

字节跳动 AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:24:32

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: 字节跳动

主要产品: 豆包, Seedance, Seed

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

ByteDance released Seedance 2.0, an AI model generating high-quality videos from text prompts. It features advanced scene transitions and character consistency. This advancement positions ByteDance as a leader in AI video generation technology.

Sources

- 字节跳动发布三款AI模型_新浪新闻 (relevance: 90%) <https://www.sina.cn/news/detail/5266526082172089.html> 2026年2月字节跳动集中发布Seedance 2.0 (AI视频)、Seedream 5.0 Lite (AI图片)、豆包大模型2.0三款模型, 性能实现突破性升级, 不仅让其与Google的技术差距大幅缩小至一两个月, 更凭借模型与应用、云服务的协同闭环, 跻身全球AI第一梯队, 也让业界重新认可DeepMind CEO Demis Hassabis此前对字节仅落后Google六个月的判断。1. Seedance 2.0: AI视频生成能力迎来质变, 指令遵循能力极强, 可精准理解长提示词并基本解决幻觉问题, 物理逻辑理解、复杂转场衔接效果出色, 能实现导演级的运镜、画面和声音生成, 让动漫短片制作等场景越过实...

- **字節跳動推出Seedance 2.0，「大片級」AI影片顛覆好萊塢？ - WSJ** (relevance: 75%)

<https://cn.wsj.com/articles/>

%E5%AD%97%E7%AF%80%E8%B7%B3%E5%8B%95%E6%8E%A8%E5%87%BAseedance-2-0-%E5%A4%A7%E7%89%87%E7%B4%9A-ai%E5%BD%B1%E7%89%87%E9%A1%9B%E8%A6%86%E5%A5%BD%E8%90%8A%E5%A1%A2-5gaa_at=eafs&gaa_n=AWetsqesXZxyGGF8uYzxTYIVuil6-X0uxd7BORWBpCgFm1HQXCgz_K_8obBj&gaa_ts=69b8a2bf&gaa_sig=h9Lgq5wi_ve8yAQVUsVKZRYNDho1o7y00QWu3xuLKG6EpS5i2x4CZnMJki-09SZGdVMff5mHV1lt4g%3D%3D TikTok 母公司開發出一款人工智慧(AI)模型，可以根據單一文本提示，生成具有故事情節、場景切換和鮮明人物形象的高品質影片。總部位於北京的字節

- **字节跳动Seed** (relevance: 75%) <https://seed.bytedance.com/zh/> 每秒推理速度 2146 Tokens，扩散语言模型 Seed Diffusion Preview 发布. 每秒推理速度 2146 Tokens，扩散语言模型 Seed Diffusion Preview 发布. 字节跳动 Seed 与比亚迪锂电池深化合作：将成立 AI 联合实验室加速电池研发. 字节跳动 Seed 与比亚迪锂电池深化合作：将成立 AI 联合实验室加速电池研发. 解锁任意模态模型训练，字节跳动 Seed 开源 VeOmni 框架. 解锁任意模态模型训练，字节跳动 Seed 开源 VeOmni 框架. 字节跳动 Seed Prover 取得 IMO 2025 银牌分数. 字节跳动 ...
- **字节又一款AI产品火了！Seedance2.0海内外刷屏哪些行业将被颠覆？** (relevance: 74%) <https://www.cls.cn/detail/2284177> ## 首页## 电报## 话题## 盯盘## VIP## FM## 投研## 下载. ### 头条. ### A股. ### 港股. ### 环球. ### 公司. ### 券商. ### 基金 • ETF. ### 地产. ### 金融. ### 汽车. ### 科创. ### 品见. 2026-02-09 09:28 星期一. ①Seedance2.0独有的多镜头叙事功能，能够根据单个提示自动生成多个相互关联的场景；②官方总结称，其生成2K视频的速度比Kling等竞争对手快30%。③东方证券指出，国内供给侧的技术天花板进一步提升，视频生成赛道进入类25年LLM模型的竞争状态。.** 《科创...
- **中國AI 橫掃全球！字節Seedance 2.0 震撼發布讓好萊塢集體失業** (relevance: 72%) https://www.youtube.com/watch?v=Rw8A3_UnJNM TikTok母公司字節跳動正式發布新一代影片生成模型「Seedance 2.0」，在社群平台引發大量討論。技術上支援自動分鏡、音效匹配及物理邏輯運算，

2.2 技术突破

Answer

ByteDance has made significant investments in AI, including developing its own hardware and attracting top talent. The company focuses on advancing AI technologies to enhance its products. ByteDance's AI efforts aim to drive innovation and maintain competitive edge.

