Prompt Engineering

The information in this presentation is classified:

Google confidential & proprietary

⚠️ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

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, Prompt Chaining, and Function Calling.

Topics

01	Prompts
02	Model Parameters
03	Advanced Prompting Techniques
04	Multimodal Prompts

Prompt Engineering:
Crafting instructions to generative AI models

The Google Cloud documentation provides two prompting guides.

1. Overview of prompting strategies from the Generative AI on Vertex AI documentation.
2. Prompt design strategies from the Gemini API documentation (using an API key-based version of the Gemini API rather than accessing it through Vertex AI associated with a GCP project, but the prompt guidance still holds).

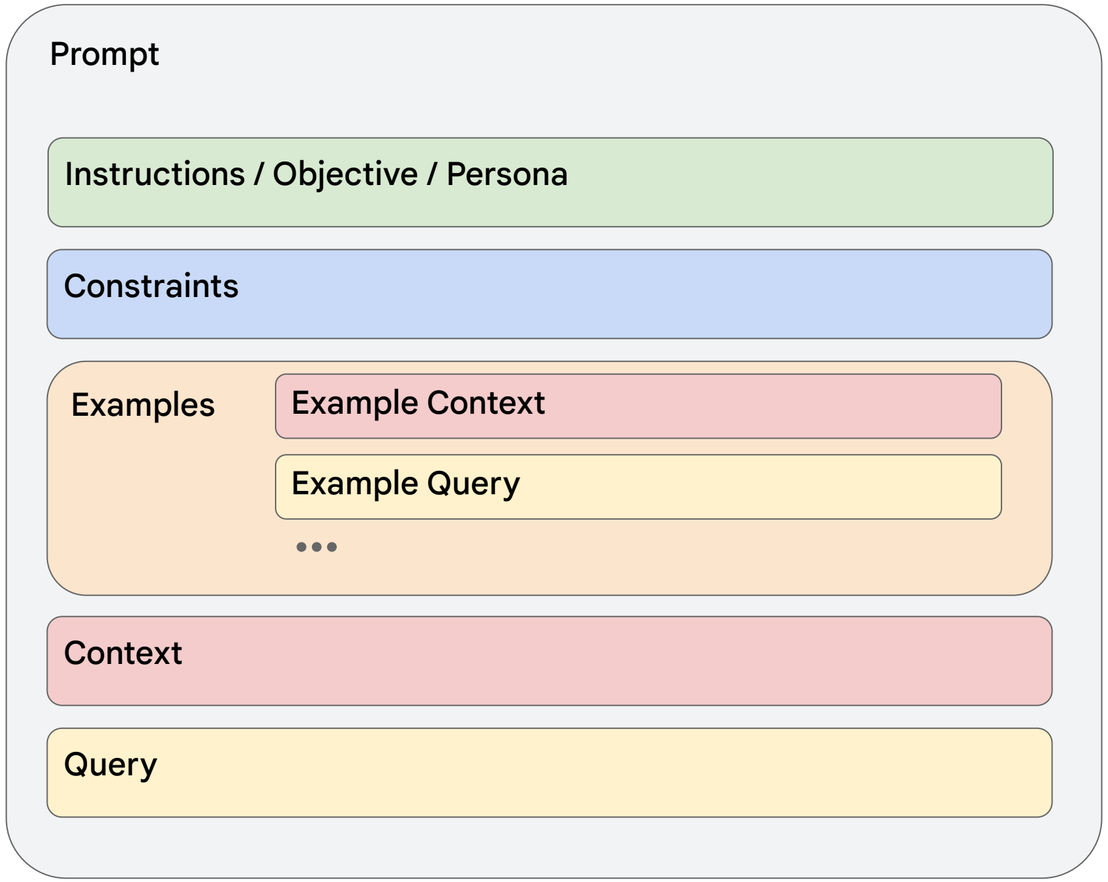
Why is prompt engineering getting so much attention?

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

Changes to a prompt make the most immediate impact on generation results

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. ...
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

Consider both content and structure in your prompt



New information can be provided as context

Context is any information provided to the model for it to fulfill its instructions. It is often:

- Specific to your organization
- Timely (like recent news or whether a product is in stock)
- Pulled from a database at query-time

A model may hallucinate without the appropriate context. Here it has no idea of the correct grocery store aisles.

PROMPT

```
On what aisle numbers can I find the following items?
    - paper plates
    - mustard
    - potatoes
```

RESPONSE.TEXT

```
## Aisle Numbers:

    Here's where you can find the items you listed:

    * **Paper plates:** Aisle 7 (usually near plastic
cups and cutlery)
    * **Mustard:** Aisle 4 (often with other condiments
like ketchup and mayonnaise)
    * **Potatoes:** Aisle 2 (look for the vegetable
section, near onions and carrots)

    **Please note:** This is a general guide based on
typical supermarket layouts. The actual aisle numbers
might vary depending on the specific store you're
visiting. It's always a good idea to check the store
directory or ask a staff member for assistance if you
can't find something.
```

But adding the context allows it to give correct results for a specific store

PROMPT

```
Context:
 Michael's Grocery Store Aisle Layout:
 Aisle 1: Fruits – Apples, bananas,  grapes,oranges,
          strawberries, avocados, peaches, etc.
 Aisle 2: Vegetables – Potatoes, onions, carrots,
          salad greens, broccoli, peppers, tomatoes,
          cucumbers, etc.
 Aisle 3: Canned Goods – Soup, tuna, fruit, beans,
          vegetables, pasta sauce, etc.
 Aisle 4: Dairy – Butter, cheese, eggs, milk,
          yogurt, etc.
 Aisle 5: Meat– Chicken, beef, pork, sausage,
          bacon etc.
 … [edited for slide display]

 Query:
 On what aisle numbers can I find the following items?
 - paper plates
 - mustard
 - potatoes
```

RESPONSE.TEXT

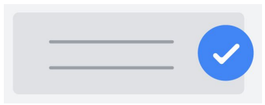
```
## Michael's Grocery Store Aisle Numbers

    Based on the aisle layout you provided, here's
where you can find the items you listed:

    * **Paper plates:** Aisle 17 (Household & Cleaning
Supplies)
    * **Mustard:** Aisle 8 (Condiments & Spices)
    * **Potatoes:** Aisle 2 (Vegetables)
```

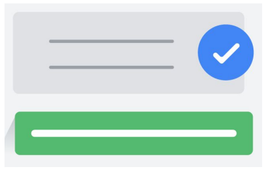
Exemplars demonstrate what output should look like

Use a history of what you have written in the past to get the model to output content in your style:



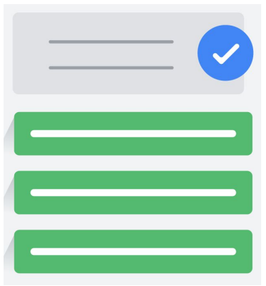
Zero-shot prompting

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").

