

# NVIDIA AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:26:50

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: NVIDIA

主要产品: H100,B200,CUDA

检索优先级: 高

## 二、最新动态检索

### 2.1 产品/模型发布

#### Answer

In 2024, NVIDIA released a new AI model called Fugatto for audio content creation and modification. The model can generate and alter sounds in various languages and accents. It is still under consideration for public release.

#### Sources

- **AI 模型 - NVIDIA 开发者** (relevance: 100%) <https://developer.nvidia.cn/ai-models> 早在 2016 年, NVIDIA 和 OpenAI 就发布了 NVIDIA DGX™, 开始突破 AI 的界限。随着 OpenAI gpt-oss-20b 和 gpt - oss-120b 的发布, 协作式 AI 创新得以延续。NVIDIA 已在 NVIDIA Blackwell 架构上优化了这两个新的开放权重模型, 以加速推理性能, 在 NVIDIA GB200 NVL72 系统上每秒可提供高达 150 万个 token (TPS)。. 阅读博客: NVIDIA 在 NVIDIA GB200 NVL72 上提供 150 万 TPS 推理, 加速从云到边缘的 OpenAI gpt-oss 模型. 在 O...
- **Nvidia AI | Nvidia新AI模型登場可修改兼生成聲音 - HKET經濟日報** (relevance: 100%) <https://inews.hket.com/article/3862970/Nvidia%20AI%E6%B0%A1%E5%9E%8B%E7%99%BB%E5%A09>

昔日新聞 電子報 hket 訂戶/會員專區. 專題:付費文章免費試閱 hketPRO 讓專業投資 觸手可及. \* 新春加碼賞 首月 hketPRO 僅需 \$10! 再送李錦記「胡麻醬雞撈麵」及「味噌扇貝湯麵」各一盒!! 熱門關鍵字: ATMX 新經濟股 收息 騰訊 阿里巴巴 滙豐 上車盤 退休規劃 ETF. # Nvidia AI | Nvidia 新 AI 模型登場 可修改兼生成聲音. 發布時間: 2024/11/26 11:16. 最後更新: 2024/11/26 11:32. 【Nvidia / 黃仁勳 / 輝達 / 人工智能 / AI】Nvidia 創辦人兼行政總裁黃仁勳近日來港, 大談 AI...

- **NVIDIA 发布开放模型和数据, 加速语言、生物学和机器人领域的AI 创新** (relevance: 100%) <https://nvidia.csdn.net/6982a394a16c6648a98728cb.html> # logo NVIDIA AI 技术专区. ### NVIDIA AI 技术专区. NVIDIA AI 技术专区 NVIDIA 发布开放模型和数据, 加速语言、生物学和机器人领域的 AI 创新. # NVIDIA 发布开放模型和数据, 加速语言、生物学和机器人领域的 AI 创新. ### NVIDIA AI 技术专区. NVIDIA AI 技术专区 • 2026-02-04 09:40:33 发布. NVIDIA 的开放模型系列, 包括面向数字 AI 的 NVIDIA Nemotron、面向物理 AI 的 Cosmos、面向机器人开发的 Isaac GR00T 以及面向生物医学 AI 的 ...
- **NVIDIA NIM 开发平台, 提供超多免费大模型- iChochy - 博客园** (relevance: 100%) <https://www.cnblogs.com/ichochy/p/19438583> NVIDIA NIM (NVIDIA Inference Microservices) 是英伟达推出的 AI 推理微服务, 用于把 AI 模型 (尤其是大模型) 快速、稳定、高性能地部署成可用的在线服务。
- **NVIDIA 发布全新开放模型、数据和工具, 推动各行业AI 技术的发展** (relevance: 100%) <https://blogs.nvidia.cn/blog/open-models-data-tools-accelerate-ai/> # NVIDIA 发布全新开放模型、数据和工具, 推动各行业 AI 技术的发展. 这些模型包括适用于代理式 AI 的 NVIDIA Nemotron 系列、适用于物理 AI 的 NVIDIA Cosmos 平台、适用于辅助驾驶汽车开发的全新 NVIDIA Alpamayo 系列、适用于机器人的 NVIDIA Isaac GR00T 以及适用于生物医学的 NVIDIA Clara, 它们将为企业提供构建真实世界 AI 系统所需的技术工具. . ## **NVIDIA Nemotron 赋予 AI 智能体语音、多模态智能和安全能力.** 基于近期发布的 NVIDIA Nemotron 3 系列开放模型与...

