Gemma3 Technical Report

A Lightweight Language Model with multimodal understanding and multilingual capabilities

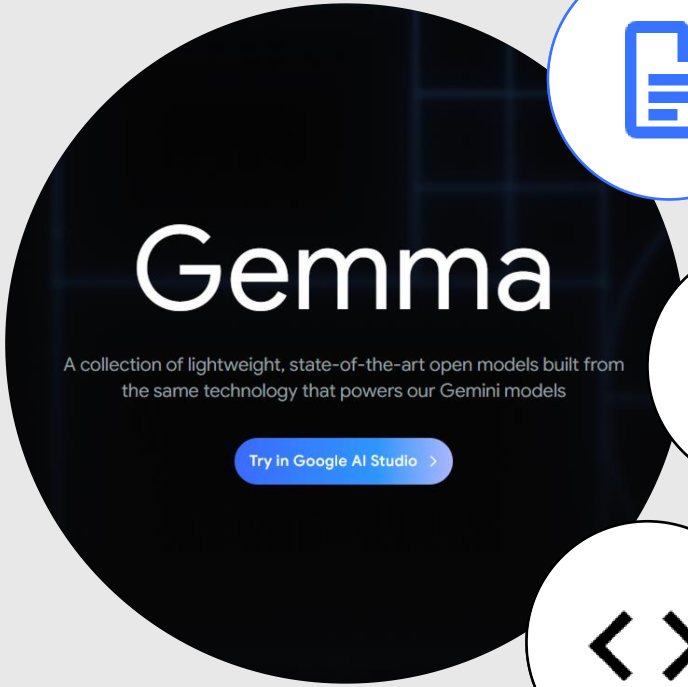
Seminar: Foundation Model Frontiers (SS 2025)
Instructor: Prof. Hinrich Schütze

Date: 13.06.2025
Presenter: Hyunji Kim

TABLE OF CONTENT

INTRODUCTION

Gemma Model Family is a set of lightweight models; Open source for developers and researchers



Gemma 1, 2 & 3

Transformer based LLM for text generation tasks + *multimodality (*only from Gemma 3)

PaliGemma

Vision-language model (VLM) for image processing, object detection, reading text embeded within images

CodeGemma

Fine-tuned version of Gemma for code completion and generation

INTRODUCTION

What's new in Gemma 3 ?

	Gemma 2	Gemma 3
Sizes	• 2B • 9B • 27B	• 1B • 4B • 12B • 27B
Context Window Length	8k	• 32k (1B) • **128k (4B, 12B, 27B)**
Multimodality	X	X (1B) ✓ **(4B, 12B, 27B)**
Multilingual Support		English (1B) + **140 languages (4B, 12B, 27B)**

