

MiniMax AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:27:11

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: MiniMax

主要产品: 海螺AI, 星野

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

MiniMax AI model M3 is set to be released in the first half of 2026, promising multi-modal capabilities and top-tier global performance. MiniMax has already released M2.5, which achieved global first in API token usage. MiniMax continues to lead in AI model innovation and market impact.

Sources

- 模型发布- MiniMax 开放平台文档中心 (relevance: 87%) <https://platform.minimaxi.com/docs/release-notes/models> ## Music-2.5+. ## MiniMax M2.5. 全新文本模型 MiniMax-M2.5 系列模型 MiniMax-M2.5 / M2.5-highspeed 正式发布，在编程、工具调用和搜索、办公等生产力场景都达到或刷新了行业的 SOTA. ## 2026 年 1 月 16 日. ## Music-2.5. ## 2025 年 12 月 22 日. ## MiniMax M2.1. 全新文本模型 MiniMax-M2.1 系列模型 MiniMax-M2.1 / M2.1-highspeed 正式发布，多语言编程专家，全面升级复杂编程体验. ## 2025 年 10 ...

- **上海AI独角兽MiniMax发布全模态“全家桶” - 新浪财经** (relevance: 85%) <https://finance.sina.com.cn/roll/2025-11-01/doc-infvwfmt7903621.shtml> # 上海AI独角兽MiniMax发布全模态“全家桶”：中国AI从跟跑到领跑的技术突围. 在人工智能技术日新月异的当下，上海AI独角兽MiniMax稀宇极智以其独特的技术路径和商业策略，在10月的最后一周掀起了一场AI技术的“全家桶”发布热潮。继开源文本大模型M2震动全球AI圈，接连发布视频模型Hailuo 2.3、语音模型Speech 2.6和音乐模型Music 2.0，标志着中国AI企业在全模态技术领域的全面突破。10月27日，新一代文本大模型MiniMax-M2正式发布和开源，这款仅有10B激活参数（总参230B）的轻量级模型在全球权威测评榜单Artificial Analysis (...)
- **上市后首次参展MiniMax携数十款AI智能硬件生态产品亮相2026AWE** (relevance: 84%) <https://finance.eastmoney.com/a/202603133671298883.html> 指数 期指 期权 个股 板块 排行 新股 基金 港股 美股 期货 外汇 黄金 自选股 自选基金. 资金流向 主力排名 板块资金 个股研报 新股申购 转债申购 北交所申购 AH股比价 年报大全 融资融券 龙虎榜 限售解禁 IPO 审核 大宗交易 估值分析. 以“AI科技慧享未来”为主题的2026年中国家电及消费电子博览会（AWE）3月12日至3月15日在上海举行。通用人工智能（AGI）科技公司MiniMax携自主研发的全模态大模型矩阵及数十款AI智能硬件生态产品参展。此次亮相，是MiniMax上市以来，首次向公众集中展现其多模态大模型等技术成果。最近OpenClaw引发的“养虾热”也蔓延到AW...
- **国产新一代大模型MiniMax 3上半年发布：多模态、全球顶级性能 - 新闻** (relevance: 82%) <http://news.17173.com/content/03032026/220044516.shtml> # 国产新一代大模型MiniMax 3上半年发布：多模态、全球顶级性能. 国产大模型MiniMax M3上半年发布！多模态、全球顶级性能，大摩看好。M2.5调用量全球第一，国产AI超越美国公司。点击了解详情！. 在AI大模型上国内厂商是一浪追着一浪高，DeepSeek V4下周就要发布，MiniMax也会在上半年发布3.0版。MiniMax系列大模型是稀宇科技推出的AI模型，2月13日发布的MiniMax M2.5，称该模型为全球首个为智能体场景原生设计的生产级旗舰模型。发布之后，只用了七天时间其调用量即突破3.07T tokens，凭借在编程和Agent工作流中的卓越性能与极低的成本...
- **全球开源大模型杭州霸榜被终结，上海Minimax M2发布即爆单 - 量子位** (relevance: 78%) <https://www.qbitai.com/2025/10/346476.html> # 全球开源大模型杭州霸榜被终结，上海Minimax M2发布即爆单，百万Tokens仅需8元人民币. 克雷西 2025-10-28 09:52:11 来源：量子位. ##### 克雷西 发自 凹非寺 量子位 | 公众号 QbitAI. 在第三方评测机构Artificial Analysis的测试中，Minimax M2以61分获得了开源模型第一，紧随Claude 4.5 Sonnet。而且经济高效，推理速度是Claude 3.5 Sonnet的两倍，API价格却只有8%。Minimax表示，智能水平、速度和成本在过去被视为“不可能三角”，但随着M2的出世，这个三角被打破了。....

