

腾讯 AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:25:16

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: 腾讯

主要产品: 混元, Hunyuan

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

Tencent released its new AI model, Hunyuan-Large, which is the largest open-source Transformer model globally. It supports long context processing and has achieved superior performance in natural language processing and general AI tasks.

Sources

- **腾讯自研AI大模型混元2.0发布：总参数406B - IT之家** (relevance: 87%) <https://www.ithome.com/0/902/856.htm> 业界 手机 电脑 测评 视频 AI 苹果 iPhone 鸿蒙 软件. 智车 数码 学院 游戏 直播 5G 微软 Win10 Win11 专题. # 腾讯自研 AI 大模型混元 2.0 发布：总参数 406B，复杂推理场景综合表现“稳居国内第一梯队”。2025/12/5 22:17:10 来源：IT之家 作者：汪淼 责编：汪淼. IT之家 12 月 5 日消息，腾讯自研 AI 大模型**混元 2.0 (Tencent HY 2.0)** 今日正式发布，包括 Tencent HY 2.0 Think 和 Tencent HY 2.0 Instruct。腾讯表示，HY 2.0 采用...

- **AI进化速递 | 腾讯混元将发布新一代生图模型 - 第一财经** (relevance: 76%) <https://www.yicai.com/news/102847061.html> 腾讯混元明日将发布新一代生图模型；苹果据悉开发类ChatGPT应用；Meta CTO称人形机器人是下一个“AR级赌注”。
- **腾讯云AI模型涨价_新浪新闻** (relevance: 75%) <https://www.sina.cn/news/detail/5275393059979944.html> 【#腾讯云AI模型涨价#】#腾讯云公告涨价# 3月11日，腾讯云发布公告称，为了持续提供稳定优质的大模型服务，腾讯云智能体开发平台将对部分模型的计费策略进行优化调整。本次调整主要涉及两类变更：模型价格调整与公测模型结束免费。根据公告内容，此次调整将从2026年3月13日00:00（北京时间）正式生效。其中，GLM 5、MiniMax 2.5、Kimi 2.5三个模型将结束限时免费公测，转为正式商用服务，根据模型调用按量计费。另一变化则是对混元系列模型Tencent HY2.0 Instruct与Tencent HY2.0 Think服务进行涨价。根据腾讯云披露的调整详情：Tencent...
- **腾讯发布全新混元大模型Hunyuan-Large：全球最大开源Transformer ...** (relevance: 74%) <https://www.51cto.com/aigc/3201.html> 这款模型不仅支持长达256K个token的超大上下文处理，还在技术层面上实现了众多创新，能够在自然语言处理及通用AI任务上取得优异的表现，甚至在某些方面超越了业界领先的模型，
- **助力“好用的AI”落地！腾讯宣布：AI能力全面开放 - 南方+** (relevance: 71%) <https://www.nfnews.com/content/O3GAj1GV00.html> # 助力“好用的AI”落地！腾讯宣布：AI能力全面开放. 9月16日，2025腾讯全球数字生态大会在深圳举行，会上公布多项AI技术和产品最新进展，并宣布全面开放腾讯AI落地能力及优势场景，助力“好用的AI”在千行百业中加速落地。不久前，腾讯发布的2025年第二季度财报显示，AI正成为腾讯新的收入增长引擎和业务基因，此次会上宣布通过腾讯云也将自身累积的技术沉淀与AI实践全面开放，打造“智能化引擎”，通过智能体解决方案、“SaaS+AI”、大模型技术三大升级，打造“好用的AI”，激发企业创新潜能。会上，腾讯首发“腾讯云智能体战略全景图”，及腾讯云智能体开发平台ADP3.0、Agent inf...

2.2 技术突破

Answer

Tencent has made significant breakthroughs in AI, including advancements in large language models and multi-modal generation technology. They have also improved their cloud infrastructure for faster processing and more open services. These innovations highlight Tencent's commitment to technological progress.

