Google Developer Group

Search Tool with Gemini 2.0

A Codelab and Introduction

Linh Nguyen - Head of AI @ Obello (Silicon Valley)

Google Developer Expert in AI/ML

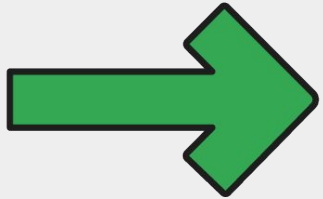
Build with AI

Agenda

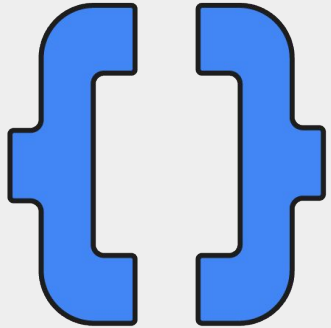
1. Gemini 2.0: Overview
2. Model Families & Benchmark
3. Features & Applications
4. Hands-on with Codelab!

Chapter One

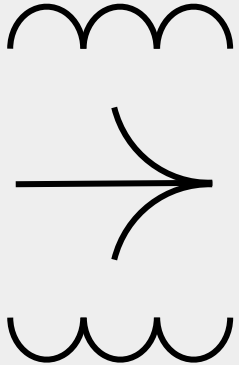
Gemini 2.0: Overview



Gemini 2.0



Gemini 2.0 represents a significant leap forward as Google DeepMind's most capable AI model yet, specifically designed for what they term the "agentic era". Building upon the foundations laid by Gemini 1.0 and 1.5, this new generation of models brings substantial advancements in multimodality, enabling it to understand and process information across **text, images, video,** and **audio** with greater proficiency.

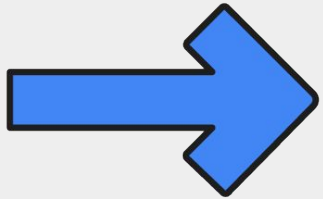


This evolution signifies a move towards creating intelligent AI agents that can **reason, plan, remember, and take actions** to assist users in more comprehensive and autonomous ways.


Gemini 2.0
Enabling the
agentic era

Chapter One

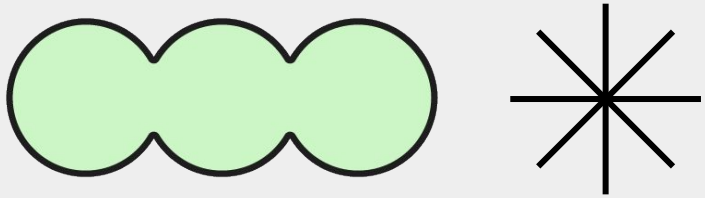
Model Families & Benchmark

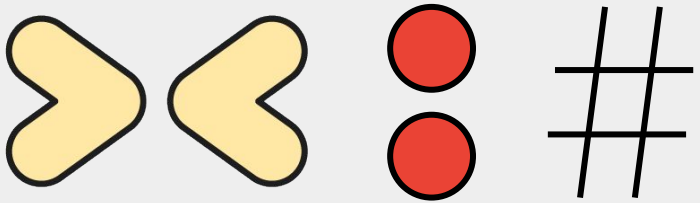


Model Families

Gemini 2.0 Flash

Powerful workhorse model with low latency and enhanced performance for agentic experiences.





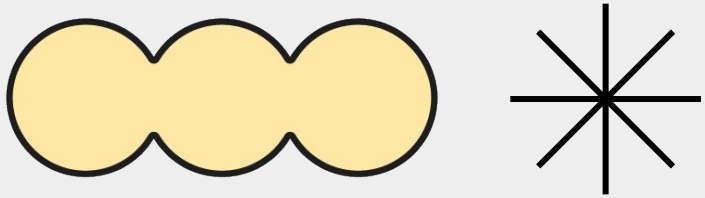
Gemini 2.0 Flash Thinking

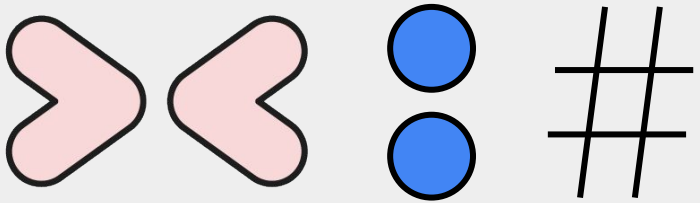
Enhanced reasoning model capable of showing its "thinking process" in Google AI Studio for improved explainability.