2.2 技术突破

Answer

MiniMax achieved significant technological breakthroughs in text, video, and voice models, establishing itself as a leader in AI innovation. The company's M2.1 model set new standards in multi-language programming. MiniMax's self-driven development approach has positioned it as a key player in global AI advancements.

Sources

- **聆訊過關！MiniMax憑什麼進入全球「第一梯隊」？ - Yahoo 財經** (relevance: 100%)
<https://hk.finance.yahoo.com/news/%E8%81%86%E8%A8%A%E9%81%8E%E9%97%9C-minimax%E6%86%91%E4%BB%80%E9%BA%BC%E9%80%B2%E5%85%A5%E5%85%A8%E7%90%E7%AC%AC-%E6%A2%AF%E9%9A%8A-064001691.html> 正是基於此技術架構，MiniMax才能實現技術迭代密集，關鍵技術突破頻率高，接連在文本、視頻與語音等模型取得突破性進展，推動其持續領先行業。筆者
 - **MiniMax发布M2.1模型：多语言编程SOTA背后的技术突破与商业逻辑** (relevance: 100%)
<https://aistudio.baidu.com/blog/detail/755304882763973> 2025年12月，MiniMax在港交所聆讯后迅速发布新一代旗舰级Coding & Agent模型M2.1，以仅10B激活参数在Multi-SWE-bench榜单中取得49.4%的成绩，超越Claude.
 - **MiniMax技术发布周展示全球首个开源混合架构模型等五大突破** (relevance: 100%)
<https://cj.sina.cn/articles/view/1735950160/67787f50040014bju?froms=ggmp&vt=4> 有分析人士认为，MiniMax的创新之路为全球AI发展提供了第二条道路。一方面，面对外部的算力限制和技术封锁，MiniMax没有选择跟随和模仿，而是坚定地走了“自主
 - **AI应用加速落地！MiniMax、云知声连日上涨商业化能力仍是考验** (relevance: 100%)
<https://www.cls.cn/detail/2274985> 艾媒咨询CEO兼首席分析师张毅表示，目前市场的上涨逻辑是行业前景、技术突破、业绩兑现等多方共振的结果。“港股18C规则对于AI企业非常友好，也有利于国际资本和科技巨头纷纷
 - **三年长成一只独角兽，MiniMax速度惊人 - 界面新闻** (relevance: 100%) <https://www.jiemian.com/article/12932715.html> 闫俊杰一直将MiniMax定位为一家“技术驱动的公司”。他曾以海螺AI对比同类型产品，表示有一些简单的功能问题不是不能解决，而是一旦去解决，精力就会
-

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析MiniMax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

MiniMax M2.5 is a leading large language model known for its cost-effectiveness and suitability for complex tasks like agent workflows. It outperforms competitors like Gemini and GPT in terms of efficiency and application in programming and tool use. MiniMax's success is attributed to its architecture and focus on practical, long-term use.

Sources

- **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> 除了国外御三家外：Gemini、GPT、Claude；国产大模型超过75%的也越来越多，最高分的MiniMax-M2.5也超过了80%。OpenAI在2月23号有发布一个文章说SWE
- **霸榜全球大模型，力压Claude、GPT，MiniMax凭什么？ - 36氪** (relevance: 100%) <https://m.36kr.com/p/3720864696070913> 它压过的，不是一些普通模型，而是Gemini、DeepSeek、Claude等旗舰模型。来源：OpenRouter LLM Leaderboard

(2026.2.12-2026.3.12) . ## **01 MiniMax为何全球第一?** . MiniMax M2.5: 输入0.27美元/百万Token, 输出0.95美元/百万Token。 . Claude Opus 4.6: 输入5美元/百万Token (约18倍)、输出25美元/百万Token (约26倍) 。 . 一位海外开发者在X上算了一笔账: “跑OpenClaw连续1小时 (100 TPS) , MiniMax M2.5只要1美元; 换Claude一天...

