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.

Gemma 3 技术报告

Gemma团队，谷歌DeepMind1

我们介绍Gemma 3，这是Gemma系列轻量级开放模型的多模态新增，参数规模从1亿到270亿不等。此版本引入了视觉理解能力，覆盖更多语言，并且支持更长的上下文——至少128K个标记。我们还改变了模型的架构，以减少在长上下文中容易爆炸的KV-cache内存。这是通过增加局部注意力层与全局注意力层的比例，并保持局部注意力的跨度较短来实现的。Gemma 3模型通过蒸馏训练，表现优于Gemma 2，无论是预训练版本还是指令微调版本。特别是，我们的新型后训练方案显著提高了数学、聊天、遵循指令和多语言能力，使Gemma3-4B-IT在基准测试中与Gemma2-27B-IT具有竞争力，而Gemma3-27B-IT与Gemini-1.5-Pro相当。我们将所有模型发布给社区。

1. 引言

我们呈现了Gemma开放语言模型的最新版本（Gemma团队，2024a），与Gemini前沿模型系列（Gemini团队，2023）共同设计。这个新版本的规模与Gemma 2（Gemma团队，2024b）相当，并增加了一个1B模型。这些模型旨在运行在标准消费级硬件上，如手机、笔记本电脑和高端GPU。这个版本为Gemma家族带来了几个新功能；即多模态、长上下文和多语言能力，同时保持或超越了之前版本的性能。

在多模态方面，大多数Gemma 3模型与SigLIP视觉编码器的定制版本兼容（Zhai等，2023）。语言模型将图像视为由SigLIP编码的软令牌序列。我们通过将视觉嵌入压缩为固定大小的256个向量来降低图像处理的推理成本。编码器在固定分辨率下工作，我们从LLaVA（Liu等，2024）中获得灵感，以通过平移和扫描（P&S）方法实现灵活的分辨率。

第二个主要的架构改进是将上下文大小增加到128K个标记，而不降低性能。长上下文的一个挑战是在推理过程中KV缓存的内存爆炸。为了减少这个问题，我们在每个全局层之间交错多个局部层，并将仅1024个标记的较小跨度分配给本地层。因此，只有全局层关注长上下文，并且每5个本地层有1个全局层。

预训练优化方案类似于Gemma 2，但在架构设计上进行了某些修改。我们使用与Gemini 2.0相同的分词器，并且我们还重新审视了我们的数据混合，以提高模型的多语言能力，同时引入图像理解。所有Gemma 3模型都采用知识蒸馏进行训练（Hinton等，2015）。

在后训练阶段，我们将精力集中在提高数学、推理和聊天能力，以及整合Gemma 3的新功能、长上下文和图像输入上。我们采用了一种新颖的后训练方法，带来了所有能力的提升，包括数学、编码、聊天、遵循指令和多语言。最终的Gemma 3指令调优模型既强大又多功能，远远超过了其前身。

在以下章节中，我们提供了对我们模型的简要概述，包括架构以及训练前和训练后的方案。我们还提供了在各种定量和定性基准上的详细评估。我们讨论了我们在安全和负责任的部署方面的方法，并概述了Gemma 3的更广泛影响、局限性和优势。

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

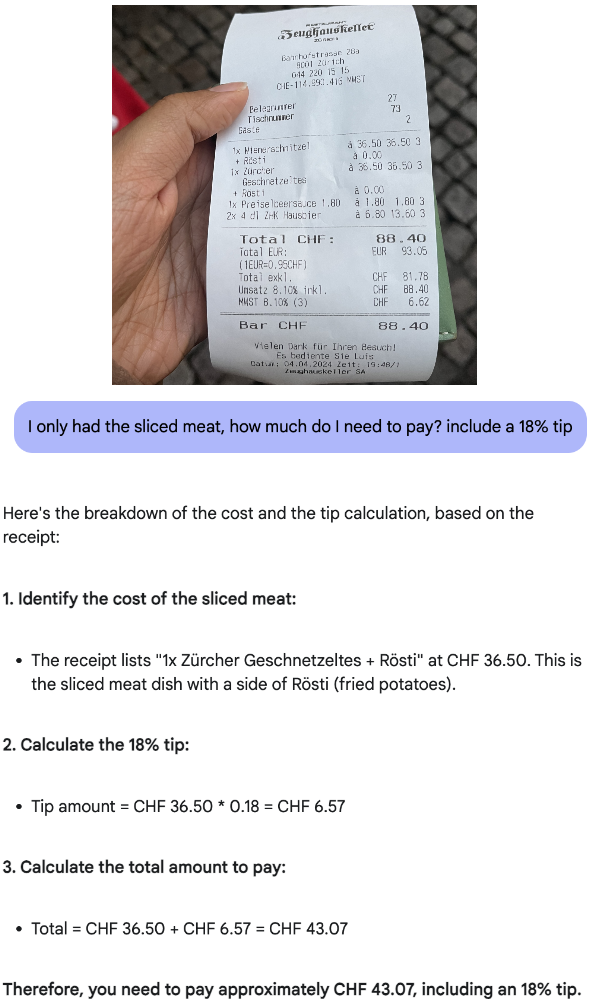

Figure 1 | Example of visual interaction with Gemma 3 27B IT model.

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 object disappeared. 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.

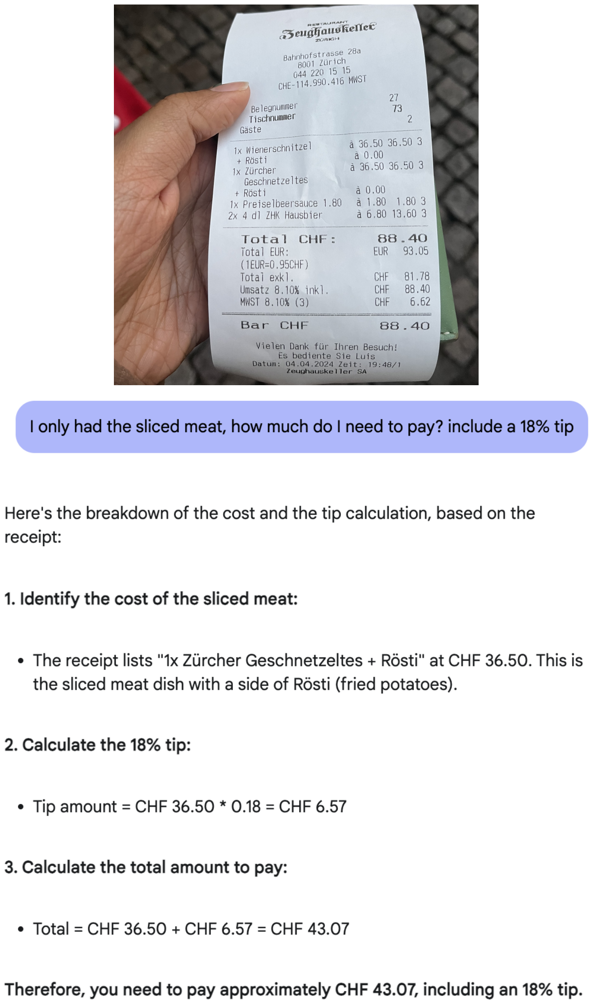

图 1 | Gemma 3 27B IT 模型的视觉交互示例。

2. 模型架构

Gemma 3 模型遵循与之前版本相同的一般解码器仅 transformer 架构（Vaswani 等，2017），大多数架构元素与前两个 Gemma 版本相似。我们使用带有后归一化和前归一化的分组查询注意力（GQA）（Ainslie 等，2023），并采用 RMSNorm（Zhang 和 Sennrich，2019）。受到 Dehghani 等（2023）、Wortsman 等（2023）和 Chameleon Team（2024）的启发，我们用 QK-norm 替换了 Gemma 2 的软上限。在本节中，我们将重点介绍与之前版本的一些关键差异。

5:1 局部/全局层的交错。我们在局部滑动窗口自注意力 (Beltagy et al., 2020) 和全局自注意力之间交替。

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

表 1 | Gemma 3 模型的参数计数。我们的词汇表有 256k 条目。

注意力（Luong 等，2015），每个全局层有 5 个局部层的模式，从局部层作为模型的第一层开始。

长上下文。Gemma 3 模型支持 128K 令牌的上下文长度，1B 模型的上下文长度为 32K。我们将全局自注意力层的 RoPE 基频从 10k 提高到 1M，并将局部层的频率保持在 10k。我们遵循与 Chen 等人（2023）类似的位置插值过程，以扩展全局自注意力层的跨度。

2.1. 视觉模态

视觉编码器。我们使用了SigLIP编码器的400M变体（Zhai等，2023），这是一个使用CLIP损失变体（Radford等，2021）训练的视觉变换器（Dosovitskiy，2020）。Gemma视觉编码器的输入为调整大小为896 x 896的正方形图像，并在视觉助手任务的数据上进行微调。为了简化，我们在我们的4B、12B和27B模型中共享视觉编码器，并在训练期间保持其冻结。

平移与扫描（P&S）。Gemma视觉编码器以固定分辨率896 × 896运行。这在处理非方形纵横比和高分辨率图像时会导致伪影，导致文本无法读取或小物体消失。我们在推理过程中通过自适应窗口算法解决了这个问题。该算法将图像分割成不重叠的相等大小的裁剪，覆盖整个图像，并将其调整为896×896像素，以便传递给编码器。此窗口化仅在必要时应用，并控制裁剪的最大数量。这是一种仅在推理时的优化，可以禁用以加快推理速度。

			Shards		
Model	Type	#Chips	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.

