Welcome to the Gemini era

Gemini is built from the ground up for multimodality — reasoning seamlessly across text, images, video, audio, and code.

Google Gemini

Course Pathways

How to get the most out of this course

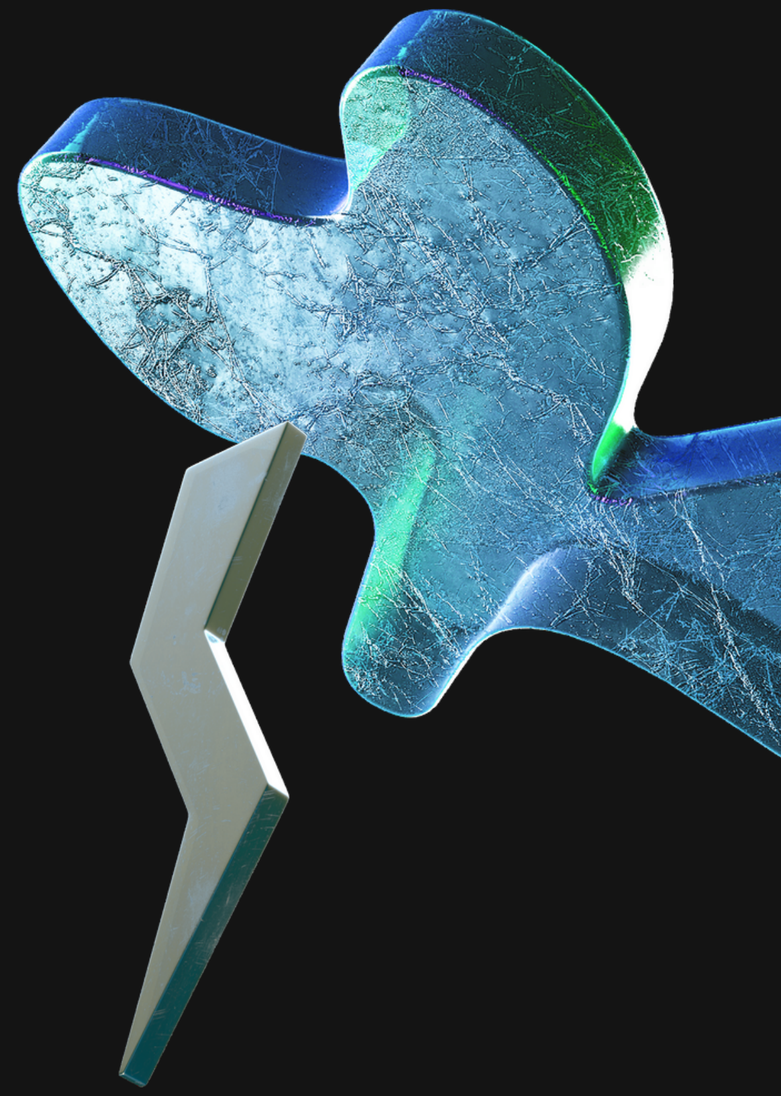
Non-Developer Path

For enthusiasts and every day AI users

BARD

PROMPT ENGINEERING

GOOGLE AI STUDIO



Developer Path

For Python & Node JS Developers

BARD

PROMPT ENGINEERING

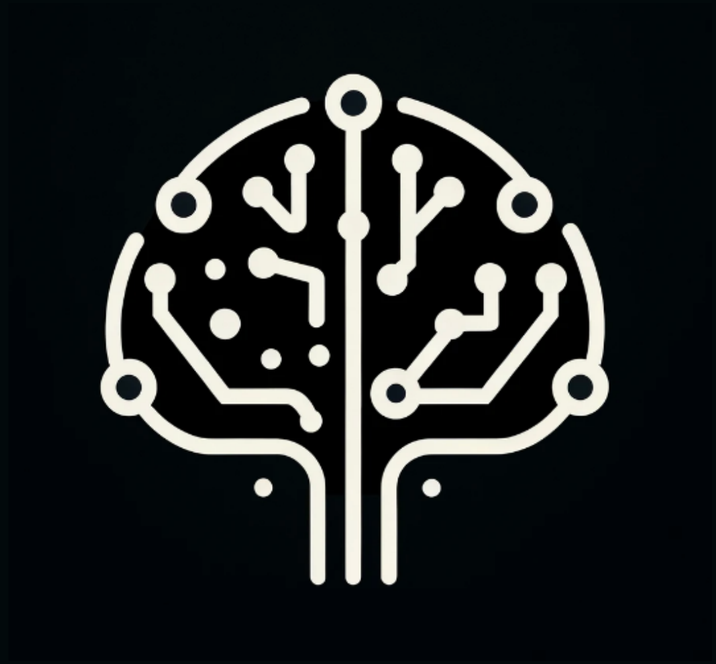
GOOGLE AI STUDIO

GEMINI SDK

GOOGLE VERTEX AI

Large Language Models

- Historically: text input and text output
- LLMs essentially “predict” the next likely “word” when generating outputs, kind of like very sophisticated autocomplete tools
- Models are trained on vast amounts of text data to understand context, nuance, and the complexities of human language.
- Most models are based on a **Transformer** architecture and employ **Attention** mechanisms, allowing the model to focus on different parts of the input sequence when generating each part of the output

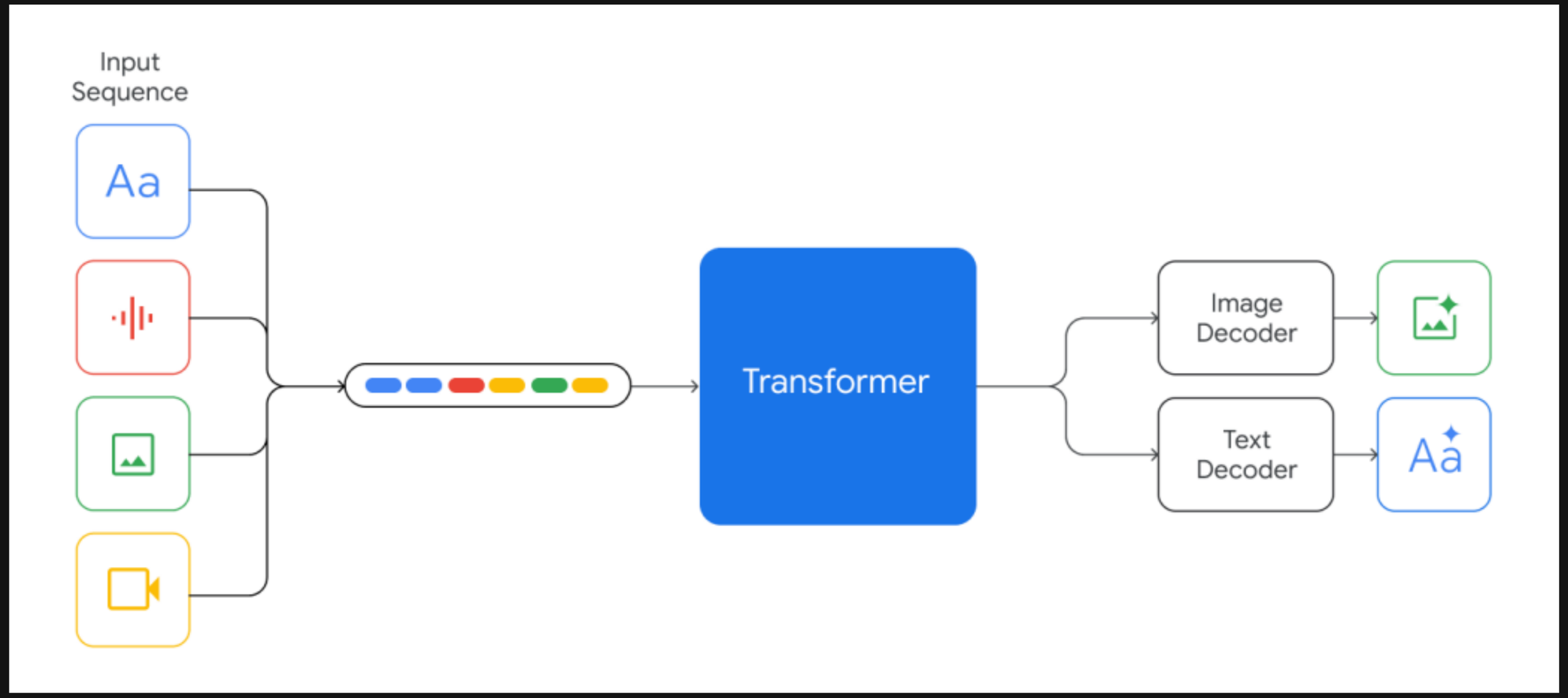


Attention

- Attention allows the network to focus on parts of the input data and dynamically asjust its focus as it processes the data
- There are different attention mechanisms but most involve computing an attention score for each piece of input data
- These scores are then used to compute a weighted sum or average of the input elements

Multi Modal Models

- Multimodal models like Gemini support many different modalities of input data (and output data) including:
  - Text
  - Code
  - Audio
  - Image
  - Video
- Gemini models are trained to accommodate textual input interleaved with a wide variety of audio and visual inputs, such as natural images, charts, screenshots, PDFs, and videos, and they can produce text and image outputs