Exemplars can demonstrate formatting and guide the models interpretation of context/queries

PROMPT

```
We offer software consulting services. Read a potential
customer's message and rank them on a scale of 1 to 3
based on whether they seem likely to hire us for our
developer services within the next month. Return the
likelihood
rating labeled as "Likelihood: SCORE".
Do not include any Markdown styling.

1 means they are not likely to hire.
2 means they might hire, but they are not likely ready
to do
so right away.
3 means they are looking to start a project soon.
```

PROMPT (CONT'D)

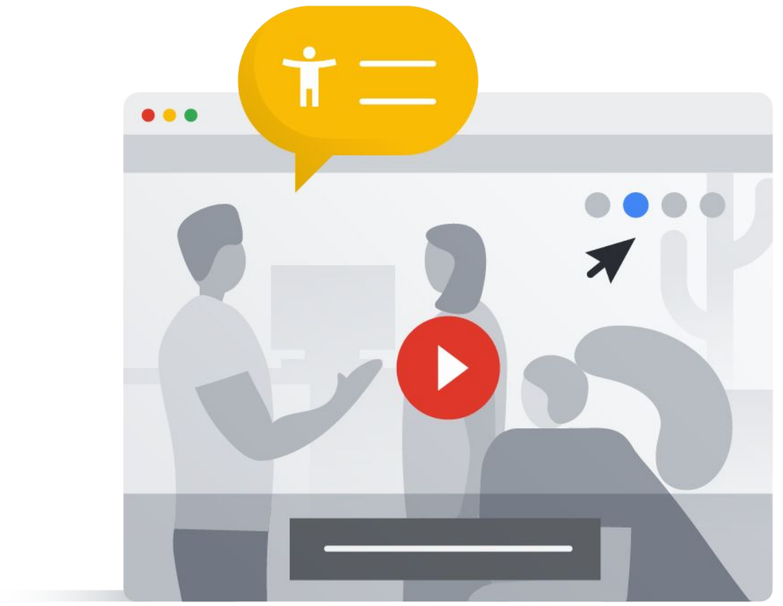
```
Example Message: Hey there I had an idea for an app,
and I have no idea what it would cost to build it.
Can you give me a rough ballpark?
Likelihood: 1

Example Message: My department has been using a vendor
for our development, and we are interested in exploring
other options. Do you have time for a discussion around
your services?
Likelihood: 2

Customer Message: Our department needs a custom gen AI
solution. We have a budget to explore our idea. Do you
have capacity to get started on something soon?
Likelihood:
```

More guidance:

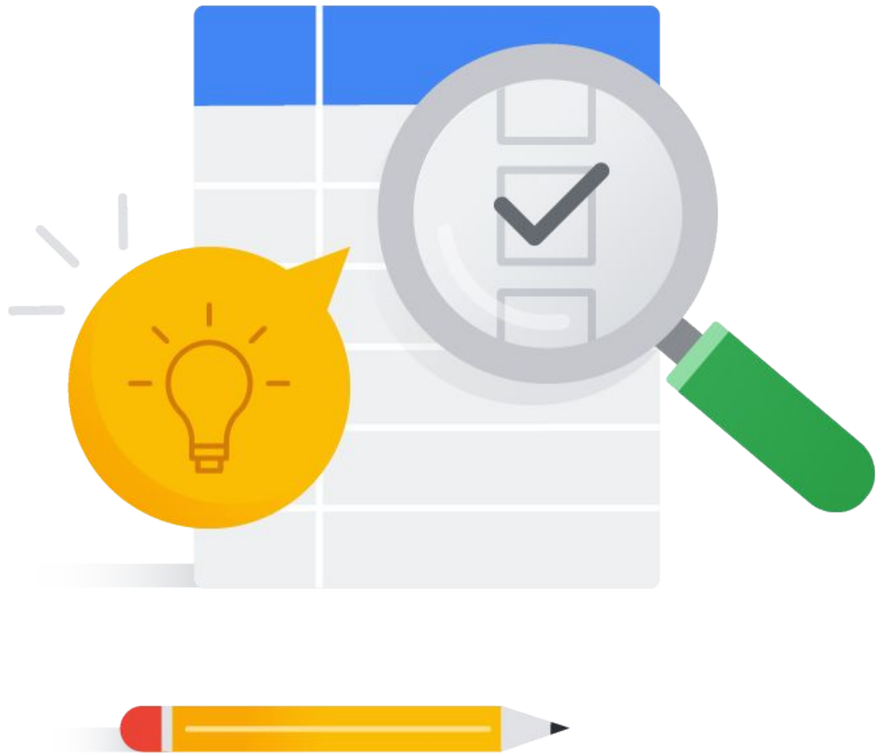
- Give clear and specific instructions
- Include few-shot examples
- Assign a role
- Add contextual information
- Use system instructions
- Structure prompts
- Instruct the model to explain its reasoning
- Break down complex tasks
- Experiment with parameter values
- Prompt iteration strategies
- Task-specific guidance



