



deepseek-ai / **DeepSeek-V4-Flash**

♡ like 1.14k

Follow DeepSeek 131k



Text Generation



Transformers



Safetensors

deepseek_v4

conversational



Eval Results



8-bit precision

fp8



License: mit



Deploy

Use this model



Model card



Files



xet



Community 32

Downloads last month

1,904,105



Safetensors

Model size

158B params

Tensor type

BF16 · I64 · F32 · F8_E8M0 · F8_E4M3 · I8

Files info

Inference Providers **NEW**



Novita



OpenAI



Text Generation

Examples



Input a message to start chatting with deepseek-ai/DeepSeek-V4-Flash.

Your prompt here...

Send

View Code

Compare providers

📁 Model tree for deepseek-ai/DeepSeek-V4-Flash

📁 Finetunes 8 models

📁 Quantizations 🧠 📄 🙌 🧐 50 models

🏠 Spaces using deepseek-ai/DeepSeek-V4-Flash 40

🚀 abdallah110/htf ✂️ crusoeai/foundry-model-arena 🏠 adesh01/civicsetu

🚀 Otter21/OPENENV_RL_01 📅 bluemoonsoldout/llm-cal + 35 Spaces

📁 Collection including deepseek-ai/DeepSeek-V4-Flash

DeepSeek-V4 📁 Collection

4 items • Updated 25 days ago • ⬆ 645

📊 Evaluation results ⓘ

Diamond on ldavidrein/gpqa ⓘ	leaderboard ➤	88.1
Swe Bench Resolved on SWE-bench/SWE-bench_Verified ⓘ	leaderboard ➤	79
General on claw-eval/Claw-Eval ⓘ	🧑 source leaderboard ➤	57.8 [*]
Multi Turn on claw-eval/Claw-Eval ⓘ	🧑 source leaderboard ➤	57.9 [*]
Mmlu Pro on TIGER-Lab/MMLU-Pro ⓘ	leaderboard ➤	86.4
Terminalbench 2 on harborframework/terminal-bench-2.0 ⓘ	leaderboard ➤	56.9
Hle on cais/hle ⓘ		34.8

🔗 DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence



Technical Report

Introduction

We present a preview version of **DeepSeek-V4** series, including two strong Mixture-of-Experts (MoE) language models — **DeepSeek-V4-Pro** with 1.6T parameters (49B activated) and **DeepSeek-V4-Flash** with 284B parameters (13B activated) — both supporting a context length of **one million tokens**.

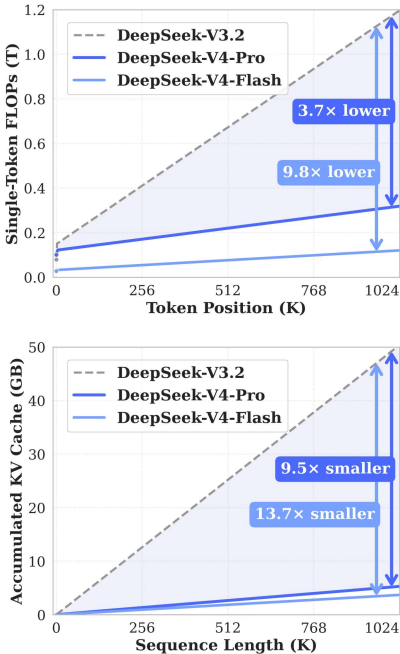
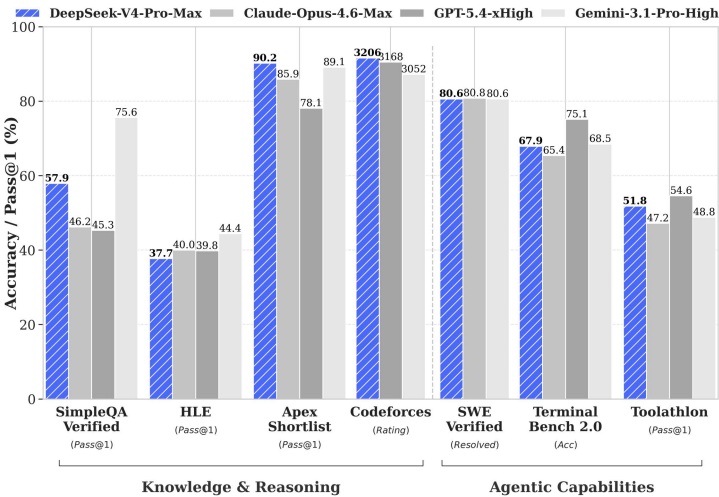
DeepSeek-V4 series incorporate several key upgrades in architecture and optimization:

1. **Hybrid Attention Architecture:** We design a hybrid attention mechanism combining Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA) to dramatically improve long-context efficiency. In the 1M-token context setting, DeepSeek-V4-Pro requires only **27% of single-token inference FLOPs** and **10% of KV cache** compared with DeepSeek-V3.2.
2. **Manifold-Constrained Hyper-Connections (mHC):** We incorporate mHC to strengthen conventional residual connections, enhancing stability of signal propagation across layers while preserving model expressivity.
3. **Muon Optimizer:** We employ the Muon optimizer for faster convergence and greater training stability.

We pre-train both models on more than **32T** diverse and high-quality tokens, followed by a comprehensive post-training pipeline. The post-training features a two-stage paradigm: independent cultivation of domain-specific experts (through SFT and RL with GRPO), followed by unified model consolidation via on-policy distillation, integrating distinct proficiencies across diverse domains into a single model.

DeepSeek-V4-Pro-Max, the maximum reasoning effort mode of DeepSeek-V4-Pro, significantly advances the knowledge capabilities of open-source models, firmly

establishing itself as the best open-source model available today. It achieves top-tier performance in coding benchmarks and significantly bridges the gap with leading closed-source models on reasoning and agentic tasks. Meanwhile, **DeepSeek-V4-Flash-Max** achieves comparable reasoning performance to the Pro version when given a larger thinking budget, though its smaller parameter scale naturally places it slightly behind on pure knowledge tasks and the most complex agentic workflows.



Model Downloads

Model	#Total Params	#Activated Params	Context Length	Precision	Download
DeepSeek-V4-Flash-Base	284B	13B	1M	FP8 Mixed	HuggingFace ModelScope
DeepSeek-V4-Flash	284B	13B	1M	FP4 + FP8 Mixed*	HuggingFace ModelScope
DeepSeek-V4-Pro-Base	1.6T	49B	1M	FP8 Mixed	HuggingFace ModelScope

Model	#Total Params	#Activated Params	Context Length	Precision	Download
DeepSeek-V4- Pro	1.6T	49B	1M	FP4 + FP8 Mixed*	HuggingFace ModelScope

**FP4 + FP8 Mixed: MoE expert parameters use FP4 precision; most other parameters use FP8.*

 Evaluation Results

 Base Model

Benchmark (Metric)	# Shots	DeepSeek-V3.2- Base	DeepSeek-V4-Flash- Base	DeepSeek-V4-Pro- Base
Architecture	-	MoE	MoE	MoE
# Activated Params	-	37B	13B	49B
# Total Params	-	671B	284B	1.6T
World Knowledge				
AGIEval (EM)	0-shot	80.1	82.6	83.1
MMLU (EM)	5-shot	87.8	88.7	90.1
MMLU-Redux (EM)	5-shot	87.5	89.4	90.8
MMLU-Pro (EM)	5-shot	65.5	68.3	73.5
MMMLU (EM)	5-shot	87.9	88.8	90.3
C-Eval (EM)	5-shot	90.4	92.1	93.1
CMMLU (EM)	5-shot	88.9	90.4	90.8

Benchmark (Metric)	# Shots	DeepSeek-V3.2-Base	DeepSeek-V4-Flash-Base	DeepSeek-V4-Pro-Base
MultiLoKo (EM)	5-shot	38.7	42.2	51.1
Simple-QA verified (EM)	25-shot	28.3	30.1	55.2
SuperGPQA (EM)	5-shot	45.0	46.5	53.9
FACTS Parametric (EM)	25-shot	27.1	33.9	62.6
TriviaQA (EM)	5-shot	83.3	82.8	85.6
Language & Reasoning				
BBH (EM)	3-shot	87.6	86.9	87.5
DROP (F1)	1-shot	88.2	88.6	88.7
HellaSwag (EM)	0-shot	86.4	85.7	88.0
WinoGrande (EM)	0-shot	78.9	79.5	81.5
CLUEWSC (EM)	5-shot	83.5	82.2	85.2
Code & Math				
BigCodeBench (Pass@1)	3-shot	63.9	56.8	59.2
HumanEval (Pass@1)	0-shot	62.8	69.5	76.8
GSM8K (EM)	8-shot	91.1	90.8	92.6
MATH (EM)	4-shot	60.5	57.4	64.5
MGSM (EM)	8-shot	81.3	85.7	84.4
CMath (EM)	3-shot	92.6	93.6	90.9

