	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

Table 6 | Performance of instruction fine-tuned (IT) models compared to Gemini 1.5, Gemini 2.0, and Gemma 2 on zero-shot benchmarks across different abilities.

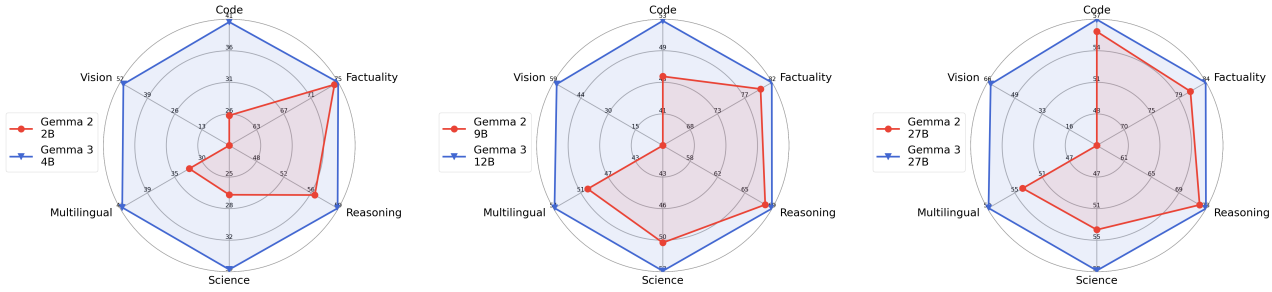

Figure 2 | Summary of the performance of different pre-trained models from Gemma 2 and 3 across general abilities. These plots are meant to give a simplified summary and details are in the appendix.

code, factuality, multilinguality, reasoning, and vision. The details of the performance across the different public benchmarks used in these plots are summarized in the appendix. Overall, we see that the new versions improve in most categories, despite the addition of vision. We particularly focus on multilinguality in this version, and this directly impacts the quality of our models. However, despite the use of decontamination techniques, there is always a risk of contamination of these probes (Mirzadeh et al., 2024), making more definitive conclusions harder to assess.

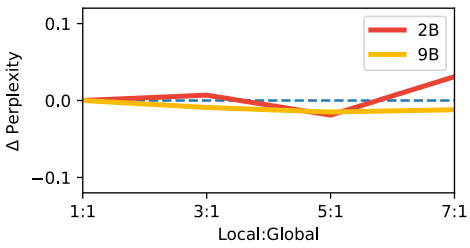

Figure 3 | **Impact of Local:Global ratio** on the perplexity on a validation set. The impact is minimal, even with 7-to-1 local to global. This ablation is run with text-only models.

5.2. Local:Global attention layers

We measure the impact of changes to local and global self-attention layers on performance and memory consumption during inference.

Local:Global ratio. In Fig. 3, we compare different ratios of local to global attention layers. 1:1 is used in Gemma 2 models, and 5:1 is used in Gemma 3. We observe minimal impact on perplexity when changing this ratio.

Sliding window size. In Fig. 4, we compare different sliding window sizes for the local at-

tention layers in different global:local ratio configurations. The sliding window can be reduced significantly without impacting perplexity.

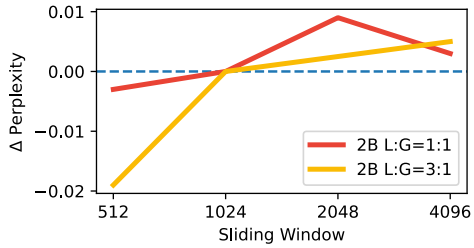

Figure 4 | **Impact of Sliding Window** size on perplexity measured on a validation set. We consider 2 2B models, with 1:1 and 1:3 local to global layer ratios. This ablation is run with text-only models.

Impact on KV cache memory. In Fig. 5, we show the balance between the memory used by the model and the KV cache during inference with a context of 32k tokens. The "global only" configuration is the standard configuration used across most dense models. The "1:1, sw=4096" is used in Gemma 2. We observe that the "global only" configuration results in a memory overhead of 60%, while this is reduced to less than 15% with 1:3 and sliding windows of 1024 ("sw=1024"). In Fig. 6, we compute the memory used by the KV cache as a function of the context length with either our 2B architecture (L:G=5:1, sw=1024) versus a "global only" 2B model.

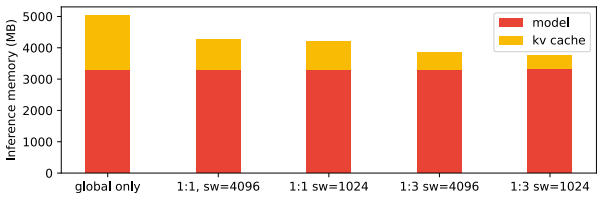

Figure 5 | **Model versus KV cache memory** during inference with a pre-fill KV cache of size 32k. We consider a 2B model with different local to global ratios and sliding window sizes (sw). We compare to global only, which is the standard used in Gemma 1 and Llama. This ablation is run with a text-only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

5.3. Enabling long context

Instead of training with 128K sequences from scratch, we pre-train our models with 32K sequences and then scale the 4B, 12B, and 27B models up to 128K tokens at the end of pre-training while rescaling RoPE (Chen et al., 2023). We find a scaling factor of 8 to work well in practice. Note that compared to Gemma 2, we have also increased the RoPE base frequency of global self-attention layers from 10k to 1M, while keeping 10k for the local self-attention layers. In Figure 7, we show the impact on perplexity for different context lengths. Our models generalize to 128K, but rapidly degrade as we continue to scale.

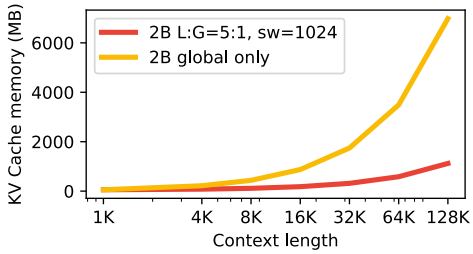

Figure 6 | **KV cache memory versus context length.** We show the memory usage of the KV cache for our architecture (L:G=5:1, sw=1024) and a transformer with global attention only – as used in LLaMa or Gemma 1.

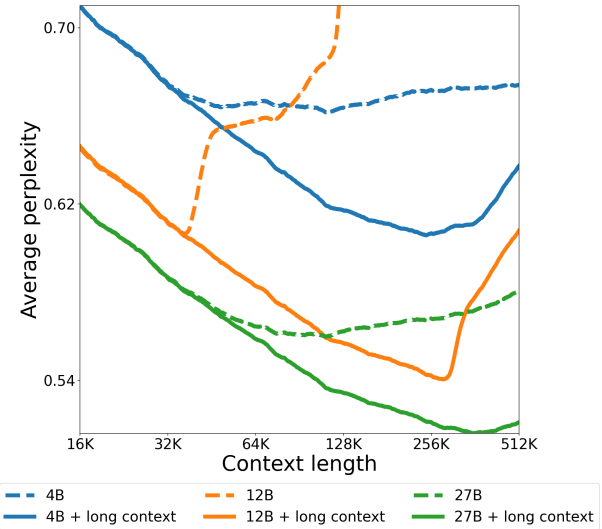

Figure 7 | **Long context** performance of pre-trained models before and after RoPE rescaling.

5.4. Small versus large teacher

A common finding is that, to train a small model, it is preferable to distill from a smaller teacher.