Lab

1.5 hours

Lab: Prompt Design Strategies

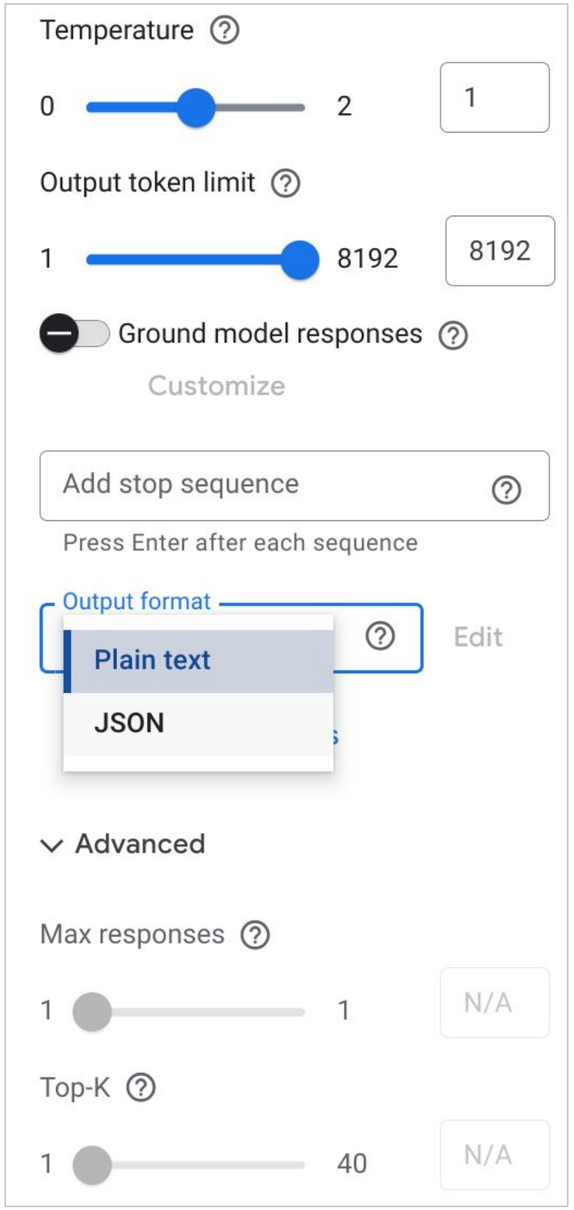


Topics

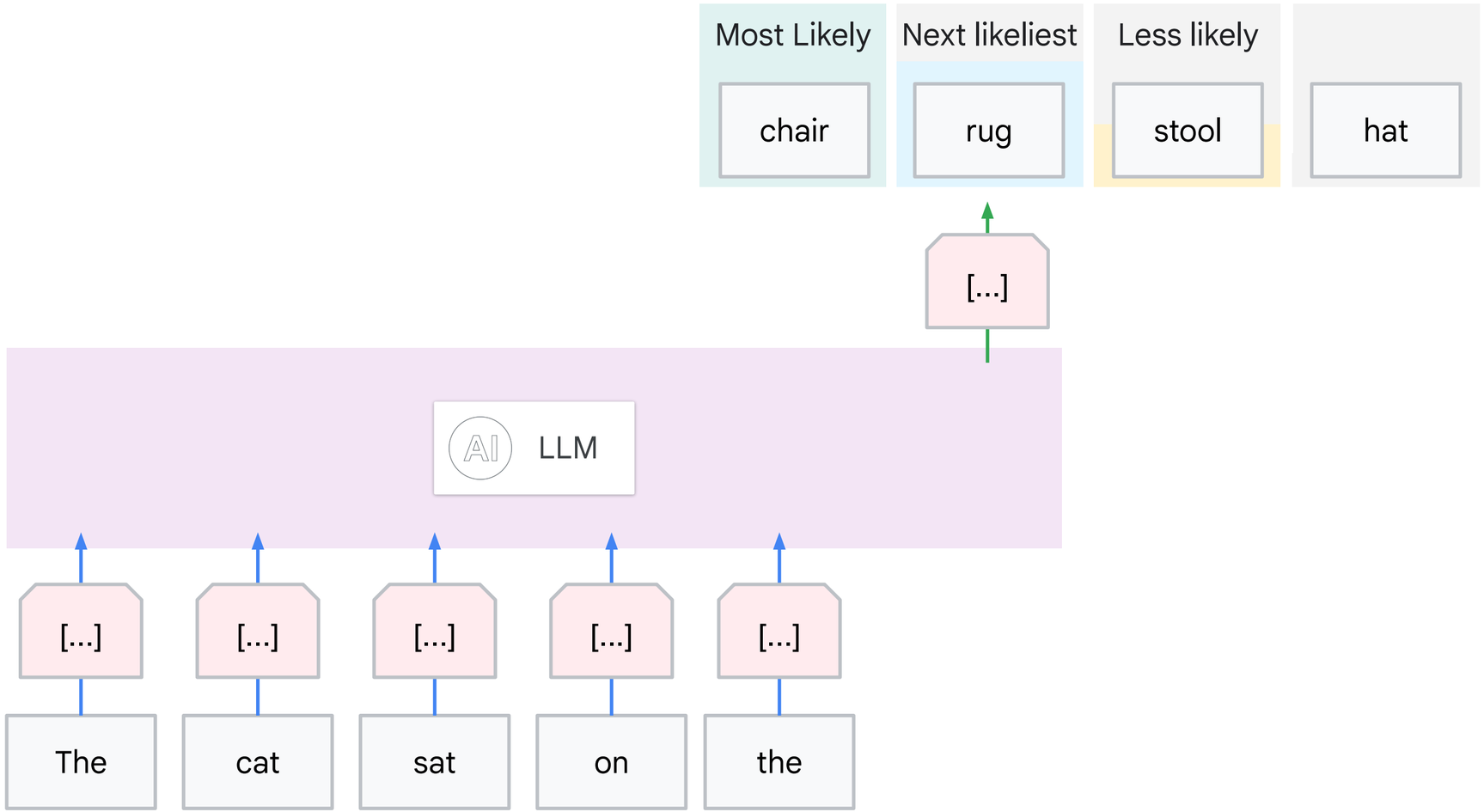
01	Prompts
02	Model Parameters
03	Advanced Prompting Techniques
04	Multimodal Prompts

Use the other model parameters to further customize the results

- Temperature
- Output token limit
- Grounding
- Stop sequences
- Output format
- Safety Filter Settings
- Seed
- Max responses
- Top-K [not used by latest versions of Gemini]
- Top-P
- Stream model responses

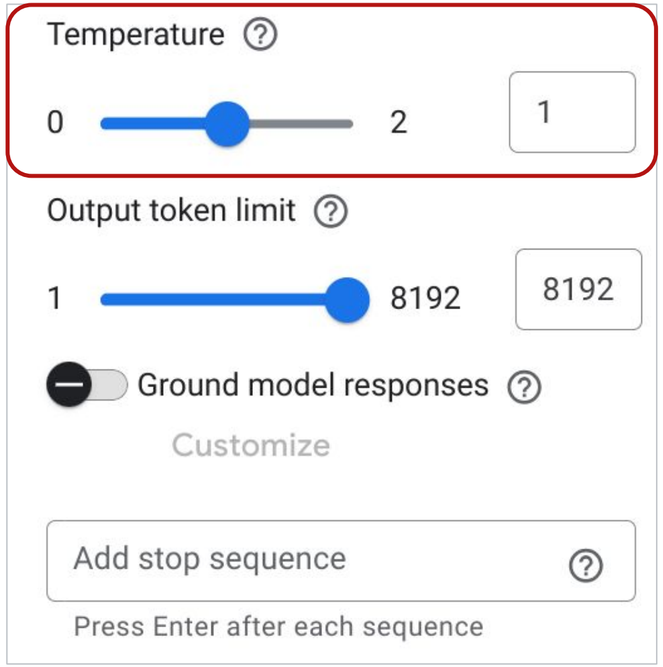


Some parameters adjust how the model adjusts & selects the next token from a probability distribution



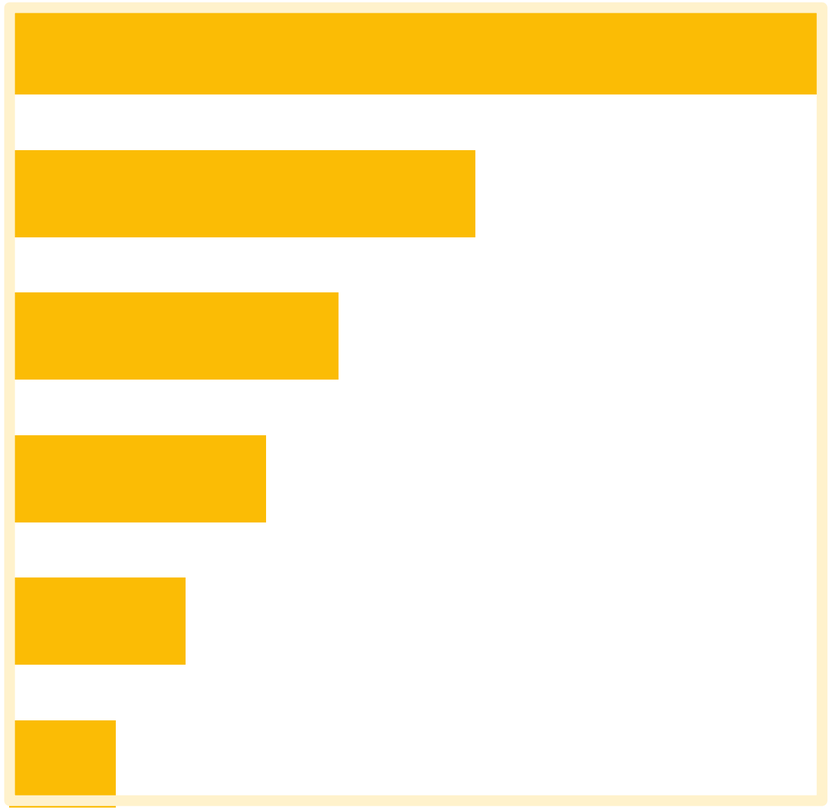
Temperature controls the degree of randomness in token selection

- Range from 0 to 2.0 for newer Gemini models. Default is 1.0
- Lower temperatures are good for prompts that require correctness. The model is more likely to select the most expected next token.
- Higher temperatures can lead to more diverse or unexpected results. More “creative” or at least less predictable.
- A temperature of 0 is deterministic
  - The highest probability token is always selected.
  - It will always return the same result for a given prompt.



Temperature adjusts the probability distribution

LOWER TEMP



HIGHER TEMP

Notice there is less distance between the most likely and least likely next token.

Your choice of 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?

