DeepSeek V4

技术文档

发布日期：4 月 27 日, 2026

更新日期: 4 月 24 日, 2026

DeepSeek V4 - 模型卡

基本信息

模型提供方	DeepSeek AI
模型名称	DeepSeek V4，包含： • DeepSeek-V4-Pro • DeepSeek-V4-Flash
发布日期	2026 年 4 月 24 日
模型依赖	不适用

模型属性

模型架构	DeepSeek V4 是基于混合专家（MoE）架构的大语言模型系列，保留了 DeepSeek-V3 的 DeepSeekMoE 框架和多 Token 预测（MTP）策略，同时引入以下关键架构创新： （1）混合注意力架构：结合压缩稀疏注意力（CSA）和深度压缩注意力（HCA）。CSA 沿序列维度压缩缓存（KV cache）并结合稀疏注意力（DSA）；HCA 以更大压缩率进行深度压缩但保持密集注意力。 （2）流形约束超连接（mHC）：残差映射约束到双随机矩阵流形（Birkhoff 多面体），增强层间信号传播的稳定性，同时保持模型表达能力。 （3）Muon 优化器：采用 Muon 优化器实现更快的收敛速度和更好的训练稳定性。
模态	文本，支持三种推理模式： • 非思考模式（Non-think）：快速、直觉式回答 • 高思考模式（Think High）：有意识的逻辑分析 • 极限思考模式（Think Max）：最大推理深度的扩展思考
上下文长度	1M
模型总参数量	Pro: 1.6T 参数量，每个 token 激活 49B 参数 Flash：285B 参数量，每个 token 激活 13B 参数

分发方式与许可证

分发渠道	DeepSeek V4 主要通过两个渠道进行分发，以满足不同的部署需求：开源代码库和应用程序接口（API）。
许可证	MIT 许可证：通过开源代码库分发的资产（包括模型权重和代码）采用 MIT 许可证，该许可证属于自由开源许可证。

使用

	API 部署	开源部署
可接受使用政策	通过 API 访问模型受 DeepSeek 开放平台服务协议的约束。	本仓库和模型权重采用 MIT 许可证授权。
预期用途	DeepSeek-V4 作为通用人工智能模型，旨在平衡推理能力与输出长度，适用于日常使用场景，如问答场景和通用智能体任务。	
模型集成的技术方式	模型集成到 AI 系统或与其他模型配合所需的技术方式在 DeepSeek API 文档中描述。	开源部署：模型集成到 AI 系统或与其他模型配合所需的技术方式在相关开源仓库（如 Hugging Face）上的模型卡片中描述。
所需硬件	不适用	请参阅 Hugging Face 上的模型卡片。
所需软件	不适用	请参阅 Hugging Face 上的模型卡片。

模型数据

DeepSeek 模型的能力建立在高质量、大规模和多样化的数据来源之上。我们高度重视并严格遵守与知识产权、商业秘密和个人隐私相关的法律法规，确保所有数据的获取和使用均在合法合规的框架内进行。

1. 预训练阶段

在预训练阶段，需要语料数据进行训练。该阶段主要使用以下两类数据：

- 公开数据：我们使用互联网上公开可用的信息来构建模型对世界知识的广泛理解。我们采用技术手段获取和过滤这些可自由获取的数据，以丰富模型的知识库。
- 授权数据：我们与第三方数据提供商合作，通过合法签署的协议获取专有数据集。我们确保所有合作均基于合法授权。

预训练阶段不需要个人信息用于训练。因此，我们不会刻意收集个人信息以将其与任何特定账户或个人关联，也不会主动使用个人信息来训练模型。我们从训练数据源中排除敏感信息、信用卡号或唯一标识信息，以最大程度降低收集任何个人信息的风险。然而，由于预训练数据规模庞大，一些公开可用的在线内容或来自其他提供商的授权数据可能偶然包含个人信息。我们采用技术手段尽可能从训练数据中筛选和删除此类信息，并在使用数据进行训练前进行测试。

2. 优化训练阶段

在优化训练阶段，我们通常需要手动或自动构建或标注一组问答对数据来训练模型。这些问答对由我们的研究团队制作，其中一小部分可能基于用户输入。如果使用用户输入来构建训练数据，我们会进行安全加密、严格去标识化和匿名化处理，使其无法与任何特定个人关联。