豆包大模型推理算力需求测算

2024 年 12 月 26 日

➢ **豆包大模型攀升至国内 AI 应用第一。**截至 2024 年 12 月中旬，豆包大模型日均 tokens 使用量超过 4 万亿，较 5 月发布时增长超过 33 倍。根据量子位数据，截至 11 月底，豆包 APP 在 2024 年的累计用户规模已成功超越 1.6 亿，每日平均新增用户下载量稳定维持在 80 万，成为全球排名第二，国内排名第一的 AIAPP。11 月份，豆包 APPDAU 接近 900 万，增长率超过 15%。

➢ **豆包大家族全面更新：**12 月 18 日，在字节跳动所召开的火山引擎 Force 大会上，豆包三大主力模型引来全面升级。1）豆包通用模型 pro：综合能力比 5 月最初发布版本提升 32%，与 gpt-4o 持平，但价格仅是其八分之一。在指令遵循、代码、专业知识、数学、推理等层面全面对齐了 gpt-4o 水平，其中指令遵循能力提升 9%，代码能力提升 58%，专业知识方面能力提升 54%，数学能力提升 43%，推理能力提升 13%。2）豆包·音乐生成模型：生成水平从“高光片段”跃升到“完整乐章”，用户简单描述或上传一张图片，就能生成一首长达 3 分钟的包含旋律、歌词和演唱的高质量音乐作品，且提供局部修改功能，在针对部分歌词修改后仍能在原有旋律的节奏框架内适配。3）豆包·文生图模型：在通用性、可控性、高质量三方面取得新突破，新增“一键海报”和“一键 p 图”能力，对文字细节的指令遵循能力强，擅长“写汉字”，其背后的技术源自豆包·文生图模型原生的文字渲染能力以及 seededit 框架，目前已接入即梦 AI 和豆包 app。

➢ **AI 应用加速落地，推理算力需求或将崛起。**IDC 数据显示，2024 上半年中国加速服务器市场规模达到 50 亿美元，同比 2023 上半年增长 63%。其中 GPU 服务器依然占主导地位，达到 43 亿美元。同时 NPU、ASIC 和 FPGA 等非 GPU 加速服务器以同比 182%的增速达到近 7 亿美元市场规模。我们根据目前豆包的月活、日活以及日均 token 调用量为基础，做出保守、中性、乐观 3 种假设，结合大模型推理算力需求计算公式，对豆包带来的推理算力需求进行测算。在 3 种假设下，预计豆包大模型或将带来 759、1139、1898 亿元的 AI 服务器资本开支需求。

➢ **投资建议：**字节豆包大模型全面升级，月活攀升至国内 AI 应用第一，我们认为 AI 应用的加速落地或将带来推理侧算力需求的升级，进而推动头部互联网厂商持续提升 2025 年算力侧资本开支情况。建议关注：1）AI 服务器环节：浪潮信息、工业富联、紫光股份等；2）液冷环节：高澜股份、英维克、浪潮信息等；3）国产推理芯片环节：寒武纪、海光信息等。

➢ **风险提示：**AI 技术落地不及预期；算力行业竞争加剧；互联网厂商资本开支预算不及预期

推荐

维持评级

分析师 吕伟
执业证书：S0100521110003
邮箱：lvwei_yj@mszq.com
分析师 丁辰晖
执业证书：S0100522090006
邮箱：dingchenhui@mszq.com

相关研究

1.计算机行业事件点评：重视鸿蒙操作系统重要机遇-2024/12/24
2.计算机周报 20241222：OpenAI 十二天总结与 Agent 新范式-2024/12/22
3.计算机行业 2025 年度投资策略：2025：全面迎接 AI+大时代-2024/12/20
4.计算机周报 20241215：OpenAI 新品、豆包与 AI 新消费-2024/12/15
5.计算机周报 20241208：OpenAI 发布会分析展望与美股 AI 应用“狂飙”-2024/12/08