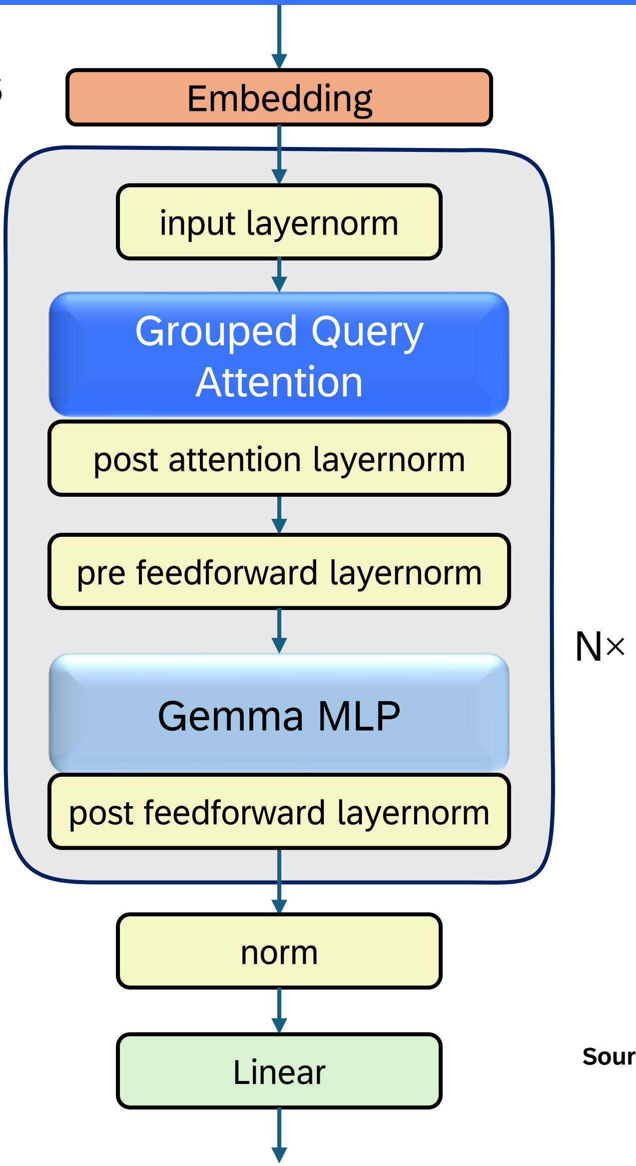
MODEL ARCHITECTURE

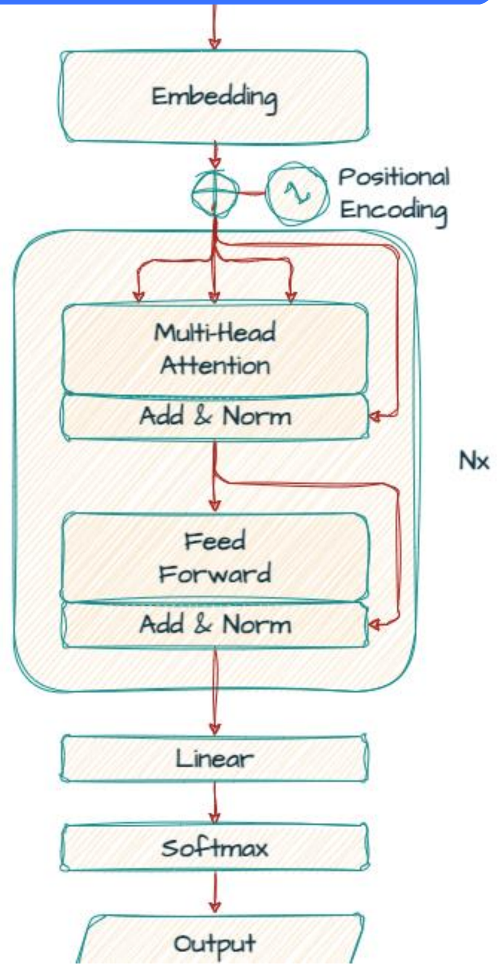
Architecture elements similar to previous versions

- **Decoder-only** transformer model
- Grouped-Query Attention (GQA) with post-norm and pre-norm with RMSNorm
- RMSNorm (Root Mean Square Layer Normalization) $\Rightarrow$ more stable and efficient training

$$\bar{a}_i = \frac{a_i}{\mathrm{RMS}(\mathbf{a})} g_i, \quad \text{where } \mathrm{RMS}(\mathbf{a}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} a_i^2}.$$

Source: Root Mean Square Layer Normalization (Zhang & Sennricht, 2019)





