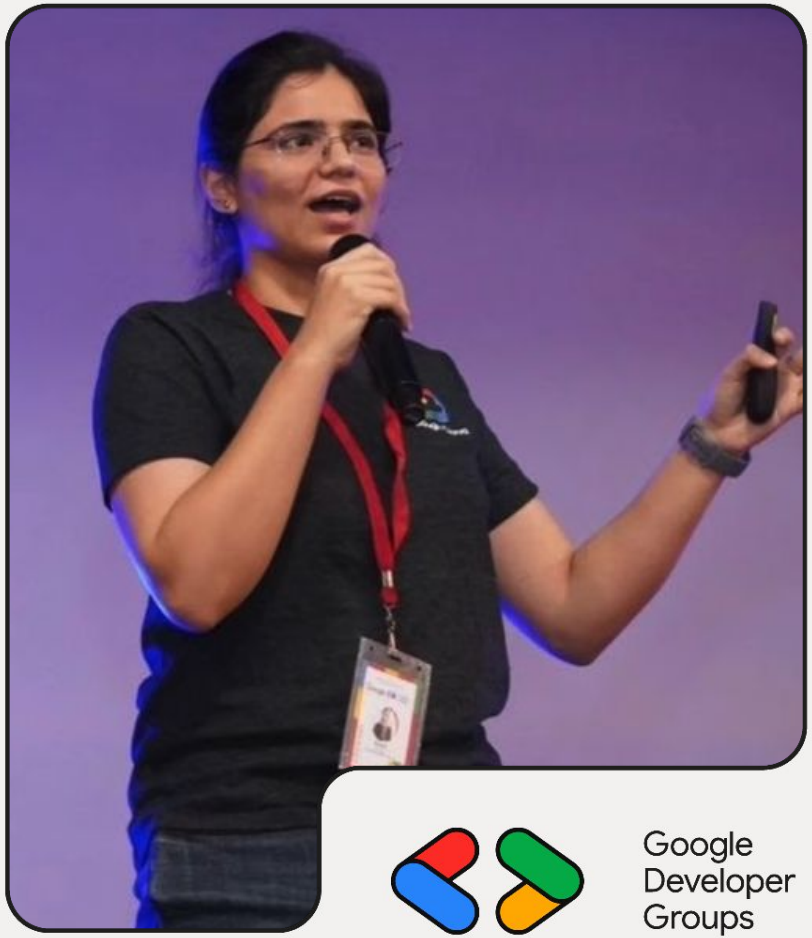
DevFest

Baroda

Unveiling the Mysteries of Large Language Models

Aashi Dutt



Google Developer Groups

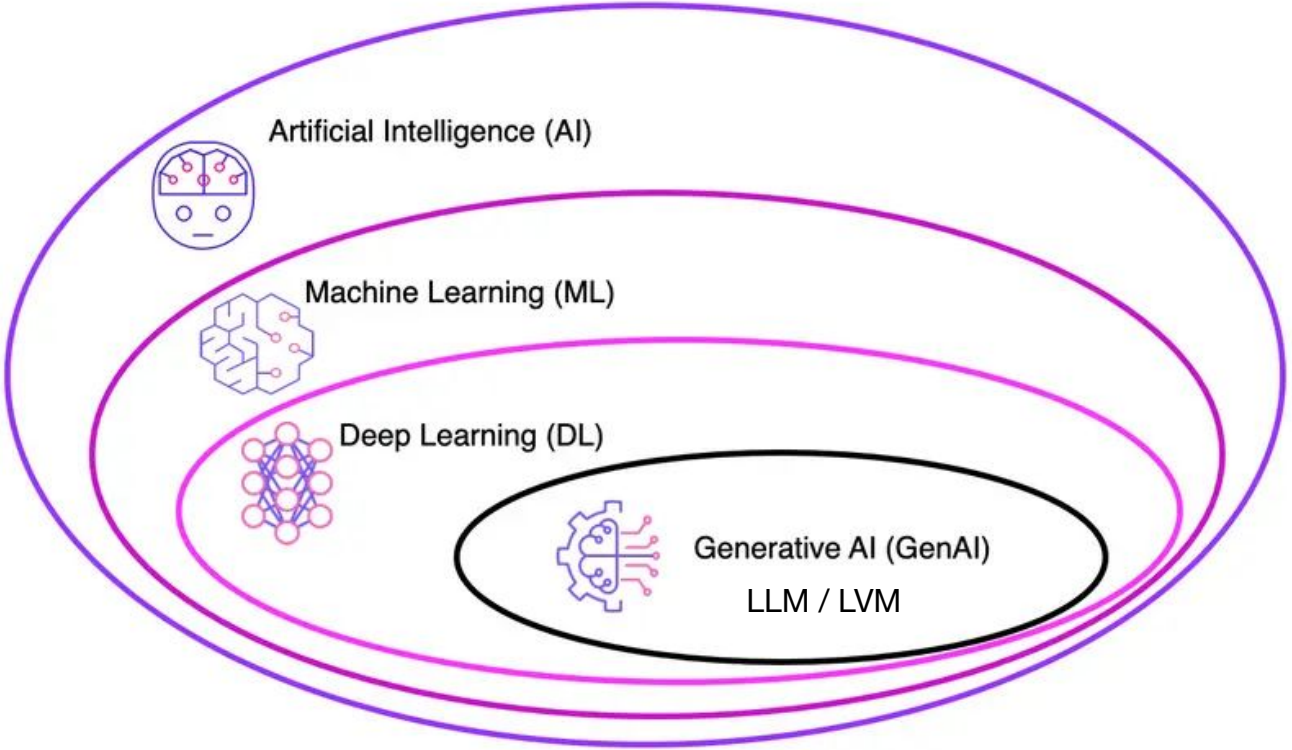
Agenda

1. Generative AI
2. LLM’s - A 10,000 ft view
3. Evolution of LLM’s
4. Transformers → Applications
5. Key Takeaways

AI
@DevFest

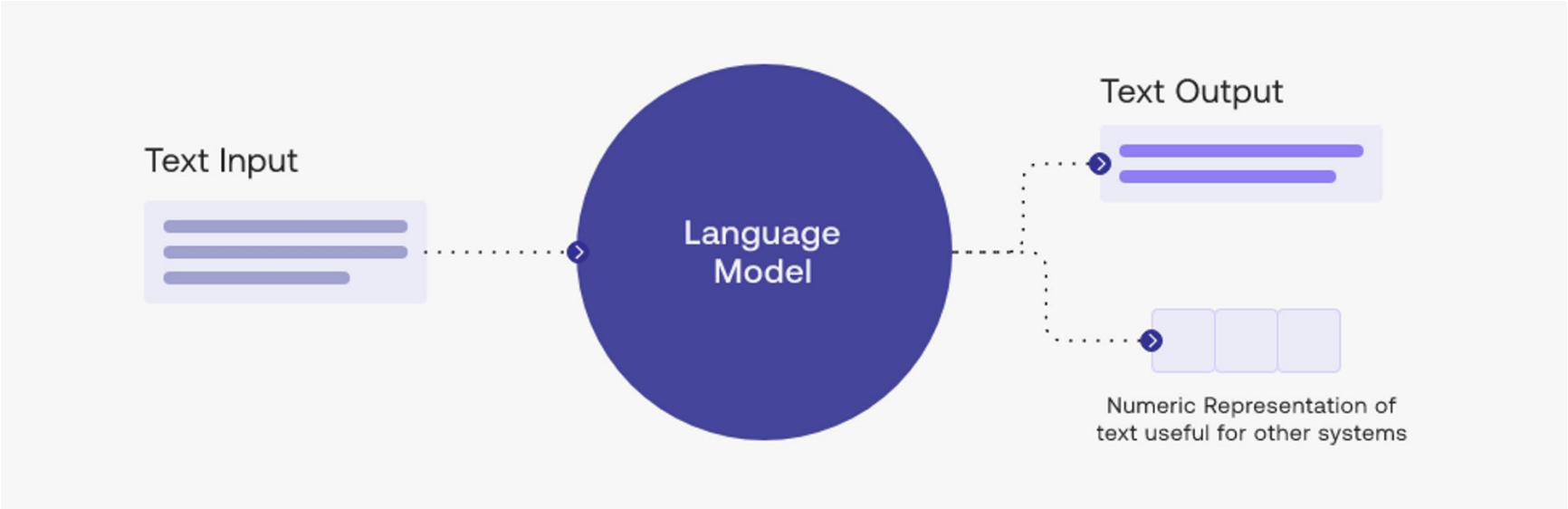
AI is Everywhere

Generative AI




LLM’s - A 10,000 ft View
AI@DevFest

Introduction to Language Models



Classic Natural Language Problems

 Entity extraction

 Classification

 Summarization

