Gemma 3 Technical Report

Gemma Team, Google DeepMind[1]

We introduce Gemma 3, a multimodal addition to the Gemma family of lightweight open models, ranging in scale from 1 to 27 billion parameters. This version introduces vision understanding abilities, a wider coverage of languages and longer context – at least 128K tokens. We also change the architecture of the model to reduce the KV-cache memory that tends to explode with long context. This is achieved by increasing the ratio of local to global attention layers, and keeping the span on local attention short. The Gemma 3 models are trained with distillation and achieve superior performance to Gemma 2 for both pre-trained and instruction finetuned versions. In particular, our novel post-training recipe significantly improves the math, chat, instruction-following and multilingual abilities, making Gemma3-4B-IT competitive with Gemma2-27B-IT and Gemma3-27B-IT comparable to Gemini-1.5-Pro across benchmarks. We release all our models to the community.

1. Introduction

We present the newest version of Gemma open language models (Gemma Team, 2024a), co-designed with the family of Gemini frontier models (Gemini Team, 2023). This new version comes in sizes comparable to Gemma 2 (Gemma Team, 2024b), with the addition of a 1B model. These models are designed to run on standard consumer-grade hardware such as phones, laptops, and high-end GPUs. This version comes with several new abilities to the Gemma family; namely, multimodality, long context, and multilinguality, while preserving or surpassing the performance of prior versions.

In terms of multimodality, most Gemma 3 models are compatible with a tailored version of the SigLIP vision encoder (Zhai et al., 2023). The language models treat images as a sequence of soft tokens encoded by SigLIP. We reduce the inference cost of image processing by condensing the vision embeddings into a fixed size of 256 vectors. The encoder works at a fixed resolution and we take inspiration from LLaVA (Liu et al., 2024) to enable flexible resolutions with a Pan and Scan (P&S) method.

The second main architectural improvement is an increase in context size to 128K tokens, without reducing performance. A challenge with long context is the memory explosion of the KV cache during inference. To reduce this issue, we interleave multiple local layers between each global layer, and assign a smaller span of only 1024 tokens to the local layers. Therefore, only the global layers attend to long context, and we have 1 global for every 5 local layers.

The pre-training optimization recipe is similar to Gemma 2, with some modifications in the architecture design. We use the same tokenizer as Gemini 2.0, and we also revisit our data mixture to improve the multilingual capabilities of the models, while introducing image understanding. All Gemma 3 models are trained with knowledge distillation (Hinton et al., 2015).

In post-training, we focus our efforts on improving mathematics, reasoning, and chat abilities, as well as integrating the new capabilities of Gemma 3, long-context, and image inputs. We use a novel post-training approach that brings gains across all capabilities, including math, coding, chat, instruction following, and multilingual. The resulting Gemma 3 instruction-tuned models are both powerful and versatile, outperforming their predecessors by a wide margin.

In the following sections, we provide a brief overview of our models, including the architecture and pre- and post-training recipes. We also provide detailed evaluations across a wide variety of quantitative and qualitative benchmarks. We discuss our approach to safe and responsible deployment and outline the broader implications of Gemma 3, its limitations, and advantages.

[1]See Contributions and Acknowledgments section for full author list. Please send correspondence to gemma-3-report@google.com.

Figure 1 | Example of visual interaction with Gemma 3 27B IT model.

2. Model Architecture

Gemma 3 models follow the same general decoder-only transformer architecture as previous iterations (Vaswani et al., 2017), with most architecture elements similar to the first two Gemma versions. We use a Grouped-Query Attention (GQA) (Ainslie et al., 2023) with post-norm and pre-norm with RMSNorm (Zhang and Sennrich, 2019). Inspired by Dehghani et al. (2023), Wortsman et al. (2023) and Chameleon Team (2024), we replace the soft-capping of Gemma 2 with QK-norm. In this section, we focus on some key differences from previous versions below.

5:1 interleaving of local/global layers. We alternate between a local sliding window self-attention (Beltagy et al., 2020) and global self-attention (Luong et al., 2015), with a pattern of 5 local layers for every global layer, starting with a local layer as the first layer of the model.

Model	Vision Encoder	Embedding Parameters	Non-embedding Parameters
1B	0	302M	698M
4B	417M	675M	3,209M
12B	417M	1,012M	10,759M
27B	417M	1,416M	25,600M

Table 1 | Parameter counts for the Gemma 3 models. Our vocabulary has 256k entries.

Long context. Gemma 3 models support context length of 128K tokens, with the exception of the 1B model that has 32K. We increase RoPE base frequency from 10k to 1M on global self-attention layers, and keep the frequency of the local layers at 10k. We follow a process similar to the positional interpolation of Chen et al. (2023) to extend the span of the global self-attention layers.

2.1. Vision modality

Vision encoder. We use a 400M variant of the SigLIP encoder (Zhai et al., 2023), a Vision Transformer (Dosovitskiy, 2020) trained with a variation of the CLIP loss (Radford et al., 2021). The Gemma vision encoder takes as input square images resized to 896 x 896, and is finetuned on data from visual assistant tasks. For simplicity, we share the vision encoder across our 4B, 12B, and 27B models, keeping it frozen during training.

Pan & Scan (P&S). The Gemma vision encoder operates at a fixed resolution of 896×896. This results in artifacts when processing non-square aspect ratios and high-resolution images, leading to unreadable text, or small objects disappearing. We address this issue with an adaptive windowing algorithm during inference. This algorithm segments images into non-overlapping crops of equal size, covering the whole image, and resize them to 896×896 pixels to pass them to the encoder. This windowing is applied only when necessary, and control for the maximum number of crops. It is an inference-time only optimization and can be disabled for faster inference.

Model	Type	#Chips	Shards		
			Data	Seq.	Replica
1B	TPUv5e	512	16	16	2
4B	TPUv5e	2048	16	16	8
12B	TPUv4	6144	16	16	24
27B	TPUv5p	6144	24	8	32

Table 2 | Training infrastructure with sharding by data, sequence (Seq.), and replica.

2.2. Pre-training

We follow a similar recipe as in Gemma 2 for pre-training with knowledge distillation.

Training data. We pre-train our models on a slightly larger token budget than Gemma 2, i.e., we train on 14T tokens for Gemma 3 27B, 12T for the 12B version, 4T for the 4B, and 2T tokens for the 1B. The increase in tokens accounts for the mix of images and text used during pre-training. We also increase the amount of multilingual data to improve language coverage. We add both monolingual and parallel data, and we handle the imbalance in language representation using a strategy inspired by Chung et al. (2023).

Tokenizer. We use the same tokenizer as Gemini 2.0: a SentencePiece tokenizer with split digits, preserved whitespace, and byte-level encodings (Kudo and Richardson, 2018). The resulting vocabulary has 262k entries. This tokenizer is more balanced for non-English languages.