Sources

- **2024年十大科技和应用趋势- Tencent 腾讯** (relevance: 100%) <https://www.tencent.com/zh-cn/articles/2201789.html> # Tencent腾讯. # 2024年十大科技和应

用趋势. 新的一年来临, 腾讯研究院邀请科学家、工程师、学者和其他专家对2024年数字科技未来发展趋势和应用前景进行了预测。我们认为, 通用人工智能渐行渐近, AI将跨行业、跨场景地驱动突破性创新, 从智慧电网到电动垂直起降飞机, 再到星地直连通信和辅助机器人。未来发电和用电模式将发生变化。一直以来, 家庭都是用电单位, 而现在, 家庭逐渐能够产生并储存电能。智能电网可以调配电动汽车的充电模式, 例如用电低谷充电省钱, 然后在用电高峰将多余的光伏电力输送回电网, 促进整个社会实现更加可持续的用电方式。飞行汽车以前只在科幻电影里才能看到。如今, 它们已经走进现...

- **巨额斥资研发, 腾讯如何砸穿AI天花板- 维科号 - OFweek** (relevance: 100%) <https://mp.ofweek.com/ai/a056714287587> 当行业还在争论AI技术哪家强时, 腾讯早已跳出单一维度, 用一场自我强化的飞轮游戏重新定义了规则—这不是简单的技术迭代, 而是一场从投入、突破到回报的闭环
- **2023年十大数字科技趋势 - Tencent 腾讯** (relevance: 100%) <https://www.tencent.com/zh-cn/articles/2201521.html> # Tencent腾讯. # 2023年十大数字科技趋势. 在今天的《2023十大数字科技前沿应用趋势》报告中, 我们的课题组描绘了四大主要技术领域, 包括: 几十年来, 算力呈指数级增长, 现在的微型芯片能够以超乎想象的速度处理海量信息。不止如此, 随着超级计算机的出现, 处理器的性能得到进一步提升。这些高性能计算结合了经典计算系统和量子计算系统, 能够以极快的速度计算和解决复杂问题。高性能计算可以带来有利于社会的科技突破, 包括开发大型语言模型、AI生成内容、自动驾驶和蛋白质结构预测等人工智能应用, 这些应用依赖于复杂的深度学习系统, 需要强大的处理能力。计算机需要操作系统才能运行。如果没有这一层关...
- **腾讯AI, 加速狂飙的这半年 - 雷锋网** (relevance: 100%) <https://m.leiphone.com/category/industrycloud/WlgZrqnpY1otb8G0.html> 除了在大语言模型领域加速追赶, 腾讯在本次大会上亮出的多模态生成技术也十分惊艳: 混元图像2.0 实现「毫秒级」生图突破, GenEval 基准测试准确率超95%, 不仅
- **腾讯邱跃鹏: 推理需求爆发, 云基础设施也要同步升级 - 华尔街见闻** (relevance: 100%) <https://wallstreetcn.com/articles/3755682> 据邱跃鹏介绍, 腾讯云已在推理加速、Agent Infra和国际化布局等方面取得突破, 并将以更加开放的姿态, 助力企业把握时代机遇。在推理加速方面, 腾讯云深入参与

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析腾讯在以下方面的进展:

- **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. LLMs like GPT, Claude, and Gemini are advanced AI models with diverse capabilities. Choose based on your specific needs for tasks like coding, long text understanding, or integration with Google services.

Sources

- **(LLM系列)什么是大语言模型？ - 腾讯云** (relevance: 100%) <https://cloud.tencent.com/developer/article/2625657> ## (LLM系列)什么是大语言模型？ . # (LLM系列)什么是大语言模型？ . ## (LLM系列)什么是大语言模型？ . 人工智能正在改变我们与技术互动的方式。大语言模型（Large Language Model，简称 LLM）作为 AI 领域最具突破性的技术之一，已经从研究实验室走向了日常应用。无论是 ChatGPT、Claude 还是 Gemini，这些工具都基于同一核心技术——大语言模型。本文将深入探讨 LLM 的工作原理，并帮助您了解如何选择最适合您需求的模型。 . ### 一、什么是大语言模型？ . 大语言模型是一种基于深度学习的人工智能系统，经过海量文本数据的训练，能够理解和生成人类...
- **最强国产多模态刚刚易主！腾讯混元把GPT-4/Claude-3.5/Gemini-1.5 ...** (relevance: 100%) <https://www.cnblogs.com/buluai/articles/18356518> 根据最新的AI行业资讯，腾讯的混元大模型在多模态能力上取得了显著的进步，甚至在某些方面超越了国际上知名的模型如GPT-4、Claude-3.5和Gemini-1.5。
- **最强国产多模态刚刚易主！腾讯混元把GPT-4/Claude-3.5/Gemini-1.5 ...** (relevance: 100%) https://blog.csdn.net/weixin_40700136/article/details/141178821 根据最新的AI行业资讯，腾讯的混元大模型在多模态能力上取得了显著的进步，甚至在某些方面超越了国际上知名的模型如GPT-4、Claude-3.5和Gemini-1.5。

