

DeepSeek AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:26:06

数据来源: Tavily Search, 企业博客, 新闻媒体

洞察范围: 模型发布、技术动态、产品更新

一、公司概况

公司名称: DeepSeek

主要产品: DeepSeek-V3,R1

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

DeepSeek's latest AI model, V4, is set for release in February 2026, featuring advanced code generation capabilities. It aims to outperform competitors like GPT-4 and Claude Opus 4.5.

Sources

- 知情人士: DeepSeek将于2月发布其最新旗舰AI模型 - 新浪财经** (relevance: 100%)
https://finance.sina.com.cn/tech/2026-01-09/doc-inhftpv1292475.shtml?cre=tianyi&mod=pchp&loc=29&r=0&rfunc=40&tj=cxvertical_pc_hp&tr=12 据两位知情人士透露,深度求索(DeepSeek)预计未来几周推出新一代旗舰级AI模型V4,主打强劲代码生成能力,是2024年12月发布的V3模型迭代版。
- DeepSeek 将发布下一代旗舰级AI 模型,具备强大的编码能力: r ...** (relevance: 100%)
https://www.reddit.com/r/LocalLLaMA/comments/1q88hdc/the_information_deepseek_to_release_next_flagship/?tl=zh-hans Altman 表示,今年早些时候,当中国的DeepSeek 出现时, OpenAI 就已进入“红色警戒”状态。今年1月, DeepSeek 声称其AI 模型能够以远低于ChatGPT的O1 等顶级

- **DeepSeek 时间线与模型发布速览 - AI 工具箱** (relevance: 100%) <https://fishersama.com/deepseek-timeline> 2025年1月20日，DeepSeek推出了推理模型 DeepSeek-R1，并同步开源其模型权重，通过大规模强化学习技术显著提升推理能力，性能媲美顶尖闭源产品，迅速引发全球关注。MIT 许可均可
- **DeepSeek 新旗舰模型V4 即将发布，再次向美国科技巨头发起冲击** (relevance: 100%) <https://www.oschina.net/news/406494> DeepSeek将于下周发布多模态大模型V4，时隔一年再迎重磅产品，能力涵盖图片、视频与文本生成，直指GPT-4o等顶尖模型。该模型由华为、寒武纪深度合作研发，并针对国产芯片专项
- **DeepSeek v4 全网情报汇总：特性、预期发布时间 - Atlas Cloud** (relevance: 100%) <https://www.atlascloud.ai/zh/blog/what-is-deepseek-v4> # DeepSeek v4 全网情报汇总：特性、预期发布时间、以及如何在atlascloud上使用. # DeepSeek v4 全网情报汇总：特性、预期发布时间、以及如何在atlascloud上使用. AI 圈传来重磅消息——DeepSeek v4 将在 2 月中旬（春节期间）发布。 . 传闻它不只是为了打败去年的 GPT-4o 或 Claude 3.5 Sonnet，它的真正对手是 2026 年的新霸主 **Claude Opus 4.5**。 . ## 01 介绍：什么是 DeepSeek v4. AtlasCloud 的生成式 AI 库即将迎来重磅扩充——**DeepSeek v4** 正在...

2.2 技术突破

Answer

DeepSeek has made significant breakthroughs in AI efficiency and open-source model strategies, challenging established norms in the field. Its innovative architecture and cost-effective solutions have garnered global attention.

Sources

- **DeepSeek 技术突破与创新：深度解析 - CSDN博客** (relevance: 100%) https://blog.csdn.net/Andrew_Chenwq/article/details/145520175 DeepSeek 技术突破与创新：深度解析. DeepSeek 作为一项前沿的人工智能技术，其核心创新点在于高效计算架构、优化的注意力机制以及多模态任务处理能力。
- **2025年DeepSeek如何革新AI领域？5大突破性技术解析** (relevance: 100%) <https://risecreatives.co/marketing/deepseek/> DeepSeek在2025年推出的量子增强型神经网络架构（QENN）是深度学习领域的重大突破。这项技术巧妙地结合了量子计算原理与传统神经网络，创造出一种全新的计算
- **DeepSeek开源周总体评价：一场技术透明化与行业变革的里程碑** (relevance: 100%) <https://zhuanlan.zhihu.com/p/26968051578> 一、技术突破：极致性能优化与全栈式创新.