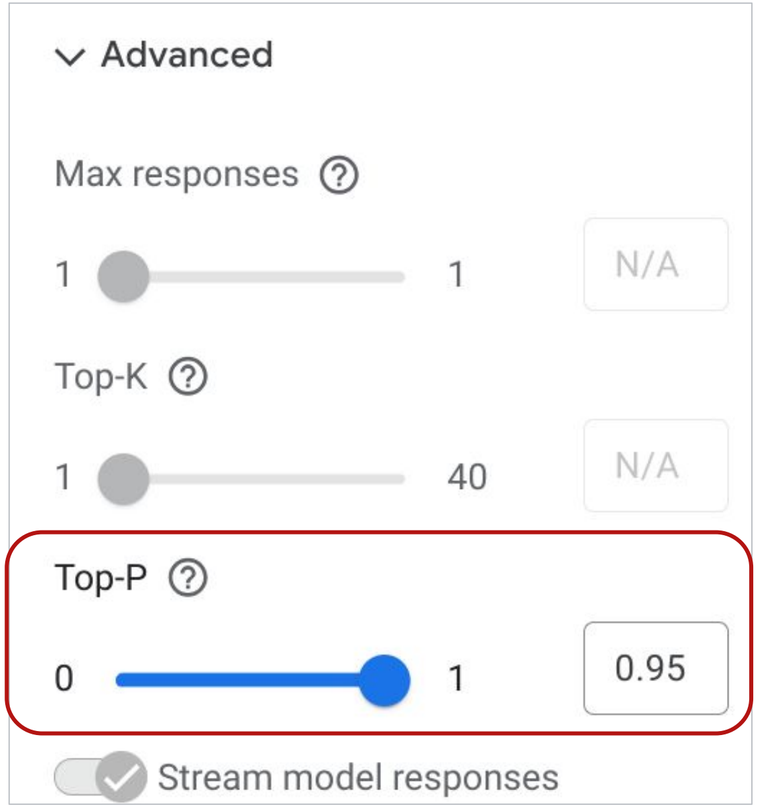
Classifying emails as Customer Service, Sales, or HR

Summarizing transcripts from an online meeting

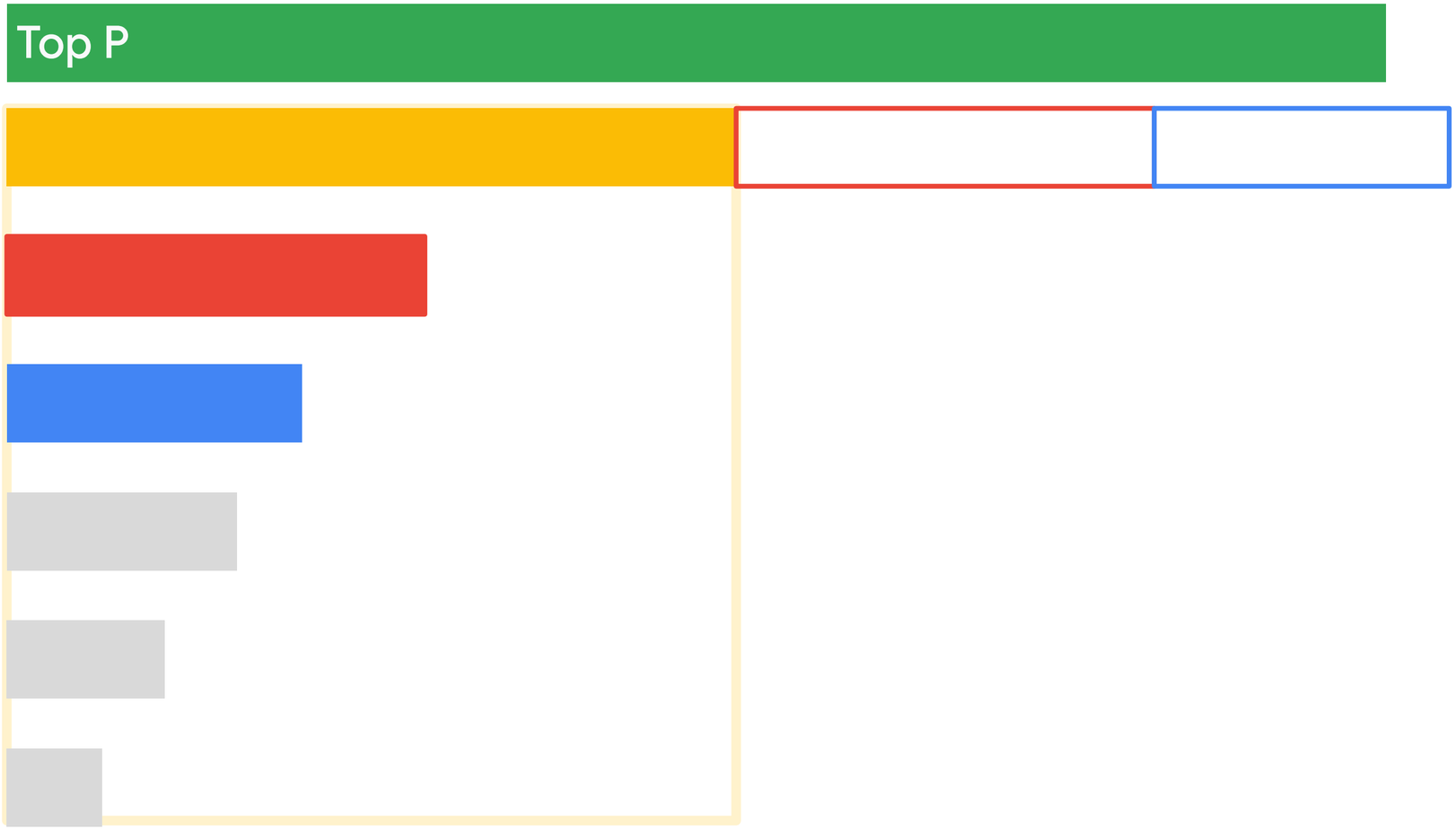
Writing social media posts for product marketing

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

- The range covers possible probabilities: 0 to 1
- The default top-P value is .8
- The higher the value, the more less-probable tokens the model will include as possibilities to select from

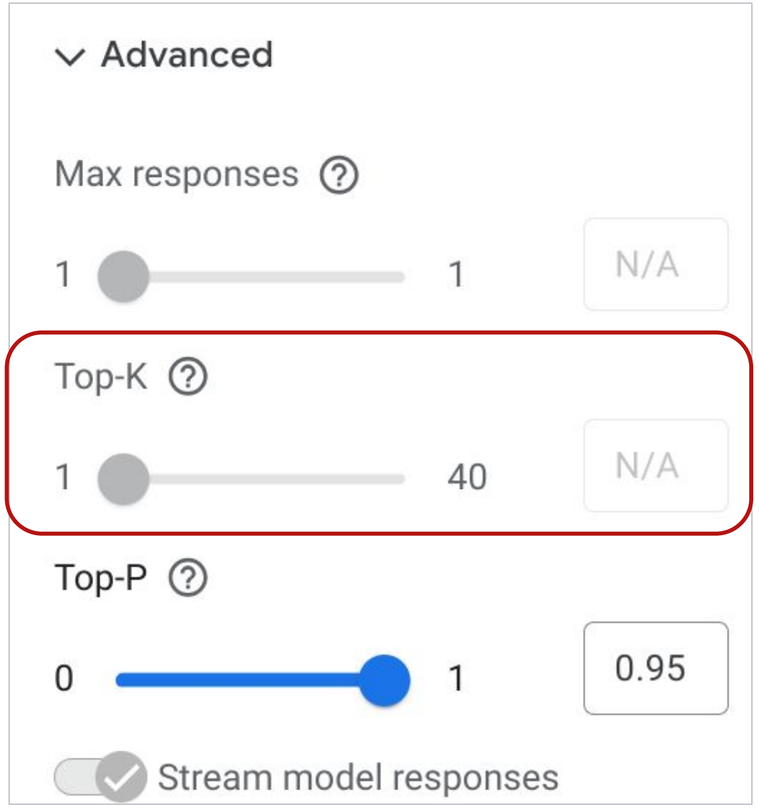


The model selects from the tokens whose probabilities sum up to at least the Top-P value