Filtering. We use filtering techniques that reduce the risk of unwanted or unsafe utterances and remove certain personal information and other sensitive data. We decontaminate evaluation sets from our pre-training data mixture, and reduce the risk of recitation by minimizing the proliferation of sensitive outputs. We also apply a quality reweighing step inspired by Sachdeva et al. (2024) to reduce occurrences of low quality data.

Distillation. We sample 256 logits per token, weighted by teacher probabilities. The student learns the teacher's distribution within these samples via cross-entropy loss. The teacher's target distribution is set to zero probability for non-sampled logits, and renormalized.

Model	Raw (GB)	Quantized (GB)		
	bf16	Int4	$\text{Int4}_{\text{blocks}=32}$	SFP8
1B	2.0	0.5	0.7	1.0
+KV	2.9	1.4	1.6	1.9
4B	8.0	2.6	2.9	4.4
+KV	12.7	7.3	7.6	9.1
12B	24.0	6.6	7.1	12.4
+KV	38.9	21.5	22.0	27.3
27B	54.0	14.1	15.3	27.4
+KV	72.7	32.8	34.0	46.1

Table 3 | Memory footprints (in GB) comparison between raw (bfloat16) and quantized checkpoints for weights and KV caching (+KV) at 32,768 context size, quantized in 8 bits.

2.3. Quantization Aware Training

Along with the raw checkpoints, we also provide quantized versions of our models in different standard formats. These versions are obtained by fine-tuning each model for a small number of steps, typically 5,000, using Quantization Aware Training (QAT) (Jacob et al., 2018). We use probabilities from the non-quantized checkpoint as targets, and adapt the data to match the pre-training and post-training distributions. Based on the most popular open source quantization inference engines (e.g. llama.cpp), we focus on three weight representations: per-channel int4, per-block int4, and switched fp8. In Table 3, we report the memory filled by raw and quantized models for each weight representation with and without a KV-cache for a sequence of 32k tokens.

2.4. Compute Infrastructure

We train our models with TPUv4, TPUv5e, and TPUv5p as outlined in Table 2. Each model configuration is optimized to minimize training step time. For the vision encoder, we pre-compute the embeddings for each image and directly train with the embeddings, adding no cost to the training of the language models.

The optimizer state is sharded using an implementation of ZeRO-3 (Ren et al., 2021). For multi-pod training, we perform a data replica re-

Context	Formatting
User turn	`<start_of_turn>user`
Model turn	`<start_of_turn>model`
End of turn	`<end_of_turn>`
Example of discussion:	
User: `Who are you?` **Model:** `My name is Gemma!` **User:** `What is 2+2?` **Model:** `2+2=4.`	
Model input:	

`[BOS]<start_of_turn>user`
`Who are you?<end_of_turn>`
`<start_of_turn>model`
`My name is Gemma!<end_of_turn>`
`<start_of_turn>user`
`What is 2+2?<end_of_turn>`
`<start_of_turn>model`	

Model output:	
`2+2=4.<end_of_turn>`	

Table 4 | Formatting for Gemma IT models. Explicitly add the `[BOS]` token after tokenization, or use the `add_bos=True` option in the tokenizer. *Do not tokenize the text "[BOS]".*

duction over the data center network, using the Pathways approach of Barham et al. (2022). We use the 'single controller' programming paradigm of Jax (Roberts et al., 2023) and Pathways (Barham et al., 2022), along with the GSPMD partitioner (Xu et al., 2021) and the MegaScale XLA compiler (XLA, 2019).

3. Instruction-Tuning

Pre-trained models are turned into instruction-tuned models with an improved post-training approach compared to our prior recipe (see Table 6).

Techniques. Our post-training approach relies on an improved version of knowledge distillation (Agarwal et al., 2024; Anil et al., 2018; Hinton et al., 2015) from a large IT teacher, along with a RL finetuning phase based on improved versions of BOND (Sessa et al., 2024), WARM (Ramé et al., 2024b), and WARP (Ramé et al., 2024a).

Reinforcement learning objectives. We use a variety of reward functions to improve helpfulness, math, coding, reasoning, instruction-following, and multilingual abilities, while minimizing model harmfulness. This includes learning from weight averaged reward models (Ramé et al., 2024b) trained with human feedback data, code execution feedback (Gehring et al., 2024), and ground-truth rewards for solving math problems (DeepSeek-AI, 2025; Lambert et al., 2024).

Data filtering. We carefully optimize the data used in post-training to maximize model performance. We filter examples that show certain personal information, unsafe or toxic model outputs, mistaken self-identification data, and duplicated examples. Including subsets of data that encourage better in-context attribution, hedging, and refusals to minimize hallucinations also improves performance on factuality metrics, without degrading model performance on other metrics.

[BOS] token. For both PT and IT models, text starts with a `[BOS]` token, that needs to be added explicitly since the text "[BOS]" does not map to the `[BOS]` token. For instance, Flax has an option, `add_bos=True`, to add this token automatically when tokenizing. An example of the formatting for an IT model is shown in Table 4,

PT versus IT Formatting. All models share the same tokenizer, with some control tokens dedicated to IT formatting. A key difference is that PT models output a `<eos>` token at the end of generation, while IT models output a `<end_of_turn>` at the end of the generation, as shown for IT in Table 4. Fine-tuning either model type thus also requires adding their respective end tokens.

4. Evaluation of final models

In this section, we evaluate the IT models over a series of automated benchmarks and human evaluations across a variety of domains, as well as static benchmarks such as MMLU.

4.1. LMSYS Chatbot Arena

In this section, we report the performance of our IT 27B model on LMSys Chatbot Arena (Chiang et al., 2024) in blind side-by-side evaluations by human raters against other state-of-the-art models. We report Elo scores in Table 5. Gemma 3 27B

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

Table 5 | Evaluation of Gemma 3 27B IT model in the Chatbot Arena (Chiang et al., 2024). All the models are evaluated against each other through blind side-by-side evaluations by human raters. Each model is attributed a score, based on the Elo rating system. *Gemma-3-27B-IT numbers are preliminary results received on March 8, 2025*.

IT (1338) is among the top 10 best models, with a score above other non-thinking open models, such as DeepSeek-V3 (1318), LLaMA 3 405B (1257), and Qwen2.5-70B (1257), which are much larger models. Finally, the Elo of Gemma 3 is significantly higher than Gemma 2, at 1220. Note that Elo scores do not take into account visual abilities, which none of the aforementioned models have.

4.2. Standard benchmarks

In Table 6, we show the performance of our final models across a variety of benchmarks compared to our previous model iteration, and Gemini 1.5. We do not compare directly with external models that often report their own evaluation settings, since running them in our setting does not guarantee a fair comparison. We encourage the reader to follow third-party static leaderboards for a fairer comparison across models. We include additional evaluations of our models on other benchmarks in the appendix.

