

Gemma

Google DeepMind releases Gemma, a series of open language models inspired by the same research and technology used to create Gemini. The Gemma model release includes 2B (trained on 2T tokens) and 7B (trained on 6T tokens) models including base and instruction-tuned checkpoints. The models are trained on a context length of 8192 tokens and generally outperform Llama 2 7B and Mistral 7B models on several benchmarks.

The Gemma model architecture is based on the transformer decoder with improvements including [multi-query attention](#) (used by the 2B model), multi-head attention (used by 7B model), [RoPE embeddings](#), [GeGLU activations](#), and [normalizer location](#).

According to the [technical report](#), Gemma 2B and 7B are trained on 2T and 6T tokens mainly consisting of web documents, mathematics, and code. Unlike Gemini, these models are not explicitly trained to support multilingual or multimodal capabilities. The vocabulary size is 256K tokens and uses a subset of the SentencePiece tokenizer of Gemini, preserves whitespace in splits digits, and relies on byte-level encodings for unknown tokens.

The instruction-tuned models are tuned using supervised fine-tuning on a mix of text-only synthetic and human-generated prompt response pairs and reinforcement learning from human feedback (RLHF) with the reward model trained on labeled preference data and the policy based on a set of high-quality prompts. Note that all the datasets used are English only. As shown in the table below, the instruction-tuned models also use specific formatting control tokens to indicate roles and turns in a conversation.

User: <start_of_turn>user
Knock knock.<end_of_turn>
<start_of_turn>model

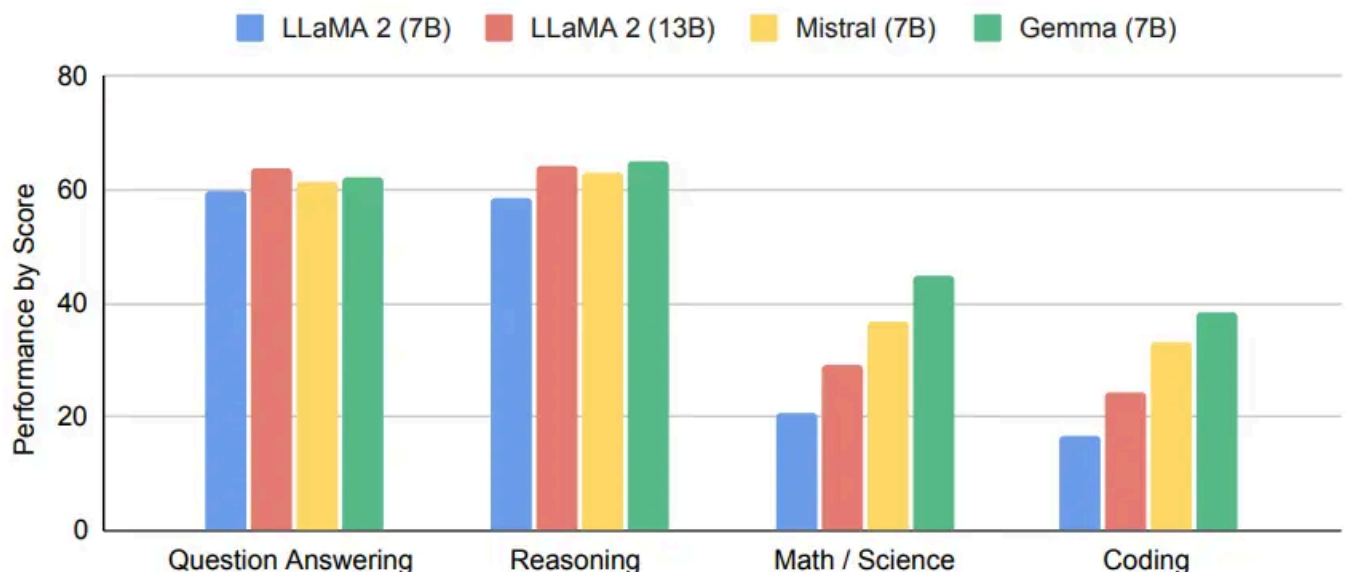
Model: Who's there?<end_of_turn>model

User: <start_of_turn>user
Gemma.<end_of_turn>
<start_of_turn>model

Model: Gemma who?<end_of_turn>model

Results

As shown in the figure below, the Gemma 7B model demonstrates strong performance on math, science, and code-related tasks. The scores correspond to the average scores on academic benchmark evaluations grouped by capability.



Gemma 7B outperforms Llama 2 7B and Mistral 7B on various academic benchmarks with notable performance on HumanEval, GSM8K, MATH, and AGIEval and improved performance on reasoning, dialogue, mathematics, and code.

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

The Gemma 7B instruction tuned models also outperform the Mistral-7B v0.2 Instruct model on safety and instruction following as evaluated by humans.



Gemma is also evaluated on several safety academic benchmarks and compared with Mistral. The technical report also mentions the use of debiasing techniques and red-

teaming to potentially mitigate common risks associated with large language models (LLMs). You can find more information on how to responsibly develop with Gemma in the [model card](#) and [Responsible Generative AI toolkit](#).



