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`.