Benchmark (Metric)	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High	K2.6 Thinking	GLM-5.1 Thinking	DS-V4-Pro Max
MMLU-Pro (EM)	89.1	87.5	91.0	87.1	86.0	87.5
SimpleQA-Verified (Pass@1)	46.2	45.3	75.6	36.9	38.1	57.9
Chinese-SimpleQA (Pass@1)	76.4	76.8	85.9	75.9	75.0	84.4
GPQA Diamond (Pass@1)	91.3	93.0	94.3	90.5	86.2	90.1
HLE (Pass@1)	40.0	39.8	44.4	36.4	34.7	37.7
LiveCodeBench (Pass@1)	88.8	-	91.7	89.6	-	93.5
Codeforces (Rating)	-	3168	3052	-	-	3206
HMMT 2026 Feb (Pass@1)	96.2	97.7	94.7	92.7	89.4	95.2
IMOAnswerBench (Pass@1)	75.3	91.4	81.0	86.0	83.8	89.8
Apex (Pass@1)	34.5	54.1	60.9	24.0	11.5	38.3
Apex Shortlist (Pass@1)	85.9	78.1	89.1	75.5	72.4	90.2
Long Context						
MRCR 1M (MMR)	92.9	-	76.3	-	-	83.5
CorpusQA 1M (ACC)	71.7	-	53.8	-	-	62.0
Agentic						
Terminal Bench 2.0 (Acc)	65.4	75.1	68.5	66.7	63.5	67.9

Benchmark (Metric)	V4-Flash Non-Think	V4-Flash High	V4-Flash Max	V4-Pro Non-Think	V4-Pro High	V4-Pro Max
SWE Multilingual (Resolved)	69.7	70.2	73.3	69.8	74.1	76.2
BrowseComp (Pass@1)	-	53.5	73.2	-	80.4	83.4
HLE w/ tools (Pass@1)	-	40.3	45.1	-	44.7	48.2
MCPAtlas (Pass@1)	64.0	67.4	69.0	69.4	74.2	73.6
GDPval-AA (Elo)	-	-	1395	-	-	1554
Toolathlon (Pass@1)	40.7	43.5	47.8	46.3	49.0	51.8

[🔗](#) Chat Template

This release does not include a Jinja-format chat template. Instead, we provide a dedicated encoding folder with Python scripts and test cases demonstrating how to encode messages in OpenAI-compatible format into input strings for the model, and how to parse the model's text output. Please refer to the [encoding](#) folder for full documentation.

A brief example:

```
from encoding_dsv4 import encode_messages, parse_message_from_completion

messages = [
    {"role": "user", "content": "hello"},
    {"role": "assistant", "content": "Hello! I am DeepSeek.", "reasoning": "Hello! I am DeepSeek."},
    {"role": "user", "content": "1+1=?"}
]

# messages -> string
prompt = encode_messages(messages, thinking_mode="thinking")
```

```
# string -> tokens
import transformers
tokenizer = transformers.AutoTokenizer.from_pretrained("deepseek-ai/DeepSeek-V4")
tokens = tokenizer.encode(prompt)
```

How to Run Locally

Please refer to the [inference](#) folder for detailed instructions on running DeepSeek-V4 locally, including model weight conversion and interactive chat demos.

For local deployment, we recommend setting the sampling parameters to temperature = 1.0, top_p = 1.0. For the Think Max reasoning mode, we recommend setting the context window to at least **384K** tokens.

License

This repository and the model weights are licensed under the [MIT License](#).

Citation

```
@misc{deepseekai2026deepseekv4,
  title={DeepSeek-V4: Towards Highly Efficient Million-Token Context Processing},
  author={DeepSeek-AI},
  year={2026},
}
```

Contact

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

 System theme

Company

TOS

Privacy

About

Careers

Website

Models

Datasets

Spaces

Pricing

Docs

