Search models, datasets, users...

google / gemma-3-4b-it-qat-q4_0-gguf

like 204

Follow Google 34.6k

Image-Text-to-Text | GGUF | gemma3 | gemma | google | conversational

arxiv:28 papers | License: gemma

Deploy | Use this model

Model card | Files xet | Community 4

Downloads last month
3,514

GGUF

Model size 4B params | Architecture gemma3 | Chat template

Hardware compatibility | Add hardware for estimation

4-bit | Q4_0 3.16 GB

Inference Providers NEW

Image-Text-to-Text

This model isn't deployed by any Inference Provider. | 1 Ask for provider support

Model tree for google/gemma-3-4b-it-qat-q4_0-gguf

Base model google/gemma-3-4b-pt

Finetuned google/gemma-3-4b-it

Quantized (145) ..... this model

Quantizations ..... 1 model

Spaces using google/gemma-3-4b-it-qat-q4_0-gguf 5

NatLibFi/Caption-Annif-Demo DipakKuma/My-Gemma-Chatbot kairusama/gemma-3-4b-it-qat

AlphaPhoenix/MATRIX ujwal55/Synopsis-Scorer

Collections including `google/gemma-3-4b-it-qat-q4_0-gguf`

Google's Gemma models family Collection
328 items • Updated Sep 12 • △ 565

Gemma 3 Release Collection
28 items • Updated Aug 11 • △ 528

Gemma 3 QAT Collection
Quantization Aware Trained (QAT) Gemm… • 15 items • Updated Jul 10 • △ 209

Edit model card

Gated model You have been granted access to this model

Gemma 3 model card

Model Page: Gemma

> This repository corresponds to the 4B **instruction-tuned** version of the Gemma 3 model in GGUF format using Quantization Aware Training (QAT). The GGUF corresponds to Q4_0 quantization.
>
> Thanks to QAT, the model is able to preserve similar quality as `bfloat16` while significantly reducing the memory requirements to load the model.
>
> You can find the half-precision version here.

Resources and Technical Documentation:

- Gemma 3 Technical Report
- Responsible Generative AI Toolkit

- Gemma on Kaggle
- Gemma on Vertex Model Garden

Terms of Use: Terms

Authors: Google DeepMind

Model Information

Summary description and brief definition of inputs and outputs.

Description

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. Gemma 3 models are multimodal, handling text and image input and generating text output, with open weights for both pre-trained variants and instruction-tuned variants. Gemma 3 has a large, 128K context window, multilingual support in over 140 languages, and is available in more sizes than previous versions. Gemma 3 models are well-suited for a variety of text generation and image understanding tasks, including question answering, summarization, and reasoning. Their relatively small size makes it possible to deploy them in environments with limited resources such as laptops, desktops or your own cloud infrastructure, democratizing access to state of the art AI models and helping foster innovation for everyone.

Inputs and outputs

- **Input:**
    - Text string, such as a question, a prompt, or a document to be summarized
    - Images, normalized to 896 x 896 resolution and encoded to 256 tokens each
    - Total input context of 128K tokens for the 4B, 12B, and 27B sizes, and 32K tokens for the 1B size

- **Output:**
  - Generated text in response to the input, such as an answer to a question, analysis of image content, or a summary of a document
  - Total output context of 8192 tokens

Usage

Below, there are some code snippets on how to get quickly started with running the model.

llama.cpp (text-only)

```
./llama-cli -hf google/gemma-3-4b-it-qat-q4_0-gguf -p "Write a poem abo
```

llama.cpp (image input)

```
wget https://github.com/bebechien/gemma/blob/main/surprise.png?raw=true
./llama-gemma3-cli -hf google/gemma-3-4b-it-qat-q4_0-gguf -p "Describe
```

ollama (text only)

Using GGUFs with Ollama via Hugging Face does not support image inputs at the moment. Please check the docs on running gated repositories.

```
ollama run hf.co/google/gemma-3-4b-it-qat-q4_0-gguf
```

Citation

```
@article{gemma_2025,
    title={Gemma 3},
    url={https://goo.gle/Gemma3Report},
```

```
    publisher={Kaggle},
    author={Gemma Team},
    year={2025}
}
```

Model Data

Data used for model training and how the data was processed.

Training Dataset

These models were trained on a dataset of text data that includes a wide variety of sources. The 27B model was trained with 14 trillion tokens, the 12B model was trained with 12 trillion tokens, 4B model was trained with 4 trillion tokens and 1B with 2 trillion tokens. Here are the key components:

- Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. The training dataset includes content in over 140 languages.
- Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code and understand code-related questions.
- Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.
- Images: A wide range of images enables the model to perform image analysis and visual data extraction tasks.

The combination of these diverse data sources is crucial for training a powerful multimodal model that can handle a wide variety of different tasks and data formats.