 Sentiment Analysis

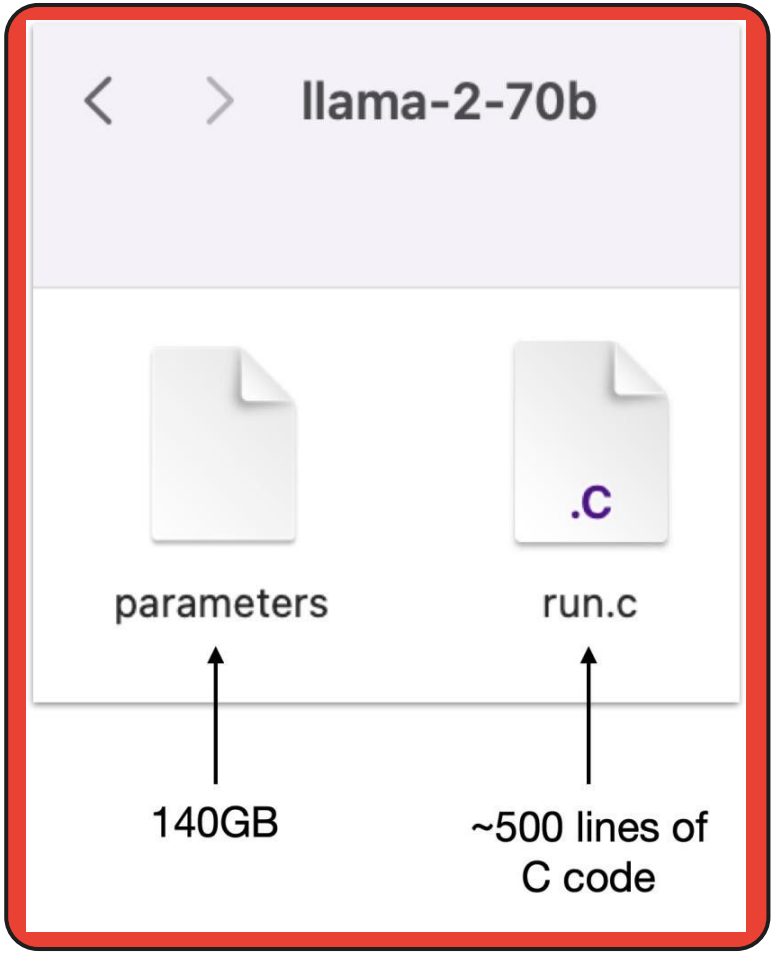
 Translation

 ...

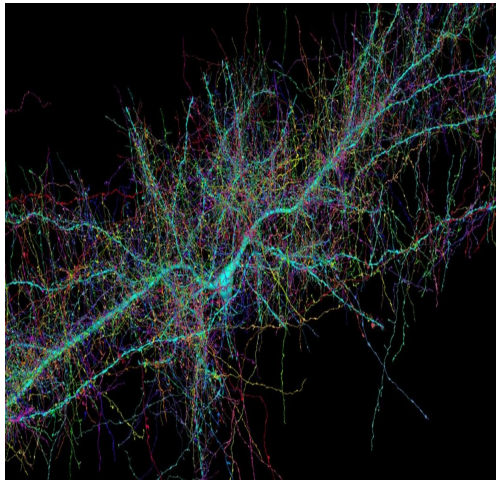
LLM in Simple Terms

Two Files

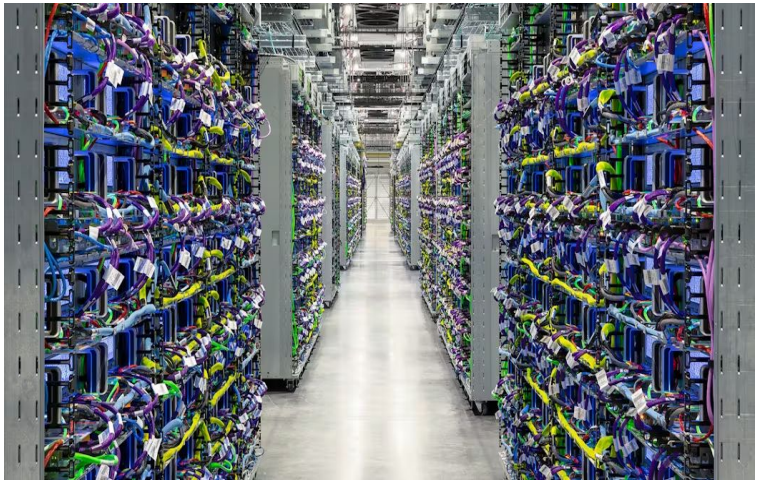
- **Parameters** - Model Weights
- **Execution Code** - run.c
- Example: LLAMA-2 70B parameter model by Meta AI Research takes 140GB space (2 Bytes, FP16 precision)



LLM Model Training



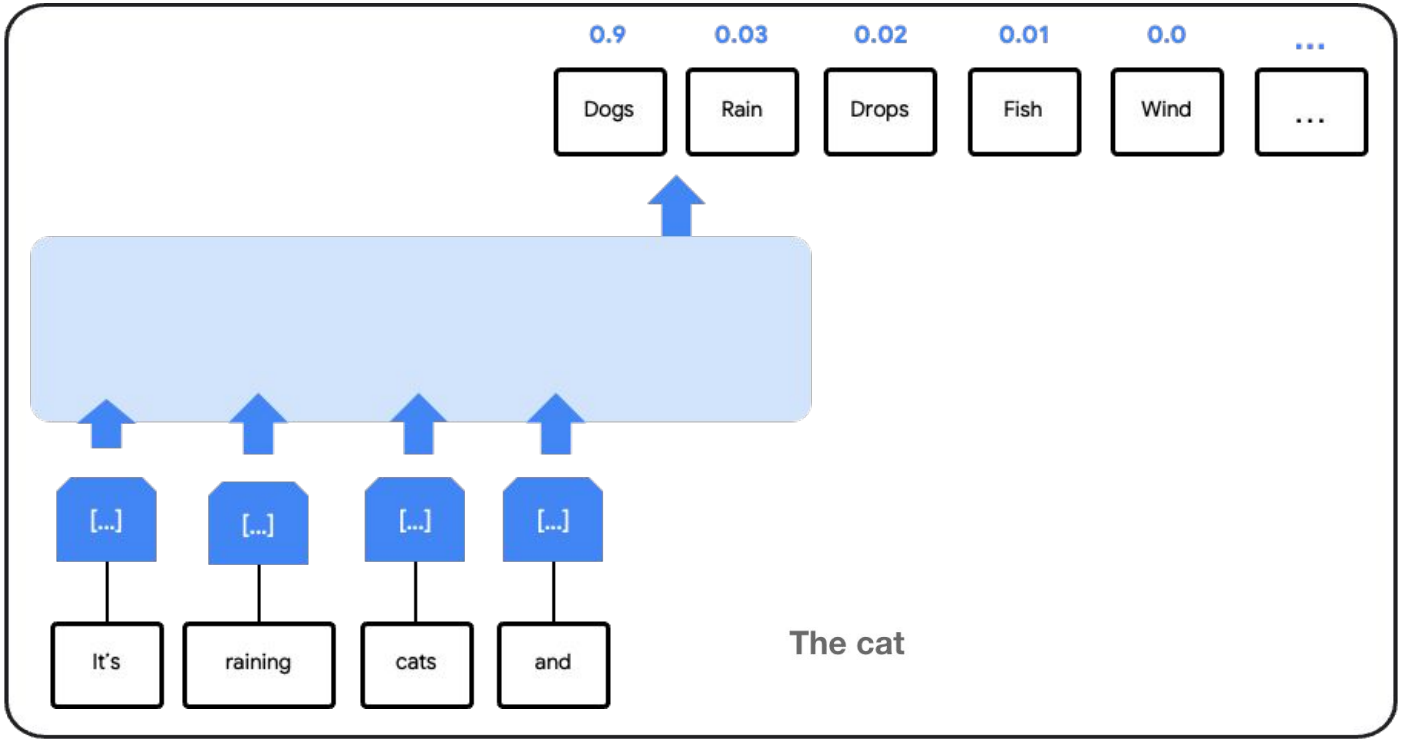
Chunk of
Internet,
~10 TB of Text



6000 GPUs for 12 days,
~$2M in Cost,
~1e24 FLOPS

~140 GB
parameters file

What is an LLM Really Learning?