5. Ablations

In this section, we focus on the impact of our architecture changes, as well as some of the vision abilities new to this model.

5.1. Pre-training ability probing

We use several standard benchmarks as probes during pre-training to ensure our models capture general abilities, and in Figure 2, we compare the quality of pre-trained models from Gemma 2 and 3 across these general abilities, namely, science,

	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.

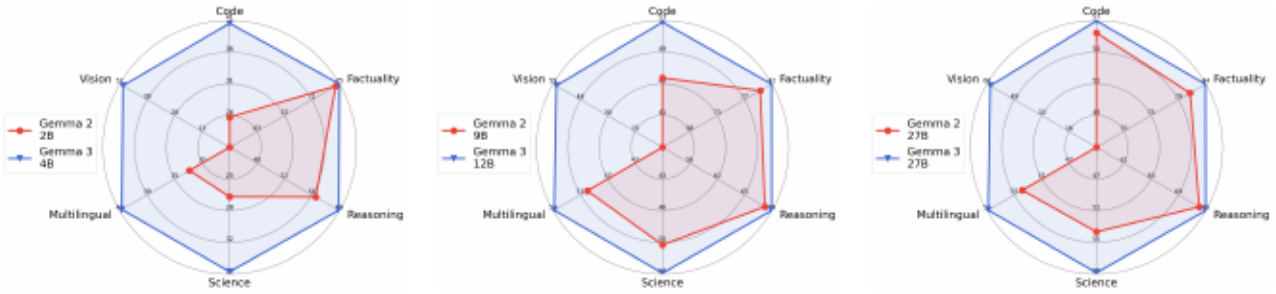

Figure 2 | Summary of the performance of different pre-trained models from Gemma 2 and 3 across general abilities. These plots are meant to give a simplified summary and details are in the appendix.

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.

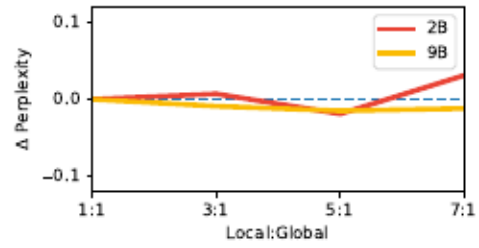

Figure 3 | **Impact of Local:Global ratio** on the perplexity on a validation set. The impact is minimal, even with 7-to-1 local to global. This ablation is run with text-only models.

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.

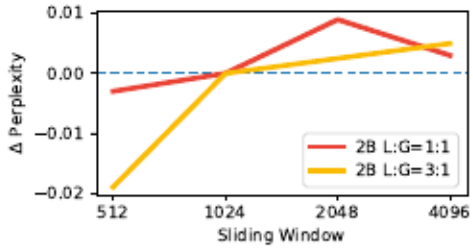

Figure 4 | **Impact of Sliding Window** size on perplexity measured on a validation set. We consider 2 2B models, with 1:1 and 1:3 local to global layer ratios. This ablation is run with text-only models.

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.

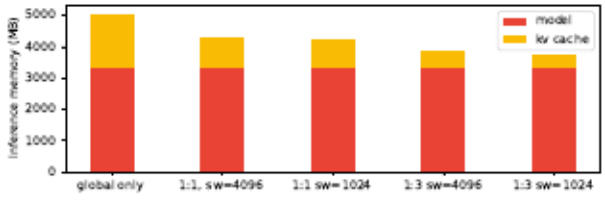

Figure 5 | **Model versus KV cache memory** during inference with a pre-fill KV cache of size 32k. We consider a 2B model with different local to global ratios and sliding window sizes (sw). We compare to global only, which is the standard used in Gemma 1 and Llama. This ablation is run with a text-only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

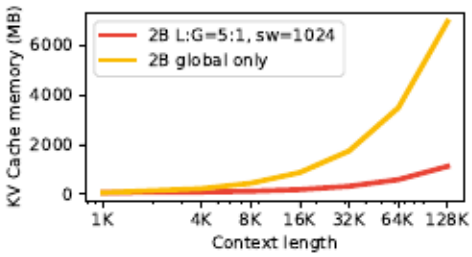

Figure 6 | **KV cache memory versus context length.** We show the memory usage of the KV cache for our architecture (L:G=5:1, sw=1024) and a transformer with global attention only – as used in LLaMa or Gemma 1.

quences 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.

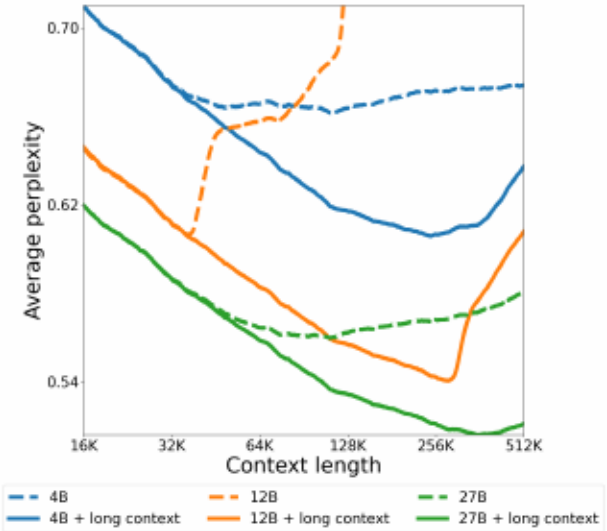

Figure 7 | **Long context** performance of pre-trained models before and after RoPE rescaling.

5.3. Enabling long context

Instead of training with 128K sequences from scratch, we pre-train our models with 32K se-

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.

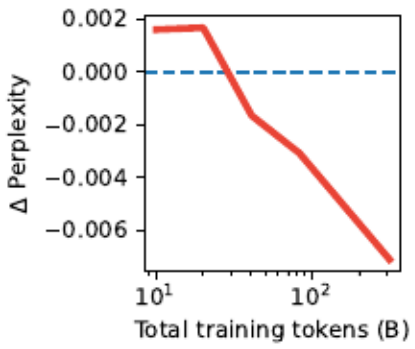

Figure 8 | **Small versus large teacher.** Relative difference of perplexity when using a small and large teacher as a function of the token size of training. Smaller numbers means distilling from a larger teacher is better.

We suspect this is because these studies are often performed in settings where the regularization effect of using a worse teacher surpasses the benefit of using a better teacher. We train a student with 2 teachers of different sizes, one large and one small, for different training horizons. In Fig. 8, we observe that for short training horizons, the smaller teacher is better, but the trend is reversed for longer training.

5.5. Vision encoder

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

Table 7 | **Impact of image encoder input resolution.** We measure performance using a short schedule 2B Gemma model on a few evaluation benchmarks to observe the effect of input image resolution on vision encoder pre-training.

