

智谱AI AI技术洞察报告

报告日期: 2026年03月17日

生成时间: 08:25:40

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: 智谱AI

主要产品: GLM, ChatGLM

检索优先级: 高

二、最新动态检索

2.1 产品/模型发布

Answer

Zhipu has released its new AI model, GLM-5, which significantly enhances programming and intelligent agent capabilities. The model's parameter size has doubled, and it has achieved top performance in open-source benchmarks. GLM-5 is now available on the chat.z.ai platform.

Sources

- [illegible]

的Claude Opus系列進行直接對標測試。智譜表示，GLM-5的參數量較上一代增加一倍以上，並將於周四正式推出。本月以來，市場對可能威脅既有商業模式的新人工智能的發布表現出更高的敏感度，受衝擊的領域涵蓋法律和合規軟件...

- **神秘模型确认！智谱发布新一代旗舰模型GLM-5 - 证券时报** (relevance: 100%) <https://www.stcn.com/article/detail/3640020.html> 智谱向证券时报记者确认，此前在全球模型服务平台OpenRouter登顶热度榜首的神秘模型“Pony Alpha”，为智谱新模型GLM-5。目前新模型已在chat.z.ai平台上线。2月6日，全球模型
- **智谱发布旗舰大模型GLM-5 引发商业化与市场涨价 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/2005560926636635588> 智谱发布旗舰大模型GLM-5，并带动价格上调与股价大涨智谱AI正式开源发布最新旗舰模型GLM-5，在编程与智能体（Agent）任务上取得开源领域领先表现，并对部分付费
- **智谱发布新一代旗舰模型GLM-5，重点提升编程与智能体能力** (relevance: 100%) <https://wallstreetcn.com/articles/3765532> Image 2: article.author.display_name李佳 02-11 17:05. 2月11日，智谱推出新一代旗舰模型GLM-5，参数规模扩展至744B，预训练数据达28.5T，集成DeepSeek稀疏注意力机制。内部评估显示，其编程任务性能较上代提升超20%，真实体验逼近Claude Opus 4.5；在BrowseComp等三项Agent评测中均获开源第一，异步强化学习为核心突破。2月11日，智谱正式推出新一代旗舰模型GLM-5，主攻编程与智能体能力，官方称已实现开源领域最优表现。这是继DeepSeek后，国产AI大模型春节档的又一重要发布。GLM-5参数规模由...
- **GLM Coding Plan - ZHIPU AI OPEN PLATFORM** (relevance: 100%) https://bigmodel.cn/special_area 智谱最新视觉推理模型，视觉理解精度达同规模SOTA，全面支持工具调用，支持128K 超长上下文，并针对Coding 场景进行了专项优化。规格1000万 tokens. 有效期3个月. 模型优势

2.2 技术突破

Answer

Zhipu AI has made significant breakthroughs in large model technology, including the development of GLM-130B and GLM-4. The company aims to achieve self-reliance in AI technology and contribute to the global AI landscape.

Sources

- **智谱AI张帆：大模型时代构建企业竞争力的四个维度 - 中国企业家网** (relevance: 100%) <http://m.iceo.com.cn/article/33bfc906-9d09-4888-8a4d-61cfd7253958> 到了2013～