	Raw (GB)	Quantized (GB)		
Model	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-

			Shards		
Model	Type	#Chips	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

表 2 | 通过数据、序列（Seq.）和副本进行分片的训练基础设施。

2.2. 预训练

我们遵循与Gemma 2中相似的配方进行知识蒸馏的预训练。

训练数据。我们在比Gemma 2稍大的标记预算上预训练我们的模型，即，我们在Gemma 3 27B上训练14T标记，在12B版本上训练12T，在4B上训练4T，在1B上训练2T标记。标记的增加考虑了在预训练期间使用的图像和文本的混合。我们还增加了多语言数据的数量，以改善语言覆盖率。我们添加了单语和双语数据，并使用受Chung等人（2023）启发的策略来处理语言表示的不平衡。

分词器。我们使用与Gemini 2.0相同的分词器：一个具有分割数字、保留空格和字节级编码的SentencePiece分词器（Kudo和Richardson，2018）。生成的词汇表有262k个条目。这个分词器对非英语语言更为平衡。

过滤。我们使用过滤技术来降低不必要或不安全言论的风险，并移除某些个人信息和其他敏感数据。我们从预训练数据混合中去除评估集的污染，并通过最小化敏感输出的扩散来降低重复的风险。我们还应用了一个质量重加权步骤，灵感来自Sachdeva等人（2024），以减少低质量数据的出现。

蒸馏。我们对每个标记采样256个logits，按教师概率加权。学生通过交叉熵损失学习教师在这些样本中的分布。教师的目标分布对未采样的logits设定为零概率，并进行重新归一化。

	Raw (GB)	Quantized (GB)		
Model	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

表 3 | 原始 (bfloat16) 和量化检查点在权重和 KV 缓存 (+KV) 下的内存占用 (以 GB 为单位) 比较，上下文大小为 32,768，量化为 8 位。

2.3. 量化感知训练

除了原始检查点，我们还提供了不同标准格式的量化版本模型。这些版本是通过对每个模型进行少量步骤的微调获得的，通常为5,000步，使用量化感知训练（QAT）（Jacob等，2018）。我们使用非量化检查点的概率作为目标，并调整数据以匹配预训练和后训练分布。基于最流行的开源量化推理引擎（例如 llama.cpp），我们专注于三种权重表示：每通道 int4、每块 int4 和切换 fp8。在表3中，我们报告了原始和量化模型在每种权重表示下的内存占用情况，包括和不包括KV缓存，针对32k个标记的序列。

2.4. 计算基础设施

我们使用TPUv4、TPUv5e和TPUv5p训练我们的模型，如表2所示。每个模型配置都经过优化，以最小化训练步骤时间。对于视觉编码器，我们预先计算每个图像的嵌入，并直接使用这些嵌入进行训练，这对语言模型的训练没有额外成本。

优化器状态使用ZeRO-3（Ren等，2021）的实现进行分片。对于多pod训练，我们执行数据副本重分配。

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).

2.5. Carbon Footprint

The carbon emissions from pre-training the Gemma 3 models is 1497.13 tCO_2eq. This is estimated based on the hourly energy usage reported from our TPU data centers and scaled to account for the additional energy expended to create and maintain the data center. Google data centers are carbon neutral, achieved through a combination of energy efficiency, renewable energy purchases, and carbon offsets. This carbon neutrality applies to our trainings and the machines running them.

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 to add their respective end token.

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>	

表4 | Gemma IT模型的格式。 在分词后显式添加[BOS]标记，或在分词器中使用add_bos=True选项。 *Do not tokenize the text "[BOS]"*。

通过数据中心网络进行引导，使用Barham等人（2022）的Pathways方法。我们使用Jax（Roberts等人，2023）和Pathways（Barham等人，2022）的“单控制器”编程范式，以及GSPMD分区器（Xu等人，2021）和MegaScale XLA编译器（XLA，2019）。

2.5. 碳足迹

Gemma 3 模型的预训练碳排放量为 1497.13 tCO_2eq。这一估算是基于我们 TPU 数据中心报告的每小时能耗，并进行了调整，以考虑创建和维护数据中心所消耗的额外能源。谷歌数据中心是碳中和的，通过提高能效、购买可再生能源和碳抵消的组合实现。这种碳中和适用于我们的训练和运行这些训练的机器。

3. 指令调优

预训练模型通过改进的后训练方法转变为指令调优模型，与我们之前的方案相比（见表6）。

技术。我们的后训练方法依赖于从大型IT教师那里改进的知识蒸馏版本（Agarwal等，2024；Anil等，2018；Hinton等，2015），以及基于改进版本的RL微调阶段，涉及BOND（Sessa等，2024）、WARM（Ramé等，2024b）和WARP（Ramé等，2024a）。

强化学习目标。我们使用多种奖励函数来提高有用性、数学、编码、推理、遵循指令和多语言能力，同时最小化模型的有害性。这包括从使用人类反馈数据训练的加权平均奖励模型（Ramé et al., 2024b）、代码执行反馈（Gehring et al., 2024）以及解决数学问题的真实奖励（DeepSeek-AI, 2025; Lambert et al., 2024）中学习。

数据过滤。我们仔细优化后训练中使用的数据，以最大化模型性能。我们过滤显示某些个人信息、不安全或有毒模型输出、错误自我识别数据和重复示例的例子。包括鼓励更好上下文归因、模糊和拒绝的子集数据，以最小化幻觉，也提高了事实性指标的表现，而不会降低模型在其他指标上的性能。

[BOS] 令牌。对于 PT 和 IT 模型，文本以 [BOS] 令牌开头，需要显式添加，因为文本 “[BOS]” 不映射到 [BOS] 令牌。例如，Flax 有一个选项 add_bos=True，可以在分词时自动添加此令牌。IT 模型的格式示例见表 4。

PT与IT格式化。所有模型共享相同的分词器，部分控制标记专用于IT格式化。一个关键区别是，PT模型在生成结束时输出一个<eos>标记，而IT模型在生成结束时输出一个<end_of_turn>，如表4中所示。因此，微调任一模型类型也需要添加其各自的结束标记。

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*.

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 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 comparisons across models. We include additional evaluations of our models on other benchmarks in the appendix.

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

表5 | Gemma 3 27B IT模型在聊天机器人竞技场中的评估（Chiang等，2024）。所有模型通过人类评审员进行盲法并排评估，相互之间进行比较。每个模型根据Elo评分系统被赋予一个分数。*Gemma-3-27B-IT numbers are preliminary results received on March 8, 2025*。

4. 最终模型的评估

在本节中，我们评估了IT模型在一系列自动基准测试和跨多个领域的人类评估中的表现，以及静态基准测试，如MMLU。

4.1. LMSYS 聊天机器人竞技场

在本节中，我们报告了我们的 IT 27B 模型在 LM Sys Chatbot Arena (Chiang et al., 2024) 中的表现，采用人类评审员进行盲测并与其他最先进的模型进行并排评估。我们在表 5 中报告了 Elo 分数。Gemma 3 27B IT (1338) 位于前 10 名最佳模型之中，得分高于其他非思考型开放模型，如 DeepSeek-V3 (1318)、LLaMA 3 405B (1257) 和 Qwen2.5-70B (1257)，这些模型的规模要大得多。最后，Gemma 3 的 Elo 显著高于 Gemma 2，达到了 1220。请注意，Elo 分数不考虑视觉能力，而上述模型都没有。

4.2. 标准基准测试

在表6中，我们展示了最终模型在各种基准测试中的表现，与我们之前的模型迭代和Gemini 1.5进行比较。我们不直接与外部模型进行比较，因为这些模型通常报告自己的评估设置，在我们的设置中运行它们并不能保证公平的比较。我们鼓励读者关注第三方静态排行榜，以便在模型之间进行更公平的比较。我们在附录中包含了我们模型在其他基准测试上的额外评估。

	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	74.0	81.0	83.4	86.5	33.0	63.4	62.3	29.9	46.9	65.2	72.1
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.

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, 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.

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

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 window 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.

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

	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	74.0	81.0	83.4	86.5	33.0	63.4	62.3	29.9	46.9	65.2	72.1
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

T表 6 | 微调（IT）模型与 Gemini 1.5、Gemini 2.0 的性能比较
Gemma 2 在不同能力的零样本基准测试中。

5. 消融实验

在本节中，我们关注架构变化的影响，以及该模型中新增加的一些视觉能力。

5.1. 预训练能力探测

我们在预训练期间使用几个标准基准作为探针，以确保我们的模型捕捉到一般能力。在图2中，我们比较了Gemma 2和3在这些一般能力（即科学、代码、事实性、多语言性、推理和视觉）上的预训练模型的质量。这些图中使用的不同公共基准的性能细节总结在附录中。总体而言，我们看到新版本在大多数类别中都有所改善，尽管增加了视觉。我们特别关注这一版本中的多语言性，这直接影响了我们模型的质量。然而，尽管使用了去污染技术，这些探针仍然存在被污染的风险（Mirzadeh等，2024），使得更明确的结论更难以评估。

5.2. 本地:全局注意力层

我们测量了在推理过程中对局部和全局自注意力层的变化对性能和内存消耗的影响。

局部:全局比率。在图3中，我们比较了局部与全局注意力层的不同比率。Gemma 2模型中使用1:1，而Gemma 3中使用5:1。我们观察到改变这个比率对困惑度的影响很小。

滑动窗口大小。在图4中，我们比较了不同全局：局部比率配置下局部注意力层的不同滑动窗口大小。滑动窗口可以显著减小，而不会影响困惑度。

对KV缓存内存的影响。在图5中，我们展示了在32k标记上下文中模型使用的内存与KV缓存之间的平衡。“仅全局”配置是大多数密集模型中使用的标准配置。“1:1, sw=4096”在Gemma 2中使用。我们观察到，“仅全局”配置导致60%的内存开销，而使用1:3和1024的滑动窗口（“sw=1024”）时，这一开销减少到不到15%。在图6中，我们计算了KV缓存使用的内存作为上下文长度的函数，比较了我们的2B架构（L:G=5:1, sw=1024）与“仅全局”2B模型。

