UNLEASHING CREATIVITY WITH VERTEX AI:
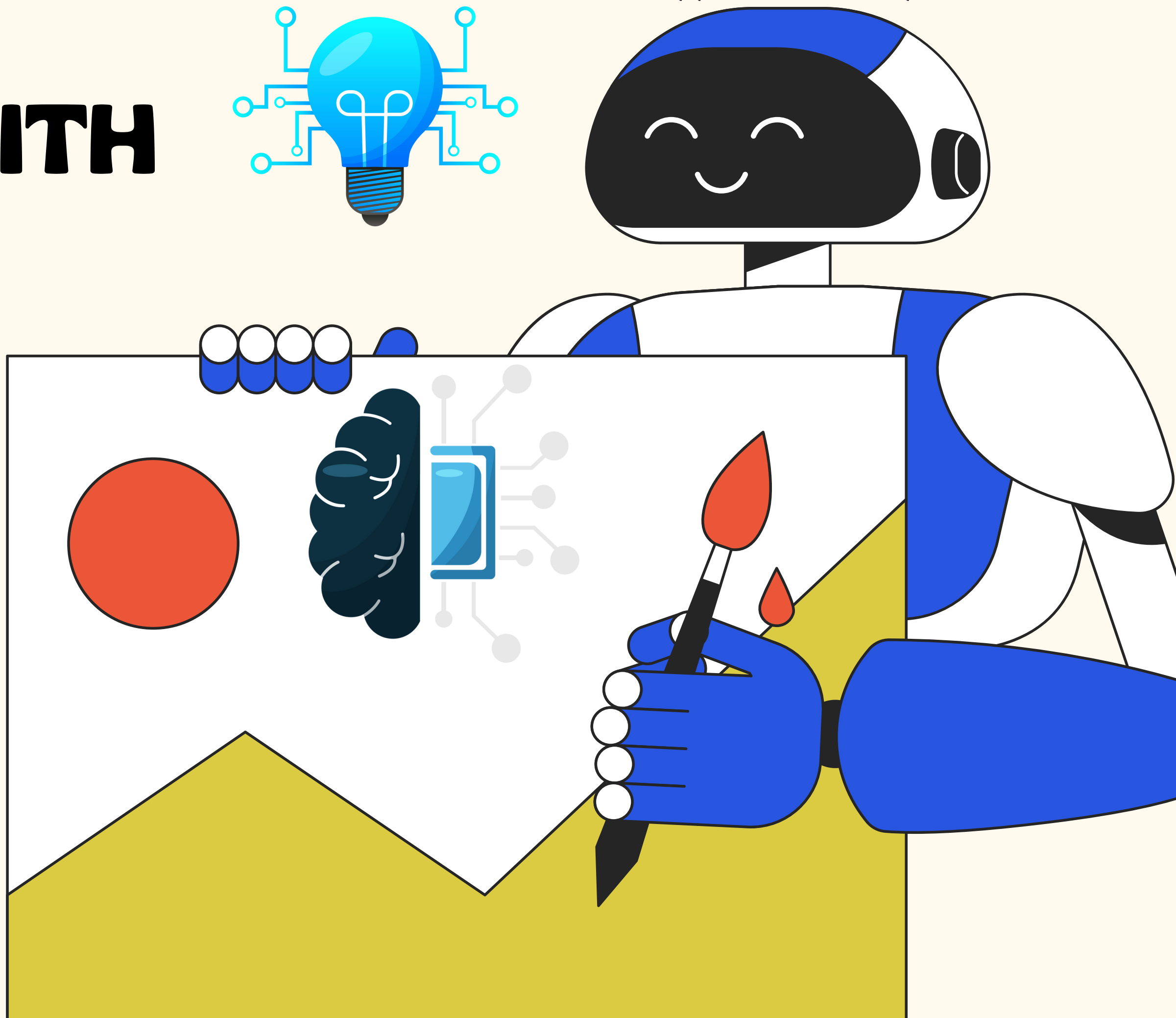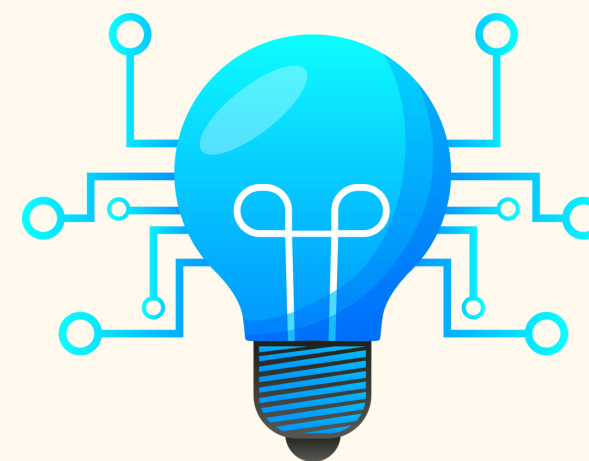
HOW TO UTILIZE GENERATIVE AI FOR YOUR CUTTING-EDGE APPLICATIONS