2014年，我们看到一个变化，就是神经网络的突破，这一突破一下子把算法层统一了，这就是AI1.0时代。大家都用一个算法，某种程度上带来了AI的第一次普

- **智谱AI：源自清华、创新成就200亿估值的AI领航者** (relevance: 100%) <https://zhuanlan.zhihu.com/p/710805313> 训练出的百亿参数模型GLM-10B和1300亿参数的超大规模模型GLM-130B，不仅在技术上取得了突破，更在国际上赢得了认可。特别是开源模型ChatGLM-6B到最新的GLM-
- **智谱GLM-5技术突破：从代码生成到系统构建 - 搜狐** (relevance: 100%) https://m.sohu.com/a/989104848_362225?scm=10001.325_13-325_13.0.0-0-0-0.5_1334 ## 智谱GLM-5技术突破：从代码生成到系统构建，开启国产算力适配新篇章. 智谱最新发布的GLM-5大模型引发行业广泛关注，其技术报告揭示了模型研发思路的重大转变——从单纯追求参数规模转向系统性工程建设。这一转变标志着中国人工智能发展进入新阶段，开始构建自主技术体系而非单纯追赶国际水平。模型能力实现质的飞跃是GLM-5最显著的突破。该模型突破传统任务边界，不仅能完成复杂软件工程任务，更具备跨文件操作、长周期规划、多轮交互等系统级能力。在Vending-Bench 2测试中，GLM-5通过模拟自动售货机全年运营的挑战，展现出接近国际顶尖水平的长期决策能力，这在开源模型中尚属首次。技...
- **挑戰OpenAI！中國AI六小虎「智譜」將發布史上最大開源模型 - 鉅亨網** (relevance: 99%) <https://news.cnyes.com/news/id/6079865> # 挑戰OpenAI！中國AI六小虎「智譜」將發布史上最大開源模型. 鉅亨網編輯林羿君2025-07-28 22:27. 中國「AI 六小虎」之一的智譜 (Zhipu, 近期英文改名為 Z.ai) 將釋出至今為止規模最大的開源模型，加入越來越多提供免費人工智慧產品的中國公司行列。cover image of news article. 挑戰OpenAI！中國AI六小虎「智譜」將發布史上最大開源模型。(圖:shutterstock). cover image of news article. 彭博引述消息人士報導，這家新創公司最快將於週一 (28 日) 旗艦模型的更新版本 GLM-4.，並挑戰...
- **智谱AI：国产全自研大模型商业化落地新解法 - 腾讯云** (relevance: 99%) <https://cloud.tencent.com/developer/article/2399737> ## 智谱AI：国产全自研大模型商业化落地新解法. # 智谱AI：国产全自研大模型商业化落地新解法. 放眼当下的科技浪潮，AI大模型无疑是一年多来持续引人注目的焦点。基于大模型的算法推理，30秒即可生成完美可用的Word、PPT文档，工作时长不必再以小时计算；真人对话一般输入简短文字，就能在30秒内得到想要的图画、代码、文本、视频……大模型让科幻小说中的场景，走进了千行百业、千家万户的现实生活。3月14日，智谱AI举办了“智领，万象新生”智谱AI媒体小型沟通会。智谱AI是国内最早入局大模型赛道的公司之一，立足当下，在过去几年中，智谱AI大模型从落地发芽到茁壮成长，展现了新的科技是如何与日...

三、技术趋势分析

3.1 模型能力演进

基于检索结果分析智谱AI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. Major language models include LLMs from Amazon, Google, and others. These models are used for various AI applications.

Sources

- **中國智譜AI最強開放模型GLM-5 是Claude 貼皮？ 一個號稱超越Google** (relevance: 100%) <https://www.facebook.com/groups/gaitech/posts/1607796527071223/> 中國智譜AI最強開放模型GLM-5 是Claude 貼皮？ 一個號稱超越Google Gemini、甚至在部分指標上逼近GPT-5.2 的頂級開源模型，在自我介紹時，竟然說出：「你
- **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

- **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...
- **GLM-4.5：推理、编码和代理能力 - bigmodel.ai** (relevance: 100%) <https://bigmodel.cn/technology-report> 智谱大模型开放平台-新一代国产自主通用AI大模型开放平台，是国内大模型排名前列的大模型网站，研发了多款LLM模型，多模态视觉模型产品，致力于将AI产品
- **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: 99%) <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——

4.推理模型

检索关键词: o1,R1,推理,思维链

Answer

DeepSeek-R1 and other Chinese models are challenging the dominance of OpenAI's o1 model in AI research. These models focus on structured reasoning and have achieved comparable performance at lower costs. China's push for self-reliance in AI technology is evident.

Sources

- **国产AI卷翻硅谷，奥特曼发文“阴阳”，类o1模型都在卷什么？** (relevance: 71%) <https://zhuanlan.zhihu.com/p/19824257286> ... R1 ... 链思维推理方面取得了SOTA成绩。性能最强的long-CoT版本Kimi k1.5，数学、代码、多模态推理能力可以达到长思考SOTA模型OpenAI o1正式版的水平。
- **国产六大推理模型激战OpenAI？ - 36氪** (relevance: 68%) <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...