- **Claude、Gemini 到国产大模型：2026 年 LLM API 聚合服务商深度 ...** (relevance: 100%) <https://juejin.cn/post/7594626381432832050> # GPT-5、Claude、Gemini 到国产大模型：2026 年 LLM API 聚合服务商深度测评与结论. ## 背景：问题已不再是“有没有模型”。到 2026 年，GPT-5、Claude、Gemini 与国产大模型已形成长期并存格局。模型能力不再稀缺，真正的挑战转向：**如何以低成本、低复杂度、可持续地使用多模型能力。** . ## 核心维度一：**模型覆盖不等于真实能力**. 几乎所有平台都能列出“支持 GPT-5 / Claude / Gemini / 国产模型”，但差异在于：. ## 核心维度二：**稳定性来自架构设计**. ## *核心维度三：低价背后的不同路径...
- **原来，这些顶级大模型都是蒸馏的-腾讯新闻** (relevance: 100%) <https://news.qq.com/rain/a/20250129A031IV00> 「除了 Claude、豆包和 Gemini 之外，知名的闭源和开源 LLM 通常表现出很高的蒸馏度。」这是中国科学院深圳先进技术研究院、北大、零一万物等机构的研究者在一篇新论文中得出的结论。蒸馏固然是一种提升模型能力的有效方法，但作者也指出，过度蒸馏会导致模型同质化，减少模型之间的多样性，并损害它们稳健处理复杂或新颖任务的能力。所以他们希望通过自己提出的方法系统地量化蒸馏过程及其影响，从而提供一个系统性方法来提高 LLM 数据蒸馏的透明度。论文链接：<https://github.com/Aegis1863/LLMs-Distillation-Quantification/blob/ma...>

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

```
{ "title": "Answering the Query", "content": "DeepSeek-R1 is an advanced reasoning model developed by DeepSeek, known for its strong performance in complex reasoning tasks.", "next_action": "final_answer" }
```

Sources

- **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: 99%) <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...

- **o1也会「想太多」？腾讯AI Lab与上海交大揭秘o1模型过度思考问题** (relevance: 98%) <https://zhuanlan.zhihu.com/p/17124737367> o1 模型通过模拟人类的深度思考过程，在思维链中运用如自我反思、纠错以及探索多种解法等推理策略，展现了强大的长时间推理（Inference-Time Scaling）性能。
- **【DeepSeek-R1背后的技术】系列六：思维链（CoT）** (relevance: 98%) <https://deepseek.csdn.net/67ab1f1979aaf67875cb9ce7.html> # logo DeepSeek技术社区. ### DeepSeek技术社区. 第12篇：分词算法Tokenizer（WordPiece，Byte-Pair Encoding（BPE），Byte-level BPE(BBPE)）. 第13篇：归一化方式介绍（BatchNorm, LayerNorm, Instance Norm 和 GroupNorm）. 论文：Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 如果我们考虑提示链（prompt chaining），它是一种更基本的 CoT 提示形式，其中...
- **2025年国内AI推理大模型效果对比：DeepSeek、通义千问、腾讯混元** (relevance: 97%) https://learnku.com/articles/90116?order_by=vote_count 社区 动态 AI作品 话题列表 社区 Wiki 优质外文 招聘求职 AI Coding 实战教程 社区文档. 幂简集成的个人博客 / 24 / 0 / 创建于 9个月前. 我们将选用同一个提示词分别对DeepSeek R1、腾讯混元hunyuan t1 latest、通义千问QWQ 32B混匀模型进行试用对比。 . ### **1、DeepSeek R1.** DeepSeek R1 是由幻方量化旗下 DeepSeek 研发的推理模型，于 2025 年 1 月 20 日发布并开源。它基于 6710 亿参数的混合专家架构，采用多阶段训练方法，包括冷启动微调...
- **g1：o1推理链开源实现，原理竟如此简单！解决60-80% 的困扰LLM ...** (relevance: 97%) <https://cloud.tencent.com/developer/article/2472870> 此外，我觉得o1不会是单纯 gpt-4o+Agent+思维链 能到达的水平，可能o1训练的数据集会存在一种tag，其作用是提供一个问题不同种类的、拆解的、细致的解决步骤，当然，这可能只是其中一种训练方式。 . 当然这一切都是一些猜测，回到正题，我们今天要介绍的是一个模仿o1思维链的开源项目——g1（在 Groq 上使用 Llama-3.1 70b 创建类似 o1 的推理链）。 . g1 是实验性的并且是开源的，以帮助激励开源社区开发新的策略来产生类似 o1 的推理。不过他主要是展示可视化步骤提示推理的力量，而不是与使用不同技术的 o1 进行比较或完全复制。 . ##### Groq 模型（如lla...