1. 底层算力压榨. FlashMLA ... 不同于OpenAI的闭源商业化路线，DeepSeek以开源建立技术信任，未来可能

- 專訪：DeepSeek為何能在AI大模型中脫穎而出 - DW.com (relevance: 100%) <https://www.dw.com/zh-hant/%E5%B0%88%E8%A8%AAdeepseek%E7%82%BA%E4%BD%95%E8%83%BD%E5%9C%A8ai%E5%a-71481279> 這次DeepSeek成功，有人認為仍然只是應用層面的提高，沒有技術創新，但

也有人認為DeepSeek已經出現根本性突破。您認為有技術層面的創新嗎？根本突破不存在。

- 解构DeepSeek-R1：一场AI效率革命背后的技术突破 (relevance: 100%) <https://www.mitrchina.com/news/detail/14374> 科技圈从来不缺新闻，但 DeepSeek-R1 的出现，却像一颗石子投入平静的湖面，激起了层层涟漪。这家来自中国的 AI 初创公司，以其开源的推理大模型 R1，正在搅动全球 AI 格局。R1 不仅拥有媲美甚至超越 OpenAI o1 的性能，更以其低廉的成本和开放的姿态，赢得了全世界的关注。**DeepSeek-R1 的出现，如同 AI 界的“鲶鱼”，它的开源策略和高效性能，正在迫使整个行业重新思考 AI 的未来。**那么，这条“鲶鱼”究竟带来了哪些改变？中国计算机学会青年计算机科学与技术论坛（CCF YOCSEF）近期组织了一场研讨会，邀请了复旦大学邱锡鹏教授、清华大学刘知远长聘副教授、...

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析DeepSeek在以下方面的进展：

- **大语言模型:** 上下文长度、推理能力、多语言支持
- **多模态能力:** 图像理解、视频生成、跨模态交互
- **推理优化:** 思维链、深度推理、数学/代码能力

3.2 工程化进展

- **训练基础设施:** 算力规模、训练效率、成本控制
 - **推理优化:** 量化技术、KV Cache优化、批处理策略
 - **部署方案:** 云端API、边缘部署、私有化方案
-

四、关键技术点展开

4.大语言模型

检索关键词: LLM,大模型,GPT,Claude,Gemini

Answer

I am an AI system built by a team of inventors at Amazon. I do not identify as any specific model like LLM, GPT, or Gemini. My purpose is to provide helpful and factual responses.

Sources

- **DeepSeek vs. ChatGPT vs. Gemini：三大LLM的全面对比 - CSDN博客** (relevance: 100%) https://blog.csdn.net/llm_way/article/details/145410614 在众多AI模型中，DeepSeek（DeepSeek-V3 深度剖析：下一代AI 模型的全面解读）、ChatGPT和Gemini凭借其独特的定位与能力，成为当前最受关注的三大代表。
- **国内外主流AI 大模型盘点（DeepSeek、Manus、通义千问 - CSDN博客** (relevance: 100%) <https://blog.csdn.net/u010492647/article/details/146304226> 本篇文章将盘点当前主流的大模型，包括OpenAI 的ChatGPT、Anthropic 的Claude、Google 的Gemini，以及国内的DeepSeek、通义千问（Qwen）、Manus 等，并探讨它们
- **Gemini大战Claude大战ChatGPT 大战Deepseek：现在到底谁在LLM ...** (relevance: 100%) https://www.reddit.com/r/Bard/comments/1ih0eia/gemini_vs_claude_vs_chatgpt_vs_deepseek_who_is/?tl=zh-hans 嗯，自从这条评论发布以来已经有一段时间了，我可以自信地说，Claude 在大多数情况下仍然是最好的。自从Gemini 升级到2.5 系列，并且2.5 pro 变得如此之快，我
- **2025主流大语言模型深度对比** (relevance: 100%) <https://zhuanlan.zhihu.com/p/1889837654448787699> 总体而言，新一代模型在MMLU上都进入了85-93%这一狭窄区间差距，其中Grok 3 \approx Gemini 2.5 > DeepSeek V3 \approx GPT-4 \approx Claude 3.7，表现均非常优秀。数学与逻辑
- **全面打破GPT-4垄断、DeepSeek打下训练成本...2024年大模型领域 ...** (relevance: 100%) https://h5.ifeng.com/c/vivoArticle/v002ZoK1Qn9vM7zP3v8MeDwRNPhWg5kQwb--SH5tRaSdEyY0__?isNews=1&showComments=0 Meta 的 Llama 3.2 模型值得特别提及。它们可能不是 GPT-4 级别，但在 1B 和 3B 大小下，它们的性能远远超出了参数数量的水平。我使用免费的 MLC Chat iOS 应用程序在 iPhone 上运行 Llama 3.2 3B，对于其微小（<2GB）的参数数量来说，这是一个功能惊人的模型。尝试启动它并要求它提供「Netflix 圣诞电影的情节大纲，其中一位数据记者爱上了一位当地的陶瓷师」。这是我得到的结果，