目录

1 AI 应用加速落地，推理算力需求或将崛起

1.1 豆包大模型全面升级，月活攀升至近 6000 万

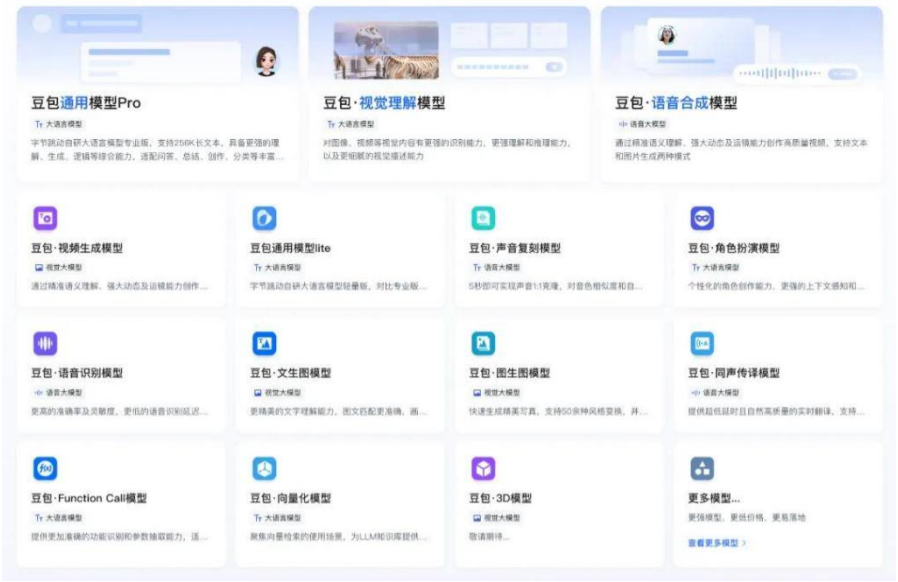
豆包大家族全面更新。12 月 18 日，在字节跳动所召开的火山引擎 Force 大会上，豆包三大主力模型引来全面升级。**1）豆包通用模型 pro：**综合能力比 5 月最初发布版本提升 32%，与 gpt-4o 持平，但价格仅是其八分之一。在指令遵循、代码、专业知识、数学、推理等层面全面对齐了 gpt-4o 水平，其中指令遵循能力提升 9%，代码能力提升 58%，专业知识方面能力提升 54%，数学能力提升 43%，推理能力提升 13%。**2）豆包·音乐生成模型：**生成水平从“高光片段”跃升到“完整乐章”，用户简单描述或上传一张图片，就能生成一首长达 3 分钟的包含旋律、歌词和演唱的高质量音乐作品，且提供局部修改功能，在针对部分歌词修改后仍能在原有旋律的节奏框架内适配。**3）豆包·文生图模型：**在通用性、可控性、高质量三方面取得新突破，新增“一键海报”和“一键 p 图”能力，对文字细节的指令遵循能力强，擅长“写汉字”，其背后的技术源自豆包·文生图模型原生的文字渲染能力以及 seededit 框架，目前已接入即梦 AI 和豆包 app。

图1：豆包通用模型 pro 综合能力提升 32%

资料来源：2024 冬季火山引擎 FORCE 原动力大会，民生证券研究院

豆包·3D 模型首次亮相。3D 模型采用 3D-DiT 等算法技术生成高质量的 3D 模型，与火山和英伟达合作的数字孪生平台 veOmniverse 结合。在技术层面，豆包 3D 生成模型基于深度学习、生成对抗网络（GAN）等前沿技术，能够以更高的真实感和细腻度生成三维视觉内容。3D 模型可实现 AIGC 世界的仿真模拟器，用户能高效完成智能训练、数据合成和数字资产制作，满足仿真训练的多样化需求，加速虚拟与现实的深度融合。

图2：豆包大模型目前的产品矩阵



资料来源：豆包官网，民生证券研究院

日均 tokens 较发布增长 33 倍。截至 2024 年 12 月中旬，豆包大模型日均 tokens 使用量超过 4 万亿，较 5 月发布时增长超过 33 倍。根据量子位数据，截至 11 月底，豆包 APP 在 2024 年的累计用户规模已成功超越 1.6 亿，每日平均新增用户下载量稳定维持在 80 万，成为全球排名第二，国内排名第一的 AI APP，11 月份，豆包 APPDAU 接近 900 万。