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 multi-modal AI video generation model by ByteDance. It supports text, images, videos, and audio inputs for high-quality video creation. It has faced legal challenges over copyright issues.

Sources

- **字节跳动发布Seedance 2.0多模态AI视频生成模型 - 腾讯新闻** (relevance: 76%) <https://news.qq.com/rain/a/20260214A037YS00> # 字节跳动发布Seedance 2.0多模态AI视频生成模型. 2026-02-14 12:08发布于北京. 科技巨头在最新AI模型竞赛中持续较量，字节跳动推出了下一代视频生成器。这家TikTok背后的中国公司在博客文章中表示，Seedance 2.0支持结合文本、图像、视频和音频的多模态提示。公司声称该模型"在生成质量上实现了实质性飞跃"，在生成包含多个主体的复杂场景以及遵循指令方面都有显著改进。用户可以通过向Seedance 2.0提供最多九张图像、三个视频片段和三个音频片段来优化文本提示。该模型能够生成长达15秒的带音频视频片段，同时考虑摄像机运动、视觉效果和动作细节。据字节跳...
- **中国AI的“DeepSeek时刻”再次来临：Seedance 2.0如何缩小多模态 ...** (relevance: 76%) <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：技术革新开启AI视频生成新纪元 - QQ.com - 腾讯** (relevance: 75%) <https://news.qq.com/rain/a/20260224A050DN00> # Seedance 2.0：技术革新开启AI视频生成新纪元. 2026-02-24 16:10发布于北京中国日报中文网官方账号. Seedance 2.0的发布成为AI视频生成领域从“单模态画面”向“多模态视听合一”质变的关键节点，其独创的“双分支扩散变换器”架构实现了全方位技术突破，不仅攻克了传统模型的诸多行业痛点，更与Sora、可灵形成差异化技术路线，重塑了行业竞争格局，推动AI视频生成从简单的素材制作向专业的工业化内容生产迈进，为行业发展带来全新变革。传统AI视频生成长期沿用“先绘画面、后配音频”的割裂制作模式，音频信号需依托已生成的画面特征进行后期匹配与拼凑，不仅极易产生明显的音...
- **[PDF] Seedance2.0：生成式视频的技术奇点与产业重构** (relevance: 72%) https://pdf.dfcfw.com/pdf/H3_AP202602211819975803_1.pdf?1771752045000.pdf 1 行业点评 (2026年2月12日) Seedance2.0：生成式视频的技术奇点与产业重构 2026年2月，字节跳动发布旗舰级AI视频生成模型Seedance 2.0。这一发布不仅是字节跳动在人工智能领域技术积累的一次集中爆发，更被视为全球生成式AI从单点工具迈向工业化深水区的标志性事件。Seedance 2.0的问世正值全球AI视频技术竞争的白热化阶段。与OpenAI的Sora 2、Google的Veo 3.1以及国内快手Kling 3.0等顶尖模型相比，Seedance 2.0凭借其独特的架构、卓越的多镜头叙事能力以及对原生音频的完美融合，...
- **Seedance 2.0：字节跳动多模态AI视频大模型功能解析** (relevance: 66%) <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

Tencent's GPU H100 and B200, along with TPU, are powerful AI chips. They are used for high-performance computing and AI model training. Their capabilities are continually being enhanced.