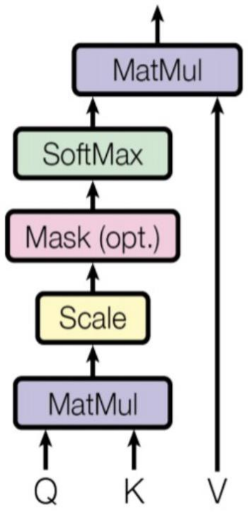
General Decoder Model

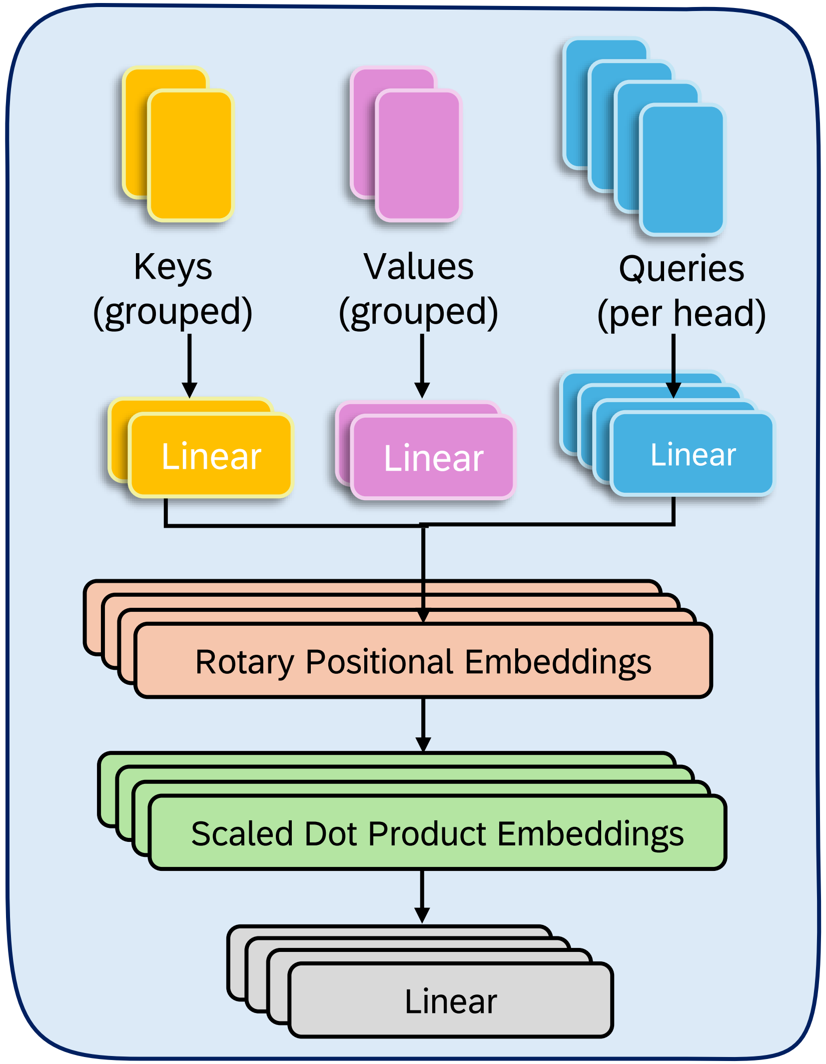
Source: Gemma explained: What's new in Gemma 2

MODEL ARCHITECTURE

Architecture elements similar to previous versions

- Grouped-Query Attention (GQA) offers good trade-off between inference speed and performance accuracy.
- The query heads (Q) are divided into **G** groups. Each shares a single key & value head
- Each group key & value head
  ⇒ mean-pooling all original heads within that group
- Quality close to Multi-Head (MHA)
  at comparable speed to Multi-Query attention (MQA)





MODEL ARCHITECTURE

Key changes for new capabilities and enhancements: multimodality, longer context, memory efficiency

Multimodality

- Custom SigLIP (Sigmoid Loss for Language-Image Pre-training) vision encoder is frozen during training
- Pan & Scan method for non-square aspect ratios or higher resolutions

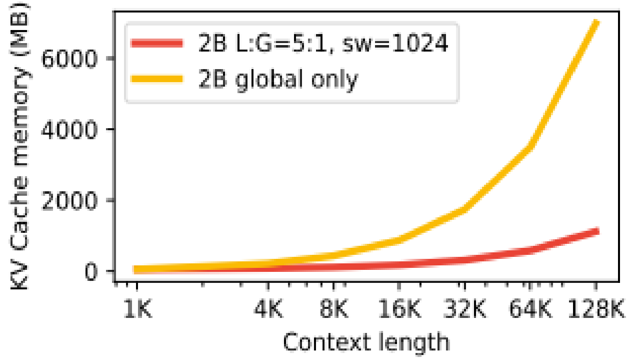
Less KV-cache memory

- 5-to-1 interleaved attention
- Smaller sliding window size sw=1024 (down from sw=4096)

Gemma 1	Gemma 2	Gemma 3
Global	Global	Global
Global	Local	Local
Global	Global	Local
Global	Local	Local
Global	Global	Local
Global	Local	Local

Longer Context

- Scaled context length to 128k



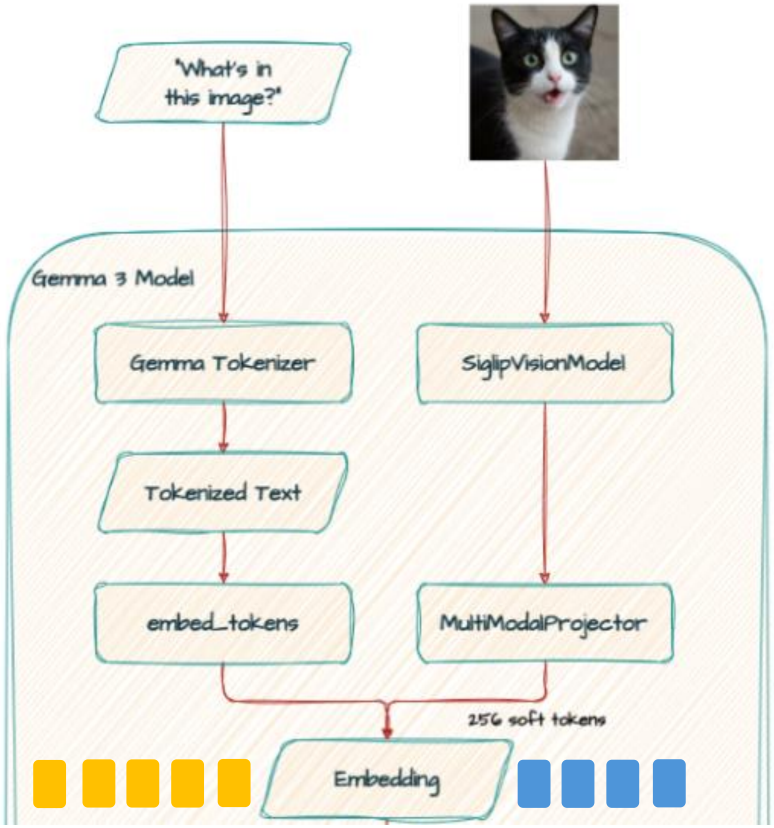
KV cache memory VS context length
Source: Gemma 3 Technical Report (Gemma Team, 2025)