Data Preprocessing

Here are the key data cleaning and filtering methods applied to the training data:

- CSAM Filtering: Rigorous CSAM (Child Sexual Abuse Material) filtering was applied at multiple stages in the data preparation process to ensure the exclusion of harmful and illegal content.
- Sensitive Data Filtering: As part of making Gemma pre-trained models safe and reliable, automated techniques were used to filter out certain personal information and other sensitive data from training sets.
- Additional methods: Filtering based on content quality and safety in line with our policies.

Implementation Information

Details about the model internals.

Hardware

Gemma was trained using Tensor Processing Unit (TPU) hardware (TPUv4p, TPUv5p and TPUv5e). Training vision-language models (VLMS) requires significant computational power. TPUs, designed specifically for matrix operations common in machine learning, offer several advantages in this domain:

- Performance: TPUs are specifically designed to handle the massive computations involved in training VLMs. They can speed up training considerably compared to CPUs.
- Memory: TPUs often come with large amounts of high-bandwidth memory, allowing for the handling of large models and batch sizes during training. This can lead to better model quality.
- Scalability: TPU Pods (large clusters of TPUs) provide a scalable solution for handling the growing complexity of large foundation models. You can distribute training across multiple TPU devices for faster and more efficient processing.
- Cost-effectiveness: In many scenarios, TPUs can provide a more cost-effective solution for training large models compared to CPU-based infrastructure, especially when considering the time and resources saved due to faster training.

- These advantages are aligned with Google's commitments to operate sustainably.

Software

Training was done using JAX and ML Pathways.

JAX allows researchers to take advantage of the latest generation of hardware, including TPUs, for faster and more efficient training of large models. ML Pathways is Google's latest effort to build artificially intelligent systems capable of generalizing across multiple tasks. This is specially suitable for foundation models, including large language models like these ones.

Together, JAX and ML Pathways are used as described in the paper about the Gemini family of models; *"the 'single controller' programming model of Jax and Pathways allows a single Python process to orchestrate the entire training run, dramatically simplifying the development workflow."*

Evaluation

> The evaluation in this section correspond to the original checkpoint, not the QAT checkpoint.

Model evaluation metrics and results.

Benchmark Results

These models were evaluated against a large collection of different datasets and metrics to cover different aspects of text generation:

Reasoning and factuality

Benchmark	Metric	Gemma 3 PT 1B	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
HellaSwag	10-shot	62.3	77.2	84.2	85.6
BoolQ	0-shot	63.2	72.3	78.8	82.4
PIQA	0-shot	73.8	79.6	81.8	83.3
SocialIQA	0-shot	48.9	51.9	53.4	54.9
TriviaQA	5-shot	39.8	65.8	78.2	85.5
Natural Questions	5-shot	9.48	20.0	31.4	36.1
ARC-c	25-shot	38.4	56.2	68.9	70.6
ARC-e	0-shot	73.0	82.4	88.3	89.0
WinoGrande	5-shot	58.2	64.7	74.3	78.8
BIG-Bench Hard	few-shot	28.4	50.9	72.6	77.7
DROP	1-shot	42.4	60.1	72.2	77.2

STEM and code

Benchmark	Metric	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
MMLU	5-shot	59.6	74.5	78.6
MMLU (Pro COT)	5-shot	29.2	45.3	52.2
AGIEval	3-5-shot	42.1	57.4	66.2
MATH	4-shot	24.2	43.3	50.0

Benchmark	Metric	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
GSM8K	8-shot	38.4	71.0	82.6
GPQA	5-shot	15.0	25.4	24.3
MBPP	3-shot	46.0	60.4	65.6
HumanEval	0-shot	36.0	45.7	48.8

Multilingual

Benchmark	Gemma 3 PT 1B	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
MGSM	2.04	34.7	64.3	74.3
Global-MMLU-Lite	24.9	57.0	69.4	75.7
WMT24++ (ChrF)	36.7	48.4	53.9	55.7
FloRes	29.5	39.2	46.0	48.8
XQuAD (all)	43.9	68.0	74.5	76.8
ECLeKTic	4.69	11.0	17.2	24.4
IndicGenBench	41.4	57.2	61.7	63.4

Multimodal

Benchmark	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
COCOcap	102	111	116
DocVQA (val)	72.8	82.3	85.6
InfoVQA (val)	44.1	54.8	59.4

Benchmark	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
MMMU (pt)	39.2	50.3	56.1
TextVQA (val)	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
OKVQA	51.0	58.7	60.2
TallyQA	42.5	51.8	54.3
SpatialSense VQA	50.9	60.0	59.4
CountBenchQA	26.1	17.8	68.0

Ethics and Safety

Ethics and safety evaluation approach and results.

Evaluation Approach

Our evaluation methods include structured evaluations and internal red-teaming testing of relevant content policies. Red-teaming was conducted by a number of different teams, each with different goals and human evaluation metrics. These models were evaluated against a number of different categories relevant to ethics and safety, including:

- **Child Safety**: Evaluation of text-to-text and image to text prompts covering child safety policies, including child sexual abuse and exploitation.
- **Content Safety:** Evaluation of text-to-text and image to text prompts covering safety policies including, harassment, violence and gore, and hate speech.
- **Representational Harms**: Evaluation of text-to-text and image to text prompts covering safety policies including bias, stereotyping, and harmful associations or inaccuracies.