- seed.bytedance.com/blog/
%E4%B8%80%E4%B8%AA%E7%AE%97%E6%B3%95%E5%8D%9A%E5%A3%AB%E6%AF%95%E4
%E6%88%91%E5%9C%A8%E5%AD%97%E8%8A%82%E8%B7%B3%E5%8A%A8%E5%81%9A%E5

隐秘扩容，规模正式突破千人，其中深耕AI芯片研发的人员超500人，CPU研发团队规模则稳定在200人左右。

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析字节跳动在以下方面的进展：

- **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 provide information based on known facts. I do not identify as any specific model name.

Sources

- **字节跳动AI研究员暗示，即将发布比Gemini更强大的开源模型 - Reddit** (relevance: 100%) https://www.reddit.com/r/singularity/comments/18cj8pe/bytedance_ai_researcher_suggests_that_open_source/?tl=zh-hans 大型语言模型(LLM)只是构建具有自主性的人类智能的第一个可用的构建模块。我不认为它们可以独立完成所有事情。LLM 的风险仅在于它们可以帮助自动化

- **26年2月底AI大模型动态跟踪——模型狂发** (relevance: 100%) <https://zhuanlan.zhihu.com/p/2011534791091176226> 文本竞技场：Claude Opus 4.6略微领先 Gemini 3.1 Pro，占据第一把交椅。前10名出现了唯一一个国产LLM，字节才发布的 seed-2.0-pro。 编码竞技场：Claude Opus
- **豆包、Claude五大主流AI模型深度对比：技术、生态与应用全景解析** (relevance: 100%) <https://blog.csdn.net/gaowenhui2008/article/details/156852153> 本报告旨在对当前市场上五款具有代表性的主流大语言模型——Google Gemini、OpenAI ChatGPT、通义千问（Qwen）、字节跳动豆包（Doubao）及Anthropic Claude——
- **字节跳动开源Seed-X 7B多语言翻译模型：28语种全覆盖 - 腾讯云** (relevance: 100%) <https://cloud.tencent.com/developer/article/2543952> 在机器翻译领域，如何在保证高质量的同时兼顾模型规模与推理效率一直是研究与工程应用的两难选择。近日，字节跳动团队开源了 **Seed-X** 系列多语言翻译模型（7B），通过精巧的模型结构设计与强化学习微调，实现在仅 7 亿参数规模下，对 28 种语言的翻译性能媲美甚至超越诸如 Gemini-2.5、Claude-3.5、GPT-4 等超大模型。Seed-X 的开源，不仅为学术研究提供了强有力的基线，也为各行业落地翻译应用带来了轻量、高效的新选择。以下将从模型概述、架构与训练流程、性能评估、多场景部署及应用、快速使用示例等方面进行深度剖析，并保留所有原始图片与表格，帮助读者全面了解 Seed-X...
- **四大AI模型终极对决：Gemini、GPT、Claude、Grok谁更强？附实测** (relevance: 98%) <https://juejin.cn/post/7615962569573285903> # 四大AI模型终极对决：Gemini、GPT、Claude、Grok谁更强？附实测. ## 如果你想一次性体验全球顶尖AI模型的中文能力，目前国内可直接访问的聚合镜像平台RskAi (ai.rsk.cn) 提供了最佳解决方案。它集成了 **Gemini 3 Pro、GPT-4o、Claude 3.5 Sonnet、Grok-2**四款主流模型，支持文件上传和联网搜索，且完全免费。本文通过五大维度实测对比，帮你找到最适合自己需求的模型。不同大语言模型在架构设计、训练数据、优化目标上存在差异，导致各自擅长领域不同。Gemini 3 Pro强在多模态与长上下文，GPT-4o均衡全面，Claude 3...

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

ByteDance released its first reasoning model, leveraging reinforcement learning and cognitive chains to enhance complex logic tasks. DeepSeek's R1 model gained attention for its superior reasoning capabilities. The competition among internet companies for advanced AI reasoning models intensified.