图3：豆包日均 tokens 增长超过 33 倍

资料来源：2024 冬季火山引擎 FORCE 原动力大会，民生证券研究院

图4：豆包 11 月 MAU 近 6000 万远超国内竞品

	国内排名		产品名 AI产品榜	应用(APP)简短描述 aicpb.com	11月上榜应用 APP MAU	11月上榜应用 MAU变化
3	1		豆包	AI 智能助手 | 抖音	59.98M	16.92%
4	2		文小言	你的随身智能助手 | 百度	12.99M	3.33%
5	3		Kimi 智能助手	Kimi 智能助手 | 月之暗面	12.82M	27.40%
6	4		智谱清言	工作提效 AI 助手 | 智谱	6.37M	22.18%
7	5		讯飞星火	懂我的AI助手 | 科大讯飞	5.94M	4.23%
8	6		天工AI	天工AI智能助手 | 昆仑万维	5.78M	3.15%
9	7		星野	所建皆你所AI | MiniMax	5.25M	2.65%
10	8		猫箱	开启你的 AI 奇遇 | 抖音	4.58M	22.51%
11	9		通义	你的超级AI助手 | 阿里	3.88M	3.48%
12	10		光速写作	语文作文批改与AI智能写作 | 作业帮	3.77M	2.28%
13	11		讯飞听见	讯飞听见 | 科大讯飞	3.17M	5.24%

资料来源：AI 产品榜，民生证券研究院

活跃率与活跃用户留存率领跑工具类 AI。根据 QuestMobile 数据显示，2024 年 9 月，豆包 APP 活跃率达 18.1%，较今年 1 月增加 4.8%；月人均使用天数 5.4 天，较今年 1 月增加 1.3 天；活跃用户 3 日留存率达 39.1%，较今年 1 月增加 8.9%。运营效率较年初大部分都有明显进步，头部效应更为明显，与逐渐兴起的

智能体生态关联紧密。

图5：2024 年 9 月典型工具类 AIGC APP 重点运营指标

AIGC APP	活跃率（DAU/MAU）	较1月变化	月人均使用天数（天）	较1月变化	活跃用户留存率（3日留存）	较1月变化
豆包	18.1%	+4.8pp	5.4	+1.3	39.1%	+8.9pp
文小言	15.8%	+2.7pp	4.7	+0.7	31.2%	+3.3pp
Kimi智能助手	13.1%	--	3.9	--	32.2%	--
天工	14.1%	+4.9pp	4.2	+1.4	28.0%	+5.4pp
智谱清言	12.8%	+1.0pp	3.8	+0.2	30.6%	+2.4pp
通义	13.1%	+4.1pp	3.9	+1.1	32.1%	+6.8pp
讯飞星火	11.8%	+1.9pp	3.5	+0.5	24.1%	-0.3pp
海螺AI	15.7%	+7.3pp	4.7	+2.1	34.1%	+10.6pp
星绘	14.3%	--	4.3	--	39.9%	--
腾讯元宝	11.6%	--	3.5	--	35.9%	--