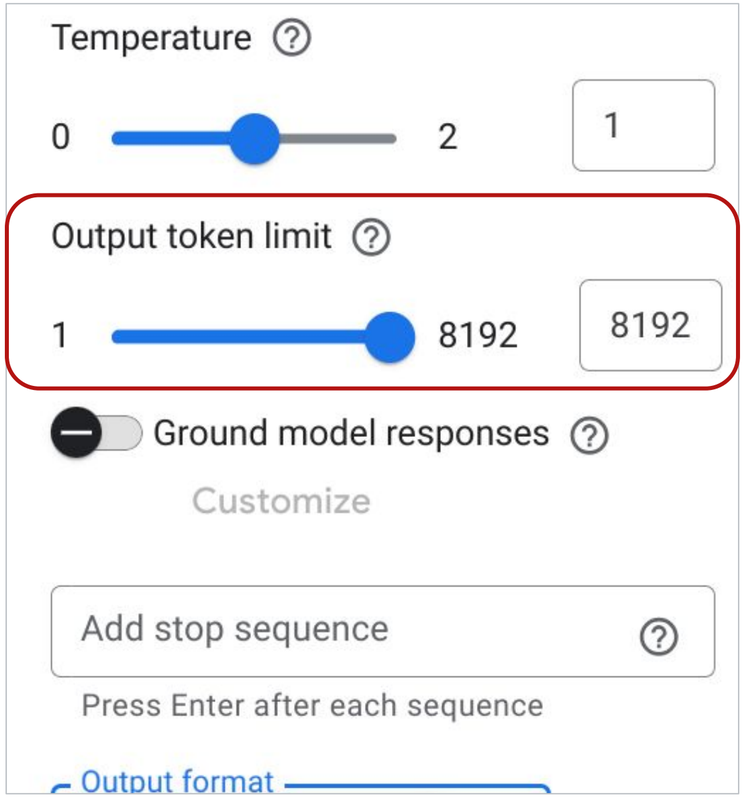
Top-K is similar, but the model selects from the top K (integer) most likely tokens

- Not used by latest Gemini models
- 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



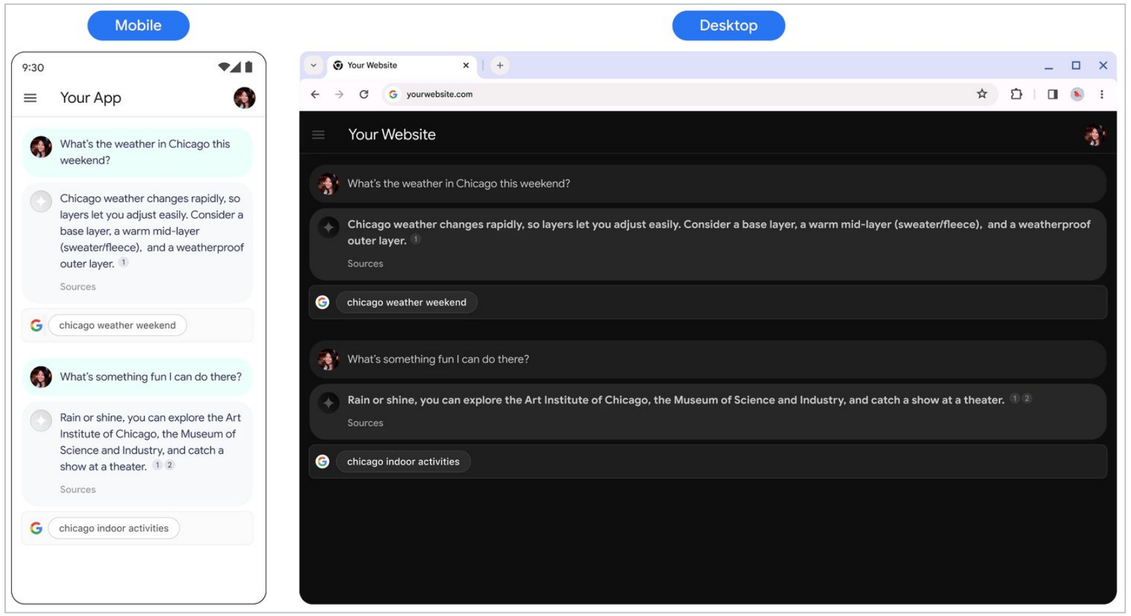
The output token limit determines the limit at which output will be truncated. The model is not aware of it.

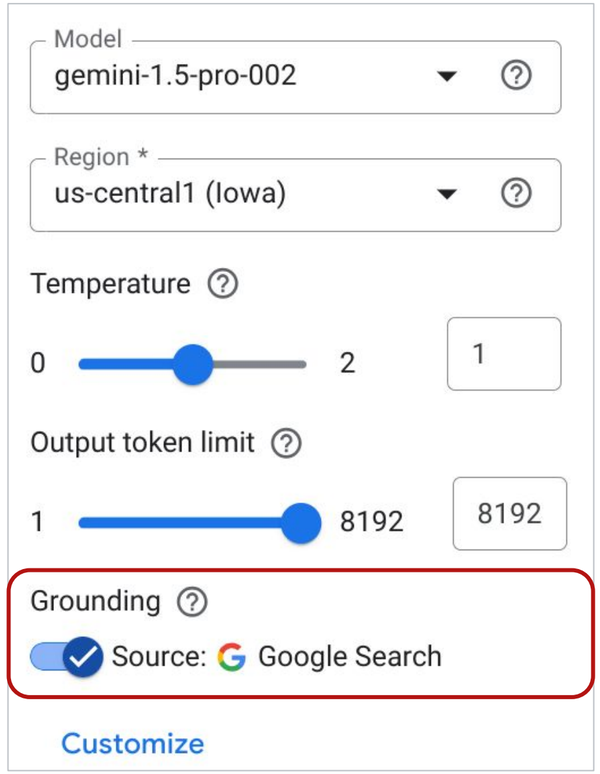
- The range is dependent on the model
- The model is **not aware of this limit**, so you should also guide the model via your prompt instructions to keep its answers under this limit.
- A token is a basic unit of text that the LLM understands
  - A token is approximately four characters
- A token represents a concept or object in GenAI
  - For example, the token "dog" represents the concept of a dog



Grounding provides your model access to Google Search or your own Vertex AI Search data store.

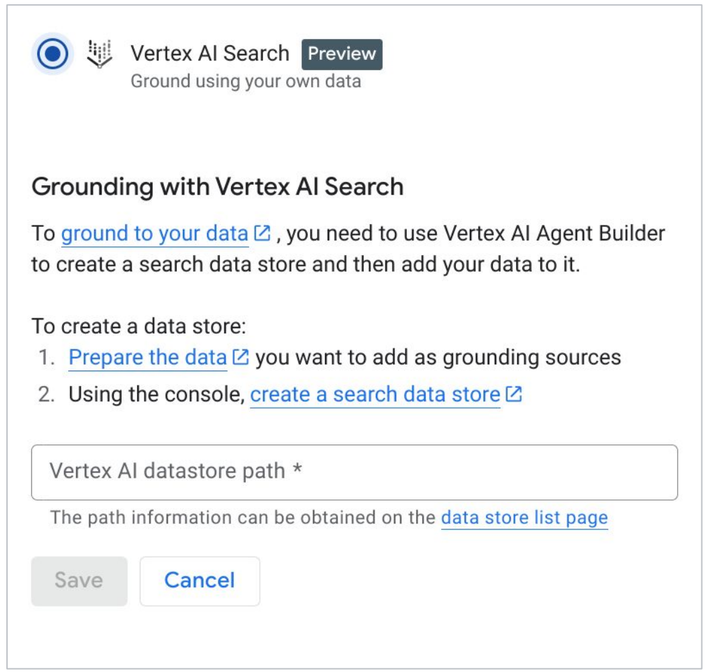
If Google Search is used, you must display the search suggestion exactly as returned, and if the user clicks, take them to the search result page for that search.