- **大语言模型-逻辑能力横评25-11月榜(Gemini 3/GPT-5.1/Opus 4.5) - 知乎** (relevance: 99%) <https://zhuanlan.zhihu.com/p/1977143847453755265> 大语言模型-逻辑能力横评25-11月榜(Gemini 3/GPT-5.1/Opus 4.5) • 本评测是个人性质, 结合自己需求和对大模型的理解, 使用滚动更新的私有题库进行长期跟踪
- **大模型选择困难症? 8款主流AI助手(GPT/Claude/GLM等)特点与适用 ...** (relevance: 99%) https://blog.csdn.net/m0_65555479/article/details/157100830 # 大模型选择困难症? 8款主流AI助手(GPT/Claude/GLM等)特点与适用场景详解, 建议收藏. 文章对比8种主流大语言模型(GPT、Claude、Gemini、GLM、Minimax、DeepSeek、Qwen和Kimi)的特点和适用场景。GPT系列全能型; Claude擅长写作; Gemini擅长资料整合; GLM中文自然; Minimax创意丰富; DeepSeek代码逻辑强; Qwen实用多场景; Kimi长文本读取能力强。为不同需求用户提供选择指南。 . ## 这两年, 大模型更新得太快了。 . GPT、Claude、Gemini、GLM、Minimax、DeepSeek、Qwen、Kimi.....
- **GLM-5 vs. MiniMax M2.5 vs. Gemini 3 深入思考: 哪個模型適合您的 ...** (relevance: 97%) <https://milvus.io/zh-hant/blog/glm5-vs-minimax-m25-vs-gemini-3-deep-think.md> 在短短兩天多的時間內, 三個主要機型接連推出: GLM-5、MiniMax M2.5 和 Gemini 3 Deep Think。三者的頭條功能相同: 編碼、深度推理和代理工作流程。

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

The MiniMax AI model o1 is a large language model designed for complex reasoning tasks. It uses a Mixture of Experts (MoE) architecture for efficient performance. It is one of the top-performing reasoning models in 2026.

Sources

- **终极指南- 2026年最佳MiniMaxAI及替代模型 - SiliconFlow** (relevance: 74%) <https://www.siliconflow.com/articles/zh-Hans/the-best-minimaxai-models-in-2025> blue pastel abstract background with subtle geometric shapes. Image height is 600 and width is 1920. # 终极指南 - 2026年最佳MiniMaxAI及替代模型. ## Elizabeth C. 我们关于2026年最佳MiniMaxAI及替代推理模型的综合指南。我们与行业专家合作, 在关键推理基准

上测试了性能，并分析了MoE架构，以揭示用于复杂推理任务的最强大AI模型。从混合注意力系统到强化学习驱动模型，这些尖端解决方案在数学推理、代码生成和长上下文理解方面表现出色——帮助开发人员和企业通过S...

- **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: 68%) <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
- **国产六大推理模型激战OpenAI? - 36氪** (relevance: 67%) <https://m.36kr.com/p/3264120214847237> # 国产六大推理模型激战OpenAI? . 离年夜饭仅剩几个小时，国内某家云服务器的工程师突然被拉入工作群，接到紧急任务，要求其快速调优芯片，以适配最新的DeepSeek-R1模型。该工程师告诉我们，“从接入到完成，整个过程不到一周”。 . 大年初二，一家从事Agent To B业务的厂商负责人电话被打爆，客户的要求简单粗暴：第一时间验证模型真实性能，尽快把部署提上日程。 . 节前大模型，节后只有DeepSeek。DeepSeek-R1就像一道分水岭，重新书写了中国大模型的叙事逻辑。 . 以2022年11月，OpenAI发布基于GPT-3.5的ChatGPT应用为起点，国内自此走上了追赶OpenAI的...
- **从o1到DeepSeek-R1，万字长文带您揭秘推理模型——及其与标准 ...** (relevance: 56%) <https://zhuanlan.zhihu.com/p/26076930125> 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的生成方式与任何其他文本序列无异。然而，这些推理轨迹表现出非常有趣
- **OpenAI 突然公开o3-mini 思维链！首秀遭质疑，实测对比DeepSeek R1** (relevance: 53%) <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，让他们意识到这将会是一件大事。 . 「o3-mini 是首个能够持续准确解答井字棋问题...