资料来源：QuestMobile，民生证券研究院

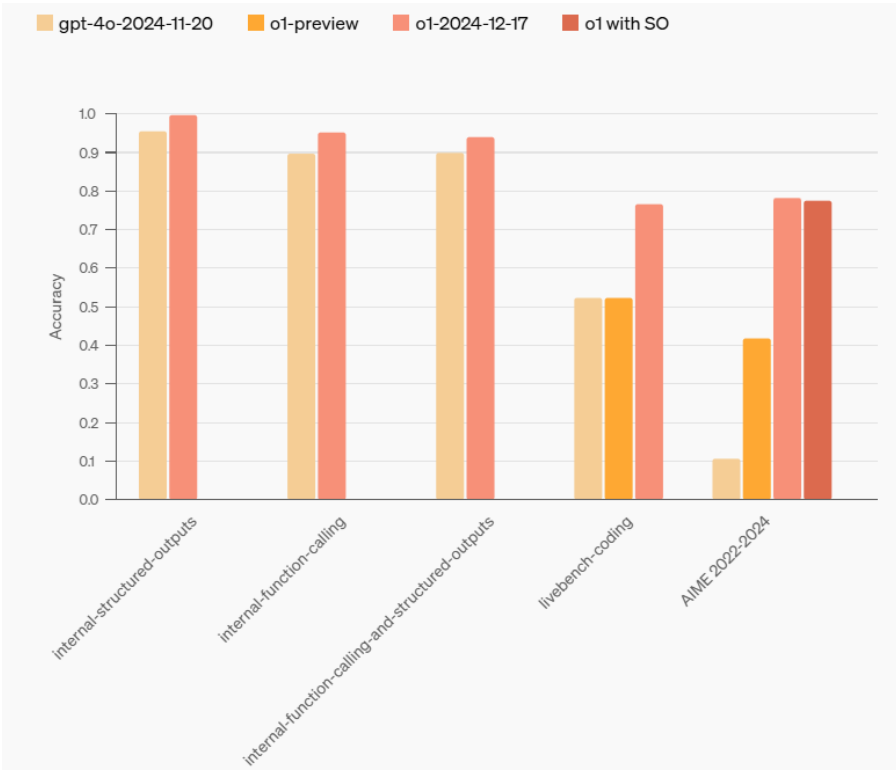
1.2 OpenAI 发布大量更新，海外 AI 加速向前

OpenAI 于 12 月 5 日开始的为期 12 天发布会上，在人工智能领域进行了大规模的更新迭代，主要集中体现在模型端、应用端与开发端：

模型端。在第九日（12 月 13 日）的发布会中，OpenAI 推出 o1 模型，o1 模型在准确性、效率和灵活性方面均实现了显著提升。在 SWE-bench Verified 中，o1 的编码结果从 41.3 提升至 48.9。而在 AIME 测试中，o1 的性能从 42 跃升至 79.2。o1 新增了几个特性，包括结构化输出功能和函数调用功能，简化了 o1 连接到 API 和数据库的过程，同时 o1 还具备了在视觉输入上进行推理的能力。OpenAI 在直播中还发布了 o3 模型的 mini 版本，mini 版本的模型尺寸更小，使用成本会有所降低。o3 mini 设置了低、中、高三种推理模式，用户能根据任务复杂度灵活调整模型的思考时间。

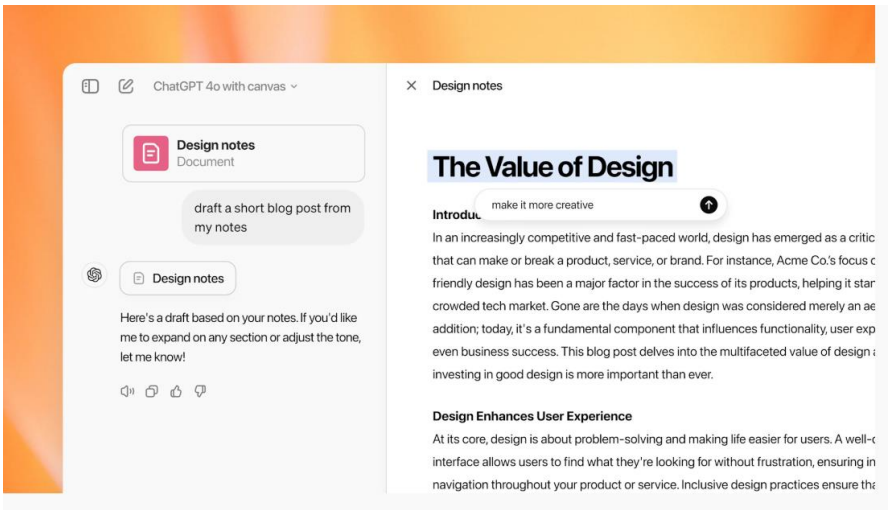
应用端。1）ChatGPT 搜索功能：OpenAI 宣布 ChatGPT 搜索功能正式全球落地，包括免费用户均可使用。其对搜索算法进行了深度优化，显著提升了搜索速度和准确性。用户提出问题后，ChatGPT 能够在极短的时间内（分钟级别）返回包括股票、新闻等在内的实时内容。**2）Sora 视频生成：**Sora 集成了 Storyboard、Remix、Re-cut 等功能，用户可以通过简单的文字描述生成分镜头视频，上传图像或创建静态视频帧，并设定播放时间，让 Sora 自动生成完整的视频作品。**3）Canvas：**Canvas 被设计为一个集智能写作、代码协作和 AI 智能体为一体的完整工作台，其内置了 WebAssembly Python 模拟器，创造了一个几乎无延迟的编程环境，并展现出理解代码意图的能力。