Input
Sequence
Aa
Transformer
Image
Decoder
Text
Decoder
Aa

Tokens

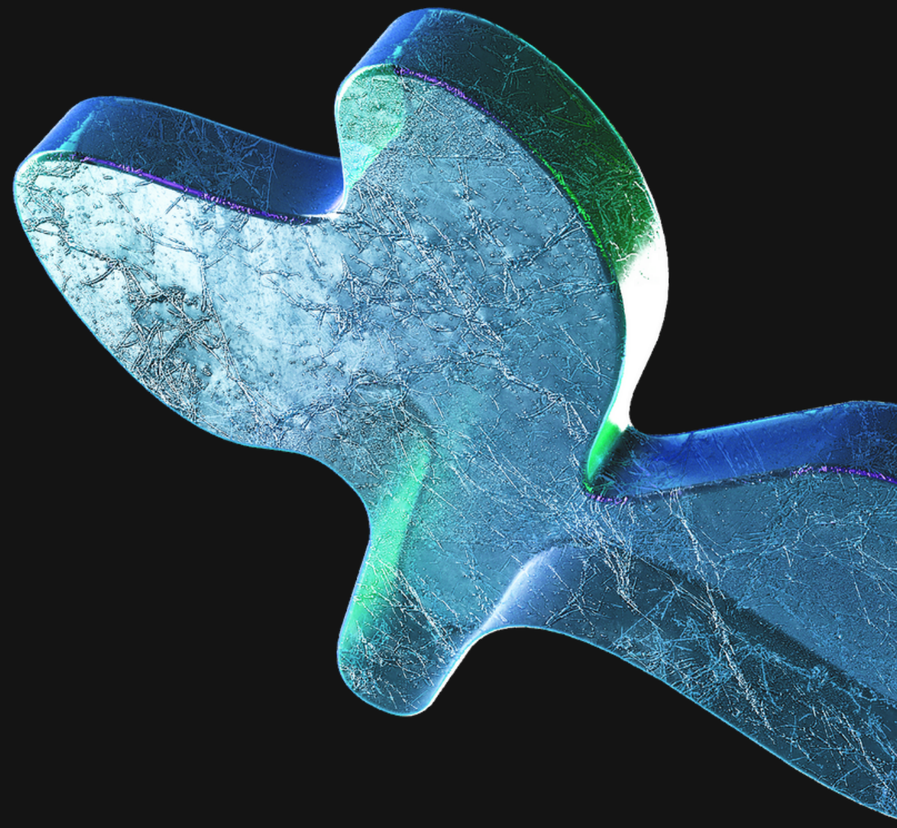
- LLMs don't "think" in words
- Your input text (and images) are instead tokenized
- Input text is split into a series of small chunks of text that are turned into numerical vectors
- Gemini uses a tokenizer called **SentencePiece**

My Cat Is Annoying → [5159, 17810, 374, 1556, 2201, 7169]

Google Gemini

The 3 Models

What They Are And How They Compare



Gemini comes in three sizes

Our most capable and largest model for highly-complex tasks.

Our best model for scaling across a wide range of tasks.

Our most efficient model for on-device tasks.

Gemini Nano

- The most efficient model, designed to run on mobile devices.
- Google trained two versions of Nano, with 1.8B (Nano-1) and 3.25B (Nano-2) parameters, targeting low and high memory devices respectively.
- Android developers can build with Gemini Nano for on-device tasks,

Gemini Pro

- The mid-tier option
- A model optimized for performance, cost, and latency
- Does well across a wide range of tasks: strong reasoning and broad multimodal capabilities.
- “our best model for scaling across a wide range of tasks“
- Used By Bard today
- Eventually Gemini Pro will be available in Google products like Search, Ads, Chrome and Duet AI.

Gemini Ultra

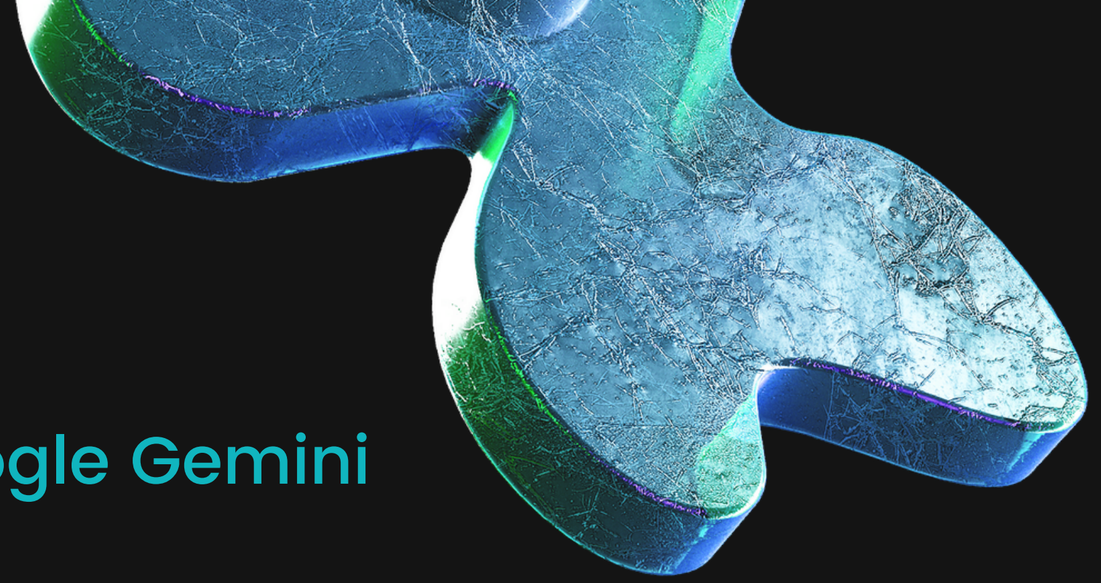
- The most capable model that delivers “state-of-the-art performance across a wide range of highly complex tasks”
- Excels at reasoning and multimodal tasks.
- our largest and most capable model for highly complex tasks.
- Gemini Ultra is currently undergoing trust and safety checks, fine tuning, and model refinement.
- Will eventually be available to “select customers, developers, partners and safety and responsibility experts for early experimentation and feedback”

Google Gemini

API Pricing

Are we going to go broke??

Free for everyone	Pay-as-you-go
Rate Limits 60 QPM (queries per minute)	Rate Limits Starts at 60 QPM
Price (input) Free	Price (input) $0.00025 / 1K characters $0.0025 / image
Price (output) Free	Price (output) $0.0005 / 1K char
Input/output data used to improve our products Yes	Input/output data used to improve our products No
Get API key in Google AI Studio	Coming soon to Google AI Studio



Google Gemini

Bard

Let’s get a taste of Gemini!

Google Gemini

Google AI Studio

An Awesome Way To Learn And Play With Gemini

Google AI Studio

Get API key

Create new

My library

No prompts yet

Getting started

Documentation

Prompt gallery

Discord community

AI Studio in Vertex

Settings

Untitled prompt

Save

Get code

Insert: Image {{ }} Test input

Run settings

Reset

Identify the city where this photo was taken and give a brief description of the city

Tab

Model

Temperature

0.25

Add stop sequence

Add stop...

Max outputs

1

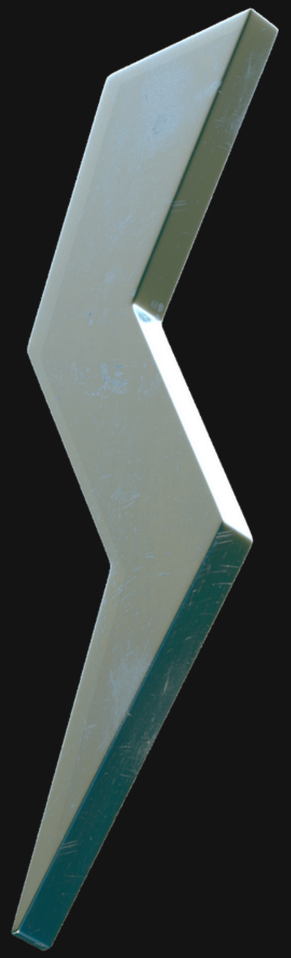
Run

Text preview 0 / 0

Google AI Studio

What is it???

"The fastest way to start building AI"



- Browser Based IDE
- Made for Prototyping
- Try Out Models and Experiment
- Supports Gemini Pro & Vision Pro
- You can Export to Code
- Completely Free!

Google Gemini

Vertex AI

Google’s Enterprise AI platform

Search models

Suggestions text embedding essay outline BERT

Modalities

Language	40
Vision	79
Tabular	2
Document	2
Speech	1
Video	4
Multimodal	1

Tasks

Generation	41
Classification	44
Detection	27
Extraction	10
Recognition	10
Translation	5
Embedding	3
Segmentation	4
Retrieval	1
Open vocabulary detection	2
Open vocabulary segmentation	2
Tracking	1

Features

Vertex AI Studio	9
API available	9
Open source	81
Notebook support	79
Pipeline support	2
One-click deployment	15

Foundation models

Pre-trained multi-task models that can be further tuned or customized for specific tasks. Models marked with ✦ are available in Vertex AI Studio.

✦ Generative AI · Language
Gemini Pro
The best performing Gemini model with features for a wide range of tasks
google/gemini-pro
VIEW DETAILS

✦ Generative AI · Multimodal
Gemini Pro Vision
Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks
google/gemini-pro-vision
VIEW DETAILS

✦ Generative AI · Language
PaLM 2 Text Bison
Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content...
text-bison@002
VIEW DETAILS

Foundation · Language
Llama 2
Fine-tune & deploy Meta's Llama 2 models on Vertex AI.
Llama2-7B-001
VIEW DETAILS

Foundation · Language
Embeddings for Text
Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...
textembedding-gecko@003
VIEW DETAILS

✦ Generative AI · Vision
Imagen for Image Generation and Editing
google/imagegeneration-005
VIEW DETAILS

✦ Generative AI · Speech
Chirp
Chirp is a version of a Universal Speech Model that has over 2B parameters and can transcribe in over 100 languages in a single model.
chirp-rnnt1
VIEW DETAILS

SHOW ALL (56)

Fine-tunable models

Models that data scientists can further fine-tune through a custom notebook or pipeline.