Example Tasks

Text Completion

Roses are red,

Roses are red,
Violets are blue,
Sugar is sweet,

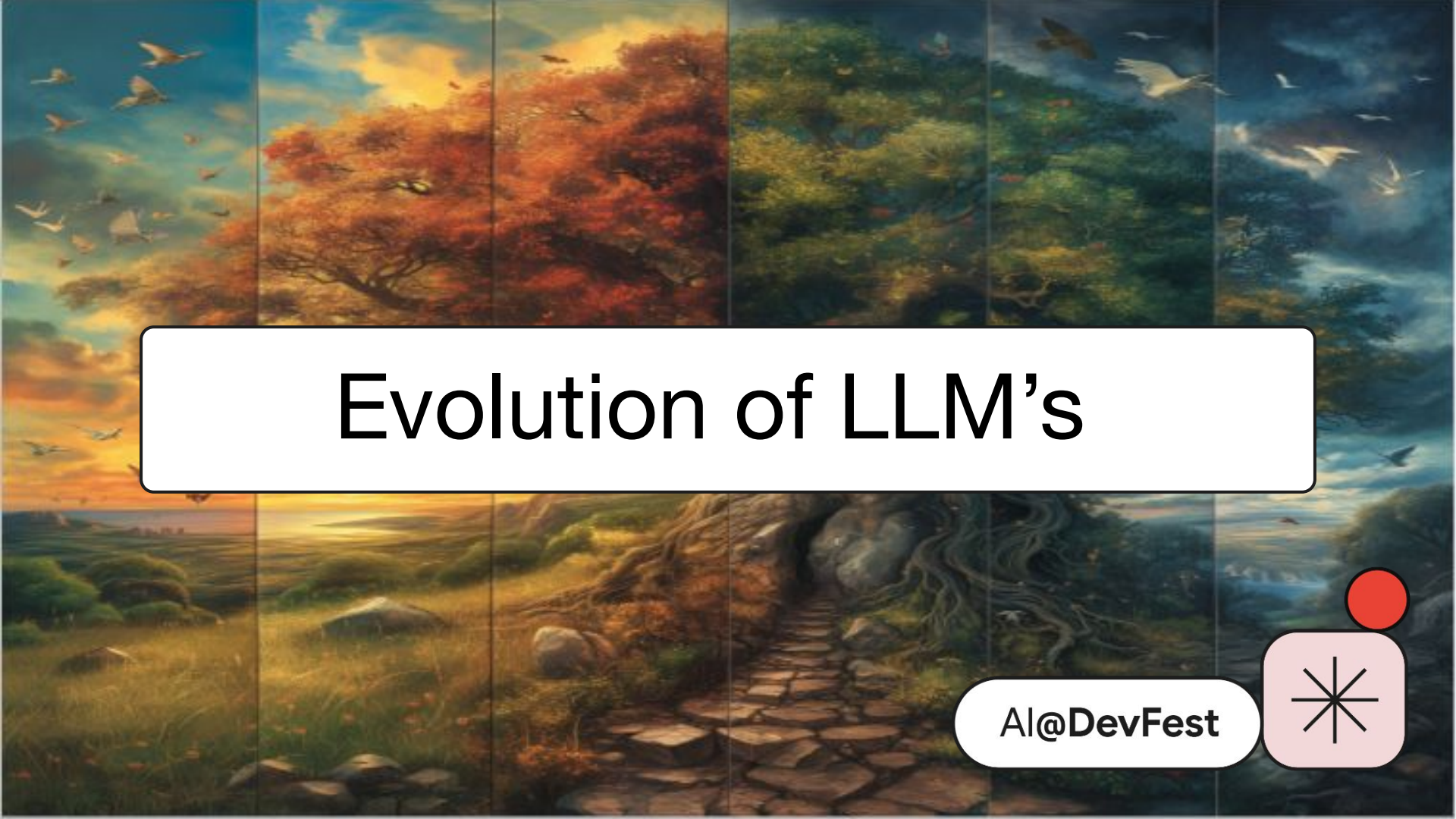
Code Completion

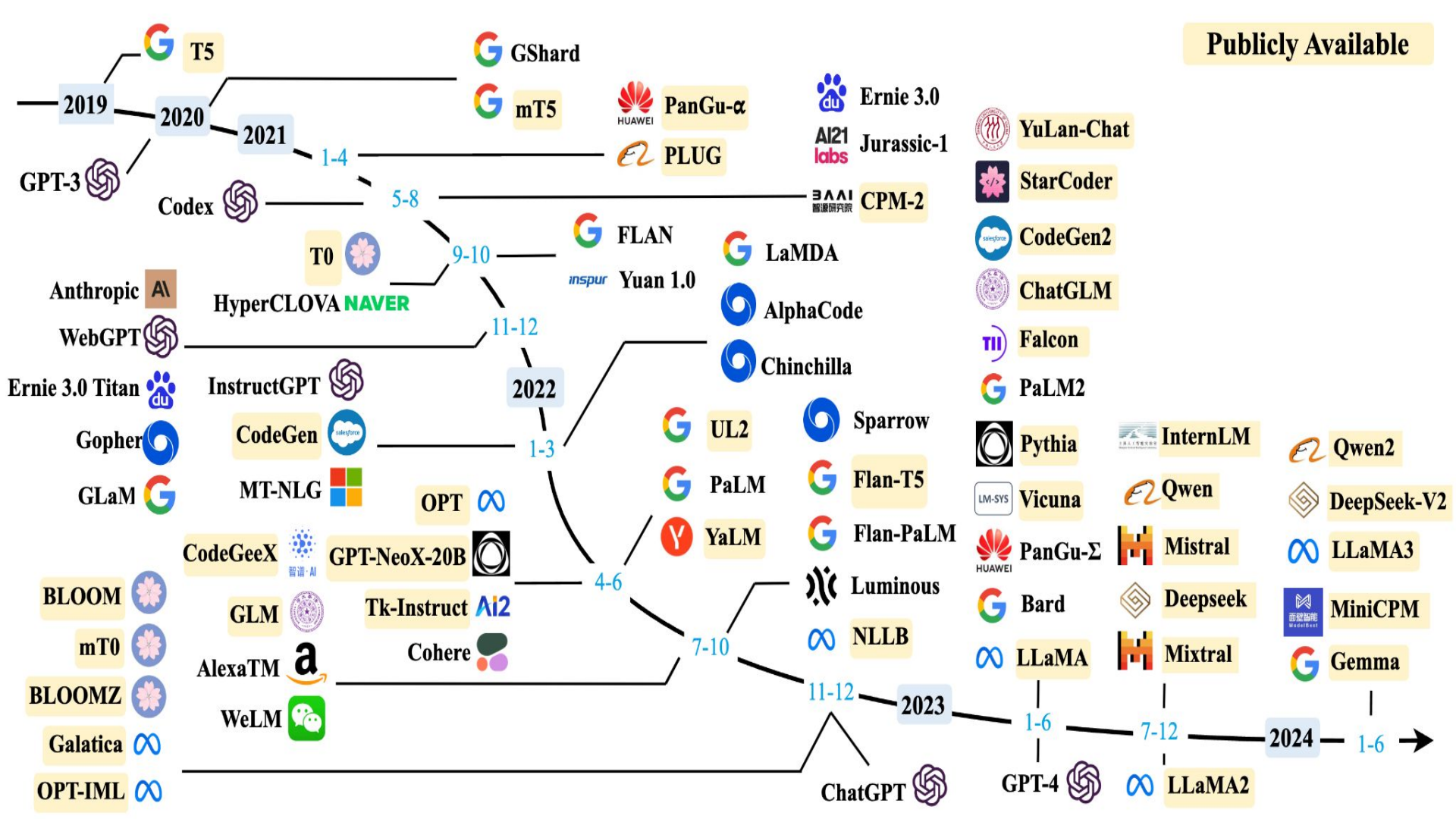
```
for(var i = 0
```

```
for(var i = 0; i <10; i++) {
```