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 leading AI video generation model with advanced multi-modal capabilities. It significantly improves video content usability and control. It has been used in high-profile applications like the Spring Festival Gala.

Sources

- **多模态大模型真能生成高质量视频吗？55%用户认为技术 ...** (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。
- **地表最强？AI视频乱真，甄别不难！** (relevance: 100%) <https://zhuanlan.zhihu.com/p/2005562602990874651> 2月12日，字节跳动Seed官方微信公众号正式宣布发布Seedance 2.0，其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，集成目前
- **字节Seedance2.0实测：多模态封神，AI视频创作彻底告别“抽 ...** (relevance: 99%) <https://cloud.tencent.com/developer/article/2633897> ## 字节Seedance2.0实测：多模态封神，AI视频创作彻底告别“抽卡式”生成. # 字节Seedance2.0实测：多模态封神，AI视频创作彻底告别“抽卡式”生成. 就在2月12日，字节跳动正式发布了新一代视频生成模型Seedance2.0，一经上线就引爆了AI创作圈——马斯克转发相关动态感叹“发展太快”，国内创作者更是连夜实测，直言它彻底打破了AI视频“好看但不好用”的困境。. 作为长期关注AI多模态技术的博主，我第一时间上手体验了已接入豆包、即梦产品，以及火山方舟体验中心的Seedance2.0（据悉2月中下旬还会上线API服务，企业用户可重点关注），今天就从核心技术亮点、工业级应...
- **Seedance 2.0春晚首秀：解锁AI视频大模型的现在与未来** (relevance: 99%) https://www.sohu.com/a/989548490_121955537 # Seedance 2.0春晚首秀：解锁AI视频大模型的现在与未来. 当Seedance 2.0登上春晚舞台，为《贺花神》《驭风歌》等节目打造惊艳视觉效果，这不仅是一次AI技术的公开亮相，更是国产AI视频大模型技术成熟度与产业化能力的全面验证。作为首款落地春晚级超高清直播场景的AI视频大模型，其首秀实现了技术层面的多重突破，重构了行业竞争的底层逻辑，也为未来1-3年AI视频大模型的发展勾勒出清晰演进方向，让市场看到了AI视频技术从实验室走向规模化产业应用的全新可能。. 春晚的高关注度、高复杂度与高稳定性要求，成为Seedance 2.0规模化应用的最佳实战检验场。在《贺花神》《驭风歌》等节目...
- **Seedance 2.0，凭什么刷屏？-腾讯新闻** (relevance: 98%) <https://news.qq.com/rain/a/20260212A044KJ00> # Seedance 2.0，凭什么刷屏？. 2026-02-12 14:16发布于河北钛媒体APP官方账号. 文 | AIX财经，作者 | 雷晶，编辑 | 金玙璠. 2月7日，字节跳动旗下AI视频生成大模型Seedance 2.0开启内测后，迅速在全球科技圈和资本市场引发关注。. 原因在于，与过往“让图片动起来”的初级生成模式不同，Seedance 2.0的运镜规划、跨镜头的角色一致性以及原生的音画同步能力，一定程度上解决了AI视频长期存在的可用率低、成本高的核心痛点。. 资本市场的反应很快。2月10日，A股传媒板块应声大涨，多达25只个股涨停。这背后是资本市场对AI视频赛道价值和潜力的重估...

4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

Answer

H100, B200, and TPUs are advanced GPU and accelerator technologies for AI. B200 offers superior memory bandwidth and efficiency. TPUs provide high-performance deep learning acceleration.

