

阿里巴巴 AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:24:52

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: 阿里巴巴

主要产品: 通义千问, Qwen

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

Alibaba released new AI models, including Qwen3 Max, claiming superior performance over competitors. The models aim to enhance capabilities in complex reasoning and natural interactions. The launch highlights Alibaba's competitive edge in AI technology.

Sources

- 阿里巴巴发布了多款新AI模型 (relevance: 88%) <https://mimai.cn/article/detail?fid=1888564856&efid=olq4v6jxs-8Moh8dh076oA> 在2025年云栖大会上, 阿里巴巴发布了多款新AI模型, 包括Qwen3-Max、Wan2.5、Qwen3-VL等, 展示了其在AI领域的强大实力和创新能力。关键点: 1. 阿里云栖大会于2025年9
- 阿里巴巴发布AI模型声称超越DeepSeek - 美国之音 (relevance: 88%) <https://www.voachinese.com/a/alibaba-releases-ai-model-it-claims-surpasses-deepseek-v3-20250129/7955186.html> ## 无障碍链接. ##### 关注我们. 中国时间 13:37 2026年3月12日 星期四. # 阿里巴巴发布AI模型 声称超越DeepSeek. 阿里巴巴透过旗下的阿里云, 发表了自家的AI语言模型通义千问Qwen 2.5-Max超大规模MoE模型, 甚至号称测试表现上优于

DeepSeek V3。(美联社：2024年5月14日). 阿里巴巴透过旗下的阿里云，发表了自家的AI语言模型通义千问Qwen 2.5-Max超大规模MoE模型，甚至号称测试表现上优于DeepSeek V3。(美联社：2024年5月14日). 中国科技公司阿里巴巴星期三(1月29日)发布了其通义千问“Qwen 2.5”人工...

- **阿里巴巴發佈Qwen3 混合推理模型樹立開源AI新標桿 - Alibaba Cloud** (relevance: 85%)
https://www.alibabacloud.com/tc/press-room/alibaba-introduces-qwen3-setting-new-benchmark?_p_lc=1 crystal.liu@alibaba-inc.com. # 阿里巴巴發佈Qwen3 混合推理模型樹立開源AI新標桿. Qwen3系列包含**六款密集模型與兩款混合專家模型（Mixture-of-Experts, MoE）**，有助開發者更靈活地開發適用於移動設備、智能眼鏡、自動駕駛汽車及機器人等場景上的應用。Qwen3全系列模型現已開源並全球上線，包含六款密集模型（參數量0.6B、1.7B、4B、8B、14B、32B）及兩款MoE模型（30B總參數、3B激活參數；235B總參數、22B激活參數）。##### 混合推理結合思考與非思考模式. Qwen3標誌著**阿里巴巴首次推出混合推理模...
- **阿里发布最新AI推理模型Qwen3 Max Thinking - 新浪财经** (relevance: 85%) <https://finance.sina.com.cn/stock/usstock/c/2026-01-27/doc-inhiseym7997912.shtml> 阿里巴巴发布最新人工智能推理模型Qwen3 Max Thinking，称该模型在准确性、复杂推理和智能体等方面较之前版本实现了“显著的性能提升”。
- **阿里巴巴发布新一代AI模型千问3，市场竞争日趋激烈** (relevance: 84%) https://cn.wsj.com/articles/%E9%98%BF%E9%87%8C%E5%B7%B4%E5%B7%B4%E5%8F%91%E5%B8%83%E6%96%B0%E4%E5%B8%82%E5%9C%BA%E7%AB%9E%E4%BA%89%E6%97%A5%E8%B6%8B%E6%BF%80%E7gaa_at=eafs&gaa_n=AWetsqeRPrChQUzAOzqvQVogL2IQCG2-FpSaLLBGEFjCoiTs9Wt_Bd4hi7kZ&gaa_ts=69b8a2d9&gaa_sig=BqxCnpz72_R4jy1xLwtg_8YHz_13vaO91RGckTpyN7WQUQHwNtl9cA_YR1Q%3D%3D 阿里巴巴(Alibaba)推出新一代大语言模型，紧随竞争对手的升级步伐。眼下中国公司正在人工智能(AI)领域展开激烈比拼。这家总部位于杭州的公司周二称，其

2.2 技术突破

Answer

Alibaba has made significant technological breakthroughs in AI, cloud computing, and big data. It focuses on advancing AI infrastructure and open collaboration. Alibaba's innovations have driven digital transformation and economic growth.