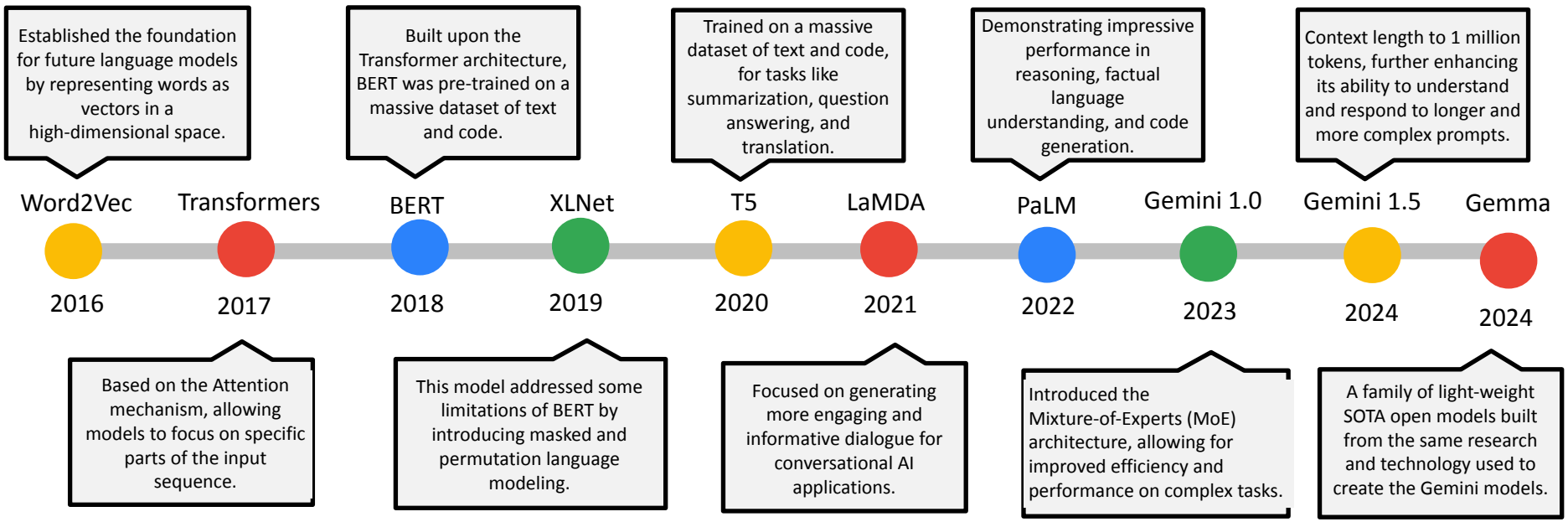
LLMs let us

prototype *fast*


Evolution of LLM’s
AI@DevFest


Publicly Available
2019
2020
2021
2022
2023
2024
T5
GPT-3
GShard
mT5
PanGu-α
PLUG
Ernie 3.0
Jurassic-1
CPM-2
YuLan-Chat
StarCoder
CodeGen2
ChatGLM
Falcon
PaLM2
Codex
1-4
5-8
9-10
11-12
1-3
4-6
7-10
11-12
1-6
7-12
1-6
T0
FLAN
Yuan 1.0
LaMDA
AlphaCode
Chinchilla
Anthropic
HyperCLOVA NAVER
WebGPT
Ernie 3.0 Titan
InstructGPT
Gopher
CodeGen
GLaM
MT-NLG
OPT
UL2
PaLM
YaLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
CodeGeeX
GPT-NeoX-20B
Tk-Instruct
GLM
Cohere
AlexaTM
WeLM
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
ChatGPT
Pythia
Vicuna
PanGu-Σ
Bard
LLaMA
GPT-4
InternLM
Qwen
Mistral
Deepseek
Mixtral
LLaMA2
Qwen2
DeepSeek-V2
LLaMA3
MiniCPM
Gemma

Language models over the years



Understanding Transformers

Self-Attention Mechanism - Allows the model to weigh the importance of different words in a sentence

- *Example*: In the sentence "The cat sat on the mat," it understands that "cat" is more related to "sat" than "mat."

Positional Encoding - Adds information about the position of words in a sentence

- *Example*: Helps the model distinguish between "John loves Mary" and "Mary loves John."

- Process entire input all at once, unlike previous models that processed word by word

- Better at understanding context and relationships between words

Understanding Transformers

Self-Attention Mechanism - Allows the model to weigh the importance of different words in a sentence

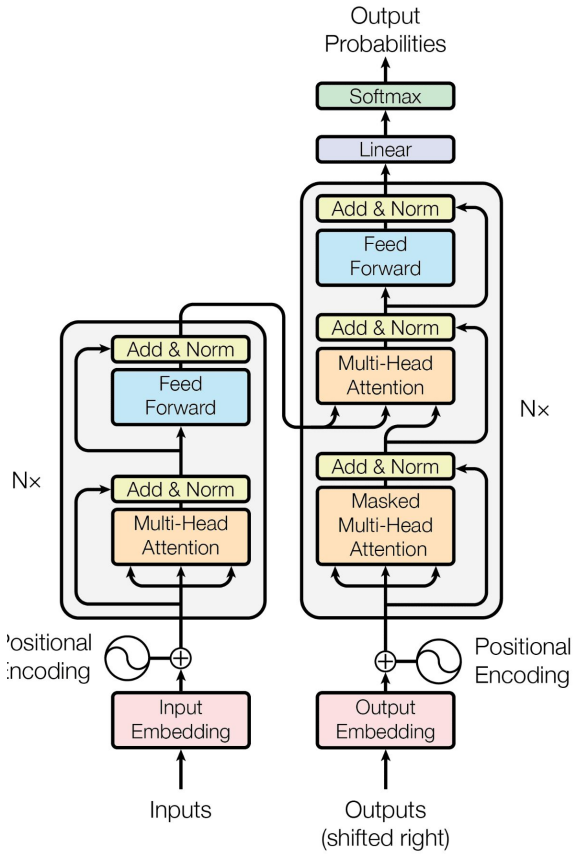
- *Example*: In the sentence "The cat sat on the mat," it understands that "cat" is more related to "sat" than "mat."

Positional Encoding - Adds information about the position of words in a sentence

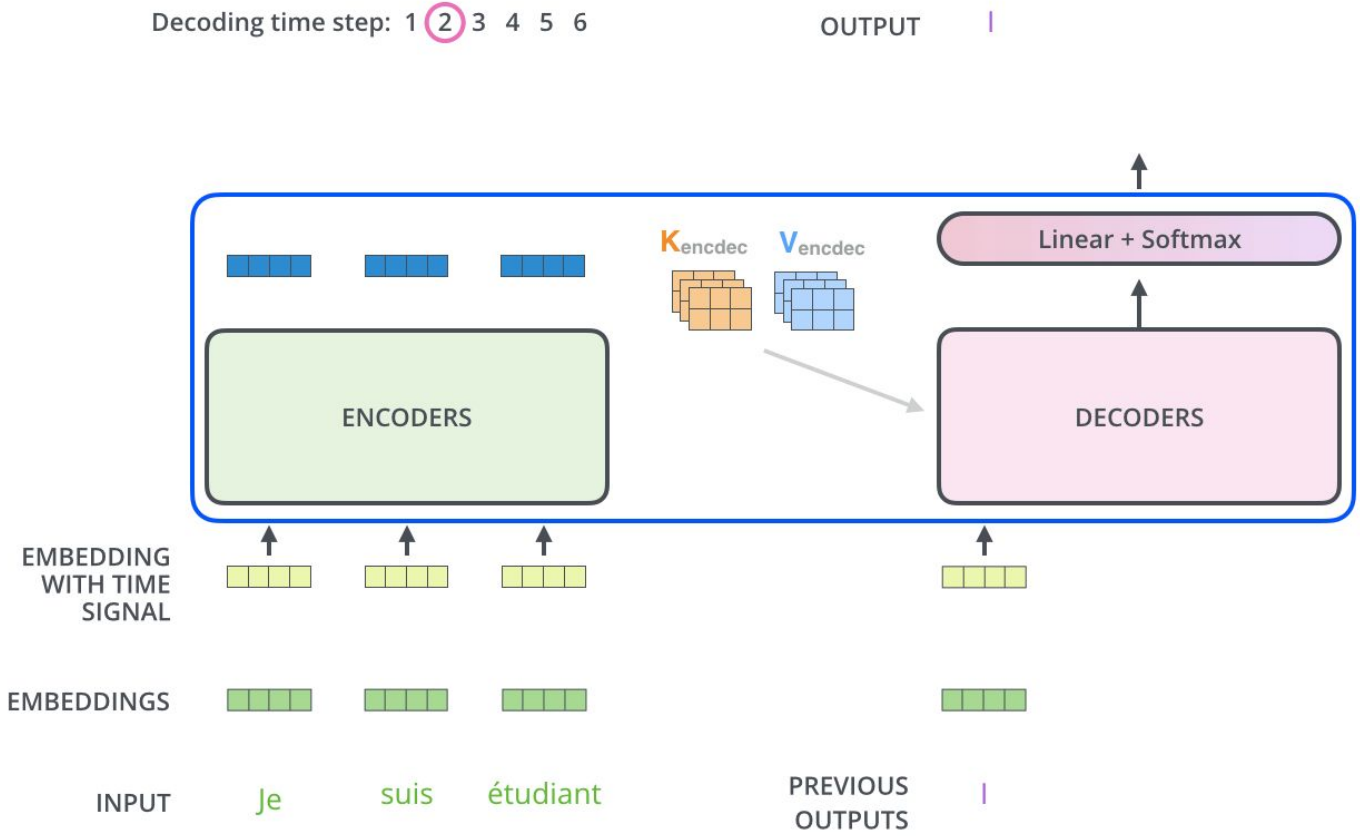
- *Example*: Helps the model distinguish between "John loves Mary" and "Mary loves John."

