

INF-MLLM2: High-Resolution Image and Document Understanding

August 19, 2024

1 Introduction

In INF-MLLM2, we have introduced significant updates, particularly in high-resolution processing and document understanding. The key improvements include the following:

- **Dynamic Image Resolution Support:** The model now supports dynamic image resolution with up to 16 image patches. This approach utilizes both a global view and sub-image patches, preserving the image’s aspect ratio and resizing dimensions to the nearest multiple of 336, allowing for resolutions up to 1344x1344 pixels.
- **Enhanced OCR Capabilities:** The model has significantly improved OCR capabilities, enabling robust document parsing, table and formula recognition, document layout analysis, and key information extraction. It also supports document analysis and extraction based on various positional information (e.g., bounding boxes, points, and colors). Additionally, it is equipped to handle Chinese PDF parsing and related document content analysis tasks.
- **Advanced Training Strategies:** We employed a progressive multi-stage training strategy along with an enhanced data mixup strategy tailored for image and document multi-task scenarios.
- **State-of-the-Art Performance:** The model not only achieves state-of-the-art (SOTA) performance on general multimodal benchmarks but also significantly outperforms OCR-free document-specific multimodal large models in various document parsing and question-answering tasks.

2 Method

As illustrated in Figure 1, the input image or document, together with the user prompt, is first tokenized and combined into a unified token sequence. This sequence is then fed into the LLM for next-token prediction training. We employ a Vision Transformer (ViT) to encode the image or document into visual tokens.

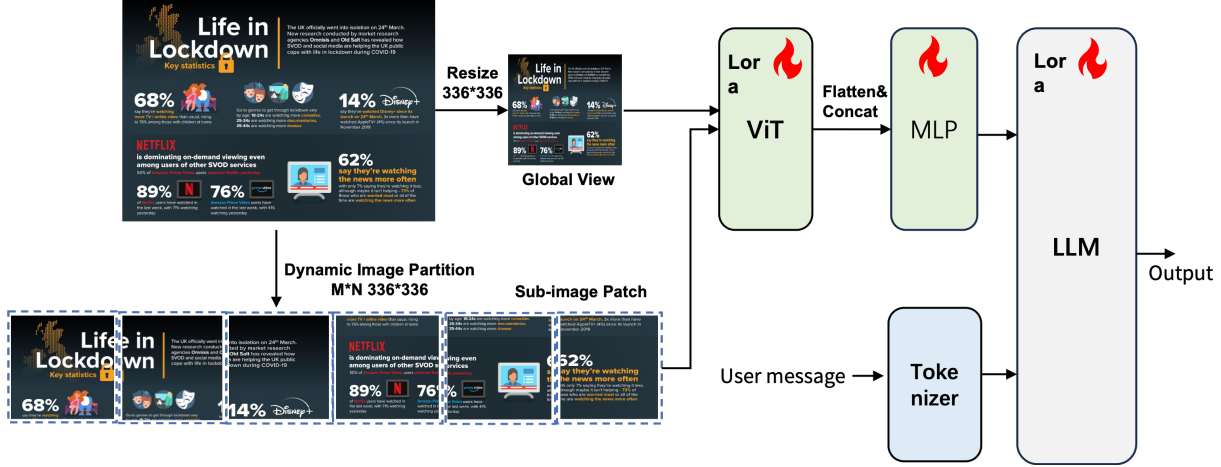


Figure 1: The Architecture of INF-MLLM2.

To accommodate high-resolution images and documents, we propose extracting both global and local visual tokens. For the local visual tokens, the input image or document is dynamically divided into several sub-patches, each with a resolution of 336×336 pixels. These sub-patches are processed independently using the ViT encoder, and the resulting visual tokens are concatenated to form the final set of local visual tokens. For the global visual tokens, the entire input image or document is resized to 336×336 pixels and processed by the ViT encoder to produce the global visual tokens. Finally, it is crucial to note that the global and local visual tokens are concatenated to create the final set of visual tokens.