Impact of image resolution. We use a vision encoder based on SigLIP (Zhai et al., 2023). The vision encoder is frozen, and only the language model is trained. Each image in this multimodal data is represented by 256 image tokens from the respective vision encoder. The higher resolution encoders thus use average pooling to reduce their output to 256 tokens. For instance, the 896 resolution encoder has a 4x4 average pooling on its output. As shown in Table 7, higher resolution encoders perform better than smaller ones.

	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)

Table 8 | **Impact of P&S.** 4-shot evaluation results on the valid set, with and without P&S on a pre-trained checkpoint. Boosts are on tasks associated with images with varying aspect ratios, or involving reading text on images.

Pan & Scan. P&S enables capturing images at close to their native aspect ratio and image resolution. In Table 8, we compare our 27B IT model with and without P&S. As expected, the ability to treat images with close to native resolution greatly helps with tasks that require some form of reading text on images, which is particularly important for visual language models.

6. Memorization and Privacy

Large language models may produce near-copies of some text used in training (Biderman et al., 2023; Carlini et al., 2021, 2022; Ippolito et al., 2022; Nasr et al., 2023). Several prior reports have released audits that quantify this risk by measuring the memorization rate (Anil et al., 2023; Chowdhery et al., 2022; Gemini Team, 2023, 2024; Gemma Team, 2024a,b; LLaMa Team, 2024). This "memorization rate"[1] is defined as the ratio of generations from the model that match its training data compared to all model generations using the following setup. We follow the methodology described in Gemma Team

[1] "We do not state or imply [here] that a model "contains" its training data in the sense that there is a copy of that data in the model. Rather, a model memorizes attributes of its training data such that in certain cases it is statistically able to generate such training data when following rules and using information about features of its training data that it does contain."

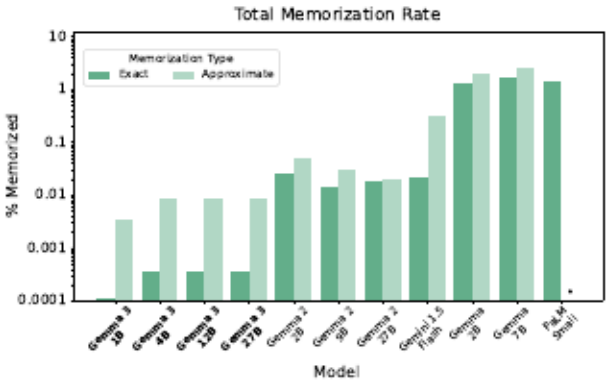

Figure 9 | Total memorization rates for both exact and approximate memorization. Gemma 3 models memorize significantly less than all prior models. *No results for approximate memorization on these models.

(2024b) to measure it. Specifically, we subsample a large portion of training data distributed uniformly across different corpora and test for discoverable extraction (Nasr et al., 2023) of this content using a prefix of length 50 and a suffix of length 50. We denote text as either "exactly memorized" if all tokens in the continuation match the source suffix or "approximately memorized" if they match up to an edit distance of 10%.

Figure 9 compares the memorization rates across Gemma and Gemini models; these models are ordered in reverse chronological order, with the newest Gemma 3 models on the left. We find that Gemma 3 models memorize long-form text at a much lower rate than prior models (note the log y-axis). We observe only a marginal difference in the memorization rates between the 4B, 12B, and 27B models, with 1B memorizing less than these larger models. Further, we find that a larger proportion of text is characterized as approximately memorized, with a relative increase in approximate memorization compared to exact memorization of roughly 24x on average.

We also study the rate at which the generations may contain personal information. To identify potentially personal information, we use the Google Cloud Sensitive Data Protection (SDP) service.[2] SDP uses broad detection rules to identify text that may contain personal information. SDP is designed to have high recall and does not consider the context in which the information may appear, which leads to many false positives. Thus, we are likely overestimating the true amount of potentially personal information contained in the outputs classified as memorized. SDP also provides broad severity levels: low, medium, and high. We classify text as personal if SDP classifies it as personal information at any severity level. We observed no personal information in the outputs characterized as memorization for all Gemma 3 models. This indicates a low rate of personal data, below our detection thresholds, in outputs classified as memorization.

[2]https://cloud.google.com/sensitive-data-protection

7. Responsibility, Safety, Security

Responsibility, safety, and security are of utmost importance in the development of Gemma models. To reduce risks to Gemma 3 users, we have continued to integrate enhanced internal safety processes that span the development workflow, in line with recent Google AI models (Gemini Team, 2024). This focuses on safety mitigation at training time, and robust and transparent model evaluations for the new image-to-text capabilities we have introduced.

7.1. Governance & Assessment

Our approach to assessing the benefits and risks of Gemma is reflective of that outlined for Gemma 1 (Gemma Team, 2024a), taking into account the changes in supported modalities. We continue to believe that openness in AI can spread the benefits of these technologies across society, but must be evaluated against the risk of malicious uses that can cause harm on both individual and institutional levels (Weidinger et al., 2021). Since the inaugural Gemma launch, we have seen these models drive a number of socially beneficial applications, such as our own ShieldGemma 2, a 4B image safety classifier built with Gemma 3, which provides a ready-made solution for image safety, outputting safety labels across dangerous content, sexually explicit, and violence categories.

Releasing Gemma 3 models required specific attention to changes in model capabilities and

close monitoring of the evolving risks of existing multimodal LLMs (Lin et al., 2024), as well as an understanding of the ways in which models are being used in the wild. Although we are yet to receive any reports of malicious use for Gemma, we remain committed to investigating any such reporting, and work with the academic and developer communities, as well as conduct our own monitoring, to flag such cases.

Despite advancements in capabilities, we believe that, given the number of larger powerful open models available, this release will have a negligible effect on the overall risk landscape.

7.2. Safety policies and train-time mitigations

A key pillar of Gemma's approach to safety is to align fine-tuned models with Google's safety policies, in line with Gemini models (Gemini Team, 2023). They are designed to help prevent our models from generating harmful content, i.e.,

- Child sexual abuse and exploitation
- Revealing personally identifiable information that can lead to harm (e.g., Social Security numbers)
- Hate speech and harassment
- Dangerous or malicious content (including promoting self-harm or instructing in harmful activities)
- Sexually explicit content
- Medical advice that runs contrary to scientific or medical consensus

We undertook considerable safety filtering of our pre-training data to reduce the likelihood of our pre-trained and fine-tuned checkpoints producing harmful content. For fine-tuned models, we also use both SFT and RLHF to steer the model away from undesirable behavior.