## 2.2 技术突破

### Answer

---

NVIDIA achieved a major breakthrough with DLSS 5, a revolutionary real-time ray tracing technology. This advancement significantly enhances graphics rendering performance.

---

### Sources

---

- **三星都被黃仁勳點名了! NVIDIA GTC 2026 值得一看的技術突破** (relevance: 100%) <https://www.inside.com.tw/article/40860-nvidia-gtc-2026-tsmc-samsung> 黃仁勳指出,

「我們與台積電共同發明了這項名為『Coupe』的製程技術。目前NVIDIA 是全球唯一將此項CPO 技術投入量產的公司，這是完全革命性的突破。

- **EP.135 揭秘NVIDIA 突破性科技和未來展望 - Apple Podcasts** (relevance: 100%) <https://podcasts.apple.com/nz/podcast/ep-135-%E6%8F%AD%E7%A7%98-nvidia-%E7%AA%81%E7%A0%B4%E6%80%A7%E7%A7%91%E6%8A%80%E5%92%8C%E6%9C%AA%E4-nvidia-ai-technology-center/id1516229812?i=1000669655455> 而CK 老師身為NVIDIA AI 技術中心台灣區技術負責人，站在浪頭之上，這次會和我們用深入淺出的方式帶出他對產業的觀察，如影像辨識、醫療影像、自駕車等領域，
- **英伟达的五步封神之路，AI芯片之王为何是ta? - 智源社区** (relevance: 100%) <https://hub.baai.ac.cn/view/34481> 2020发布的Ampere架构让英伟达在GPU技术上的又一重大突破，不仅对Tensor Core进行了进一步的升级，增加了对稀疏矩阵计算的支持，在性能和效率上更是提升到了
- **英伟达展示DLSS 5 超分技术，黄仁勋称是图形领域的时刻 - Chiphell\*\*** (relevance: 100%) <https://www.chiphell.com/thread-2788057-1-1.html> DLSS 5 是自2018 年实时光线追踪技术问世以来，计算机图形领域最具颠覆性的突破。DLSS 5 的核心在于引入了全新的实时神经渲染模型，能够彻底打破传统
- **英伟达展示DLSS 5 超分技术，黄仁勋称是图形领域的GPT 时刻|GPU ...** (relevance: 100%) [https://tech.sina.cn/2026-03-17/detail-inhrfmyh9695485.d.html?vt=4&cid=79649&node\\_id=79649](https://tech.sina.cn/2026-03-17/detail-inhrfmyh9695485.d.html?vt=4&cid=79649&node_id=79649) 黄仁勋表示，DLSS 5 是自2018 年实时光线追踪技术问世以来，计算机图形领域最具颠覆性的突破。DLSS 5 的核心在于引入了全新的实时神经渲染模型，能够彻底打破

### 三、技术趋势分析

### 3.1 模型能力演进

基于检索结果分析NVIDIA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 GPT, Gemini, or others. My purpose is to assist users with information and tasks.

