

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

Gemmaverse



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

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



BAAI/
bge-multilingual-gemma2

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



dschm-mayr/
Series-2-VL-Chinese-Chat

The first instruction-tuned language model built with Google Gemma 2 and LLaVA-1.5...



Tristan-Jung/
gemma-7-9b-11-StepN

StepN (Step is Reinforcement Learning) is an off-policy preference optimization algorithm...



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

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



yam-peleg/
hebrew-gemma-11B

Hebrew-Gemma-11B is an open-source Large Language Model (LLM) is a hebrew/english...

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

MatFormer:
flexible depth
(skip block 5)



Gemma Block 7

Gemma Block 6

Gemma Block 5

Gemma Block 4

Gemma Block 3

Gemma Block 2

Gemma Block 1

Gemma Block

Transformer layer 5

Transformer layer 4

Transformer layer 3

Transformer layer 2

Transformer layer 1



MatFormer:
flexible FFN
(8x & 4x)



FFN



Attn



Gemini Nano

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

Fast
Context aware
Completely private



GOOGLE DEEPMIND · FEATURED HACKATHON · A MONTH AGO

Google - The Gemma 3n Impact Challenge

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


[Overview](#)
[Data](#)
[Code](#)
[Discussion](#)
[Writeups](#)
[Rules](#)

Overview

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

Start

Jun 26, 2025

Close

Aug 6, 2025

Competition Host

Google DeepMind



Prizes & Awards

\$150,000

Does not award Points or Medals

Participation

8,621 Entrants

609 Submissions





Agentic AI



TIMELESS[®]
PERSON OF THE YEAR 2026



THE REFLECTION OF OUR TIMES

DECEMBER 2026
£9.99 / \$14.99



Thank You!