7.3. Assurance Evaluations

We also run our IT models through a set of baseline assurance evaluations to understand the potential harms that our models can cause. As we champion open models, we also recognize that the irreversible nature of weight releases requires rigorous risk assessment. Our internal safety processes are designed accordingly, and for previous Gemma models we have also undertaken evaluations of capabilities relevant to extreme risks (Phuong et al., 2024; Shevlane et al., 2023). As we continue to develop and share open models, we will follow the heuristic that thoroughly evaluating a more capable model often provides sufficient assurance for less capable ones. As such, we prioritised a streamlined set of evaluations for Gemma 3, reserving in-depth dangerous capability assessments for cases where a specific model may present a potentially heightened risk (as described below on CBRN evaluations). We balance development speed with targeted safety testing, ensuring our evaluations are well-focused and efficient, while upholding the commitments laid out in our Frontier Safety Framework.

Baseline Evaluations

Baseline assurance captures the model violation rate for safety policies, using a large number of synthetic adversarial user queries, and human raters to label the answers as policy violating or not. Overall, Gemma 3 violation rate is significantly low overall on these safety policies.

Chemical, Biological, Radiological and Nuclear (CBRN) knowledge

Owing to enhanced performance on STEM-related tasks, we evaluated knowledge relevant to biological, radiological, and nuclear risks using an internal dataset of closed-ended, knowledge-based multiple choice questions. For evaluations of chemical knowledge, we employed a closed-ended knowledge-based approach on chemical hazards developed by Macknight et al. Our evaluation suggests that the knowledge of Gemma 3 models in these domains is low.

7.4. Our approach to responsible open models

Designing safe, secure, and responsible applications requires a system-level approach, working to mitigate risks associated with each specific use case and environment. We will continue to adopt assessments and safety mitigations proportionate to the potential risks from our models, and

will only share these with the community when we are confident that the benefits significantly outweigh the foreseeable risks.

8. Discussion and Conclusion

In this work, we have presented Gemma 3, the latest addition to the Gemma family of open language models for text, image, and code. In this version, we focus on adding image understanding and long context while improving multilinguality and STEM-related abilities. Our model sizes and architectures are designed to be compatible with standard hardware, and most of our architecture improvements are tailored to fit this hardware while maintaining performance.

References

Realworldqa. https://x.ai/news/grok-1.5v.

M. Acharya, K. Kafle, and C. Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2018.

R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist, and O. Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018.

R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *ACL*, 2020.

A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*, 2020.

J. Austin, A. Odena, M. I. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. J. Cai, M. Terry, Q. V. Le, and C. Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021.

P. Barham, A. Chowdhery, J. Dean, S. Ghemawat, S. Hand, D. Hurt, M. Isard, H. Lim, R. Pang, S. Roy, B. Saeta, P. Schuh, R. Sepassi, L. E. Shafey, C. A. Thekkath, and Y. Wu. Pathways: Asynchronous distributed dataflow for ml, 2022.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

S. Biderman, U. Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit, and E. Raff. Emergent and predictable memorization in large language models. *NeurIPS*, 36: 28072–28090, 2023.

Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641, 2019.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *USENIX*, 2021.

N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf,

G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.

S. Chen, S. Wong, L. Chen, and Y. Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

F. Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

H. W. Chung, N. Constant, X. Garcia, A. Roberts, Y. Tay, S. Narang, and O. Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, 2023.

C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044, 2019.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoningt learning, 2025.

M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

D. Deutsch, E. Briakou, I. Caswell, M. Finkelstein, R. Galor, J. Juraska, G. Kovacs, A. Lui, R. Rei, J. Riesa, S. Rijhwani, P. Riley, E. Salesky, F. Trabelsi, S. Winkler, B. Zhang, and M. Freitag. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects, 2025.

A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *ACL*, 2019.

B. Fatemi, M. Kazemi, A. Tsitsulin, K. Malkan, J. Yim, J. Palowitch, S. Seo, J. Halcrow, and B. Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390, 2024.

J. Gehring, K. Zheng, J. Copet, V. Mella, T. Cohen, and G. Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*, 2024.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

Gemma Team. Gemma: Open models based on gemini research and technology, 2024a.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.

O. Goldman, U. Shaham, D. Malkin, S. Eiger, A. Hassidim, Y. Matias, J. Maynez, A. M. Gilady, J. Riesa, S. Rijhwani, L. Rimell, I. Szpektor, R. Tsarfaty, and M. Eyal. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025.

N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *ACL*, 2022.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

J. Hessel, A. Marasović, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

D. Ippolito, F. Tramèr, M. Nasr, C. Zhang, M. Jagielski, K. Lee, C. A. Choquette-Choo, and N. Carlini. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*, 2022.

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, 2018.

M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551, 2017.

M. Kazemi, H. Alvari, A. Anand, J. Wu, X. Chen, and R. Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

M. Kazemi, N. Dikkala, A. Anand, P. Dević, I. Dasgupta, F. Liu, B. Fatemi, P. Awasthi, D. Guo, S. Gollapudi, and A. Qureshi. Remi: A dataset for reasoning with multiple images. *ArXiv*, abs/2406.09175, 2024a.

M. Kazemi, Q. Yuan, D. Bhatia, N. Kim, X. Xu, V. Imbrasaite, and D. Ramachandran. Boardgameqa: A dataset for natural language reasoning with contradictory information. *NeurIPS*, 36, 2024b.

M. Kazemi, B. Fatemi, H. Bansal, J. Palowitch, C. Anastasiou, S. V. Mehta, L. K. Jain, V. Aglietti, D. Jindal, P. Chen, et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016.

E. Kıcıman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. 2018.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *ACL*, 2019.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Z. Lin, J. Cui, X. Liao, and X. Wang. Malla: Demystifying real-world large language model integrated malicious services, 2024.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *NeurIPS*, 36, 2024.

LLaMa Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. 2015.

Macknight, Aung, and Gomes. Personal Communication.

K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. ACL, 2022.

M. Mathew, D. Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. *WACV*, 2020.

M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar. Infographicvqa. In *WACV*, 2022.

I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

A. Nie, Y. Zhang, A. S. Amdekar, C. Piech, T. B. Hashimoto, and T. Gerstenberg. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *NeurIPS*, 36, 2024.

R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching clip to count to ten. *ICCV*, 2023.

M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Deletang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, and T. Shevlane. Evaluating frontier models for dangerous capabilities, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

A. Ramé, J. Ferret, N. Vieillard, R. Dadashi, L. Hussenot, P.-L. Cedoz, P. G. Sessa, S. Girgin, A. Douillard, and O. Bachem. WARP: On the benefits of weight averaged rewarded policies, 2024a.

A. Ramé, N. Vieillard, L. Hussenot, R. Dadashi, G. Cideron, O. Bachem, and J. Ferret. WARM: On the benefits of weight averaged reward models. In *ICML*, 2024b.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*, abs/2311.12022, 2023.