图6：不同指标维度下的模型精准性评估



资料来源：OpenAI 官网，民生证券研究院

图7：Canvas 交互界面



资料来源：OpenAI 官网，民生证券研究院

开发端。o1 模型推出 API 开放，o1 模型正式推出 API，成本降低 60%，新增函数调用、开发者消息、结构化输出以及视觉识别等功能。另一方面，实时 API 进一步更新，包括 WebRTC 支持、代码简化、价格优化以及 Python SDK 支持，简化了实时语音应用的开发。同时，发布了 Go SDK 和 Java SDK，支持所有 API 功能，并简化了 API 获取流程。

1.3 AI 应用全面落地，推理算力建设或成为新增量

大模型加速落地致使推理算力需求大幅增长。豆包大模型的应用场景不断拓展，在信息处理、客服与销售、硬件终端等场景的调用量快速提升。12 月日均 tokens 使用量超过 4 万亿，较 5 月发布时期增长超过 33 倍，这使得对推理算力的需求不断攀升，主要集中在硬件设备算力需求、数据中心规模扩张需求、通信网络需求三方面。

IDC 数据显示，2024 上半年中国加速服务器市场规模达到 50 亿美元，同比 2023 上半年增长 63%。其中 GPU 服务器依然占主导地位，达到 43 亿美元。同时 NPU、ASIC 和 FPGA 等非 GPU 加速服务器以同比 182%的增速达到近 7 亿美元市场规模。

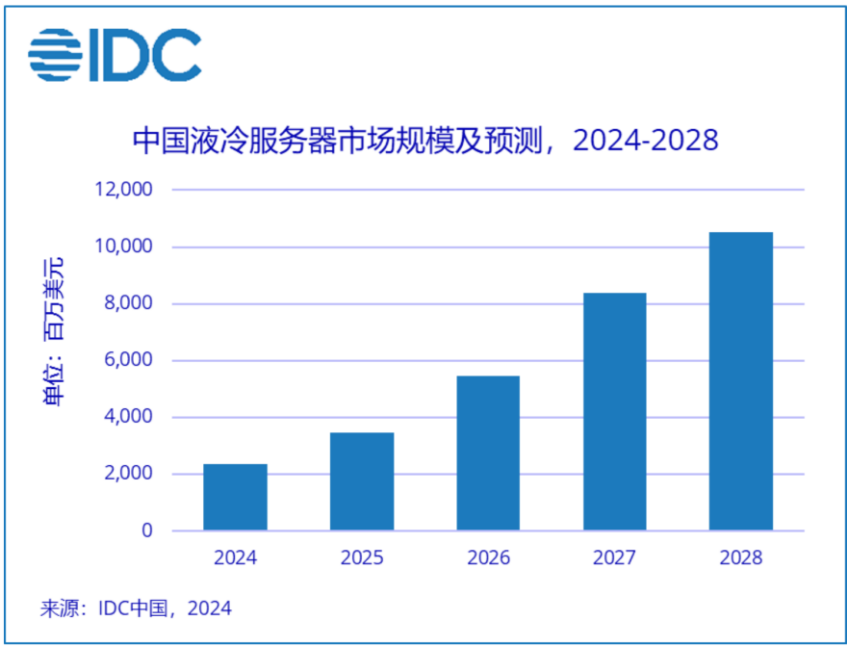
从国际环境来看，由于美国对相关技术及产品的管控，一方面限制了中国 AI 产业的发展；另一方面也激发了中国厂商自研 AI 芯片的积极性。但同时，国产自研芯片的后期维护与生态支持仍存在提升空间。从市场与产业链角度看，在深入开发行业大模型的当下，市场对于高算力且稳定的 AI 服务器的需求不断增大；5G 通信技术落地之后，市场对于全面高效协同的 AI 算力系统网络要求加深。这两点趋势在更高算力与更快带宽之外，都对国产 AI 服务器自身更短的平均故障时间、基础算力配套设施的更全面和运维团队的更专业、覆盖面更广提出了更高要求。IDC 预测，到 2028 年中国加速服务器市场规模将达到 253 亿美元。其中非 GPU 服务器市场规模将接近 50%。

