The future of on-device intelligence

Gemma and Gemini Nano

Tris Warkentin | Director of Product Management | Google Deepmind

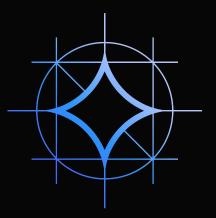




Tris Warkentin Director, Product Management Google DeepMind

Google DeepMind





Gemma



BGE-Multilingual-Gemma2 is a LLM-base multilingual embedding model...

HuggingFaceH4/
zephyr-7b-gemma-v0.1

Zephyr is a series of language models that are trained to act as helpful assistants...



Gemmaverse

Princeton-equi
genne-2-90-11-Sin-20

5 n-2-3 ling o Port-conce Cod master is a
office protector optimization algorithm.

yam-peleg/ hebrew-gemma-11B

> ebrew-Gemma-11B is an open-source Large anguage Model (LLM) is a hebrew/english...

Skywork/
 skywork-reward-gemma-2-27b

Trained using the Skywork Reward Data Collection containing only 80K high-quality preference...

Variants

PaliGemma

CodeGemma

ShieldGemma

RecurrentGemma

SignGemma

MedGemma

T₅Gemma

DataGemma

. . .

Main Models

Gemma

Gemma 2

Gemma 3

Gemma 3n

Tools

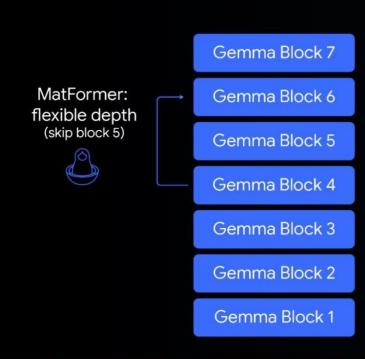
Gemma Scope

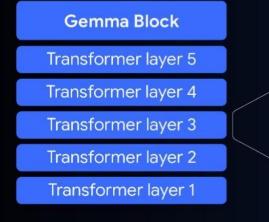
Gemma.cpp





MatFormer in Gemma 3n







Gemini Nano



	Prefix Speed - Gemini	Prefix Speed - Gemini	Prefix Speed - Gemini
	nano-v2 on Pixel 9	nano-v2* on Pixel 10	nano-v3 on Pixel 10
	Pro	Pro	Pro
Text-to-text	510 tokens/second	610 tokens/second	940 tokens/second
lmage-to-text	510 tokens/second +	610 tokens/second +	940 tokens/second +
	0.8 seconds for image	0.7 seconds for image	0.6 seconds for image
	encoding	encoding	encoding



Fast Context aware Completely private







Home

Competitions

Datasets

Models

Benchmarks

Code

Discussions

Learn

More

Q Search

Sign In

Register

GOOGLE DEEPMIND · FEATURED HACKATHON · A MONTH AGO

Accept Rules

Google - The Gemma 3n Impact Challenge

Explore the newest Gemma model and build your best products for a better world



Overview Code Discussion Writeups

Overview

Your mission is to leverage the unique capabilities of Gemma 3n to create a product that addresses a significant realworld challenge. Think bigger than a simple chatbot. How can a private, offline-first, multimodal model make a tangible difference in people's lives?

Start

Close Aug 6, 2025 Jun 26, 2025

Competition Host

Google DeepMind



Prizes & Awards

\$150,000 Does not award Points or Medals

Participation

8,621 Entrants 609 Submissions











Thank You!