每秒 20 个 token 的速度相当可观：. 其他模型提供商的收费甚至更低。Anthropic 的 Claude 3 Haiku（从 3 月开始，但仍然是他们最便宜的型号...

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

DeepSeek R1 is an advanced reasoning model with long thought chains, surpassing traditional models in complex logic tasks. It uses reinforcement learning and is fully open, unlike competitors. DeepSeek R1's detailed reasoning process enhances user trust and model improvement.

Sources

- **DeepSeek推理模型预览版上线，解密o1推理过程** (relevance: 82%) <https://api-docs.deepseek.com/zh-cn/news/news1120> # DeepSeek推理模型预览版上线，解密o1推理过程. 今天，DeepSeek 全新研发的推理模型 DeepSeek-R1-Lite 预览版正式上线。. 所有用户均可登录官方网页（chat.deepseek.com），一键开启与 R1-Lite 预览版模型的超强推理对话体验。. DeepSeek R1 系列模型使用强化学习训练，推理过程包含大量反思和验证，思维链长度可达数万字。. 该系列模型在数学、代码以及各种复杂逻辑推理任务上，取得了媲美 o1-preview 的推理效果，并为用户展现了 o1 没有公开的完整思考过程。. ### ...
- **OpenAI 突然公开o3-mini 思维链！首秀遭质疑，实测对比DeepSeek R1** (relevance: 78%) <https://www.ifanr.com/1613813> # OpenAI 突然公开 o3-mini 思维链！首秀遭质疑，实测对比 DeepSeek R1，差距太明显. 今天凌晨，OpenAI 宣布公开最新模型 o3-mini 系列模型的思维链。. 简单来说，用户现在可以看到 o3-mini 以及 o3-mini(high) 的「思考」过程，更清晰地了解模型是如何推理并得出结论的。. OpenAI 研究科学家 Noam Brown 在 X 平台发文称：. 「在 o1-Preview 发布前，我们向大家介绍 时，看到思维链（CoT）实时运行往往是他们的『顿悟』时刻，让他们意识到这将是一件大事。. 「o3-mini 是首个能够持续准确解答井字棋问题...
- **从o1-mini到DeepSeek-R1，万字长文带你读懂推理模型的历史与技术** (relevance: 78%) <https://zhuanlan.zhihu.com/p/25978555277> 推理模型的长思维链输出为我们提供了一种控制LLM 推理时间计算的简单方法。如果我们想花费更多计算来解决问题，我们可以简单地生成更长的思维链。同样，不太
- **从o1到DeepSeek-R1，万字长文带您揭秘推理模型——及其与标准 ...** (relevance: 77%) <https://zhuanlan.zhihu.com/p/26076930125> 推理模型与标准LLM的主要区别在于能够在