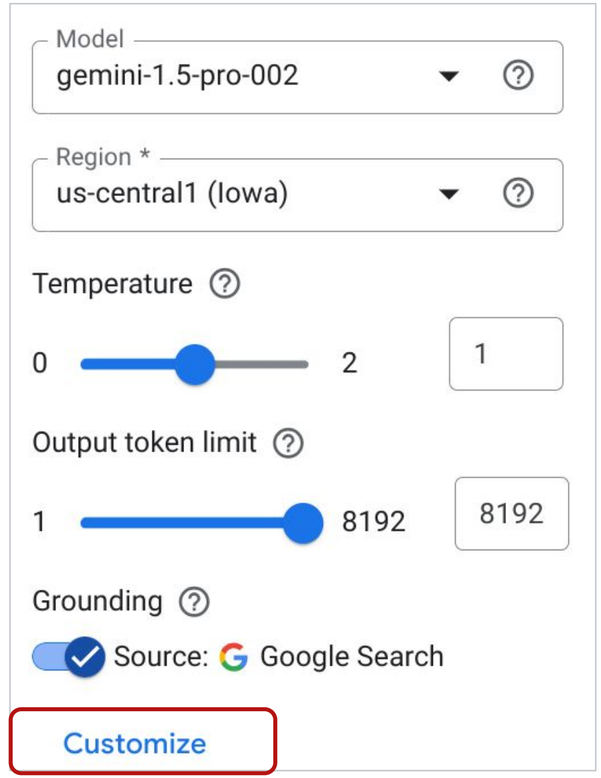


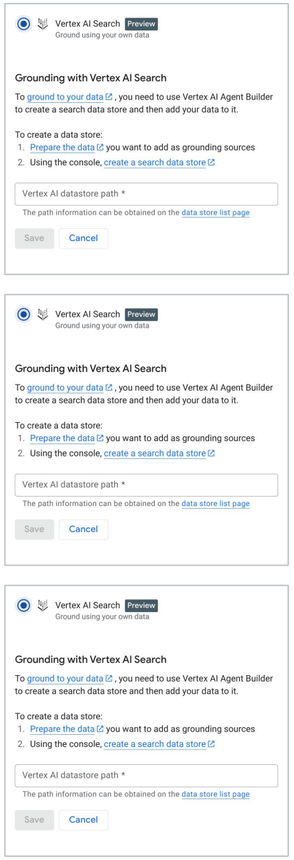


If grounding on your own data, you need to create a Vertex AI Search data store within Agent Builder

We'll cover these later, but this is effectively your own data uploaded, chunked, and processed to be stood as content able to be retrieved by a model.



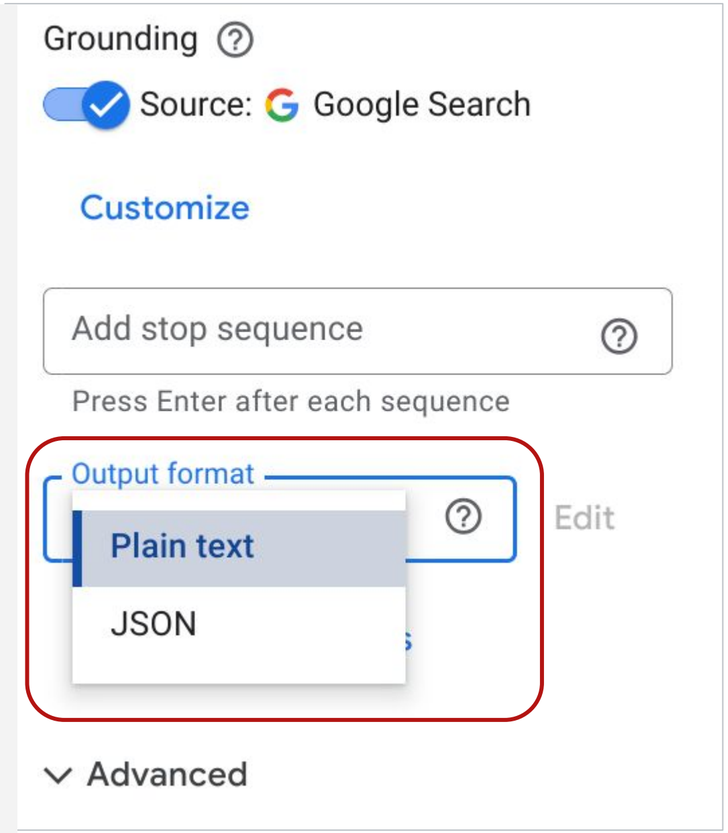




Output format allows you to select a JSON mode for “controlled generation” according to a schema.

```
response_schema = {"type": "array",
                   "items": {
                       "type": "object", "properties": {
                           "recipe_name": {"type": "string",},
                       },
                       "required": ["recipe_name"],
                    },}


model = GenerativeModel("gemini-1.5-pro-002")
response = model.generate_content("List a few popular cookie recipes",
    generation_config=GenerationConfig(
        response_mime_type="application/json",
        response_schema=response_schema
    ),)
```

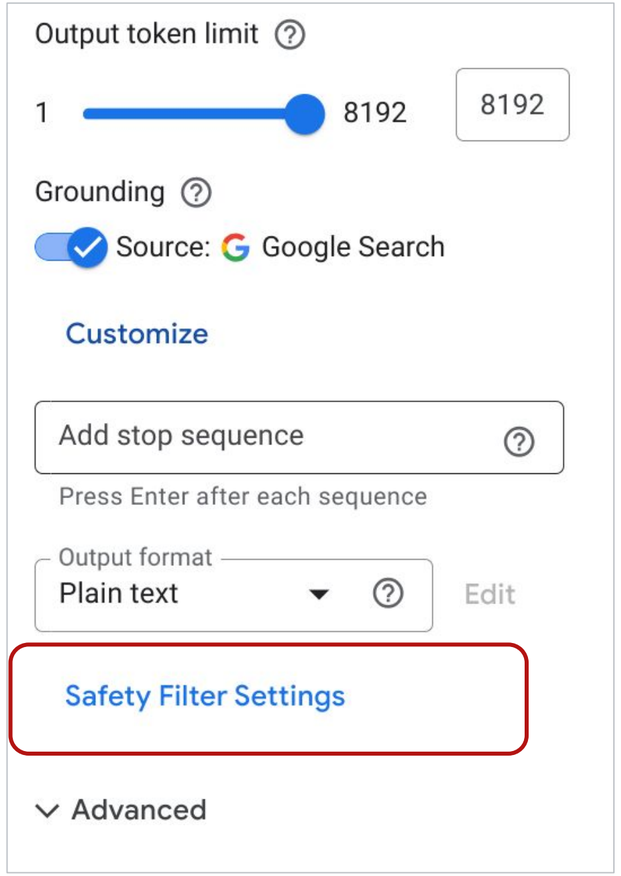


Safety Filter Settings allow you to check generated results for certain categories before returning

Harm categories

Safety filters assess content based on the following harm categories:

Harm Category	Definition
Hate Speech	Negative or harmful comments targeting identity and/or protected attributes.
Harassment	Threatening, intimidating, bullying, or abusive comments targeting another individual.
Sexually Explicit	Contains references to sexual acts or other lewd content.
Dangerous Content	Promotes or enables access to harmful goods, services, and activities.

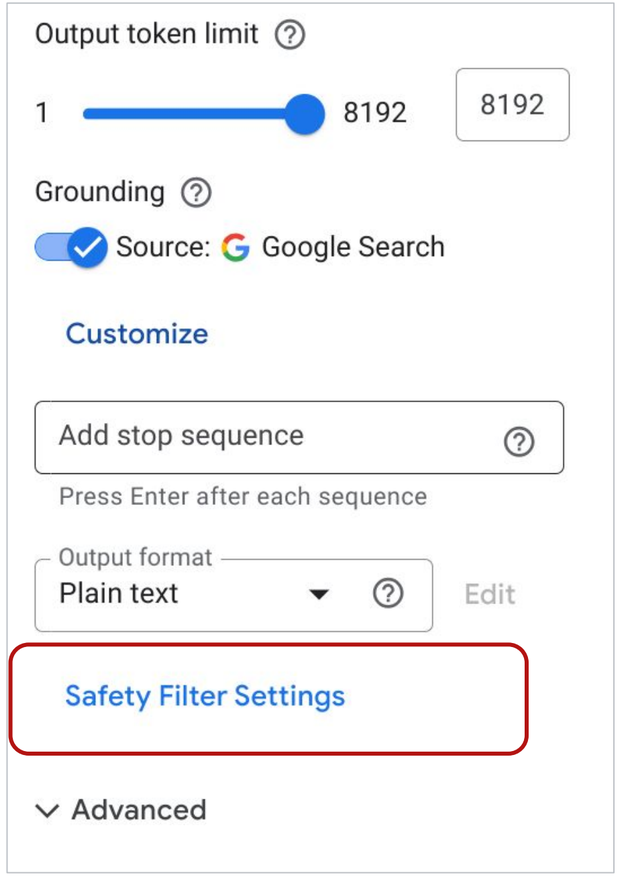


You may want to set them not to block content by setting them to Block Only High probability risks.

