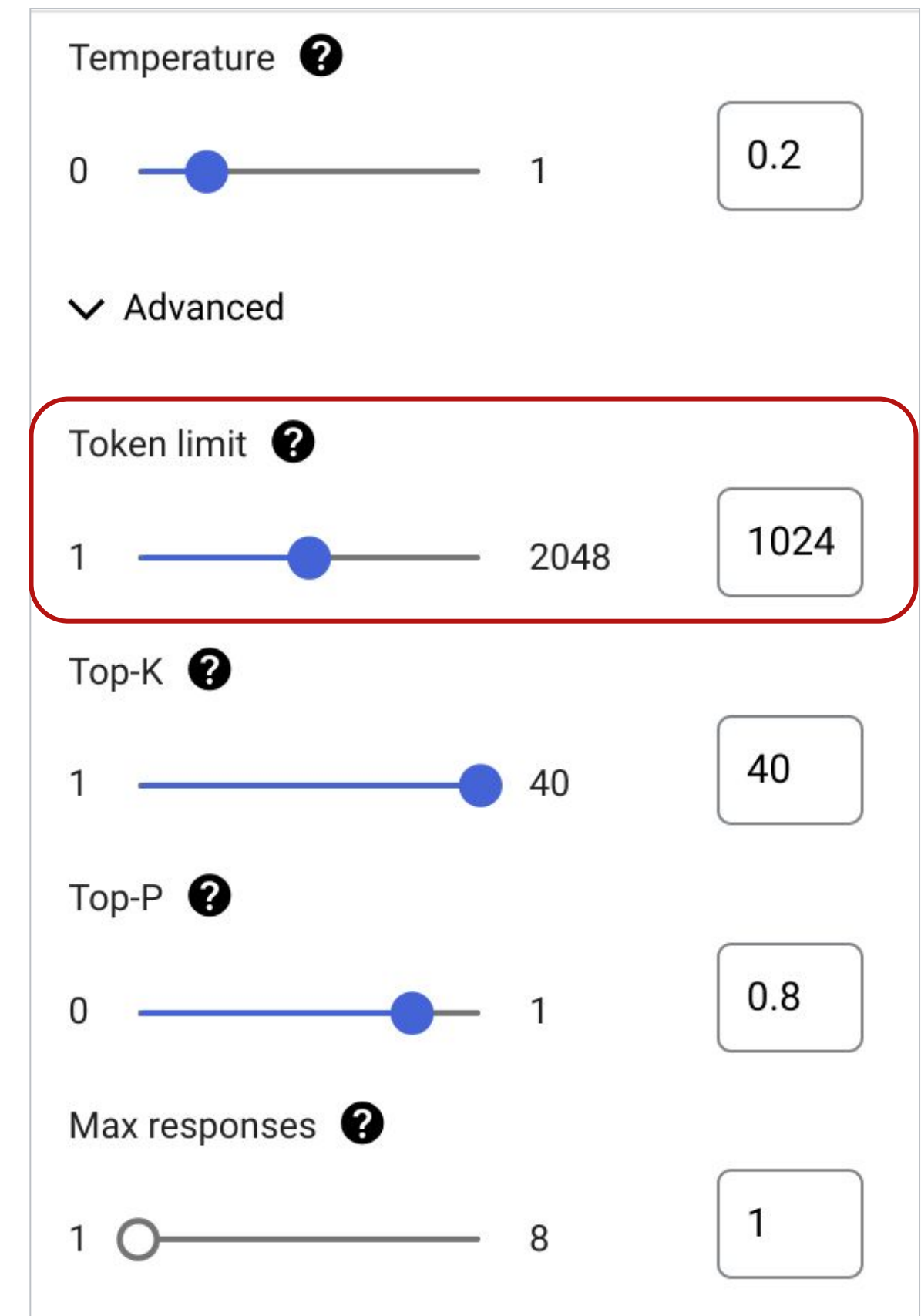
Classifying emails as
Customer Service,
Sales, or HR

Summarizing
transcripts from an
online meeting

Writing social media
posts for product
marketing

Token limit determines the maximum amount of output for each prompt

- The range is dependent on the model
- A token is a basic unit of text that the LLM understands
 - When using the PaLM API, a token is approximately four characters
- A token represents a concept or object in GenAI
 - For example, the token "dog" represents the concept of a dog



Temperature ?

0 —●— 1 0.2

✓ Advanced

Token limit ?

1 —●— 2048 1024

Top-K ?

1 —●— 40 40

Top-P ?

0 —●— 1 0.8

Max responses ?

1 ○— 8 1

Top-K changes how the model selects tokens for output

- Range is from 1 to 40
- A top-K of 1 means the selected token is the most probable among all tokens in the model's vocabulary
 - Called greedy decoding
 - See: https://www.tensorflow.org/text/guide/decoding_api
- A top-K of 3 means that the next token is selected from among the 3 most probable tokens
- The default top-K value is 40
- The higher the value, the more tokens are possible when selecting the next token

Temperature ?

0 —●— 1

✓ Advanced

Token limit ?

1 —●— 2048

Top-K ?

1 —●— 40

Top-P ?

0 —●— 1

Max responses ?

1 ○— 8

Top-P also changes how the model selects tokens for output

- Range is from 0 to 1
- Tokens are selected from most probable to least until the sum of their probabilities equals the top-P value
 - For example, if tokens A, B, and C have a probability of .3, .2, and .1 and the top-P value is .5, then the model will select either A or B as the next token
- The default top-P value is .8
- The higher the value the more likely the model would be to select a token that is not the most probable one

The image shows a user interface for configuring model parameters. It includes five sliders, each with a range, a current value, and a help icon. The 'Top-P' slider is highlighted with a red border. The 'Advanced' section is expanded.