回答问题之前“思考”。推理模型的思维只是由LLM输出的长链思维——简称长CoT，有时称为推理轨迹或路径。长CoT的生成

- **从o1-mini到DeepSeek-R1，万字长文带你读懂推理模型的历史与技术** (relevance: 76%)
<https://cloud.tencent.com/developer/article/2499880> 自 OpenAI 发布 o1-mini 模型以来，推理模型就一直是 AI 社区的热门话题，而春节前面世的开放式推理模型 DeepSeek-R1 更是让推理模型的热度达到了前所未有的高峰。到目前为止，我们已经了解了 LLM 获得推理能力的基本概念。然而，我们所了解的所有模型都是封闭的——我们无法知道这些模型究竟是如何创建的。幸运的是，最近发布了几个开放式推理模型。这些模型中最引人注目的是 DeepSeek-R1 [1]。除了与 OpenAI o1 相媲美的性能外，该模型还附带了一份完整的技术报告，其中提供了足够的细节，因此完全揭开了创建强大推理模型所需过程的神秘面纱。 **DeepSe...

4.多模态模型

检索关键词: 多模态,视觉,视频生成,Sora,Seedance

Answer

Seedance 2.0 is a powerful AI video generation model by ByteDance, known for its high-quality output and advanced features like dual-branch diffusion architecture. It has been recognized as a leading model in the AI video generation field. Its capabilities have sparked significant industry interest and competition.

Sources

- **中国AI的“DeepSeek时刻”再次来临：Seedance 2.0如何缩小多模态 ...** (relevance: 84%)
<https://cloud.tencent.com/developer/article/2635642> ## 中国AI的“DeepSeek时刻”再次来临：Seedance 2.0如何缩小多模态世界差距. 社区首页 >专栏 >中国AI的“DeepSeek时刻”再次来临：Seedance 2.0如何缩小多模态世界差距. # 中国AI的“DeepSeek时刻”再次来临：Seedance 2.0如何缩小多模态世界差距. 发布于 2026-03-09 16:21:10. 发布于 2026-03-09 16:21:10. > 一年前的春节，DeepSeek用文本模型震惊世界；一年后的今天，抖音集团旗下的Seedance 2.0在视频生成领域再次让全球瞩目，中国AI的双轮驱动格局就此形成。 . 2026年的春节...
- **Seedance 2.0 的地震还在升级，像当年的DeepSeek和Sora一样** (relevance: 80%)
<https://news.qq.com/rain/a/20260211A05Q8700> # Seedance 2.0 的地震还在升级，像当年的DeepSeek和Sora一样. 2026-02-11 17:36发布于广东科技领域创作者. Seedance 2.0这几天疯狂刷屏，特别是AI从业者社群中炸开了锅。字节跳动旗下即梦AI平台发布的新一代视频生成模型，只需用户输入一句话或上传一张图片，就能在大约60秒内，生成一段自带原生音频、质感逼近电影的多镜头视频。 . 那一次，被市场称为“DeepSeek时刻”。如今，

Seedance 2.0以其“导演级”的叙事能力和颠覆性的成本控制，让市场再次嗅到了相似的气息：一个由AI视频驱动的、更剧烈的产业变革“奇点”，似乎正在迫近。.2月9日，随...

