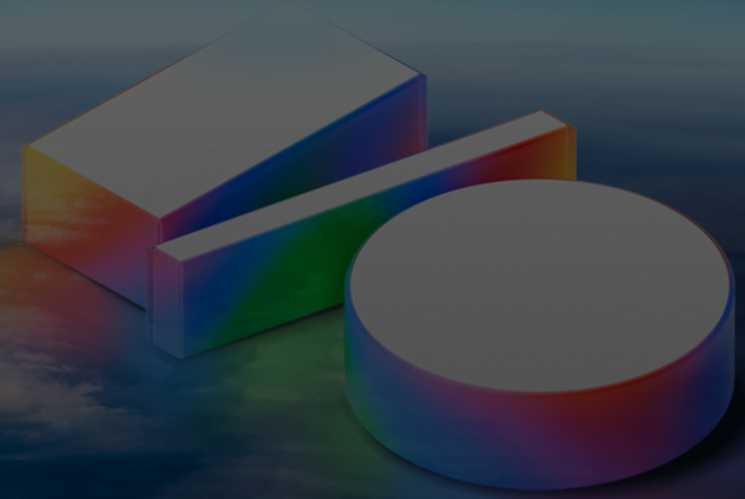
Gemma 4 : New Models



Google I/O Extended

- Martin Andrews = Gemma4 @ mdda.net

27-June-2026

Google I/O Extended Singapore 2026

About Me

- ML / Startups / Finance
  - NYC → Singapore in Sep-2013
- 2014 = 'fun' :
  - Machine Learning, DL, NLP
  - Robots, drones
- Since 2015 = 'serious'
  - NLP + deep learning
    - Including Papers...
  - & GDE ML
    - ML-Singapore co-organiser...
  - & Red Dragon AI...

MLSG
MACHINE LEARNING
SINGAPORE

Outline

- Gemma Models overview
- Gemma 4 12B
  - how the multi-modal interfaces work
  - demo!
- Gemma 4 Diffusion
  - why diffusion?
  - how diffusion works
  - demo!
- Wrap-up & QR-code

MLSG
MACHINE LEARNING
SINGAPORE

Gemma Model Overview

Google's LLM releases

- Pre-Gemma
  - BERT, T5, ...
- Gemma series of models
  - License = "Gemma Terms of Use"
- Gemma (1): February 2024
  - 2B and 7B (8k context)
    - arrived after Llama 2
- Gemma 2: June 2024
  - 2B, 9B and 27B (8k context)
  - PaliGemma 1/2 (add 400M SigLIP image encoder)
- Gemma 3: March 2025
  - 1B (non-vision), 4B, 12B and 27B (131k context)
  - MedGemma, ShieldGemma, TranslateGemma

MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 Launch

MLSG
MACHINE LEARNING
SINGAPORE

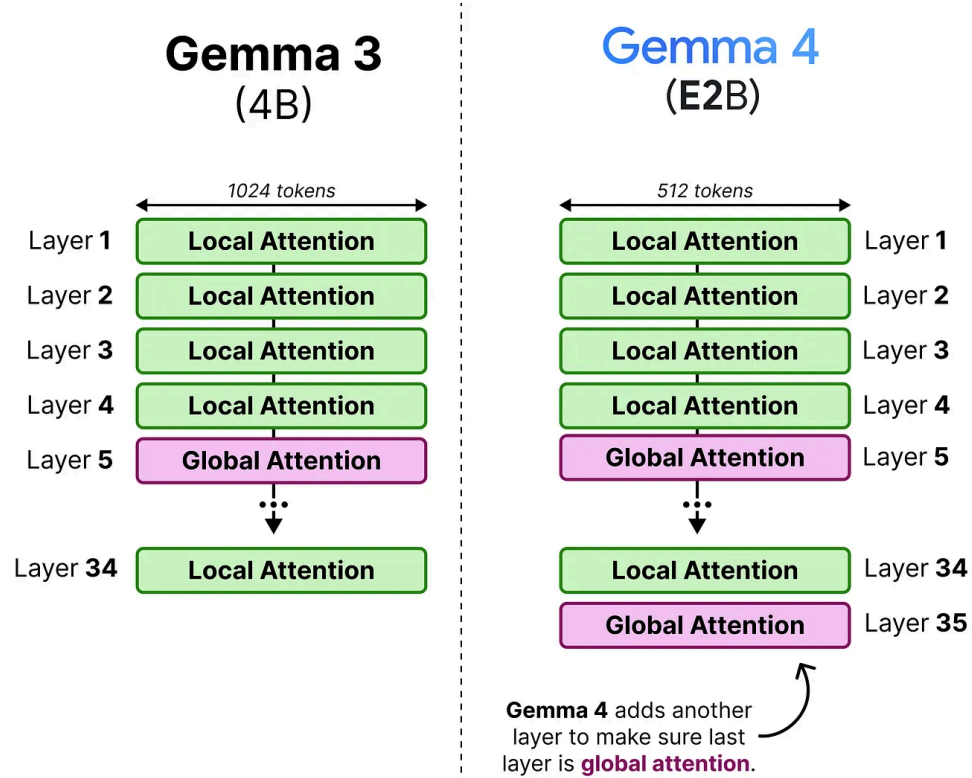
Gemma 4 Launch

- Gemma 4: April 2026
  - Apache 2.0
- Edge models: (128k context)
  - Gemma 4 E2B, Gemma 4 E4B
- Medium models: (256k context)
  - Gemma 4 26B A4B "MoE"
  - Gemma 4 31B "Dense"
- Innovations explored by branches of series
  - edge and server models share few design decisions