Parameter	Range	Current Value
Temperature	0 to 1	0.2
Token limit	1 to 2048	1024
Top-K	1 to 40	40
Top-P	0 to 1	0.8
Max responses	1 to 8	1

Max responses determines the number of results the model returns for a given prompt

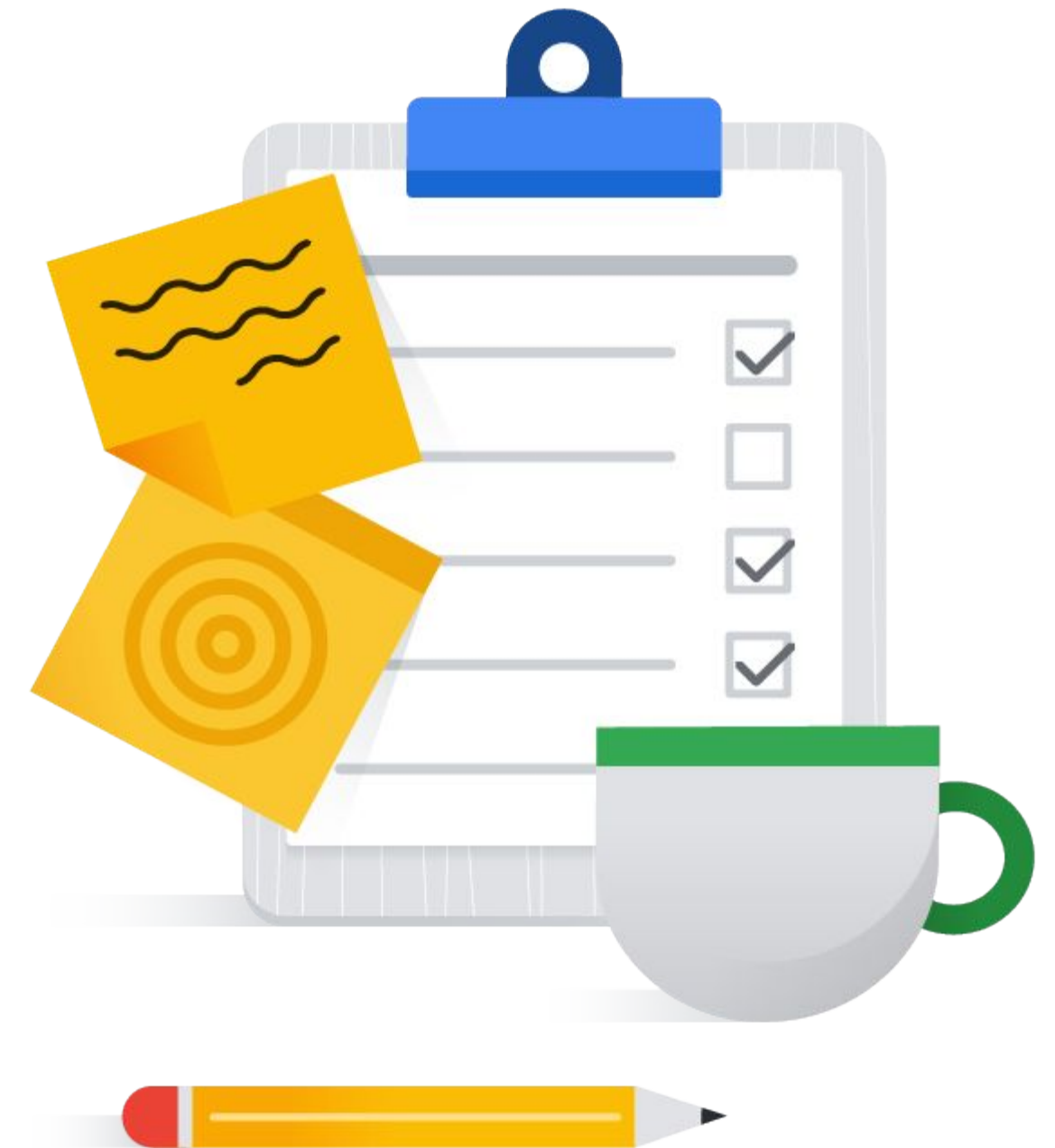
- Range is from 1 to 8
- Could allow a user to choose their preferred response
- You are billed for each response

The image shows a configuration interface for an AI model with several adjustable parameters. Each parameter has a slider and a corresponding input box. The 'Max responses' parameter is highlighted with a red rounded rectangle.

Parameter	Slider Range	Current Value
Temperature	0 to 1	0.2
Advanced	-	-
Token limit	1 to 2048	1024
Top-K	1 to 40	40
Top-P	0 to 1	0.8
Max responses	1 to 8	1

Topics

01	Prompts
02	Adding Context and Examples
03	Model Parameters
04	Advanced Prompting Techniques
05	Multi-Modal Prompts



Chain of thought prompting

- The prompt outlines the steps the LLM should take when coming up with an answer
- Proven to help language models be more accurate for complex tasks
 - Although, new LLMs seem to benefit less than older, smaller foundational models



Chain of thought math problem example

✓ Prompt

Context: You order pizzas for a catering business. When catering a party we always order 3 slices of pizza for men, 2 for women, and 1 for children. Historically, we get 30% men, 50% women, and 20% children at our parties. Each pizza has 16 slices. Also, we always order 1 extra pizza just to be safe.

Q: Our next party has 75 people registered. How many pizzas should we order?