3 Training

Stage-1	Training data	CC3M/CC12, Laion-115M, CapsFusion-120M	
	Trainable module	MLP	
Stage-2	Training data	Image PDF OCR	ShareGPT4V-1M, LLava-ReCap-3M, ALLaVA-4V-Caption Doc2md-2M(arxiv), Chinese-Doc2md-600k synthdog-en/ch, AnyText-3M, DocOwl-1.5
	Trainable module	ViT+MLP+LLM	
Stage-3	Training data	General multimodal SFT	LLava-1.5-665k, ShareGPT4V-100k, LVIS, ALLava-Instruction, Cambrian
		OCR	OCR, text grounding, OCR VQA
		Doc QA	DocQA, ChartQA, InfoQA, TextVQA
		Doc/Table/Formula parsing	DocQA-GPT4V, Doc2md-334k, im2latex, pubtabnet, Doc2md-110k-chn, MMTab-instruct
	Trainable module	Layout analysis/KIE	CDLA, FUNSD&XFUND, SROIE, WildReceipt, DocLayNet, CROHME
In house dataset		Chinese pdf/table parsing data	

Table 1: Training data and trainable parameters across three stages

INF-MLLM2 undergoes a three-stage training process, starting from coarse pretraining and progressing to SFT (Supervised Fine-Tuning). The specific training data and trainable parameters used in each stage are detailed in Table 1. In this study, we trained a total of three models of varying sizes, ranging from 1.5B to 14B parameters, as illustrated in Table 2.

	ViT model size	LLM model size
INF-MLLM2-1.5B	300 M	1.5B
INF-MLLM2-7B		7B
INF-MLLM2-14B		14B

Table 2: Parameter configurations of INF-MLLM2 models of different sizes.

4 Results

4.1 Comparison with SOTA General multimodal LVM

Model	LLM size	OCR Benchmark					General Benchmark			
		DocVQA [1]	ChartQA [2]	TextVQA [3]	InfoVQA [4]	OCRBench	MMMU [5]	MMBench [6] en	MMBench ch	MME [7]
DocOwl 1.5 [8]	7B	82.2	70.2	68.6	50.7	599	-	-	-	-
TextMonkey [9]	7B	66.7	59.9	64.3	28.6	561	-	-	-	-
Llava next [10]	8B	78.2					41.7			1971.5
CogVLM2 [11]	8B	88.4	74.7	85.0	-	780	42.8			
InternLM-XComposer2 4KHD [12]	7B	90.0	81.0	77.2	68.6	675	39.7	80.2	77.7	2,204.9
MiniCPM-V 2.5 [13]	8B	84.8	-	76.6	-	725	45.8	77.2	74.2	2024.6
INF-MLLM2-1.5B	1.5B	87.30	72.44	73.83	63.79	714.0	37.00	68.90	63.40	1695.62
INF-MLLM2-7B	7B	92.20	81.24	77.17	73.65	772.0	44.60	76.97	75.171	2099.04

Table 3: Comparison with SOTA general multimodal LVM.

Table 3 presents a comparison between our model and general multimodal LLMs across various OCR and general benchmarks.

4.2 Comparison with SOTA OCR-Free LVM for Visual Document Understanding

The comparison with OCR-free multimodal LLMs for visual document understanding is organized as follows, divided into three parts: content parsing of documents, tables, and formulas; key information extraction (KIE); and document-related question answering (QA) tasks.

4.2.1 Document Parsing

The document parsing comparison covers three key aspects: document text extraction, table recognition, and formula recognition. For the document text extraction task (Doc2Markdown), both English and Chinese scenarios are considered, with data sourced from Arxiv and the construction process referencing the Nougat [14] method. The table extraction task involves the end-to-end generation of HTML representations from input table images in both English and Chinese scenarios. The English scenario utilizes the pubTable validation dataset, while the Chinese scenario employs a self-constructed dataset based on financial contexts. The formula recognition task outputs LaTeX representations of input formula images, using the im2latex validation test set. The comparison methods include the specialized document parsing model Nougat and the open-source model Vary [15]. The comparison results are shown in Table 4.

Model	Doc2Markdown		Table2Html		Math2Latex
	pdf2md Eng	pdf2md Chn	PubTableNet Eng	FinTable Chn	im2latex-100k
Nougat-Base [14]	0.1076	0.7296	0.211	-	0.128
Vary-base [15]	0.106	-	0.280	-	0.082
INF-MLLM2-1.5B	0.11209	0.08807	0.0717	0.1023	0.07120
INF-MLLM2-7B	0.09278	0.065	0.0566	0.091	0.0580

Table 4: Comparison of document parsing using Edit Distance as a metric.