All Gemma 4 Models

- Local vs Global attention mix
  - Alternating local sliding-window and global full-context attention layers
  - Edge models use sliding windows of 512 tokens
    - larger models use 1024 tokens
  - Dual RoPE configurations:
    - standard RoPE for sliding layers
    - pruned RoPE for global layers
      - (to enable longer context)

- Quantisation-Aware Training
  - Rather than quantising a finished model
    - ... QAT quantises while training
  - QAT cuts memory requirements and maximizes on-device performance
  - Gemma 4 Quantization-Aware Training (QAT) checkpoints on Hugging Face



MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 Edge models

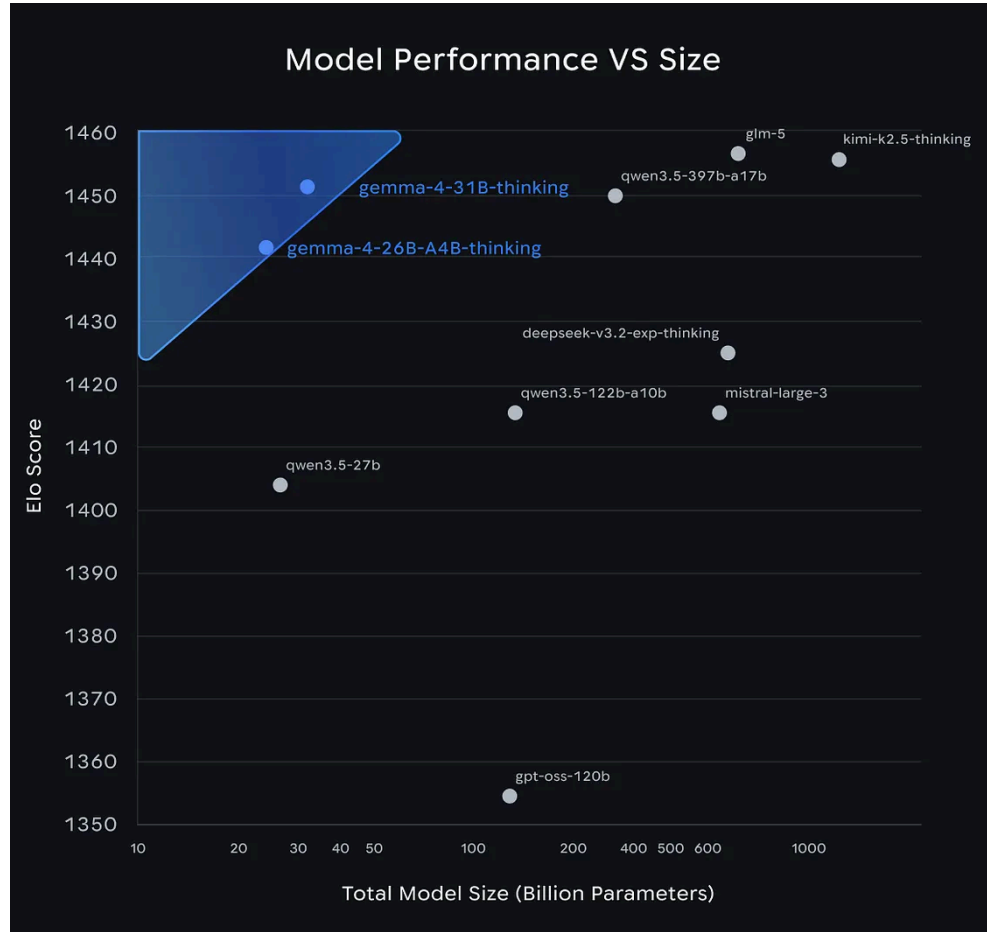
- KV cache compression
  - skip new W_k and W_v for late layers
    - but increase FFN sizes these layers
  - saves on saving these values in cache
- Audio encoder (E2B, E4B)
  - "Universal Speech Model"-style conformer
    - the same base architecture as the one in Gemma-3n

- Per-Layer Embeddings (PLE):
  - a second embedding table that feeds a small residual signal into *every* decoder layer
  - E2B blows nearly half its parameter budget (46%) on flash-based lookup tables
    - that 4.7GB footprint is effectively free
  - shares some characteristics with DeepSeek Engram idea

MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 Larger models

- Multi-Token Prediction (MTP)
  - Unsloth : runs 2x faster with MTP GGUFs
- MoE vs Dense
  - MoE : 26B params, 4B active
    - 128-experts per layer, 8 active
  - Dense : 31B params, 31B active...
- Thinking is quite efficient
  - compared to (for instance) Qwen models



MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 Launch Benchmarks

Benchmark		Gemma 4 31B IT Thinking	Gemma 4 26B A4B IT Thinking	Gemma 4 E4B IT Thinking	Gemma 4 E2B IT Thinking	Gemma 3 27B IT
Arena AI (text) As of 4/2/26		**1452**	1441	—	—	1365
MMMLU Multilingual Q&A	No tools	**85.2%**	82.6%	69.4%	60.0%	67.6%
MMMU Pro Multimodal reasoning		**76.9%**	73.8%	52.6%	44.2%	49.7%
AIME 2026 Mathematics	No tools	**89.2%**	88.3%	42.5%	37.5%	20.8%
LiveCodeBench v6 Competitive coding problems		**80.0%**	77.1%	52.0%	44.0%	29.1%
GPQA Diamond Scientific knowledge	No tools	**84.3%**	82.3%	58.6%	43.4%	42.4%
τ2-bench Agentic tool use	Retail	**86.4%**	85.5%	57.5%	29.4%	6.6%