- Process entire input all at once, unlike previous models that processed word by word

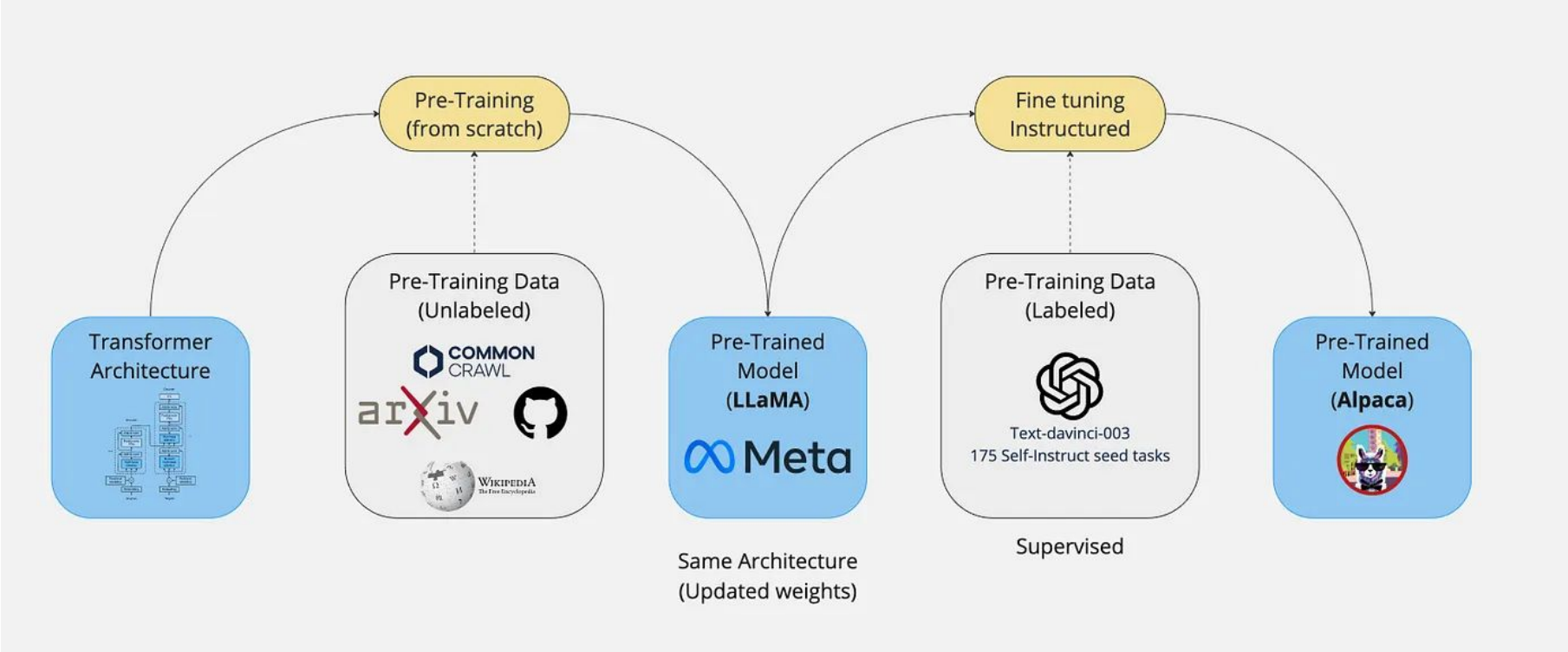
- Better at understanding context and relationships between words



Transformer Illustration for Translation Task



Pre-training vs Fine-tuning




Transformers → Applications
AI@DevFest

Gemma Open Models

A family of lightweight, state-of-the art open models built from the same research and technology used to create the Gemini models

How are these models trained?

These models were trained on a dataset of text data that includes a wide variety of sources, totaling 6 trillion tokens, including data from:

Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content.

Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions.

Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.

Performance with Gemma

Responsible by design

Incorporating comprehensive safety measures, these models help ensure responsible and trustworthy AI solutions through curated datasets and rigorous tuning.

Unmatched performance at size

Gemma models achieve exceptional benchmark results at its 2B, 7B, 9B, and 27B sizes, even outperforming some larger open models.

Framework flexible

With Keras 3.0, enjoy seamless compatibility with JAX, TensorFlow, and PyTorch, empowering you to effortlessly choose and switch frameworks depending on your task.

New release

DataGemma

DataGemma are the first open models designed to connect LLMs with extensive real-world data drawn from Google's Data Commons.

Gemma 1

Gemma models are lightweight, text-to-text, decoder-only large language models, trained on a massive dataset of text, code, and mathematical content for a variety of natural language processing tasks.

RecurrentGemma

RecurrentGemma is a technically distinct model that leverages recurrent neural networks and local attention to improve memory efficiency.

PaliGemma

PaliGemma is an open vision-language model inspired by PaLI-3, leveraging SigLIP and Gemma, designed as a versatile model for transfer to a wide range of vision-language tasks.

CodeGemma

Harnessing the foundation of our original pre-trained Gemma models, CodeGemma brings powerful code completion and generation capabilities in sizes fit for your local computer.

Gemma 2

Breakthrough performance and efficiency

- ***Class Leading Performance:*** At 9 and 27 billion parameters, Gemma 2 delivers performance comparable to Llama 3 70B at less than half the size for the latter.

- ***Reduced Deployment Costs:*** Gemma 2's efficient design allows it to fit on less than half the compute of comparable models. The 27B model is optimized to run on NVIDIA’s GPUs or can run efficiently on a single TPU host in Vertex AI, making deployment more accessible and cost-effective for a wider range of users.

- ***Versatile Tuning Toolchains:*** From cloud-based solutions like Google Cloud to popular community tools fine-tuning Gemma 2 will be easier. Integration with Hugging Face and NVIDIA TensorRT-LLM, along with JAX and Keras, ensures you can optimize performance and efficiently deploy across various hardware configurations.

PaliGemma 2

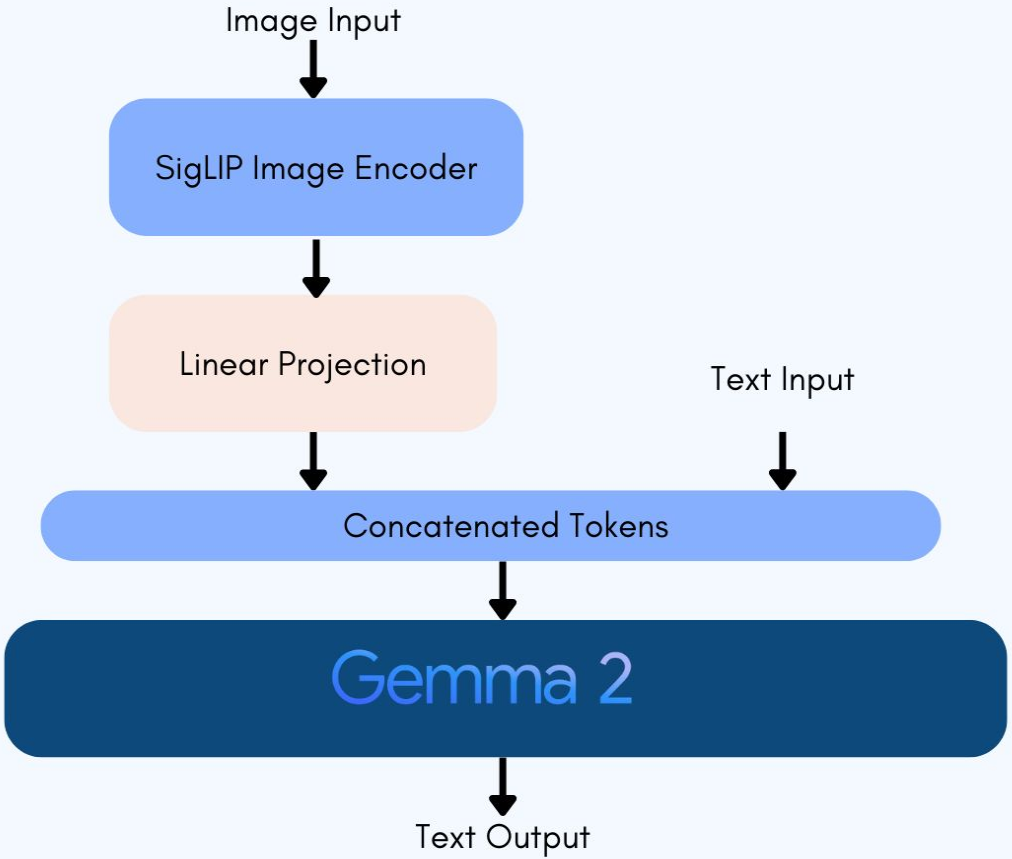


Image Input
SigLIP Image Encoder
Linear Projection
Text Input
Concatenated Tokens
Gemma 2
Text Output

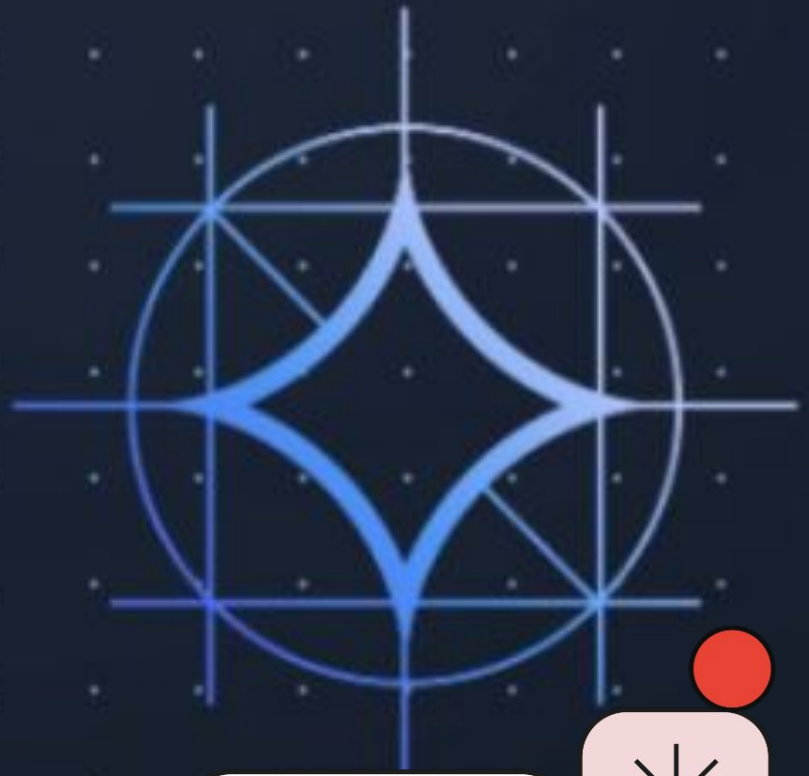
PaliGemma 2

- *Scalable performance:* Optimize performance for any task with PaliGemma 2's multiple model sizes (3B, 10B, 28B parameters) and resolutions (224px, 448px, 896px).