MODEL ARCHITECTURE

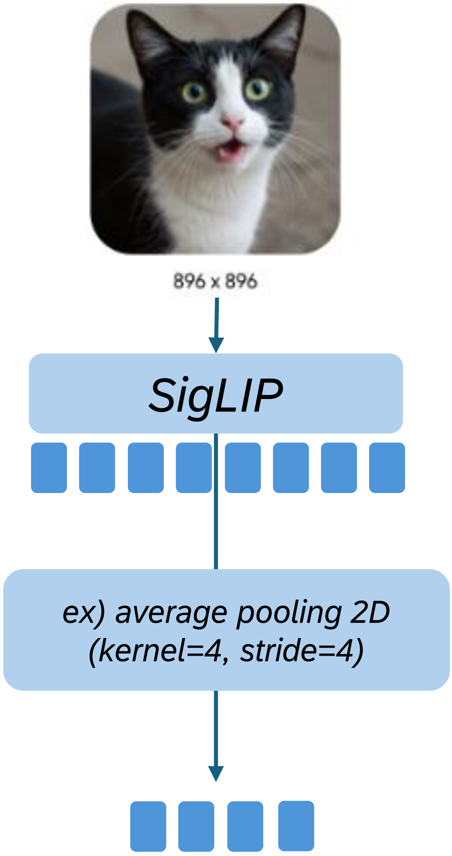
SentencePiece tokenizer for text input & Vision encoder for image input

- SentencePiece – same as previous Gemma models and Gemini 2.0
- 262k vocabulary size (previously 256k)
- Improved encoding for languages like Chinese, Japanese, and Korean

- SigLIP Vision Transformer Layers
- Fixed input resolution 896 × 896 pixels
- ViT normally uses 16 × 16 pixel patches (896/16 = 56) => grid of 56 × 56 patches
- Average pooling to reduce output to 256 tokens (16 × 16)



Source: Gemma explained: What's new in Gemma 3 (J. Ji & R. Kumar, 2025)