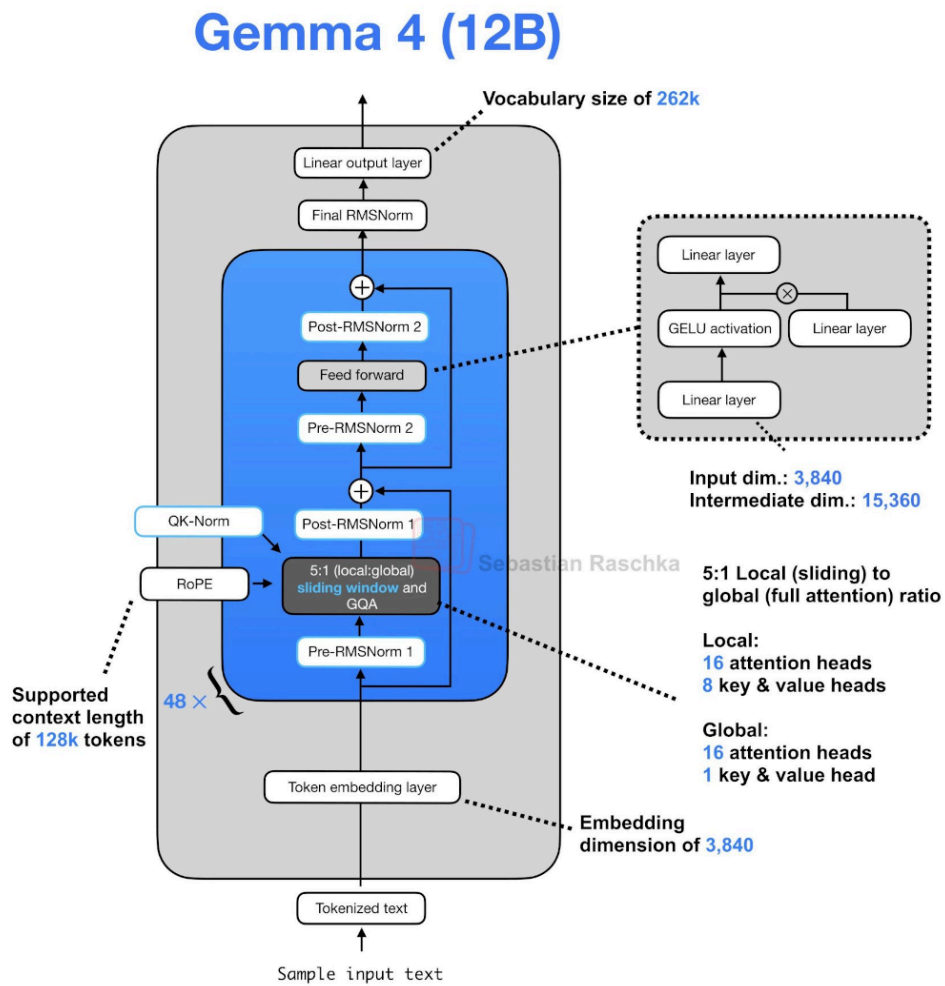
These models were evaluated against a large collection of datasets and metrics to cover different aspects of text generation. See additional benchmarks in model card.

MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 12B

Gemma 4 12B Architecture

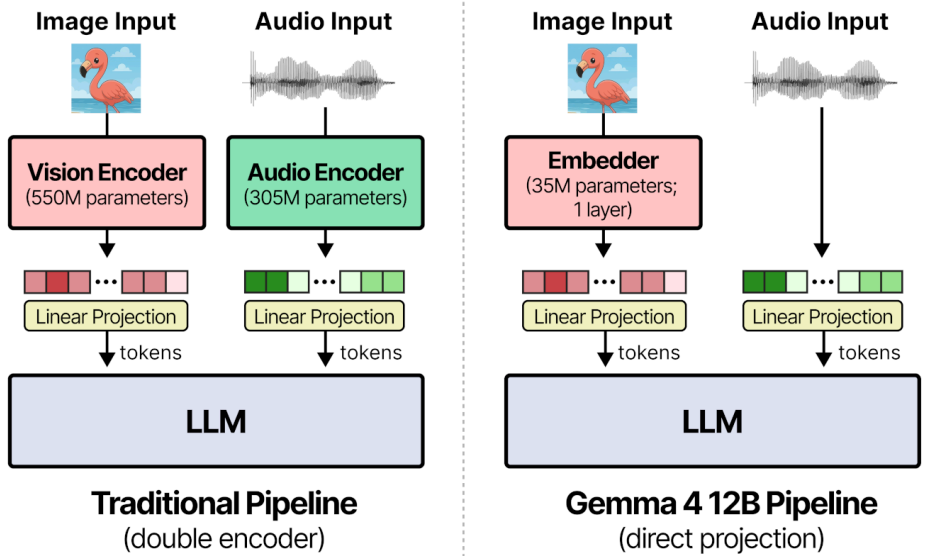
- Apache 2.0 license
- Comes equipped with Multi-Token Prediction (MTP) drafters to reduce latency
- Delivers performance nearing our larger 26B MoE model on standard benchmarks
  - but at less than half the total memory footprint



MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 12B Multimodal

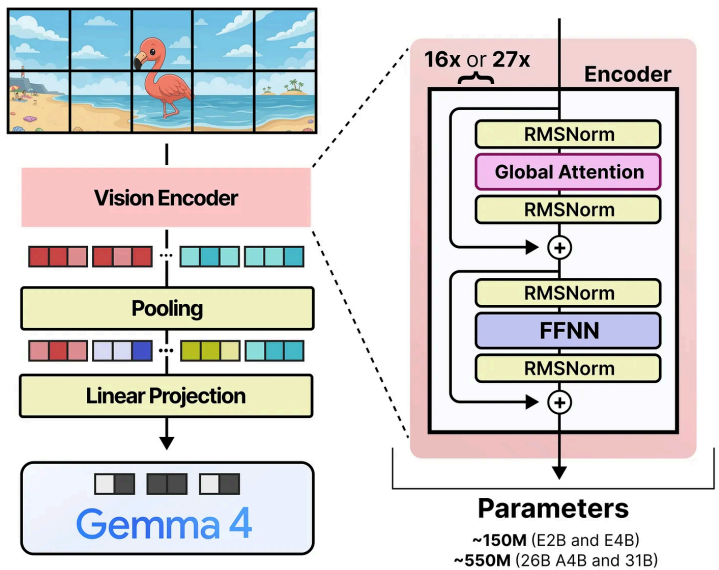
- Gemma 4 12B radically simplifies Multimodal input
  - need for pipeline of models for vision / audio removed



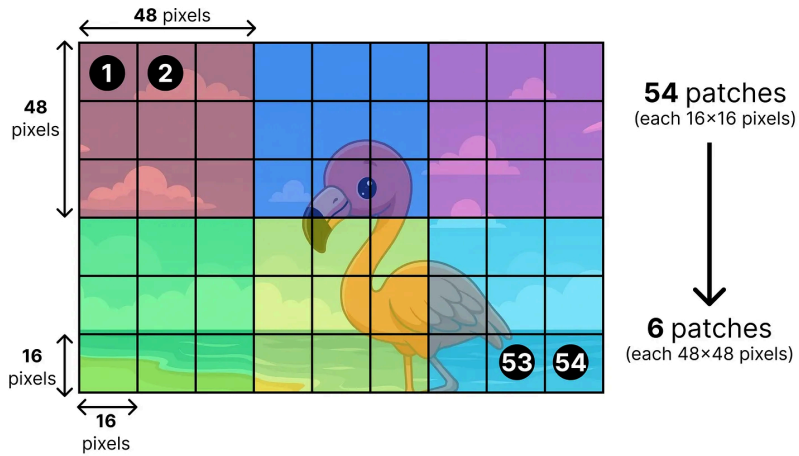
MLSG
MACHINE LEARNING
SINGAPORE

How the New Vision Input Works

- Gemma 4 Previous models



- Gemma 4 12B

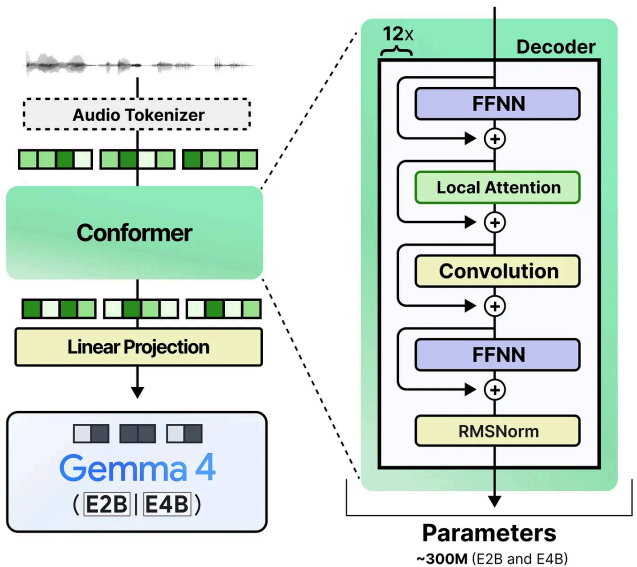


- 35M param vision embedder
  - 48x48x3 image patches projected to LLM hidden dim
    - with a single matmul
    - and add learned X/Y positional embedding to each patch
  - savings : 550M ⟶ 35M parameters
    - = a 15x reduction