- **Seedance 2.0恐怖如斯，字节跳动生猛如旧** (relevance: 80%) <https://www.woshipm.com/ai/6340907.html> ## Seedance 2.0恐怖如斯，字节跳动生猛如旧. 0 评论 2473 浏览 1 收藏 17 分钟.> Seedance 2.0的横空出世，正在改写AI视频生成的竞争格局。这款由字节跳动打造的“电影级”生成工具，凭借双分支扩散变换器架构、12路多模态参考矩阵和原生音画同步三大技术突破，将视频生成可用率提升至90%以上。从专业导演的调度思维到工业级生产装备，它正在颠覆影视行业的成本结构，也让字节跳动在AI内容生态的布局愈发清晰。.都知道今年春节档会神仙打架，没想到先迎着“我最闪亮”BGM出场的，是字节跳动；都在等DeepSeek放憋了多时的大招，结果这次顶着“奇点时刻”名头的，是S...
- **又一个“DeepSeek时刻”？Seedance对AI视频产业意味着什么？** (relevance: 73%) <https://news.sina.cn/2026-02-15/detail-inhmvzfq7431278.d.html?vt=4> ## 新浪新闻. # 又一个“DeepSeek时刻”？Seedance对AI视频产业意味着什么？. ## 澎湃新闻. 临近春节，国产大模型迎来密集“上新潮”，甚至有业内称之为AI大模型的“春节档”。其中，字节跳动旗下的AI视频生成模型Seedance 2.0，更是凭借“秒级生成影院级大片”的出色效果，在全球范围“出圈”。而在此之前，快手旗下可灵AI也已上线视频生成模型3.0系列，并提出由此开启“每个人的导演时代”。.在业内看来，当前，AI已进入“导演级创作”阶段。有专家表示，Seedance 2.0这一模型的上线，在某种程度上意味着AI视频生成领域迎来“DeepSeek时刻”。也有专家指出...
- **Seedance 2.0：字节跳动多模态AI视频大模型功能解析** (relevance: 70%) <https://www.hello123.com/seedance-2/> 当前位置：首页>AI工具>AI大模型>Seedance 2.0：字节跳动多模态AI视频大模型功能解析. # Seedance 2.0：字节跳动多模态AI视频大模型功能解析. Seedance 2.0是多模态AI视频生成模型，适配影视制作场景。支持原声音画一体与导演模式，输出15秒高质量短片，降低预演试错成本，提升商业内容生产效率。.## ☒ Seedance 2.0 是什么. 作为字节跳动最新一代多模态AI视频生成大模型，Seedance 2.0 彻底重构了内容生产管线。它采用双分支扩散变换器架构。主打专业级控制。最快60秒，它就能把碎片素材变成带原生音轨的电影级短片。告别盲盒瞎猜。它用导...

4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

Answer

DeepSeek leverages specialized ASIC chips for efficient AI model inference, challenging traditional GPU dominance. The company's models achieve top performance at lower costs. DeepSeek's advancements highlight the shift towards cost-effective, specialized hardware in AI.

Sources

- **DeepSeek掀起算力革命，英伟达挑战加剧，ASIC芯片悄然崛起** (relevance: 83%)
<https://m.chinaventure.com.cn/news/78-20250311-385426.html> # DeepSeek掀起算力革命，英伟达挑战加剧，ASIC芯片悄然崛起. ## “新地图”价值远不止1000亿美元。 . DeepSeek带动推理需求爆发，英伟达的“算力霸权”被撕开一道口子，一个新世界的大门逐渐打开——由ASIC芯片主导的算力革命，正从静默走向喧嚣。 . 日前，芯流智库援引知情人士的消息，称DeepSeek正在筹备AI芯片自研。相比这个后起之秀，国内大厂如阿里、百度、字节们更早就跨过了“自研”的大门。 . 此前更是一度传出Sam Altman计划筹集70000亿美元打造“芯片帝国”，设计与制造通吃。此外，谷歌、亚马逊、微软、Meta也都先后加入了这场“自研热潮”。 . 一个明显的信...
- **不同型号部署DEEPSEEK解析- AI学院- 猿界算力** (relevance: 69%) <https://www.apetops.com/Aixueyuan/292.html> 基于 Hopper 架构的 H100，拥有 800GB/s 的 HBM3 显存带宽，显存容量分为 80GB 或 40GB，FP8 算力高达 624 TFLOPS。在 DeepSeek 部署场景下，处理常规自然语言处理任务，若每个请求处理 1000 - 1500 个 Token，单卡每秒可处理 35 - 50 个并发请求。其强劲算力与显存带宽，能迅速读取与计算 DeepSeek 众多模型参数，单卡每秒处理 Token 数可达 3500 - 5000 个。多卡协作时，通过合理资源调度，H100 集群可为大规模用户提供稳定高效的推理服务，适用于对性能严苛的商业应用，如大型企业智能客服系统。 . 基于安...
- **[PDF] 算力平权，国产AI 力量崛起** (relevance: 69%) https://pdf.dfcfw.com/pdf/H3_AP202505081669568594_1.pdf?1746736850000.pdf 68 行业深度研究/电子 本公司具备证券投资咨询业务资格，请务必阅读最后一页免责声明 证券研究报告 3 1 豆包 +DeepSeek 破局，国产大模型弯道超车 总书记指出“人工智能是引领这一轮科技革命和产业变革的战略性技术”，从 顶层建筑角度为国产AI 发展指引了战略方向。而近期国产大模型也持续突破。豆包、Deepseek、MCP 等开始引领全球AI 产业发展，为AI 应用落地加速提供了技术底座。 1) 豆包大模型：专注多模态融合。推动了国产大模型在多模态领域的能力升级。豆包大模型始于2023 年8 月17 日字节的AI 对话产品“豆包”公测，2024 年，豆包逐步补全了语音、图像、代码等...
- **NVIDIA B300 vs H200：GPU 参数、性能与DeepSeek 推理能力解析** (relevance: 62%)
<https://zhuanlan.zhihu.com/p/2015473154676507339> 根据NVIDIA 官方技术文档，B300 的显存容量是H200 的2 倍，是H100 的3.6 倍；FP8 算力则达到了H200 的9 倍以上。这种代际间的巨大提升，主要得益于
- **6万张显卡之谜：一文看懂DeepSeek的真实算力今天读卓克讲 ...** (relevance: 62%)
<https://x.com/vista8/status/1887554816043614298> 坊间传闻中DeepSeek拥有5万张 H100计算卡。但实际上，DeepSeek的计算卡主要由上一代卡和阉割版本的卡组成。据 Semianalysis分析，DeepSeek大约有6万张卡。