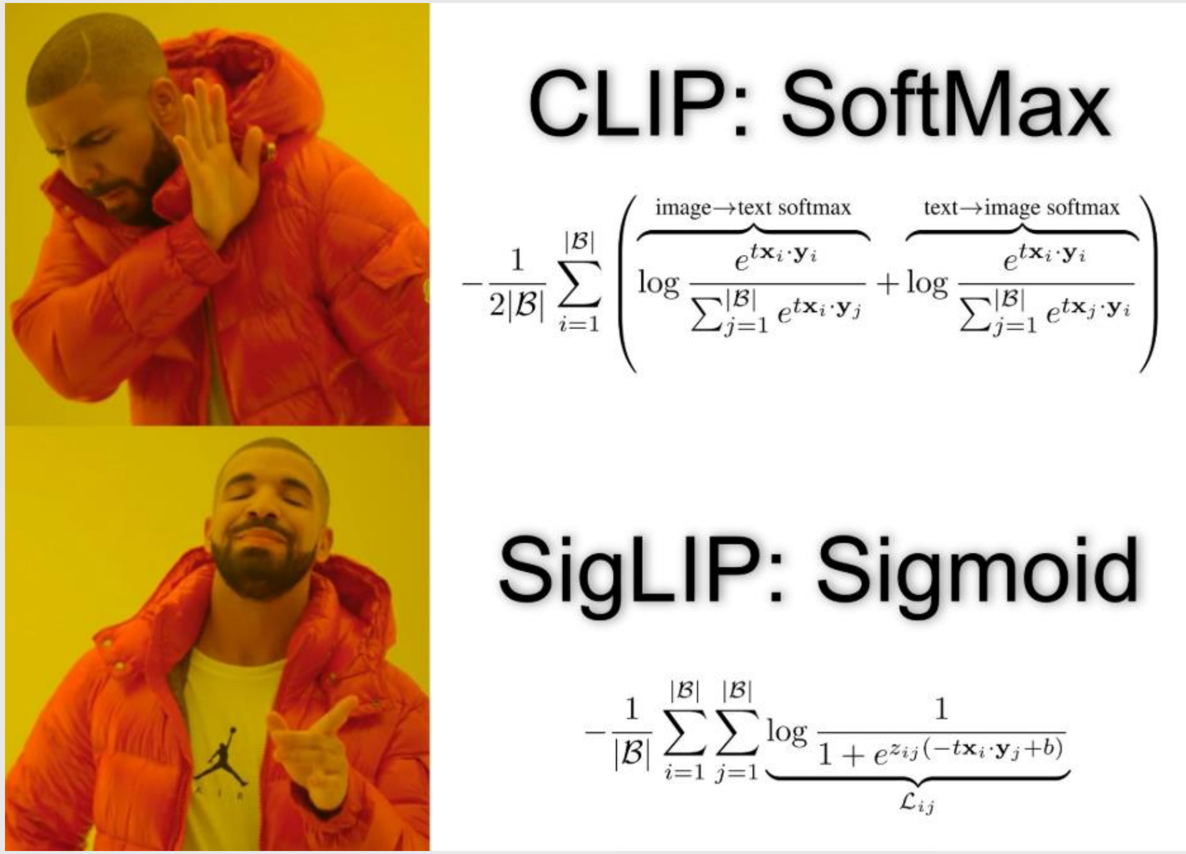


SigLIP (Sigmoid Loss for Language Image Pre-training)

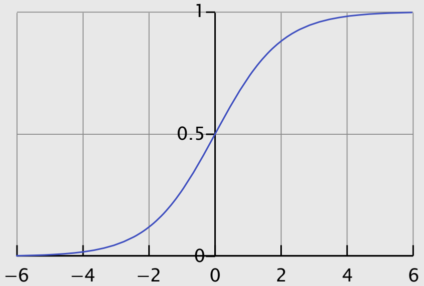
- CLIP (Contrastive Language-Image Pre-training) : positive image-text pairs & negative image-text pairs
- CLIP with large batch size $\Rightarrow$ $O(n^2)$ memory complexity ↑

$$\mathcal{B} = \{(I_1, T_1), (I_2, T_2), \dots\}$$

- **SigLIP** operates on image-text pair **independently**
- SigLIP allows scaling up the batch size & better generalization
- Performs better at smaller batch sizes and in zero-shot image classification.

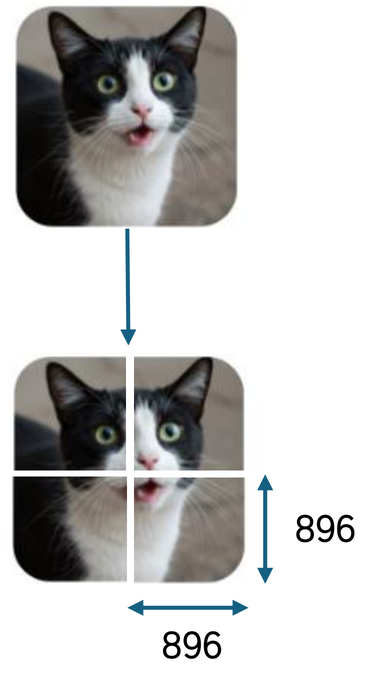


Source: https://twitter.com/giffmana/status/1692641733459267713



MULTI-MODALITY

Pan & Scan method



- Crop the image → Resize each crop to 896 × 896 → Encode with SigLIP
- Flexible to process non-standard ratios or high-resolution images
- Segments images into non-overlapping crops of equal size
- Improves performance when detailed information is critical (e.g. reading text on images)

	DocVQA	InfoVQA	TextVQA
4B	72.8	44.1	58.9
4B w/ P&S	81.0	57.0	60.8
Δ	(+8.2)	(+12.9)	(+1.9)
27B	85.6	59.4	68.6
27B w/ P&S	90.4	76.4	70.2
Δ	(+4.8)	(+17.0)	(+1.6)

Impact of P&S
Source: Gemma 3 Technical Report (Gemma Team, 2025)

Resolution	DocVQA	InfoVQA	TextVQA
256	31.9	23.1	44.1
448	45.4	31.6	53.5
896	59.8	33.7	58.0

Impact of image encoder input resolution
Source: Gemma 3 Technical Report (Gemma Team, 2025)