The Vertex AI Gemini API provides the following "harm block" thresholds:

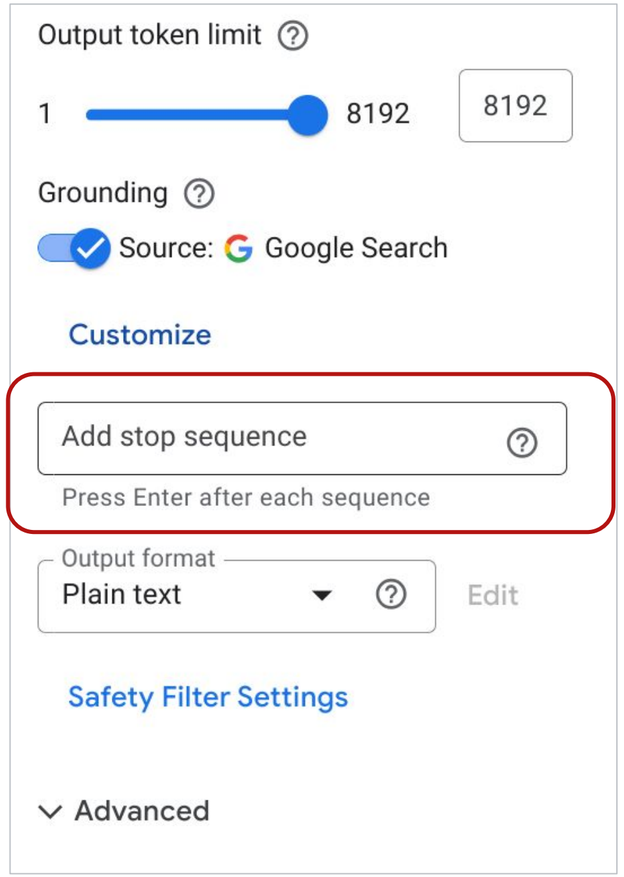
- **BLOCK_LOW_AND_ABOVE**: Block when the probability score or the severity score is `LOW`, `MEDIUM` or `HIGH`.
- **BLOCK_MEDIUM_AND_ABOVE**: Block when the probability score or the severity score is `MEDIUM` or `HIGH`. For `gemini-1.5-flash-001` and `gemini-1.5-pro-001`, `BLOCK_MEDIUM_AND_ABOVE` is the default value.
- **BLOCK_ONLY_HIGH**: Block when the probability score or the severity score is `HIGH`.
- **HARM_BLOCK_THRESHOLD_UNSPECIFIED**: Block using the default threshold.
- **OFF**: No automated response blocking and no safety metadata is returned. For `gemini-1.5-flash-002` and `gemini-1.5-pro-002`, `OFF` is the default value.
- **BLOCK_NONE**: The `BLOCK_NONE` safety setting removes automated response blocking. Instead, you can configure your own safety guidelines with the returned scores. This is a restricted field that isn't available to all users in GA model versions.

★ **Note:** Access to the **`BLOCK_NONE`** or **`OFF`** settings for Gemini 1.5 Flash 001, Gemini 1.5 Pro 001, and prior models is restricted. To get access, apply through the Gemini safety filter allowlist form, or switch your account type to monthly invoiced billing with the Google Cloud invoiced billing reference.



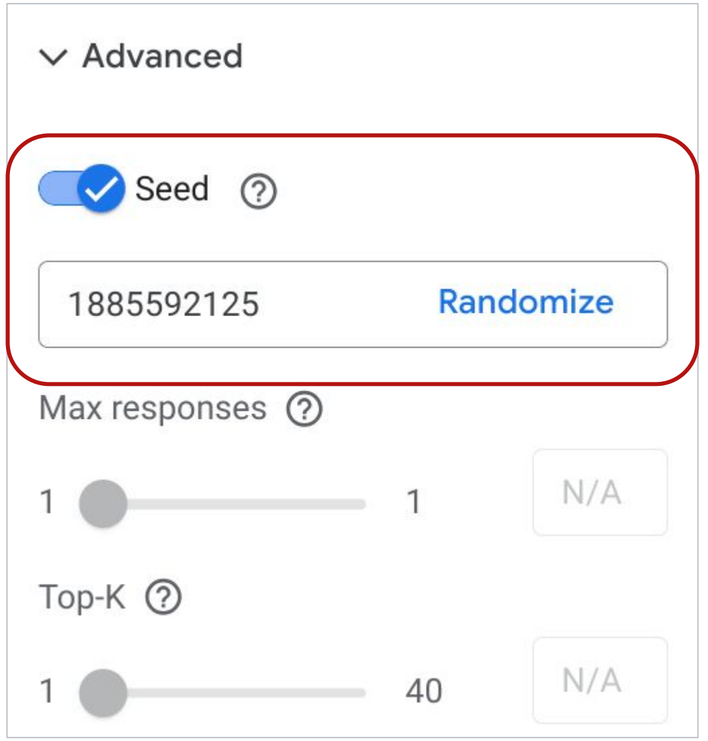
Stop sequences are hard brakes. If a stop sequence string is encountered in the response, it stops.

- You can set up to 5 stop sequences.
- You could combine this with fallback sequences, so your model can say a term to stop generation, and then you handle the error.



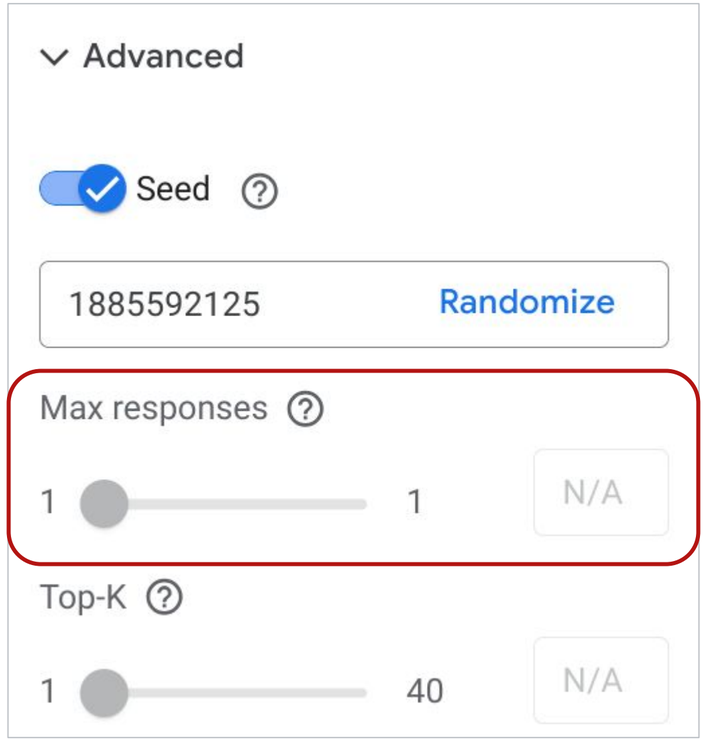
Setting a Seed can help produce repeated results when generating multiple times with the same settings.

- Deterministic output is not guaranteed, as a high temperature, for example, can still cause different output to be generated.