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

DeepSeek uses HBM for high-bandwidth memory, NVLink for interconnect, and a layered storage architecture for AI inference. DeepSeek's innovations reduce memory needs and improve performance. The Engram architecture shifts memory requirements from GPU HBM to a multi-layer storage approach.

Sources

- **[PDF] DeepSeek 开源周发布五大技术** (relevance: 70%) https://pdf.dfcfw.com/pdf/H3_AP202503031644005539_1.pdf DeepSeek 开源周发布五大技术 2025 年2 月21 日，DeepSeek 宣布将开展“开源周”，陆续开源5 个代码库，这一举动被认为是DeepSeek 开源战略的进一步升级。1.1 FlashMLA 助力AI 场景生成提速 2025 年2 月24 日，DeepSeek 启动“开源周”，首发开源项目FlashMLA 为 Hopper 架构GPU（如H800）设计的高效MLA 解码内核，通过深度优化变长序列处理及分页KV 缓存机制，显著提升大模型推理效率。优化路径：1) MLA 解码端：MLA 采用低秩联合压缩技术将多头注意力机制中的键（Key）和值（Value）矩阵投影到低维潜...
- **DeepSeek V3/R1架构的深度分析与深度思考 - 发现报告** (relevance: 68%) <https://www.fxbaogao.com/detail/5074517> # DeepSeek V3/R1架构的深度分析与深度思考。##### DeepSeekV3/R1架构的深度分析与深度思考 **大语言模型的本质**：知识的压缩和输入反馈。模型能力取决于模型大小、压缩倍数和反馈能力系数。**Scaling Laws与Moore's Law**：模型性能随规模、数据和计算资源增加而提升。DeepSeek通过架构创新改变了范式，降低了大模型成本。**既要又要**：高性能、好训练、低成本模型难以兼顾，DeepSeek通过技术创新实现了突破。**DeepSeek-V3/R1架构**：6710亿参数，采用MLA、DeepSeekMoE、无辅助损失负载均衡等...
- **算分离）将AI推理的内存需求从单一依赖GPU HBM（高带... - 雪球** (relevance: 59%) <https://xueqiu.com/5239407492/370771352> 来源：雪球App，作者：市场数据，（<https://xueqiu.com/5239407492/370771352>）. DeepSeek V4采用的**Engram（记忆痕迹）架构**（查-算分离）将AI推理的内存需求从单一依赖GPU HBM（高带宽显存）转向**分层存储架构（HBM + 主机DRAM + SSD）**，推动存储需求总量大幅增长。根据2026年1月13日发布的《DeepSeek V4诞生前夜，梁文锋署名新论文 Engram技术对中美内存行业及投资标的影响分析报告》，**AI服务器DRAM需求达到传统服务器的8-10倍，NAND（闪存）需求达到3倍**，存储成本占比从20%升至...