Sources

- **阿里巴巴双11技术演进与突破：技术创新与架构优化原创 - CSDN博客** (relevance: 76%) https://blog.csdn.net/weixin_31620365/article/details/149795639 简介：本书深入探讨了阿里巴巴在双11背后的科技创新与架构优化，揭示了面对极端业务挑战时的技术解决方案。介绍了从单体架构向微服务架构的转变，以及在
 - **《麻省理工科技评论》全球十大突破性技术，阿里巴巴正研究其中4项** (relevance: 74%) <https://developer.aliyun.com/article/495290> 中国科技巨头阿里巴巴成为给所有人的人工智能（云端AI）、对抗性神经网络、传感城市、材料的量子飞跃四项技术主要研究者。“全球十大突破性技术”是科技领域的权威榜单，至今
 - **2023's Top 10 Tech Trends By Alibaba's DAMO Academy** (relevance: 70%) <https://www.alibabagroup.com/document-1549860542816976896> 2023's Top 10 Tech Trends By Alibaba's DAMO Academy. 科技发展日新月异，对社会发展影响深远。阿里巴巴达摩院在新年伊始发布2023十大科技趋势预测，涵盖人工智能、云计算等领域的前沿科技，这些趋势或将重塑不同行业的未来。· **阿里巴巴达摩院院长张建锋**表示：“展望2023年，多元技术的协同并进，将驱动计算与通讯、硬体和软体的融合；科技的广泛应用意味着将有更多 AI与其他数字技术在各个垂直市场推出，促进公私营企业及个人在安全技术与管理上的协作。由科技进步及产业应用驱动的创新已成为不可逆转的宏大趋势。”· 达摩院通过分析公开论文和专利申请等基础数...
 - **阿里巴巴以先进技术服务社会千行百业 - 世界互联网大会** (relevance: 64%) https://cn.wicinternet.org/2025-04/25/content_37990515.htm # 阿里巴巴以先进技术服务社会千行百业. 阿里巴巴集团（以下简称“阿里巴巴”）创立于1999年，是一家以“让天下没有难做的生意”为使命的科技公司。成立25年以来，阿里巴巴通过持续的技术创新与业务拓展，构建起了一个繁荣的互联网平台生态。集团现拥有淘宝天猫、国际数字商业、云智能、本地生活、菜鸟和大文娱等业务部门，以及钉钉、飞猪、灵犀互娱等多家业务公司。阿里巴巴电商板块服务10亿中国消费者和超过3亿海外消费者；阿里云是亚太第一、全球前列的云计算服务商，全国80%的科技公司、60%的A股上市企业和65%的“专精特新”企业都在使用阿里云的服务。· 阿里巴巴始终把创新作为核心战略，坚持高强度投入研发，致力...
 - **MIT评出全球十大突破性技术阿里巴巴连摘两项 - 科技** (relevance: 60%) <https://tech.huanqiu.com/article/9CaKrnK0I5t> 今年《麻省理工科技评论》公布的十大突破性技术榜单依次为强化学习、360°自拍、基因疗法2.0、细胞图谱、自动驾驶货车、刷脸支付、太阳能热光伏电池、实用
-

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析阿里巴巴在以下方面的进展：

- **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 my training data. I do not identify as any specific model name.

Sources

- **2025年主流大模型全景对比：Grok、Claude、ChatGPT与Gemini的 ...** (relevance: 70%) <https://www.cnblogs.com/gccbuaa/p/19264126> # gccbuaa. # 2025年主流大模型全景对比：Grok、Claude、ChatGPT与Gemini的战场 - 教程. 在人工智能技术突飞猛进的2025年，大语言模型（LLM）已成为驱动企业数字化转型的核心引擎。本文聚焦Grok、Claude、ChatGPT和Gemini四大代表性模型，从技能架构、性能特点到适用场景进行全面解析，助您精准选择适配业务需求的AI解决方案。 . Gemini是谷歌DeepMind团队研发的原生多模态模型，采用单一架构统一处理文本、图像、音频和视频，实现跨模态隐式对齐，幻觉率降低35%。其核心优势在于实时搜索增强，可调用Google Search材料补全时效性...