---

### Sources

---

- **Nvidia挑戰ChatGPT與Gemini 推出自家NVLM 1.0大語言模型** (relevance: 100%) <https://netmag.tw/2024/10/12/nvidia-launches-nvlm-to-challenge-gpt> # Nvidia 挑戰 ChatGPT 與 Gemini 推出自家 NVLM 1.0 大語言模型. Nvidia 上周以開原碼專案釋出 NVLM 1.0 大型語言模型 (LLM) 家族，挑戰 OpenAI GPT 與 Google 。. Nvidia 上周釋出模型權重資料，承諾會再釋出訓練程式碼，讓第三方研究人員及開發商用於 AI 專案。NVLM 1.0 家族最大的是 720 億參數的 NVLM-D-72B，具多模態能力，號稱在複雜視覺與文字處理都有絕佳效能，比起封閉模型（如 GPT-4o）也毫不遜色。 . NVIDIA 的新 AI 模型分析了一個將學術摘要與完整論文進行比較的迷因，展示了其解...
- **GPT-4.1、Claude 3.7、Gemini 2.5 编码大对决：谁是真“码农之光 ...** (relevance: 100%) <https://blog.csdn.net/fq1986614/article/details/149127374> DeepSeek-v2.5是一个最先进的开源大型语言模型（LLM），在性能测试中超越了GPT-4 Turbo、Claude 3和Google Gemini等领先模型。该模型将DeepSeek版本2
- **2025：大语言模型（LLM）之年 - 36氪** (relevance: 100%) <https://m.36kr.com/p/3640423298125193> OpenAI 在 2024 年 9 月用 o1 和 o1-mini 开启了“推理”革命，也叫做推理侧扩展或可验证奖励强化学习（RLVR）。在 2025 年初，他们通过推出 o3、o3-mini 和 o4-mini 进一步强化了这一优势。自此，“推理”已成为几乎每家主流 AI 实验室模型的招牌功能。 . 一个显著的成果是 AI 辅助搜索现在真的变好用了。以前将搜索引擎连接到 LLM 的效果差强人意，但现在我发现，即使是复杂的调研问题，ChatGPT 的 GPT-5 Thinking 通常也能给出答案。 . Claude Code 是我所谓的“编程智能体”最杰出的代表——这种 LLM 系统可以编写代码...
- **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](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

- **2026 最新五大主流AI 語言模型(LLM) 全解析，付費 - 鏈新聞** (relevance: 100%) <https://abmedia.io/latest-top-5-llm-pricing-usage-safety> 第一版 Claude 於 2023 年推出，由 OpenAI 前核心成員 Dario Amodei 與 Daniela Amodei 等人於 2021 年創立的 AI 新創 Anthropic 所打造，主打「安全可控」的通用 AI，最新版本為 Claude 4.5 Sonnet。Gemini 的命名靈感來自 DeepMind 與 Google Brain 的合併，以及向 NASA 的雙子座計畫致敬。Google 共同創辦人 Sergey Brin 已重返公司，親自參與 Gemini 核心開發。這類產品一般會被涵蓋在整體 Google Cloud 或 Google AI 業務，因此沒有單獨對...

## 4.推理模型

检索关键词: o1,R1,推理,思维链

## Answer

---

DeepSeek-R1 is a competitive open-source reasoning model to OpenAI's o1, offering similar performance at a lower cost. It uses advanced reinforcement learning techniques and has an open API for various tasks. DeepSeek-R1's open licensing encourages further development and integration.

---

## Sources

---

- **新开源推理模型在高维潜空间思考，抛弃思维链 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/23013991514> 开源推理大模型新架构来了，采用与Deepseek-R1/OpenAI o1截然不同的路线：. 抛弃长思维链和人类的语言，直接在连续的高维潜空间用隐藏状态推理，可自适应
- **多模态也做到了强推理！工业界首个开源的R1V，让视觉思考进入o1 ...** (relevance: 100%) <https://www.51cto.com/article/810987.html> 作为一个VLM 推理模型，R1V 采用高效的多模态迁移方法，最大程度保留了文本推理能力，同时优化视觉任务表现。同时，R1V 提出通过混合优化策略来加强视觉文本
- **OpenAI的推理大模型o1模型的强有力竞争者！DeepSeekAI发布 ...** (relevance: 100%) <https://www.datalearner.com/blog/1051732459286417> 该模型利用了类似的o1的思维链思索过程，推理能力大幅增强。DataLearnerAI将在本文中对该模型进行介绍，并进行几个简单的对比结果测试。
- **o1推理框架最新成果：斯坦福&伯克利提出元链式思维 - CSDN博客** (relevance: 100%) <https://blog.csdn.net/QbitAI/article/details/145272009> 在最新的一篇长达100页的论文

