



openbmb/MiniCPM-Llama3-V-2\_5

♡ like

138

Visual Question Answering

Transformers

Safetensors

HaoyeZhang/RLAIF-V-Dataset

English

Chinese

minicpmv

feature-extraction

custom\_code

Train ▾

Use this model ▾

Model card

Files

Community



Safetensors ⓘ

Model size

8.54B params

Tensor type

FP16

Visual Question Answering

Inference API (serverless) does not yet support model repos that contain custom code.

Dataset used to train openbmb/MiniCPM-Llama3-V-2\_5

HaoyeZhang/RLAIF-V-Dataset

Updated about 13 hours ago • ♡ 7

Collection including openbmb/MiniCPM-Llama3-V-2\_5

MiniCPM-2B





Collection

The MiniCPM family of LLMs and VLLMs. • 17 items • Updated about 2 hours ago • △ 7


[Edit model card](#)[GitHub](#) | [Demo](#)

## [🔗](#) MiniCPM-Llama3-V 2.5

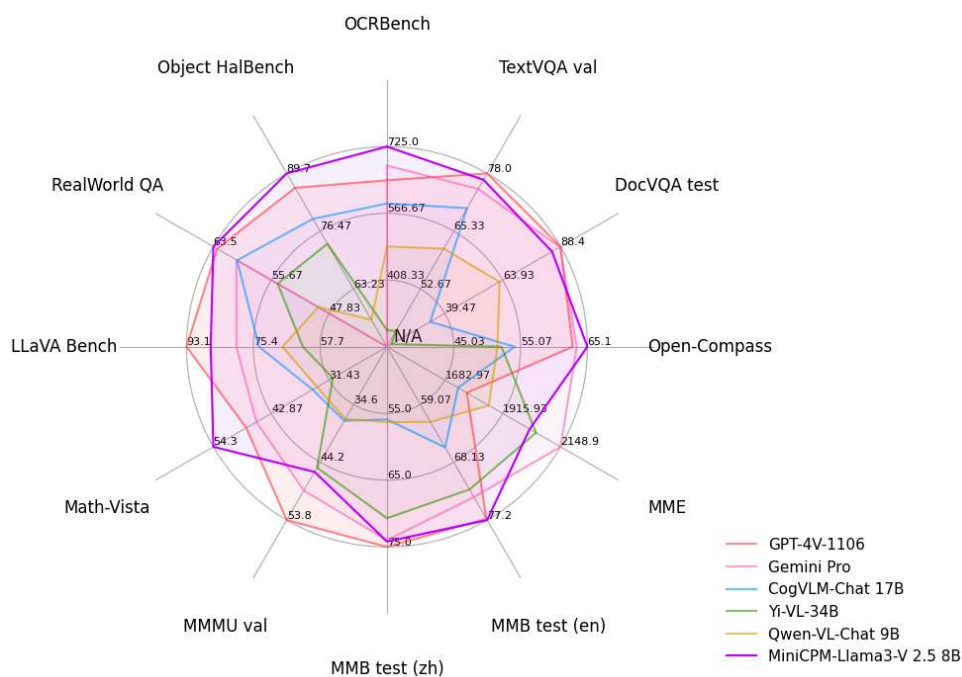
**MiniCPM-Llama3-V 2.5** is the latest model in the MiniCPM-V series. The model is built on SigLip-400M and Llama3-8B-Instruct with a total of 8B parameters. It exhibits a significant performance improvement over MiniCPM-V 2.0. Notable features of MiniCPM-Llama3-V 2.5 include:

-  **Leading Performance.** MiniCPM-Llama3-V 2.5 has achieved an average score of 65.1 on OpenCompass, a comprehensive evaluation over 11 popular benchmarks. **It surpasses widely used proprietary models like GPT-4V-1106, Gemini Pro, Qwen-VL-Max and Claude 3 with 8B parameters, greatly outperforming other multimodal large models built on Llama 3.**
-  **Strong OCR Capabilities.** MiniCPM-Llama3-V 2.5 can process images with any aspect ratio up to 1.8 million pixels, achieving an **700+ score on OCRBench, surpassing proprietary models such as GPT-4o, GPT-4V-0409, Qwen-VL-Max and Gemini Pro.** Based on recent user feedback, MiniCPM-Llama3-V 2.5 has now enhanced full-text OCR extraction, table-to-markdown conversion, and other high-utility capabilities, and has further strengthened its instruction-following and complex reasoning abilities, enhancing multimodal interaction experiences.
-  **Trustworthy Behavior.** Leveraging the latest [RLAIF-V](#) method (the newest technology in the [RLHF-V](#) [CVPR'24] series), MiniCPM-Llama3-V 2.5 exhibits trustworthy multimodal behavior. It achieves **10.3%** hallucination rate on Object HalBench, lower than GPT-4V-1106 (13.6%), achieving the best level within the open-source community.
-  **Multilingual Support.** Thanks to Llama 3's robust multilingual capabilities and VisCPM's cross-lingual generalization technology, MiniCPM-Llama3-V 2.5 extends

its foundational bilingual (Chinese-English) multimodal capabilities to support 30+ languages including German, French, Spanish, Italian, Russian etc. We achieve this extension through only minimal instruction-tuning with translated multimodal data. [All Supported Languages](#).

-  **Efficient Deployment.** MiniCPM-Llama3-V 2.5 systematically employs **model quantization, CPU optimizations, NPU optimizations and compilation optimizations** as acceleration techniques, achieving high-efficiency deployment on edge devices. For mobile phones with Qualcomm chips, we have integrated the NPU acceleration framework QNN into llama.cpp for the first time. After systematic optimization, MiniCPM-Llama3-V 2.5 has realized a **150-fold acceleration in multimodal large model edge-side image encoding** and a **3-fold increase in language decoding speed**.

## 🔗 Evaluation

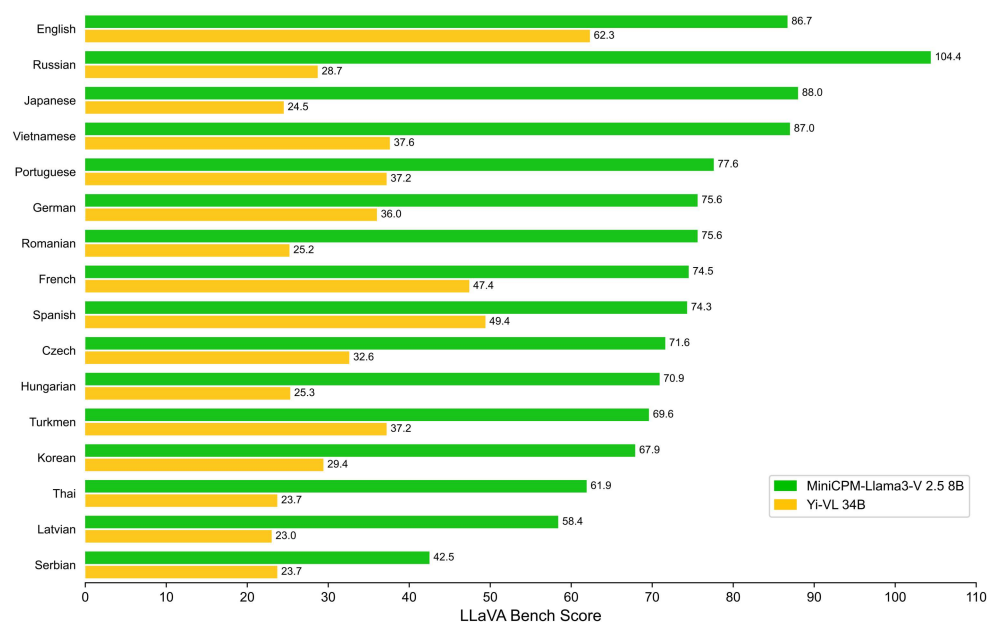


Results on TextVQA, DocVQA, OCRBench, OpenCompass, MME, MMBench, MMMU, MathVista, LLaVA Bench, RealWorld QA, Object HalBench.