- **阿里巴巴的Qwen (通義千問) 是我心目中頭五大LLM 模型。分別是 ...** (relevance: 67%)
<https://www.threads.com/@ar.shek/post/DCiqCk9tQj0/> 阿里巴巴的Qwen (通義千問) 是我心目中頭五大LLM 模型。分別是Claude, GPT, Gemini, Mixtral, Qwen。其中只有 Gemini 和Qwen 剛推出的2.5 Turbo
- **ChatGPT成為各家模型挑戰對象2025年獨領風騷的10個大語言模型** (relevance: 67%)
<https://www.cmmedia.com.tw/home/articles/52148> # ChatGPT成為各家模型挑戰對象 2025年獨領風騷的10個大語言模型. ## 輝達執行長黃仁勳在今年CES展上談到大型語言模型 (Large language model) 是AI產業發展的指標，從矽谷到杭州，各家ai公司今年都將推出最新的LLM模型。(圖片來源／Gemini的AI生成圖片). 輝達執行長黃仁勳在今年CES展上談到大型語言模型 (Large language model) 是AI產業發展的指標，從矽谷到杭州，各家ai公司今年都將推出最新的LLM模型。(圖片來源／Gemini的AI生成圖片). 大型語言模型 (Large language model) 是AI產業發展的指標，2024...
- **2026年十大最佳大型語言模型 (LLM) - Botpress** (relevance: 62%) <https://botpress.com/tw/blog/best-large-language-models> | GPT-4o |  128K | 輸入\$5／輸出\$15 |. | Claude 4 Sonnet |  200K | 輸入\$3／輸出\$15 |. | Grok 3 |  131K | 輸入\$3／輸出\$15 |. ## 最佳推理型 LLM. | OpenAI o3 |  200K | \$10 輸入 / \$40 輸出 |. | Claude 4 Opus |  200K | \$15 輸入 / \$75 輸出 |. | Gemini 2.5 Pro |  100 萬 | \$1.25 輸入 / \$10 輸出 |. | DeepSeek R1 |  128K ...
- **2025：大语言模型 (LLM) 之年 - 36氪** (relevance: 60%) <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...

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

DeepSeek R1系列模型使用强化学习训练，具备长思维链推理能力，在复杂逻辑任务上表现优异。其推理过程展示了详细的思考过程，性能超越多个基准测试。

Sources

- **类o1系列模型大盘点：QwQ - DeepSeek技术社区** (relevance: 100%) <https://deepseek.csdn.net/67ab1e2879aaf67875cb9ab6.html> 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
- **国内“推理模型”卷疯了！类o1推理模型，谁更强？** (relevance: 100%) <https://www.icviews.cn/semiCommunity/postDetail/7426> DeepSeek-R1-Lite 是深度求索推出的新一代AI推理模型，用强化学习训练，具备长思维链推理能力，能实时展示推理思考过程，性能在多个基准测试中超越GPT-4 等
- **旺精通~阿里START：自学工具调用的长思维链推理模型** (relevance: 100%) <https://zhuanlan.zhihu.com/p/32078743531> 随后，以OpenAI-o1（OpenAI，2024b）和DeepSeek-R1（DeepSeek-AI，2025）为代表的强化学习取得突破，建立了一种名为长思维链的新范式，该范式模仿人类的认知策略，
- **模型思考效率评测最佳实践 - EvalScope** (relevance: 100%) https://evalscope.readthedocs.io/zh-cn/v0.12.1/best_practice/think_eval.html 特别是对于R1类模型，其输出通常包含较长的思维链，输出token数量往往超过1万。使用高效的推理框架部署模型可以显著提高推理速度。下面以DeepSeek-R1-Distill-Qwen-7B为例，
- **一种先进的人工智能形式——于去年9月由美国公司OpenAI ...** (relevance: 100%) <https://x.com/dotey/status/1883700968501469535> 世界上首个“推理模型”（reasoning model）——一种先进的人工智能形式——于去年9月由美国公司OpenAI 发布，名为“o1”。该模型采用“思维链”（chain of

4.多模态模型

检索关键词: 多模态,视觉,视频生成,Sora,Seedance

Answer

Alibaba's Seedance 2.0 is a leading AI video generation model known for its multi-modal capabilities and high-quality video output. It integrates visual and audio generation with narrative control. Seedance 2.0 excels in complex motion and scene consistency.