- **国产AI 最卷一夜！大模型黑马DeepSeek、Kimi 硬刚OpenAI o1 - 爱范儿** (relevance: 68%) <https://www.ifanr.com/1612733> 前脚 DeepSeek-R1 正式发布，号称性能对标 OpenAI o1 正式版，后脚 k1.5 新模型也正式登场，表示性能做到满血版多模态 o1 水平。·如果再加上此前强势登场的智谱 GLM-Zero，阶跃星辰推理模型 Step R-mini，星火深度推理模型 X1，年末上大分的国产大模型拉开了真刀真枪的帷幕，也给以 OpenAI 为代表的海外模型狠狠上了一波压力。·* DeepSeek-R1：在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版. 昨晚率先发布的 DeepSeek-R1 现在已经上架 DeepSeek 官网与 App，打开就能用。·面对弱智吧难题...
- **国产六大推理模型激战OpenAI？ - 投中网** (relevance: 66%) <https://www.chinaventure.com.cn/news/78-20250425-386044.html> # 国产六大推理模型激战 OpenAI？. ## 2025年的AI叙事主题呼之欲出：“六大推理模型迎战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...
- **ChatGPT两周年，国产o1大模型们紧追不舍 - 新浪财经** (relevance: 64%) <https://finance.sina.com.cn/wm/2024-12-08/doc-incytuqv3719453.shtml> 和OpenAI的o1相比，R1毫无保留地放出了大模型思考的完整过程。官方表示，R1的思维链长度可达数万字。从官方测试结果来看，在AIME（美国数学竞赛）、部分编程

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 is a leading AI video generation model, supporting multi-modal inputs for high-quality video creation. It integrates visual, textual, and audio elements for professional-grade outputs. Seedance 2.0 has gained significant attention for its advanced capabilities in the AI video generation field.

Sources

- **AI应用层迎技术变革，智谱一日暴涨36.22%，领跑港股热潮 - 时代在线** (relevance: 100%)
<https://www.time-weekly.com/post/327430> 字节跳动旗下即梦平台日前宣布，正内测一款AI视频生成模型——Seedance 2.0。该模型支持同时上传图片、视频、音频等多达12类参考文件，凭借多模态参考
- **Seedance：字节跳动的AI视频生成技术突破与行业变革 - 新浪财经** (relevance: 100%)
<https://finance.sina.com.cn/stock/roll/2026-02-13/doc-inhmrnzk2295391.shtml>
Seedance是字节跳动旗下Seed团队研发的多模态AI视频生成平台，是当前全球领先的AI视频生成技术之一。其核心定位是通过扩散模型（Diffusion Model
- **Seedance 2.0 完整指南：字节跳动革命性的AI 视频生成器(2026)** (relevance: 100%)
<https://www.nxcode.io/zh/resources/news/seedance-2-0-complete-guide-ai-video-generation-2026> 掌握Seedance 2.0，字节跳动下一代具备原生音频同步、多镜头叙事和2K 电影级输出的AI 视频生成器。包含API 集成、定价以及与Sora、Runway 和Kling 对比
- **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
- **实测字节Seedance 2.0视频生成，多模态创作神器01. 免费体验入口 ...** (relevance: 99%)
<https://umaax.com/shicezijieseedance20shipinshengchengduomotaichuangzuoshenqi0/> 2B 2C只是渠道，品牌才是终局 —— 跨境 | 运营 | 营销 | 资讯 | 学习 | 服务 | 品牌 |. # 实测字节Seedance 2.0视频生成，多模态创作神器01. 发表评论 / 14 2 月, 2026. 整个互联网，甚至全球的创作者圈子，都被字节跳动的“新玩具”——Seedance 2.0给炸了锅了。从2月7日上线即梦，开始小范围内测，2月9日话题直接冲上微博热搜，随后几天，关于Seedance 2.0的讨论、评测、惊艳案例，席卷所有内容平台。网友称：“去年DS今年SD”，“中国Sora来了”，“导演失业，剪辑师下岗” … 你可以把Seedance 2.0想象成一…

4.算力卡

检索关键词: GPU,H100,B200,TPU,算力

Answer

H100 and B200 are high-performance GPUs used for AI and HPC tasks; H100 offers significant computational power and efficiency, while B200 targets large-scale AI model training.

