行业研究

行业专题报告

证券研究报告

电子

从训练到推理：算力芯片需求的华丽转身——算力专题研究二

投资要点：

- **推理算力：算力芯片行业的第二重驱动力**

我们在此前外发报告《如何测算文本大模型 AI 训练端算力需求？》中，对未来三年 AI 训练卡需求持乐观态度。我们认为，推理侧算力对训练侧算力需求的承接不意味着训练需求的趋缓，而是为算力芯片行业贡献第二重驱动力。当前推理算力市场已然兴起，24 年 AI 推理需求成为焦点。据 Wind 转引英伟达 FY24Q4 业绩会纪要，公司 2024 财年数据中心有 40%的收入来自推理业务。如何量化推理算力需求？与训练算力相比，推理侧是否具备更大的发展潜力？我们整理出 AI 推理侧算力供给需求公式，并分类讨论公式中的核心参数变化趋势，以此给出我们的判断。

- **Scaling Laws&长文本趋势：推理需求的核心驱动力**

根据 OpenAI《Scaling Laws for Neural Language Models》，并结合我们对于推理算力的理解，我们拆解出云端 AI 推理算力需求≈2×模型参数量×数据规模×峰值倍数。由 Scaling Laws 驱动的参数量爆发是训练&推理算力需求共同的影响因素；而对于推理需求，更为复杂的是对数据规模的量化。我们将数据规模（tokens）拆解为一段时间内用户对于大模型的访问量与单次访问产生的数据规模（tokens）的乘积，其中，单次访问产生的数据规模（tokens）可以进一步拆解为单次提问的问题与答案所包含的 token 数总和乘以单次访问提出的问题数。通过层层拆解，我们发现单次问答所包含的 token 数是模型中的重要影响因素，其或多或少会受到大模型上下文窗口（Context Window）的限制。而随着上下文窗口瓶颈的快速突破，长文本趋势成为主流，有望驱动推理算力需求再上新台阶。

- **结论：**

我们首先根据前述逻辑测算得到 AI 大模型推理所需要的计算量，随后通过单 GPU 算力供给能力、算力利用率等数值的假设，逐步倒推得到 GPU 需求数量。若以英伟达当代&前代 GPU 卡供给各占 50%计算，我们认为 2024-2026 年 OpenAI 云端 AI 推理 GPU 合计需求量为 148/559/1341 万张。

- **建议关注**

- 算力芯片：寒武纪 海光信息 龙芯中科

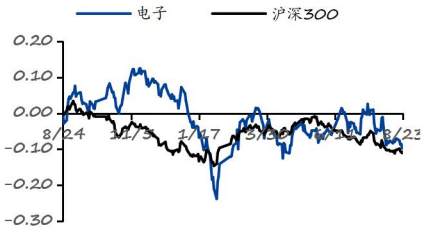
- 服务器产业链：工业富联 沪电股份 深南电路 胜宏科技

- **风险提示**

AI 需求不及预期风险、Scaling Law 失效风险、长文本趋势发展不及预期风险、GPU 技术升级不及预期的风险、测算模型假设存在偏差风险。

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **陈海进(S0210524060003)**
chj30590@hfzq.com.cn
分析师： **徐巡(S0210524060004)**
xx30511@hfzq.com.cn
联系人： **李雅文(S0210124040076)**
lyw30508@hfzq.com.cn

相关报告

1、AMD 宣布收购服务器供应商，英伟达强力加持“黑神话”游戏体验-算力周跟踪——2024.08.22
2、苹果领军 AI 端侧创新，消费电子长期量价上行空间打开——2024.08.21
3、20240818 周报：关注折叠屏手机形态演进及新机发布——2024.08.19

正文目录

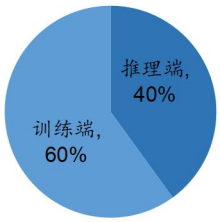
图表目录

1 如何测算文本大模型 AI 推理端算力需求？

我们在此前外发报告《如何测算文本大模型 AI 训练端算力需求？》中，对未来三年 AI 训练卡需求持乐观态度，经过测算，以英伟达 Hopper/Blackwell/下一代 GPU 卡 FP16 算力衡量，我们认为 2024-2026 年全球文本大模型 AI 训练侧 GPU 需求量为 271/592/1244 万张。由此我们认为，**推理侧算力对训练侧算力需求的承接不意味着训练需求的趋缓，而是为算力芯片行业贡献第二重驱动力。**