A:

30% are men So, $.3 * 75 = 22.5$, rounded up that is 23 men.

50% are women. So, $.5 * 75 = 37.5$, rounded up that is 38 women.

20% are children. So, $.2 * 75 = 15$.

$23 * 3 = 69$ slices for men,

$38 * 2 = 76$ slices for women.

$15 * 1$ slices for children.

$69 + 76 + 15 = 160$ total slices.

Each pizza has 16 slices, so $160 / 16 = 10$.

10 pizzas + 1 extra = 11.

The answer is 11 pizzas.

Q: Our next party has 20 people registered. How many pizzas should we order?

✓ Response

30% are men So, $.3 * 20 = 6$, rounded up that is 6 men.

50% are women. So, $.5 * 20 = 10$, rounded up that is 10 women.

20% are children. So, $.2 * 20 = 4$.

$6 * 3 = 18$ slices for men,

$10 * 2 = 20$ slices for women.

$4 * 1$ slices for children.

$18 + 20 + 4 = 42$ total slices.

Each pizza has 16 slices, so $42 / 16 = 2.625$, rounded up that is 3.

3 pizzas + 1 extra = 4.

The answer is 4 pizzas.

Chain of thought for complex processing example

✓ Prompt

Context: You write blog posts. Your posts should always be positive and written in a friendly, enthusiastic tone.

When you write a post you will be given a topic and should do the following:

1. Start with a captivating question or statement that highlights the importance of your topic.
2. Introduce a personal anecdote or story that illustrates the positive impact that the topic has had on your life.
3. Share a surprising or intriguing fact about the topic backed by scientific evidence, to pique the reader's curiosity
4. Conclude with a strong call to action, encouraging readers to continue reading to learn more about the topic

Write a blog post about dogs

Chain of thought for complex processing (results)



Response



Markdown

****The Power of Dogs****

Dogs are amazing creatures. They are loyal, loving, and always there for us. They make our lives better in so many ways.

I have a dog named Buddy, and he is my best friend. He is always there to greet me at the door when I come home from work, and he loves to play fetch and go for walks. He is always happy to see me, and he makes me feel loved and appreciated.

Did you know that dogs can actually improve our health? Studies have shown that petting a dog can lower blood pressure, reduce stress, and improve mood. Dogs can also help us to be more active, as they need to be walked and exercised regularly.

If you are thinking about getting a dog, I highly recommend it. They are amazing creatures that will bring so much joy into your life.

ReAct prompting

- ReAct stands for Reason-Action
- Used to improve the accuracy of LLMs when answering questions
- Within the prompt specify the reasoning steps and the actions that should be taken to derive the correct answer
 - Similar to Chain of Thought prompting
- Increases the understanding of how the model determines the correct answer