ABLATIONS

Local:Global=5:1 attention layers & Sliding window size (sw=1024)

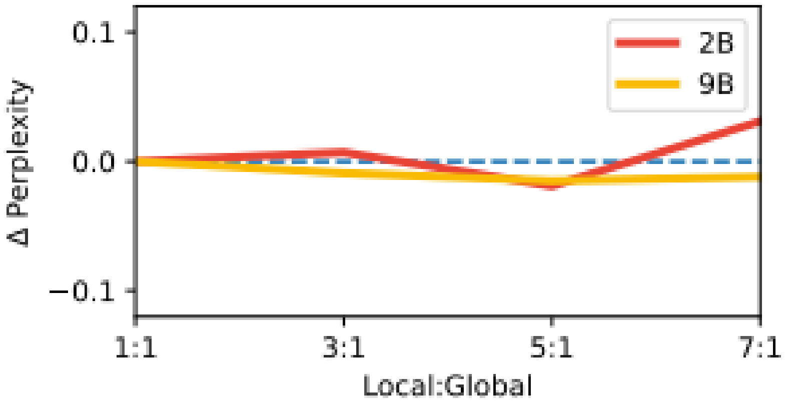


Figure 3 | **Impact of Local:Global ratio** on the perplexity on a validation set. The impact is minimal, even with 7-to-1 local to global. This ablation is run with text-only models.

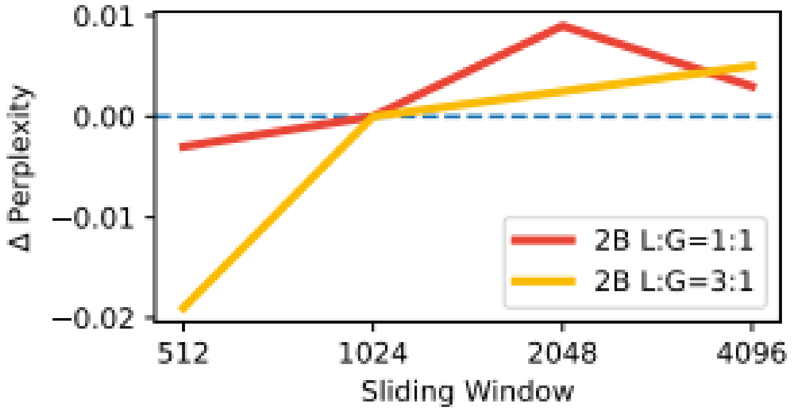


Figure 4 | **Impact of Sliding Window** size on perplexity measured on a validation set. We consider 2 2B models, with 1:1 and 1:3 local to global layer ratios. This ablation is run with text-only models.

Source: Gemma explained: What's new in Gemma 3 (J. Ji & R. Kumar, 2025)

ABLATIONS

Local:Global=5:1 attention layers & Sliding window size (sw=1024)

Source: Gemma explained: What's new in Gemma 3 (J. Ji & R. Kumar, 2025)

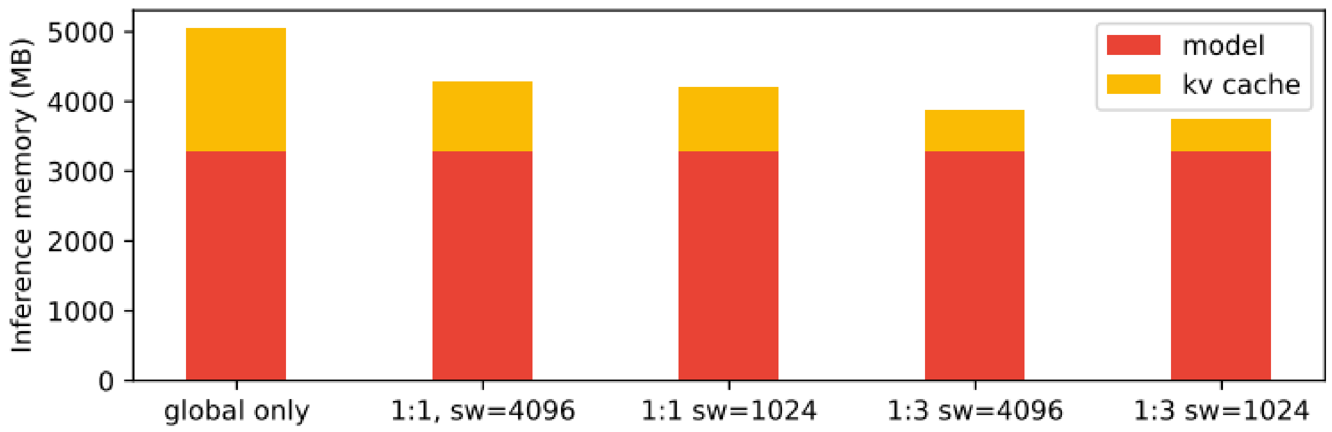


Figure 5 | **Model versus KV cache memory** during inference with a pre-fill KV cache of size 32k. We consider a 2B model with different local to global ratios and sliding window sizes (sw). We compare to global only, which is the standard used in Gemma 1 and Llama. This ablation is run with a text-only model.

