DeepSeek V4

Technical Documentation

Publication date: April 27, 2026

Updated date: April 24, 2026

DeepSeek V4 - Model Card

General Information

Model Provider	DeepSeek AI
Model name	DeepSeek V4, including: • DeepSeek-V4-Pro • DeepSeek-V4-Flash
Release date	April 24, 2026
Model dependencies	N/A

Model properties

Model architecture	DeepSeek V4 is a Mixture-of-Experts (MoE) language model series that retains the DeepSeekMoE framework and Multi-Token Prediction (MTP) strategy from DeepSeek-V3, while introducing several key architectural innovations: (1) Hybrid Attention Architecture: combines Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA) to dramatically improve long-context efficiency. CSA compresses KV caches along the sequence dimensionand applies DeepSeek Sparse Attention (DSA); HCA applies heavier compressionwith dense attention. (2) Manifold-Constrained Hyper-Connections (mHC): constrains residual mapping onto the manifold of doubly stochastic matrices (Birkhoff polytope), enhancing signal propagation stability while preserving model expressivity. (3) Muon Optimizer: employs the Muon optimizer for faster convergence and improved training stability.
Modalities	Text, three reasoning modes available: • Non-think: Fast, intuitive responses • Think High: Deliberate logical analysis • Think Max: Extended reasoning at maximum capacity
Context length	1M
Total model size	Pro: 1.6T parameters, of which 49B are activated for each token Flash: 285B parameters, of which 13B are activated for each token

Methods of distribution and licenses

Distribution channels	DeepSeek V4 is distributed primarily through two channels to accommodate different deployment needs: Open-source Repositories and application program interface(API).
License	MIT License: The assets distributed via open-source repositories (including model weights and code) are licensed under the MIT License as a free and open-source license.

Use

	API deployment	Open-source deployment
Acceptable Use Policy	Access to the model via the API is governed by the DeepSeek Open Platform Terms of Service.	This repository and the model weights are licensed under the MIT License.
Intended uses	DeepSeek-V4, as a general-purpose AI model, aims to balance reasoning capabilities with output length, making it suitable for everyday use, such as question-answering scenarios and general agent tasks.	
Technical means for model integration	The technical means required for the model to be integrated into AI Systems or with models are described in the DeepSeek API Docs.	Open-source deployment: The technical means required for the model to be integrated into AI Systems or with models are described in the model card available on the relevant open-source repository such as Hugging Face.
Required hardware	N/A	See the model card available on Hugging Face.
Required software	N/A	See the model card available on Hugging Face.

Model data

The capabilities of DeepSeek models are built on high-quality, large-scale, and diverse data sources. We place great emphasis on and strictly comply with laws and regulations related to intellectual property, trade secrets, and personal privacy, ensuring that all data acquisition and usage occur within a legal and compliant framework.

1. Pre-training Phase

During the pre-training phase, corpus data is required for training. This phase primarily uses the following two categories of data:

- Public Data: We use publicly available information on the internet to build the model's broad understanding of world knowledge. We employ technical methods to acquire and filter these freely accessible data to enrich the model's knowledge base.

- Licensed Data: We collaborate with third-party data providers to obtain proprietary datasets through legally signed agreements. We ensure all collaborations are based on lawful authorization.

The pre-training phase does not require personal information for training. Therefore, we do not intentionally collect personal information to associate with any specific account or individual, nor do we proactively use it to train our models. We exclude sensitive information, credit card numbers, or unique identification information from our training data sources to minimize the risk of collecting any personal information. However, due to the vast scale of pre-training data, some publicly available online content or licensed data from other providers may incidentally contain personal information. We employ technical measures to screen and remove such information from the training data as much as possible and conduct tests before using the data for training.

2. Optimization Training Phase

During the optimization training phase, we typically need to construct or annotate a set of question-answer pair data manually or automatically to train the model. These question-answer pairs are produced by our research team, with a small portion potentially based on

user input. If user input is used to construct training data, we apply secure encryption, strict de-identification, and anonymization to make it cannot be linked to any specific individual.