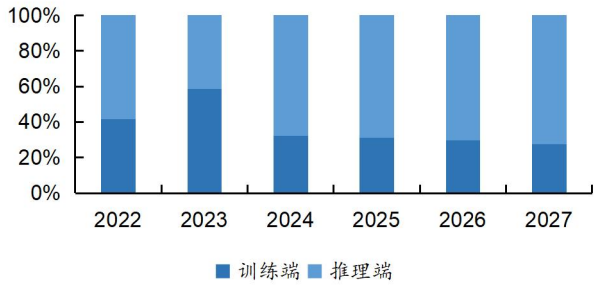
推理算力市场已然兴起，24 年 AI 推理需求成为焦点。据 Wind 转引英伟达 FY24Q4 业绩会纪要，公司 2024 财年数据中心有 40%的收入来自推理业务。放眼国内，IDC 数据显示，我国 23H1 训练工作负载的服务器占比达到 49.4%，预计全年的占比将达到 58.7%。随着训练模型的完善与成熟，模型和应用产品逐步投入生产，推理端的人工智能服务器占比将随之攀升，预计到 2027 年，用于推理的工作负载将达到 72.6%。

图表 1：英伟达 FY2024 数据中心推理与训练占比



来源：英伟达财报电话会议纪要，Wind，华福证券研究所
注：按销售收入口径

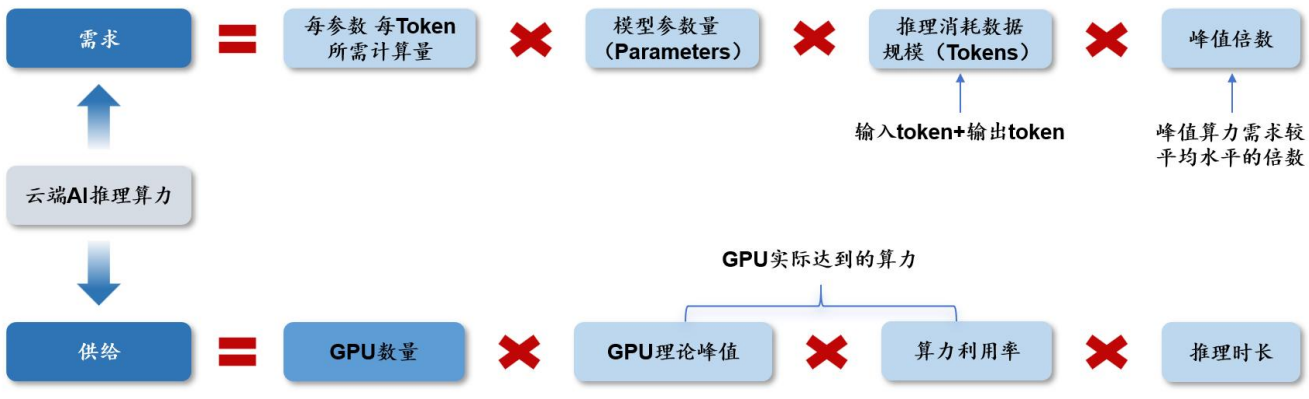
图表 2：中国人工智能服务器负载及预测



来源：IDC，《2023-2024 中国人工智能计算力发展评估报告》，华福证券研究所

如何量化推理算力需求？与训练算力相比，推理侧是否具备更大的发展潜力？我们整理出 AI 推理侧算力供给需求公式，并分类讨论公式中的核心参数变化趋势，以此给出我们的判断。需要说明的是，本文将视角聚焦于云端 AI 推理算力，端侧 AI 算力主要由本地设备自带的算力芯片承载。**基于初步分析，我们认为核心需要解决的问题聚焦于需求侧**——推理消耗的数据规模如何测算？而**供给侧**，GPU 性能提升速度、算力利用率等，我们认为与 AI 训练大致无异。

