Gemma 3

Gemma 3 is the latest family of open-source, lightweight models from Google DeepMind. The big deal with this version is that they're multimodal, meaning they can handle both text and images. They're built to be efficient and run on standard hardware.

Gemma 3 , Gemma 3 Technical Report

Overview

Developer: Google DeepMind

Release Date: March 2025

Type: Family of open, lightweight, multimodal (text and image) models

Architecture

Type: Decoder-only Transformer

Attention: Uses a 5:1 ratio of local (sliding window) to global self-attention layers. The local attention span is 1024 tokens.

Context Length: 128K tokens for 4B, 12B, and 27B models; 32K for the 1B model

Tokenizer: SentencePiece tokenizer with a 262k vocabulary

Vision

Encoder: A 400M variant of SigLIP vision encoder

Input Resolution: 896x896 pixels

Inference Method: Includes a “Pan & Scan” adaptive windowing algorithm to handle high-resolution and non-square images.

Model Family

Model	Total Parameters	Vision Encoder
1B	1 billion	None
4B	4.3 billion	417M
12B	12.2 billion	417M
27B	27.4 billion	417M