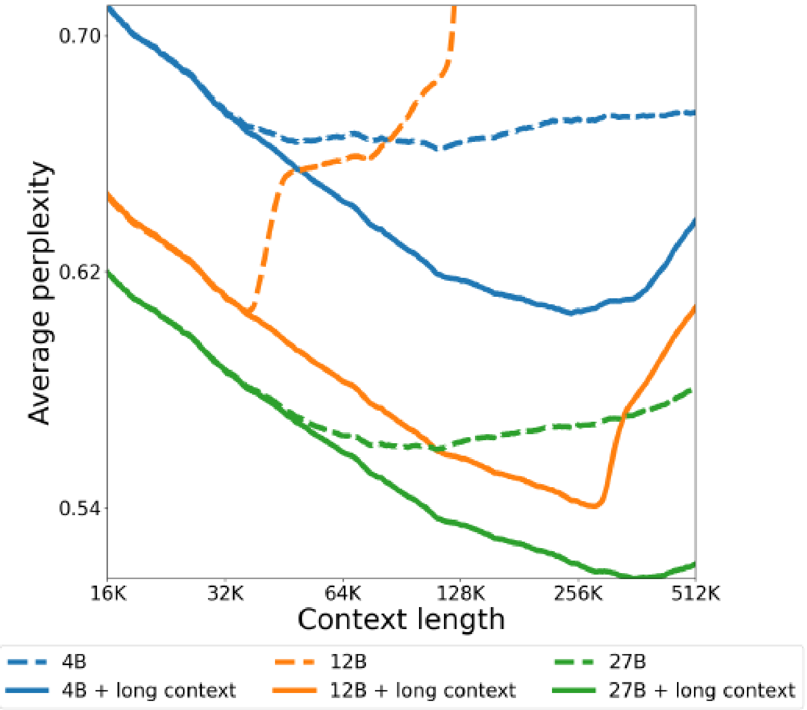


Figure 7 | **Long context** performance of pre-trained models before and after RoPE rescaling.

TRAINING STRATEGIES

Knowledge Distillation

- Pre-training optimization recipe with knowledge distillation
- Sampling 256 logits per token, which is a subset of teacher's output probabilities

Training Data

- Multilingual data to improve language coverage (both monolingual and parallel data)
- Slightly larger token budget than Gemma 2; 27B trained on 14T tokens

Instruction Tuning

- Post-training approach with knowledge distillation from a large IT teacher
- RL objectives to improve helpfulness, math, coding, reasoning, multilingualism

Data Filtering

- Filtering out private, unsafe or toxic content in post-training dataset
- Aligned with Google's safety policies

PERFORMANCE EVALUATION

LMSYS Chatbot Arena

- Outperformed much larger open models, such as DeepSeek-V3 and LLaMA 3 405B
- Elo scores are significantly higher than Gemma 2

  *Elo rating system: widely-used in chess, rating updates based on performance on a particular tasks (Text, Vision, Code, etc.)

- New LMArena Leaderboard ([Overview Leaderboard | LMArena])

Model 206 / 206	Overall	Hard Prompts	Coding
gemini-2.0-flash-001	21	23	27
gemma-3-27b-it	21	27	30
deepseek-v3	27	29	26
qwen3-235b-a22b	27	22	14