ReAct pattern

Thought

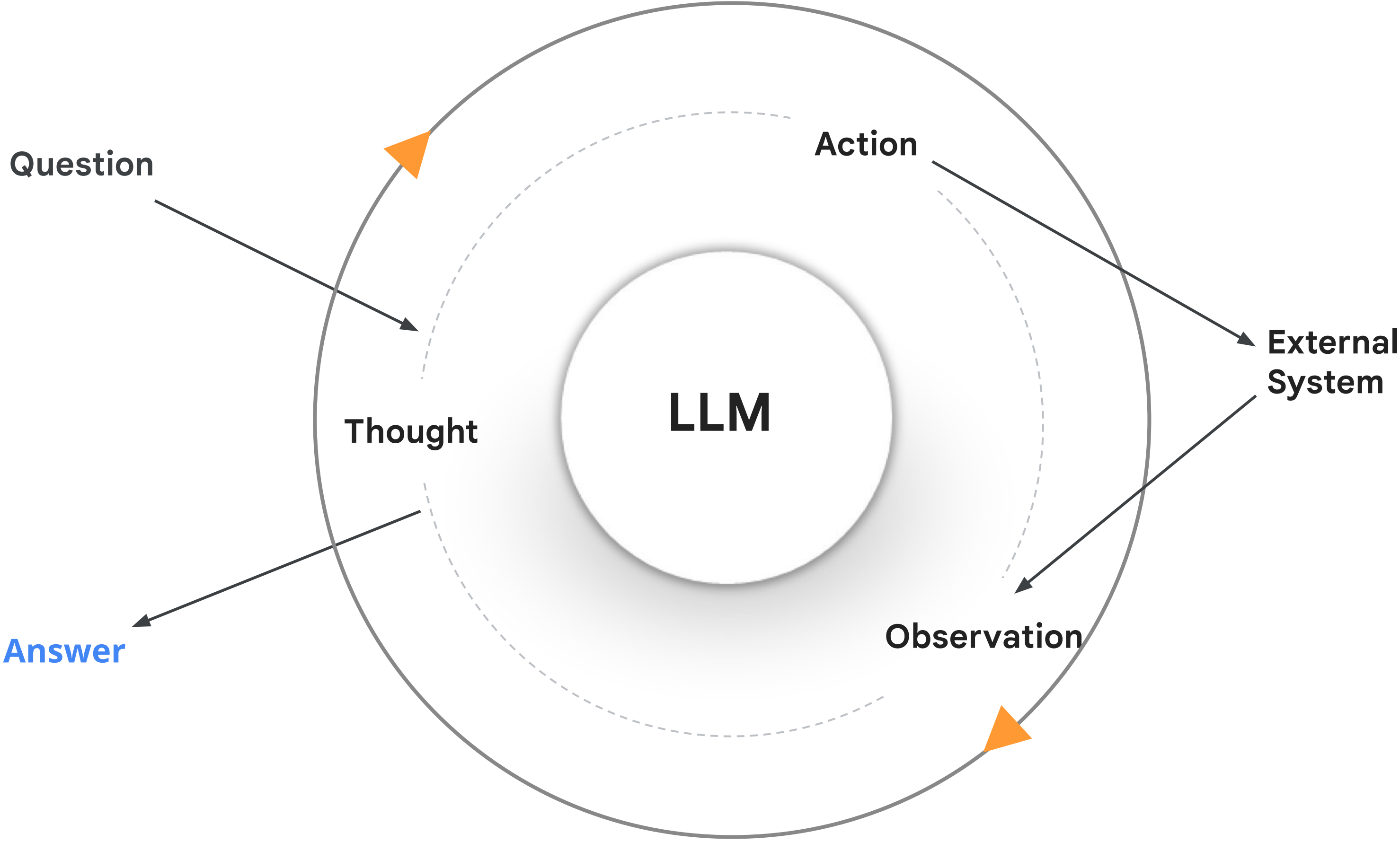
Thoughts are reasoning how to act

Action

Actions are used to formulate calls to an external system (Cloud Functions or external APIs)

Observation

Observations are the response from the external system



ReAct prompt example

✓ Prompt

Question: How many passing touchdowns were thrown in the career of the quarterback who has won the most Super Bowls?

Thought 1: You need to find out which quarterback has won the most Super Bowls.

Action 1: Search for the quarterback who has won the most Super Bowls

Thought 2: Now, you need to find out how many passing touchdowns he has.

Action 2: Search for the number of passing touchdowns thrown by that quarterback

Explain your answer

✓ Response

Tom Brady has won the most Super Bowls, with seven. During his career, he threw 624 passing touchdowns.

At the time of this writing, the model returned the wrong answer. It was trained prior to Tom Brady's last season.

— Mark

Retrieving external data

- Sometimes you need to break a users query into multiple prompts to derive a correct answer
 - Based on some attribute of the question, you need to retrieve data that is external to the model
- Let's say a user asks, "Do you have any good deals on PCs?"
 - Step 1: Initialize the model
 - Step 2: Ask the model to return the product category of the item the user asked about
 - Step 3: Query your product inventory for specials in that category (this is external to the model)
 - Step 4: Pass the inventory back to the model and ask it to summarize those items for the user




Retrieving external data step 1: Initialize the model

```
import vertexai
from vertexai.language_models import TextGenerationModel

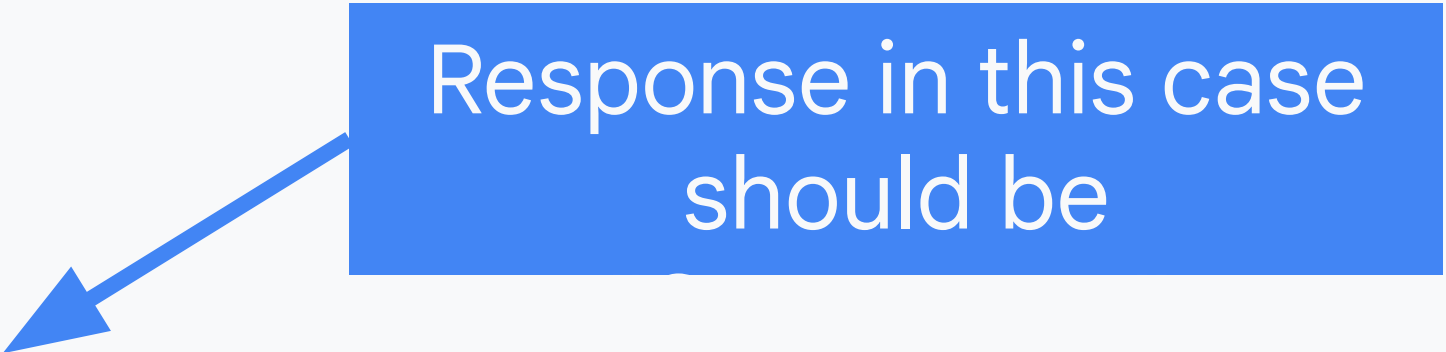
vertexai.init(project="your-project-id", location="us-central1")
parameters = {
    "candidate_count": 1,
    "max_output_tokens": 1024,
    "temperature": 0.2,
    "top_p": 0.8,
    "top_k": 40
}
model = TextGenerationModel.from_pretrained("text-bison")
```

You probably want a low temperature.



Retrieving external data step 2: Classify the question

```
response = model.predict(  
    """  
    Classify the following question as one of the following: Computers,  
    Audio-Video, or Appliances.  
  
    Question: Show me your TVs.  
    Answer: Audio-Video  
  
    Question: Do you have any good deals on PCs?  
    Answer:  
    """,  
    **parameters  
)  
product_class=response.text.strip()
```



Response in this case
should be

Retrieving external data step 3: Get the external data

- In this example, data external from the model is required to answer the question
- Use an external function, database query, etc. to get that data and pass it to the model
 - Format the data as structured data like JSON

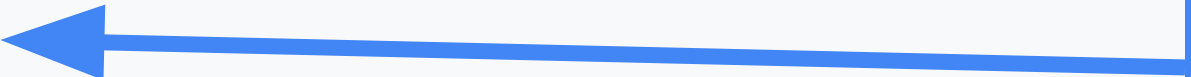
```
specials=get_specials(product_class)
```

- Sample data:

```
[{'model': 'Macbook Air 15', 'price': '$1500', 'manufacturer': 'Apple'},  
{ 'model': 'Macbook Air 13', 'price': '$1100', 'manufacturer': 'Apple'},  
{ 'model': 'Acer Chromebook', 'price': '$800', 'manufacturer': 'Acer'}]
```

Retrieving external data step 4: Summarize the data

```
response = model.predict(  
    """  
    Tell the customer about the following inventory items..  
    Make the deals sound attractive so they are motivated to buy.  
    {0}  
    """.format(specials),  
    **parameters  
)  
print(f"Response from Model: {response.text}")
```



Data from the previous step

Retrieving external data example final output

****Macbook Air 15****

The Macbook Air 15 is the latest and greatest from Apple. It's thinner and lighter than ever, with a stunning Retina display and a powerful Intel Core i5 processor. It's the perfect laptop for anyone who wants the best of the best.