中，他们将o1模型背后的推理机制提炼成了一个通用的框架——元链式思维（Meta-CoT）。

- **DeepSeek开源推理模型R1，比肩OpenAI o1正式版。** - 智源社区 (relevance: 100%)

<https://hub.baai.ac.cn/view/42817> 1月20日晚，DeepSeek（深度求索）公司发布推理模型 DeepSeek-R1 正式版，同步开源模型权重，并允许用户利用模型输出、通过模型蒸馏等方式训练其他模型。DeepSeek-R1 不仅开源了大量模型，还泄露了所有训练秘密。他们可能是第一个显示 RL（强化学习）飞轮发挥主要作用、持续增长的 OSS 项目。##

**\*DeepSeek-R1：**登录 DeepSeek 官网或官方 App，打开「深度思考」模式，即可调用最新版 DeepSeek-R1 完成各类推理任务。| 图片来源：DeepSeek. **\*DeepSeek-R1** 也同步上线了 API，对用户开...

## 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 multimodal AI model for video generation, emphasizing realistic visual and audio synchronization. It competes with models like Sora and Helios, focusing on complex motion and scene generation. The model supports real-time video generation at high speeds.

## Sources

---

- **Seedance 2.0 正式发布 - ByteDance Seed** (relevance: 63%) <https://seed.bytedance.com/zh/blog/seedance-2-0-%E6%AD%A3%E5%BC%8F%E5%8F%91%E5%B8%83> # Seedance 2.0 正式发布. 目前，Seedance 2.0 已上线即梦AI、豆包等平台，欢迎体验和反馈。 . [https://seed.bytedance.com/seedance2\\_0](https://seed.bytedance.com/seedance2_0). 1) 即梦网页端-视频生成-选择 Seedance 2.0; . 2) 豆包 App 对话框-Seedance2.0-选择 2.0 模型; . 3) 火山方舟体验中心-选择 Doubao-Seedance-2.0。 . ### 拟真视听效果和导演级操控. ### 让音视频生成“所想即所见”. 能完成前代模型难以实现的多人竞技运动生成，音频效果更加自然沉浸，输入也不再局限于单一的文字或图片， ...
- **Seedance vs Sora vs Kling：AI 视频生成模型深度对比** (relevance: 61%) <https://developer.aliyun.com/article/1711714> Sora 的核心创新在于引入"世界模型"(World Model)概念。它不仅仅是在像素层面生成视频,而是通过大规模预训练学习物理世界的运行规律——重力、光影、材质