- *Long captioning:* PaliGemma 2 generates detailed, contextually relevant captions for images, going beyond simple object identification to describe actions, emotions, and the overall narrative of the scene.

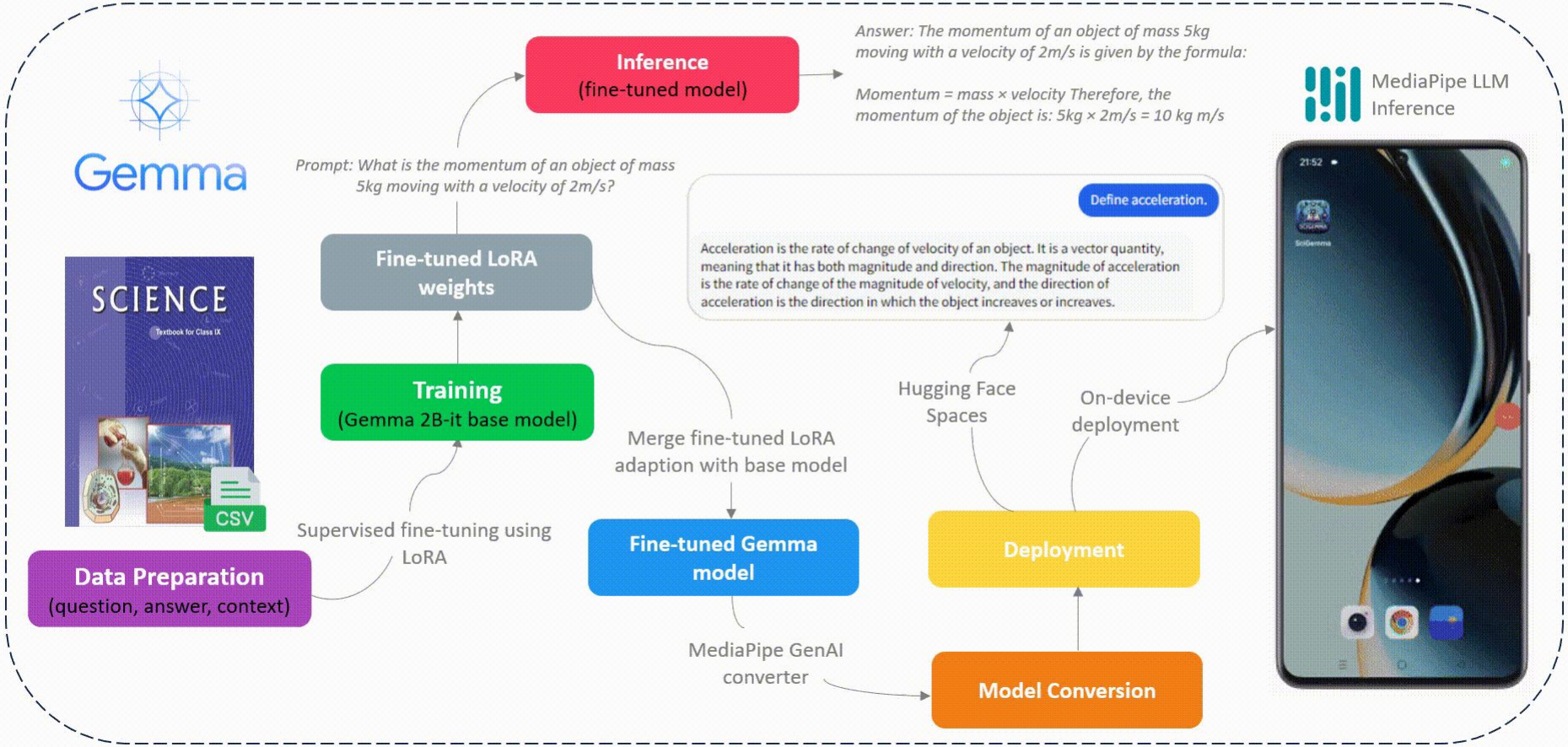
- *Expanding to new horizons:* Research demonstrates leading performance on chemical formula recognition, music score recognition, spatial reasoning, and chest X-ray report generation.

Built with Gemma



AI@DevFest

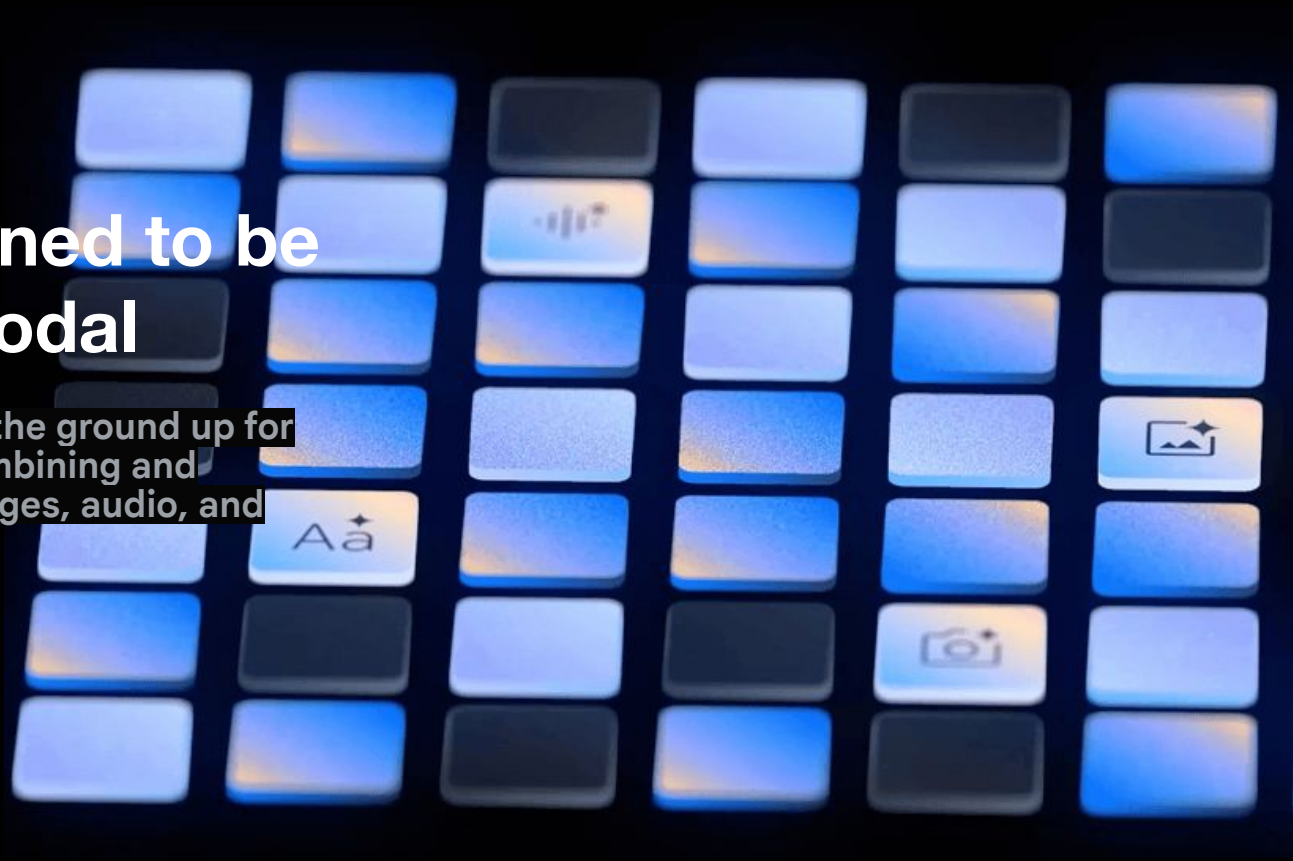
SciGemma: Fine-tuning and deploying Gemma on Android - Pipeline