Rank	Model	Elo	95% CI	Open	Type	#params/#activated
1	Grok-3-Preview-02-24	1412	+8/-10	-	-	-
1	GPT-4.5-Preview	1411	+11/-11	-	-	-
3	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-5	-	-	-
3	Gemini-2.0-Pro-Exp-02-05	1380	+5/-6	-	-	-
3	ChatGPT-4o-latest (2025-01-29)	1377	+5/-4	-	-	-
6	DeepSeek-R1	1363	+8/-6	yes	MoE	671B/37B
6	Gemini-2.0-Flash-001	1357	+6/-5	-	-	-
8	o1-2024-12-17	1352	+4/-6	-	-	-
9	**Gemma-3-27B-IT**	**1338**	**+8/-9**	**yes**	**Dense**	**27B**
9	Qwen2.5-Max	1336	+7/-5	-	-	-
9	o1-preview	1335	+4/-3	-	-	-
9	o3-mini-high	1329	+8/-6	-	-	-
13	DeepSeek-V3	1318	+8/-6	yes	MoE	671B/37B
14	GLM-4-Plus-0111	1311	+8/-8	-	-	-
14	Qwen-Plus-0125	1310	+7/-5	-	-	-
14	Claude 3.7 Sonnet	1309	+9/-11	-	-	-
14	Gemini-2.0-Flash-Lite	1308	+5/-5	-	-	-
18	Step-2-16K-Exp	1305	+7/-6	-	-	-
18	o3-mini	1304	+5/-4	-	-	-
18	o1-mini	1304	+4/-3	-	-	-
18	Gemini-1.5-Pro-002	1302	+3/-3	-	-	-
...						
28	Meta-Llama-3.1-405B-Instruct-bf16	1269	+4/-3	yes	Dense	405B
...						
38	Llama-3.3-70B-Instruct	1257	+5/-3	yes	Dense	70B
...						
39	Qwen2.5-72B-Instruct	1257	+3/-3	yes	Dense	72B
...						
59	Gemma-2-27B-it	1220	+3/-2	yes	Dense	27B

Source: Gemma 3 Technical Report (Gemma Team, 2025)

PERFORMANCE EVALUATION

Performance comparison of instruction fine-tuned (IT) models

- Zero-shot benchmarks across different abilities
- Consistent improvement over STEM (science, technology, engineering, and mathematics) related tasks and Code

	Gemini 1.5		Gemini 2.0		Gemma 2			Gemma 3			
	Flash	Pro	Flash	Pro	2B	9B	27B	1B	4B	12B	27B
MMLU-Pro	67.3	75.8	77.6	79.1	15.6	46.8	56.9	14.7	43.6	60.6	67.5
LiveCodeBench	30.7	34.2	34.5	36.0	1.2	10.8	20.4	1.9	12.6	24.6	29.7
Bird-SQL (dev)	45.6	54.4	58.7	59.3	12.2	33.8	46.7	6.4	36.3	47.9	54.4
GPQA Diamond	51.0	59.1	60.1	64.7	24.7	28.8	34.3	19.2	30.8	40.9	42.4
SimpleQA	8.6	24.9	29.9	44.3	2.8	5.3	9.2	2.2	4.0	6.3	10.0
FACTS Grounding	82.9	80.0	84.6	82.8	43.8	62.0	62.4	36.4	70.1	75.8	74.9
Global MMLU-Lite	73.7	80.8	83.4	86.5	41.9	64.8	68.6	34.2	54.5	69.5	75.1
MATH	77.9	86.5	90.9	91.8	27.2	49.4	55.6	48.0	75.6	83.8	89.0
HiddenMath	47.2	52.0	63.5	65.2	1.8	10.4	14.8	15.8	43.0	54.5	60.3
MMMU (val)	62.3	65.9	71.7	72.7	-	-	-	-	48.8	59.6	64.9

Source: Gemma 3 Technical Report (Gemma Team, 2025)

CONCLUSION

Key Improvements Gemma 3 Multimodal Language Model

Vision understanding	Treating images as sequences of 256 soft tokens encoded by a tailored SigLIP vision encoder, reducing inference cost of image processing
KV-cache Memory Reduction	Increasing ratio of local to global attention layers to 5:1 to reduce KV-cache memory explosion with long contexts
Longer Context Length	Longer context length of 128k tokens without reducing performance by increasing RoPE base frequency on global self-attention layers
Multilingual Support	Improved multilingual capabilities through increased pre-training data mixture, along with novel post-training recipe

DISCUSSION

Gemma 3 outperformed much larger language models. Different paradigms can be employed, including model distillation, while maintaining/improving its performance.

- Does LLM size matter?
- Trade-off between computational resources and model performance. Which matters more?

REFERENCES

1. Gemma 3 Technical Report (Gemma Team, 2025)
2. Gemma explained: What's new in Gemma 3 (J. Ji & R. Kumar, 2025)
3. Gemma explained: An overview of Gemma model family architectures (J. Ji & R. Kumar, 2024)
4. Root Mean Square Layer Normalization (Zhang & Sennrich, 2019)
5. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints (Ainslie et al., 2023)
6. Sigmoid Loss for Language Image Pre-Training (Zahi et al., 2023)
7. Sigmoid Loss for Language Image Pre-Training | by Ahmed Taha | Medium
8. Welcome Gemma 3: Google's all new multimodal, multilingual, long context open LLM (Gosthipaty et al., 2025)

Thank you!

Questions?