Model	Size	OCRBench	TextVQA val	DocVQA test	Open-Compass	MME	MMB test (en)	MMB test (cn)	MMMU val	Math-Vista	LLaVA Bench	RealWorld QA	Object HalBench
Proprietary													
Gemini Pro	-	680	74.6	88.1	62.9	2148.9	73.6	74.3	48.9	45.8	79.9	60.4	-
GPT-4V (2023.11.06)	-	645	78.0	88.4	63.5	1771.5	77.0	74.4	53.8	47.8	93.1	63.0	86.4
Open-source													
Mini-Gemini	2.2B	-	56.2	34.2*	-	1653.0	-	-	31.7	-	-	-	-
Qwen-VL-Chat	9.6B	488	61.5	62.6	51.6	1860.0	61.8	56.3	37.0	33.8	67.7	49.3	56.2
DeepSeek-VL-7B	7.3B	435	64.7*	47.0*	54.6	1765.4	73.8	71.4	38.3	36.8	77.8	54.2	
Yi-VL-34B	34B	290	43.4*	16.9*	52.2	2050.2	72.4	70.7	45.1	30.7	62.3	54.8	79.3
CogVLM-Chat	17.4B	590	70.4	33.3*	54.2	1736.6	65.8	55.9	37.3	34.7	73.9	60.3	73.6
TextMonkey	9.7B	558	64.3	66.7	-	-	-	-	-	-	-	-	-
IDEFICS2-8B	8.0B	-	73.0	74.0	57.2	1847.6	75.7	68.6	45.2	52.2	49.1	60.7	-
Bunny-Llama-3-8B	8.4B	-	-	-	54.3	1920.3	77.0	73.9	41.3	31.5	61.2	58.8	-
LLaVA-NeXT Llama-3-8B	8.4B	-	-	-	-	1971.5	-	-	41.7	-	80.1	60.0	-
MiniCPM-V 1.0	2.8B	366	60.6	38.2	47.5	1650.2	64.1	62.6	38.3	28.9	51.3	51.2	78.4
MiniCPM-V 2.0	2.8B	605	74.1	71.9	54.5	1808.6	69.1	66.5	38.2	38.7	69.2	55.8	85.5
MiniCPM-Llama3-V 2.5	8.5B	725	76.6	84.8	65.1	2024.6	77.2	74.2	45.8	54.3	86.7	63.5	89.7

\* We evaluate the officially released checkpoint by ourselves.

Results of LLaVABench in multiple languages.



[Examples](#)

## Unesco announces its newest geoparks around the world

9 April 2024  
By Lynn Brown, Features correspondent

**Please identify the text in the picture.**

**MiniCPM-Llama3-V 2.5**

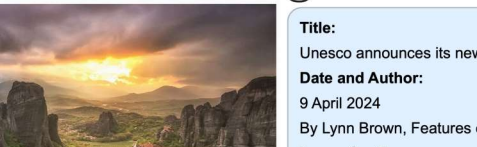
**Title:**  
Unesco announces its newest geoparks around the world

**Date and Author:**  
9 April 2024  
By Lynn Brown, Features correspondent

**Image Caption:**  
Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis)

**Body Text:**  
From dinosaur fossils in Brazil to the soaring meteors of Meteora, these 18 new geological sites highlight a particular region's natural, cultural and intangible heritage.  
  
Travellers interested in deep dives into geology, culture and sustainability have several new destinations to place on their bucket list. Unesco just announced the designation of 18 new Global Geoparks, bringing the total to 213 sites representing 48 countries around the world.  
  
From crystal-clear karst lakes to an archipelago formed during the last ice age, these parks are more than just unique geological features; they also serve to highlight a particular region's natural, cultural and intangible heritage. Unesco Global Geoparks represent landscapes where internationally significant geological heritage intertwines with human history and supports rich biodiversity," explains António Abreu, director of the Division of Ecological Sciences at Unesco. "Each new geopark tells a unique story, using geology as its lens."