Sources

- **比H100快20倍还更便宜！英伟达的“掘墓人”出现了？ - 腾讯** (relevance: 100%) <https://news.qq.com/rain/a/20240627A0489000> Etched公司表示，Sohu速度比英伟达的最新一代Blackwell架构的B200 GPU还要快一个数量级，而且价格更便宜。可以帮助客户构建GPU无法实现的产品。 . “通过将Transformer架构集成到我们的芯片中，虽然无法运行大多数传统的AI模型，比如为Instagram广告提供支持的DLRM、像AlphaFold 2这样的蛋白质折叠模型、像Stable Diffusion 2这样的旧的图像模型，以及CNN、RNN或LSTM等模型，但是针对基于Transformer架构的大模型，Sohu将是有史以来最快的AI芯片，没有哪个芯片能够与之匹敌。” Etched公司说...
- **显卡基础知识 | 英伟达算力开挂的GPU！ - 腾讯云** (relevance: 100%) <https://cloud.tencent.com/developer/article/2616538> 本文深入解析AI模型训练与推理的显卡算力选择策略，涵盖浮点精度、英伟达架构、超级芯片设计及多GPU扩展方案。重点对比FP64到INT4不同精度算力表现，
- **10万卡H100集群有多难搭？一文解析算力集群技术要点 - 腾讯新闻** (relevance: 100%) <https://news.qq.com/rain/a/20240623A037GP00> H100的小型集群通常只使用多模收发器，通过一层或两层的交换机，以400G的速度将每个GPU连接在一起。如果是大型集群，则需要增加更多层的交换机，光纤设备也会
- **AI算力芯片天下一分为四，中国实力渐显** (relevance: 100%) <https://ee.ofweek.com/2023-11/ART-8440-2800-30618101.html>  5 |||| 3D3 | e |  7 | . ǎʀ | DZ | VR | ý |||  ² || . | ǎ N豸 | ee5| |  || 7 | A | PCB. * Öã ı v eZ 7 LED  P . H100 ЫúĴ9 GPUĭġ AI

HPC 4h A100 Tensor Core GPU 100 μs 100 I/O A100 L 6 h AI HPC 201...

- **中信建投：各厂商AI芯片向大功率不断迭代 - 腾讯** (relevance: 98%) <https://news.qq.com/rain/a/20260117A05SHX00> # 中信建投：各厂商AI芯片向大功率不断迭代，推升单机柜功率迅速增长. 2026-01-17 22:03发布于北京腾讯财经《研报精选》栏目官方账号. 以英伟达为首的AI芯片厂商不断迭代升级其最新AI芯片的功率。英伟达已从Maxwell时代单颗芯片250W提升至B200芯片的超过1000W；谷歌、微软、Meta等自研芯片也向单颗1000W以上迈进。通过NVL72（英伟达）、Superpod（谷歌）等多芯片设计，单机柜功率进一步向MW级迈进。我们认为，至2028年北美新增AI数据中心功率可达71GW。· 本报告尝试详尽地梳理AIDC电源技术发展趋势，以及给各环节设备带来的变化。投资机会包括四大类： ...

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

Tencent utilizes HBM, NVLink for high-speed GPU data transfer, and advanced storage solutions for AI applications. NVLink technology significantly boosts GPU-to-GPU data exchange speed. Tencent's storage solutions support large-scale AI model training.

Sources

- **显卡基础知识 | 英伟达算力开挂的GPU! - 腾讯云** (relevance: 100%) <https://cloud.tencent.com/developer/article/2616538> 高HBM大小和带宽，配合高FLOPS/OPS能够显著提升GPU处理数据的能力，可以更快速地处理大型模型数据，在训练深度学习模型时表现突出。 · NVLink带宽决定了多个
- **存储芯片本轮涨价能走多远？一文看懂产业链 - 腾讯** (relevance: 99%) <https://news.qq.com/rain/a/20260226A034KK00> # 存储芯片本轮涨价能走多远？一文看懂产业链. 2026-02-26 11:18发布于北京北京融中传媒科技有限公司官方账号. 存储芯片是芯片行业的第二大产业，仅次于CPU、GPU等逻辑芯片。得益于上游SK海力士、三星等存储晶圆原厂主动控制产出，存储芯片价格从2023年下半年开始反转，进入第五个上行周期。· 存储芯片是芯片行业的第二大产业，仅次于CPU、GPU等逻辑芯片。本轮存储芯片市场的热潮，源于全球范围内供需关系的深刻调整。需求端方面，人工智能基础设施建设的激增，导致对高端内存的需求前所未有，供给端方面，美光科技等国际大厂已预警短缺状态将持续。· 存储芯片一般3-4年为一个周期，得益于上...