Model Families

Gemini Pro

Best model yet for coding performance and complex prompts.





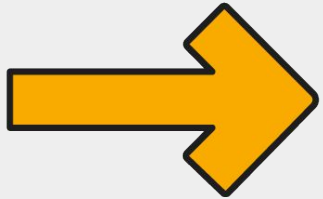
Gemini 2.0 Flash-Lite

Low latency and cost-effective.

CAPABILITY	BENCHMARK	DESCRIPTION	Gemini 1.5 Flash 002	Gemini 1.5 Pro 002	Gemini 2.0 Flash Experimental
General	MMLU-Pro	Enhanced version of popular MMLU dataset with questions across multiple subjects with higher difficulty tasks	67.3%	75.8%	**76.4%**
Code	Natural2Code	Code generation across Python, Java, C++, JS, Go . Held out dataset HumanEval-like, not leaked on the web	79.8%	85.4%	**92.9%**
	Bird-SQL (Dev)	Benchmark evaluating converting natural language questions into executable SQL	45.6%	54.4%	**56.9%**
	LiveCodeBench (Code Generation)	Code generation in Python. Code Generation subset covering more recent examples: 06/01/2024 - 10/05/2024	30.0%	34.3%	**35.1%**
Factuality	FACTS Grounding	Ability to provide factuality correct responses given documents and diverse user requests. Held out internal dataset	82.9%	80.0%	**83.6%**
Math	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	77.9%	86.5%	**89.7%**
	HiddenMath	Competition-level math problems, Held out dataset AIME/AMC-like, crafted by experts and not leaked on the web	47.2%	52.0%	**63.0%**
Reasoning	GPQA (diamond)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry	51.0%	59.1%	**62.1%**
Long context	MRCR (1M)	Novel, diagnostic long-context understanding evaluation	71.9%	**82.6%**	69.2%
Image	MMMU	Multi-discipline college-level multimodal understanding and reasoning problems	62.3%	65.9%	**70.7%**
	Vibe-Eval (Reka)	Visual understanding in chat models with challenging everyday examples. Evaluated with a Gemini Flash model as a rater	48.9%	53.9%	**56.3%**
Audio	CoVoST2 (21 lang)	Automatic speech translation (BLEU score)	37.4	**40.1**	39.2
Video	EgoSchema (test)	Video analysis across multiple domains	66.8%	71.2%	**71.5%**

Chapter One

Features & Applications



Features & Under the Hood

Optimized

Built on an optimized Transformer architecture, especially for multimodal understanding