4.2.2 Key Information Extraction

For the document key information extraction task, the FUNSD, SROIE, and POIE datasets are used, with Accuracy as the evaluation metric. Among the comparison models, InternLM-XComposer2 represents a state-of-the-art general multimodal large model, while TextMonkey, StructText V3, and LayTextLLM are specialized multimodal large models optimized for document understanding. The comparison results are shown in Table 5.

	FUNSD [16]	SROIE [17]	POIE [18]
InternLM-XComposer2 [12]	15.3	34.2	49.3
TextMonkey [9]	42.9	46.2	32.0
StructText V3	42.4	-	55.9
LayTextLLM [19]	49.4	86.7	66.1
INF-MLLM2-1.5B	63.59	70.89	82.58
INF-MLLM2-7B	70.02	81.55	86.03

Table 5: Performance comparison of key information extraction.

4.2.3 Document-Oriented VQA

The comparison models include chart multimodal large models (ChartLlama, ChartAst), document multimodal large models (DocOwl 1.5, Vary-base, TextMonkey, LayTextLLM, StructText V3), and general multimodal large models (Qwen-VL, InternLM-XComposer2 4KHD, MiniCPM-V 2.5). The results in the table show that our model achieves state-of-the-art (SOTA) performance in several document-oriented VQA tasks.

4.3 Visualization

In this section, we present multi-turn dialogue results involving images and documents generated by our INF-MLLM2 model. The visualizations of these results are provided in Figures 2, 3, and 4.

	LLM Size	DocVQA ANLS	ChartVQA Average	TextVQA	InfoVQA ANLS
ChartLlama [20]	13B	-	69.6	-	-
ChartAst [21]	13B	-	75.1	-	-
Qwen-VL [22]	7B	65.1	65.7	-	35.4
DocOwl 1.5 [8]	7B	82.2	70.2	68.6	50.7
Vary-base [15]	7B	78.1	66.1	-	-
TextMonkey [9]	7B	66.7	59.9	64.3	28.6
InternLM-XComposer2 4KHD [12]	7B	90.0	81.0	77.2	68.6
MiniCPM-V 2.5 [13]	8B	84.8	-	76.6	-
LayTextLLM [19]	7B	77.4	-	-	66.1
StructText v3	1.8B	72.8	70.8	-	28.9
INF-MLLM2-1.5B	1.5B	87.30	72.44	73.83	63.79
INF-MLLM2-7B	7B	92.20	81.24	77.17	73.65

Table 6: Comparison of document-oriented VQA tasks.

References

- [1] M. Mathew, D. Karatzas, and C. V. Jawahar, “Docvqa: A dataset for vqa on document images,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2021. [Online]. Available: <http://dx.doi.org/10.1109/WACV48630.2021.00225>
- [2] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, “Chartqa: A benchmark for question answering about charts with visual and logical reasoning,” 2022.
- [3] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, “Towards vqa models that can read,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2019.00851>
- [4] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar, “Infographicvqa,” in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2022. [Online]. Available: <http://dx.doi.org/10.1109/WACV51458.2022.00264>
- [5] 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,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.16502>
- [6] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, “Mmbench: Is your multi-modal model an all-around player?” 2023.



Q: Please give a detail caption for this image.

INF Mllm2: A cat gazes intently at a bird perched on a wooden stand, set against the backdrop of a frosted window. The interplay of light and shadow accentuates the natural curiosity and potential predator-prey dynamic between the two creatures.

Q: Why this image is interesting ?.

INF Mllm2: This image captures a moment of interaction between a cat and a bird, which is an interesting subject as it highlights the natural curiosity and predatory instincts of both animals. The cat's focused gaze on the bird suggests a moment of tension or anticipation, while the bird's calm demeanor adds a layer of contrast to the scene. The use of sunlight and shadows in the image creates a dramatic effect that enhances the visual appeal and emotional impact of the scene.



Q: Please give a detail caption for this image.