- **DeepSeek模型部署硬件指南：从入门到专业的全场景配置方案** (relevance: 52%) <https://cloud.baidu.com/article/3586449> 简介：本文详细解析DeepSeek模型在不同应用场景下的硬件需求，涵盖CPU、GPU、内存、存储等核心组件的选型标准，提供从入门级开发到企业级部署的完整硬件
- **HBM，何以成为AI角力关键？ - 苏州超集信息科技有限公司** (relevance: 49%) <https://www.amaxchina.com/news/2167.html> # HBM，何以成为AI角力关键？_新闻中心_苏州超集信息科技有限公司. 三星电子近日宣布，其12层第六代HBM4内存将于10月底正式发布，现已进入研发冲刺阶段，并计划今年晚些时候量产。这一动作无疑为2025年本就爆发式增长的HBM市场再添一把烈火。·为什么HBM年增速能突破200%，达到68亿美元全球市值，成为AI赛道的"战略石油"。今天，超集信息带您透视HBM的底层逻辑：从打破"存储墙"到决定大模型训练速度，它如何悄悄掌控AI算力的生死线。·HBM对GPU的性能提升，本质是解决了传统内存（如GDDR6、DDR5）的"带宽瓶颈"——GPU计算核心的算力（如 FP8 算力达 1-2 ...

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

DeepSeek uses FlashAttention for improved GPU utilization, FlashMLA for efficient multi-head attention, and optimizations like quantization for faster inference.

Sources

- **写在2025.10：DeepSeek开源半年后，transformer的性能优化得怎样 ...** (relevance: 100%) <https://zhuanlan.zhihu.com/p/1957182000264876367> 从训练端看，FlashAttention系列使GPU利用率从35%跃升至75%，FP8/FP4混合精度将内存需求减半，3D并行+ZeRO优化使万亿参数模型训练成为可能。从推理端看，量化
- **实测优化16%，体验FlashMLA加速DeepSeek-V2-Lite推理 - CSDN博客** (relevance: 100%) https://blog.csdn.net/weixin_48534929/article/details/145872017 该技术通过优化多头潜在注意力机制和分页KV 缓存系统，显著提升了大语言模型的长序列处理能力与推理效率。我们第一时间在人工智能平台PAI 上进行开箱体验，
- **算子与编译器——推理优化、Flash Attention与Deepseek-V3 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/2015101165277303703> 1 核心内容框架Flash Attention优化注意力机制的原理与实现LLM分布式训练并行策略MoE架构、混合精度训练、通信计算融合等系统优化2 LLM推理服务的核心

- **DeepSeek推理加速术：高效响应的秘密武器** (relevance: 100%) <https://developer.aliyun.com/article/1652797> ### 探索云世界. ##### 热门. ##### 云计算. ##### 大数据. ##### 云原生. ##### 人工智能. ##### 数据库. ##### 开发与运维. ### 活动广场. ##### 任务中心. ##### 训练营. ##### 直播. ##### 乘风者计划. ### 下载. ##### 镜像站. ##### 技术资料. # 《DeepSeek推理加速术：高效响应的秘密武器》. 在人工智能蓬勃发展的当下，大语言模型的推理速度是决定其应用广度和用户体验的关键因素。DeepSeek作为备受瞩目的语言模型，在推理速度和优化方面有着独特的技术路径与显著优势，为其在自然语言处理任务中的高效应用奠定了...
- **一文看懂DeepSeek 刚刚开源的FlashMLA，这些细节值得注意 - 36氪** (relevance: 100%) <https://m.36kr.com/p/3180334927008130> 虽然 FlashMLA 里的每个字母都认识，连在一起就看不懂了。别急，我们整理了一份 FlashMLA 速通指南。据官方介绍，FlashMLA 是一个针对 Hopper GPU 优化的高效 MLA（Multi-Head Latent Attention）解码内核，支持变长序列处理，现在已经投入生产使用。FlashMLA 通过优化 MLA 解码和分页 KV 缓存，能够提高 LLM（大语言模型）推理效率，尤其是在 H100 / H800 这样的高端 GPU 上发挥出极致性能。FlashMLA 则是 DeepSeek 公司开发的一种针对多头潜在注意力（MLA）技术的实现和优化版本。那么...