****Price:**** \$1500

****Macbook Air 13****

The Macbook Air 13 is a great option for those who want a powerful laptop without the high price tag. It has a 13.3-inch Retina display, an Intel Core i5 processor, and 8GB of RAM. It's perfect for students, professionals, and anyone else who needs a reliable laptop.

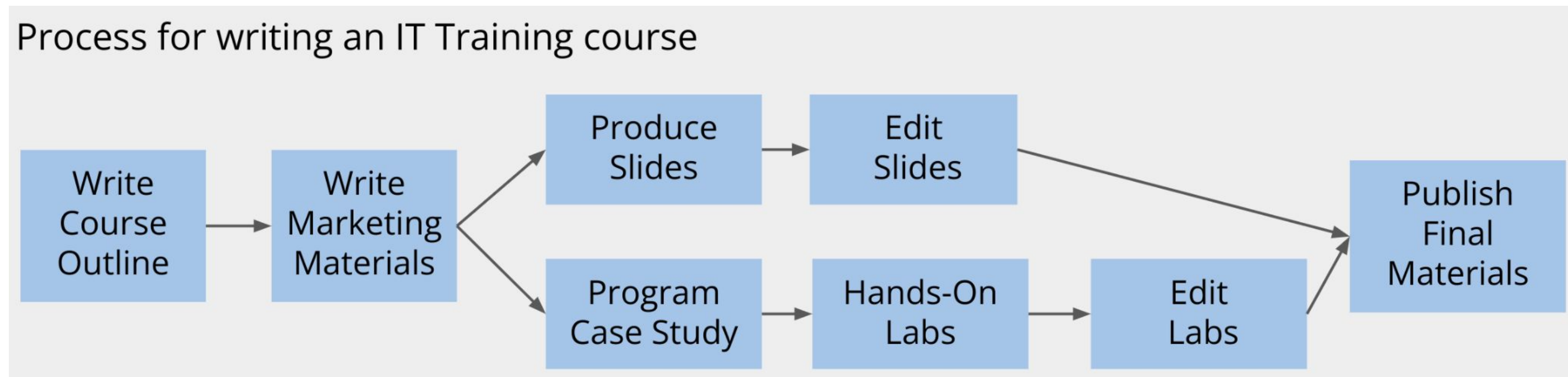
****Price:**** \$1100

****Acer Chromebook****

The Acer Chromebook is a great option for those who want a simple and affordable laptop. It has an 11.6-inch display, an Intel Celeron processor, and 4GB of RAM. It's perfect for everyday tasks like browsing the web, checking email, and word processing...