5.3. 启用长上下文

与其从头开始使用128K序列进行训练，我们先用32K序列对模型进行预训练，然后在预训练结束时将4B、12B和27B模型扩展到128K个标记。

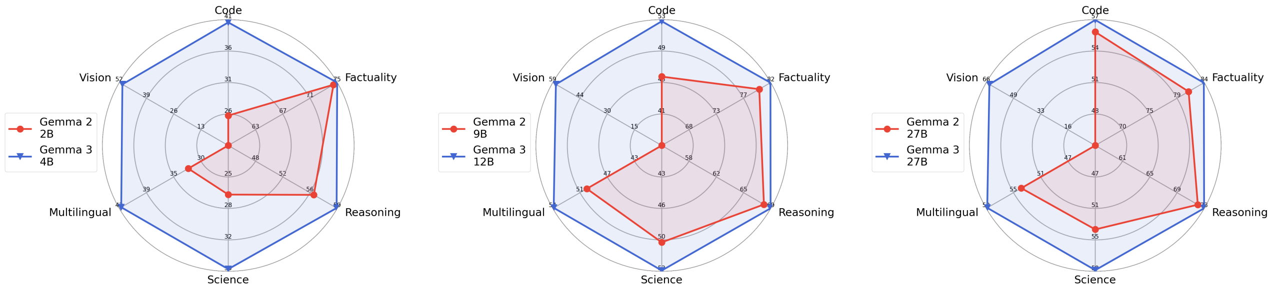

Figure 2 | Summary of the performance of different pre-trained models from Gemma 2 and 3 across general abilities. This plots are meant to give an simplified summary and details are in the appendix.

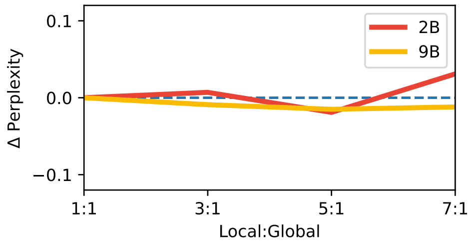

Figure 3 | **Impact of Local:Global ratio** on the perplexity on a validation set. The impact is minimal, even with 7-to-1 local to global. This ablation is run with text-only models.

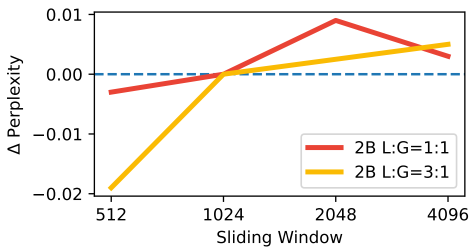

Figure 4 | **Impact of Sliding Window** size on perplexity measured on a validation set. We consider 2 2B models, with 1:1 and 1:3 local to global layer ratios. This ablation is run with text-only models.

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.

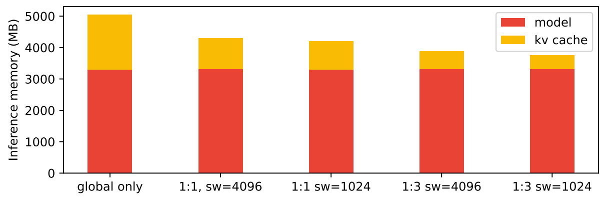

Figure 5 | **Model versus KV cache memory** during inference with a pre-fill KV cache of size 32k. We consider a 2B model with different local to global ratios and sliding window sizes (sw). We compare to global only, which is the standard used in Gemma 1 and Llama. This ablation is run with a text-only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

5.4. Small versus large teacher

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

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 resolu-

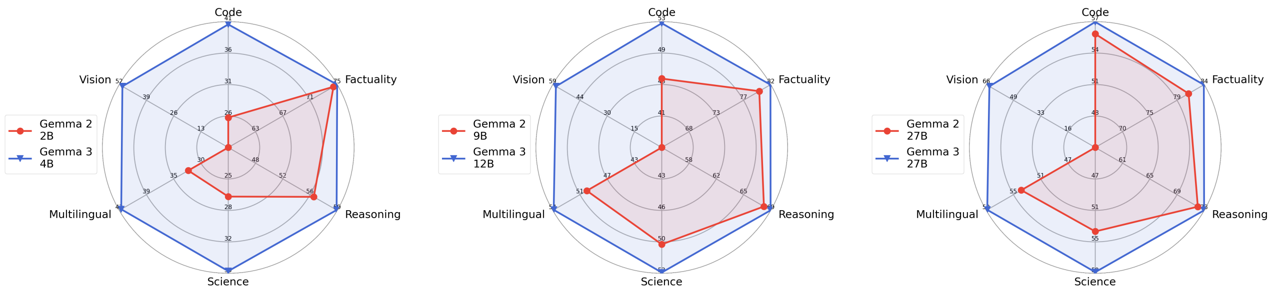

图 2 | 不同预训练模型在 Gemma 2 和 3 中的整体能力表现总结。此图旨在提供简化的总结，详细信息见附录。

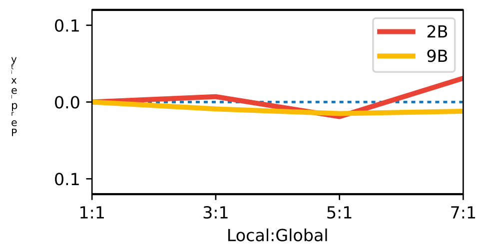

图3 | 本地：全球比率对验证集困惑度的影响。影响很小，即使是7比1的本地与全球比率。此消融实验使用仅文本模型进行。

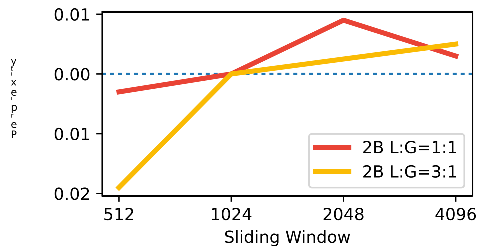

图4 | 滑动窗口大小对在验证集上测量的困惑度的影响。我们考虑2 2B模型，具有1:1和1:3的局部到全局层比例。此消融实验使用仅文本模型进行。

在重新缩放RoPE（Chen等，2023）时，我们发现8的缩放因子在实践中效果良好。请注意，与Gemma 2相比，我们还将全局自注意力层的RoPE基频从10k提高到1M，同时将局部自注意力层保持在10k。在图7中，我们展示了不同上下文长度对困惑度的影响。我们的模型可以推广到128K，但随着我们继续扩展，性能迅速下降。

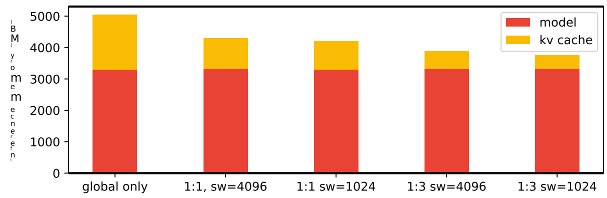

图5 | 模型与 KV 缓存内存在推理期间的对比，使用大小为 32k 的预填充 KV 缓存。我们考虑一个具有不同本地到全局比率和滑动窗口大小 (sw) 的 2B 模型。我们与仅使用全局的标准进行比较，这也是 Gemma 1 和 Llama 中使用的标准。此消融实验使用的是仅文本模型。

5.4. 小型教师与大型教师

一个常见的发现是，为了训练一个小模型，从一个较小的教师中提取知识是更可取的。我们怀疑这是因为这些研究通常是在使用较差教师的正则化效果超过使用较好教师的好处的情况下进行的。我们用两个不同规模的教师训练一个学生，一个大教师和一个小教师，针对不同的训练时间。在图8中，我们观察到对于短训练时间，较小的教师效果更好，但对于较长的训练时间，趋势则相反。

5.5. 视觉编码器

图像分辨率的影响。我们使用基于SigLIP（Zhai等，2023）的视觉编码器。视觉编码器被冻结，仅训练语言模型。该多模态数据中的每个图像由来自相应视觉编码器的256个图像标记表示。更高的分辨率

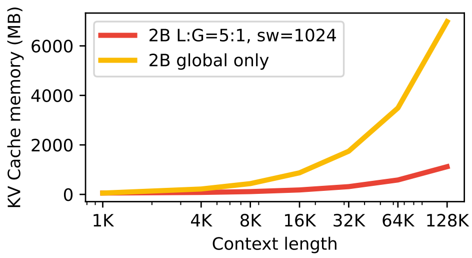

Figure 6 | **KV cache memory versus context length.** We show the memory usage of the KV cache for our architecture (L:G=5:1, sw=1024) and a transformer with global attention only – as used in LLaMa or Gemma 1.

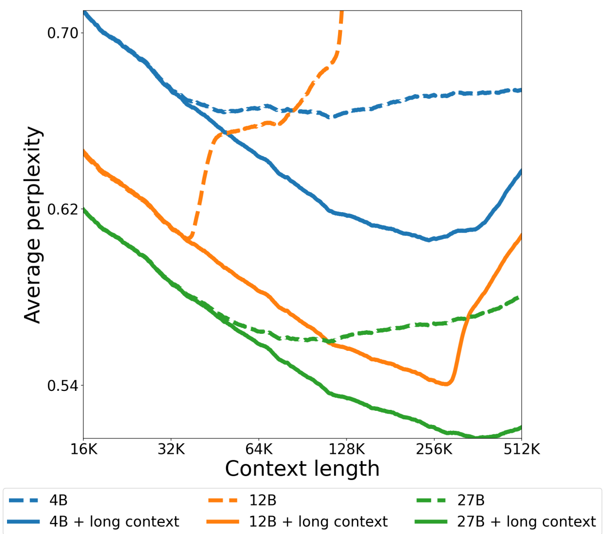

Figure 7 | **Long context** performance of pre-trained models before and after RoPE rescaling.

tion 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 than smaller ones.