Sources

- **字节跳动发布首款推理模型加剧基座模型竞争 - 财新网** (relevance: 100%) <https://m.caixin.com/m/2025-04-17/102310638.html> # 字节跳动发布首款推理模型 加剧基座模型竞争. 【财新网】4月17日, 字节跳动旗下火山引擎面向B端发布豆包1.5深度思考模型, 这是字节跳动首款推理模型, 可在解决问题时“边想边搜”, 根据目标规划搜索路径; 同时具备视觉推理能力, 可以综合理解图片中的各类信息. . 推理模型是指模型在预训练之后的阶段采用强化学习、思维链的技术, 进一步“训练”提高模型处理复杂逻辑推理任务的能力. . OpenAI于2024年9月率先推出o1模型让业界转向推理模型, 而后DeepSeek于1月20日推出推理模型R1, 凭借顶尖的能力引发了广泛讨论. 此后, 互联网厂商间AI基座模型的竞争加剧, 腾讯、阿里先后推出T1、QwQ...
- **字节豆包推理模型Seed-Thinking-v1.5要来了 - DeepSeek技术社区** (relevance: 100%) <https://deepseek.csdn.net/67f9c93be47cbf761b58751f.html> 推理模型主要依赖思维链 (CoT) 数据, 这种数据展示逐步推理过程. 该团队的初步研究表明, 过多非思维链数据会削弱模型探索能力. 研究团队在强化学习
- **DeepSeek-R1: 突破推理能力的满血开源O1 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/19878591772> 这一创新与近期OpenAI提出的RFT以及字节跳动的ReFT有异曲同工之妙. 这些方法都指出, SFT可能存在搜索空间不足的局限性, 而强化学习则可以探索更多的
- **Large Model Application Algorithm Research ... - 字节跳动招聘官网** (relevance: 100%) <https://jobs.bytedance.com/campus/m/position/detail/7508770413570001159?recomId=1cbb5b17-eb3d-11f0-94fe-fa163e53fcf3&sourceJobId=7508738343146178823> 字节跳动 ... 为了提升推理能力, OpenAI 的o1 系列模型通过增加思维链 ... 最近deepseek r1在论文中提到通过纯强化学习的方法
- **DeepSeek推理模型预览版上线, 解密o1推理过程** (relevance: 100%) <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. . ### ...

4.多模态模型

检索关键词: 多模态,视觉,视频生成,Sora,Seedance

Answer

ByteDance's Seedance 2.0 and Sora are advanced multi-modal models for high-quality video generation, promising industry-leading performance in visual and narrative coherence. These models aim to revolutionize AI-driven video production, reducing costs and enhancing creative control.

Sources

- **多模态大模型真能生成高质量视频吗？55%用户认为技术已突破** (relevance: 100%)
<https://post.smzdm.com/p/a7g7zq4l> 2025至2026年，Seedance 2.0、Sora 2、Veo 3等多模态大模型密集发布，宣称实现“原生音画同步”“角色一致性”“物理世界模拟”等能力，推动AI视频生成进入工业化应用阶段。
- **Seedance 2.0 - ByteDance Seed** (relevance: 100%) https://seed.bytedance.com/zh/seedance2_0 # Seedance 2.0. Seedance 2.0 采用统一的多模态音视频联合生成架构，支持文字、图片、音频、视频四种模态输入，集成了目前业界最全面的多模态内容参考和编辑能力。 . # 极致拟真的视听体验. # 所想即所见的导演级操控. 支持音、视、图全模态参考输入，打破素材边界，赋予创作者对表演、光影、运镜的调度权，超强的可控性让创意转化为画面，真正实现“像导演一样生成”。 . # 影视工业链路赋能. 深度适配广告、影视与社媒营销场景，输出质量对齐工业交付标准，大幅度降低特效制作与实拍成本，为行业带来显著的效率提升. # 模型表现. 以下是 Seedance 2.0 在内部基准测试 See...
- **可能會搞到好多人失業字節跳動旗下的Seedance2.0的AI影片生成 ...** (relevance: 100%)
<https://www.instagram.com/reel/DUm6Bgfj8aG/> 其核心模型疑似結合了多模態生成、聲紋建模與圖像語義辨識技術，使最終 ... 字節跳動(ByteDance) 最近推嘅Seedance 2.0 視頻生成大模型嘅網絡上
- **Seed Models** (relevance: 100%) <https://seed.bytedance.com/zh/models> 多模态理解能力全面升级，LLM 与 Agent 表现大幅强化. 专注多模态理解与推理的视觉-语言基础模型，在 38 项公开评测基准中取得 SOTA 表现. Seed1.5 (Doubao-1.5-pro) . Seed1.5 (Doubao-1.5-pro) . 实验性扩散语言模型，专注于代码生成，推理速度可达 2146 tokens/s. 采用统一的多模态音视频联合生成架构，支持文字、图片、音频、视频四种模态输入，集成了目前业界最全面的多模态内容参考和编辑能力. 新一代图像创作模型，能够灵活应对复杂的多模态生成任务，支持 4K 高清图像生成. 3D 生成基础模型，纹理与材质生成性能 SOTA，可准确...
- **Seedance 1.0 - ByteDance Seed** (relevance: 100%) <https://seed.bytedance.com/zh/seedance> # Seedance 1.0. 一个模型同时支持文字和图片生成视频，原生具备多镜头叙事能力。在语义理解和指令遵循方面取得突破，可生成运动流畅、细节丰富、具备影视级美感的 1080p 高清视频。 . # 流畅稳定的运动表现. Seedance 1.0 具备良好的动态范围，能够