Here are five of the newest geoparks on Unesco's list:



(Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis))

Chris Karagelis

[illegible]

模型名称	发布时间	发布机构	语言	参数量规	Tokens-规模	模型结构	是否开源
T5	2019-10	Google	英	13B		T5-style	√
GPT-3	2020-05	OpenAI	英	175B	300B	GP-Tstyle	×
LaMDA	2021-05	Google	英	137B	2.8T	GP-Tstyle	×
Jurassic	2021-08	AI21	英	178B	300B	GP-Tstyle	×
MT-NLG	2021-10	Microsoft, NVIDIA	英	530B	270B	GP-Tstyle	×
ERNIE 3.0	2021-12	Baidu	中	260B	300B	Multi-task	×
Gopher	2021-12	DeepMind	英	280B	300B	GP-Tstyle	×
Chinchilla	2022-04	DeepMind	英	70B	1.4T	GP-Tstyle	×
PaLM	2022-04	Google	多语言	540B	780B	GP-Tstyle	×
OPT	2022-05	Meta	英	125M-175B	180B	GP-Tstyle	√
BLOOM	2022-07	BigScience	多语言	176B	366B	GP-Tstyle	√
GLM-130B	2022-08	Tsinghua	中、英	130B	400B	GLM-style	√
LLaMA	2023-02	Meta	多语言	7B-65B	1.4T	GP-Tstyle	√

**How does UltraEval work?**

The workflow is divided into three main stages:

- Data preparation:** Raw data (Knowledge, Code, ...) is collected from Official data sources (Hugging Face, GitHub, ...).
- Evaluation process:** The data is converted into Prompt input, which is then used for Model deployment. The output is processed (Post-process) and then Metric calculation is performed, resulting in a Score.
- Model support:** The model is supported by various APIs (AI, API, ...), Local URLs, vLLM/Torch, HF Model, and Custom Model.

A user icon is shown next to the question "How does it work?"



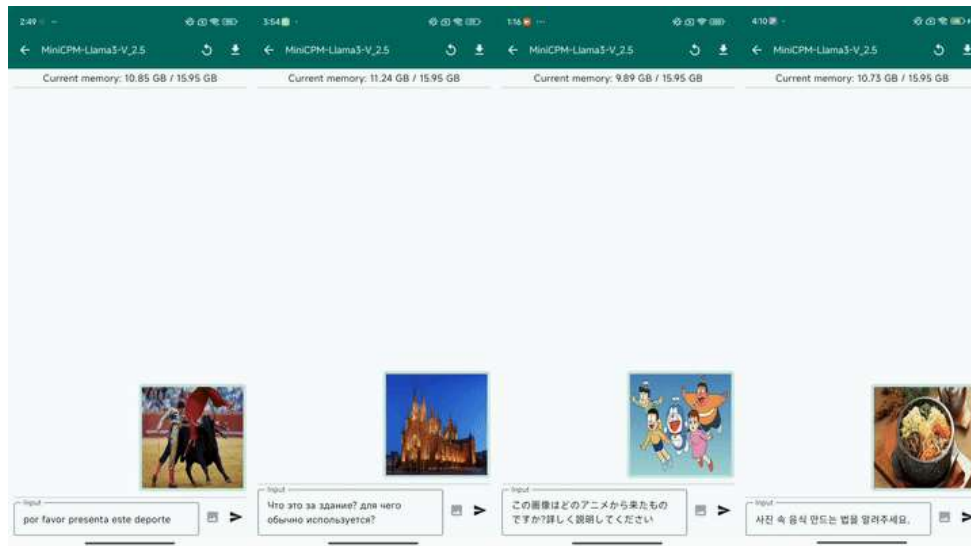
UltraEval appears to be a system or process designed to evaluate data, likely in the context of a scientific or research project. The flowchart outlines the steps involved in this process:

- 1.Data Preparation:** This stage involves collecting and organizing raw data, which can come from official data sources such as GitHub or Hugging Face. These are represented by icons indicating their origin.
- 2.Evaluation Process:** After the data is prepared, it undergoes an evaluation process that includes several steps:
  - Prompt Input:** This could involve providing prompts or instructions for the model.
  - Model Deployment:** Here, the model is likely trained or configured to perform evaluations based on the input.
  - Local URL:** This step may involve accessing or using a local URL, possibly for deploying the model or retrieving results.
  - Post-process:** This could refer to further processing of the model's output, such as cleaning, filtering, or formatting data.
- 3. Metric Calculation:** Following the post-process, metrics are calculated to evaluate the performance of the model. This is indicated by a graph icon, suggesting quantitative analysis.
- 4. Score:** The final stage is to determine the score, which is likely the outcome of the metric calculation. This score would reflect the model's performance or the quality of the data after evaluation.
- 5. Model Support:** Throughout the process, there is support for various models, including vLLM/Torch, HF Model, and Custom Model. These models are likely used at different stages of the evaluation process.