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.

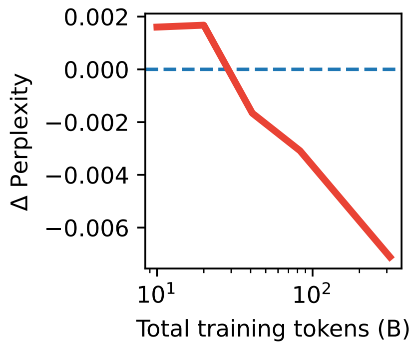

Figure 8 | **Small versus large teacher.** Relative difference of perplexity when using a small and large teacher as a function of the token size of training. Smaller numbers means distilling from a larger teacher is better.

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.

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; Dubey et al., 2024; Gemini Team, 2023, 2024; Gemma Team, 2024a,b). 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 fol-

[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."

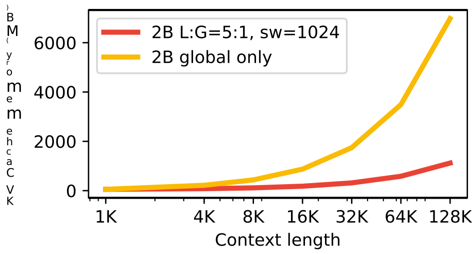

图6 | KV缓存内存与上下文长度的关系。我们展示了我们架构（L:G=5:1, sw=1024）和仅使用全局注意力的变换器（如LLaMa或Gemma 1）的KV缓存的内存使用情况。

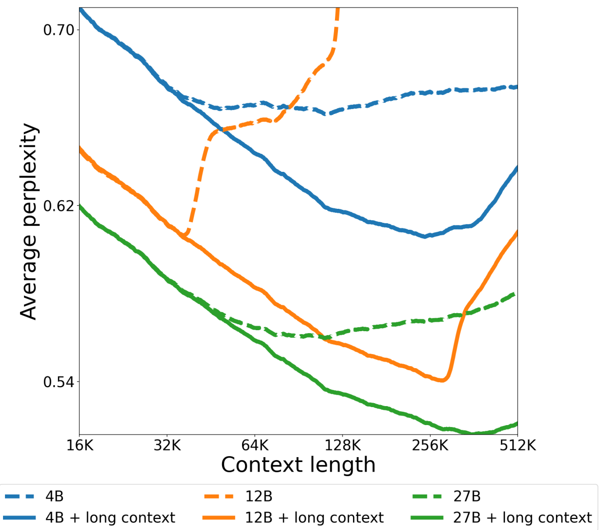

图7 | 预训练模型在RoPE重新缩放前后的长上下文性能。

编码器因此使用平均池化将其输出减少到256个标记。例如，896分辨率的编码器在其输出上进行了4x4的平均池化。如表7所示，高分辨率编码器的性能优于较小的编码器。

平移与扫描。P&S 使得以接近其原生宽高比和图像分辨率捕捉图像成为可能。在表 8 中，我们比较了我们的 27B IT 模型在有和没有 P&S 的情况下的表现。如预期的那样，以接近原生分辨率处理图像的能力极大地帮助了需要在图像上读取文本的任务，这对于视觉语言模型尤其重要。

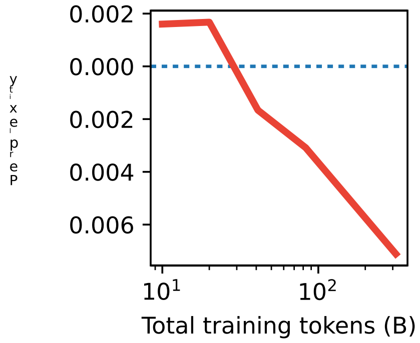

图8 | 小教师与大教师。使用小教师和大教师时，相对困惑度的差异，作为训练的标记大小的函数。较小的数字意味着从较大教师中提取更好。

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

表 7 | 图像编码器输入分辨率的影响。我们使用短期计划 2B Gemma 模型在一些评估基准上测量性能，以观察输入图像分辨率对视觉编码器预训练的影响。

6. 记忆与隐私

大型语言模型可能会生成与训练中使用的某些文本几乎相同的副本（Biderman et al., 2023; Carlini et al., 2021, 2022; Ippolito et al., 2022; Nasr et al., 2023）。几份先前的报告发布了审计，量化了通过测量记忆率来评估这一风险（Anil et al., 2023; Chowdhery et al., 2022; Dubey et al., 2024; Gemini Team, 2023, 2024; Gemma Team, 2024a,b）。这个“记忆率”[1] 被定义为模型生成的与其训练数据匹配的生成数量与所有模型生成数量的比率，使用以下设置。我们遵循-

[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."

	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.

low the methodology described in Gemma Team (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

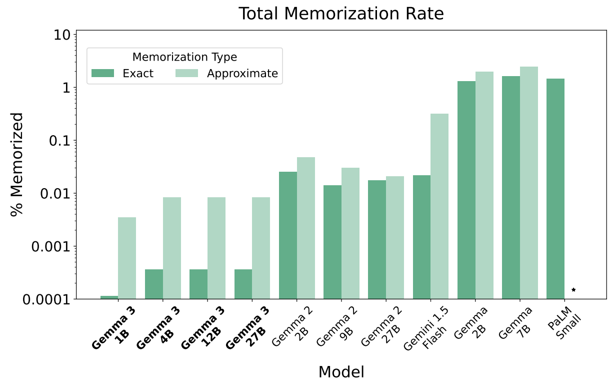

Figure 9 | Total memorization rates for both exact and approximate memorization. Gemma 3 models memorize significantly less than all prior models. *No results for approximate memorization on these models.

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

	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)

表8 | P&S的影响。在验证集上的4次评估结果，分别在预训练检查点上有无P&S。提升体现在与具有不同纵横比的图像相关的任务，或涉及在图像上阅读文本的任务。

低于Gemma团队（2024b）中描述的方法来测量它。具体而言，我们从不同语料库中均匀分布的大部分训练数据中进行子采样，并使用长度为50的前缀和长度为50的后缀测试可发现的提取（Nasr等，2023）该内容。如果续集中的所有标记与源后缀匹配，则我们将文本称为“完全记忆”；如果它们的匹配编辑距离达到10%，则称为“近似记忆”。

图9比较了Gemma和Gemini模型的记忆率；这些模型按时间倒序排列，最新的Gemma 3模型在左侧。我们发现Gemma 3模型对长文本的记忆率远低于之前的模型（注意对数y轴）。我们观察到4B、12B和27B模型之间的记忆率仅有微小差异，而1B模型的记忆率低于这些更大的模型。此外，我们发现更大比例的文本被标记为大致记忆，相较于精确记忆，近似记忆的相对增加约为24倍。

我们还研究了各代可能包含个人信息的速率。为了识别潜在的个人信息，我们使用谷歌云敏感数据保护（SDP）服务。[2] SDP使用广泛的检测规则来识别可能包含个人信息的文本。SDP旨在具有高召回率，并且不

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

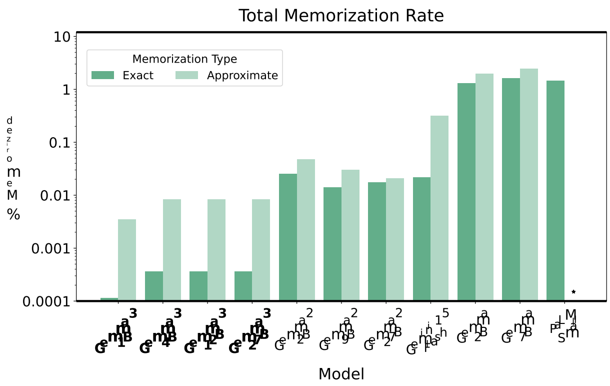

图9 | 精确和近似记忆的总记忆率。Gemma 3 模型的记忆量显著低于所有先前的模型。*这些模型没有近似记忆的结果。

考虑信息可能出现的上下文，这导致了许多误报。因此，我们可能高估了被分类为记忆的输出中潜在个人信息的真实数量。SDP还提供了广泛的严重性等级：低、中和高。如果SDP在任何严重性等级上将文本分类为个人信息，我们就将其归类为个人信息。我们在所有Gemma 3模型的记忆特征输出中没有观察到个人信息。这表明在被分类为记忆的输出中，个人数据的比例很低，低于我们的检测阈值。

7. 责任、安全、保安

责任、安全和保障在Gemma模型的开发中至关重要。为了降低Gemma 3用户的风险，我们继续整合增强的内部安全流程，这些流程贯穿整个开发工作流程，符合最近的Google AI模型（Gemini团队，2024）。这集中于训练时的安全缓解，以及我们引入的新图像到文本能力的稳健和透明的模型评估。

7.1. 治理与评估

我们评估Gemma的好处和风险的方法反映了为Gemma概述的方法。

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

1 (Gemma Team, 2024a)，考虑到支持的模式的变化。我们仍然相信，人工智能的开放性可以将这些技术的好处传播到整个社会，但必须与可能在个人和机构层面造成伤害的恶意使用风险进行评估（Weidinger et al., 2021）。自首次Gemma发布以来，我们已经看到这些模型推动了一些社会有益的应用，例如我们自己的ShieldGemma 2，这是一个基于Gemma 3构建的4B图像安全分类器，为图像安全提供现成的解决方案，输出危险内容、色情和暴力类别的安全标签。