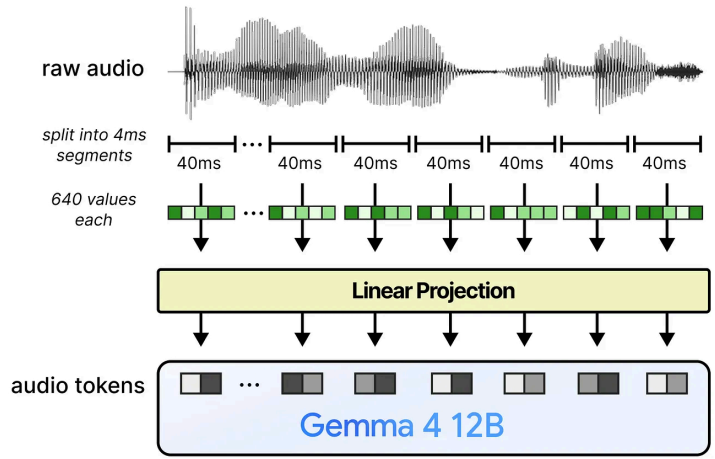
MLSG
MACHINE LEARNING
SINGAPORE

How the New Audio Input Works

- Gemma 4 Previous models

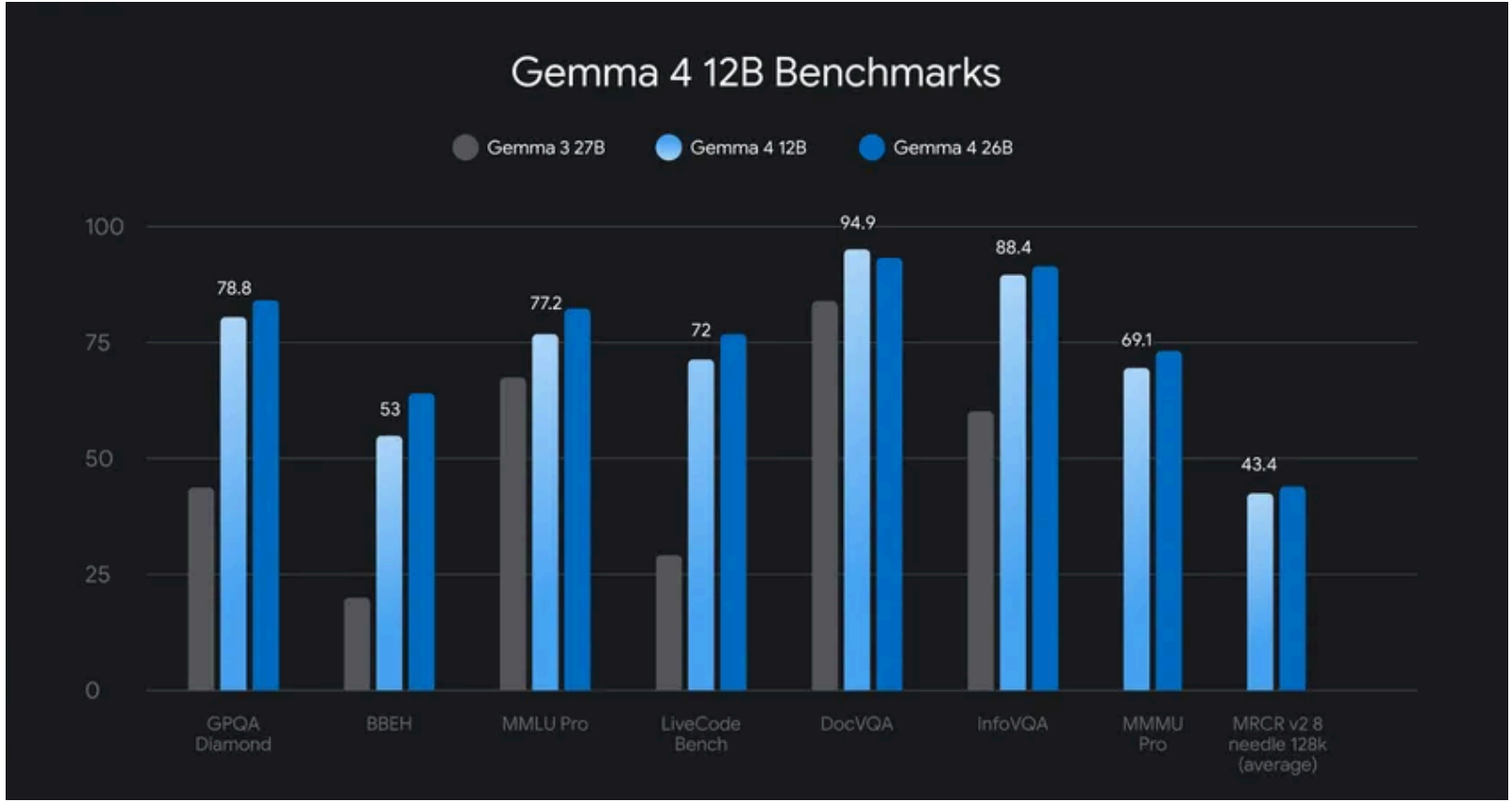


- Gemma 4 12B



- No audio encoder!
  - Raw 16 kHz audio signals are sliced into 40ms frames
    - (640 floats each)