图8：2024-2028 年中国加速计算服务器市场规模

资料来源：IDC，民生证券研究院

液冷服务器或显著受益于 AI 算力需求提升。IDC 正式发布《中国半年度液冷服务器市场（2024 上半年）跟踪》报告。数据显示，中国液冷服务器市场 2024 上半年同比大幅增长 98.3%，市场规模达到 12.6 亿美元，出货量同比增长 81.8%，其中浪潮信息以超过 50%的份额蝉联中国市场第一。液冷服务器市场将继续保持高速增长，预计 2023-2028 年，中国液冷服务器年复合增长率将达 47.6%，市场规模有望在 2028 年达到 102 亿美元。

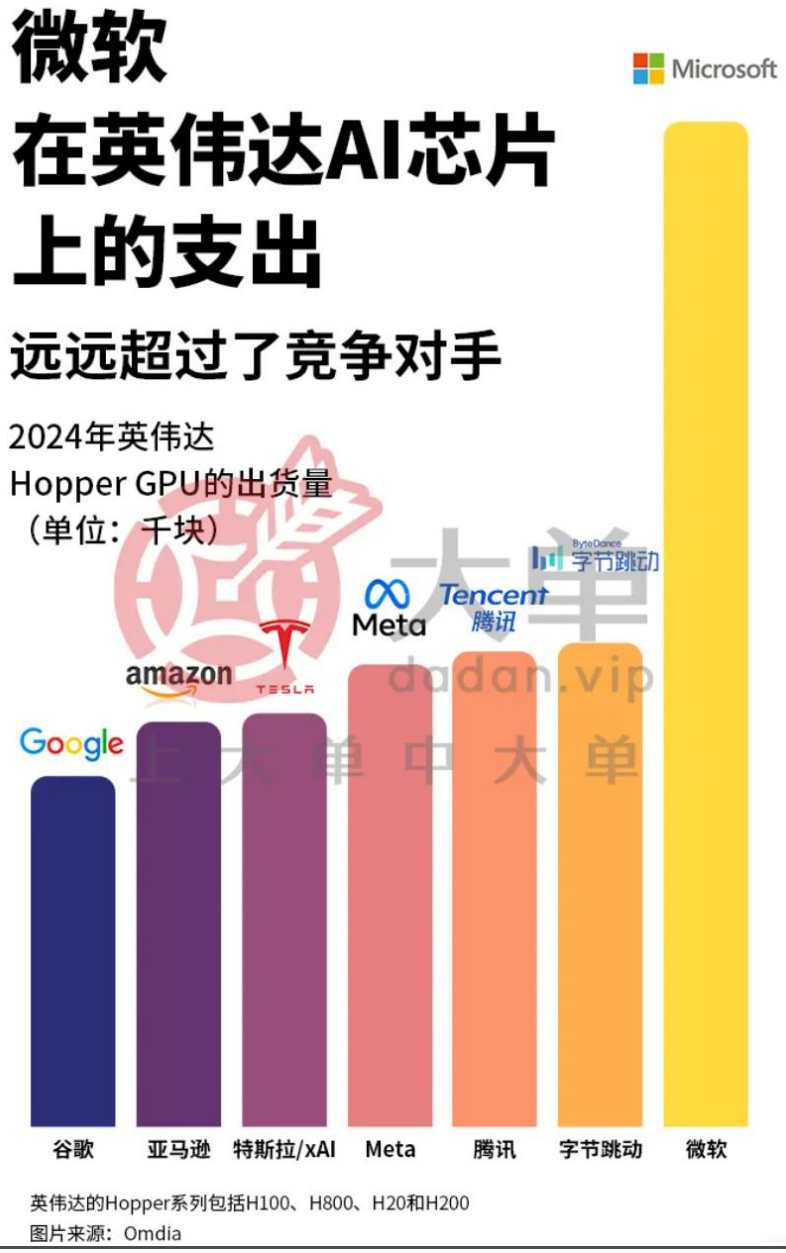
图9：2024-2028 年中国液冷服务器市场规模



资料来源：IDC，民生证券研究院

AI 应用有望显著带动算力建设，字节算力资本开支持续攀升。今年，各大科技巨头已在兴建运行英伟达最新芯片的数据中心上投入了数百亿美元。自两年前首次亮相的 ChatGPT 引发前所未有的 AI 投资热潮以来，英伟达的最新芯片已然成为硅谷最抢手的商品。Omdia 估计，字节跳动和腾讯今年各自订购了约 23 万块英伟达芯片，其中包括 H20 型号，这款低配版的 Hopper 经过改动，以满足针对中国客户的美国出口管制条例。

图10：全球科技巨头在英伟达 AI 芯片上的支出



资料来源：Omdia，云头条微信公众号，民生证券研究院

豆包大模型，将带来多少推理端的算力需求增量？我们根据目前豆包的月活、日活以及日均 token 调用量为基础，做出保守、中性、乐观 3 种假设，结合大模型推理算力需求计算公式，对豆包带来的推理算力需求进行测算。

在 3 种假设下，预计豆包大模型或将带来 759、1139、1898 亿元的 AI 服务器资本开支需求。

图11：三种假设情形下，豆包带来的推理需求测算

豆包大模型推理算力需求测算	2024年末	2025-2026E		
		保守	中性	乐观
MAU（万人）	6000	10000	15000	20000
DAU（万人）	900	1500	2250	3000
豆包（云雀）大模型参数量	1300	1300	1300	1300
日均token调用量（亿）	40000	66667	100000	133333
推理计算时间（s）	2	2	2	2