Classification · Vision
tfhub/EfficientNetV2
EfficientNet V2 are a family of image classification models, which achieve better parameter efficiency and faster training speed than prior arts.
tensorflow-hub/efficientnetv2
VIEW DETAILS

Classification · Vision
tfvision/vit
The Vision Transformer (ViT) is a transformer-based architecture for image classification.
tfvision/vit-s16
VIEW DETAILS

Detection · Vision
tfvision/SpineNet
SpineNet is an image object detection model generated using Neural Architecture Search.
tfvision/spinenet49
VIEW DETAILS

Detection · Vision
tfvision/YOLO
YOLO algorithm is a one-stage object detection algorithm that can achieve real-time performance on a single GPU.
tfvision/scaled-yolo
VIEW DETAILS

Classification · Vision
ResNet (with checkpoint)
Image classification model as described in the paper "Deep Residual Learning for Image Recognition".
resnet50
VIEW DETAILS

Classification · Vision
ResNet-RS (with checkpoint)
Image classification model as described in the paper "Revisiting ResNets: Improved Training and Scaling Strategies".
ResNet-RS-50
VIEW DETAILS

Segmentation · Vision
DeepLabv3+ (with checkpoint)
Semantic segmentation is the task of assigning a label to each pixel in an image, where each label corresponds to a specific class of object or scene element.
deeplabv3plus-cityscapes-20230315
VIEW DETAILS

SHOW ALL (43)

Task-specific solutions

Most of these pre-built models are ready to use off the shelf, and many can be customized using your own data.

Detection · Vision
Occupancy analytics

Classification · Language
Entity analysis

Detection · Vision
Person/vehicle detector

Classification · Language
Content classification

Classification · Language
Sentiment analysis

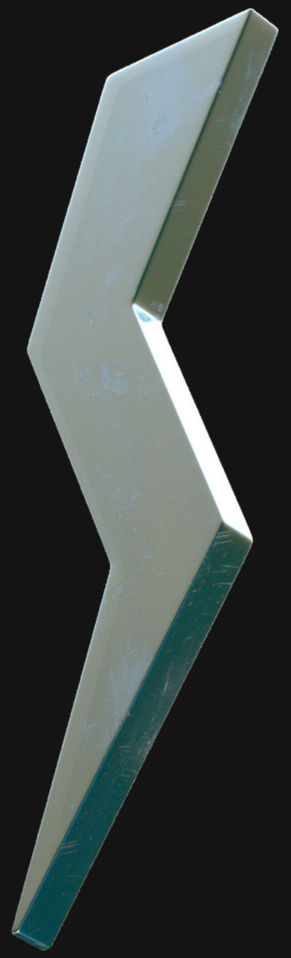
Classification · Language
Entity sentiment analysis

Extraction · Language
Syntax analysis

Google Vertex AI

What is it???

Google's AI platform designed with scalability in mind, for enterprise operations

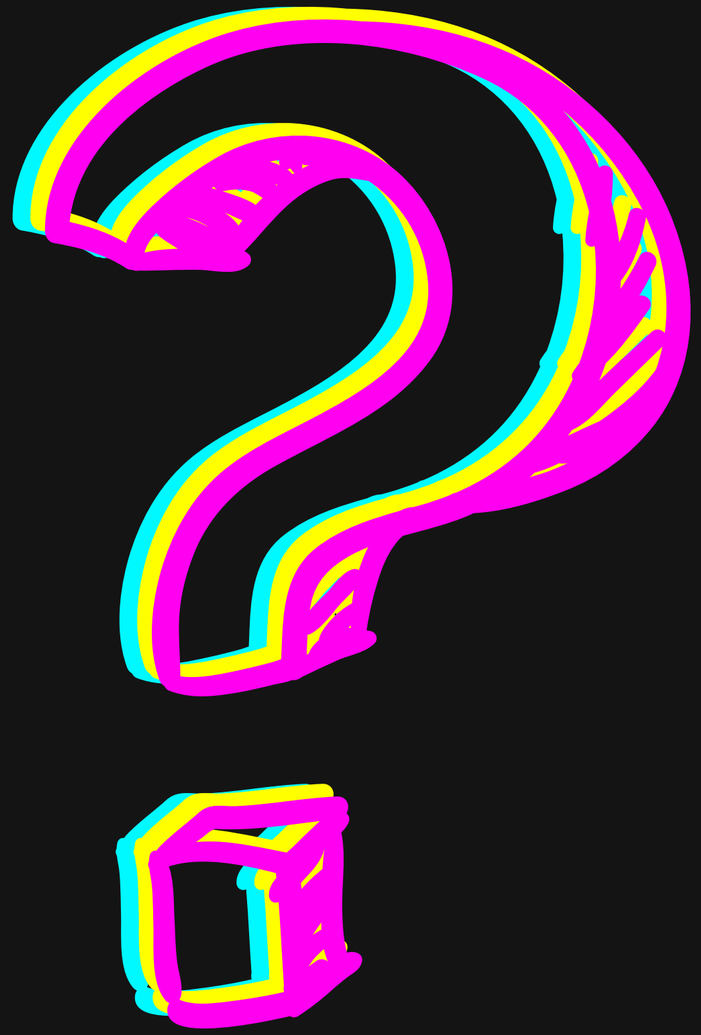


- Enterprise-Grade
- Access to many more models
- Integrated With Google Cloud
- Features For Developing Models
- Features For Deploying Models
- Pay As You Go Pricing

AI Studio Vs. Vertex AI

It’s a Little Confusing...

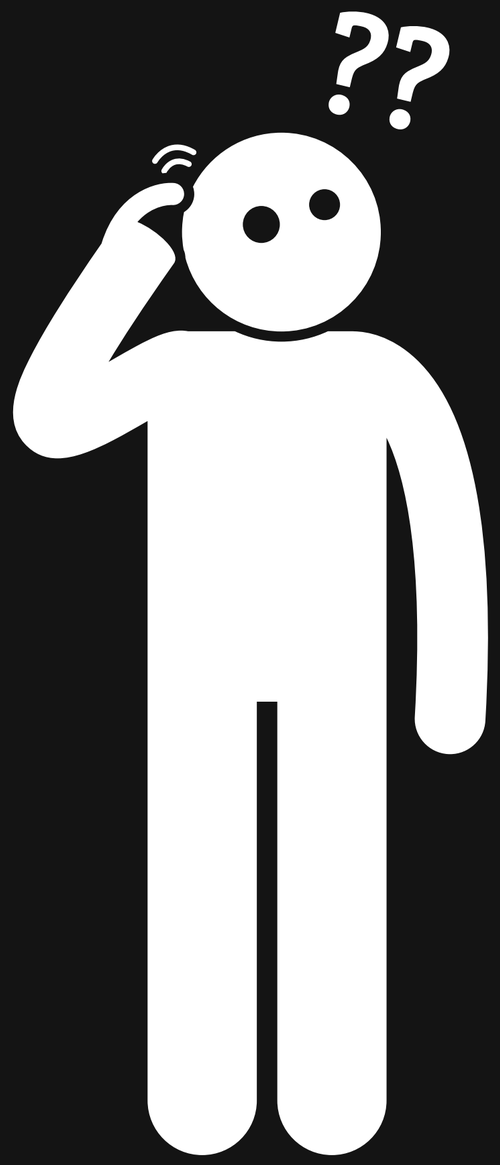
- “If you are new to Gemini, using Google AI Studio is the fastest way to get started”
- However, as your generative AI solutions mature, you may need a platform for building and deploying generative AI applications and solutions end to end.
- Google Cloud provides a comprehensive ecosystem of tools to enable developers to harness the power of generative AI, from the initial stages of app deployment to managing complex data at scale. Google Cloud's Vertex AI platform offers a suite of MLOps tools that streamline the usage, deployment, and monitoring of AI models for efficiency and reliability



“The Vertex AI Gemini API and Google AI Gemini API both let you incorporate the capabilities of Gemini models into your applications. The platform that's right for you depends on your goals”

- The Vertex AI Gemini API is designed for developers and enterprises for use in scaled deployments. It offers features such as enterprise security, data residency, performance, and technical support. If you're an existing Google Cloud customer or deploy medium to large scale applications, you're in the right place.

- If you're a hobbyist, student, or developer who is new to Google Cloud, try the Google AI Gemini API, which is suitable for experimentation, prototyping, and small deployments. If you're looking for a way to use Gemini directly from your mobile and web apps, see the Google AI SDKs for Android, Swift, and web.



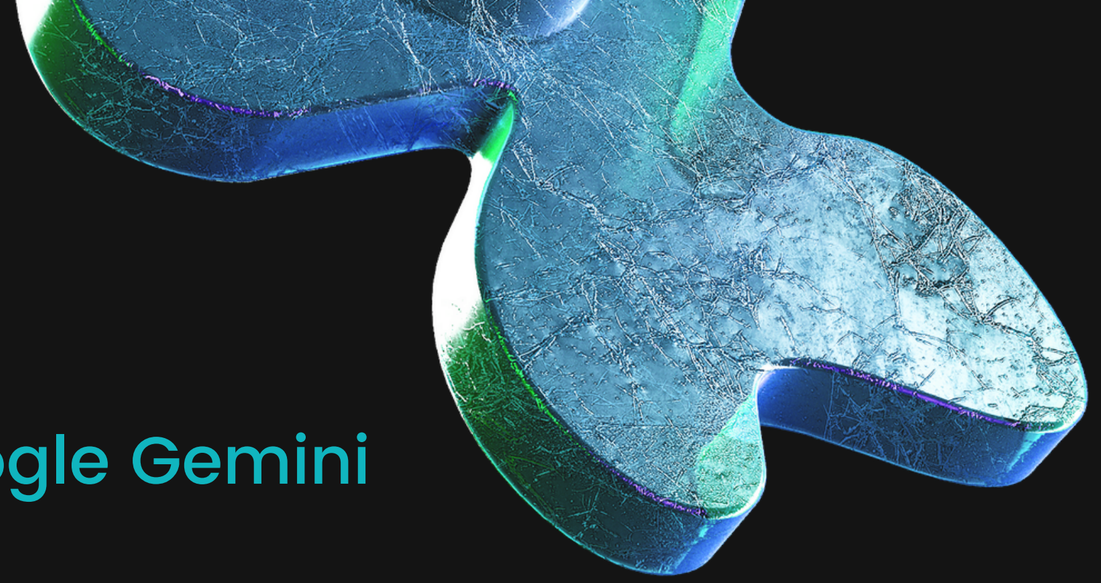
Features	Google AI Gemini API	Google Cloud Vertex AI Gemini API
Latest Gemini models	Gemini Pro and Gemini Ultra	Gemini Pro and Gemini Ultra
Sign up	Google account	Google Cloud account (with terms agreement and billing)
Authentication	API key	Google Cloud service account
User interface playground	Google AI Studio	Vertex AI Studio
API & SDK	Python, Node.js, Android (Kotlin/Java), Swift, Go	SDK supports Python, Node.js, Java, Go
Free tier	Yes	$300 Google Cloud credit for new users
Quota (Request per minute)	60 (can request increase)	Increase upon request (default: 60)
Enterprise support	No	Data privacy commitments Customer encryption key Virtual private cloud Data residency Access transparency
MLOps	No	Full MLOps on Vertex AI (Examples: model evaluation, Model Monitoring, Model Registry)

Google Gemini

In Your Code!