Gemma 7B Prompt Format

The Gemma base models don't use any specific prompt format but can be prompted to perform tasks through zero-shot/few-shot prompting. The Gemma Instruct model uses the following format:

```
<start_of_turn>user
Generate a Python function that multiplies two numbers <end_of_turn>
<start_of_turn>model
```

Here is a table showing the relevant formatting control tokens available in Gemma:

Context	Relevant Token
User turn	user
Model turn	model
Start of conversation turn	<start_of_turn>

Context	Relevant Token
End of conversation turn	<end_of_turn>

You can also use the special control tokens in the context of a multi-turn user prompt as follows:

```
<start_of_turn>user
What is a good place for travel in the US?<end_of_turn>
<start_of_turn>model
California.<end_of_turn>
<start_of_turn>user
What can I do in California?<end_of_turn>
<start_of_turn>model
```

How to Prompt Gemma 7B

Prompting Gemma 7B effectively requires being able to use the prompt template properly. In the following examples, we will cover a few examples that demonstrate the use effective use of the prompt template of Gemma 7B Instruct for various tasks.

Zero-shot Prompting

As with any model, you can leverage Gemma's zero-shot capabilities by simply prompting it as follows:

```
<start_of_turn>user
Explain why the sky is blue<end_of_turn>
<start_of_turn>model
```

Zero-shot Prompting with System Prompt

Adding a system role or system prompt helps to steer LLMs better. While there is no explicit system role in Gemma, you can add additional instructions as part of the prompt as follows:

```
<start_of_turn>user
Answer the following question in a concise and informative manner:

Explain why the sky is blue<end_of_turn>
```

```
<start_of_turn>model
```

In the example above, we added "Answer the following question in a concise and informative manner:" as an additional instruction or system prompt to steer the model better.

Role Playing

We can also test Gemma for role-playing capabilities, which is a useful ability to not only personalize model responses but also to build more complex domain-specific [LLM agents](#).

Prompt:

```
<start_of_turn>user
You are a helpful 2nd-grade teacher. Help a 2nd grader to answer questions in a short and
clear manner.
Explain why the sky is blue<end_of_turn>
<start_of_turn>model
```

Model response:

```
Sure, here is an explanation for why the sky is blue:
The sky appears blue because of a phenomenon called Rayleigh Scattering.
Here's the science behind it:
Sunlight is made up of different colors of light, like red, green, blue, and violet.
When sunlight hits the Earth's atmosphere, the different colors of light scatter in different
directions.
Blue light scatters more easily than other colors because it has shorter wavelengths.
The scattered blue light travels in all directions, including towards your eyes.
That's why we see the sky as blue.
It's important to note that this explanation is simplified and does not account for all
scientific details. However, it gives you a good understanding of the basic science behind
why the sky appears blue.
```

Note that we haven't checked the response for accuracy or whether it's suitable for a 2nd grader. However, you can compare the above response with the response obtained from the chain-of-thought prompt that follows.

Reasoning

You can elicit Gemma's reasoning capabilities by using [zero-shot chain-of-thought prompting](#) as follows:

Prompt:

```
<start_of_turn>user
Think and write your step-by-step reasoning before responding.

Explain why the sky is blue.<end_of_turn>
<start_of_turn>model
```

Model response (truncated for brevity):

```
Step 1: Scattering of Light
The sky appears blue due to the phenomenon of scattering of light. When sunlight hits the
Earth's atmosphere, it interacts with the particles of air, primarily nitrogen and oxygen
molecules.
Step 2: Rayleigh Scattering
Rayleigh scattering is the specific scattering mechanism that occurs when light interacts
with particles much smaller than the wavelength of light. In this process, shorter
wavelengths (blue light) are scattered more effectively than longer wavelengths (red light).
Step 3: Blue Scattered Light Travels in All Directions
...
Conclusion
The sky appears blue due to the scattering of light, specifically Rayleigh scattering.
Shorter wavelengths of light are scattered more effectively, and this scattered light travels
in all directions, with the scattered light that travels opposite to the sun's rays being
most visible to our eyes.
```

Something to note about the Gemma model is that it can output markdown format automatically. We have edited the text format a bit for demonstration and simplicity purposes but the content is exactly what the model responded with. Note that we also didn't assess the accuracy of the response or if the model hallucinates.

Resources and Integrations

Here are several resources and integrations that were part of the Gemma release:

- [Colab](#) and [Kaggle](#) notebooks
- [Hugging Face models](#)
- [MaxText](#)
- [NVIDIA NeMo](#)

- [TensorRT-LLM](#)
- Gemma 7B is available in the [NVIDIA AI Playground](#)

According to the official [blog release](#), the [Terms of Use](#) permit responsible commercial usage and distribution for all organizations, regardless of size.

References

- [Gemma: Introducing new state-of-the-art open models](#)
- [Gemma: Open Models Based on Gemini Research and Technology](#)
- [Responsible Generative AI Toolkit](#)
- [Fast Transformer Decoding: One Write-Head is All You Need](#)
- [Roformer: Enhanced transformer with rotary position embedding](#)
- [GLU variants improve transformer](#)
- [Root mean square layer normalization](#)

Last updated on September 19, 2024