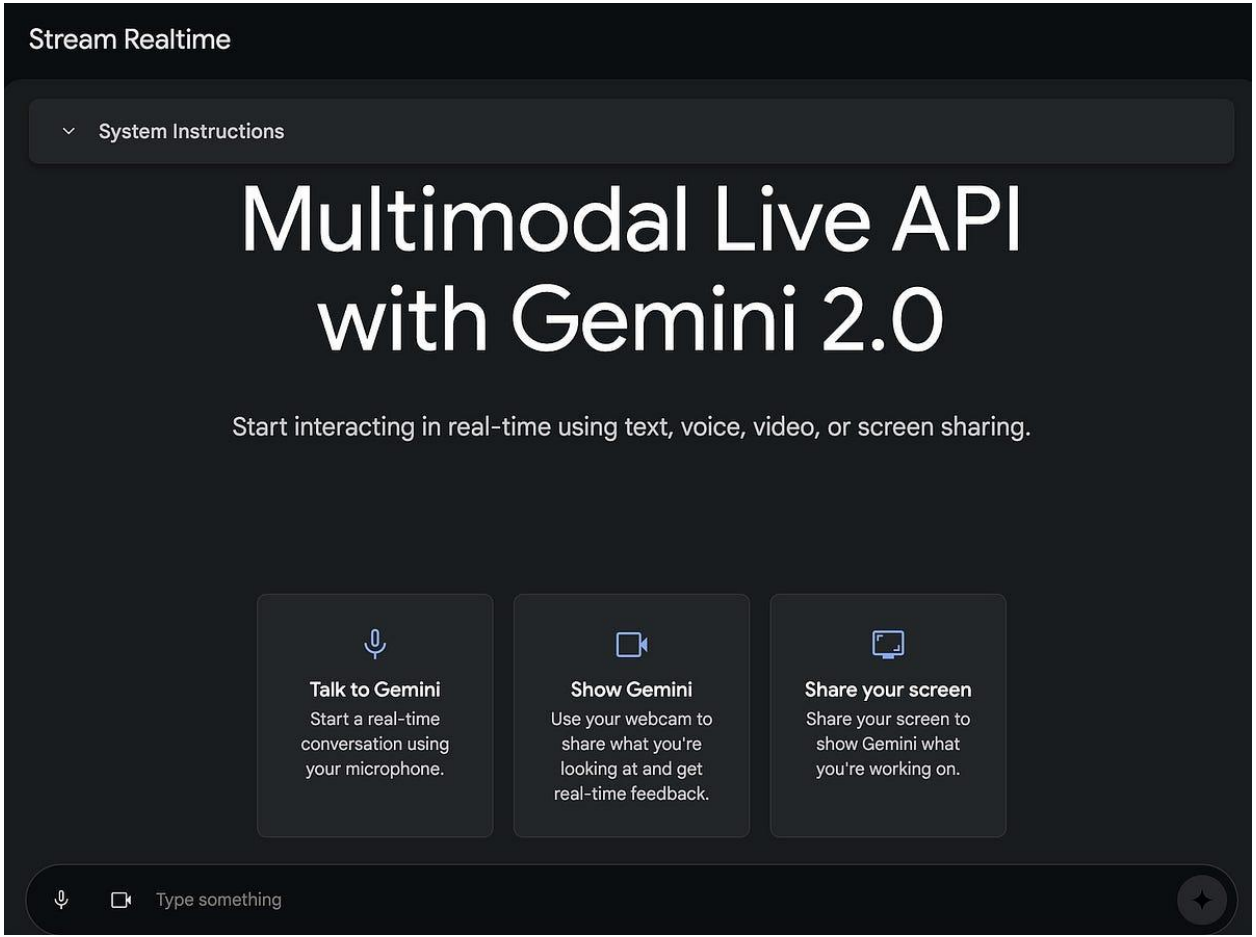
6th's Gen TPU

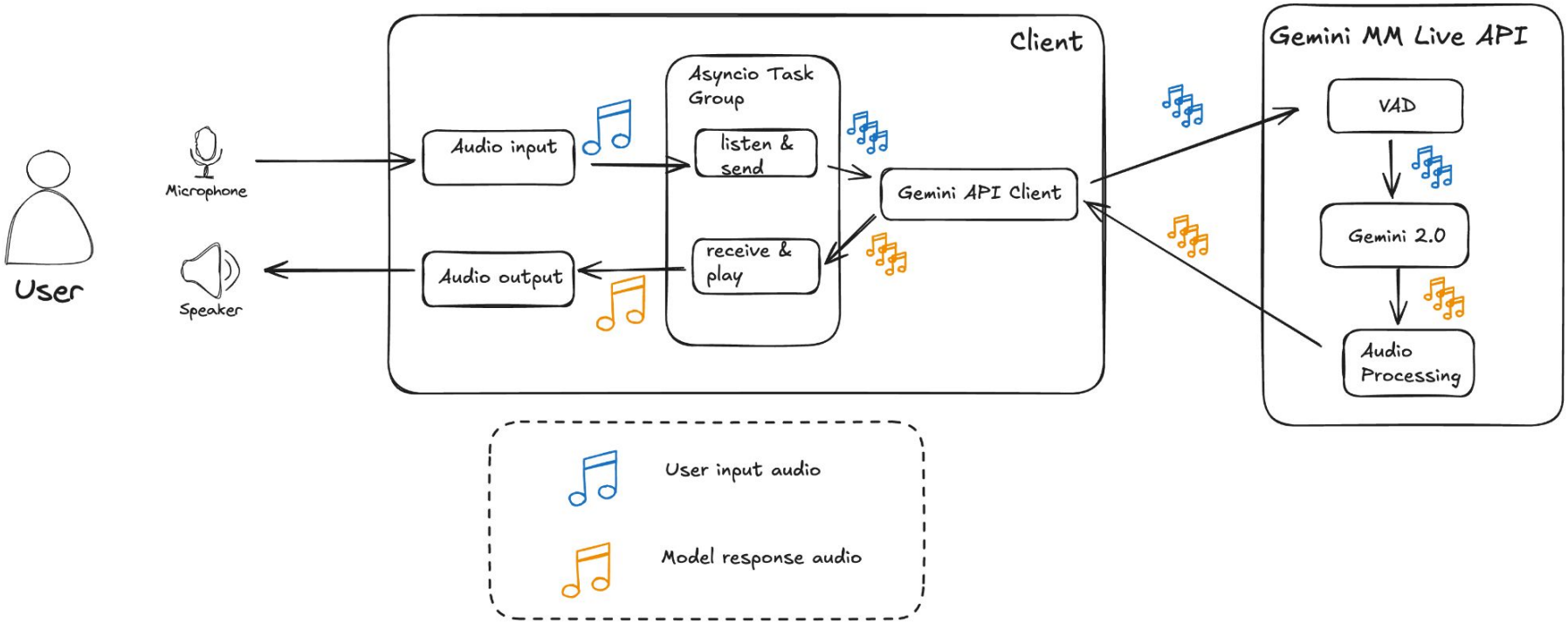
Leverages Google's Sixth-Generation Tensor Processing Units (TPUs) (Trillium) for accelerated training and inference