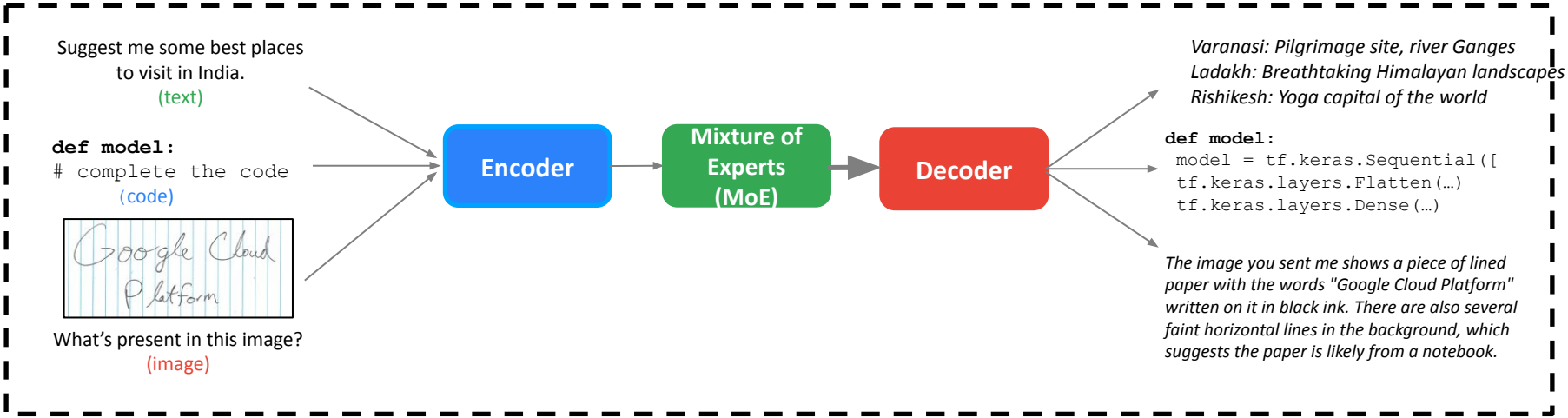

Gemini
Natively multimodal
Sophisticated reasoning
Advanced Coding

Gemini is designed to be natively multimodal

Gemini models are built from the ground up for multimodality, seamlessly combining and understanding text, code, images, audio, and video.



Architecture of Gemini



Encoder: Takes input formats like text, code, images, or audio and converts them into a common internal representation that the decoder can understand.

Decoder: Based on the encoded information and the specific task at hand, the decoder generates outputs in different modalities, such as text, code, or translated languages.

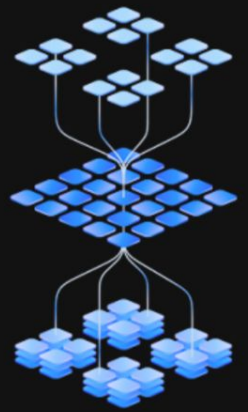
Mixture of Experts: This component employs a network of smaller, specialized models ("experts") instead of a single, monolithic model. The input is routed to the most relevant expert based on its characteristics for improved efficiency and performance.


Gemini 2.0
Built for the agentic era

PERFORMANCE

Gemini 2.0 is our most capable model yet, building on the strengths of our previous generations.

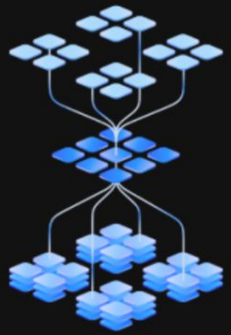
Gemini Model Family



1.0 Ultra

Our largest model for highly complex tasks.

Learn more



1.5 Pro

Our best model for reasoning across large amounts of information.

Learn more

2.0 Flash Experimental

Our workhorse model with low latency and enhanced performance, built to power agentic experiences.

Learn more

1.0 Nano

Our most efficient model for on-device tasks.

Learn more

Introducing Gemini 2.0 Flash Experimental

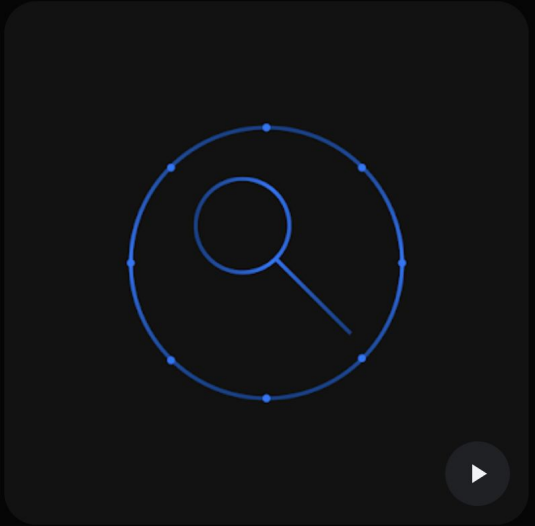
Our workhorse model with low latency and enhanced performance. 2.0 Flash Experimental introduces improved capabilities like native tool use and for the first time, Gemini can also natively create images and generate speech.

Native image generation

Create or edit images and seamlessly blend them with text.

Native text-to-speech

Easily steer Gemini's speaking style to match any mood.



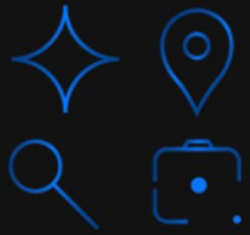
Native tool use

Build agents that use Google Search, code execution and more.

AI Agents

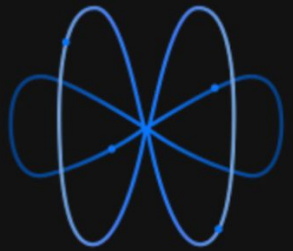
Taking action

Agents can follow instructions and take helpful actions under your supervision.



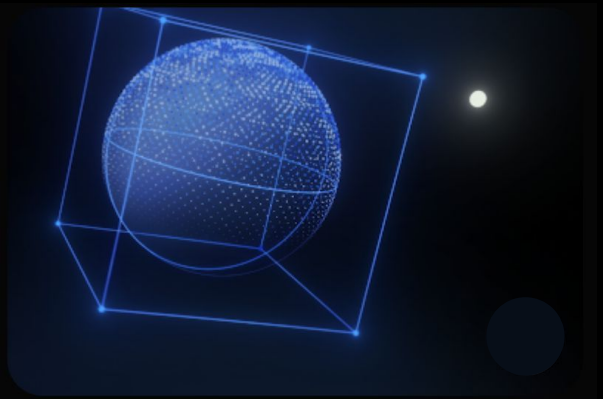
Tool use

Agents can search for information, look up reviews, translate and more.



Real-time streaming

Agents respond seamlessly to live audio and video input.



Spatial understanding

Ask Gemini to give you the locations of objects, text, and more.

Video understanding

Outline key moments, or summarize with an overview in a paragraph, or even a haiku.

Function calling with Maps API

Ask questions based on geography, or choose a pre-populated topic to watch the map travel to different locations using Google Maps.

Gemini helps you in daily life

- Content Generation
- Gmail Q & A
- Code Generation
- Ask Photos

Fit | Background | Layout

Gemini for Education

- Gemini can generate creative text formats, like poems, code, scripts, musical pieces, email, letters, etc., and answer your questions in an informative way, even if they are open ended, challenging, or strange.
- Gemini can be used to tutor students in various subjects.
- Gemini can help teachers create lesson plans.
- Gemini can create personalized learning experiences for students.

Close | Retry | Insert | Sources

Click to add speaker notes

Gemini

the Gemini Pro Vision model to identify medicines and provide details like usage, dosage, and diagnosis.

- **FarmScan:** An Android implementation of the Google Pro Vision model API that helps farmers assess the freshness of produce, estimate market value and shelf life, and gain other insights for cultivation and sales planning.

Sources (1)

Create a slide about "Gemini in Finance"