图表 3：文本大模型云端 AI 推理算力供给需求公式



来源：OpenAI《Scaling Laws for Neural Language Models》，NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，新智元，CIBA 新经济，思瀚产业研究院，极市平台，华福证券研究所

2 Scaling Laws&长文本趋势：推理需求的核心驱动力

根据 OpenAI《Scaling Laws for Neural Language Models》，并结合我们对于推理算力的理解，**我们拆解出云端 AI 推理算力需求≈2×模型参数量×数据规模×峰值倍数。**

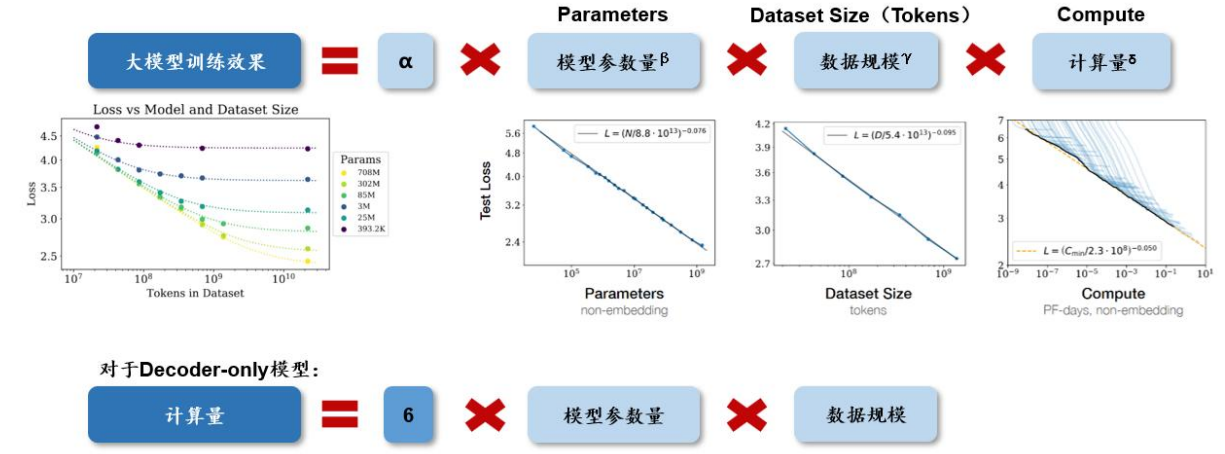
图表 4：云端 AI 推理需求公式拆解

来源：OpenAI《Scaling Laws for Neural Language Models》，思瀚产业研究院，极市平台，华福证券研究所

2.1 关于模型参数量：Scaling Laws 仍为核心

我们在此前外发报告《如何测算文本大模型 AI 训练端算力需求？》中已详细介绍过 Scaling Law 的基本原理及其对大模型参数量的影响，主要观点为：模型的最终性能主要与计算量、模型参数量和数据大小三者相关，而与模型的具体结构（层数/深度/宽度）基本无关。如下图所示，对于计算量、模型参数量和数据规模，（1）当不受其他两个因素制约时，模型性能与每个因素都呈现幂律关系。（2）如模型的参数固定，无限堆数据并不能无限提升模型的性能，模型最终性能会慢慢趋向一个固定的值。**因此，为了提升模型性能，模型参数量和数据大小需要同步放大。Scaling Law 仍然是当下驱动行业发展的重要标准。**

图表 5：大模型训练的 Scaling Law



来源：OpenAI《Scaling Laws for Neural Language Models》，PaperWeekly，Expara Academy，华福证券研究所

我们也详细统计了海内外主流 AI 大模型训练情况，过去几年来 AI 大模型参数量呈现快速增长。以 OpenAI 为例，GPT-3 到 GPT-4 历时三年从 175B 参数快速提升到 1.8T 参数（提升 9 倍）。目前国内主流 AI 大模型也逐步突破了千亿参数大关，乃至采用万亿参数进行预训练。