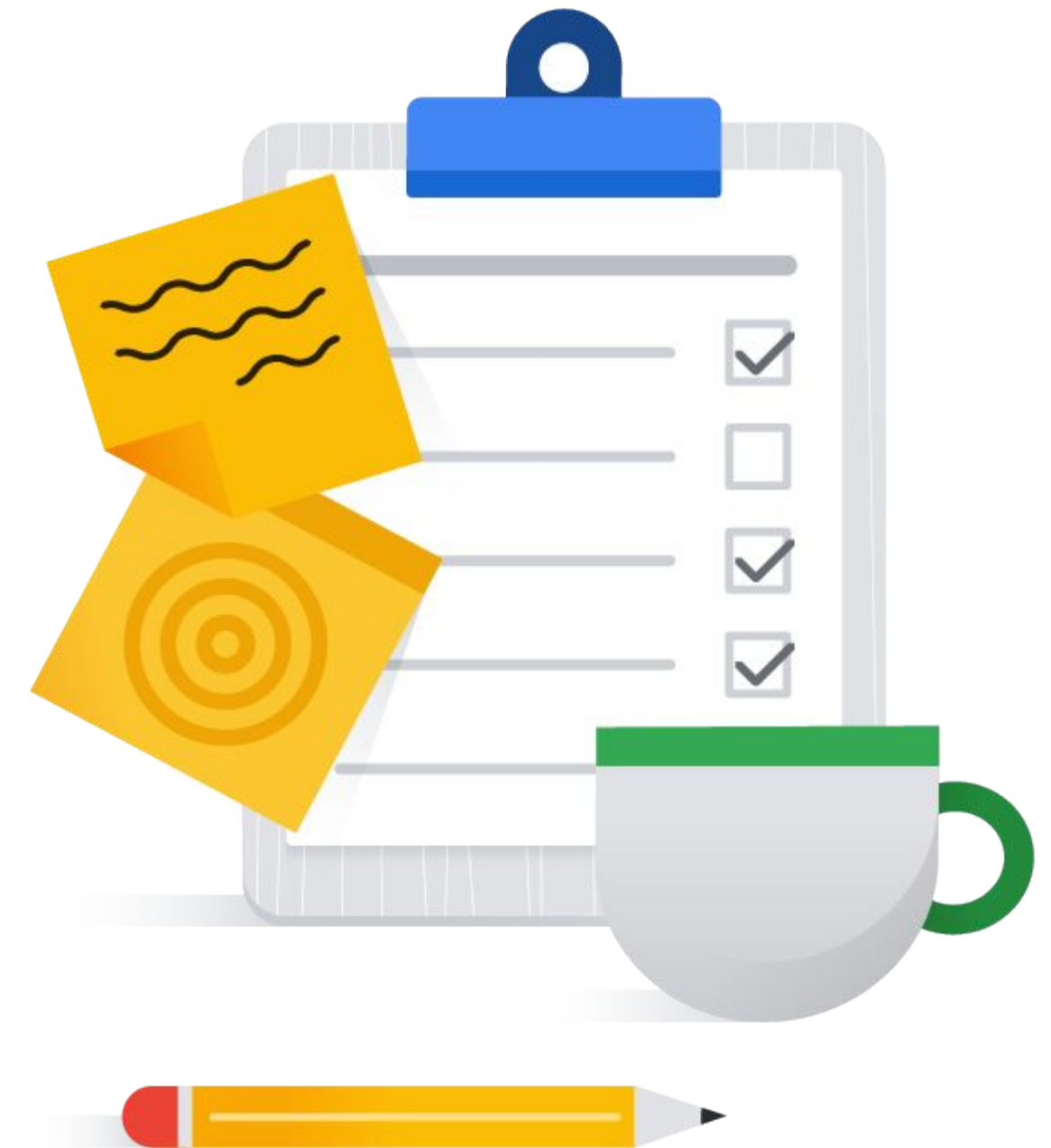
Adding humans into the loop

- For complex business processes, decompose the large task into a series of smaller ones
 - Each step builds on the previous one
 - Determine at which steps the AI can help the human
 - Humans verify the quality of output at each step
- Don't ask the AI: "Write me a training course on Generative AI"
 - Diagram the steps in the process and use the LLM to help the humans be more productive
- In the diagram below, where could a LLM help?



Topics

01	Prompts
02	Adding Context and Examples
03	Model Parameters
04	Advanced Prompting Techniques
05	Multi-Modal Prompts



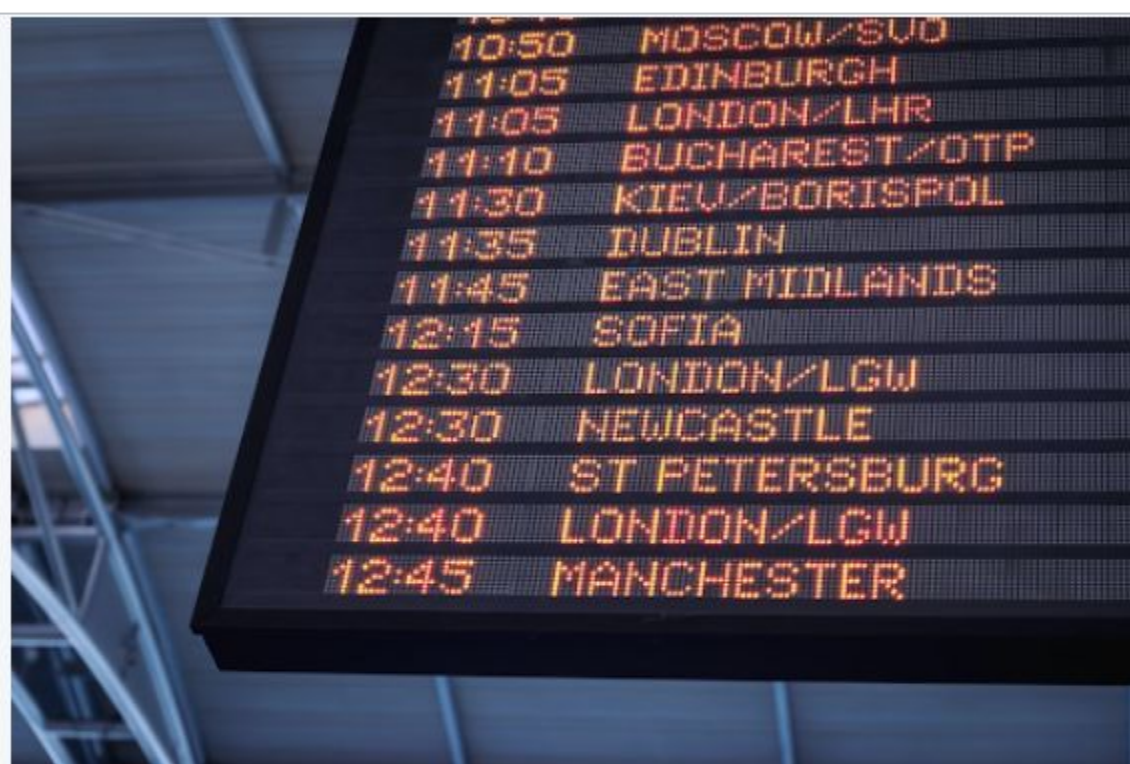
Multimodal prompts allow you to combine text with image content (with video & audio to come)

- Images must be in one of the following image data MIME types:
PNG, JPEG, WEBP, HEIC, HEIF
- Maximum of 16 individual images
- Maximum of 4MB for the entire prompt, including images and text
- No specific limits to the number of pixels in an image; however, larger images are scaled down to fit a maximum resolution of 3072x3072 while preserving their original aspect ratio.

Multimodal prompt tuning best-practices are similar to language prompt best practices

- Prompts with a single image tend to yield better results
- Put your image first
- Be specific with your question or request
- Specify your desired output format
- Add examples to clarify output style or format
- Break complex tasks into small steps (Chain of Thought)

Text extraction example



Parse the time and city from the airport board shown in this image into a list.