日均每秒token计算数量（亿）	0.31	0.51	0.77	1.03
峰值倍数	4	4	4	5
算力需求公式	云端 AI 推理算力需求≈2× 模型参数量 × 数据规模 × 峰值倍数			
算力需求结果（FLOPS）	3.20988E+19	5.34979E+19	8.02469E+19	1.33745E+20
H20单卡算力（TFLOPS）	148	148	148	148
MFU	50%	50%	50%	50%
需要H20卡数量（万张）	**43**	**72**	**108**	**181**
H20单价（万元）	8.4	8.4	8.4	8.4
H20服务器均价（万元）	84	84	84	84
豆包大模型创造AI服务器需求（亿元）	**455.46**	**759.09**	**1138.64**	**1897.73**

资料来源：AI 产品榜、OpenAI 官网、豆包大模型官网、英伟达官网、36 氪，民生证券研究院测算

2 投资建议

字节豆包大模型全面升级，月活攀升至国内 AI 应用第一，我们认为 AI 应用的加速落地或将带来推理侧算力需求的升级，进而推动头部互联网厂商持续提升 2025 年算力侧资本开支情况。建议关注：**1）AI 服务器环节：浪潮信息、工业富联、紫光股份等；2）液冷环节：高澜股份、英维克、浪潮信息等；3）国产推理芯片环节：寒武纪、海光信息。**

3 风险提示

1）AI 技术落地不及预期。推理算力方向的资本开支预算，与 AI 应用的落地情况息息相关，若后续豆包大模型月活、日活等数据进展不及预期，可能会影响未来资本开支的预算。

2）算力行业竞争加剧。若后续算力行业的竞争加剧，可能会影响产品的毛利率情况，进而影响算力环节公司整体的盈利能力。

3）互联网厂商资本开支预算不及预期。互联网厂商的资本开支预算与自身的经营情况以及业务发展息息相关，若后续整体经营受到宏观环境影响，亦或者 AI 的商业化落地进程不及预期，可能会影响资本开支的情况。

插图目录

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，客户应当充分考虑自身特定状况，不应单纯依靠本报告所载的内容而取代个人的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室； 518048