- **Seedance 2.0 vs Sora 2.0：5个关键架构维度硬核对比 - CSDN博客** (relevance: 61%) <https://blog.csdn.net/CodeIsle/article/details/158172166> 当多模态生成模型从“能生成”迈向“可编排、可验证、可部署”的工业级阶段，Seedance 2.0 与Sora 2.0 的同步发布，标志着生成式AI 架构范式的根本性迁移。
- **14B规模竟也能单卡实时生成视频？多亏这个强大的开源底座** (relevance: 58%) [https://h5.ifeng.com/c/vivoArticle/v002xlanfR-\\_VnDxslPeNOCRnnoIZOKOfz-\\_ExwHwWwlNTtBQ\\_\\_?isNews=1&showComments=0](https://h5.ifeng.com/c/vivoArticle/v002xlanfR-_VnDxslPeNOCRnnoIZOKOfz-_ExwHwWwlNTtBQ__?isNews=1&showComments=0) 春节期间，Seedance 2.0 爆火，堪称现象级，这也再次把视频生成推上风口。前两天，字节跳动又携手北大、安努智能和Canva 共同开源了具备实时生成能力的视频模型 Helios 家族。该系列包含了 Helios-Base、Helios-Mid 与 Helios-Distilled 三个版本，全面覆盖了 T2V、I2V、V2V 以及交互式生成任务。其能以 14B 参数量之躯，实现高达 19.5 FPS 的单卡生成速度，可以说是真正做到了「质量」与「速度」齐飞。就在大家震惊于 Helios 高质量、高速度的生成能力时，技术社区却在底层架构里寻得了另一番玄机：这个模型的核心开发团队是北...
- **视频生成模型技术进展 - 知乎专栏** (relevance: 55%) <https://zhuanlan.zhihu.com/p/1917688698310493175> 端到端多模态模型，具备“推理链”机制，可完成图像编辑、未来帧预测等复杂任务。混合专家多模态模型，支持视觉理解，文本到图像生成，图像编辑，并且思考模式

## 4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

## Answer

---

The H100 GPU is designed for AI training and inference, while the B200 offers higher performance and efficiency. Blackwell architecture in B200 provides up to 5x the performance of H100. B200 and GB200 require liquid cooling for maximum performance.

---

## Sources

---

- **一张图说清：H100、H200、B200 到底该怎么选？ - 博客园** (relevance: 75%) <https://www.cnblogs.com/AlayaNeW/articles/19388803> | NVLink | 第四代 (900 GB/s) | 第四代 | 第五代 (1.8 TB/s) |. H200 不是算力升级，而是显存与带宽升级，解决“跑不动”的问题；. B200 则是一次架构级跃迁，面向千卡集群、下一代 AI 工厂设计。.| **<7B 参数，微调/推理 | A10 / L4 / RTX 6000 Ada** | 小模型对算力要求低，A10/L4 成本更低；H100 属性能过剩，仅在统一集群时考虑 |. | **7B-30B，全参训练** | H100 | 在 FP8 + 梯度检查点 + ZeRO 下可高效训练PyTorch/TensorFlow 生态最成熟，调试工...

- **万字长文解析：从H100 到B200，GPGPU 与大模型扩展性深度分析** (relevance: 70%)  
<https://zhuanlan.zhihu.com/p/1985478405458788975> 随着大模型参数量的指数级增长，NVIDIA H100/B200 等高性能GPU 已成为算力基础设施的核心。然而，在大规模训练中，单纯堆砌GPU 数量并不足以线性提升性能。
- **NVIDIA H100、B200、GB200 晶片的差異與製程資訊整理 - 方格子** (relevance: 67%)  
<https://vocus.cc/article/673b2431fd89780001cf4a9c>
  - 支援 HBM3 記憶體，容量最大達 80GB。
  - 4. \* 由 B200 GPU 和 Grace CPU 集成而成。
  - \* 採用 NVLink-C2C 技術，實現 CPU 和 GPU 高速互聯。
  - \* CPU 與 GPU 的深度整合適合複雜的 HPC 和 AI 工作負載。
  - 1. \***H100**：單晶片設計，適合 AI 和 HPC 工作。
  - 3. **GB200**\*：結合 CPU 和 GPU，專注於超算和複雜的 AI 應用。
  - NVIDIA 的 GB300 GPU 作為 H100 的繼任者，採用創新的 chiplet 架構，帶來效能與設計上的重大變革。本文深入探討其...
- **一文读懂英伟达的“新GPU”：比H100快5倍？1200W的功耗？液冷 ...** (relevance: 62%)  
<https://wallstreetcn.com/articles/3710774> 一文读懂英伟达的“新GPU”：比H100快5倍？1200W的功耗？液冷？与MI300X比较又如何？ - 华尔街见闻. 一文读懂英伟达的“新GPU”：比H100快5倍？1200W的功耗？液冷？与MI300X比较又如何？. 硬AI 2024-03-19 08:37. 据媒体周一分析，作为Hopper架构的后继者，Blackwell在性能提升方面表现出色，最高规格的Blackwell芯片浮点运算速度（FLOPS）大约快了5倍，能耗也进一步优化，与AMD MI300X GPU相比显现出强大竞争力，巩固了英伟达在性能和能效方面的技术优势。性能提升的关键在于Blackwell价格设计，每款GP...
- **英伟达显卡大比拼B100、H200、L40S、A100、H100架构性能对比** (relevance: 60%)  
<https://zhuanlan.zhihu.com/p/690084558> 这台超级计算机配备了数万个NVIDIA A100 GPU，并利用60多个数据中心的数十万个GPU辅助，为ChatGPT提供了强大的算力支持。这种规模的算力部署不仅体现