发布Gemma 3模型需要特别关注模型能力的变化，并密切监测现有多模态LLM（Lin et al., 2024）不断演变的风险，以及了解模型在实际使用中的方式。尽管我们尚未收到关于Gemma的恶意使用报告，但我们仍然致力于调查任何此类报告，并与学术界和开发者社区合作，同时进行我们自己的监测，以标记此类案例。

尽管能力有所提升，我们相信，考虑到可用的大型强大开放模型的数量，此次发布对整体风险格局的影响将微乎其微。

7.2. 安全政策和列车运行时间的缓解措施

Gemma安全方法的一个关键支柱是将精细调整的模型与谷歌的安全政策对齐，符合Gemini模型（Gemini团队，2023）。它们旨在帮助防止我们的模型生成有害内容，即，

- 儿童性虐待和剥削
- 揭露可能导致伤害的个人身份信息（例如，社会安全号码）
- 仇恨言论和骚扰
- 危险或恶意内容（包括促进自残或指导有害活动）
- 性暗示内容
- 与科学或医学共识相悖的医疗建议

我们对预训练数据进行了大量的安全过滤，以减少我们的预训练和微调检查点产生有害内容的可能性。对于微调模型，我们还使用了SFT和RLHF来引导模型远离不良行为。

7.3. 保障评估

我们还通过一系列基线保障评估来运行我们的IT模型，以了解我们的模型可能造成的潜在危害。作为开放模型的倡导者，我们也认识到权重发布的不可逆性要求进行严格的风险评估。我们的内部安全流程是相应设计的，对于之前的Gemma模型，我们还进行了与极端风险相关的能力评估（Phuong等，2024；Shevlane等，2023）。随着我们继续开发和分享开放模型，我们将遵循这样一个启发式原则：彻底评估一个更强大的模型通常能为较弱的模型提供足够的保障。因此，我们为Gemma 3优先考虑了一套简化的评估，将深入的危险能力评估保留给可能呈现潜在更高风险的特定模型（如下文所述的CBRN评估）。我们在开发速度与针对性的安全测试之间取得平衡，确保我们的评估聚焦且高效，同时遵循我们在前沿安全框架中提出的承诺。

Baseline Evaluations

基线保证捕捉了安全政策的模型违规率，使用大量合成对抗用户查询和人工评估者对答案进行标记，判断其是否违反政策。总体而言，Gemma 3在这些安全政策上的违规率显著较低。

Chemical, Biological, Radiological and Nuclear (CBRN) knowledge

由于在STEM相关任务上的表现提升，我们使用内部数据集评估了与生物、辐射和核风险相关的知识，该数据集包含封闭式的基于知识的多项选择题。对于评估

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.

在化学知识方面，我们采用了Macknight等人开发的关于化学危害的封闭式知识基础方法。我们的评估表明，Gemma 3模型在这些领域的知识水平较低。

7.4. 我们对负责任的开放模型的看法

设计安全、可靠和负责任的应用程序需要一种系统级的方法，旨在减轻与每个特定用例和环境相关的风险。我们将继续采用与我们模型潜在风险相称的评估和安全缓解措施，并且只有在我们确信收益显著超过可预见风险时，才会与社区分享这些信息。

8. 讨论与结论

在这项工作中，我们介绍了Gemma 3，这是Gemma家族中最新的开放语言模型，适用于文本、图像和代码。在这个版本中，我们专注于增加图像理解和长上下文，同时提高多语言能力和STEM相关能力。我们的模型大小和架构旨在与标准硬件兼容，并且我们大多数架构改进都是为了适应这些硬件，同时保持性能。

参考文献

真实世界问答。
`https://x.ai/news/grok-1.5v`。

M. Acharya, K. Kafle 和 C. Kanan. Tallyqa: 回答复杂的计数问题. 在 *AAAI*, 2018. R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist 和 O. Bachem. 基于策略的语言模型蒸馏: 从自生成的错误中学习. 在 *ICLR*, 2024. J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón 和 S. Sanghai. Gqa: 从多头检查点训练广义多查询变换器模型. *arXiv preprint arXiv:2305.13245*, 2023.

R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl 和 G. E. Hinton. 通过在线蒸馏进行大规模分布式神经网络训练.
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 等. Palm 2 技术报告. *arXiv preprint arXiv:2305.10403*, 2023. M. Artetxe, S. Ruder 和 D. Yogatama. 关于单语表示的跨语言可转移性. 在 *ACL*, 2020. A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi 和 H. Hajishirzi. Xor qa: 跨语言开放检索问答.
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 和 C. Sutton. 使用大型语言模型进行程序合成. *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 和 Y. Wu. Pathways: 用于机器学习的异步分布式数据流, 2022. I. Beltagy, M. E. Peters 和 A. Cohan. Longformer: 长文档变换器. *arXiv preprint arXiv:2004.05150*, 2020. S. Biderman, U. Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit 和 E. Raff. 大型语言模型中的新兴和可预测记忆. *NeurIPS*, 36: 28072–28090, 2023. Y. Bisk, R. Zellers, R. L. Bras, J. Gao 和 Y. Choi. PIQA: 关于自然语言中的物理常识推理. *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 等.
从大型语言模型中提取训练数据. 在 *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 comprehen-

N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, 和 C. Zhang. 量化神经语言模型中的记忆化。*arXiv preprint arXiv:2202.07646*，2022年。Chameleon团队。Chameleon：混合模态早期融合基础模型。*arXiv preprint arXiv:2405.09818*，2024年。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, 和 W. Zaremba. 评估在代码上训练的大型语言模型。*CoRR*，abs/2107.03374，2021年。S. Chen, S. Wong, L. Chen, 和 Y. Tian. 通过位置插值扩展大型语言模型的上下文窗口。*arXiv preprint arXiv:2306.15595*，2023年。X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, 和 C. L. Zitnick. Microsoft coco captions：数据收集和评估服务器。*ArXiv*，abs/1504.00325，2015年。W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, 和 I. Stoica. 聊天机器人竞技场：一个通过人类偏好评估大型语言模型的开放平台，2024年。F. Chollet. 关于智能的衡量。*arXiv preprint arXiv:1911.01547*，2019年。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, 和 N. Fiedel。Palm：通过路径扩展语言建模，2022年。H. W. Chung, N. Constant, X. Garcia, A. Roberts, Y. Tay, S. Narang, 和 O. Firat。Unimax：更公平和更有效的大规模多语言预训练语言采样，2023年。C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, 和 K. Toutanova。Boolq：探索自然是/否问题的惊人难度。*CoRR*，abs/1905.10044，2019年。K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, 和 J. Schulman。训练验证器解决数学文字问题。*CoRR*，abs/2110.14168，2021年。DeepSeek-AI。Deepseek-r1：激励推理学习，2025年。M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, 等等。将视觉变换器扩展到220亿参数。在*ICML*，2023年。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, 和 M. Freitag。Wmt24++：将wmt24的语言覆盖范围扩展到55种语言和方言，2025年。A. Dosovitskiy。一张图像价值16x16个单词：用于大规模图像识别的变换器。*arXiv preprint arXiv:2010.11929*，2020年。D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, 和 M. Gardner。DROP：阅读理解。

sion benchmark requiring discrete reasoning over paragraphs. In *ACL*, 2019.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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.

sion基准需要对段落进行离散推理。在 *ACL*，2019年。A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan 等人。llama 3 模型群。*arXiv preprint arXiv:2407.21783*，2024年。B. Fatemi, M. Kazemi, A. Tsitsulin, K. Malkan, J. Yim, J. Palowitch, S. Seo, J. Halcrow 和 B. Perozzi。时间的考验：评估 llms 在时间推理上的基准。*arXiv preprint arXiv:2406.09170*，2024年。X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma 和 R. Krishna。Blink：多模态大型语言模型可以看见但无法感知。*ArXiv*，abs/2404.12390，2024年。J. Gehring, K. Zheng, J. Copet, V. Mella, T. Cohen 和 G. Synnaeve。Rlef：通过强化学习将代码 llms 基于执行反馈进行基础化。*arXiv preprint arXiv:2410.02089*，2024年。

双子团队。双子：一系列高能力的多模态模型，2023。

双子团队。双子1.5：解锁跨越数百万个上下文标记的多模态理解，2024。Gemma团队。Gemma：基于双子研究和技术的开放模型，2024a。Gemma团队。Gemma 2：在实用规模上改进开放语言模型。*arXiv preprint arXiv:2408.00118*，2024b。O. Goldman, U. Shaham, D. Malkin, S. Eiger, A. Hassidim, Y. Matias, J. Maynez, A. M. Gi- lady, J. Riesa, S. Rijhwani, L. Rimell, I. Szpektor, R. Tsarfaty, 和 M. Eyal。Eclektic：用于评估跨语言知识转移的新挑战集，2025。N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, 和 A. Fan。用于低资源和多语言机器翻译的flores-101评估基准。*ACL*，2022。

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, 和 D. Parikh. 让 VQA 中的 V 变得重要：提升图像理解在视觉问答中的作用。在 *CVPR*，2017。

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, 和 J. Steinhardt. 测量大规模多任务语言理解. *CoRR*, abs/2009.03300, 2020.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, 和 J. Steinhardt. 使用数学数据集测量数学问题解决能力。 *NeurIPS*, 2021.

J. Hessel, A. Marasovi , J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, 和 Y. Choi. 安卓会嘲笑电羊吗？来自《纽约客》标题比赛的幽默“理解”基准。*arXiv preprint arXiv:2209.06293*, 2022.