INF Mllm2: The image is a serene natural landscape, likely taken in a mountainous region given the presence of snow-capped peaks and dense forest. The wooden dock serves as a focal point leading the viewer's eye towards the tranquil lake and distant mountains. This could be a location for recreational activities such as fishing or simply enjoying nature. The lighting suggests it may be either early morning or late afternoon, times when the light is softer and more diffused.

Q: What are the things I should be cautious about when I visit here?

INF Mllm2: When visiting such a location, it's important to be cautious about the following:

1. Weather conditions: The mountainous terrain can change rapidly with weather, so check forecasts before your visit.
2. Wildlife: Be aware of local fauna and do not disturb them.
3. Safety on the dock: Ensure you have proper footwear and that the dock is stable before stepping onto it.
4. Environmental impact: Respect the natural environment by leaving no trace, disposing of waste properly, and avoiding harmful activities like littering or damaging vegetation.
5. Navigation: If the lake is deep or if there are currents, ensure you know how to navigate safely.

Figure 2: Visualization of image-based multi-turn dialogue

- [7] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2306.13394>
- [8] A. Hu, H. Xu, J. Ye, M. Yan, L. Zhang, B. Zhang, C. Li, J. Zhang, Q. Jin, F. Huang, and J. Zhou, "mplug-docowl 1.5: Unified structure learning for ocr-free document understanding," 2024.
- [9] Y. Liu, B. Yang, Q. Liu, Z. Li, Z. Ma, S. Zhang, and X. Bai, "Textmonkey: An ocr-free large multimodal model for understanding document," 2024.
- [10] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li, "Llava-next: Stronger llms supercharge multimodal capabilities in the wild," May 2024. [Online]. Available: <https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/>

- [11] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, “Cogvlm: Visual expert for pretrained language models,” 2023.
- [12] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao, W. Zhang, Y. Li, H. Yan, Y. Gao, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang, “Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model,” 2024.
- [13] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, X. Zhang, Z. L. Thai, K. Zhang, C. Wang, Y. Yao, C. Zhao, J. Zhou, J. Cai, Z. Zhai, N. Ding, C. Jia, G. Zeng, D. Li, Z. Liu, and M. Sun, “Minicpm: Unveiling the potential of small language models with scalable training strategies,” 2024.
- [14] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, “Nougat: Neural optical understanding for academic documents,” 2023.
- [15] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, and X. Zhang, “Vary: Scaling up the vision vocabulary for large vision-language models,” 2023.
- [16] G. Jaume, H. Kemal Ekenel, and J.-P. Thiran, “Funsd: A dataset for form understanding in noisy scanned documents,” in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. IEEE, Sep. 2019. [Online]. Available: <http://dx.doi.org/10.1109/ICDARW.2019.10029>
- [17] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, “Icdar2019 competition on scanned receipt ocr and information extraction,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, Sep. 2019. [Online]. Available: <http://dx.doi.org/10.1109/ICDAR.2019.00244>
- [18] J. Kuang, W. Hua, D. Liang, M. Yang, D. Jiang, B. Ren, and X. Bai, “Visual information extraction in the wild: Practical dataset and end-to-end solution,” 2023.
- [19] J. Lu, H. Yu, Y. Wang, Y. Ye, J. Tang, Z. Yang, B. Wu, Q. Liu, H. Feng, H. Wang, H. Liu, and C. Huang, “A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding,” 2024.
- [20] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang, “Chartl-lama: A multimodal llm for chart understanding and generation,” 2023.
- [21] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo, “Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning,” 2024.
- [22] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” 2023.

Benchmark	metric	LLaMA-2		Mistral		Gemma	
		7B	13B	7B	2B	7B	
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	64.3	
HellaSwag	0-shot	77.2	80.7	81.0	71.4	81.2	
PIQA	0-shot	78.8	80.5	82.2	77.3	81.2	
SIQA	0-shot	48.3	50.3	47.0*	49.7	51.8	
Boolq	0-shot	77.4	81.7	83.2*	69.4	83.2	
Winogrande	partial scoring	69.2	72.8	74.2	65.4	72.3	
CQA	7-shot	57.8	67.3	66.3*	65.3	71.3	
OBQA		58.6	57.0	52.2	47.8	52.8	
ARC-e		75.2	77.3	80.5	73.2	81.5	
ARC-c		45.9	49.4	54.9	42.1	53.2	
TriviaQA	5-shot	72.1	79.6	62.5	53.2	63.4	
NQ	5-shot	25.7	31.2	23.2	12.5	23.0	
HumanEval	pass@1	12.8	18.3	26.2	22.0	32.3	
MBPP [†]	3-shot	20.8	30.6	40.2*	29.2	44.4	
GSM8K	maj@1	14.6	28.7	35.4*	17.7	46.4	
MATH	4-shot	2.5	3.9	12.7	11.8	24.3	
AGIEval		29.3	39.1	41.2*	24.2	41.7	
BBH		32.6	39.4	56.1*	35.2	55.1	
Average		47.0	52.2	54.0	44.9	56.4	