图表 6：海外主流 AI 大模型训练侧算力供给需求情况

	单位	GPT	GPT2	GPT3	GPT4	Gopher	PaLM	PaLM 2
基本信息								
发布机构		OpenAI	OpenAI	OpenAI	OpenAI	DeepMind	Google	Google
发布时间		2018-06	2019	2020-05	2023-03	2021-12	2022-04	2023-05
AI训练								
大模型算力需求								
参数量	亿	1	15	1746	18000	2800	5400	3400
预训练数据规模（token）	万亿			0.3	13.0	0.3	0.8	3.6
GPU算力供给								
GPU产品				V100	A100	TPU v3	TPU v4	
GPU数量	片			10000	25000	4096	6144	
理论峰值FP16 TC	TFlops			125	312			
算力利用率				21%	34%	33%	46%	

来源： OpenAI《Language Models are Few-Shot Learners》，Google《PaLM : Scaling Language Modeling with Pathways》，英伟达，谷歌研究院，腾讯科技，机器之心，中关村在线，河北省科学技术厅，华福证券研究所
注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计
注 2：GPT4 算力利用率在 32-36%区间，本文取中值粗略计算
注 3：英伟达 V100 理论峰值为官网所示“深度学习 |NVLink 版本”性能

图表 7：国内主流 AI 大模型训练侧算力供给需求情况

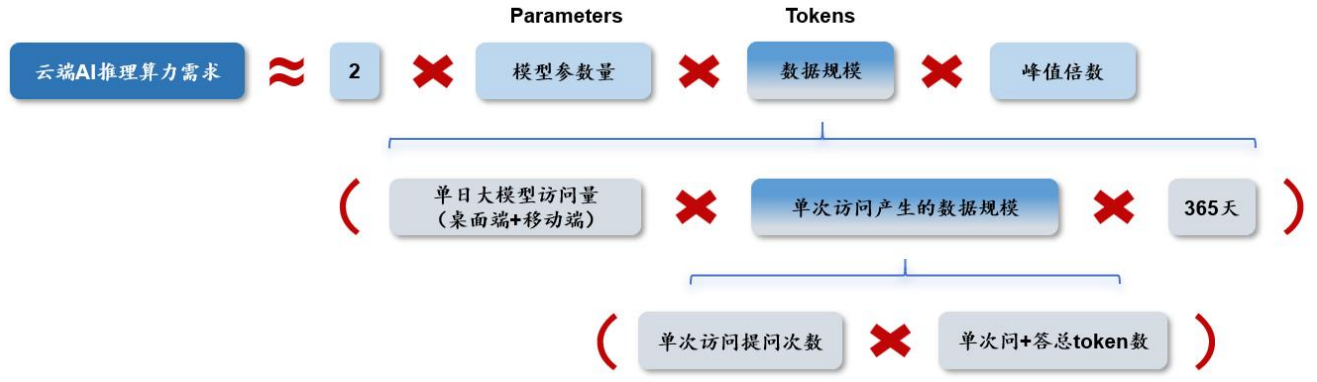
	单位	混元	混元	Qwen-72B	Qwen1.5-110B	商量2.0	商量5.0	DeepSeek v2
基本信息								
发布机构		腾讯	腾讯	阿里	阿里	商汤	商汤	幻方
发布时间		2023-09	2024-04	2023-11	2024-04	2023-07	2024-04	2024-05
AI训练								
大模型算力需求								
参数量	亿	超千亿	万亿	720	1100	1040	6000	2360
预训练数据规模（token）	万亿	2.0		3.0		1.6	10.0	8.1

来源：腾讯云，通义千问公众号&GitHub 网页，新闻晨报，市界，IT 之家，华尔街见闻，新浪科技，钛媒体，华福证券研究所
注 1：由于各公司对于大模型的训练数据披露口径不一，以上为本文非完全统计
注 2：腾讯混元参数量披露口径较为模糊，分别为超千亿参数/万亿参数，在本图中不涉及左侧第二列单位