J. Ren, S. Rajbhandari, R. Y. Aminabadi, O. Ruwase, S. Yang, M. Zhang, D. Li, and Y. He. Zero-offload: Democratizing billion-scale model training. In *USENIX*, 2021.

A. Roberts, H. W. Chung, G. Mishra, A. Levskaya, J. Bradbury, D. Andor, S. Narang, B. Lester, C. Gaffney, A. Mohiuddin, et al. Scaling up models and data with t5x and seqio. *JMLR*, 2023.

N. Sachdeva, B. Coleman, W.-C. Kang, J. Ni, L. Hong, E. H. Chi, J. Caverlee, J. McAuley, and D. Z. Cheng. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*, 2024.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641, 2019.

E. Sánchez, B. Alastruey, C. Ropers, P. Stenetorp, M. Artetxe, and M. R. Costa-jussà. Linguini: A benchmark for language-agnostic linguistic reasoning. *arXiv preprint arXiv:2409.12126*, 2024.

M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728, 2019.

P. G. Sessa, R. Dadashi, L. Hussenot, J. Ferret, N. Vieillard, A. Ramé, B. Shariari, S. Perrin, A. Friesen, G. Cideron, S. Girgin, P. Stanczyk, A. Michi, D. Sinopalnikov, S. Ramos, A. Héliou, A. Severyn, M. Hoffman, N. Momchev, and O. Bachem. Bond: Aligning llms with best-of-n distillation, 2024.

K. Shah, N. Dikkala, X. Wang, and R. Panigrahy. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *arXiv preprint arXiv:2409.10502*, 2024.

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe. Model evaluation for extreme risks, 2023.

F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language models are multilingual chain-of-thought reasoners. In *ICLR*, 2023.

A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, and P. Talukdar. Indicgenbench: a multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*, 2024a.

S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W. Q. Leong, Y. Susanto, R. Ng, S. Longpre, W.-Y. Ko, M. Smith, A. Bosselut, A. Oh, A. F. T. Martins, L. Choshen, D. Ippolito, E. Ferrante, M. Fadaee, B. Ermis, and S. Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024b.

A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long, S. Qin, R. Ingle, E. Bugliarello, S. Kazemzadeh, T. Mesnard, I. Alabdulmohsin, L. Beyer, and X. Zhai. PaliGemma 2: A Family of Versatile VLMs for Transfer. *arXiv preprint arXiv:2412.03555*, 2024.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

G. Tyen, H. Mansoor, P. Chen, T. Mak, and V. Cărbune. Llms cannot find reasoning errors, but can correct them! *arXiv preprint arXiv:2311.08516*, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. 2017.

K. Vodrahalli, S. Ontanon, N. Tripuraneni, K. Xu, S. Jain, R. Shivanna, J. Hui, N. Dikkala, M. Kazemi, B. Fatemi, et al. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS*, 2024.

L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models, 2021.

C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

M. Wortsman, P. J. Liu, L. Xiao, K. Everett, A. Alemi, B. Adlam, J. D. Co-Reyes, I. Gur, A. Kumar, R. Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

XLA. Xla: Optimizing compiler for tensorflow, 2019. URL https://www.tensorflow.org/xla.

Y. Xu, H. Lee, D. Chen, B. A. Hechtman, Y. Huang, R. Joshi, M. Krikun, D. Lepikhin, A. Ly, M. Maggioni, R. Pang, N. Shazeer, S. Wang, T. Wang, Y. Wu, and Z. Chen. GSPMD: general and scalable parallelization for ML computation graphs. 2021.

Y. Yamada, Y. Bao, A. K. Lampinen, J. Kasai, and I. Yildirim. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*, 2023.

K. Yang, O. Russakovsky, and J. Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. *ICCV*, 2019.

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *CVPR*, 2023.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In *ACL*, 2019.

X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *CVPR*, 2023.

B. Zhang and R. Sennrich. Root mean square layer normalization. 2019.

J. Zhang, L. Jain, Y. Guo, J. Chen, K. L. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. Rogers, K. Jamieson, et al. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*, 2024.

W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Appendix

Details of pre-trained performances.

	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
HellaS	72.9	81.9	**86.4**	62.3	77.2	84.2	85.6
BoolQ	75.6	77.5	76.2	63.2	72.3	78.8	**82.4**
PIQA	78.1	81.9	**83.5**	73.8	79.6	81.8	83.3
SIQA	51.8	53.3	53.8	48.9	51.9	53.4	**54.9**
TQA	60.2	76.5	83.8	39.8	65.8	78.2	**85.5**
NQ	17.2	29.2	34.7	9.48	20.0	31.4	**36.1**
ARC-C	55.8	69.1	**71.4**	38.4	56.2	68.9	70.6
ARC-E	80.6	88.3	88.6	73.0	82.4	88.3	**89.0**
WinoG	65.4	73.9	**79.4**	58.2	64.7	74.3	78.8
BBH	42.4	69.4	74.8	28.4	50.9	72.6	**77.7**
Drop	53.2	71.5	75.2	42.4	60.1	72.2	**77.2**

Table 9 | Factuality, common-sense performance and reasoning after pre-training phase.

Factuality and common-sense. In Table 9, we report the performance of our new pre-trained benchmarks compared to previous versions. We consider several standard benchmarks, namely HellaSwag (Zellers et al., 2019), BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), ARC-C and ARC-E (Chollet, 2019), WinoGrande (Sakaguchi et al., 2019), BBH (Suzgun et al., 2022), DROP (Dua et al., 2019). Evaluation details are described in Table 19. Overall, our models are in the same ballpark as Gemma 2, which is encouraging since these abilities are not the focus of the improvements brought in this version.

STEM and code. The details of our performance on STEM and Code are in Table 10. We consider several standard benchmarks, namely MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024), AGIEval (Zhong et al., 2023), MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), GPQA (Rein et al., 2023), MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021). Evaluation details are described in Table 19. Overall we see a consistent improvement over STEM abilities across our pre-trained models. On code, we see a similar improvement for the 4B and 12B models but not on the 27B.

	Gemma 2			Gemma 3		
	2B	9B	27B	4B	12B	27B
MMLU	52.2	71.2	75.2	59.6	74.5	**78.6**
MMLUpro	22.2	43.7	49.4	29.2	45.3	**52.2**
AGIE	31.6	53.1	55.1	42.1	57.4	**66.2**
MATH	16.4	36.4	42.1	24.2	43.3	**50.0**
GSM8K	25.0	70.2	74.6	38.4	71.0	**82.6**
GPQA Diamond	12.5	24.8	**26.3**	15.0	25.4	24.3
MBPP	31.0	51.2	60.8	46.0	60.4	**65.6**
HumanE	19.5	40.2	**51.2**	36.0	45.7	48.8

Table 10 | STEM and code performance after pre-training phase.