Benchmark	metric	LLaMA-2		Mistral		Gemma	
		7B	13B	7B	2B	7B	
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	64.3	
HellaSwag	0-shot	77.2	80.7	81.0	71.4	81.2	
PIQA	0-shot	78.8	80.5	82.2	77.3	81.2	
SIQA	0-shot	48.3	50.3	47.0*	49.7	51.8	
Boolq	0-shot	77.4	81.7	83.2*	69.4	83.2	
Winogrande	partial scoring	69.2	72.8	74.2	65.4	72.3	
CQA	7-shot	57.8	67.3	66.3*	65.3	71.3	
OBQA		58.6	57.0	52.2	47.8	52.8	
ARC-e		75.2	77.3	80.5	73.2	81.5	
ARC-c		45.9	49.4	54.9	42.1	53.2	
TriviaQA	5-shot	72.1	79.6	62.5	53.2	63.4	
NO	5-shot	25.7	31.2	23.2	12.5	23.0	
HumanEval	pass@1	12.8	18.3	26.2	22.0	32.3	
MBPP [†]	3-shot	20.8	30.6	40.2*	29.2	44.4	
GSM8K	maj@1	14.6	28.7	35.4*	17.7	46.4	
MATH	4-shot	2.5	3.9	12.7	11.8	24.3	
AGIEval		29.3	39.1	41.2*	24.2	41.7	
BBH		32.6	39.4	56.1*	35.2	55.1	
Average		47.0	52.2	54.0	44.9	56.4	

Please convert the table into html format.

$$\begin{aligned} \mathbf{v}(\mathbf{x}, t) &= -\frac{1}{2}\beta_t \mathbf{x} + \left(-\frac{1}{2}\beta_t(1 - e^{-\int_0^t \beta_s ds}) - \frac{1}{2}\beta_t e^{-\int_0^t \beta_s ds}\right) \mathbf{s}(\mathbf{x}, t) \\ &= -\frac{1}{2}\beta_t \mathbf{x} - \frac{1}{2}\beta_t \mathbf{s}(\mathbf{x}, t) \end{aligned}$$



```
\begin{array}{l} \{ 1 \} \{ \mathbf{v} \} \{ \mathbf{x} \} \{ t \} = - \frac{1}{2} \{ 1 \} \{ 2 \} \beta_t \{ t \} \mathbf{x} + ( - \frac{1}{2} \{ 1 \} \{ 2 \} \beta_t \{ t \} ( 1 - e^{\wedge \{ - \int_0^t \{ \beta_s \} \{ s \} \} } ) - \frac{1}{2} \{ 1 \} \{ 2 \} \beta_t \{ t \} e^{\wedge \{ - \int_0^t \{ \beta_s \} \{ s \} \} } ) \mathbf{s} \{ \mathbf{x} \} \{ t \} ) \setminus \{ = - \frac{1}{2} \{ 1 \} \{ 2 \} \beta_t \{ t \} \mathbf{x} - \frac{1}{2} \{ 1 \} \{ 2 \} \beta_t \{ t \} \mathbf{s} \{ \mathbf{x} \} \{ t \} \} \end{array}
```

Please convert the formula into latex format.