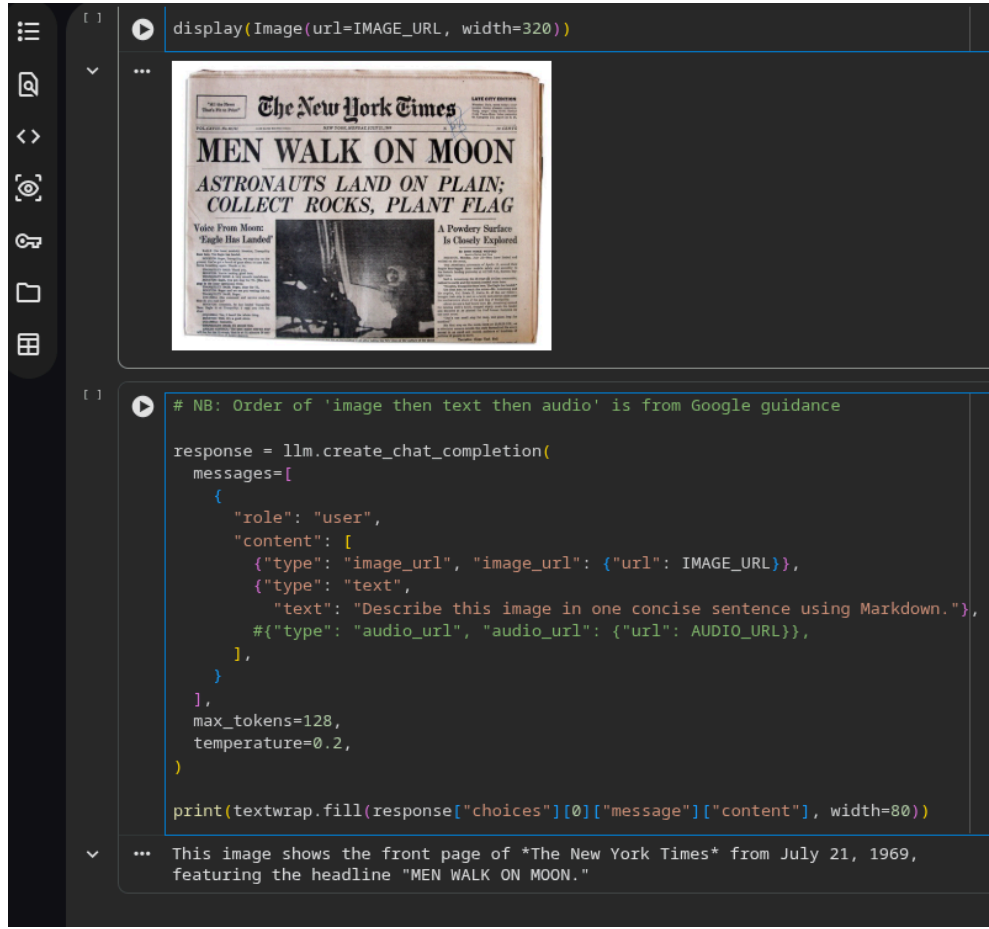
MLSG
MACHINE LEARNING
SINGAPORE



MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 12B Demo

- Guidance from Google : Gemma 4 was trained with a specific convention to interleave input modalities:
  - Image content goes before the text in your prompt
  - Audio content goes after the text in your prompt
- Gemma 4 12B QAT - run this Colab for Free!
- Can also use `llama.cpp` directly :
  - `llama serve -hf ggml-org/gemma-4-12B-it-GGUF`
    - llama.cpp just added video input support
- Can also run on 8Gb VRAM GPU!
  - See: My Blog Post



MLSG
MACHINE LEARNING
SINGAPORE

Gemma 4 Diffusion

MLSG
MACHINE LEARNING
SINGAPORE

Why do Diffusion?

- Google Blog: DiffusionGemma
- Key message : 4x faster text generation
  - parallel denoising generates roughly 15–20 tokens per forward pass
  - per-user generation speeds exceeding 1100 tokens/second at low batch sizes
    - assuming : H100, FP8
- Bi-directional attention
  - generates 256 tokens in parallel with each forward pass
    - every token to attends to all others
  - significant advantages for non-linear domains such as
    - in-line editing, code infilling, amino acid sequences or mathematical graphs
- Also: Adaptive Computation (sexy topic)
  - inference compute can be adaptive too!
    - simpler prompts and structured tasks like code need fewer denoising steps
    - so tokens-per-second scales with task complexity

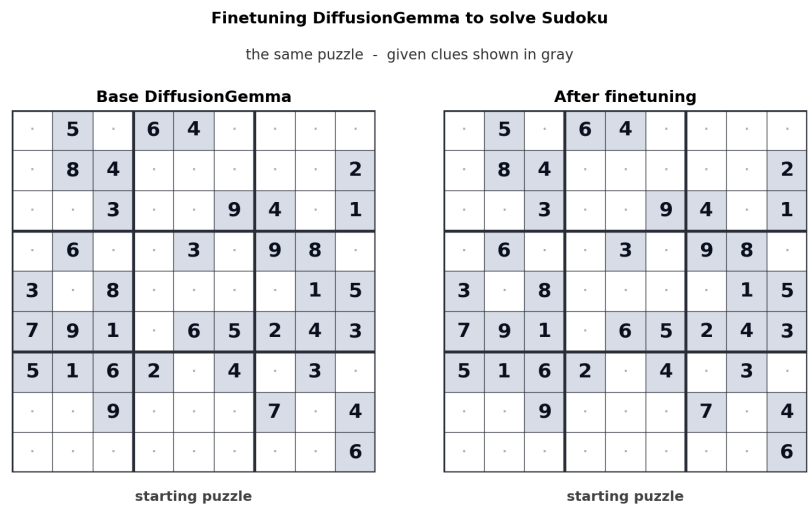
How diffusion works

- Model applies bidirectional attention over a "canvas" of 256 tokens
  - Trained to denoise tokens *in parallel*
  - Takes multiple steps to resolve ambiguities - but overall less 'computation'

Gemma Diffusion Details

- Gemma Diffusion is built on Gemma 4 26B A4B MoE foundation
  - "fits comfortably within 18GB VRAM limits of high-end dedicated consumer GPUs when quantized"
  - supports up to 256K context
- Modalities
  - takes text and image inputs
  - generates text
- Apache 2.0 license

- Gemma Diffusion is first *mainstream* diffusion model
  - Unsloth (for instance) supports fine-tuning, etc
- You can train it yourself!