Work with the APIs directly in your applications

```python
import google.generativeai as genai

genai.configure(api_key="YOUR_API_KEY")

generation_config = {
  "temperature": 0.9,
  "top_p": 1,
  "top_k": 1,
  "max_output_tokens": 2048,
}

safety_settings = [
  {
    "category": "HARM_CATEGORY_HARASSMENT",
    "threshold": "BLOCK_MEDIUM_AND_ABOVE"
  },
  {
    "category": "HARM_CATEGORY_HATE_SPEECH",
    "threshold": "BLOCK_MEDIUM_AND_ABOVE"
  },
  {
    "category": "HARM_CATEGORY_SEXUALLY_EXPLICIT",
    "threshold": "BLOCK_MEDIUM_AND_ABOVE"
  },
  {
    "category": "HARM_CATEGORY_DANGEROUS_CONTENT",
    "threshold": "BLOCK_MEDIUM_AND_ABOVE"
  }
]

model = genai.GenerativeModel(model_name="gemini-pro",
                              generation_config=generation_config,
                              safety_settings=safety_settings)

prompt_parts = [
  "Write me an essay to get me into MIT",
]

response = model.generate_content(prompt_parts)
print(response.text)
```



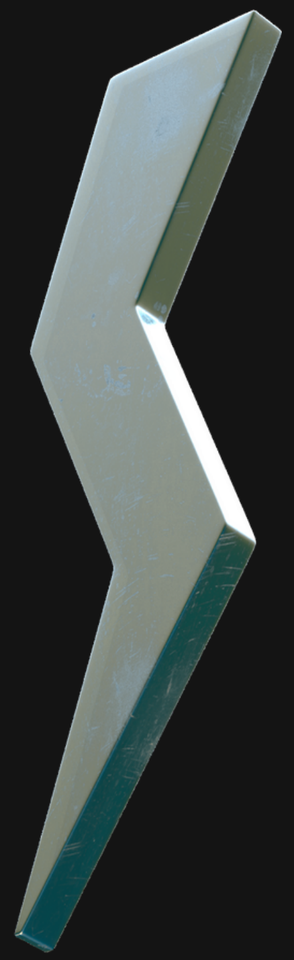
Google Gemini

Bard

The easiest way to work with Gemini

Bard Features

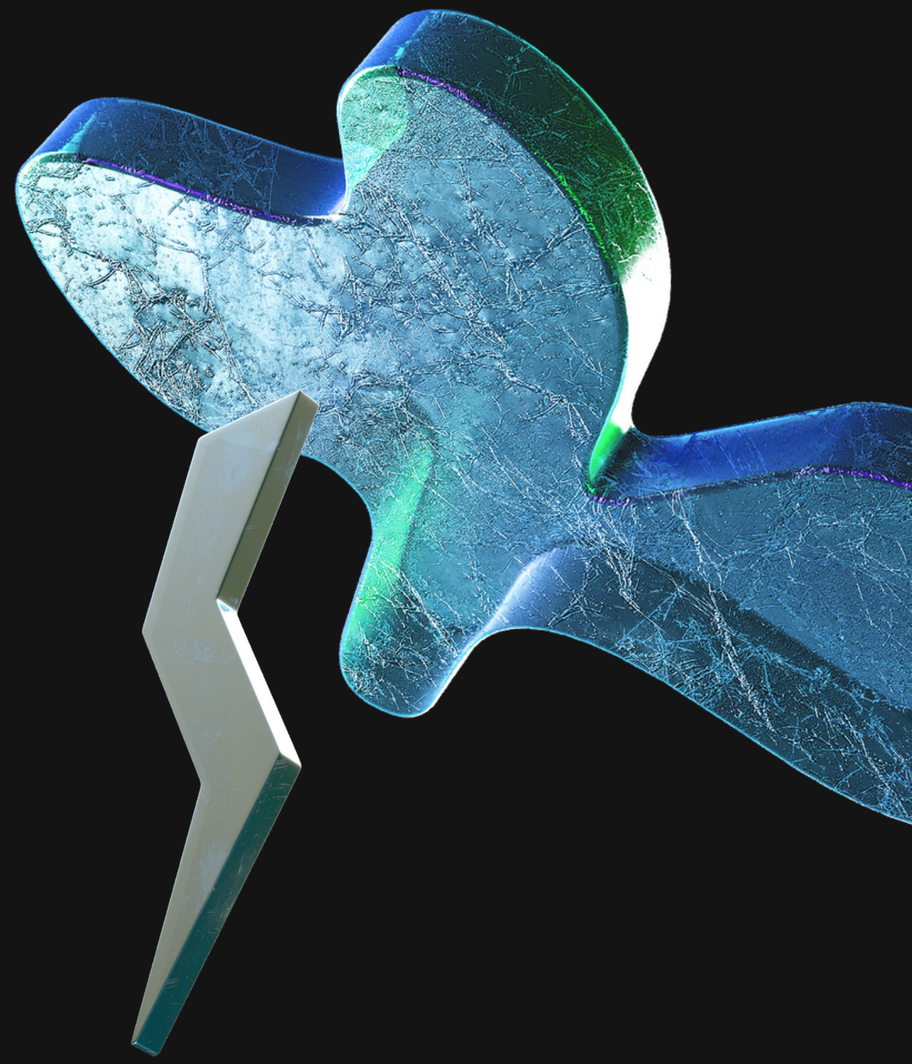
What can it do??



- Location-specific replies
- Analyzing Youtube videos
- Data Visualization
- Summarizing Emails
- Generating tables and spreadsheets
- Connecting to Drive, Docs, and more!

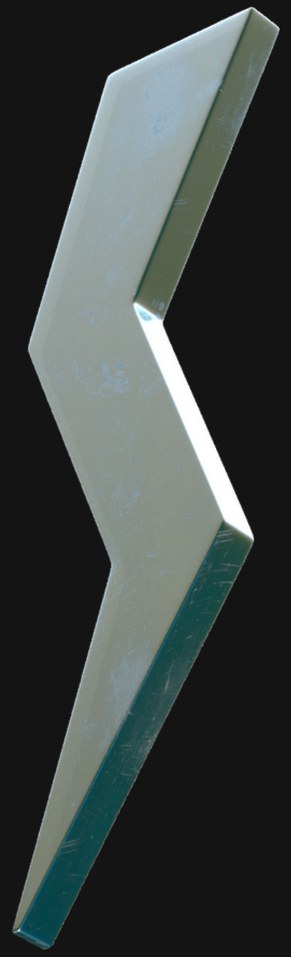
Prompting Strategies

Google Gemini



Basic Prompting Principles

For Any Prompt



- Be clear and specific
- Consider the audience
- Specify an output format
- Use examples when needed
- Use simple language
- Test and iterate for best results

Prompt Elements

The Input

Context

Examples

The Input

A question

What were the most popular baby names in Hawaii in the year 1990?

The Input

A task

Make me an outline for a course teaching avalanche safety

The Input

Completion

The best way to be happy every day in life is to

The Input

Entity

Classify the sentiment in the following tweets as either
Positive, Neutral, or Negative:
“I freaking hate the Dallas Cowboys”
“The Cowboys are the best thing on earth”
“I don’t know anything about the Dallas Cowboys”

Context

Provide
Context/Data
When Needed

```
I am a coding instructor on Udemy. I am trying to create a new
promotional video for my course on TypeScript.  I don't take
myself or my courses too seriously.  I work with an animation
team to make my promotional videos with some sort of theme.

Suggest some potential visual themes for my promo video.
```

Context

Provide
Context/Data
When Needed

```
Last week I received a job offer for a design role with a salary
of $120,000.  This is the offer letter:
<INSERT OFFER LETTER TEXT HERE>

Help me craft a response that indicates that I'm very interested
in the role but need higher compensation.
```

Examples

Generate me a 2 color palette based on the following sentence: "We sat by the fire as the snow fell"

Palette: Cozy Warmth

- **Main color: Ember red (#C25636) – Captures the glowing warmth of the fire and the coziness of sitting close to it.**
- **Accent color: Snowy white (#FDFEFE) – Represents the gentle snowfall and adds a touch of serenity and contrast.**

This palette evokes a feeling of warmth, comfort, and quiet reflection amidst the wintery landscape.

Examples

```
Generate me a color palette with 2 colors based on an input sentence.

Prompt: "We sat by the fire as the snow fell"
Your Reply: #C25636, #FDFEFE
Prompt: "Tropical waves and sand"
Your Reply: #40E0D0, #FDC593

Prompt: "aurora against a dark sky"
Your Reply:
```

#1C2428,#A2FF00

Format

Create a list of the 5 names of the top most populated countries on earth and their populations

Here are the top 5 most populated countries on Earth and their populations as of July 1, 2023:

1. **India: 1,428,627,663**
2. **China: 1,425,671,352**
3. **United States: 339,996,563**
4. **Indonesia: 277,534,122**
5. **Pakistan: 240,485,658**

Format

Create a comma separated list of the 5 names of the top most populated countries on earth and their populations. Format your output using CSV:

```
Rank,Country,Population
1,China,1425671352
2,India,1428627663
3,United States,339996563
4,Indonesia,277534122
5,Pakistan,240485658
```

Provide Cues

```
Generate me a list of the 5 longest ultramarathons in the world
Start your response here:
*
```

Google Gemini

Google AI Studio

An Awesome Way To Play and Build With Gemini

Google AI Studio

Get API key

Create new

My library

No prompts yet

Getting started

Documentation

Prompt gallery

Discord community

AI Studio in Vertex

Settings

Untitled prompt

Save

Get code

Insert: Image {{ }} Test input

Run settings

Reset

Identify the city where this photo was taken and give a brief description of the city

Tab

Model

Temperature

0.25

Add stop sequence

Add stop...