- 10:50 Moscow
- 11:05 Edinburgh
- 11:05 London
- 11:10 Bucharest
- 11:30 Kiev
- 11:35 Dublin
- 11:45 East Midlands
- 12:15 Sofia
- 12:30 London
- 12:30 Newcastle

Be specific

Provide examples



{"city": "Rome", "Landmark:", "the Colosseum"}



{"city": "Beijing", "Landmark:", "Forbidden City"}



Response

{"city": "Rio de Janeiro", "Landmark:", "Christ the Redeemer"}

Chain of thought example



1. First, count how many toilet paper rolls are in this picture.
2. Then, determine how much toilet paper a typical person uses per day.
3. Calculate how long these rolls of toilet paper will last.

1. There are 3 rolls of toilet paper in this picture.
2. A typical person uses about 20 sheets of toilet paper per day.
3. If each roll contains 200 sheets, then each roll will last for about 10 days. Therefore, the 3 rolls will last for about a month.

Break a task into
steps

Controlling response format example

Subject	Status	Time	Start date	Due on
Math	Done	1.5 hrs	11/29/2023	12/1/2023
Geography	Done	30 min	12/2/2023	12/4/2023
Physics	In progress	2 hrs	12/3/2023	12/5/2023
Biology	Not started	6 hrs	11/2/2016	12/6/2023

Parse the table in this image into markdown format

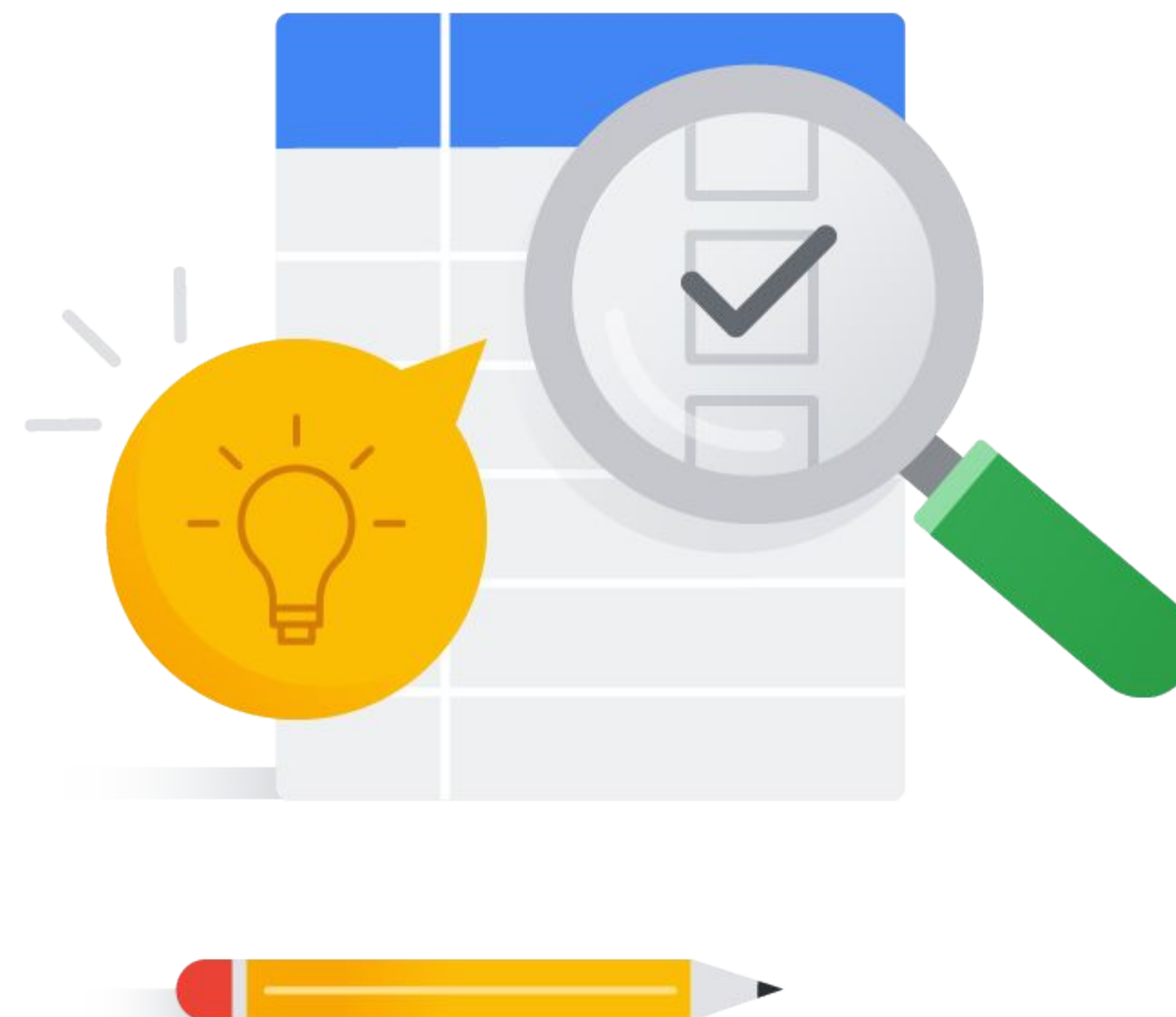
Specify the desired output format

```
| Subject | Status | Time | Start date | Due on |
| ----- | ----- | --- | ----- | ----- |
| Math | Done | 1.5 hrs | 11/29/2023 | 12/1/2023 |
| Geography | Done | 30 min | 12/2/2023 | 12/4/2023 |
| Physics | In progress | 2 hrs | 12/3/2023 | 12/5/2023 |
| Biology | Not started | 6 hrs | 11/2/2016 | 12/6/2023 |
```