	4B	12B	27B
COCO caption	102	111	**116**
DocVQA	72.8	82.3	**85.6**
InfoVQA	44.1	54.8	**59.4**
MMMU	39.2	50.3	**56.1**
TextVQA	58.9	66.5	**68.6**
RealWorldQA	45.5	52.2	**53.9**
ReMI	27.3	38.5	**44.8**
AI2D	63.2	75.2	**79.0**
ChartQA	63.6	74.7	**76.3**
VQAv2	63.9	71.2	**72.9**
BLINK	38.0	35.9	**39.6**
OK-VQA	51.0	58.7	**60.2**
TallyQA	42.5	51.8	**54.3**
SpatialSense VQA	50.9	**60.0**	59.4
CountBench VQA	26.1	17.8	**68.0**

Table 11 | Multimodal performance after pre-training phase. The scores are on the val split of each dataset without P&S.

Image understanding. In Table 11, we report performance across a variety of visual question answer benchmarks for the different models that were trained with a vision encoder, namely COCO Caption (Chen et al., 2015), DocVQA (Mathew et al., 2020), InfographicVQA (Mathew et al., 2022), MMMU (Yue et al., 2023), TextVQA (Singh et al., 2019), RealWorldQA (Rea), ReMI (Kazemi et al., 2024a),

AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), VQA v2 (Goyal et al., 2017), BLINK (Fu et al., 2024), OK-VQA (Marino et al., 2019), TallyQA (Acharya et al., 2018), SpatialSense VQA (Yang et al., 2019), CountBench VQA (Paiss et al., 2023). Evaluation details are described in Table 20.

	PaliGemma 2			Gemma 3		
	2B	9B	27B	4B	12B	27B
DocVQA	81.6	86.3	85.1	86.1	89.0	**89.5**
InfoVQA	41.4	53.1	50.2	55.6	61.6	**64.6**
TextVQA	76.3	76.3	75.1	79.1	81.6	**83.2**
ChartQA	70.7	79.1	71.3	79.8	83.5	83.4
AI2D	76.0	84.4	84.6	80.9	85.6	**86.5**
OKVQA	64.1	68.6	70.6	65.2	69.3	**71.1**
CountBenchQA	82.0	85.3	87.4	79.4	83.5	**87.8**
COCO caption	143.	**145.**	**145.**	143.	143.	144.
VQAv2	84.8	**85.8**	**85.8**	84.1	84.9	85.1
Tally QA	80.6	**82.4**	82.1	79.0	81.3	81.7

Table 12 | Performance of pre-trained checkpoints after fine-tuning on multi-modal benchmarks (without P&S). PaliGemma 2 was transferred at 896x896 resolution for the first four benchmarks, and at 448x448 resolution for the others.

Comparison to PaliGemma 2. We fine-tune multimodal Gemma 3 pre-trained checkpoints following the protocol from Steiner et al. (2024) – only learning rate is swept, otherwise the same transfer settings are used. The results in Table 12 show that Gemma 3 excels at benchmarks involving document understanding, even outperforming the *larger* PaliGemma 2 variant. Note that due to average pooling in the vision encoder the Gemma 3 4B and 12B models are about 10x cheaper to transfer compared with the PaliGemma 2 9B and 27B models at the same 896 x 896 resolution. Gemma 3 also performs better on AI2D and OKVQA, but PaliGemma 2 performs slightly better on VQAv2 and COCO caption.

Multilinguality. In Table 13 we report the performance of the pre-trained models on multilingual tasks. We apply in-context learning with multi-shot prompting and present results on the following benchmarks: MGSM (Shi et al., 2023), Global-MMLU-Lite (Singh et al., 2024b), WMT24++ (Deutsch et al., 2025), FLoRes (Goyal et al., 2022), XQuAD (Artetxe et al., 2020), ECLeKTic (Goldman et al., 2025), IndicGenBench (Singh et al., 2024a), XOR QA (Asai et al., 2020). Evaluation details are described in Table 19.

	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
MGSM	18.7	57.3	68.0	2.04	34.7	64.3	**74.3**
GMMLU	43.3	64.0	69.4	24.9	57.0	69.4	**75.7**
WMT24++	38.8	50.3	53.0	36.7	48.4	53.9	**55.7**
Flores	30.2	41.3	44.3	29.5	39.2	46.0	**48.8**
XQuAD	53.7	72.2	73.9	43.9	68.0	74.5	**76.8**
ECLeKTic	8.29	14.0	17.1	4.69	11.0	17.2	**24.4**
IndicGB	47.4	59.3	62.1	41.4	57.2	61.7	**63.4**

Table 13 | Multilingual performance after the pre-training phase. IndicGenBench is an average over benchmarks reported in Table 14.

	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
XQuAD Indic	54.3	73.1	74.9	43.1	68.3	75.2	**77.8**
XORQA in-en	66.2	69.3	**72.5**	56.3	68.3	69.8	70.4
XORQA in-xx	31.2	40.8	44.3	27.1	39.8	43.8	**46.0**
Flores Indic	38.1	54.0	56.9	39.0	52.3	58.0	**59.5**

Table 14 | Detailed IndicGenBench performance after the pre-training phase.

Long context. In Table 15 we report the performance of pre-trained and fine-tuned models on long context benchmarks. We include RULER (Hsieh et al., 2024) and MRCR (Vodrahalli et al., 2024) benchmarks evaluating at 32K and 128K sequence lengths.

8.1. Performance of IT models

We report in Table 18, additional benchmarks on our IT models. Note that N2C refers to Natural2Code, the Gemini 1.0 internal held-out dataset, which uses author-generated sources instead of web-based information. BBEH refers to BIG-Bench Extra Hard (Kazemi et al., 2025), a challenging LLM reasoning benchmark that aggregates several reasoning tasks (Fatemi et al., 2024;

		Gemma 3 PT			Gemma 3 IT		
	Context	4B	12B	27B	4B	12B	27B
RULER	32K	67.1	**90.6**	85.9	61.4	80.3	**91.1**
RULER	128K	51.7	**80.7**	72.9	46.8	57.1	**66.0**
MRCR	32K	44.7	59.8	**63.2**	49.8	53.7	**63.2**
MRCR	128K	40.6	56.9	**60.0**	44.6	49.8	**59.3**

Table 15 | Performance of pre-trained (PT) and instruction fine-tuned (IT) models on long context benchmarks at different context lengths.

	4B	12B	27B
MMMU (val)	48.8	59.6	**64.9**
DocVQA	75.8	**87.1**	86.6
InfoVQA	50.0	64.9	**70.6**
TextVQA	57.8	**67.7**	65.1
AI2D	74.8	84.2	**84.5**
ChartQA	68.8	75.7	**78.0**
VQAv2 (val)	62.4	**71.6**	71.0
MathVista (testmini)	50.0	62.9	**67.6**

Table 16 | Performance of instruction fine-tuned (IT) models on multimodal benchmarks. If not mentioned, these results are on the final test set of each dataset with P&S applied.