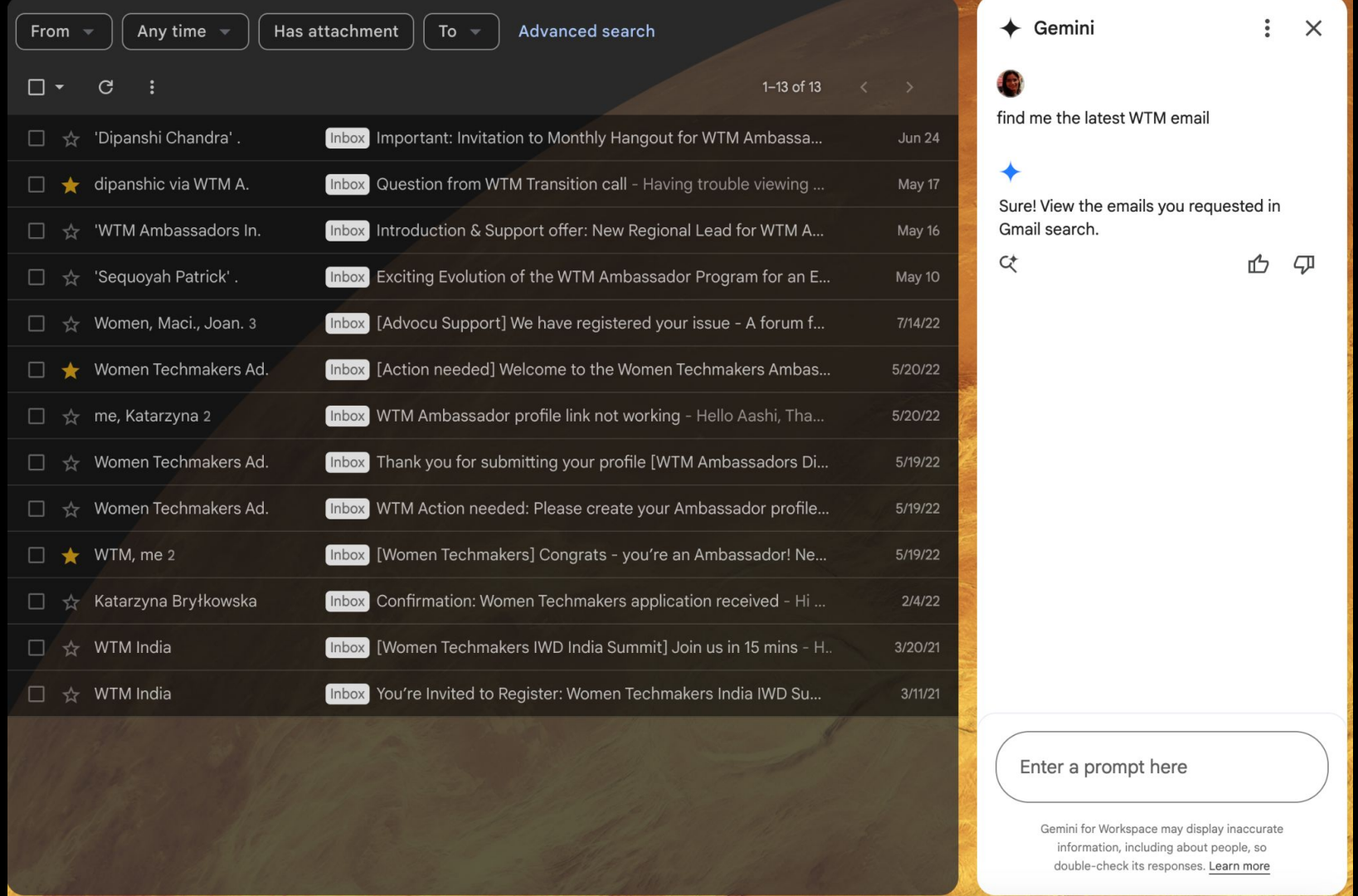
Create a slide about "Gemini for Education"

Create a slide about "Gemini for Education"

Show more suggestions

Enter a prompt here

Gemini for Workspace may display inaccurate information, including about people, so double-check its responses. Learn more


From
Any time
Has attachment
To
Advanced search
1–13 of 13
'Dipanshi Chandra' . Inbox Important: Invitation to Monthly Hangout for WTM Ambassa... Jun 24
dipanshic via WTM A. Inbox Question from WTM Transition call - Having trouble viewing ... May 17
'WTM Ambassadors In. Inbox Introduction & Support offer: New Regional Lead for WTM A... May 16
'Sequoyah Patrick' . Inbox Exciting Evolution of the WTM Ambassador Program for an E... May 10
Women, Maci., Joan. 3 Inbox [Advocu Support] We have registered your issue - A forum f... 7/14/22
Women Techmakers Ad. Inbox [Action needed] Welcome to the Women Techmakers Ambas... 5/20/22
me, Katarzyna 2 Inbox WTM Ambassador profile link not working - Hello Aashi, Tha... 5/20/22
Women Techmakers Ad. Inbox Thank you for submitting your profile [WTM Ambassadors Di... 5/19/22
Women Techmakers Ad. Inbox WTM Action needed: Please create your Ambassador profile... 5/19/22
WTM, me 2 Inbox [Women Techmakers] Congrats - you're an Ambassador! Ne... 5/19/22
Katarzyna Bryłkowska Inbox Confirmation: Women Techmakers application received - Hi ... 2/4/22
WTM India Inbox [Women Techmakers IWD India Summit] Join us in 15 mins - H.. 3/20/21
WTM India Inbox You're Invited to Register: Women Techmakers India IWD Su... 3/11/21
Gemini
find me the latest WTM email
Sure! View the emails you requested in Gmail search.
Enter a prompt here
Gemini for Workspace may display inaccurate information, including about people, so double-check its responses. Learn more

Gemini_demo.ipynb

File Edit View Insert Runtime Tools Help Last edited on 2 July

Comment Share

+ Code + Text

Connect Gemini

```
!pip install -q -U google-generativeai
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 163.9/163.9 kB 1.5 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 718.3/718.3 kB 9.9 MB/s eta 0:00:00
```

```
[ ] from google.colab import userdata
    import google.generativeai as genai
    from PIL import Image
```

```
[ ] API_KEY = userdata.get('gemini')

    # Configure the API key.
    genai.configure(api_key=API_KEY)
```

```
[ ] # List all available models.
    for m in genai.list_models():
      if 'generateContent' in m.supported_generation_methods:
        print(m.name)
```

```
models/gemini-1.0-pro
models/gemini-1.0-pro-001
models/gemini-1.0-pro-latest
models/gemini-1.0-pro-vision-latest
models/gemini-1.5-flash
models/gemini-1.5-flash-001
models/gemini-1.5-flash-latest
models/gemini-1.5-pro
models/gemini-1.5-pro-001
models/gemini-1.5-pro-latest
models/gemini-pro
models/gemini-pro-vision
```

Gemini Pro 1.0 (Text input)

Gemini

Aashi Dutt

Show me a list of publicly available datasets

Gemini

There are a number of publicly available datasets available in Google Colab. To access them, simply navigate to the left sidebar and click on "Files". Then, click on the folder icon with a down arrow, which will open up a tab called "Sample Data". In this tab, you will find a number of datasets such as:

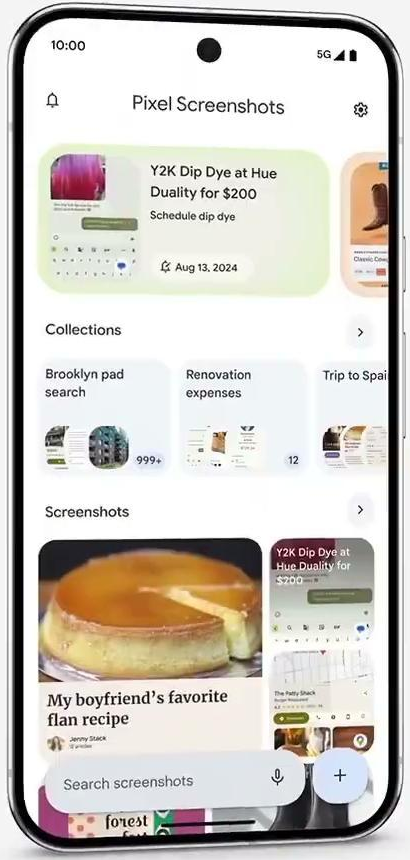
- California Housing
- MNIST
- Shakespeare

Rate this answer

Gemini access in Colab is available to unsubscribed users for a limited time only. Upgrade to Colab Pro

Enter a prompt here

0 / 400


10:00
5G
Pixel Screenshots
Y2K Dip Dye at Hue Duality for $200
Schedule dip dye
Aug 13, 2024
Collections
Brooklyn pad search
999+
Renovation expenses
12
Trip to Spai
Screenshots
My boyfriend's favorite flan recipe
Jenny Stack
Y2K Dip Dye at Hue Duality for $200
Search screenshots
forest

Built with Gemini

PharmaScan : Extract medicine details using Gemini

PharmaScan is an Android application that leverages the Gemini Pro Vision model to identify medicines and provide their details such as usage, dosage, diagnosis, etc. on the go.

Check out on GitHub:
https://github.com/NSTiwari/Medicine-Scan-with-Gemini

Try it on Hugging Face 🤗:
https://huggingface.co/spaces/Aashi/Medicine-Prescription-with-Gemini
Watch the complete demo on YouTube:
https://www.youtube.com/watch?v=Q06ABLwFGTQ

FarmScan : Farmer's Digital Assistance built with Gemini

FarmScan is an implementation of the Google Pro Vision model API on Android to recognize the freshness of fruits/vegetables, their approximate market value, shelf life, and a lot more insights to help farms plan the cultivation/selling of crops better.

Check out on GitHub:
https://github.com/NSTiwari/FarmScan-using-Gemini

Try it on Hugging Face 🤗:
https://huggingface.co/spaces/Aashi/FarmScan


IIT'S DEMO
TIME!

Gemini 2.0 Flash Experimental Realtime

I’m just a Chill Guy

All I need is

Key Takeaways

- Transformers Revolutionized NLP
- LLMs as Compressed Knowledge Banks
- Fine-Tuning for Specific Use-Cases
- Future Lies in Adaptability
- Real-World Applications are Expanding


Thank You
AI@DevFest