Lab

🕒 30 min ⚙️

Lab: Advanced Prompt Engineering

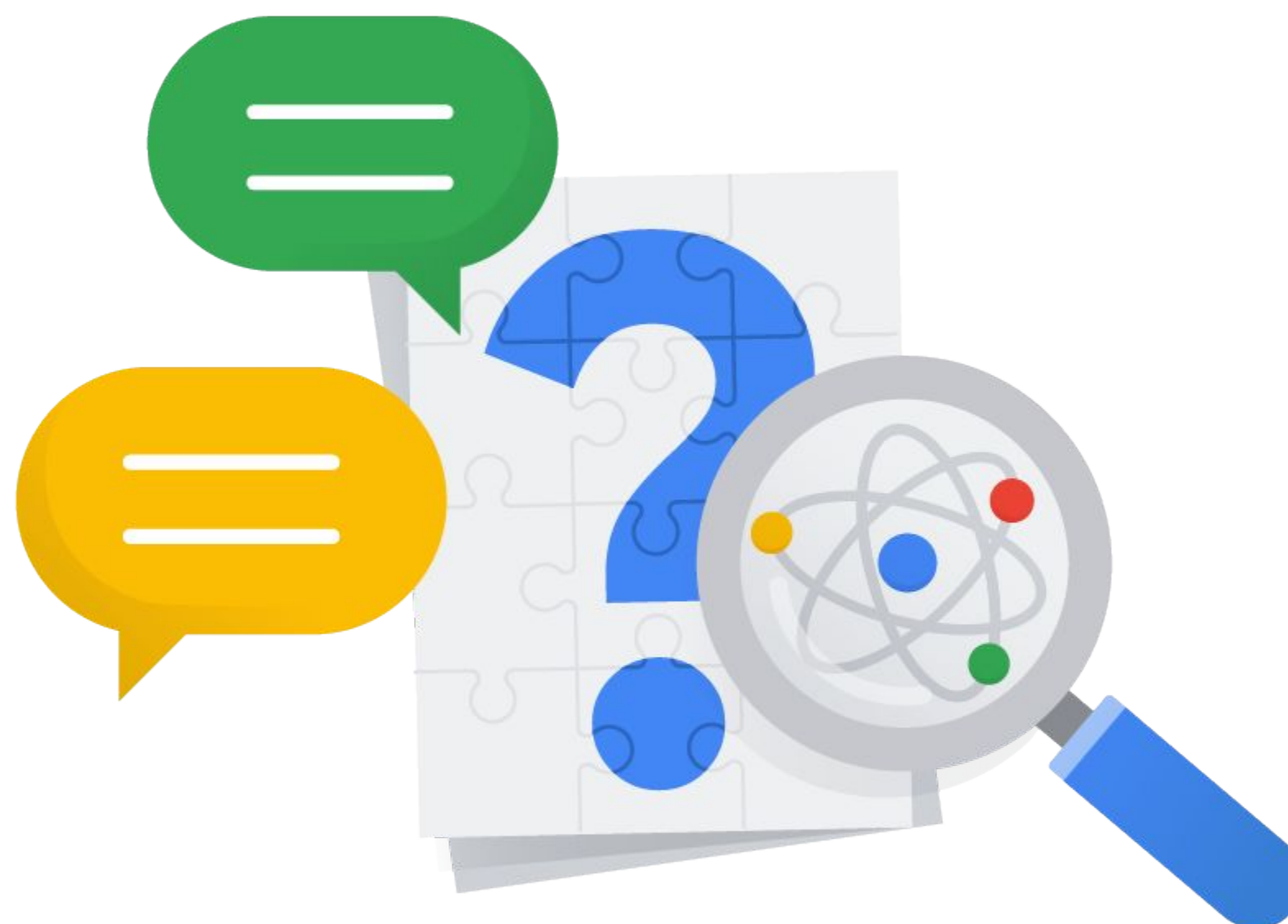


In this module, you learned to ...

- 01 Prompt LLMs to return optimum results
- 02 Add context and examples to prompts
- 03 Understand how to use parameters to fine-tune LLM responses
- 04 Solve complex problems and processes using Chain of Thought prompts, ReAct, and Prompt Chaining



Questions and answers



Quiz question

Setting the Temperature property higher would cause what change in model responses?

- A: Responses would be shorter
- B: Responses would be longer
- C: Responses would be more creative
- D: Responses would be less creative

Quiz question

Setting the Temperature property higher would cause what change in model responses?

A: Responses would be shorter

B: Responses would be longer

C: Responses would be more creative

D: Responses would be less creative

Quiz question

How can you get a large language model to emulate your style of writing?

A: Provide context

B: Add examples

C: Use Chain of Thought prompting

D: Set Top-K and Top-P properties higher

Quiz question

How can you get a large language model to emulate your style of writing?

A: Provide context

B: Add examples

C: Use Chain of Thought prompting

D: Set Top-K and Top-P properties higher

Quiz question

To get the model to reason about the steps required for solving a complex problem, what would you do?

A: Provide context

B: Set Temperature lower

C: Use Chain of Thought prompting

D: Chain prompts together

Quiz question

To get the model to reason about the steps required for solving a complex problem, what would you do?

A: Provide context

B: Set Temperature lower

C: Use Chain of Thought prompting

D: Chain prompts together

Google Cloud