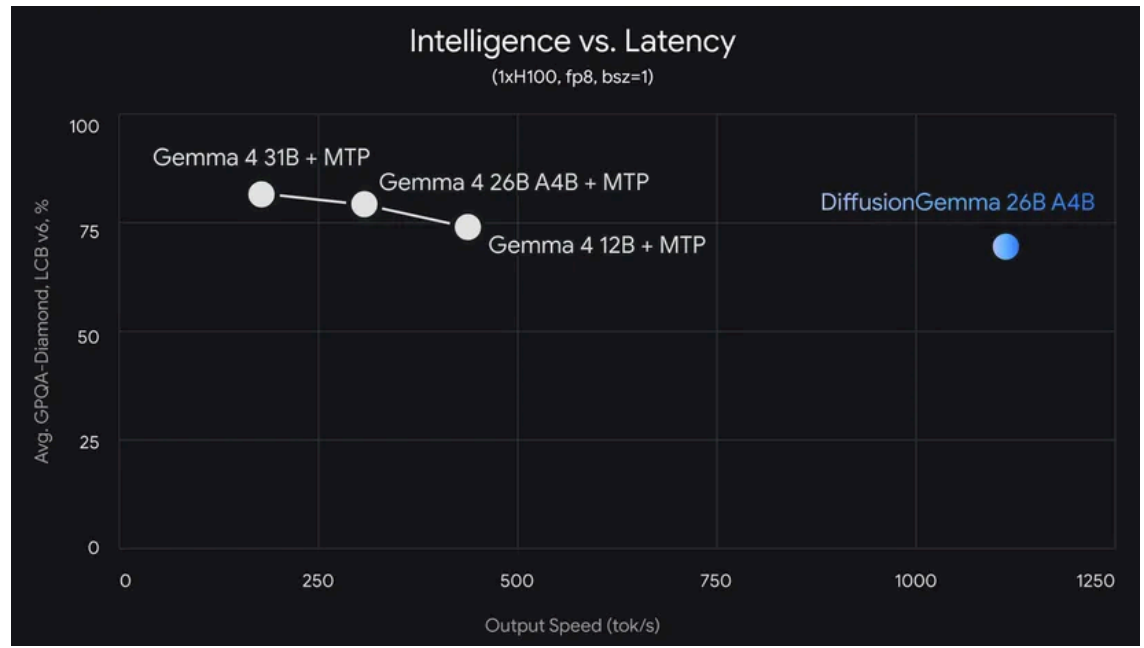


MLSG
MACHINE LEARNING
SINGAPORE

Diffusion Speed Chart

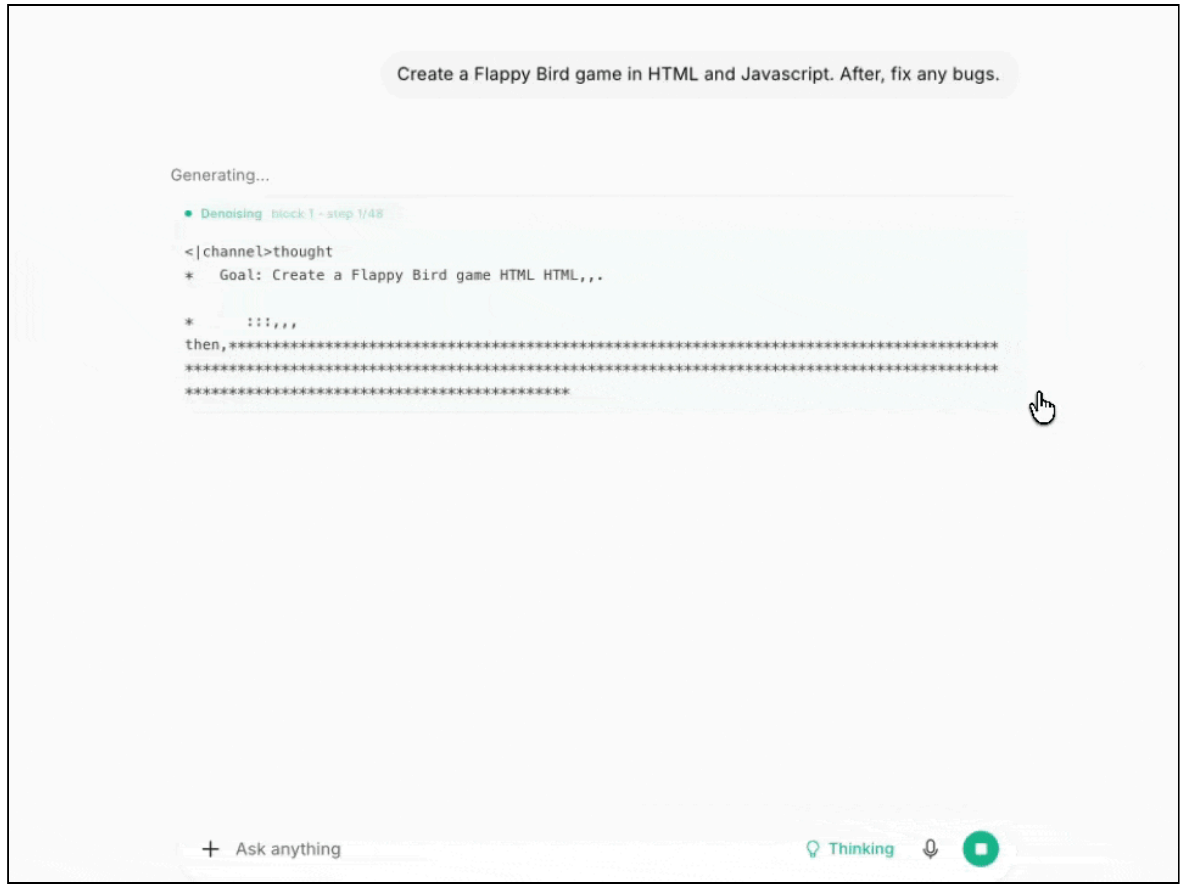
- DiffusionGemma trails the autoregressive 26B A4B on most tasks
  - for example:
    - `MMLU Pro....` : 77.6% vs 82.6%
    - `AIME 2026 ...` : 69.1% vs 88.3%
    - `GPQA Diamond` : 73.2% vs 82.3%
  - edging ahead on a few
    - `HLE no tools` : 11.0% vs 8.7%

- But focus on the Speed!



MLSG
MACHINE LEARNING
SINGAPORE

Gemma Diffusion Demo ...