2.2 关于数据规模（tokens）：长文本趋势已确立

根据公式，云端 AI 推理需求公式中的数据规模（tokens），可以拆解为一段时间内**用户对于大模型的访问量**与**单次访问产生的数据规模（tokens）**的乘积。

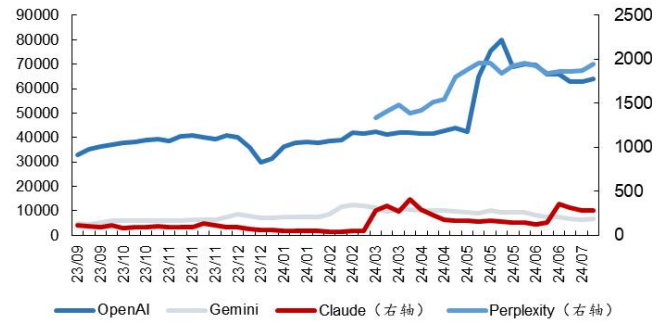
图表 8：云端 AI 推理需求公式进一步拆解



来源：OpenAI《Scaling Laws for Neural Language Models》，思瀚产业研究院，极市平台，华福证券研究所

1、从大模型访问量来看，我们认为需要覆盖到不同流量入口的访问量之和，包括（1）桌面端，（2）移动端。由于 Similarweb 数据统计了桌面端+移动端所有流量之和，我们以此为基础测算推理应用产生的 token 数据规模。

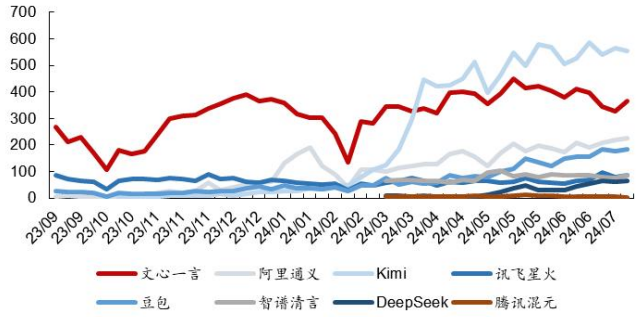
图表 9：文本大模型网站访问量周度数据(单位:万次)



来源：Similarweb，华福证券研究所

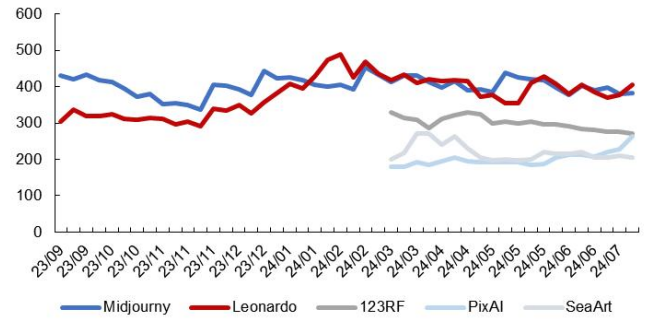
注：受网页改版影响，OpenAI 统计口径为 openai.com 和 chatgpt.com 之和