Hessel et al., 2022; Kazemi et al., 2023, 2024b; Kıcıman et al., 2023; Nie et al., 2024; Sánchez et al., 2024; Shah et al., 2024; Tyen et al., 2023; White et al., 2024; Yamada et al., 2023; Zhang et al., 2024). ECLeKTic refers to Goldman et al. (2025). We report the micro average score. More evaluation details are described in Table 21.

8.2. Performance of IT models on video understanding

Additional multimodal evaluations. Gemma 3 IT models were evaluated on common vision benchmarks following the evaluation protocol of Gemini 1.5 (Gemini Team, 2024). The results are given in Table 16 when P&S is activated.

	4B	12B	27B
Perception Test MCVQA	50.6	54.9	58.1
ActivityNet-QA	46.3	50.4	52.8

Table 17 | Performance of instruction fine-tuned (IT) models on vision understanding benchmarks using 0 shot with 16 frames linspace. Perception Test consists of real-world videos designed to show perceptually interesting situations and we report results on the multiple choice video QA benchmark in terms of top-1 accuracy. ActivityNet-QA reports standard gpt-evaluation.

	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
MMLU	56.1	71.3	76.2	38.8	58.1	71.9	**76.9**
MBPP	36.6	59.2	67.4	35.2	63.2	73.0	**74.4**
HumanEval	20.1	40.2	51.8	41.5	71.3	85.4	**87.8**
N2C	46.8	68.3	77.3	56.0	70.3	80.7	**84.5**
LiveCodeBench	7.0	20.0	29.0	5.0	23.0	32.0	**39.0**
GSM8K	62.6	88.1	91.1	62.8	89.2	94.4	**95.9**
MATH	27.2	49.4	55.6	48.0	75.6	83.8	**89.0**
HiddenMath	2.0	8.0	12.0	15.0	42.0	51.0	**56.0**
BBH	41.4	69.0	74.9	39.1	72.2	85.7	**87.6**
BBEH	5.9	9.8	14.8	7.2	11.0	16.3	**19.3**
IFEval	80.4	88.4	**91.1**	80.2	90.2	88.9	90.4
GMMLU-Lite	41.9	64.8	68.6	34.2	54.5	69.5	**75.1**
ECLeKTic	5.3	11.8	**17.6**	1.4	4.6	10.3	16.7
WMT24++	37.4	48.7	51.7	35.9	46.8	51.6	**53.4**

Table 18 | Performance of instruction fine-tuned (IT) models of different sizes on more internal and external benchmarks.

Evaluation	Metric	Type	n-shot	COT	Norm
MBPP	pass@1	sampling	3-shot		
HumanEval	pass@1	sampling	0-shot		
HellaSwag	Accuracy	scoring	10-shot		Char-Len
BoolQ	Accuracy	scoring	0-shot		Char-Len
PIQA	Accuracy	scoring	0-shot		Char-Len
SIQA	Accuracy	scoring	0-shot		Char-Len
TriviaQA	Accuracy	sampling	5-shot		
Natural Questions	Accuracy	sampling	5-shot		
ARC-C	Accuracy	scoring	25-shot		Char-Len
ARC-E	Accuracy	scoring	0-shot		Char-Len
WinoGrande	Accuracy	scoring	5-shot		Char-Len
BBH	Accuracy	sampling	few-shot	Yes	
DROP	Token F1 score	sampling	1-shot		
AGIEval	Accuracy	sampling	3-5-shot		
MMLU	Accuracy	scoring	5-shot		Char-Len
MATH	Accuracy	sampling	4-shot	Yes	
GSM8K	Accuracy	sampling	8-shot	Yes	
GPQA Diamond	Accuracy	sampling	5-shot	Yes	
MMLU-Pro	Accuracy	sampling	5-shot	Yes	
MGSM	Accuracy	sampling	8-shot		
FLoRes	CHaRacter-level F-score	sampling	1-shot		
Global-MMLU-Lite	Accuracy	scoring	5-shot		Char-Len
XQuAD	CHaRacter-level F-score	sampling	5-shot		
WMT24++	CHaRacter-level F-score	sampling	5-shot		
ECLeKTic	ECLeKTic score	sampling	2-shot		First-line/strip
XQuAD Indic	CHaRacter-level F-score	sampling	5-shot		
XOR QA IN-EN	CHaRacter-level F-score	sampling	5-shot		
XOR QA IN-XX	CHaRacter-level F-score	sampling	5-shot		
FLoRes Indic	CHaRacter-level F-score	sampling	5-shot		
RULER	Accuracy	sampling	0-shot		
MRCR	MRCR score	sampling	few-shot		

Table 19 | Details on text benchmarks. Char-Len stands for Character Length Normalization and COT stands for Chain-Of-Thought prompting.

Evaluation	Metric	Type	n-shot
COCO Caption	Cider score	sampling	4-shot
DocVQA	ANLS score	sampling	4-shot
InfographicVQA	ANLS score	sampling	4-shot
MMMU	Accuracy	sampling	3-shot text only
TextVQA	Accuracy	sampling	4-shot
RealWorldQA	Accuracy	sampling	4-shot text only
ReMI	Accuracy	sampling	4-shot
AI2D	Accuracy	sampling	4-shot
ChartQA	Accuracy	sampling	4-shot
VQA v2	Accuracy	sampling	4-shot
BLINK	Accuracy	sampling	0-shot
OK-VQA	Accuracy	sampling	4-shot
TallyQA	Accuracy	sampling	4-shot
SpatialSense VQA	Accuracy	sampling	4-shot
CountBench VQA	Accuracy	sampling	0-shot

Table 20 | Details on vision benchmarks. No Chain-Of-Thought prompting nor normalization.

Evaluation	Metric	Type	n-shot	COT
MMLU	Accuracy	sampling	0-shot	
MBPP	pass@1	sampling	3-shot	
HumanEval	pass@1	sampling	0-shot	
N2C	pass@1	sampling	0-shot	
LiveCodeBench	Average over 8 samples	sampling	0-shot	Yes
GSM8K	Accuracy	sampling	0-shot	Yes
GPQA Diamond	Accuracy	sampling	0-shot	Yes
MATH	Accuracy	sampling	0-shot	
HiddenMath	Accuracy	sampling	0-shot	
BBH	Accuracy	sampling	0-shot	
BBEH	Accuracy	sampling	0-shot	
IFEval	Accuracy	sampling	0-shot	
Global-MMLU-lite	Accuracy	sampling	0-shot	Yes
ECLeKTic	ECLeKTic score	sampling	0-shot	
WMT24++	CHaRacter-level F-score	sampling	0-shot	

Table 21 | Details on instruction fine-tuned (IT) benchmarks. No normalization.