G. Hinton, O. Vinyals 和 J. Dean. 提炼神经网络中的知识. *arXiv preprint arXiv:1503.02531*, 2015.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, 和 B. Ginsburg. Ruler: 你的长上下文语言模型的真实上下文大小是多少? *arXiv preprint arXiv:2404.06654*, 2024.

D. Ippolito, F. Tramèr, M. Nasr, C. Zhang, M. Jagielski, K. Lee, C. A. Choquette-Choo, 和 N. Carlini. 防止语言模型的逐字记忆会给人一种虚假的隐私感。 *arXiv preprint arXiv:2210.17546*, 2022.

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, 和 D. Kalenichenko. 用于高效整数算术推理的神经网络量化和训练。在 *CVPR*，2018。

M. Joshi, E. Choi, D. S. Weld 和 L. Zettlemoyer. Triviaqa：一个大规模远程监督的阅读理解挑战数据集。*CoRR*，abs/1705.03551，2017。

M. Kazemi, H. Alvari, A. Anand, J. Wu, X. Chen, 和 R. Soricut. Geomverse: 大型模型在几何推理中的系统评估. *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.

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,

M. Kazemi, N. Dikkala, A. Anand, P. Devi , I. Das gupta, F. Liu, B. Fatemi, P. Awasthi, D. Guo, S. Gol lapudi, 和 A. Qureshi. Remi: 一个用于多图像推理的数据集。 *ArXiv*, abs/2406.09175, 2024a.

M. Kazemi, Q. Yuan, D. Bhatia, N. Kim, X. Xu, V. Imbrasaite 和 D. Ramachandran. Boardgameqa: 一个用于自然语言推理与矛盾信息的数据集. *NeurIPS*, 36, 2024b.

M. Kazemi, B. Fatemi, H. Bansal, J. Palowitch, C. Anastasiou, S. V. Mehta, L. K. Jain, V. Aglietti, D. J indal, P. Chen, 等人。大基准超难。 *arXiv preprint arXiv:2502.19187*，2025年。

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Haj ishirzi 和 A. Farhadi. 一张图胜过十张图片。 *ArXiv*, abs/1603.07396, 2016.

E. K c man, R. Ness, A. Sharma 和 C. Tan. 因果推理与大型语言模型：为因果关系开辟新领域。 *arXiv preprint arXiv:2305.00050*, 2023.

T. Kudo 和 J. Richardson. SentencePiece：一种简单且与语言无关的子词标记器和去标记器，用于神经文本处理。2018。

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Colli ns, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelce y, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le 和 S. Petrov. 自然问题：问答研究的基准。*ACL*，20 19。

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. I vison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dzir i, S. Lyu 等人。T\" ulu 3：推动开放语言模型后训练的前沿。*arXiv preprint arXiv:2411.15124*, 2024。

Z. Lin, J. Cui, X. Liao, 和 X. Wang. Malla: 揭示现实世界大型语言模型集成恶意服务的奥秘, 2024.

H. Liu, C. Li, Q. Wu, 和 Y. J. Lee. 视觉指令调优. *NeurIPS*, 36, 2024.

M. Luong, H. Pham 和 C. D. Manning. 基于注意力的神经机器翻译的有效方法. 2015.

麦克奈特、昂和戈梅斯。个人通讯。

K. Marino, M. Rastegari, A. Farhadi 和 R. Mot-tagh i. Ok-vqa：一个需要外部知识的视觉问答基准。在 *CVPR*，2019。

A. Masry, X. L. Do, J. Q. Tan, S. Joty, 和 E. Hoque. ChartQA：一个关于图表的问答基准，涉及视觉和逻辑推理。ACL, 2022.

M. Mathew, D. Karatzas, R. Manmatha 和 C. V. Ja wahar. Docvqa：一个用于文档图像的视觉问答数据集。*WACV*，2020。

M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Val-veny, 和 C. Jawahar. 信息图表vqa. 在 *WACV*, 2022.

I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, 和 M. Farajtabar. Gsm-symbolic: 理解大型语言模型中数学推理的局限性。 *arXiv preprint arXiv:2410.05229*, 2024.

M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. C ooper, D. Ippolito, C. A. Choquette-Choo, E. Wallac e, F. Tramèr, 和 K. Lee. 从（生产）语言模型中可扩展地提取训练数据。*arXiv preprint arXiv:2311.17035*，2023。

A. Nie, Y. Zhang, A. S. Amdekar, C. Piech, T. B. H ashimoto 和 T. Gerstenberg. Moca: 在因果和道德判断任务中测量人类语言模型的一致性. *NeurIPS*, 36, 2024.

R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, 和 T. Dekel. 教学视频：数到十。 *ICCV*, 202 3.

M. Phuong, M. Aitchison, E. Catt, S. Co- gan, A. Ka skasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Ass ael, S. Hodkinson, H. Howard, T. Lieberum, R. Ku mar, 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

L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Dele- tang, A. Ruoss, S. El-Sayed, S. Brown, A. Dra- gan, R. Shah, A. Dafoe, 和 T. Shevlane. 评估具有危险能力的前沿模型, 2024. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, 等. 从自然语言监督中学习可转移的视觉模型. 在 *ICML*, 页码 8748–8763. PMLR, 2021. A. Ramé, J. Ferret, N. Vieillard, R. Dadashi, L. Hussenot, P.-L. Cedoz, P. G. Sessa, S. Girgin, A. Douillard, 和 O. Bachem. WARP: 权重平均奖励策略的好处, 2024a. A. Ramé, N. Vieillard, L. Hussenot, R. Dadashi, G. Cideron, O. Bachem, 和 J. Ferret. WARM: 权重平均奖励模型的好处. 在 *ICML*, 2024b. D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, 和 S. R. Bow- man. Gpqa: 一个研究生级别的谷歌防护问答基准. *ArXiv*, abs/2311.12022, 2023. J. Ren, S. Rajbhandari, R. Y. Aminabadi, O. Ruwase, S. Yang, M. Zhang, D. Li, 和 Y. He. Zero-offload: 民主化十亿规模模型训练. 在 *USENIX*, 2021. A. Roberts, H. W. Chung, G. Mishra, A. Levskaya, J. Bradbury, D. Andor, S. Narang, B. Lester, C. Gaffney, A. Mohiuddin, 等. 使用 t5x 和 seqio 扩大模型和数据. *JMLR*, 2023. N. Sachdeva, B. Coleman, W.-C. Kang, J. Ni, L. Hong, E. H. Chi, J. Caverlee, J. McAuley, 和 D. Z. Cheng. 如何训练数据高效的 llms. *arXiv preprint arXiv:2402.09668*, 2024. K. Sakaguchi, R. L. Bras, C. Bhagavatula, 和 Y. Choi. WINOGRANDE: 一个大规模的对抗性 Winograd 语法挑战. *CoRR*, abs/1907.10641, 2019. E. Sánchez, B. Alastruey, C. Ropers, P. Stenetorp, M. Artetxe, 和 M. R. Costa-jussà. Linguini: 一个语言无关的语言学基准推理。*arXiv preprint arXiv:2409.12126*，2024。

M. Sap, H. Rashkin, D. Chen, R. L. Bras, 和 Y. Choi. Socialiqa: 关于社会互动的常识推理. *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 和 O. Bachem. Bond: 将 llms 与最佳 distillation 对齐, 2024.

K. Shah, N. Dikkala, X. Wang 和 R. Panigrahy。因果语言模型可以引发逻辑难题的搜索和推理能力。*arXiv preprint arXiv:2409.10502*，2024。

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Mar- chal, M. Anderljung, N. Kolt, L. Ho, D. Sid- darth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, 和 A. Dafoe. 极端风险的模型评估, 2023.

F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Sri-vats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, 和 J. Wei. 语言模型是多语言的链式思维推理者。在 *ICLR*，2023。

A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, 和 M. Rohrbach. 朝着能够阅读的 vqa 模型迈进。在 *CVPR*，2019。

H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, 和 P. Talukdar. Indicgenbench：一个多语言基准，用于评估 llms 在印度语言上的生成能力。*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, 和 S. Hooker. 全球 mmlu：理解和应对

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

多语言评估中的文化和语言偏见，2024b。A. Steiner, A. S. Pinto, M. Tschannen, D. Key- sers, X. Wang, Y. Bitton, A. Gritsenko, M. Min- derer, A. S herbondy, S. Long, S. Qin, R. In- gle, E. Bugliarello , S. Kazemzadeh, T. Mes- nard, I. Alabdulmohsin, L . Beyer, 和 X. Zhai。PaliGemma 2：用于迁移的多功能 VLM 家族。
arXiv preprint arXiv:2412.03555，2024。M. Suz gun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Z hou, 和 J. Wei。挑战大基准任务以及思维链是否能解决它们，2022。G. Tyen, H. Mansoor, P. Che n, T. Mak, 和 V. Cărbune。大型语言模型无法发现推理错误，但可以纠正它们！*arXiv preprint arXiv:2311.08516*，2023。A. Vaswani, N. Shazee r, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L . Kaiser, 和 I. Polo- sukhin。注意力是你所需要的一切。2017。K. Vodrahalli, S. Ontanon, N. Tripur aneni, K. Xu, S. Jain, R. Shivanna, J. Hui, N. Dikkal a, M. Kazemi, B. Fatemi 等。米开朗基罗：通过潜在结构查询进行超越干草堆的长上下文评估。
arXiv preprint arXiv:2409.12640，2024。Y. Wa ng, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang 等。Mmlu-pro：一个更强大且具有挑战性的多任务语言理解基准。在 *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. B rown, W. Hawkins, T. Stepleton, C. Biles, A. Birhan e, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, 和 I. Gabriel。语言模型的伦理和社会风险，2021。C. White, S. Dooley, M. R oberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saiful- lah, S. Naidu 等。Livebench：一个挑战-