Sources

- **全球五大巨头GPU总量曝光！2025年等效H100或超1240万块 - 网易** (relevance: 100%) <https://www.163.com/dy/article/JIDGTKBG0511ABV6.html> ... H100算力。据称这包括35万块H100，剩余部分很可能是H200，以及少量 ... 在全面了解各家手握多少GPU/TPU算力之后，下一个问题是，这些算力将用在
- **NVIDIA GPU 全面对比：A 系/ H 系/ B 系 - 知乎专栏** (relevance: 100%) <https://zhuanlan.zhihu.com/p/1939012368395903238> 性能代差巨大. A100 → H100 的FP8 性能提升超过3 倍; H100 → B100 再提升约2.5 倍，且显存翻倍至192GB; B200 双芯片直接面向万亿参数模型，是AI 工厂级别的怪兽卡. 2
- **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 | | | . ǎ | DZ | VR | ŷ ||| ² ||. | ǎ | N 豸 | ee5 | | | | | A | PCB. * Ö ǎ ı v eZ ı LED | P . H100 Ĩ ŭ ħ 9 GPU ĩ ğ AI HPC ĥ Ĩ ŭ A100 Tensor Core GPU | | H100 μ | | H100 ! | A100 L 6 h AI HPC ص ǒ ı ı ? . 201...
- **全球AI算力报告出炉，LLM最爱A100！谷歌坐拥超100万H100等效算力** (relevance: 100%) <https://hub.baai.ac.cn/view/43400> # 全球AI算力报告出炉，LLM最爱A100！谷歌坐拥超100万H100等效算力. 新智元 2025-02-16 21:40 分享. ### --- **新智元报道**. #####
【新智元导读】全球有多少AI算力？算力增长速度有多快？在这场AI「淘金热」中，都有哪些新「铲子」？AI初创企业Epoch AI发布了最新全球硬件估算报告。. AI的物质基础是机器学习硬件，例如图形处理单元（GPU）和张量处理单元（TPU）. . 除了传统硬件厂商英伟达、AMD等纷纷推出加速卡，一些新兴势力开始「造芯」，算力持续提升. . 除了GPU，硬件类型也丰富了起来。比如，出现了专门处理张量计算的TPU (...)
- **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> 评估模型在GPU 上运行效率的核心指标是算力强度(Arithmetic Intensity),即每次内存访问所对应的计算量。H100 指标:算力990 TFLOPs (BF16),带宽3.35 TB/s

4.数据存储

检索关键词: HBM,显存,存储,NVLink

Answer

HBM is crucial for AI due to its high bandwidth, enabling faster GPU performance. NVLink enhances GPU-GPU communication, boosting multi-GPU training speeds. AI demand drives growth in DRAM and HBM markets.

Sources

- **HBM，何以成为AI角力关键？_新闻中心** (relevance: 63%) <https://www.amaxchina.com/news/2167.html> # HBM，何以成为AI角力关键？_新闻中心_苏州超集信息科技有限公司。三星电子近日宣布，其12层第六代HBM4内存将于10月底正式发布，现已进入研发冲刺阶段，并计划今年晚些时候量产。这一动作无疑为2025年本就爆发式增长的HBM市场再添一把烈火。为什么HBM年增速能突破200%，达到68亿美元全球市值，成为AI赛道的"战略石油"。今天，超集信息带您透视HBM的底层逻辑：从打破"存储墙"到决定大模型训练速度，它如何悄悄掌控AI算力的生死线。HBM对GPU的性能提升，本质是解决了传统内存（如GDDR6、DDR5）的"带宽瓶颈"——GPU计算核心的算力（如FP8 算力达 1-2 ...
- **国产算力产业走向「闭环」** (relevance: 56%) <https://news.pedaily.cn/202601/559710.shtml> # 国产算力产业走向「闭环」。1月8日，上海天数智芯半导体股份有限公司（下称“天数智芯”，09903.HK）在港股挂牌上市，其此次公开发售获超400倍认购，充分显示了资本市场的热情。此前不久，另外两家国产GPU头部企业也已接连登陆科创板：2025年12月5日，摩尔线程（688795.SH）上市首日股价一度上涨468.78%，总市值突破3055亿元；2025年12月17日，沐曦股份（688802.SH）上市首日涨幅达到692.95%，市值站上3300亿元关口。与GPU厂商的二级市场热潮相呼应，国产存储芯片厂商也在2025年末完成了关键一跃。2025年12月30日，长鑫科技集团股份有限公...
- **算力需求爆发，优刻得“国产智算+GPU切分”破解AI资源荒** (relevance: 53%) <https://zhuanlan.zhihu.com/p/1986742895798157513> Level 0（HBM）：作为高带宽内存，HBM直接集成于GPU封装内部，凭借超高的数据传输带宽，可有效消除GPU的数据等待瓶颈，是AI高算力运算的核心存储支撑。
- **2026内存涨价深度解析：AI驱动下的DRAM/HBM产业逻辑与 ...** (relevance: 52%) https://blog.csdn.net/2401_89353317/article/details/156766843 NVLink vs PCIe：8卡训练速度实测对比. 在多卡训练系统中，NVLink相比PCIe能显著提升训练速度。实测数据显示，NVLink全互联架构可将通信开销从20-30%降
- **高性能GPU服务器硬件拓扑与集群组网** (relevance: 51%) <https://www.eet-china.com/mp/a303588.html> 最下面一排是18根Gen4 NVLink；双向总带宽18 lanes * 25GB/s/lane = 900GB/s；中间蓝色的是L2 cache；左右两侧是HBM芯片，即显存。