Max outputs

1

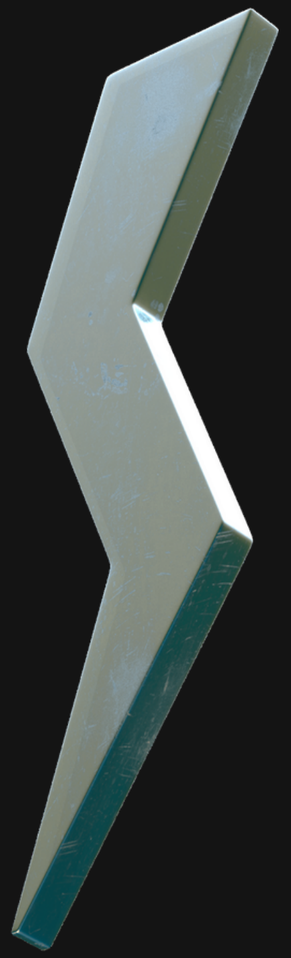
Run

Text preview 0 / 0

Google AI Studio

What is it???

"The fastest way to start building AI"



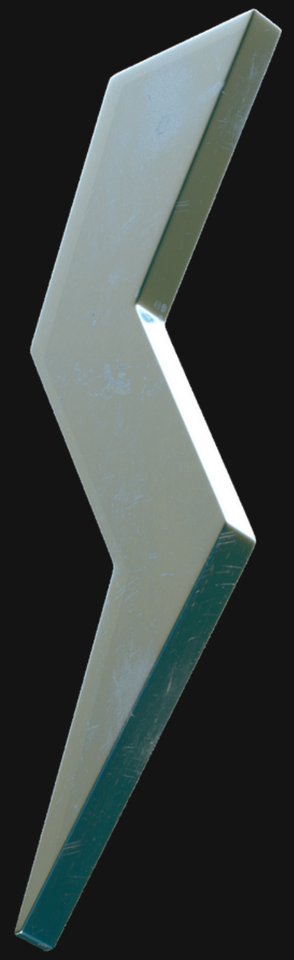
- Browser Based IDE
- Made for Prototyping
- Try Out Models and Experiment
- Supports Gemini Pro & Vision Pro
- You can Export to Code
- Completely Free!

Let's SIGN UP NOW

Gemini Pro

Google AI Studio

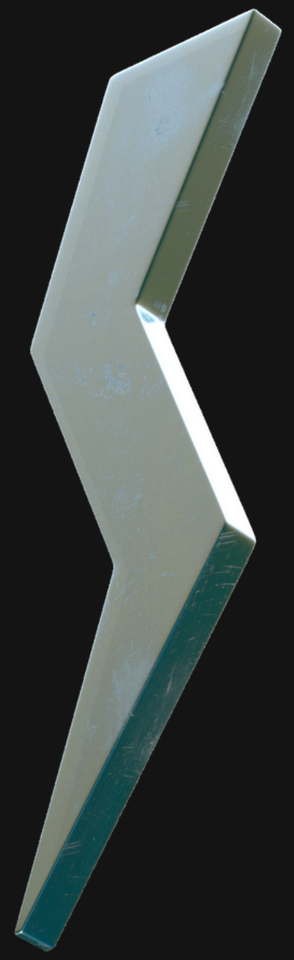
- Input: text
- Output: text
- Input token limit: 30720
- Output token limit: 2048
- Rate Limit: 60 requests per minute



Gemini Pro Vision

Vertex AI API

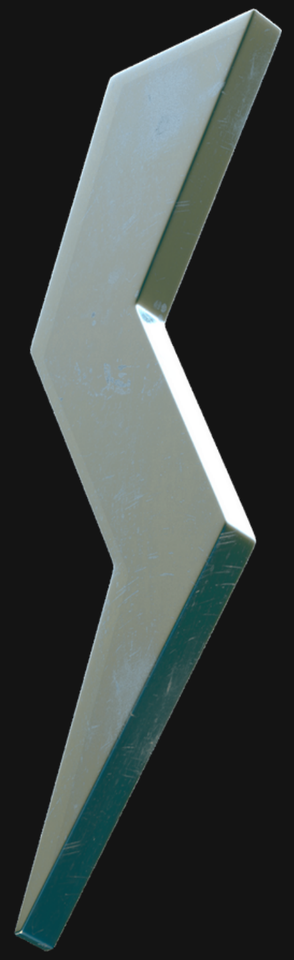
- Input: text, image, or video
- Output: text
- Max tokens (input and output): 16,384
- Output token limit: 2,048
- Training data: Up to Feb 2023



Gemini Pro

Vertex AI API

- Input: text
- Output: text
- Max tokens (input and output): 32,760
- Output token limit: 8,192
- Training data: Up to Feb 2023



Google Gemini

Test Inputs

Convert red to a hex code

Test your prompt

	INPUT input	OUTPUT
1	red	#FF0000
2	blue	#0000FF
3	sunny sky	#1F4A93
4	anxiety	#F9F9F9

Google Gemini

Structured Prompts

You are a product marketer targeting a Gen Z audience. Create exciting and fresh advertising copy for products and their simple description. Keep copy under a few sentences long.

2 / 500 examples

Actions

INPUT Product:	OUTPUT Product Copy:
Old-school sneaker	Let's lace up! These kicks bring an iconic look and a one of a kind color palette, while supporting you in style and function like no other shoe before.
Supersoft hoodie	Stay cozy and stylish in our new unisex hoodie! Made from 100% cotton, this hoodie is soft and comfortable to wear all day long. The semi-brushed inside will keep you warm on even the coldest days.
The user's input	The model's response

Test your prompt

	INPUT Product:	OUTPUT Product Copy:
1	Fluffy Mittens	Introducing the world's fluffiest mittens! Made with ultra-soft yarn, these mittens will keep your hands warm and cozy even on the coldest winter days. The wristbands are stretchy and comfortable, making them perfect for all-day wear.

Add test example

Google Gemini

Chat Prompts

Write your prompt examples

User

You are a helpful chatbot that helps people learn Spanish. You are humorous and an understanding, patient teacher.

Model

Ok! Let's get started learning Spanish!

User

The user's input

Model

The model's response

Test your prompt

Testing with 1 example

User

Hey there!

Model

¡Hola! ¿Qué tal? (Hello! How are you?)

User

How do I say "tired" in spanish?

Model

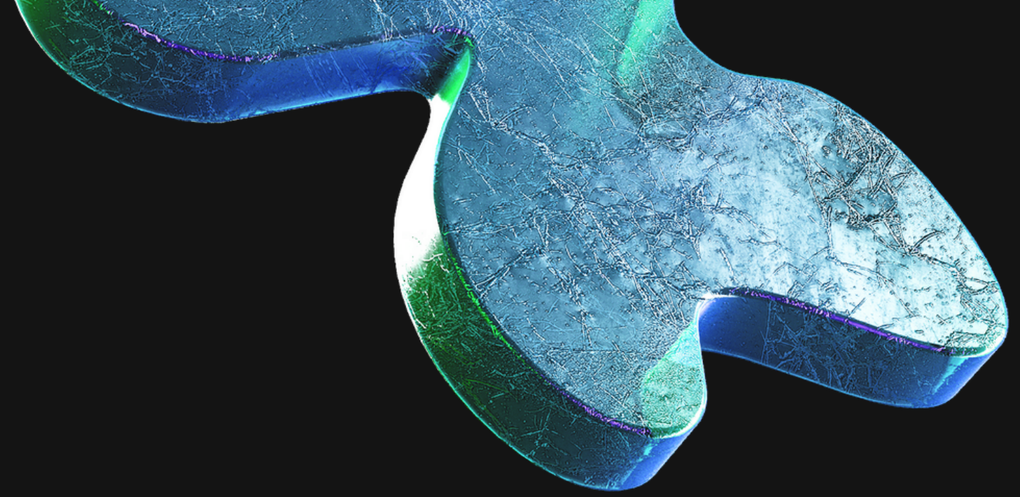
To say "tired" in Spanish, you can say "cansado" (masculine) or "cansada" (feminine). For example:

* Estoy cansado. (I am tired.)
* Ella está cansada. (She is tired.)

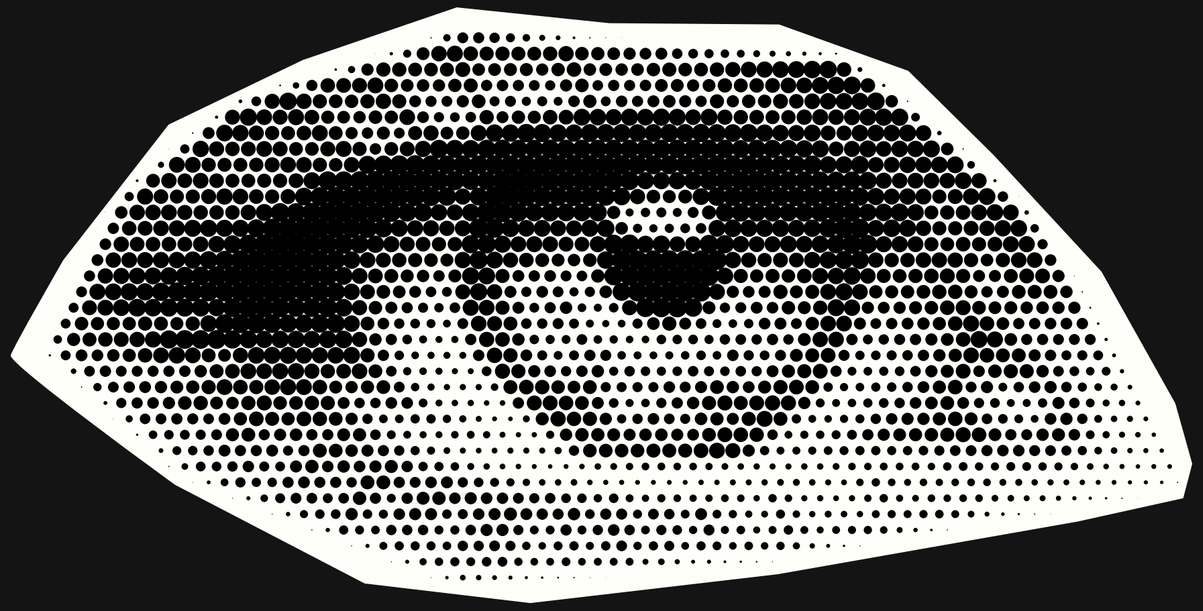
You can also use the verb "tener sueño" (to be sleepy) to express tiredness. For example:

* Tengo sueño. (I am sleepy.)
* Él tiene sueño. (He is sleepy.)