无污染的llm基准测试。*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, 等人。大规模变压器训练不稳定性的微型代理。
arXiv preprint arXiv:2309.14322，2023。XLA。Xla：用于 TensorFlow 的优化编译器，2019。网址 `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. Maggio ni, R. Pang, N. Shazeer, S. Wang, T. Wang, Y. Wu, 和 Z. Chen. GSPMD：用于机器学习计算图的通用和可扩展并行化。2021。

Y. Yamada, Y. Bao, A. K. Lampinen, J. Kasai, 和 I. Yildirim. 评估大型语言模型的空间理解能力。*arXiv preprint arXiv:2310.14540*, 2023.

K. Yang, O. Russakovsky 和 J. Deng. Spa- tialsense：一个针对空间关系识别的对抗众包基准。*ICCV*，2019。

X. 岳, Y. 倪, K. 张, T. 郑, R. 刘, G. 张, S. 史蒂文斯, D. 江, W. 任, Y. 孙, C. 魏, B. 余, R. 袁, R. 孙, M. 尹, B. 郑, Z. 杨, Y. 刘, W. 黄, H. 孙, Y. 苏, 和 W. 陈. Mmmu: 一个针对专家 AGI 的大规模多学科多模态理解与推理基准. *CVPR*, 2023.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, 和 Y. Choi. HellaSwag: 机器真的能完成你的句子吗? 在 *ACL*，2019。

X. Zhai, B. Mustafa, A. Kolesnikov 和 L. Beyer. 用于语言图像预训练的 sigmoid 损失. 在 *CVPR*, 2023.

B. Zhang 和 R. Sennrich. 均方根层归一化. 2019.

J. Zhang, L. Jain, Y. Guo, J. Chen, K. L. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. Rogers, K. Jamieson 等人。人工智能中的幽默：大规模

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.

扩展众包偏好和漫画字幕的基准。
arXiv preprint arXiv:2406.10522，2024。

W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wan
g, A. Saied, W. Chen, 和 N. Duan. Agieval: 一个以
人为中心的基准，用于评估基础模型，2023。

Core contributors
Aishwarya Kamath*
Johan Ferret*
Shreya Pathak*
Nino Vieillard*
Ramona Merhej*
Sarah Perrin*
Tatiana Matejovicova*
Alexandre Ramé*
Morgane Rivière*
Louis Rouillard*
Thomas Mesnard*
Geoffrey Cideron*
Jean-bastien Grill*
Sabela Ramos*
Edouard Yvinec*
Michelle Casbon*
Etienne Pot
Ivo Penchev
Gaël Liu
Francesco Visin
Kathleen Kenealy
Lucas Beyer
Xiaohai Zhai
Anton Tsitsulin
Robert Busa-Fekete
Alex Feng
Noveen Sachdeva
Benjamin Coleman
Yi Gao
Basil Mustafa
Iain Barr
Emilio Parisotto
David Tian
Matan Eyal
Colin Cherry
Jan-Thorsten Peter
Danila Sinopalnikov
Surya Bhupatiraju
Rishabh Agarwal
Mehran Kazemi
Dan Malkin
David Vilar
Idan Brusilovsky
Jiaming Luo
Andreas Steiner

Contributors (alphabetical order)
Abe Friesen
Abhanshu Sharma
Abheesht Sharma
Adi Mayrav Gilady
Adrian Goedeckemeyer
Alex Feng
Alexander Kolesnikov
Alexei Bendebury
Alvin Abdagic
Amit Vadi
André Susano Pinto
Anil Das
Ankur Bapna
Antoine Miech
Antoine Yang
Antonia Paterson
Ashish Shenoy
Ayan Chakrabarti
Bilal Piot
Bo Wu
Bobak Shahriari
Bryce Petrini
Charlie Chen
Charline Le Lan
Christopher A. Choquette-Choo
CJ Carey
Cormac Brick
Daniel Deutsch
Danielle Eisenbud
Dee Cattle
Derek Cheng
Dimitris Paparas
Divyashree Shivakumar Sreepathihalli
Doug Reid
Dustin Tran
Dustin Zelle
Eric Noland
Erwin Huizenga
Eugene Kharitonov
Frederick Liu
Gagik Amirkhanyan
Glenn Cameron
Hadi Hashemi
Hanna Klimczak-Plucińska
Harsh Mehta
Harshal Tushar Lehri
Hussein Hazimeh
Ian Ballantyne

* co-first authors.

核心贡献者 Aishwarya Kamath[*] Johan Ferret[*] Shreya Pathak[*] Nino Vieillard[*] Ramona Merhej[*] Sarah Perrin[*] Tatiana Matejovicova[*] Alexandre Ramé[*] Morgane Rivière[*] Louis Rouillard[*] Thomas Mesnard[*] Geoffrey Cideron[*] Jean-bastien Grill[*] Sabela Ramos[*] Edouard Yvinec[*] Michelle Casbon[*] Etienne Pot Ivo Penchev Gaël Liu Francesco Visin Kathleen Kenealy Lucas Beyer Xiaohai Zhai Anton Tsitsulin Robert Busa-Fekete Alex Feng Noveen Sachdeva Benjamin Coleman Yi Gao Basil Mustafa Iain Barr Emilio Parisotto David Tian Matan Eyal Colin Cherry Jan-Thorsten Peter Danila Sinopalnikov Surya Bhupatiraju Rishabh Agarwal Mehran Kazemi Dan Malkin David Vilar Idan Brusilovsky Jiaming Luo Andreas Steiner

贡献者（按字母顺序）
阿贝·弗里森 阿班舒·夏尔马 阿比什特·夏尔马 阿迪·梅拉夫·吉拉迪 阿德里安·戈德克迈耶 亚历克斯·冯 亚历山大·科列斯尼科夫 阿列克谢·本德布里 阿尔文·阿布达吉奇 阿米特·瓦迪 安德烈·苏萨诺·平托 阿尼尔·达斯 安库尔·巴普纳 安托万·米赫 安托万·杨 安东尼亚·帕特森 阿希什·申诺伊 阿扬·查克拉巴尔提 比拉尔·皮奥特 博·吴 博巴克·沙赫里亚里 布莱斯·佩特里尼 查理·陈 查琳·勒·兰 克里斯托弗·A乔奎特-楚 CJ·凯里 科尔马克·布里克 丹尼尔·德意志 丹妮尔·艾森布德 迪·卡特尔 德里克·程 迪米特里斯·帕帕拉斯 迪维亚什里·希瓦库马尔·斯里帕提哈利 道格·里德 达斯汀·陈 达斯汀·泽尔 埃里克·诺兰 埃尔温·惠岑加 尤金·哈里托诺夫 弗雷德里克·刘 加吉克·阿米尔哈尼扬 格伦·卡梅伦 哈迪·哈谢米 汉娜·克林查克-普鲁辛斯卡 哈尔什·梅赫塔 哈尔沙尔·图沙尔·莱赫里 侯赛因·哈齐梅 伊恩·巴兰泰因

[*] co-first authors.

Idan Szpektor
Ivan Nardini
Jean Pouget-Abadie
Jetha Chan
Joe Stanton
John Wieting
Jonathan Lai
Jordi Orbay
Joseph Fernandez
Josh Newlan
Ju-yeong Ji
Jyotinder Singh
Kat Black
Kathy Yu
Kevin Hui
Kiran Vodrahalli
Klaus Greff
Linhai Qiu
Marcella Valentine
Marina Coelho
Marvin Ritter
Matt Hoffman
Matthew Watson
Mayank Chaturvedi
Michael Moynihan
Min Ma
Nabila Babar
Natasha Noy
Nathan Byrd
Nick Roy
Nikola Momchev
Nilay Chauhan
Noveen Sachdeva
Oskar Bunyan
Pankil Botarda
Paul Kishan Rubenstein
Phil Culliton
Philipp Schmid
Pier Giuseppe Sessa
Pingmei Xu
Piotr Stanczyk
Pouya Tafti
Rakesh Shivanna
Ravin Kumar
Renjie Wu
Renke Pan
Reza Rokni
Rob Willoughby
Rohith Vallu
Ryan Mullins
Sammy Jerome
Sara Smoot
Sertan Girgin
Shariq Iqbal
Shashir Reddy
Shruti Sheth
Siim Põder
Sijal Bhatnagar
Sindhu Raghuram Panyam
Sivan Eiger
Susan Zhang
Tianqi Liu
Trevor Yacovone
Tyler Liechty
Uday Kalra
Utku Evci
Vedant Misra
Vincent Roseberry
Vlad Feinberg
Vlad Kolesnikov
Woohyun Han
Woosuk Kwon
Yinlam Chow
Zichuan Wei
Zoltan Egyed

Support
Victor Cotruta
Minh Giang
Phoebe Kirk
Anand Rao
Kat Black
Nabila Babar
Jessica Lo
Erica Moreira
Luiz Gustavo Martins
Omar Sanseviero
Lucas Gonzalez
Zach Gleicher
Tris Warkentin