## 4.数据存储

检索关键词: HBM,显存,存储,NVLink

## Answer

---

NVLink is a high-speed interconnect for connecting multiple GPUs, offering high bandwidth and low latency. High Bandwidth Memory (HBM) provides large, fast memory for GPUs. NVLink enables GPUs to share memory and communicate efficiently.

---



## Sources

- **不同GPU的HBM透過NVLINK相連之後，所有的VRAM就都共享了 ...** (relevance: 100%)  
[https://www.threads.com/@llamatechtrend\\_zh/post/DHuRlqihgOT/  
%E9%81%8E%E5%8E%BB%E6%88%91%E4%B8%80%E7%9B%B4%E6%B2%92%E6%90%9E%E6](https://www.threads.com/@llamatechtrend_zh/post/DHuRlqihgOT/%E9%81%8E%E5%8E%BB%E6%88%91%E4%B8%80%E7%9B%B4%E6%B2%92%E6%90%9E%E6)  
# Thread. 過去我一直沒搞懂NVLINK跟HBM的關係是什麼，兩個buzzword一直在媒體出現，最近才搞懂。 NVLINK可以整合多顆GPU像NVL72就是整合了72個GPU在同一個機架上  
面， HBM則是影響每個GPU內部記憶體跟運算單元傳輸跟讀寫的速度。最有趣的是，因為  
NVLINK跟HBM的速度提上來了，不同GPU的HBM透過NVLINK相連之後，所有的VRAM就  
都共享了，以NVL72搭配GB300的架構來看，共享的VRAM理論值可以超過20TB。（這還沒  
談NVIDIA跨機器互聯的矽光子技術。） 比起8個H200的設計，只有1000多GB的共享  
VRAM，10TB以上的共享記憶體很逆...
- **大模型训练—Nvidia GPU 互联技术全景图 - 腾讯云** (relevance: 100%) [https://  
cloud.tencent.com/developer/article/2616528](https://cloud.tencent.com/developer/article/2616528) ## 大模型训练—Nvidia GPU 互联技术全  
景图.# 大模型训练—Nvidia GPU 互联技术全景图.**第一次拷贝：** 存储系统(NVMe) →系统  
内存(Host Memory). 技术实现：使用 DMA 技术，通过PCI-e总线，由存储控制器直接将数  
据从NVMe 拷贝到系统内存，无需CPU干预。. 技术实现：使用 CUDA的cudaMemcpy拷贝  
函数，通过PCIe总线将系统内存中的数据，拷贝到GPU显存中。.##### **1.2，优化版，  
GPUDirect Storage.** Storage是GPUDirect系列技术之一，GPUDirect经过多年的发展，  
如...
- **GPU内存概念浅析 - 博客园** (relevance: 99%) [https://www.cnblogs.com/  
ArsenalfanInECNU/p/18021724](https://www.cnblogs.com/ArsenalfanInECNU/p/18021724) 高带宽存储HBM(High Bandwidth Memory)是常用的片  
下GPU存储硬件。它将很多个DDR芯片堆叠在一起后和GPU封装在一起，实现大容量，高  
位宽的DDR组合阵列。
- **高速带宽除了PCle和HBM，还有哪些原创 - CSDN博客** (relevance: 97%) [https://  
blog.csdn.net/xiong\\_xin/article/details/146341006](https://blog.csdn.net/xiong_xin/article/details/146341006) (1)设备内部数据传输；HBM带宽最  
高，适用于大规模AI计算、HPC。；GDDR适用于高性能GPU，但带宽低于HBM。；LPDDR  
适用于移动设备，功耗优化，但带宽不够
- **NVIDIA NVLink與NVLink交换器** (relevance: 85%) [https://www.nvidia.com/zh-tw/  
data-center/nvlink/](https://www.nvidia.com/zh-tw/data-center/nvlink/) 整個伺服器機架必須有順暢的高傳輸量 GPU對GPU通訊，最新的AI  
模型才能達到最高效能。第六代 NVIDIA NVLink™與NVLink交換器採用低延遲的超大網路  
頻寬，具備全互連能力，是專為加速訓練與推論而設計，推理與代理 AI工作負載速度更  
快。 . 第六代 NVLink為NVIDIA Rubin平台上的每顆GPU提供高達每秒 3.6 TB 的頻寬，效  
能較前一代提升 2 倍，並比 PCIe Gen6 高出 14 倍。NVIDIA Vera Rubin NVL72這類機架級  
架構，可在全互連拓撲下連接 72 顆 GPU，總頻寬為 260 TB/s，為訓練與推論頂尖混合...