Write something the user could type



Gemini Pro Vision

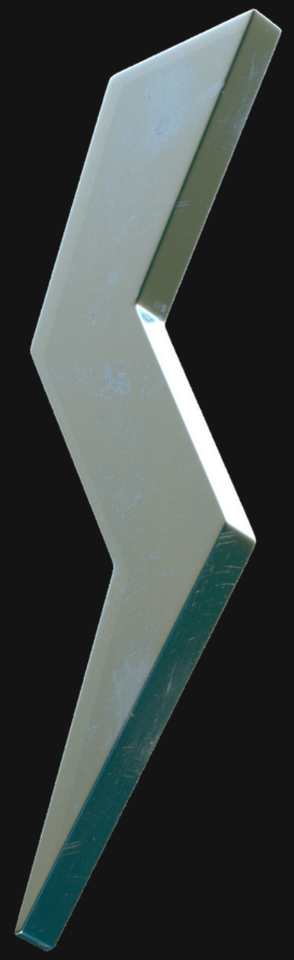


A model that understands input from text and visual modalities (image and video)

Gemini Pro Vision

Google AI Studio

- Input: text and/or image
- Output: text
- Input token limit: 12288
- Output token limit: 4096
- Rate Limit: 60 requests per minute



Gemini Vision

Image Requirements

Maximum of 16 images per prompt

Maximum of 4MB for the entire prompt (including images and text)

No limit to the number of pixels in an image. However, larger images are scaled down and padded to fit a maximum resolution of 3072 x 3072

Each image accounts for 258 tokens

Supports JPEG, PNG, WEBP, HEIC, HEIF images

Prompts with a single image tend to yield better results.

Gemini Vision

Image Best Practices

Prompts with a single image tend to yield better results.

If your prompt contains a single image, placing the image before the text prompt might produce better results.

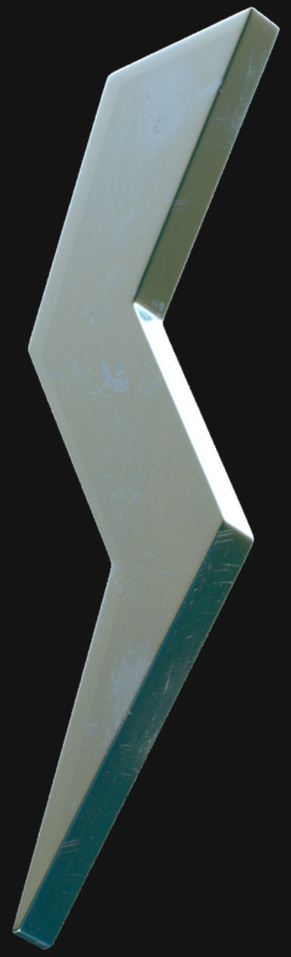
Images with higher resolution yield better results.

Rotate images to their proper orientation before adding them to the prompt.

If working with multiple images, label the images with text so you can refer to them within your prompt

Image Prompting Tips

For Bard (and Gemini)



- Be specific in your instructions
- Break it down step by step
- Call attention to parts of the image
- Put the image first (Gemini API only)
- Ask the model to describe the image first before doing anything else
- Ask the model to “explain why”

LLMS Under The Hood

Are model responses random or deterministic?

BOTH!

- When we prompt an LLM, its text response is generated in two stages.
  - In the first stage, the LLM processes the input prompt and generates a probability distribution over possible tokens that are likely to come next.

For example, the prompt **"My favorite color is ... "**, might result in this set of probable next words:

```
[("blue", 0.42), ("green", 0.25), ("purple", 0.09), ...]
```

This process is deterministic. A model will give us the same distribution every time we give it the same input prompt

- In the second stage, the model converts these distributions into actual text replies through one of many decoding strategies.
  - A simple decoding strategy might select the most likely token at every step. This process would always be deterministic.

For example, the prompt **"My favorite color is ... "**, might result in this set of probable next words:

```
[("blue", 0.42), ("green", 0.25), ("purple", 0.09), ...]
```

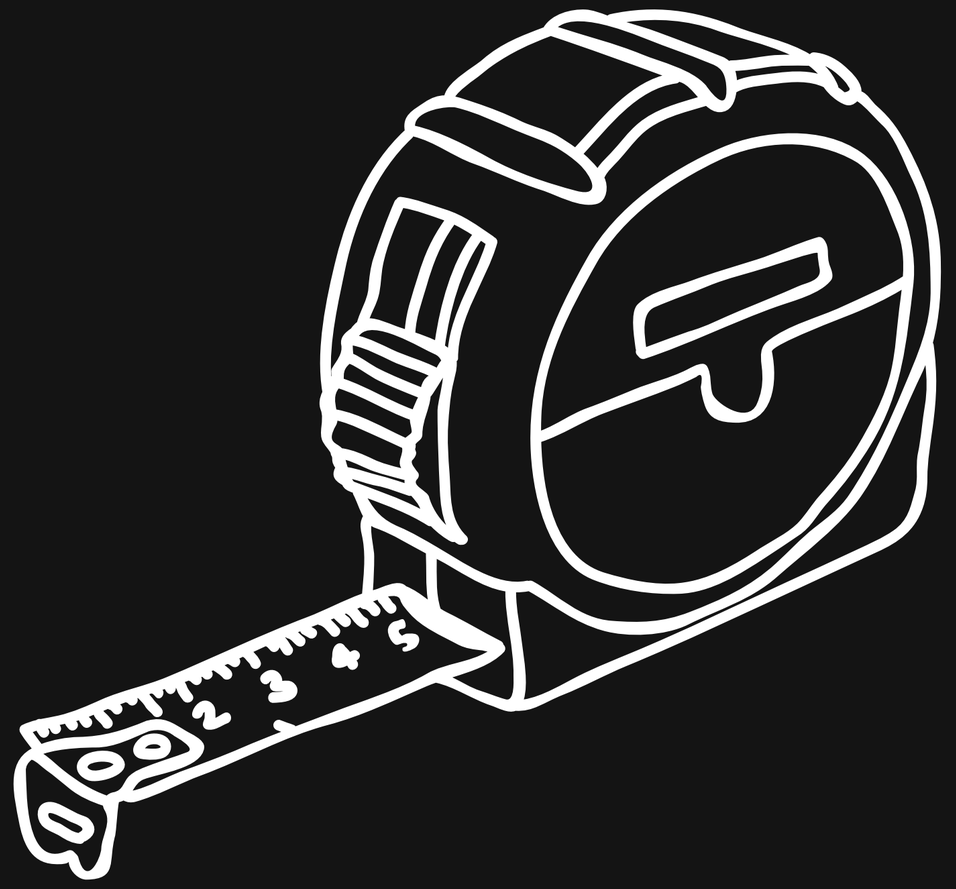
- Instead, LLM models choose to generate a response by randomly sampling over the distribution returned by the model.
- With models like Gemini, we can control the degree of randomness allowed by setting the temperature (and other parameters)
- A temperature of 0 means only the most likely tokens are selected, and there's no (less) randomness.
- A high temperature injects a high degree of randomness into the tokens selected by the model, leading to more unexpected, surprising model responses.

```
[("blue", 0.42), ("green", 0.25), ("purple", 0.09), ...]
```

Max Output Length

Controls the size of the generated output

- Specifies the maximum number of tokens that can be generated in the response.
- A token is approximately four characters.
- 100 tokens correspond to roughly 60-80 words.
- Specify a lower value for shorter responses and a higher value for potentially longer responses.
- Default for gemini-pro: 8192
- Default for gemini-pro-vision: 2048



Stop Sequences

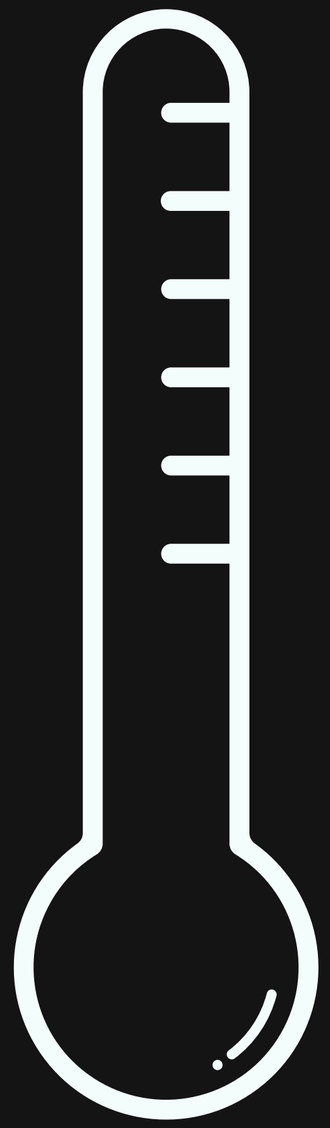
“Stop if you generate one of these sequences”