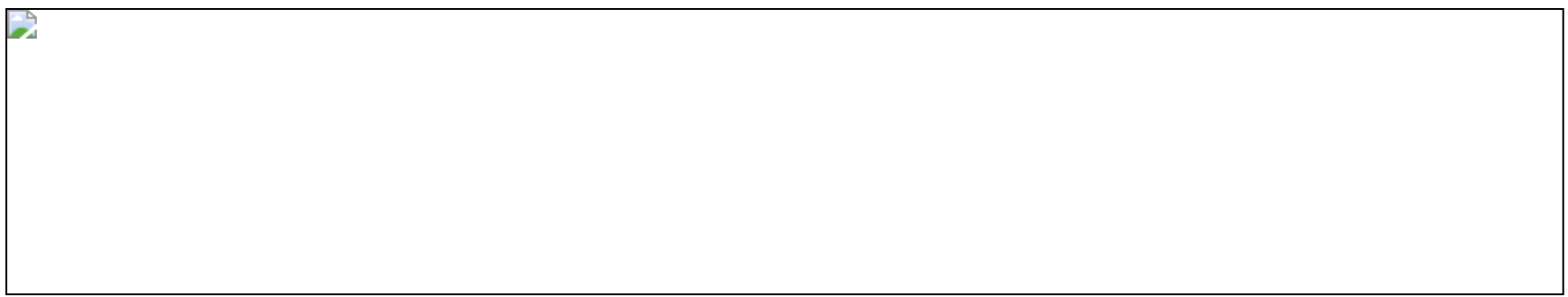


MLSG
MACHINE LEARNING
SINGAPORE

Diffusion Demo ...

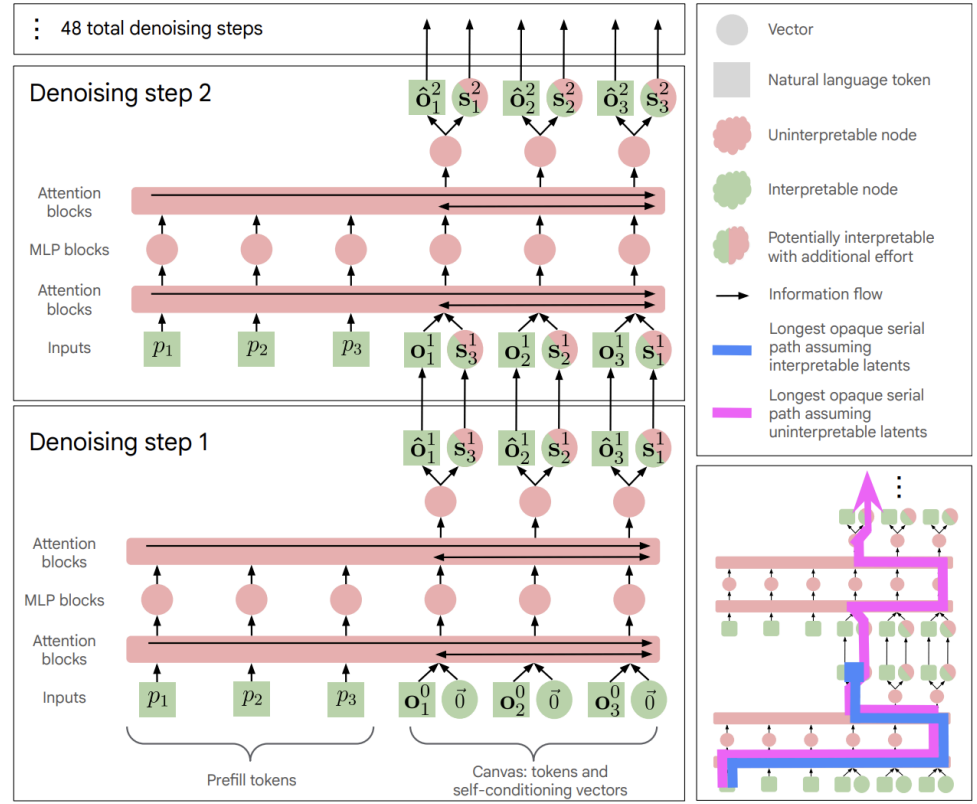
- Apparently it is *possible* to run Gemma Diffusion on smaller GPUs :
  - "diffusiongemma-26B-A4B-it-Q4_K_M doesn't fit on my RTX 3060 12GB out of the box, but ..."
    - using the new WIP `llama.cpp` diffusion branch
    - running with `-ngl 15`, multi-block generation shows signs of working...

- But you can also train and run a diffusion model!
  - Have a look in `https://github.com/mdda/tiny-diffusion-jax`



Diffusion Model Interpretability

- How Transparent is DiffusionGemma? - Engels *et al* (2026)
  - Google DeepMind paper!
  - Sadly, no code repo...
- Twitter/X thread by first Author
  - "We find that the intermediates are interpretable"
    - bonus animation available in thread
  - "This recovers many of the benefits of CoT!"



MLSG MACHINE LEARNING SINGAPORE

Wrap-Up

- Google continues to extend Gemma 4 series
  - a huge win for everyone
- Gemini benefits from Gemma 'experiments'
  - this is a good synergy
    - a huge win for Google
- Looking forwards to what comes next!

MLSG
MACHINE LEARNING
SINGAPORE

Machine Learning SG MeetUp Group

- Next Meeting = TBD @ Google
- Topic : "TBD"
- Typical Contents :
  - Talk for people starting out
  - Something from the bleeding-edge
  - Lightning Talks
- MeetUp.com / Machine-Learning-Singapore

MLSG
MACHINE LEARNING
SINGAPORE

Link to Slides

- https://bit.ly/2026-06_IO

MLSG
MACHINE LEARNING
SINGAPORE

See You Next Time !

Please add yourself to the MLSG Calendar on Luma!

END