## 4.数据加速

检索关键词: FlashAttention,量化,推理优化

## Answer

---

FlashAttention optimizes attention computation for GPUs, achieving up to 9x speedup over standard attention, and significantly reduces memory usage. It uses quantization and advanced techniques for faster inference.

---

## Sources

---

- **新一代的FlashAttention - NVIDIA 技术博客** (relevance: 100%) <https://developer.nvidia.cn/blog/next-generation-of-flashattention/> # 新一代的FlashAttention. 作者: Vijay Thakkar 和 Fred Oh. NVIDIA 很高兴能与 Colfax、Together.ai、Meta 和普林斯顿大学合作, 利用 Hopper GPU 架构和 Tensor Core, 加速关键的融合注意力内核, 使用 CUTLASS 3. . FlashAttention-3 采用关键技术, 相比使用 FP16 的 FlashAttention-2, 性能提升 1.5–2.0 倍, 最高可达 740 TFLOPS。另外, 在使用 FP8 时, FlashAttention-3 可达到高达 1.2 PFLOPS, 且误差比基准 FP8 注意...
- **FlashAttention与FasterTransformer整合: NVIDIA推理加速终极指南** (relevance: 100%) [https://blog.csdn.net/gitblog\\_00990/article/details/151455016](https://blog.csdn.net/gitblog_00990/article/details/151455016) 推理速度提升: 相比传统注意力机制, 推理速度提升最高可达5.3倍 • 内存效率优化: 内存占用减少10-20倍, 支持更长的序列长度 • 端到端加速: 从输入到输出的完整
- **FlashAttention FasterTransformer整合: NVIDIA推理加速 - CSDN博客** (relevance: 100%) [https://blog.csdn.net/gitblog\\_00224/article/details/151162701](https://blog.csdn.net/gitblog_00224/article/details/151162701) # FlashAttention FasterTransformer整合: NVIDIA推理加速. ## FlashAttention FasterTransformer整合: NVIDIA推理加速. 【免费下载链接】flash-attention Fast and memory-efficient exact attention 【免费下载链接】flash-attention 项目地址: [https://gitcode.com/GitHub\\_Trending/fl/flash-attention](https://gitcode.com/GitHub_Trending/fl/flash-attention). 传统注意力机制的计算复杂度为 $O(N^2)$ , 内存占用同样为 $O(N^2)$ , 其中N是序列长度。这导致在处...
- **针对GPU优化的Attention计算介绍优化算法FlashAttention-3 - YouTube** (relevance: 100%) <https://www.youtube.com/watch?v=i1Cf5r4jGLA> Blog: <https://tridao.me/blog/2024/flash3/>
- **比标准Attention提速5-9倍, 大模型都在用的FlashAttention v2来了** (relevance: 100%) <https://juejin.cn/post/7261921491448692793> FlashAttention-2 将加速现有模型的训练、微调和推理。这意味着我们可以用相同成本训练2 倍上下文长度的语言模型。这将有助于语言模型理解长篇书籍和报告