Scalability

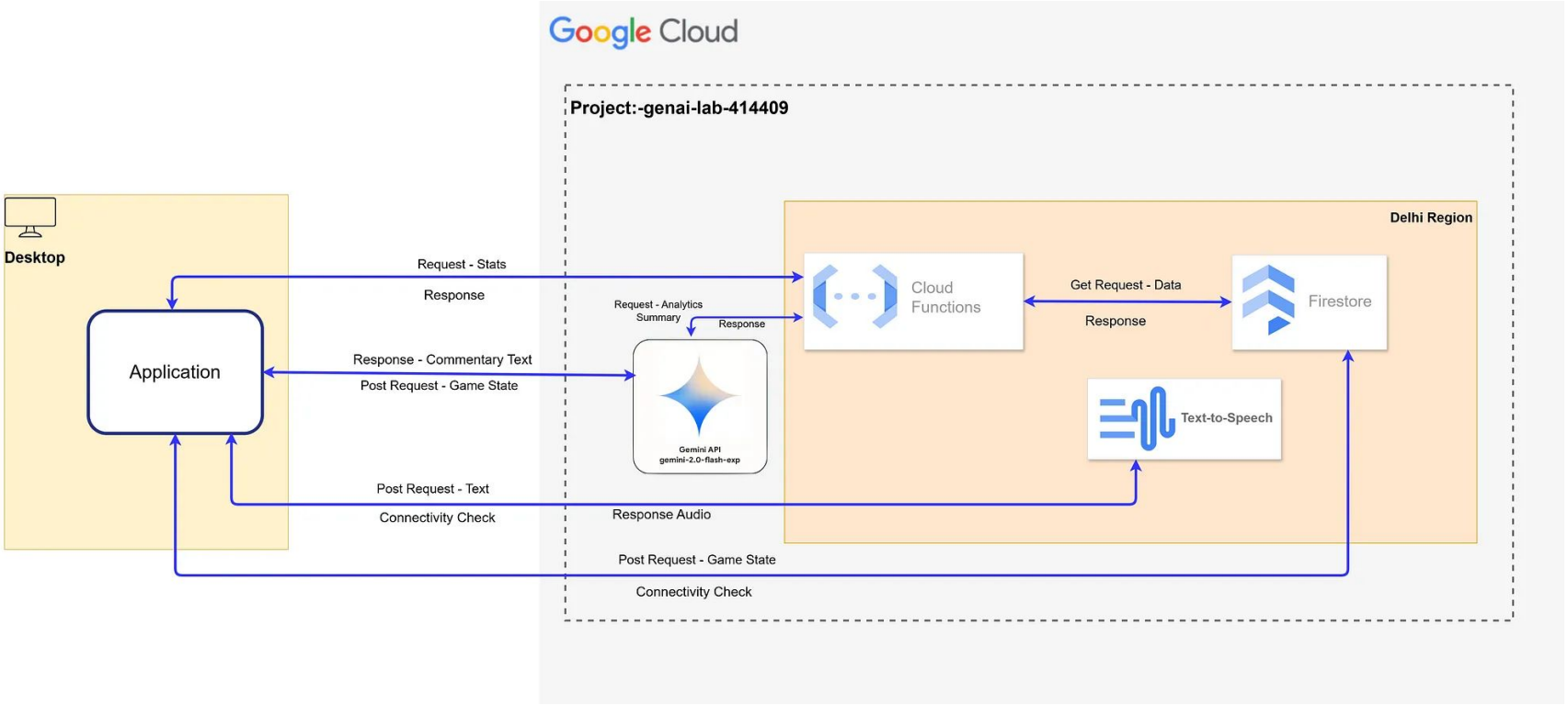
Utilizes the JAX/XLA framework for scalability and computational efficiency

