deepseek-ai / DeepSeek-V3.1-Base

like 881

Follow DeepSeek 90.4k

Text Generation | Transformers | Safetensors | deepseek_v3 | conversational | custom_code | text-generation-inference | fp8 | arxiv:2412.19437 | License: mit

Train | Deploy | Use this model

Model card | Files xet | Community 53

Downloads last month
12,528

Safetensors

Model size: 685B params | Tensor type: BF16 · F8_E4M3 · F32 | Chat template | Files info

Inference Providers NEW

Text Generation

This model isn't deployed by any Inference Provider. 100 Ask for provider support

Model tree for deepseek-ai/DeepSeek-V3.1-Base

- **Finetunes** — 4 models
- **Merges** — 1 model
- **Quantizations** — 7 models

Spaces using deepseek-ai/DeepSeek-V3.1-Base 5

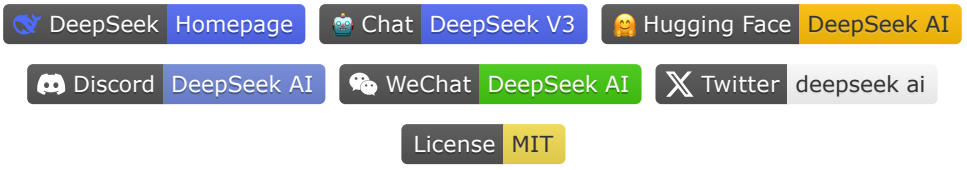
umint/ai | ajay5364747/Ajay | qualybittech/bharath | juliusNice/deepseekv3.1 | Xavernox/deepseek-ai-DeepSeek-V3.1-Base

Collection including deepseek-ai/DeepSeek-V3.1-Base

DeepSeek-V3.1 Collection

3 items • Updated 3 days ago • △ 202

DeepSeek-V3.1



Introduction

DeepSeek-V3.1 is a hybrid model that supports both thinking mode and non-thinking mode. Compared to the previous version, this upgrade brings improvements in multiple aspects:

- **Hybrid thinking mode**: One model supports both thinking mode and non-thinking mode by changing the chat template.
- **Smarter tool calling**: Through post-training optimization, the model's performance in tool usage and agent tasks has significantly improved.
- **Higher thinking efficiency**: DeepSeek-V3.1-Think achieves comparable answer quality to DeepSeek-R1-0528, while responding more quickly.

DeepSeek-V3.1 is post-trained on the top of DeepSeek-V3.1-Base, which is built upon the original V3 base checkpoint through a two-phase long context extension approach, following the methodology outlined in the original DeepSeek-V3 report. We have expanded our dataset by collecting additional long documents and substantially extending both training phases. The 32K extension phase has been increased 10-fold to

630B tokens, while the 128K extension phase has been extended by 3.3x to 209B tokens. Additionally, DeepSeek-V3.1 is trained using the UE8M0 FP8 scale data format to ensure compatibility with microscaling data formats.

Model Downloads

Model	#Total Params	#Activated Params	Context Length	Download
DeepSeek-V3.1-Base	671B	37B	128K	HuggingFace | ModelScope
DeepSeek-V3.1	671B	37B	128K	HuggingFace | ModelScope

Chat Template

The details of our chat template is described in `tokenizer_config.json` and `assets/chat_template.jinja`. Here is a brief description.

Non-Thinking

First-Turn

Prefix: `<｜begin▁of▁sentence｜>{system prompt}<｜User｜>{query}<｜Assistant｜></think>`

With the given prefix, DeepSeek V3.1 generates responses to queries in non-thinking mode. Unlike DeepSeek V3, it introduces an additional token `</think>`.

Multi-Turn

Context: `<｜begin▁of▁sentence｜>{system prompt}<｜User｜>{query}<｜Assistant｜></think>{response}<｜end▁of▁sentence｜>...<｜User｜>{query}<｜Assistant｜></think>{response}<｜end▁of▁sentence｜>`

Prefix: `<｜User｜>{query}<｜Assistant｜></think>`

By concatenating the context and the prefix, we obtain the correct prompt for the query.

Thinking

First-Turn

Prefix: `<｜begin_of_sentence｜>{system prompt}<｜User｜>{query}<｜Assistant｜><think>`

The prefix of thinking mode is similar to DeepSeek-R1.

Multi-Turn

Context: `<｜begin_of_sentence｜>{system prompt}<｜User｜>{query}<｜Assistant｜></think>{response}<｜end_of_sentence｜>...<｜User｜>{query}<｜Assistant｜></think>{response}<｜end_of_sentence｜>`

Prefix: `<｜User｜>{query}<｜Assistant｜><think>`

The multi-turn template is the same with non-thinking multi-turn chat template. It means the thinking token in the last turn will be dropped but the `</think>` is retained in every turn of context.