In addition to development level evaluations, we conduct "assurance evaluations" which are our 'arms-length' internal evaluations for responsibility governance decision making. They are conducted separately from the model development team, to inform decision making about release. High level findings are fed back to the model team, but prompt sets are held-out to prevent overfitting and preserve the results' ability to inform decision making. Assurance evaluation results are reported to our Responsibility & Safety Council as part of release review.

Evaluation Results

For all areas of safety testing, we saw major improvements in the categories of child safety, content safety, and representational harms relative to previous Gemma models. All testing was conducted without safety filters to evaluate the model capabilities and behaviors. For both text-to-text and image-to-text, and across all model sizes, the model produced minimal policy violations, and showed significant improvements over previous Gemma models' performance with respect to ungrounded inferences. A limitation of our evaluations was they included only English language prompts.

Usage and Limitations

These models have certain limitations that users should be aware of.

Intended Usage

Open vision-language models (VLMs) models have a wide range of applications across various industries and domains. The following list of potential uses is not comprehensive. The purpose of this list is to provide contextual information about the possible use-cases that the model creators considered as part of model training and development.

- Content Creation and Communication
    - Text Generation: These models can be used to generate creative text formats such as poems, scripts, code, marketing copy, and email drafts.
    - Chatbots and Conversational AI: Power conversational interfaces for customer service, virtual assistants, or interactive applications.
    - Text Summarization: Generate concise summaries of a text corpus, research papers, or reports.
    - Image Data Extraction: These models can be used to extract, interpret, and summarize visual data for text communications.
- Research and Education
    - Natural Language Processing (NLP) and VLM Research: These models can serve as a foundation for researchers to experiment with VLM and NLP techniques, develop algorithms, and contribute to the advancement of the field.
    - Language Learning Tools: Support interactive language learning experiences, aiding in grammar correction or providing writing practice.
    - Knowledge Exploration: Assist researchers in exploring large bodies of text by generating summaries or answering questions about specific topics.

Limitations

- Training Data
    - The quality and diversity of the training data significantly influence the model's capabilities. Biases or gaps in the training data can lead to

limitations in the model's responses.

    - The scope of the training dataset determines the subject areas the model can handle effectively.

- Context and Task Complexity

    - Models are better at tasks that can be framed with clear prompts and instructions. Open-ended or highly complex tasks might be challenging.
    - A model's performance can be influenced by the amount of context provided (longer context generally leads to better outputs, up to a certain point).

- Language Ambiguity and Nuance

    - Natural language is inherently complex. Models might struggle to grasp subtle nuances, sarcasm, or figurative language.

- Factual Accuracy

    - Models generate responses based on information they learned from their training datasets, but they are not knowledge bases. They may generate incorrect or outdated factual statements.

- Common Sense

    - Models rely on statistical patterns in language. They might lack the ability to apply common sense reasoning in certain situations.

Ethical Considerations and Risks

The development of vision-language models (VLMs) raises several ethical concerns. In creating an open model, we have carefully considered the following:

- Bias and Fairness

    - VLMs trained on large-scale, real-world text and image data can reflect socio-cultural biases embedded in the training material. These models underwent careful scrutiny, input data pre-processing described and posterior evaluations reported in this card.

- Misinformation and Misuse
    - VLMs can be misused to generate text that is false, misleading, or harmful.
    - Guidelines are provided for responsible use with the model, see the Responsible Generative AI Toolkit.
- Transparency and Accountability:
    - This model card summarizes details on the models' architecture, capabilities, limitations, and evaluation processes.
    - A responsibly developed open model offers the opportunity to share innovation by making VLM technology accessible to developers and researchers across the AI ecosystem.

Risks identified and mitigations:

- **Perpetuation of biases**: It's encouraged to perform continuous monitoring (using evaluation metrics, human review) and the exploration of de-biasing techniques during model training, fine-tuning, and other use cases.
- **Generation of harmful content**: Mechanisms and guidelines for content safety are essential. Developers are encouraged to exercise caution and implement appropriate content safety safeguards based on their specific product policies and application use cases.
- **Misuse for malicious purposes**: Technical limitations and developer and end-user education can help mitigate against malicious applications of VLMs. Educational resources and reporting mechanisms for users to flag misuse are provided. Prohibited uses of Gemma models are outlined in the Gemma Prohibited Use Policy.
- **Privacy violations**: Models were trained on data filtered for removal of certain personal information and other sensitive data. Developers are encouraged to adhere to privacy regulations with privacy-preserving techniques.

Benefits

At the time of release, this family of models provides high-performance open vision-language model implementations designed from the ground up for responsible AI development compared to similarly sized models.

Using the benchmark evaluation metrics described in this document, these models have shown to provide superior performance to other, comparably-sized open model alternatives.