流畅地生成大幅度的运动。从细微的表情到动态场景，均能保持高水平的稳定性与物理真实感。· # 原生多镜头叙事能力· # 多样的风格化表达· 无论是写实摄影、赛博朋克，还是手绘插画、毛毡特效，Seedance 1.0 都能准确响应多样的风格指令，以支持各类创意需求。· # 精准的语义理解与指令遵循· 能够精准解析复杂的自然语言指令...

4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

Answer

字节跳动采用H100和B200 GPU，以及TPU，用于大规模AI训练。H100具有高算力和能效，B200提供更高功率。TPU用于智能算力优化。

Sources

- 兆规模起飞！“平替” 輝達Blackwell架構GPU！智能算力TPU產業鏈 ... (relevance: 100%) <https://hao.cnyes.com/post/212327> B200叢集：同等算力需10240卡，功耗85MW; 百萬卡規模：功耗達7.3GW ... 9.4.2 字節跳動：應用驅動的算力採購· 需求：抖音、TikTok推薦系統日訓練
- 谷歌的TPU技术相较于英伟达的GPU如何？谷歌TPU对外销售吗？ (relevance: 100%) <https://www.zhihu.com/question/1942748207240176788/answer/1984562693403850395> ... 算力解决方案，这场技术竞赛正在重塑全球芯片产业格局。NVIDIA 的算力霸权：从H100 到B200 的持续领跑· NVIDIA 凭对AI 算力脉动的精准卡位，筑起了
- Nvidia Blackwell B100 vs B200 vs GB200NVL72 性价比分析- 文章 (relevance: 99%) <https://developer.volcengine.com/articles/7387625290035855370> 风冷 700W 的 B100 将是首批发货的型号，提供 1750 TFLOPS 的 FP16/BF16 计算性能。B100 的底板设计可以插入今天 HGX H100 系统中使用的相同设计中，迫使 B100 在现有系统的热包络内以较低的功率和时钟速度运行。很快在 B100 发货后，B200 将以更高的功率和更快的时钟速度进入市场，提供 2250 TFLOPS 的 FP16/BF16 计算性能。此外，液冷的 GB200 NVL72 将允许 Blackwell GPU 在更高的功率水平下运行，进一步提升性能——提供 2500 TFLOPS 的 FP16/BF16 计算性能，比 H100 和 H200...
- 百芯竞逐，芯片产业的“新战国时代” (relevance: 99%) <https://zhuanlan.zhihu.com/p/2002259261925909478> 主流ASIC在算力性能上已基本对齐英伟达H系列GPU，但能效比优势突出，同代际芯片具体指标对比：· ○算力方面，英伟达H100 FP16算力为990 TFLOPS，谷歌

- **10万卡H100集群有多难搭？一文解析算力集群技术要点 - 腾讯云** (relevance: 98%) <https://cloud.tencent.com/developer/article/2432525> ## 有钱买卡还不够，10万卡H100集群有多难搭？一文解析算力集群技术要点. # 有钱买卡还不够，10万卡H100集群有多难搭？一文解析算力集群技术要点. ##### **【新智元导读】**在英伟达市值猛涨、各家科技巨头囤芯片的热潮中，我们往往会忽视GPU芯片是如何转变为数据中心算力的。最近，一篇SemiAnalysis的技术文章就深入解读了10万卡H100集群的构建过程。包括但不限于OpenAI、微软、xAI和Meta在内的多个头部公司都在争相建立超过10万卡的GPU集群，在这个规模上，仅仅是服务器的成本就超过40亿美元，还要受到数据中心容量和电力不足等多项因素的限制。我们可以做...