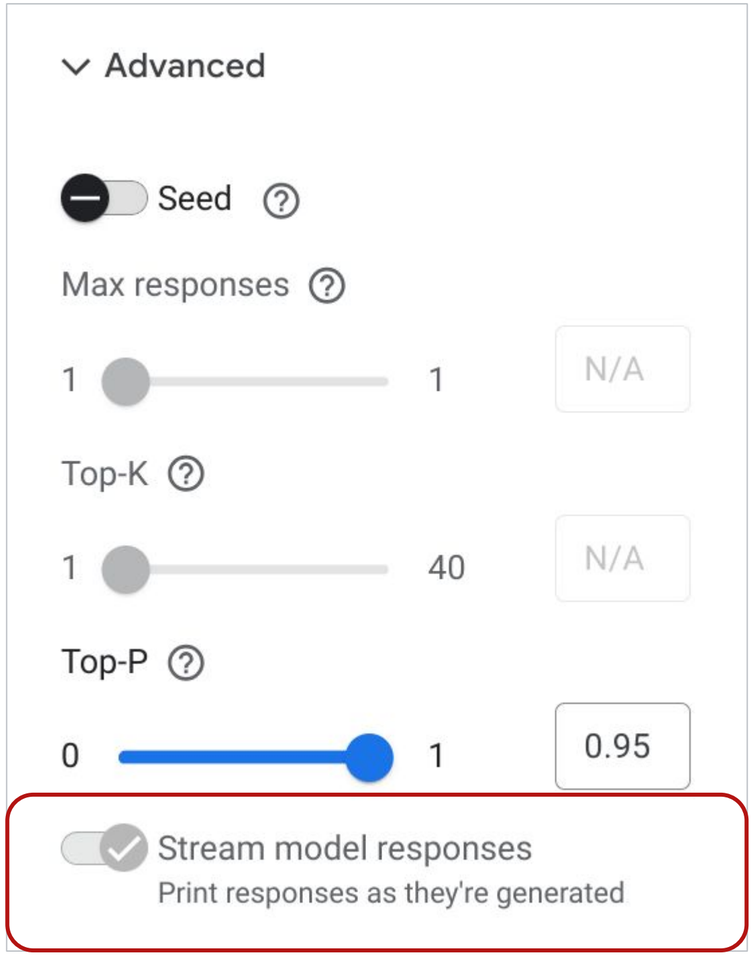
Max responses determines the number of results the model returns for a given prompt

- Range is from 1 to 8, depending on the model.
- You could generate multiple responses in a batch and allow a user to choose their preferred response
- You are billed for each response
- Not used with current Gemini models



Stream enabled means the model sends responses in small chunks rather than all at once.

- It's on by default in the Vertex AI Studio UI, but can be disabled by removing it or setting it to `false` in the code.
- This is useful in chat applications to give the user the sense of the model actively responding.



Topics

01	Prompts
02	Model Parameters
03	Advanced Prompting Techniques
04	Multimodal Prompts

Chain of Thought prompting

- The prompt provides examples of the steps the LLM should take when coming up with an answer
- Proven to help language models be more accurate for complex tasks
- Newer versions of Gemini are good at exhibiting this behavior without needing the examples

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:

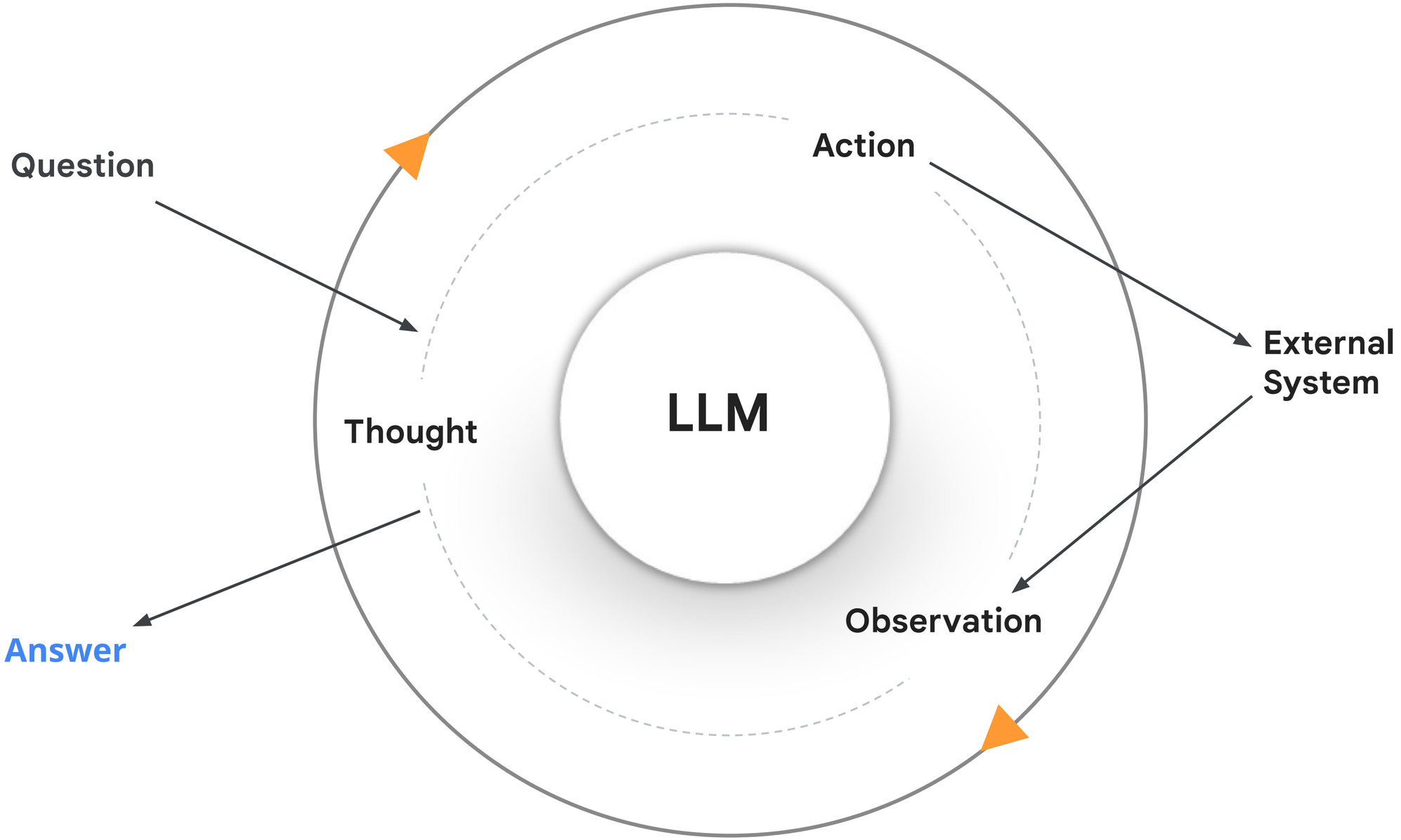
1. Start with a captivating question or statement that highlights the importance of your topic.
2. Introduce a personal anecdote or story that illustrates the positive impact that the topic has had on your life.
3. Share a surprising or intriguing fact about the topic backed by scientific evidence, to 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 response

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 is seen in many more complex generative AI patterns

Two examples of ReAct in wide use:

- RAG
- Function Calling

Topics

01	Prompts
02	Model Parameters
03	Advanced Prompting Techniques
04	Multimodal Prompts

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"}

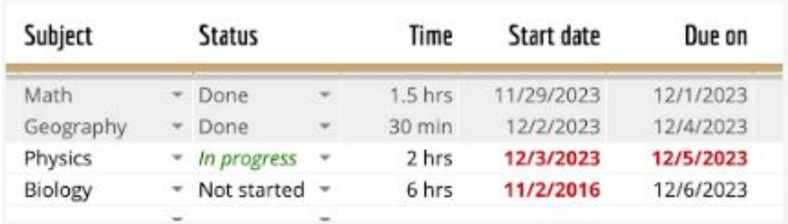
Chain of thought example

1. 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.

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.

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 |

Multimodal prompts allow you to combine text with image content (with video and 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 engineering best-practices are similar to language prompts, with some extra advice:

- 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)

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, Prompt Chaining, and Function Calling.

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