Sponsors
Vahab Mirrokni
Evan Senter
Eli Collins
Joelle Barral
Zoubin Ghahramani

伊丹·斯佩克托尔 伊万·纳尔迪尼 让·普盖-阿巴迪 杰莎·陈 乔·斯坦顿 约翰·维廷 乔纳森·赖 乔尔迪·奥尔巴伊 约瑟夫·费尔南德斯 乔什·纽兰 朱永志 乔廷德·辛格 卡特·布莱克 凯西·余 凯文·辉基兰·沃德拉哈利 克劳斯·格雷夫 林海·邱 马塞拉·瓦伦丁 玛丽娜·科埃略 马文·里特 马特·霍夫曼 马修·沃森 马扬克·查图尔维迪 迈克尔·莫伊尼汉 闵马 纳比拉·巴巴尔 娜塔莎·诺伊 纳森·伯德 尼克·罗伊 尼古拉·莫姆切夫 尼莱·乔汉 诺维恩·萨赫德瓦 奥斯卡·巴尼安 潘基尔·博塔尔达 保罗·基尚·鲁本斯坦 菲尔·卡利顿 菲利普·施密德 皮尔·朱塞佩·塞萨 平梅·徐 皮奥特·斯坦齐克 普亚·塔夫提 拉凯什·希瓦纳 拉文·库马尔 任杰·吴 任克·潘 雷扎·罗克尼 罗布·威洛比 罗希特·瓦卢

瑞安·穆林斯 萨米·杰罗姆 萨拉·斯穆特 塞尔坦·吉尔金 沙里克·伊克巴尔 沙希尔·雷迪 舒尔提·谢特 西姆·波德尔 西贾尔·巴特纳加尔 辛杜·拉古拉姆 潘亚姆·西万 艾格尔·苏珊 张 天齐·刘 特雷弗·雅科沃内 泰勒·莱赫提 乌代·卡尔拉 乌特库·埃夫奇 维丹特·米斯拉 文森特·罗斯伯里 弗拉德·费因伯格 弗拉德·科列斯尼科夫 伍贤·韩 伍硕·权 银岚·周 子川·魏 佐尔坦·埃吉德

支持
维克多·科特鲁塔 明江 菲比·柯克 阿南德·拉奥 卡特·布莱克 纳比拉·巴巴尔 杰西卡·洛 埃里卡·莫雷拉 卢伊斯·古斯塔沃·马丁斯 奥马尔·桑塞维罗 卢卡斯·冈萨雷斯 扎克·格莱彻 特里斯·瓦肯廷

赞助商
瓦哈布·米罗克尼 埃文·森特 埃利·柯林斯 乔埃尔·巴拉尔 祖宾·加哈拉马尼

Raia Hadsell
D. Sculley
Slav Petrov
Noah Fiedel
Noam Shazeer
Oriol Vinyals
Jeff Dean
Demis Hassabis
Koray Kavukcuoglu
Clement Farabet

Technical advisors
Elena Buchatskaya
Jean-Baptiste Alayrac
Rohan Anil
Dmitry (Dima) Lepikhin
Sebastian Borgeaud
Olivier Bachem

Lead
Armand Joulin

Technical leads
Alek Andreev
Cassidy Hardin
Robert Dadashi
Léonard Hussenot

Raia Hadsell D. Scul
ley Slav Petrov Noa
h Fiedel Noam Shaz
eer Oriol Vinyals Jef
f Dean Demis Hassa
bis Koray Kavukcuo
glu Clement Farabet

技术顾问
埃琳娜·布哈茨卡娅 让-
巴蒂斯特·阿拉伊拉克 罗
汉·阿尼尔 德米特里（迪
玛）·列皮欣 塞巴斯蒂安
·博尔乔 奥利维耶·巴赫
姆

铅阿尔芒·朱林

技术负责人：阿列
克·安德烈耶夫，
卡西迪·哈丁，罗
伯特·达达希，莱
昂纳德·于塞诺

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	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),

附录

预训练性能的详细信息。

	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**

表9 | 事实性、常识表现和预训练阶段后的推理
。

事实性和常识。在表9中，我们报告了我们新的预训练基准与之前版本的性能比较。我们考虑了几个标准基准，即HellaSwag（Zellers等，2019），BoolQ（Clark等，2019），PIQA（Bisk等，2019），SIQA（Sap等，2019），TriviaQA（Joshi等，2017），自然问题（Kwiatkowski等，2019），ARC-C和ARC-E（Chollet，2019），WinoGrande（Sakaguchi等，2019），BBH（Suzgun等，2022），DROP（Dua等，2019）。评估细节在表19中描述。总体而言，我们的模型与Gemma 2处于同一水平，这令人鼓舞，因为这些能力并不是本版本改进的重点。

STEM 和代码。我们在 STEM 和代码上的表现细节见表 10。我们考虑了几个标准基准，即 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)。评估细节在表 19 中描述。总体而言，我们看到 STEM 能力有了一致的提升。

	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	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

表10 | STEM和编码在预训练阶段后的表现。

预训练模型。在代码上，我们看到4B和12B模型有类似的改进，但27B模型没有。

	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**

表11 | 预训练阶段后的多模态性能。分数是在每个数据集的验证分割上，未使用P&S。

图像理解。在表11中，我们报告了使用视觉编码器训练的不同模型在各种视觉问答基准上的表现，具体包括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 transfered 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 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;

AI2D（Kembhavi 等，2016），ChartQA（Masry 等，2022），VQA v2（Goyal 等，2017），BLINK（Fu 等，2024），OK-VQA（Marino 等，2019），TallyQA（Acharya 等，2018），SpatialSense VQA（Yang 等，2019），CountBench VQA（Paiss 等，2023）。评估细节在表 20 中描述。

	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

表12 | 在多模态基准上微调后预训练检查点的性能（不包括P&S）。PaliGemma 2 在前四个基准上以896x896分辨率传输，其他基准则以448x448分辨率传输。

与 PaliGemma 2 的比较。我们根据 Steiner 等人（2024）的协议对多模态 Gemma 3 预训练检查点进行了微调——仅调整学习率，其他转移设置保持不变。表 12 中的结果显示，Gemma 3 在涉及文档理解的基准测试中表现出色，甚至超越了 *larger* PaliGemma 2 变体。请注意，由于视觉编码器中的平均池化，Gemma 3 的 4B 和 12B 模型在相同的 896 x 896 分辨率下，转移成本比 PaliGemma 2 的 9B 和 27B 模型低约 10 倍。Gemma 3 在 AI2D 和 OKVQA 上的表现也更好，但 PaliGemma 2 在 VQAv2 和 COCO 标题上表现略优。

多语言性。在表13中，我们报告了预训练模型在多语言任务上的表现。我们应用了多次提示的上下文学习，并在以下基准上呈现结果：MGSM（Shi等，2023），Global-MMLU-Lite（Singh等，2024b），WMT24++ (Deutsch等，2025)，FLoRes（Goyal

	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**

表13 | 预训练阶段后的多语言性能。IndicGenBench 是表14中报告的基准的平均值。

et al., 2022），XQuAD（Artetxe et al., 2020），ECLeKTic（Goldman et al., 2025），IndicGen-Bench（Singh et al., 2024a），XOR QA（Asai et al., 2020）。评估细节在表19中描述。

	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**

表14 | 预训练阶段后 IndicGenBench 性能详细信息。

长上下文。在表15中，我们报告了预训练和微调模型在长上下文基准测试上的表现。我们包括了在32K和128K序列长度下评估的RULER（Hsieh等，2024）和MRCR（Vodra-halli等，2024）基准测试。

8.1. IT模型的性能

我们在表18中报告了我们IT模型的额外基准测试。请注意，N2C指的是Natural2Code，即Gemini 1.0内部保留的数据集，它使用作者生成的来源，而不是基于网络的信息。BBEH指的是BIG-Bench Extra Hard（Kazemi等，2025），这是一个具有挑战性的LLM推理基准，汇集了多个推理任务（Fatemi等，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 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**

表 15 | 预训练 (PT) 和指令微调 (IT) 模型在不同上下文长度的长上下文基准测试中的表现。

	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**

表16 | 指令微调（IT）模型在多模态基准上的表现。如果未提及，这些结果是在每个数据集的最终测试集上应用P&S后的结果。

Hessel 等人，2022；Kazemi 等人，2023，2024b；K c man 等人，2023；Nie 等人，2024；Sánchez 等人，2024；Shah 等人，2024；Tyen 等人，2023；White 等人，2024；Yamada 等人，2023；Zhang 等人，2024）。ECLeKTic 指的是 Goldman 等人（2025）。我们报告微平均分数。更多评估细节见表 21。

8.2. IT模型在视频理解上的表现

额外的多模态评估。Gemma 3 IT 模型在遵循 Gemini 1.5 的评估协议的情况下，在常见的视觉基准上进行了评估（Gemini 团队，2024）。当激活 P&S 时，结果如表 16 所示。

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

表 17 | 在视觉理解基准上使用 0 次和 16 帧线性间隔的微调（IT）模型的性能。感知测试由旨在展示感知上有趣情况的真实世界视频组成，我们报告在多项选择视频问答基准上的结果，以 top-1 准确率为标准。ActivityNet-QA 报告标准的 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
Global-MMLU	33.0	63.4	62.3	29.9	46.9	65.2	**72.0**
ECLeKTic	4.45	6.00	13.0	1.20	4.20	10.5	**16.2**
WMT24++	37.4	50.2	51.7	28.9	46.8	51.6	**53.4**

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

	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
Global-MMLU	33.0	63.4	62.3	29.9	46.9	65.2	**72.0**
ECLeKTic	4.45	6.00	13.0	1.20	4.20	10.5	**16.2**
WMT24++	37.4	50.2	51.7	28.9	46.8	51.6	**53.4**

表18 | 不同规模的指令微调（IT）模型在更多内部和外部基准上的表现。

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	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	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	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		

表19 | 文本基准的详细信息。Char-Len 代表字符长度标准化，COT 代表思维链提示。

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
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	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.

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

表20 | 视觉基准的详细信息。没有链式思维提示或归一化。

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
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	Accuracy	sampling	0-shot	Yes
ECLeKTic	ECLeKTic score	sampling	0-shot	
WMT24++	CHaRacter-level F-score	sampling	0-shot	

表21 | 指令微调（IT）基准的详细信息。无归一化。