**Only for Gemini 2.0 Flash: Optimized for low latency with reduced time-to-first-token (TTFT)*

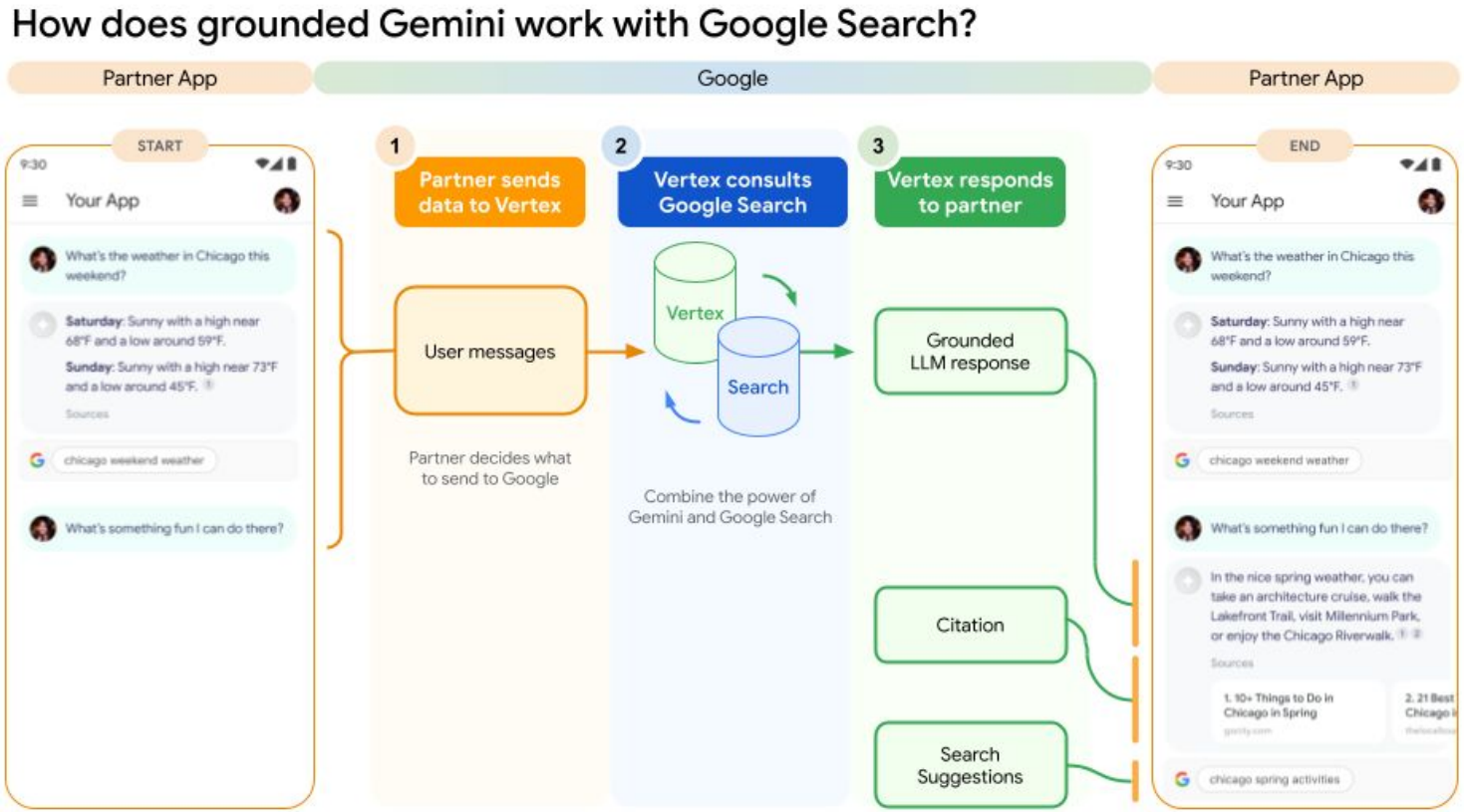

Stream Realtime
System Instructions
Multimodal Live API with Gemini 2.0
Start interacting in real-time using text, voice, video, or screen sharing.
Talk to Gemini
Start a real-time conversation using your microphone.
Show Gemini
Use your webcam to share what you're looking at and get real-time feedback.
Share your screen
Share your screen to show Gemini what you're working on.
Type something



Talking and interacting live with Gemini 2.0



Live commentary while playing game



Revolutionize your Search workflow with Gemini 2.0, and ground it with search grounding

Codelab

Scan this QR to the codelab

Or click here to get to the lab

Google Developer Group

Thank you!

Nguyen Khanh Linh (she/her)

https://linhkid.github.io

Build with AI