In summary, UltraEval seems to be a structured approach to evaluating data using machine learning models, with a focus on performance metrics and customization options for different types of models.

We deploy MiniCPM-Llama3-V 2.5 on end devices. The demo video is the raw screen recording on a Xiaomi 14 Pro at double speed.





## [Demo](#)

Click here to try out the Demo of [MiniCPM-Llama3-V 2.5](#).

## [Deployment on Mobile Phone](#)

Coming soon.

## [Usage](#)

Inference using Huggingface transformers on NVIDIA GPUs. Requirements tested on python 3.10:

```
Pillow==10.1.0
torch==2.1.2
torchvision==0.16.2
transformers==4.40.0
sentencepiece==0.1.99
```

```
# test.py
import torch
from PIL import Image
```

```
from transformers import AutoModel, AutoTokenizer

model = AutoModel.from_pretrained('openbmb/MiniCPM-Llama3-V-2_5', trust_remote_code=True)
model = model.to(device='cuda')

tokenizer = AutoTokenizer.from_pretrained('openbmb/MiniCPM-Llama3-V-2_5', trust_remote_code=True)
model.eval()

image = Image.open('xx.jpg').convert('RGB')
question = 'What is in the image?'
msgs = [{'role': 'user', 'content': question}]

res = model.chat(
    image=image,
    msgs=msgs,
    tokenizer=tokenizer,
    sampling=True,
    temperature=0.7
)
print(res)
```

Please look at [GitHub](#) for more detail about usage.

## Int4 quantized version

Download the int4 quantized version for lower GPU memory usage: [MiniCPM-Llama3-V-2\\_5-int4](#).

## MiniCPM-V 2.0

Please see the info about MiniCPM-V 2.0 [here](#).

## License

## Model License



- The code in this repo is released according to [Apache-2.0](#)
- The usage of MiniCPM-Llama3-V 2.5's parameters is subject to "[General Model License Agreement - Source Notes - Publicity Restrictions - Commercial License](#)"
- The parameters are fully open to academic research
- Please contact [cpm@modelbest.cn](mailto:cpm@modelbest.cn) to obtain a written authorization for commercial uses. Free commercial use is also allowed after registration.

### Statement

- As a LLM, MiniCPM-Llama3-V 2.5 generates contents by learning a large amount of texts, but it cannot comprehend, express personal opinions or make value judgement. Anything generated by MiniCPM-Llama3-V 2.5 does not represent the views and positions of the model developers
- We will not be liable for any problems arising from the use of the MiniCPM-V open Source model, including but not limited to data security issues, risk of public opinion, or any risks and problems arising from the misdirection, misuse, dissemination or misuse of the model.

### Other Multimodal Projects from Our Team

[VisCPM](#) | [RLHF-V](#) | [LLaVA-UHD](#) | [RLAIF-V](#)

### Citation

If you find our work helpful, please consider citing the following papers

```
@article{yu2023rlhf,  
  title={Rlhf-v: Towards trustworthy mllms via behavior alignment from  
  author={Yu, Tianyu and Yao, Yuan and Zhang, Haoye and He, Taiwen and  
  journal={arXiv preprint arXiv:2312.00849},  
  year={2023}  
}  
  
@article{viscpm,
```

```
title={Large Multilingual Models Pivot Zero-Shot Multimodal Learning}
author={Jinyi Hu and Yuan Yao and Chongyi Wang and Shan Wang and Yizhe
journal={arXiv preprint arXiv:2308.12038},
year={2023}
}

@article{xu2024llava-uhd,
title={{LLaVA-UHD}: an LMM Perceiving Any Aspect Ratio and High-Resolution}
author={Xu, Ruyi and Yao, Yuan and Guo, Zonghao and Cui, Junbo and Ning}
journal={arXiv preprint arXiv:2403.11703},
year={2024}
}
```



Company

TOS

Privacy

About

Jobs

Website

Models

Datasets

Spaces

Pricing

Docs

© Hugging Face