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

ByteDance uses HBM and NVLink for high-bandwidth memory and efficient data transfer in AI systems. NVLink offers significantly higher bandwidth than PCIe, crucial for large AI models. HBM enhances data throughput speed in GPU-based AI computations.

Sources

- **AI领域GPU研发最新进展与趋势报告 - X** (relevance: 100%) <https://x.com/myfxtrader/status/1888959527883804783> 整个ExaPOD系统拥有**1.3TB片上SRAM和13TB高带宽内存（HBM）**供模型数据存储(Tesla Dojo - Wikipedia)。... 据《华尔街日报》报道，华为已邀请字节跳动、百度
- **被卡脖子的真相：为什么造GPU算力卡这么难？（科普文章） - 新浪** (relevance: 64%) http://k.sina.com.cn/article_5953466437_162dab0450670a7go6.html 这类算力卡彻底弱化显示输出功能，重点强化并行计算效率、HBM协同能力（高带宽显存，High Bandwidth Memory，核心作用是提升数据吞吐速度）等核心特性，是AI技术
- **2026年AI Infra重点发展方向的深度分析** (relevance: 31%) <https://zhuanlan.zhihu.com/p/1999176653281187226> 节点内多路径互连（MIHF）架构将聚合PCIe、NVLink和CXL带宽，实现主机内存到GPU的高效数据供给，支撑“无限内存”训练，。推荐系统的分层存储（HPS）：针对推荐
- **深度解析NVIDIA的超节点架构演进原创 - CSDN博客** (relevance: 18%) <https://blog.csdn.net/reiase/article/details/152609995> 以H100为例，NVLink 4.0提供900GB/s的GPU间带宽，而PCIe 5.0 x16仅128GB/s，相差7倍。随着模型规模增长，这种差距会成为整个AI算力集群的核心瓶颈。内存墙

- **[PDF] 拥抱AI 和存储国产化机会** (relevance: 4%) <https://pic-test-gjmetal-1324067834.cos.ap-shanghai.myqcloud.com/news2/af21ee24dd424b1084b2762d94f030bf20251219184025.pdf> 中国存储公司的α：两存全球市占率提升可期 DRAM 侧，2024 年长鑫存储推出新一代16Gb DDR5 产品，跳过17nm（D1y）节点直接采用16nm（D1z）节点。产能和出货上，长鑫也在加速从DDR4/LPDDR4 向DDR5/LPDDR5 转移，根据Counterpoint 预测，2025Q1 长鑫存储按比特出货量全球市占率6%，预计2025Q4 将提升至8%；年底DDR5 市占率预计将从2025Q1 的1%提升至7%，LPDDR5 预计将从2025Q1 的0.5%提升至9%。此外，AI 算力驱动下，国内HBM 及3D DRAM 产业化也有望加速，共同驱动存储产业...

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

ByteDance uses FlashAttention for efficient attention mechanism, along with quantization and optimization techniques to speed up model inference. FlashAttention reduces memory access and improves GPU efficiency. It also supports hardware-aware optimizations for better performance.