Sources

- **Seedance vs Sora vs Kling：AI 视频生成模型深度对比** (relevance: 100%) <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一骑绝尘背后：中国AI春节前为何“杀疯了”？ - 新湖南** (relevance: 100%)
<https://m.voc.com.cn/xhn/news/202602/31575800.html> 在国产AI全产业链自主化方面，Seedance实现核心算法、训练框架与关键技术的自主可控，坚持以真实产业需求为导向，牵引多模态理解、长视频生成、高效算力调度
- **Seedance 原理全解：从架构设计到核心代码，这篇彻底讲清楚了** (relevance: 100%)
<https://developer.aliyun.com/article/1714692> ### 探索云世界. #### 热门. # Seedance 原理全解：从架构设计到核心代码，这篇彻底讲清楚了. Seedance 1.5 Pro做了一个极其激进的决定：在模型层面，直接把音频和视频当成一回事来处理。 . fal_client.api_key = os.getenv(“FAL_KEY”) . def generate_video_from_text(prompt, duration= “5” , resolution= “720p”):. “fal-ai/bytedance/seedance/v1.5/pro/text-to-video” , . “duration” : dura...
- **Seedance 2.0 正式发布 - ByteDance Seed** (relevance: 100%) <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. 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， ...
- **[PDF] Seedance2.0：生成式视频的技术奇点与产业重构** (relevance: 100%) 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， ...

4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

Answer

Alibaba's "平头哥" is developing competitive GPUs to rival foreign brands like Nvidia. These GPUs aim to improve performance and efficiency to compete in AI and computing markets. The focus is on achieving cost-effective solutions for large-scale deployments.

Sources

- **百芯大戰 - 富途牛牛财经新闻** (relevance: 70%) <https://news.futunn.com/hk/post/68242330#> 百芯大戰. 一年前，我們在《DeepSeek掀起算力革命，英偉達挑戰加劇，ASIC芯片悄然崛起》一文中，更多的是看好ASIC帶來類似博通和晶圓代工的產業機會。一年後的今天，ASIC的發展速度遠超預期。尤其近半年以來，ASIC甚至逐漸成為AI競爭的勝負手：國內外大廠開年以來股價表現最好的分別是百度、谷歌和阿里。谷歌TPU+自研模型+雲+內部應用的王炸，已經讓其立於不敗之地；國內互聯網大廠，近期被重估的只有自研ASIC芯片拆分獨立IPO的百度（計劃拆分崑崙芯IPO）和阿里（計劃拆分平頭哥IPO）。ASIC (Application Specific Integrated Circui...
- **[PDF] AI系列专题报告（一） - 算力** (relevance: 68%) https://pdf.dfcfw.com/pdf/H3_AP202506121689781660_1.pdf AI系列专题报告（一）算力：算力基建景气度高，国产AI芯片发展势头良好 证券研究报告 分析师：陈福栋S1060523070003（证券投资咨询） 分析师：闫磊 S1060519100002（证券投资咨询） 平安证券研究所电子信息团队 2025年6月12日 请务必阅读正文后免责条款 电子行业强于大市（维持） 核心摘要 ☑AIGC蓬勃发展，对底层智能算力产生强劲需求。行业前期，训练是算力需求的主力，大量大模型训练需要海量算力支撑。2024年末，DeepSeek重磅发布，其轻量化、低成本、高性能特征大幅拉低了AI应用门槛，有望成为各类推理场景爆发的契机，推理算力市场需求潜力巨大。在此背景下...
- **美股懂哥解读小米、谷歌、英伟达、阿里巴巴与GPU/TPU关系研报** (relevance: 67%) <https://www.tiktok.com/@user5860790033378/video/7586618237295545618> 2015年推出首款TPU，2025年发布的TPU v7峰值算力达4614 TFLOPS，能效超英伟达B200，已获Meta数十亿美元采购订单。TPU专为自家TensorFlow框架及Gemini大模型
- **一张图说清：H100、H200、B200 到底该怎么选？ - 博客园** (relevance: 65%) <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...