Google DeepMind

2024-02-21

Gemma: Open Models Based on Gemini Research and Technology

Gemma Team, Google DeepMind¹

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

Abstract

This work introduces Gemma, a family of lightweight, state-of-the-art open models built from the research and technology used to create Gemini models. Gemma models demonstrate strong performance across academic benchmarks for language understanding, reasoning, and safety. We release two sizes of models (2 billion and 7 billion parameters), and provide both pretrained and fine-tuned checkpoints. Gemma outperforms similarly sized open models on 11 out of 18 text-based tasks, and we present comprehensive evaluations of safety and responsibility aspects of the models, alongside a detailed description of model development. We believe the responsible release of LLMs is critical for improving the safety of frontier models, and for enabling the next wave of LLM innovations.

Introduction

We present Gemma, a family of open models based on Google's Gemini models (Gemini Team, 2023).

We trained Gemma models on up to 6T tokens of text, using similar architectures, data, and training recipes as the Gemini model family. Like Gemini, these models achieve strong generalist capabilities in text domains, alongside state-of-the-art understanding and reasoning skills at scale. With this work, we release both pre-trained and fine-tuned checkpoints, as well as an open-source codebase for inference and serving.

Gemma comes in two sizes: a 7 billion parameter model for efficient deployment and development on GPU and TPU, and a 2 billion parameter model for CPU and on-device applications. Each size is designed to address different computational constraints, applications, and developer requirements. At each scale, we release raw, pre-trained checkpoints, as well as checkpoints fine-tuned for dialogue, instruction-following, helpfulness, and safety. We thoroughly evaluate the shortcomings of our models on a suite of quantitative and qualitative benchmarks. We believe the release of both pretrained and fine-tuned checkpoints will enable thorough research and investigation into the impact of current instruction-tuning regimes, as well as the development of increasingly safe and responsible model development methodologies.

Gemma advances state-of-the-art performance relative to comparable-scale (and some larger), open models (Almazrouei et al., 2023; Jiang et al., 2023; Touvron et al., 2023a,b) across a wide range of domains including both automated benchmarks and human evaluation. Example domains include question answering (Clark et al., 2019; Kwiatkowski et al., 2019), commonsense reasoning (Sakaguchi et al., 2019; Sugun et al., 2022), mathematics and science (Cobbe et al., 2021; Hendrycks et al., 2020), and coding (Austin et al., 2021; Chen et al., 2021). See complete details in the Evaluation section.

Like Gemini, Gemma builds on recent work on sequence models (Sutskever et al., 2014) and transformers (Vaswani et al., 2017), deep learning methods based on neural networks (LeCun et al., 2015), and techniques for large-scale training on distributed systems (Barham et al., 2022; Dean et al., 2012; Roberts et al., 2023). Gemma also builds on Google's long history of open models and ecosystems, including Word2Vec (Mikolov et al., 2013), the Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2018), and T5 (Raffel et al., 2019) and TSX (Roberts et al., 2022).

We believe the responsible release of LLMs is critical for improving the safety of frontier models, for ensuring equitable access to this breakthrough technology, for enabling rigorous evaluation and analysis of current techniques, and for enabling

© 2024 Google DeepMind. All rights reserved.



Gemma: Open Models Based on Gemini Research and Technology

Gemma Team, Google DeepMind¹

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

Abstract

This work introduces Gemma, a family of lightweight, state-of-the-art open models built from the research and technology used to create Gemini models. Gemma models demonstrate strong performance across academic benchmarks for language understanding, reasoning, and safety. We release two sizes of models (2 billion and 7 billion parameters), and provide both pretrained and fine-tuned checkpoints. Gemma outperforms similarly sized open models on 11 out of 18 text-based tasks, and we present comprehensive evaluations of safety and responsibility aspects of the models, alongside a detailed description of model development. We believe the responsible release of LLMs is critical for improving the safety of frontier models, and for enabling the next wave of LLM innovations.

Gemma Team, Google DeepMind

Introduction

We present Gemma, a family of open models based on Google's Gemini models (Gemini Team, 2023). We trained Gemma models on up to 6T tokens of text, using similar architectures, data, and training recipes as the Gemini model family. Like Gemini, these models achieve strong generalist capabilities in text domains, alongside state-of-the-art understanding and reasoning skills at scale. With this work, we release both pre-trained and fine-tuned checkpoints, as well as an open-source codebase for inference and serving.