ToolCall

Toolcall is supported in non-thinking mode. The format is:

`<｜begin_of_sentence｜>{system prompt}\n\n{tool_description}<｜User｜>{query}<｜Assistant｜></think>` where the tool_description is

```
## Tools
You have access to the following tools:

### {tool_name1}
Description: {description}

Parameters: {json.dumps(parameters)}
```

```
IMPORTANT: ALWAYS adhere to this exact format for tool use:
<｜tool▁calls▁begin｜><｜tool▁call▁begin｜>tool_call_name<｜tool▁sep｜>to

Where:
- `tool_call_name` must be an exact match to one of the available tools
- `tool_call_arguments` must be valid JSON that strictly follows the to
- For multiple tool calls, chain them directly without separators or sp
```

Code-Agent

We support various code agent frameworks. Please refer to the above toolcall format to create your own code agents. An example is shown in `assets/code_agent_trajectory.html`.

Search-Agent

We design a specific format for searching toolcall in thinking mode, to support search agent.

For complex questions that require accessing external or up-to-date information, DeepSeek-V3.1 can leverage a user-provided search tool through a multi-turn tool-calling process.

Please refer to the `assets/search_tool_trajectory.html` and `assets/search_python_tool_trajectory.html` for the detailed template.

Evaluation

Category	Benchmark (Metric)	DeepSeek V3.1-NonThinking	DeepSeek V3 0324	DeepSeek V3.1-Thinking	DeepSeek R1 0528
General					
	MMLU-Redux (EM)	91.8	90.5	93.7	93.4
	MMLU-Pro (EM)	83.7	81.2	84.8	85.0

Category	Benchmark (Metric)	DeepSeek V3.1-NonThinking	DeepSeek V3 0324	DeepSeek V3.1-Thinking	DeepSeek R1 0528
	GPQA-Diamond (Pass@1)	74.9	68.4	80.1	81.0
	Humanity's Last Exam (Pass@1)	-	-	15.9	17.7
Search Agent					
	BrowseComp	-	-	30.0	8.9
	BrowseComp_zh	-	-	49.2	35.7
	Humanity's Last Exam (Python + Search)	-	-	29.8	24.8
	SimpleQA	-	-	93.4	92.3
Code					
	LiveCodeBench (2408-2505) (Pass@1)	56.4	43.0	74.8	73.3
	Codeforces-Div1 (Rating)	-	-	2091	1930
	Aider-Polyglot (Acc.)	68.4	55.1	76.3	71.6
Code Agent					
	SWE Verified (Agent mode)	66.0	45.4	-	44.6
	SWE-bench Multilingual (Agent mode)	54.5	29.3	-	30.5
	Terminal-bench (Terminus 1	31.3	13.3	-	5.7

Category	Benchmark (Metric)	DeepSeek V3.1-NonThinking	DeepSeek V3 0324	DeepSeek V3.1-Thinking	DeepSeek R1 0528
	framework)				
Math					
	AIME 2024 (Pass@1)	66.3	59.4	93.1	91.4
	AIME 2025 (Pass@1)	49.8	51.3	88.4	87.5
	HMMT 2025 (Pass@1)	33.5	29.2	84.2	79.4

Note:

- Search agents are evaluated with our internal search framework, which uses a commercial search API + webpage filter + 128K context window. Seach agent results of R1-0528 are evaluated with a pre-defined workflow.
- SWE-bench is evaluated with our internal code agent framework.
- HLE is evaluated with the text-only subset.

Usage Example

```
import transformers

tokenizer = transformers.AutoTokenizer.from_pretrained("deepseek-ai/Dee

messages = [
    {"role": "system", "content": "You are a helpful assistant"},
    {"role": "user", "content": "Who are you?"},
    {"role": "assistant", "content": "<think>Hmm</think>I am DeepSeek"}
    {"role": "user", "content": "1+1=?"}
]

tokenizer.apply_chat_template(messages, tokenize=False, thinking=True,
# '<｜begin_of_sentence｜>You are a helpful assistant<｜User｜>Who are y
```

```
tokenizer.apply_chat_template(messages, tokenize=False, thinking=False,
# '<｜begin▁of▁sentence｜>You are a helpful assistant<｜User｜>Who are y
```

How to Run Locally

The model structure of DeepSeek-V3.1 is the same as DeepSeek-V3. Please visit DeepSeek-V3 repo for more information about running this model locally.

License

This repository and the model weights are licensed under the MIT License.

Citation

```
@misc{deepseekai2024deepseekv3technicalreport,
      title={DeepSeek-V3 Technical Report},
      author={DeepSeek-AI},
      year={2024},
      eprint={2412.19437},
      archivePrefix={arXiv},
      primaryClass={cs.CL},
      url={https://arxiv.org/abs/2412.19437},
}
```

Contact

If you have any questions, please raise an issue or contact us at service@deepseek.com.

TOS

Privacy

About

Jobs

Website

Models

Datasets

Spaces

Pricing

Docs