- **揭秘！阿里隐藏大招曝光，一文读懂超火GPU - 36氪** (relevance: 65%) <https://eu.36kr.com/de/p/3658910978372356> # 阿里放出隐藏大招？一文读懂大火的GPU. 近期，阿里又放出一枚重磅炸弹：阿里巴巴集团已决定，支持旗下芯片公司“平头哥半导体”未来独立上市。· 据说，平头哥在2025年推出的通用GPU芯片（PPU），综合性能可以对标英伟达H20，升级版性能则可以比肩A100。· 这个A100，就是前段时间刚刚解禁的，英伟达H200的上一代产品，也是当下中小规模的AI训练，性价比最高的产品之一。· 12月5日，被称为“小英伟达”的摩尔线程登陆科创板，高开涨幅468%，中一签就是小27万元；· 一边是市场的火热，另一边，是对手千方百计的阻挠，这个GPU赛道，究竟有着怎样的含金量？· ## **到底什么是GPU？ ...

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

Alibaba focuses on AI data storage with HBM, DRAM, and NVLink technology to enhance computing power. AI-driven demand drives market growth for high-speed interconnects and storage solutions.

Sources

- **AI拿走HBM，手机与PC啃剩饭？ - 科技- 新浪** (relevance: 100%) <https://tech.sina.cn/2026-01-30/detail-inhkauvi0528308.d.html?vt=4> 以往存储周期可能会出现 A “涨” B “稳”，或者C “涨” D “平”的分化，而本次涨价的异常之处在于，它几乎席卷了全球所有存储细分市场：HBM、DRAM、NAND、HDD等全品类存储
- **国内外AI 芯片概述 - CSDN博客** (relevance: 100%) <https://blog.csdn.net/fuhanghang/article/details/135328443> NVIDIA GPU 显存有两种类型，GDDR 和HBM，每种也有不同的型号。针对显存我们通常会关注两个指标：显存大小和显存带宽。HBM 显存通常可以提供更高的显存
- **是超越英伟达了吗？和阿里的平头哥算力芯片比如何？ - 知乎** (relevance: 99%) <https://www.zhihu.com/question/1951971387632751207> 此外，NVLink还支持原子操作和同步，从而在精细粒度上实现数据一致性。利用这些能力，NVLink可以实现以下关键功能：· 内存池化（Memory Pooling）：NVLink允许将
- **一文解读：阿里云AI基础设施的演进与挑战 - 知乎专栏** (relevance: 98%) <https://zhuanlan.zhihu.com/p/694443801> 可以看到，像NVIDIA也在Grace Hopper架构上推出了NVlink C2C方案，能够大幅提升整个数据传输的速率。第三个是通讯墙。尤其对于训练来说，分布式训练规模还是

- **中金：AI服务器产业链拆解 - 华尔街见闻** (relevance: 97%) <https://wallstreetcn.com/articles/3685834> 中金彭虎等 2023-04-06 07:56. 1.AI云端算力市场规模的测算：我们预计2023~2025年训练型和推理型AI加速芯片可实现的增量市场规模分别为72亿美元和168亿美元，对应服务器的出货增量分别为7.5万台和17.5万台，对应服务器的市场规模分别为149亿美元和348亿美元。考虑到AI应用的持续推广和活跃用户数的大幅提升，长期来看，我们认为推理型AI加速芯片和推理型服务器仍有望保持高增长。 . 2.AI服务器产业链拆解：AI服务器核心组件按价值量由高到低依次为GPU、DRAM、SSD、CPU、网卡、PCB、高速互联芯片和散热模组等，按7.5万台训练型和17.5万台推理型服务器测算...

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

FlashAttention accelerates large language model inference, optimizes memory usage, and improves speed. It uses techniques like block computation and caching. Alibaba's Flash-LLM leverages this for faster model processing.

Sources

- **深度解析！一起扒扒阿里Qwen3背后的技术细节 - 51CTO** (relevance: 73%) <https://www.51cto.com/article/814669.html> vLLM框架：支持FlashAttention-2加速，推理速度提升37%。 • 昇腾适配：与昇腾910B芯片协同，千亿模型推理能耗下降55%。 • 量化工具：INT8量化后0.6B模型可在树莓
- **125_训练加速：FlashAttention集成- 推导注意力优化的独特内存节省** (relevance: 68%) <https://developer.aliyun.com/article/1684066> 内存优化原理：FlashAttention通过分块计算、计算重排和利用高速缓存，将注意力机制的内存复杂度从 $O(n^2)$ 降低到 $O(n\sqrt{M})$ ，其中M是GPU高速缓存大小。 • 数学公式重
- **广告深度学习计算：阿里妈妈大模型服务框架HighService - 智源社区** (relevance: 67%) <https://hub.baai.ac.cn/view/43319> 算子加速：主要使用了FlashAttention算法加速推理过程。 • Batching推理优化：AIGB场景单次推理的计算量，相较于AIGC场景较小，单次推理无法用满GPU的算力资源，
- **Flash-LLM：阿里开源的大模型推理加速库 - CSDN博客** (relevance: 46%) https://blog.csdn.net/gitblog_01020/article/details/145007852 Flash-LLM 是由阿里巴巴集团开源的一个大型语言模型（LLM）推理加速库，旨在通过无结构化模型剪枝技术提高推理效率。该项目主要使用Cuda、Python、C++、C