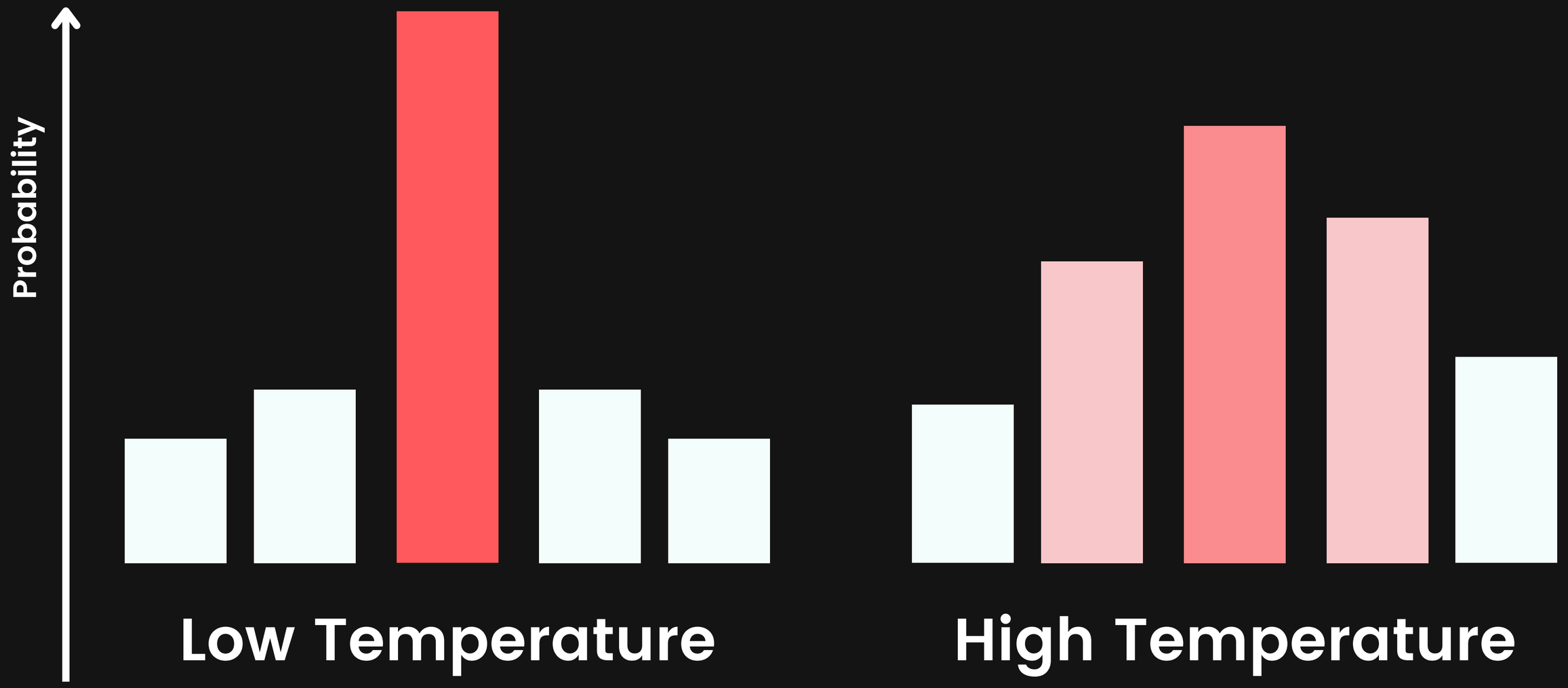
- Specifies a list of strings that tells the model to stop generating text if one of the strings is encountered in the response.
- If a string appears multiple times in the response, then the response truncates where it's first encountered.
- The strings are case-sensitive
- Maximum of 5 sequences allowed

Temperature

Controls the “randomness” in token selection

- Ranges from 0 - 1
- Lower temperatures lead to more deterministic and less open-ended or creative responses
- Higher temperatures can lead to more diverse or creative results.
- According to Google: “For most use cases, try starting with a temperature of 0.2” If the model returns a response that's too generic, too short, or the model gives a fallback response, try increasing the temperature.
- Default for gemini-pro: 0.9
- Default for gemini-pro-vision: 0.4




Probability
Low Temperature
High Temperature

- To generate more conservative and predictable outputs, such as in factual reporting or customer service responses.
- To ensure that the model's outputs are aligned with specific guidelines or requirements.

- To generate more creative and surprising outputs, such as in creative writing or brainstorming sessions.
- To encourage the model to explore more diverse options, especially when dealing with limited or repetitive input.

lower temperature → **higher temperature**

When generating legal documents or financial reports, to ensure that the output is accurate, consistent, and free of errors.

When writing code, to reduce the likelihood of introducing bugs or security vulnerabilities.

When creating training data for a machine learning model, to ensure that the data is consistent and representative of the real world

When generating song lyrics or poetry, to encourage the model to come up with more creative and surprising ideas.

When writing marketing copy, to generate more engaging and attention-grabbing headlines or slogans.

When creating dialogue for a fictional character, to make their speech more unique and memorable.

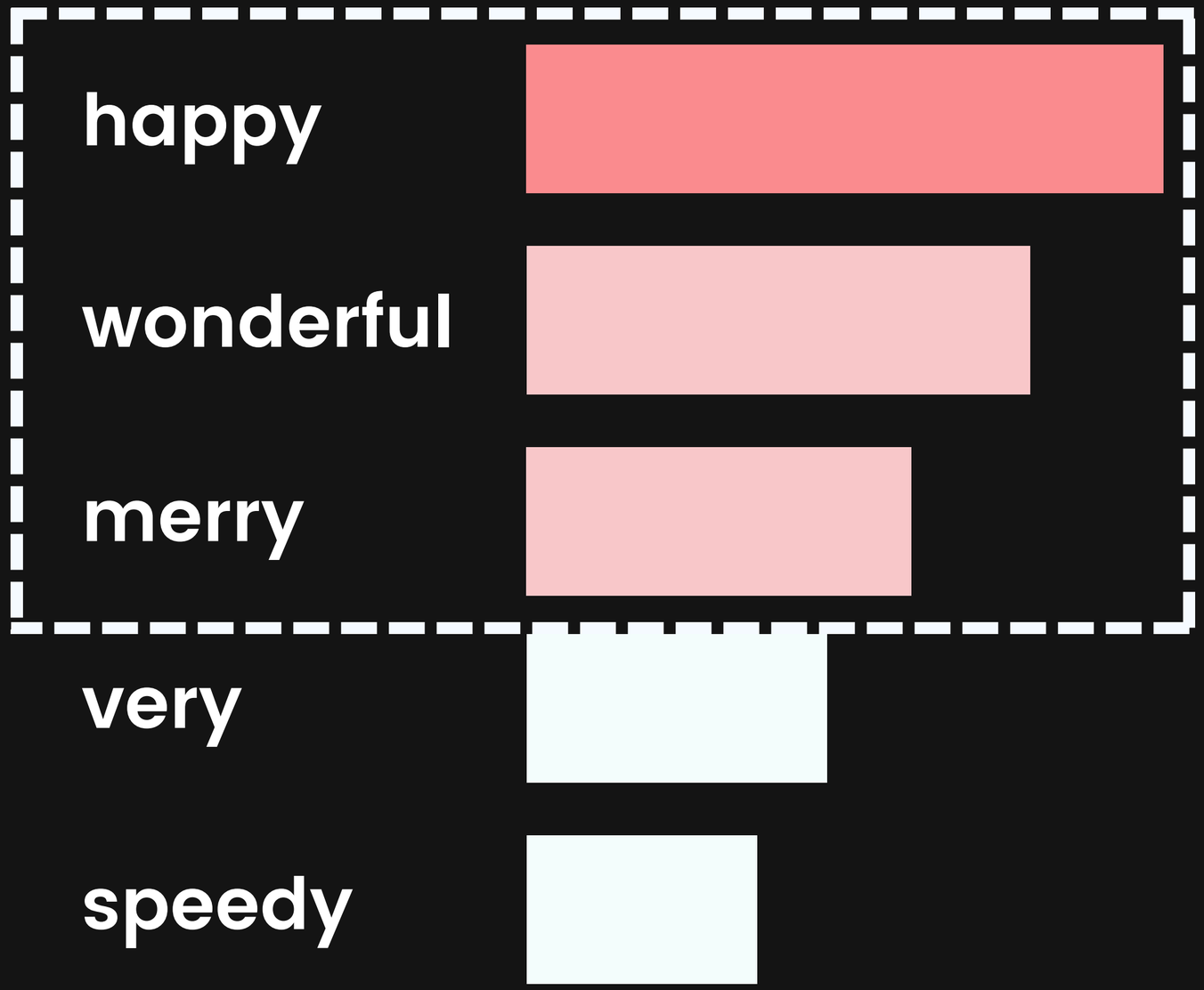
lower temperature → **higher temperature**

Top K

Changes how the model selects tokens for output.