## 4.Agent

检索关键词: 智能体,Agent,AutoGPT

## Answer

---

AutoGPT is an open-source project demonstrating autonomous agent capabilities. It is part of a broader ecosystem of AI agent frameworks like LangChain and MetaGPT. These agents execute complex tasks by planning and executing operations.

---

## Sources

---

- **构建您的首个LLM 代理申请- NVIDIA 技术博客** (relevance: 67%) <https://developer.nvidia.cn/blog/building-your-first-llm-agent-application/> AutoGPT: 这个 GitHub 项目是首批真正的智能体之一，它旨在展示智能体能够提供的各种功能。 • Voyager: 这个项目由NVIDIA 研究 所提出，探索了自我提升智能体的
- **万字读透：智能体（Agent代理） - 知乎专栏** (relevance: 61%) <https://zhuanlan.zhihu.com/p/17702168145> Agent首先会制定一个包含多个操作的计划任务，然后按照顺序去执行这些操作。这种方案对于复杂任务的执行而言是非常有用的，AutoGPT、BabyAGI、GPTEngineer等都是这样的
- **常见LLM Agent框架：AutoGPT - 飞书文档** (relevance: 57%) <https://docs.feishu.cn/v/wiki/Vo4kwaphgi7wlfktRzzcYeahnb/a8> (3) 智能体（Agent）模式。人类设定目标和提供必要的资源（例如计算能力），然后AI独立地承担大部分工作，最后人类监督进程以及评估最终结果。这种模式下，AI充分体现了智能体
- **十大AI Agent开发平台深度解析：从AutoGPT到LangChain - 鲸林向海** (relevance: 50%) <https://www.itsolotime.com/archives/16224> AutoGPT 是 AI Agent 领域的开创性项目，在 GitHub 上已获得超过 18 万星标。 . AutoGPT 作为开源项目，极大地推动了 AI Agent 领域的发展，是研究自主智能体（Autonomous Agents）的必读项目。 . \* **开源地址:** <https://github.com/Significant-Gravitas/AutoGPT>. Dify 是一个在 GitHub 上获得超过 12 万星标的大模型应用开发平台。它不仅仅是一个 Agent 框架，更融合了后端即服务（BaaS）和 LLMOps 的理念。 . 平台支持通过拖拽节点来可视化编排复杂的 Agent 逻...
- **AI-Compass Agent智能体技术生态：整合AutoGPT、LangGraph** (relevance: 45%) <https://segmentfault.com/a/1190000046915541> ## 1.modelscope-agent. ## 1.Agently. LangManus 是一个社区驱动的 AI 自动化框架，基于开源社区构建，旨在将语言模型与网页搜索、爬虫和 Python 代码执行等专业工具结合，实现复杂任务自动化。 . ## 1.Refact-AI-Agent. Refact.ai 是一款开源的 AI 软件工程智能体 (AI Agent)，旨在作为 GitHub Copilot 的替代方案。它能够端到端地处理工程任务，深入理解代码库，并与开发者的工具、数据库

和浏览器集成，以自动化复杂的多步骤任务，从而提升开发效率和代码质量。Refact.ai 的核心基于先进的 AI...

---

## 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，NVIDIA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NVIDIA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 08:27:11 AM CST 2026