- **14-PagedAttention、FlashAttention与投机采样：推理优化三大技术** (relevance: 42%) <https://juejin.cn/post/7612562930848841754> PagedAttention、FlashAttention与投机采样：推理优化三大技术大模型推理的三大瓶颈在上一章中，我们学习了KV Cache如何通过缓存已计算的K和V来加速推理

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 advanced AI capabilities and tools for developing intelligent agents. These agents integrate large language models, planning, feedback, and tool use for autonomous task execution. They aim to reduce human intervention in complex workflows.

Sources

- **【AI agent】_AI agent问题与内容精选-阿里云** (relevance: 76%) <https://www.aliyun.com/sswb/519515.html> ## spring AI agent. * 文章 | Spring AI Alibaba Admin 开源！以数据为中心的 Agent 开发平台. * 文章 | Spring AI Alibaba实践 | 后台定时 Agent. * 阿里云文档 | 【新功能/规格】RDS AI应用上线RAG Agent功能. ## agent AI. * 文章 | 智能体（Agent）：AI不再只是聊天，而是能替你干活. * 阿里云文档 | 【新功能/规格】RDS AI应用上线RAG Agent功能. * 文章 | RL 和 Memory 驱动的 Personal Agent，实测 Macaron AI. * 文章...
- **零失误搭建Agent！阿里AgentScope+AutoGPT双框架实战 - CSDN博客** (relevance: 72%) https://blog.csdn.net/m0_59164520/article/details/151762721 AI大模型新应用：阿里巴巴推出AgentScope多智能体开发平台. 在多智能体应用开发的浪潮中，阿里巴巴通义实验室近日开源了一款创新的编程框架与开发平台
- **AutoGPT如何用大模型重构AI workflow LLM Agent智 - 稀土掘金** (relevance: 66%) <https://juejin.cn/post/7595413459200049188> # LLM Agent智能体引爆未来！深度解析 AgentGPT、AutoGPT如何用大模型重构AI workflow. jimeng-2026-01-16-6875-扁平化动画风格，科技海报设计，技术博客封面图，极简主义构图，科技感十足的背景元素....png. jimeng-2026-01-16-6875-扁平化动画风格，科技海报设计，技术博客封面图，极简主义构图，科技感十足的背景元素....png. ## LLM Agent智能体引爆未来！深度解析AgentGPT、AutoGPT如何用大模型重构AI workflow. **摘要：** 本文深入剖析LLM Agent智能体技术革命，聚焦AgentGPT与AutoGP...

- 科技巨头狂卷“智能体”，大模型上终于长出了“大家伙”？ - 36氪 (relevance: 64%)
<https://m.36kr.com/p/2928513942232455> # 科技巨头狂卷“智能体”，大模型上终于长出了“大家伙”？. 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 ...
- 智能体应用- 大模型服务平台百炼- 阿里云 - Alibaba Cloud (relevance: 54%) <https://www.alibabacloud.com/help/zh/model-studio/single-agent-application> # 大模型服务平台百炼：智能体应用. 大语言模型（Large Language Model, LLM）无法直接访问专有知识库或获取实时动态信息。针对这一瓶颈，阿里云百炼提供了智能体（Agent）应用。智能体支持以零代码方式，将大模型与外部工具进行集成，从而扩展模型的能力边界。控制台访问限制：仅面向**2025年4月21日**前创建过阿里云百炼应用的**国际版用户**开放**应用开发**页签（如下图）的访问权限。API调用限制：仅面向**2025年4月21日**前创建过阿里云百炼应用的**国际版用户**开放**应用数据、知识库及Prompt工程**功能的接口的调用权限。## ...

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，阿里巴巴的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

预测阿里巴巴在未来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 08:25:16 AM CST 2026