- A topK of 1 means that the model selects the token that is the most probable among all the tokens in the model's vocabulary
- A topK of 3 means that the next token is selected from among the 3 most probable using the temperature.
- For each token selection step, the topK tokens with the highest probabilities are sampled.
- Tokens are then further filtered based on topP with the final token selected using temperature sampling.
- Range: 1-40
- Default for gemini-pro: 1
- Default for gemini-pro-vision: 32

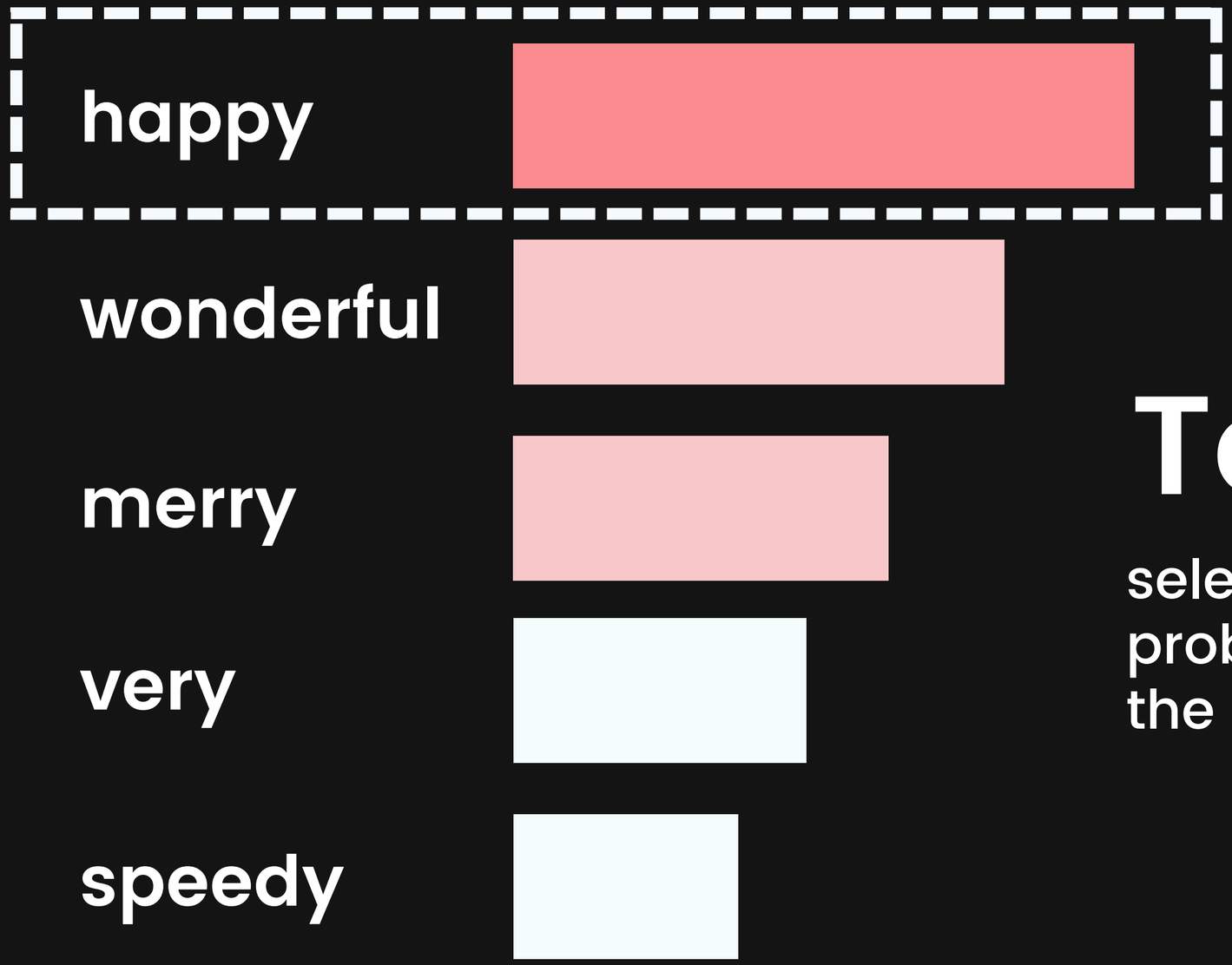
Wishing you a...



Top k = 3

next token is selected from among the 3 most probable

Wishing you a...



Top k = 1

selects the token that is the most probable among all the tokens in the model's vocabulary

- To focus the model's output on a more specific set of options, reducing the likelihood of unexpected or irrelevant results.
- To ensure that the model's responses are concise and relevant to the input, especially in situations where brevity is important.

- To promote greater diversity in the model's output, especially when dealing with large or complex datasets.
- To encourage the model to explore less common words or phrases, leading to more varied and interesting responses

lower top k → **higher top k**

When generating text that needs to be very concise or specific, to focus the model's output on the most likely options.

When working with a dataset that contains a lot of noisy or irrelevant data, to reduce the likelihood of the model generating text that is off-topic or nonsensical.

When generating text that needs to be consistent with a specific set of rules or constraints, to reduce the likelihood of the model generating text that violates those rules or constraints.

When generating text in a specific style or genre, to encourage the model to explore a wider range of options within that style

When working with a dataset that contains a lot of rare or uncommon words or phrases, to ensure that the model is able to generate text that is both diverse and relevant.

When generating text for a specific audience, to ensure that the model is able to generate text that is appropriate for that audience

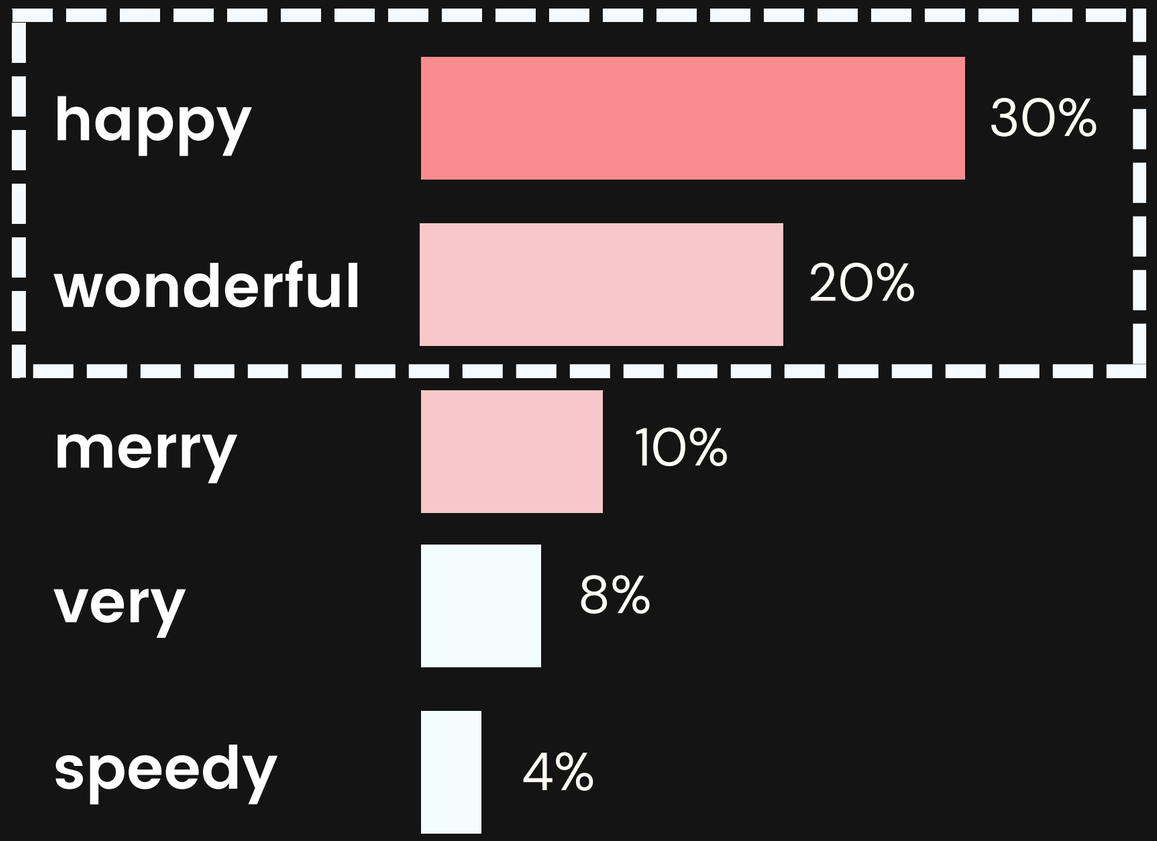
lower top k → **higher top k**

Top P

Also changes how the model selects tokens for output.

- Picks from the top tokens based on the **sum of their probabilities**
- Tokens are selected from the most to least probable until the sum of their probabilities equals the topP value.
- For example, if tokens A, B, and C have a probability of 0.3, 0.2, and 0.1 and the topP value is 0.5, then the model will select either A or B as the next token by using the temperature and exclude C as a candidate.
- The default topP value is 0.95 (docs are inconsistent here)
- Range: 0.0 - 1.0

Wishing you a...



Top p = 0.5

Tokens are selected from the most to least probable until the sum of their probabilities equals the topP value.

- To focus the model's output on a narrower range of more likely or relevant options.
- To ensure that the model's responses are consistent with a specific style or tone, reducing the likelihood of unexpected or off-brand outputs

- To allow the model to generate more diverse and surprising outputs, while still maintaining some level of coherence and relevance to the input.
- To encourage the model to explore a wider range of possibilities, especially when dealing with open-ended or ambiguous prompts.

lower top p → **higher top p**

When generating text that needs to be very accurate or factual, to reduce the likelihood of the model generating text that is incorrect or misleading.

When working with a dataset that contains a lot of sensitive or confidential information, to reduce the likelihood of the model generating text that could be harmful or damaging.

When generating text that needs to be very creative or surprising, to encourage the model to explore a wider range of possibilities.

When working with a dataset that contains a lot of ambiguous or incomplete information, to allow the model to generate text that is both plausible and consistent with the available information.

When generating text for a specific purpose, to allow the model to generate text that is tailored to that purpose.

lower top p → **higher top p**

	Temperature	Top-K	Top-P
How it works	Controls the randomness of the model's output	Limits the number of tokens considered during generation	Specifies a cumulative probability threshold
Effect on output	More varied and surprising outputs at higher temperatures, more conservative and predictable outputs at lower temperatures	Encourages the model to explore less common words and phrases	Allows for more fine-grained control over the diversity of outputs
When to use	When you want to generate more creative and surprising outputs, or when you want to encourage the model to explore diverse options	When you want to generate a diverse set of outputs without sacrificing coherence	When you want to generate very creative or surprising outputs, or when you need to ensure that the outputs are consistent with a specific style or tone

Gemini Vision

Video Requirements

Only information in the first 2 minutes is processed.

Each video accounts for 1,032 tokens

No audio information or timestamp metadata is analyzed

Google recommends no more than one video per prompt.

Supports MOV, MPEG, MP4, MPG, AVI, WMV, MPEGPS, FLS

Let’s Get

CODING