Sources

- **万字长文解析：从H100到B200，GPGPU与大模型扩展性深度分析** (relevance: 100%)
<https://zhuanlan.zhihu.com/p/1985478405458788975> GPU 算力增速远超带宽提升,数据并行的临界Batch Size 从H100 的2500 tokens/GPU 激增到B200 的5625 tokens/GPU; 应对策略包括FP8/FP4 量化、MoE 稀疏
- **2026年中国智能算力规模将翻倍，推理算力占比首超50%_CPU** (relevance: 100%)
<https://post.smzdm.com/p/azz2860p> Google TPU v6e、AMD MI300X 与NVIDIA H100/B200的推理成本大比拼：根据Artificial Analysis最新硬件基准测试，NVIDIA在「每百万输入输出token成本」指标上，较TPU
- **【深度万文】10万卡H100集群的尽头……** (relevance: 100%) <https://www.eeworld.com.cn/emp/XSY/a382186.aspx> 与2万卡A100集群相比，AI训练算力可提升31.5倍。资料来源... 算力能力的GPU。（译者注：智算中心在建设时，一般会配置一定比例的备用
- **11.4 硬件选型：GPU、TPU与专用加速器|大模型原理与架构 - GitBook** (relevance: 100%) https://yeasy.gitbook.io/llm_internals/di-san-bu-fen-tui-li-yu-bu-shu-pian/11_serving/11.4_hardware ## hashtag 11.4.1 NVIDIA GPU：当前的主流选择. NVIDIA GPU 凭借 CUDA 生态的成熟度和 Tensor Core 的强大性能，是 LLM 训练和推理的首选。关键指标对比：. 表 11-2：NVIDIA 主要 GPU 型号对比（*为 FP8 Tensor Core 算力，FP16 密集算力约为一半）。对于推理场景，**显存带宽**通常比算力更重要（10.1 节已解释生成阶段是访存密集型的）。H200 的 141 GB 显存使其可以在单卡上运行 70B 的 INT4 量化模型，极大降低了部署复杂度。而 B200 的 192 GB 显存和 8 TB/s 带...
- **一张图说清：H100、H200、B200到底该怎么选？ - 稀土掘金** (relevance: 99%) <https://juejin.cn/post/7586892413890068526> | NVLink | 第四代（900 GB/s）| 第四代 | 第五代（1.8 TB/s）|. * 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，调试工具完善 |. | **30B-70B，推理或 LoRA/QLoRA...

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

MiniMax 数据存储使用HBM技术，提供高带宽和低延迟，显存和NVLink结合提升GPU计算效率。

Sources

- **AI芯片与数据中心关键技术 - 知乎专栏** (relevance: 21%) <https://zhuanlan.zhihu.com/p/3914664135> 模型训练和推理过程中，显存不仅决定了模型的大小和数据处理能力，还直接影响GPU的计算效率。HBM主要是由DRAM进行多层封装实现的，因此正好满足了GPU对计算时高带宽访存的
- **新浪GPU热点小时报 | 2026年03月07日14时** (relevance: 9%) <https://cj.sina.com.cn/articles/view/7857201856/1d45362c001902ygb8?from=ggmp> 据介绍，南亚科技开发的AI内存采用UltraWIO（IT之家注：Ultra Wide I/O）架构，能够与客户的AI运算引擎紧密结合，概念上类似HBM与GPU之间的运作模式，通过

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

MiniMax 数据加速通过FlashAttention和量化技术提升推理效率，前者直接加速注意力计算，后者降低内存和计算开销。FlashAttention通过分块计算显著减少内存复杂度。量化技术在保持精度的情况下进一步优化推理速度。

Sources

- **MiniMax-M1混合架构与vLLM的结合：长上下文支持与高效推理- 知乎** (relevance: 100%) <https://zhuanlan.zhihu.com/p/1924246503872008841> 量化技术在精度损失最小化的情况下降低了内存和计算开销，而FlashAttention则直接加速了注意力计算，使实际应用中的推理速度显著提升。vLLM中的Lightning