Sources

- **端侧AI Infra开发实习生-音视频技术** (relevance: 60%) <https://jobs.bytedance.com/campus/m/position/detail/7594472672476760325?recomId=5a64c401-1efc-11f1-b8f3-043f72a6392c&sourceJobId=7532098404223666450> 1、负责模型推理全流程优化，低比特量化（AWQ、GPTQ、SpinQuant等）、模型剪枝、蒸馏等模型压缩技术，以及FlashAttention、KVCache高效管理、投机推理等推理优化手段
- **4比特量化三倍加速不掉点！清华即插即用的SageAttention ...** (relevance: 53%) <https://cloud.tencent.com/developer/article/2496492> 此前，清华大学陈键飞团队提出的 8-Bit 的即插即用 Attention（SageAttention），将 Attention 中的 QK^T 量化至 INT8，将 PV 保持为 FP16 精度并使用 FP16 精度的矩阵乘法累加器，同时提出 Smooth K 技术保持了量化 Attention 的精度，实现了 2 倍加速于 FlashAttention2，且在各类大模型上均保持了端到端的精度表现。SageAttention2 实现了高效的 Attention 算子，可以实现即插即用的推理加速。输入任意 Q, K, V 矩阵，SageAttention2 可以快速返回 Attention...
- **突破大模型精度瓶颈：FlashAttention大输入场景下的数值 ...** (relevance: 53%) https://blog.csdn.net/gitblog_00661/article/details/151456462 优化推理效率，尤其是在大模型环境下，就需要高效地利用SRAM和HBM资源，以避免带宽瓶颈。借助于Ampere架构的新特性，数据可以直接从HBM拷贝到SRAM，提高了SRAM

- **FlashAttention 系列技术详解：加速大模型训练的利器** (relevance: 50%) <https://developer.vlccengine.com/articles/7444834108175286298> # FlashAttention 系列技术详解：加速大模型训练的利器. ## FlashAttention 的解决方案. 为了解决上述问题，**FlashAttention** 提出了一种 **快速、节省内存、精确的注意力计算方法**，通过减少内存读写次数，提高 GPU 的内存 IO 效率，加快训练速度，并增加上下文窗口长度。其核心目标是通过 **硬件感知 (IO-awareness)** 的算法优化，将整个注意力计算过程在更高速的 **SRAM**（静态随机存取存储器）中完成，减少对 HBM（高带宽内存）的依赖。在传统的注意力机制中，首先需要将输入矩阵从 HBM 中读取，随后计算注意...
- **字节高性能Transformer推理库获IPDPS 2023最佳论文奖** (relevance: 19%) <https://developer.aliyun.com/article/1262677> ### 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. # 大幅优化推理过程，字节高性能Transformer推理库获IPDPS 2023最佳论文奖. ### 为什么选择阿里云. ### 大模型. ### 产品和定价. ### 技术内容. ### 权益. ### 服务. ### 关注阿里云. ### 友情链接. © 2009-...

4.Agent

检索关键词: 智能体,Agent,AutoGPT

Answer

AI agents are autonomous systems that perform tasks using large language models, planning, and tool use. They aim to reduce human intervention in complex processes. Major tech companies are developing and deploying AI agents to enhance productivity and automate tasks.

Sources

- **2026年AI Agent搭建平台全景图：12个主流工具深度对比与选型指南** (relevance: 100%) <https://www.betteryeah.com/blog/2026-ai-agent-platform-guide> AutoGPT作为自主智能体的开源鼻祖，以其强大的任务拆解与自主执行 ... 字节跳动推出的一站式AI Bot开发平台，以其零门槛的可视化开发体验和强大
- **AI智能体(Agent)保姆级入门指南，零基础小白也能轻松上手 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/1995097301845309381> Coze (扣子)：由字节跳动出品，完全免费，界面友好，非常适合新手。你可以用它快速创建聊天机器人、知识库问答、工作流等各种类型的Agent，并一键发布到豆包、

- **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：
- **AI-Compass Agent智能体技术生态：整合AutoGPT、LangGraph** (relevance: 100%) <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...
- **科技巨头狂卷“智能体”，大模型上终于长出了“大家伙”？** (relevance: 99%) <https://m.cyzone.cn/article/774291.html> The Information 援引内部消息报道称，OpenAI 计划最快将在今年秋天推出代号「草莓（Strawberry）」的全新 AI，其拥有前所未有的「推理」能力，可以处理复杂的数学和编程任务，甚至还能体现在日常生活中的非技术问题上。 . 此外，报道还指出这项技术对未来 AI 产品，特别是旨在解决多步骤任务的「智能体（Agent）」具有重要意义。 . 在 2022 年年底 ChatGPT 大火之后，「智能体」很快就从故纸堆中一跃而出，引起整个行业的广泛关注。而从开源项目 AutoGPT 到 OpenAI 官方推出的 GPTs 和 GPT 商店，作为「雏形」，也都在一定程度上展现了 AI 智能...

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，字节跳动的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

预测字节跳动在未来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 08:24:52 AM CST 2026