Gemma comes in two sizes: a 7 billion parameter model for efficient deployment and development on GPU and TPU, and a 2 billion parameter model for CPU and on-device applications. Each size is designed to address different computational constraints, applications, and developer requirements. At each scale, we release raw, pre-trained checkpoints, as well as checkpoints fine-tuned for dialogue, instruction-following, helpfulness, and safety. We thoroughly evaluate the shortcomings of our models on a suite of quantitative and qualitative benchmarks. We believe the release of both pretrained and fine-tuned checkpoints will enable thorough research and investigation into the impact of current instruction-tuning regimes, as well as the development of increasingly safe and responsible model development methodologies.

Gemma advances state-of-the-art performance relative to comparable-scale (and some larger), open models (Almazrouei et al., 2023; Jiang et al., 2023; Touvron et al., 2023, 2023a) across a wide range of domains including both automated benchmarks and human evaluation. Example domains include question answering (Clark et al., 2019; Kwiatkowski et al., 2019), commonsense reasoning (Sakaguchi et al., 2019; Sugun et al., 2022), mathematics and science (Cobbe et al., 2021; Hendrycks et al., 2020), and coding (Austin et al., 2021; Chen et al., 2021). See complete details in the Evaluation section.

Like Gemini, Gemma builds on recent work on sequence models (Sutskever et al., 2014) and transformers (Vaswani et al., 2017), deep learning methods based on neural networks (LeCun et al., 2015), and techniques for large-scale training on distributed systems (Barham et al., 2022; Dean et al., 2012; Roberts et al., 2023). Gemma also builds on Google's long history of open models and ecosystems, including Word2Vec (Mikolov et al., 2013), the Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2018), and T5 (Raffel et al., 2019) and TSX (Roberts et al., 2022).

We believe the responsible release of LLMs is critical for improving the safety of frontier models, for ensuring equitable access to this breakthrough technology, for enabling rigorous evaluation and analysis of current techniques, and for enabling

Please convert the document to markdown.

Document + Instruction input

Prediction output

Figure 3: Visualization of document content extraction

Google DeepMind

2024-02-21

Gemma: Open Models Based on Gemini Research and Technology

Gemma Team, Google DeepMind¹

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

This work introduces Gemma, a family of lightweight, state-of-the-art open models built from the research and technology used to create Gemini models. Gemma models demonstrate strong performance across academic benchmarks for language understanding, reasoning, and safety. We release two sizes of models (2 billion and 7 billion parameters), and provide both pre-trained and fine-tuned checkpoints. Gemma outperforms similarly sized open models on 11 out of 18 text-based tasks, and we present comprehensive evaluations of safety and responsibility aspects of the models, alongside a detailed description of model development. We believe the responsible release of LLMs is critical for improving the safety of frontier models, and for enabling the next wave of LLM innovations.

Introduction

We present Gemma, a family of open models based on Google's Gemini models (Gemma Team, 2023).

We trained Gemma models on up to 6T tokens of text, using similar architectures, data, and training recipes as the Gemini model family. Like Gemini, these models achieve strong generalist capabilities in text domains, alongside state-of-the-art understanding and reasoning skills at scale. With this work, we release both pre-trained and fine-tuned checkpoints, as well as an open-source codebase for inference and serving.

Gemma comes in two sizes: a 7 billion parameter model for efficient deployment and development on GPU and TPU, and a 2 billion parameter model for CPU and on-device applications. Each size is designed to address different computational constraints, applications, and developer requirements. At each scale, we release raw, pre-trained checkpoints, as well as checkpoints fine-tuned for dialogue, instruction-following, helpfulness, and safety. We thoroughly evaluate the shortcomings of our models on a suite of quantitative and qualitative benchmarks. We believe the release of both pre-trained and fine-tuned checkpoints will enable thorough research and investigation into the impact of current instruction-tuning regimes, as well as the development of increasingly safe and responsible model development methodologies.

Gemma advances state-of-the-art performance relative to comparable-scale (and some larger), open models (Almazrouei et al., 2023; Jiang et al., 2023; Touvron et al., 2023a,b) across a wide range of domains including both automated benchmarks and human evaluation. Example domains include question answering (Clark et al., 2019; Kwiatkowski et al., 2019), commonsense reasoning (Sakaguchi et al., 2019; Sugum et al., 2022), mathematics and science (Cobbe et al., 2021; Hendrycks et al., 2020), and coding (Austin et al., 2021; Chen et al., 2021). See complete details in the Evaluation section.