- **125_训练加速: FlashAttention集成- 推导注意力优化的独特内存节省** (relevance: 100%) <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
- **Minimax Link伙伴日学习笔记: LLM与多模态 - 飞书文档** (relevance: 100%) <https://docs.feishu.cn/v/wiki/QVQRwOii8iHOwwkrVlPclmKMn0b/aj> 在实际操作中, 优化MFU的关键在于针对这三项指标进行监督和改进。尤其是对于不成熟的硬件系统或新模型结构, Matmul和FlashAttention算子在特殊情况下可能表现不佳, 非融合算子
- **2025年LLM核心架构优化盘点 - 知乎专栏** (relevance: 25%) <https://zhuanlan.zhihu.com/p/1980660967881851337> 论文提出了Tensor Product Attention (TPA) 机制, 通过将注意力机制中的Q、K、V 进行动态的上下文低秩分解, 有效地解决了LLM 推理时的KV Cache 内存瓶颈。
- **从DeepSeek-V3到MiniMax-M2, 现代大模型架构设计全解析_人工智能** (relevance: 21%) <https://adg.csdn.net/694d04e95b9f5f31781ad1cd.html> 总而言之, DeepSeek-V3 是一个拥有 6710 亿参数的巨型模型, 发布时性能超越了包括 4050 亿参数 Llama 3 在内的其他开源模型。尽管规模更大, 但得益于其 MoE 架构 (每次推理仅激活约 370 亿参数), 它在推理时反而更加高效。另一个关键的区别特征是 DeepSeek-V3 使用了多头潜在注意力 (MLA) 而非分组查询注意力 (GQA)。MLA 和 GQA 都是标准多头注意力 (M. DeepSeek-V3 MoE 设计的一个显著特点是使用了**共享专家** (shared expert) ——一个对每个 token 始终激活的专家。这一想法并非全新, 早在 DeepSeek 202...

4.Agent

检索关键词: 智能体,Agent,AutoGPT

Answer

MiniMax Agent is an advanced AI tool with strong coding and intelligent agent capabilities. It supports full-stack development and has a high performance-to-cost ratio. It integrates multiple modalities for comprehensive AI applications.

Sources

- **一个悄然崛起的Agent, 正火爆全球! 原创 - CSDN博客** (relevance: 100%) <https://blog.csdn.net/c406495762/article/details/148767213> MiniMax Agent: AI智能体开发新时代摘要: MiniMax Agent是一款革命性的 ... **AutoGPT**: 一个开源项目, 能够通过API创建完整的项目, 自主完成

- **2025 AI Agent迷局：谁在玩真的，谁在演戏？ - 创业邦** (relevance: 100%) <https://m.cyzone.cn/article/786888> 这个实验让人们第一次对AI Agent（智能体）产生了期待——具有自主意识和决策能力的 ... 举例：AutoGPT通过CoT技术分解复杂问题，动态选择最优解决路径。面对一个
- **讲真，这次真的碾压了OpenAI！ | MiniMax Agent 发布30天后** (relevance: 100%) <https://m.aitntnews.com/newDetail.html?newId=16474> # 讲真，这次真的碾压了OpenAI！ | MiniMax Agent 发布30天后，迎来首个重大更新. 讲真，这次真的碾压了OpenAI！ | MiniMax Agent 发布30天后，迎来首个重大更新. 进入 2025 这个被誉为「Agent 元年」的下半场，我们见证着一众 AI Agent 的智能水平正朝着同一个新目标快速发展：. 就在上周，有媒体报道 AI 独角兽 MiniMax 完成了新一轮融资，公司估值突破 40 亿美元关卡。这家成立三年多的「年轻」公司再一次获得资本青睐。. 而就在上个月的 MiniMax Week，他们做了一系列开源动作：从超长上下文的 MiniMax-M1 推理模型，...
- **AI Agent行业深度：框架拆解、应用方向** (relevance: 100%) <https://zhuanlan.zhihu.com/p/676245844> 发展历程：AI Agent经历了符号智能体、反映型智能体、基于强化学习的智能 ... AGENT开发热潮的AutoGPT。实操性应用更加强调与实际场景的适配
- **M2模型杀回Coding和Agent领域，MiniMax想要「普惠智能」 - 腾讯云** (relevance: 100%) <https://cloud.tencent.com/developer/article/2594019> 最近，MiniMax 发布并开源全新的 M2 模型，正是这一方向的典型实践：不仅在权威测评中跻身全球第一梯队，更以极致性能与性价比的双重突破，再次印证了其在大模型下半场竞争中的领先身位。. 真实开发者的使用数据最具说服力，因为 M2 模型开源后官方 API 和 Agent 限时免费，在全球极具影响力的 AI 模型聚合与调用平台 OpenRouters 上数据显示: 仅仅开源后的第一天，MiniMax-M2 的模型调用量就冲到了全球前十。. 1 Artificial Analysis 榜单全球前五，MiniMax-M2 凭什么. 相比此前的 M1，MiniMax-M2 定位轻量级模型，但综合性能...

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，MiniMax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MiniMax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 08:27:31 AM CST 2026