4.数据加速

检索关键词: FlashAttention,量化,推理优化

Answer

FlashAttention speeds up attention computation in large language models, reduces memory usage, and uses hardware-aware optimizations. It combines quantization and

mixed-precision techniques for further efficiency. FlashAttention can achieve up to 10x speedup compared to traditional attention mechanisms.

Sources

- **FlashAttention算法详解 - 智源社区** (relevance: 70%) <https://hub.baai.ac.cn/view/29867> # FlashAttention算法详解. 深度学习技术前沿 2023-08-31 15:50 分享. 这篇文章的目的是详细的解释Flash Attention, 为什么要解释FlashAttention呢? 因为FlashAttention 是一种重新排序注意力计算的算法, 它无需任何近似即可加速注意力计算并减少内存占用。所以作为目前LLM的模型加速它是一个非常好的解决方案, 本文介绍经典的V1版本, 最新的V2做了其他优化我们这里暂时不介绍。因为V1版的FlashAttention号称可以提速5-10倍, 所以我们来研究一下它到底是怎么实现的。 . ## 介绍. “FlashAttention:...
- **又快又准, 即插即用! 清华8比特量化Attention, 两倍加速于FlashAttention2, 各端到端任务均不掉点! - 知乎** (relevance: 69%) <https://zhuanlan.zhihu.com/p/2418851068> 对 Q, K 进行分块 INT8 量化。对于矩阵 Q, K, SageAttention 采用了以 FlashAttention 的分块大小为粒度的 INT8 量化。这是因为: 1. 对 Q, K 矩阵进行 INT8 量化相比于进行 FP8 量化, 注意力的精度更高。2. 在一些常用卡上, 比如 RTX4090, INT8 矩阵乘法 (INT32 为累加器) 的速度是 FP8 (FP32 为累加器) 的两倍。 • 对 P, V 采用 FP16 数据类型的矩阵乘法累加器。对于矩阵 P, V, SageAttention 采用了保留 P, V 为 FP16 的类型, 但进行矩阵乘法时采用 FP16 数据类型的累加器。这...
- **混合精度量化: FlashAttention实现模型压缩与加速的核心技术-CSDN博客** (relevance: 66%) https://blog.csdn.net/gitblog_01067/article/details/152772343 October 31, 2025 - 推理阶段: 根据硬件选择最优精度 (H100用FP8, A100用BF16, 消费级GPU用FP16) • 通过训练脚本中的配置文件可以设置不同阶段的精度策略, 例如: • # training/configs/model/gpt2.yaml model: type: GPT params: dtype: bf16 attn_pdrop: 0.0 resid_pdrop: 0.0 use_flash_attn: true flash_attn_kwargs: causal: true softmax_scale: 1.0 • FlashAttention通过精细化的...
- **使用 FlashAttention-2 加速微调与推理_flash attention 2-CSDN博客** (relevance: 66%) <https://blog.csdn.net/xuner1213/article/details/149072083> July 4, 2025 - 文章浏览阅读 1k次, 点赞9次, 收藏9次。FlashAttention 是一种优化 Transformer 中注意力计算的流行方法, 它能显著加速大型语言模型(LLM)的推理和微调过程。目前 FlashAttention 已在多种框架中实现, 包括 Hugging Face Transformers 和
- **揭秘 AI 推理优化: 从模型压缩到 GPU 加速的低延迟架构实战_牛客网-具身智能开发者生态** (relevance: 59%) <https://i-newcar.com/index.php?a=index&aid=4801&c=View&m=home> GPU 内存在优化模型架构...们会将数据移动到速度较慢的系统内存中。 • 优化模型架构的几个好处包括节省时间和金钱。首先, 从密集

Transformer 切换到 LoRA 优化或基于 FlashAttention 的变体，每个查询的响应时间可减少 200 到 400 毫秒，这在聊天机器人和游戏等场景中至关重要。此外，量化模型（如

4.Agent

检索关键词: 智能体,Agent,AutoGPT

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 or name.