CHALLENGE SOLVED : IN THIS DISCUSSION

- OVERVIEW OF THE VERTEXAI PLATFORM : FULL FEATURE ML PLATFORM

  - MODEL GARDEN

  - WORK BENCH

  - PIPELINES

  - GENERATIVE AI STUDIO

- UNDERSTAND THE MODELS AND THEIR PRICING INSIDE VERTEXAI PLATFORM

- INTRODUCE THE MODELS AND THEIR CAPABILITIES

- UNDERSTAND HOW THE MODEL PROMPTING WORKS, AND THE DEFINITION OF PARAMETERS

- TRYING OUT THE 5 TASKS USING THE TEXT-BISON-001 MODEL

PEEK @ PRICING

Model	Type	Region	Function	Resource charge	Price	Pricing unit
text-bison-001	Input	Global	• Online requests • Batch requests	0%	$0.0010	1000 characters
	Output	Global	• Online requests • Batch requests	0%	$0.0010	1000 characters
chat-bison-001	Input	Global	• Online requests • Batch requests	0%	$0.0005	1000 characters
	Output	Global	• Online requests • Batch requests	0%	$0.0005	1000 characters
textembedding-gecko-001	Input	Global	• Online requests • Batch requests	0%	$0.0001	1000 characters

MODEL & CAPABILITIES

Model name	Description	Model properties
text-bison-001 ↗	Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: • Classification • Sentiment Analysis • Entity extraction • Extractive Question Answering • Summarization • Re-writing text in a different style • Ad copy generation • Concept ideation	Max input token: 8,192 Max output tokens: 1,024 Training data: Up to Feb 2023
textembedding-gecko-001 ↗ (Model tuning not supported)	Returns model embeddings for text inputs.	3,072 input tokens and outputs 768-dimensional vector embeddings.
chat-bison-001 ↗ (model tuning not supported)	Fine-tuned for multi-turn conversation use cases.	Max input token: 4,096 Max output tokens: 1,024 Training data: Up to Feb 2023 Max turns : 2,500

IMPORTANT DEFINITIONS

prompt	Text input to generate model response. Prompts can include preamble, questions, suggestions, instructions, or examples.
temperature	The temperature is used for sampling during the response generation, which occurs when topP and topK are applied. For most use cases, try starting with a temperature of 0.2.
maxOutputTokens	Maximum number of tokens that can be generated in the response.A token may be smaller than a word. A token is approximately four characters. 100 tokens correspond to roughly 60-80 words.
topK	Top-k changes how the model selects tokens for output. A top-k of 1 means the selected token is the most probable among all tokens in the model's vocabulary. If 3, then 3 probable tokens are selected
topP	Top-p changes how the model selects tokens for output. Tokens are selected from most K (see topK parameter) probable to least until the sum of their probabilities equals the top-p value. The default top-p value is 0.95.Specify a lower value for less random responses and a higher value for more random responses.

USECASE : TASKS

Summarization	Support Rep chat Summary, News Article Summary, Chat Agent Summary, Chat Agent follow up, Transcript summarization, Dialog summarization, Hashtag Tokenisation, Title generation
Classification	sentiment analysis on new article, Customer request classification, Text Classification Few Shots, Article Classification, Headline classification, Object Classification, Sentiment Analysis
Extraction	Technical Specification Writing, Troubleshooting, Contract analysis, extractive QA
Writing	Marketing Writeup, Ad Copy Generation, Essay Outline, Correct Grammar
ideation	reading comprehension, Meme Generation, Interview Questions, Naming, General tips & advice

JOIN ME : LETS EXPLORE FREELY

Language **PREVIEW**

GET STARTED   MY PROMPTS   TUNING

Get started

Design and test your own prompts

Design prompts for tasks relevant to your business use case including code generation

+ CREATE PROMPT

Start a conversation

Have a freeform chat with the model, which tracks what was previously said and responds based on context
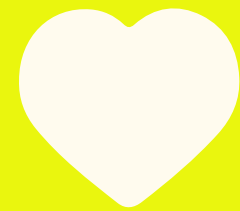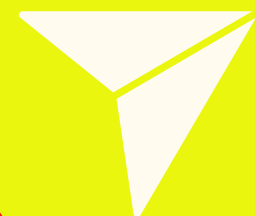
+ CREATE CHAT PROMPT

Tune a model

Tune a model so it's better equipped for your use case, then deploy to an endpoint to get predictions or test it in prompt design
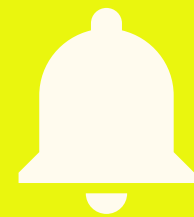
+ CREATE TUNED MODEL

THANKS FOR WATCHING

LIKE

SHARE

SUBSCRIBE