Like Gemini, Gemma builds on recent work on sequence models (Sutskever et al., 2014) and transformers (Vaswani et al., 2017), deep learning methods based on neural networks (LeCun et al., 2015), and techniques for large-scale training on distributed systems (Barham et al., 2022; Dean et al., 2012; Roberts et al., 2023). Gemma also builds on Google's long history of open models and ecosystems, including Word2Vec (Mikolov et al., 2013), the Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2018), and T5 (Raffel et al., 2019) and T5X (Roberts et al., 2022).

We believe the responsible release of LLMs is critical for improving the safety of frontier models, for ensuring equitable access to this breakthrough technology, for enabling rigorous evaluation and analysis of current techniques, and for enabling

© 2024 Google DeepMind. All rights reserved

What is the title and topic of this document?

The title of the document is "Gemma: Open Models Based on Gemini Research and Technology," and the topic is about a family of lightweight, state-of-the-art open models built from research and technology used to create Gemini models.

Are there any authors attributed to this work, and if so, list them.

Yes, there are authors attributed to this work. The authors listed are Gemma Team and Google DeepMind.

How many Gemma models released in this work?

Two Gemma models were released in this work.

How does Gemma models perform compared to other models?

Gemma models demonstrate strong performance across academic benchmarks for language understanding, reasoning, and safety. They outperform similarly sized open models on 11 out of 18 text-based tasks and present comprehensive evaluations of safety and responsibility aspects of the models. Gemma models also provide a detailed description of model development and enable the next wave of Large Language Models (LLM) innovations.

Gemma: Open Models Based on Gemini Research and Technology

■ LLaMA 2 (7B) ■ LLaMA 2 (13B) ■ Mistral (7B) ■ Gemma (7B)

Task	LLaMA 2 (7B)	LLaMA 2 (13B)	Mistral (7B)	Gemma (7B)
Question Answering	60	65	62	68
Reasoning	60	65	62	68
Math / Science	30	35	32	38
Coding	20	25	22	28

Figure 1 | Language understanding and generation performance of Gemma 7B across different capabilities compared to similarly sized open models. We group together standard academic benchmark evaluations by capability and average the respective scores; see Table 6 for a detailed breakdown of performance.

the development of the next wave of innovations. While thorough testing of all Gemma models has been conducted, testing cannot cover all applications and scenarios in which Gemma may be used. With this in mind, all Gemma users should conduct rigorous safety testing specific to their use case before deployment or use. More details on our approach to safety can be found in section Responsible Deployment.

In this technical report, we provide a detailed overview of the model architecture, training infrastructure, and pretraining and fine-tuning recipes for Gemma, followed by thorough evaluations of all checkpoints across a wide-variety of quantitative and qualitative benchmarks, as well as both standard academic benchmarks and human-preference evaluations. We then discuss in detail our approach to safe and responsible deployment. Finally, we outline the broader implications of Gemma, its limitations and advantages, and conclusions.

Model Architecture

The Gemma model architecture is based on the transformer decoder (Vaswani et al., 2017). The core parameters of the architecture are summarized in Table 1. Models are trained on a context length of 8192 tokens.

We also utilize several improvements proposed after the original transformer paper. Below, we list the included improvements:

Multi-Query Attention (Shazeer, 2019). No-

Can you describe the figure shown in the image?

The figure in the image is labeled "Figure 1 | Language understanding and generation performance of Gemma 7B across different capabilities compared to similarly sized open models." It is a bar chart comparing the performance of Gemma 7B with other models in terms of language understanding and generation. There are four bars representing different models: Gemma 7B, LLaMA 2 (7B), LLaMA 2 (13B), and Mistral (7B). Each bar is color-coded to represent the model and the bars are labeled with the model names. The x-axis represents the language understanding and generation performance, with a scale from 0 to 100.

what are the feedforward hidden dims of 2B and 7B models respectively.

The feedforward hidden dims of 2B and 7B models are 32768 and 49152 respectively.

List the embedding parameters of 2B and 7B models.

The embedding parameters of model 2B are 524,550,144 and for model 7B, they are 786,825,216.

Figure 4: Visualization of multi-turn dialogue results for documents