- **冯明宪博士演讲：AI时代HBM产业发展探讨——华强北视角（一）** (relevance: 97%) <https://news.qq.com/rain/a/20240425A080JM00> # 冯明宪博士演讲：AI时代HBM产业发展探讨——华强北视角（一）。2024-04-25 18:12发布于广东。2024年4月12日，由华强电子网主办的“2024半导体产业发展趋势大会暨2023年度（第十六届）华强电子网优质供应商&电子元器件行业优秀国产品牌颁奖盛典”在深圳南山华侨城洲际大酒店圆满举办！冯明宪博士受邀参加芯榜承办的分论坛之“芯片与智能硬件融合创新论坛”，并做“AI时代HBM产业发展探讨”主题报告。在报告中，冯明宪博士提出：HBM是AI半导体性能进一步提升并突破算力瓶颈的关键，随着AI技术加速创新应用，HBM市场规模也将急剧增长，2026年有望达到230亿美元，进而带动存...
- **被卡脖子的真相：为什么造GPU算力卡这么难？（科普文章）** (relevance: 96%) <https://finance.sina.com.cn/roll/2026-03-16/doc-inhrcpie9841385.shtml> HBM显存：作为AI数据吞吐核心，2026年主流为HBM3E，制造精度要求极高，需12层芯片微尺度堆叠，依赖数百万个TSV硅通孔互连。目前全球仅三家企业能量产
- **转载：【AI系统】NVLink 原理剖析 - 腾讯云** (relevance: 82%) <https://cloud.tencent.com/developer/article/2475996> 英伟达的 NVLink 技术则为 GPU 之间提供了更高速度的数据交换能力，其传输速度远超传统的 PCIe 连接，使得数据在 GPU 之间的传输更加高效。此外，NVSwitch 技术进一步扩展了这种能力，允许多达数十个 GPU 之间实现高速、高带宽的直接连接。这种先进的互连技术极大地提高了大规模 GPU 集群处理复杂模型时的数据交换效率，降低了通信延迟，从而使得万亿级别的模型训练成为可能。在英伟达推出其创新的 NVLink 和 NVSwitch 互联技术之前，构建强大计算节点的常规方法是通过 PCIe 交换机将多个 GPU 直接连接到 CPU，如下图所示。这种配置方式依赖于 PCIe 标准...

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

FlashAttention optimizes attention mechanisms for faster inference, while SageAttention3 offers significant speedups and supports 8-bit training. FlashAttention-3 leverages hardware features for optimal performance, achieving up to 16x speedup over standard attention.

Sources

- **LLM推理优化技术：从理论到实践 - 腾讯云** (relevance: 64%) <https://cloud.tencent.com/developer/article/2611322> 2025年大语言模型(LLM)推理优化技术取得重大突破，涵盖模型压缩、硬件加速、算法优化和系统优化四大方向。最新技术如4位量化、FlashAttention、连续批