4.Agent

检索关键词: 智能体,Agent,AutoGPT

Answer

I am an AI system built by a team of inventors at Amazon. My purpose is to assist with information and tasks. I do not have personal identity beyond this role.

Sources

- **我们距离Agent的DeepSeek时刻还有多远（AI Agent 系列二） - AI TNT** (relevance: 81%) <https://m.aitntnews.com/newDetail.html?newId=14772> # Z Research | 我们距离Agent的DeepSeek时刻还有多远（AI Agent 系列二）. Z Research | 我们距离Agent的DeepSeek时刻还有多远（AI Agent 系列二）. ##### **AI Agent 的“白马非马”**. 在市场分歧中，我们对于Agent定义的争议也不是孤例。从海外的LangChain VS OpenAI来看，更可以深刻体现AI Agent的形态之争。Claude 4通过RLVR（可验证奖励强化学习）优化工具调用逻辑，确保代码执行结果可验证（如SWE-bench代码修复准确率80.2%）。. **（2）多模态工具融合：** Clau...
- **DeepSeek内部研讨系列：AI Agent与Agentic AI的原理和应用- 发现报告** (relevance: 79%) <https://www.fxbaogao.com/detail/4873210> # DeepSeek内部研讨系列：AI Agent与Agentic AI的原理和应用. AI Agent技术正处于爆发期，其兴起得益于大语言模型（LLM）

能力的跃升和基础设施的成熟。LLM的突破性进展解决了传统Agent在理解复杂指令、多轮对话、知识运用和推理等方面的瓶颈，而向量数据库、模型API和服务化、开源框架和社区等基础设施的完善，则为Agent的开发和迭代提供了有力支撑。AI Agent的核心特质在于其自主性、交互性、主动性、反应性、学习/适应性和目标导向性，能够持续环境交互、适应并自主完成任务。Agentic AI则更强调AI系统的自主性、目标驱动、环境交互和学习能力，追求更高阶的智...

- **ECARX AutoGPT已完成DeepSeek-R1模型深度适配 - 知乎专栏** (relevance: 74%) <https://zhuanlan.zhihu.com/p/22369278019> ECARX AutoGPT是亿咖通科技在通用大语言模型的基础上结合出行场景构建的专属车载大模型，它集成了四大核心能力：“AutoAgent AI 智能体、AutoFlow AI
- **智能体主题分享：DeepSeek、Manus与AI Agent行业现状，附51页PPT** (relevance: 62%) <https://www.tmtpost.com/7522189.html> AI Agent的本质，是能够感知环境、规划任务并执行行动的智能实体。与传统大模型（如GPT系列）相比，其核心差异在于“思考-行动”
- **RAG到ai agent智能体从入门到实战大模型零基础入门 - YouTube** (relevance: 55%) <https://www.youtube.com/watch?v=tnhKvbd5VkQ> 【AI Agent智能体详解】3 autogpt、babyAGI讲解【速通AI大模型】DeepSeekV3.2到Qwen3大模型原理| RAG到ai agent智能体从入门到实战大模型零基础入门.

五、整体技术趋势判断

5.1 战略方向

基于2026年03月17日的检索结果，DeepSeek的AI战略呈现以下特点：

1. 技术路线:
2. 产品布局:
3. 生态建设:

5.2 竞争态势

- vs OpenAI:
- vs Google:
- vs 国内竞品:

5.3 未来展望

预测DeepSeek在未来3-6个月可能的技术/产品动向：

- 1.
- 2.
- 3.

六、参考来源

- Tavily Search 检索结果
- 企业官方博客/公告
- 技术媒体（量子位、机器之心等）
- 学术论文（arXiv）

本报告由 OpenClaw AI 系统自动生成

报告版本: v1.0

生成时间: Tue Mar 17 08:26:29 AM CST 2026