Sources

- **AI也能接单挣钱？智谱发布最新AI智能体：一边进行复杂思考 - 时代在线** (relevance: 76%) <https://www.time-weekly.com/post/319972> 首页 要闻 区域 财经 金融 公司 科技 地产 Time News 汽车 商研院 消费观察 大健康. 评论 文娱 动态 天下 数据新闻 创业圈 ESG. # AI也能接单挣钱？智谱发布最新AI智能体：一边进行复杂思考，一边执行操作. 3月31日，智谱在 2025 中关村论坛上发布最新 Agent 产品 AutoGLM 沉思。作为首个集深度研究能力和操作能力于一体的 Agent，AutoGLM 沉思能一边进行复杂思考，一边执行操作。例如能打开并浏览网页，完成从数据检索、分析到生成报告一系列动作。. AI智能体（Agent）是指使用AI技术，能够自主感知环境、作出决策并执行行动的智能实体。...
- **中美AI竞争加剧：OpenAI对手智谱发布智能体应用 - 新浪财经** (relevance: 75%) <https://finance.sina.com.cn/cj/2025-08-20/doc-infmrakh0454766.shtml> # 中美AI竞争加剧：OpenAI对手智谱发布智能体应用，奥尔特曼称美国低估中国AI威胁. 就在刚刚，OpenAI曾点名的中国竞争对手智谱（Z.ai，原Zhipu）发布全新AI智能体应用AutoGLM。8月20日消息，智谱今天发布全球首个手机Agent智能体应用产品AutoGLM 2.0版本，基于GLM-4.5、GLM-4.5V等纯国产模型驱动，具备推理、代码与多模态的全能能力，拥有iOS版、安卓版和网页版等全平台版本，支持Agent+云手机等新技术，突破硬件限制，能在任何设备、任何场景下运行，帮助用户Agent智能体操作。会前媒体沟通会上，智谱CEO张鹏表示，此前发布的GLM 4.5...
- **智谱推出Agentic GLM 系列矩阵，全栈布局AI智能体生态** (relevance: 69%) <http://www.bilibili.com/read/cv41183673/> 今天，智谱在中关村论坛上正式发布「AutoGLM沉思」，这一全新智能体不仅具备深度研究能力（Deep Research），还能实现实际操作（Operator），真正推动AI Agent进入「边想边干」的阶段。「AutoGLM沉思」的技术演进路径包括：GLM-4基座模型 → GLM-Z1推理模型 → GLM-Z1-Rumination沉思模型 → AutoGLM模型。其中核心链路的模型和技术，我们将于4月14日正式开源，以推动行业生态发展。「让机器像人一样思考」，智谱始终专注于AGI的基座模型研发，目前已经探索到L3-Age...

- **智谱发布免费的超级Agent：像Manus一样干活** (relevance: 67%) <https://zhuanlan.zhihu.com/p/1890366621287166337> 智谱将AutoGLM 沉思定位为一个能探究开放式问题，并根据结果执行操作的自主智能体（AI Agent），它能够模拟人类的思维过程，完成从数据检索、分析到生成报告。
- **继Manus撤离中国后智谱发布全球首个手机智能体AutoGLM 2.0** (relevance: 67%) <https://www.stcn.com/article/detail/3202109.html> 国产AI智能体新进展：继Manus撤离中国后 智谱发布全球首个手机智能体AutoGLM 2.0. 来源：证券时报网作者：聂英好 2025-08-20 14:06. 8月20日，智谱正式发布AutoGLM 2.0。该产品由纯国产模型GLM-4.5与GLM-4.5V驱动，具备推理、代码以及多模态处理能力，可在多种设备和场景中运行，现已面向普通用户开放。·值得一提的是，与常见的移动端AI助手不同，AutoGLM 2.0定位于能够在设备上执行具体操作的智能体。· **AutoGLM 2.0可操作手机、电脑**. AutoGLM是智谱推出的智能体产品，支持一句话实现云端操作与自动执行。据智谱介绍...

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，智谱AI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

预测智谱AI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 08:26:06 AM CST 2026