- **比国外竞品计算性能快5倍，清华团队提出微缩版FP4注意力机制** (relevance: 63%)
<https://news.qq.com/rain/a/20250529A08KXW00> 近日，清华大学团队打造了首个用于推理加速的微缩版 FP4 注意力机制——SageAttention3，在英伟达 RTX5090 上实现了 1038TOPS 的计算性能。**相比此前在英伟达 RTX5090 上计算性能最快的、由美国斯坦福大学提出的 FlashAttention，SageAttention3 的计算性能快了 5 倍。**实验表明，SageAttention3 能够加速各种模型，并且不会导致端到端质量指标的下降。由于注意力机制的时间复杂度是 n^2 ，因此注意力机制的效率非常重要。为此，他们通过两个关键贡献提高了注意力的效率：首先，研究团队利用英伟达 Blackwell GPU 中...
- **清华SageAttention3，FP4量化5倍加速！且首次支持8比特训练** (relevance: 62%)
<https://news.qq.com/rain/a/20250618A07JGZ00> 随着大型模型需要处理的序列长度不断增加，注意力运算（Attention）的时间开销逐渐成为主要开销。此前，清华大学陈键飞团队提出的即插即用的 SageAttention 和 SageAttention2 已经被业界及社区广泛的使用于各种开源及商业的大模型中，比如 Vidu，CogvideoX，Mochi，Wan，HunyuanVideo，Flux，Llama3，Qwen 等。近日，清华大学陈键飞团队进一步提出了针对 BlackWell 架构的首个全 FP4 量化的即插即用注意力算子（**SageAttention3**）。实现了**5 倍**相比于 FlashAttention 的**即插...
- **英伟达亲自下场FlashAttention三代升级，比标准注意力快16倍-36氟** (relevance: 55%)
<https://m.36kr.com/p/2858732550490752> # H100利用率飙升至75%，英伟达亲自下场 FlashAttention三代升级，比标准注意力快16倍。**训练速度提升1.5-2倍**，FP16下计算吞吐量高达740TFLOPs/s，达理论最大吞吐量**75%**，更充分利用计算资源，此前只能做到 35%。而且这一次，不再是一作Tri Dao单打独斗，FlashAttention-3直接和**英伟达**、Meta、谷歌等合作，针对最强芯片H100专门做优化。同时和前作一样，FlashAttention-3也将开源，PyTorch和Hugging Face中都集成。前Stable Diffusion老板Emad也非常关注这一进展...
- **FlashAttention - 3 精解：硬件感知Attention 优化-腾讯云开发者社区** (relevance: 51%)
<https://cloud.tencent.com/developer/article/2538735> ## FlashAttention - 3 精解：硬件感知 Attention 优化。
FlashAttention-3 作为这一领域的闪耀新星，带着硬件感知的优化理念横空出世。它不再仅仅局限于算法层面的优化，而是深入到硬件底层，与硬件特性紧密相连，挖掘出硬件隐藏的性能潜力，为注意力机制的高效执行开辟了一条全新的道路。FlashAttention-3 针对传统 Attention 的痛点，带来了以下关键革新：FlashAttention-3 的理论基础源于多篇重要的研究工作。例如，Huang 等人在《...

4.Agent

检索关键词: 智能体,Agent,AutoGPT

Answer

An AI system built by a team of inventors at Amazon provides concise, factual responses. Agents are AI systems that perform tasks autonomously. Auto-GPT is an open-source AI agent framework.

Sources

- **智能体(Agent)开发全攻略，从AutoGPT到“伐谋”，让AI不再“嘴炮”直接 ...** (relevance: 100%) https://blog.csdn.net/m0_56255097/article/details/156112175 智能体技术的演进历程清晰可见。2023年3月，AutoGPT框架的发布标志着智能体技术从理论走向实践，实现了大模型的外推能力；同年11月，OpenAI推出的Assistant
 - **实用至上：智能体/Agent 是什么-腾讯新闻** (relevance: 100%) <https://news.qq.com/rain/a/20240331A05EXY00> # 实用至上：智能体/Agent 是什么. 我算比较资深的 Agent 开发者：ChatGPT中，用量最大的 Plugin 和用量最大的华人捏的 Bot，可能都是我做的。. 之前写过一篇实操教程：《保姆级教程：Coze 打工你躺平》，今天想从 Agent 的发展脉络，来更深入谈谈。## Agent 的起源. ## 现在的 Agent. 时至今日，对于 Agent 是什么，可能还没有一个标准的定义。一个常见的观点是，Agent 是一种让 AI 以类似人的工作和思考方式，来完成一系列的任务。一个 Agent 可以是一个 Bot，也可以是多个 Bot 的协同。就像是职场里，简单的工作独立完成...
 - **【单Agent框架】01-AutoGPT：以ChatGPT为核心的自治AI智能体- 知乎** (relevance: 100%) <https://zhuanlan.zhihu.com/p/668234147> 同年4月，Auto GPT成为国内外的热门话题，那AutoGPT到底是什么呢？其实，AutoGPT是一个AI agent（智能体），也是开源的应用程序，结合了GPT-4和GPT-
 - **RAG到ai agent智能体从入门到实战大模型零基础入门 - YouTube** (relevance: 100%) <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.
 - **单智能体框架AutoGPT有哪些优缺点？ - 飞书文档** (relevance: 100%) <https://docs.feishu.cn/v/wiki/Uwh7wnNN0iWbwqknboocnmMlnbe/ah> AI Agent 阶段性总结与创投观察 • 1. 智能体：在上面单独定义的基础上，在多智能体系统中的智能体协同工作，每个智能体都具备独特有的LLM、观察、思考、行动和记忆； • 2. 环境：
-

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，腾讯的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

预测腾讯在未来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 08:25:40 AM CST 2026