图表 10：文本大模型网站访问量周度数据（单位：万次）



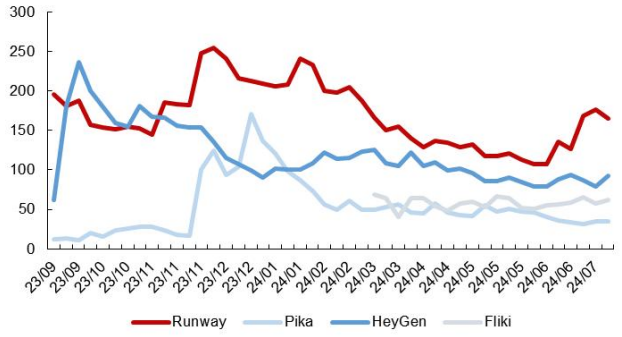
来源：Similarweb，华福证券研究所

图表 11：图片大模型网站访问量周度数据（单位：万次）



来源：Similarweb，华福证券研究所

图表 12：视频大模型网站访问量周度数据（单位：万次）



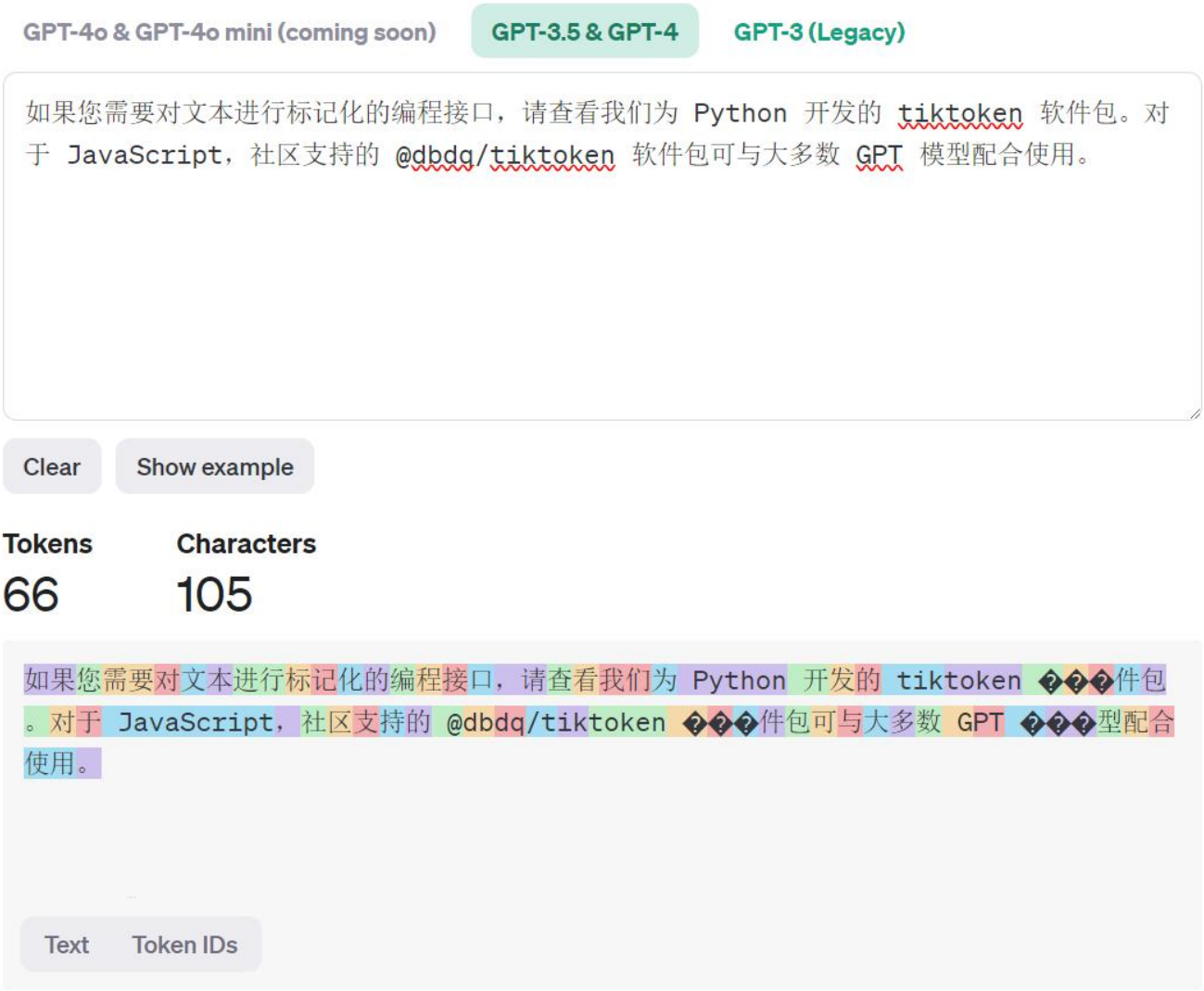
来源：Similarweb，华福证券研究所

2、从单次访问产生的数据规模来看，运算公式可以拆解为单次提问的问题与答案所包含的 token 数总和乘以单次访问提出的问题数，其中单次问答所包含的 token 数取决于**字数&每个字对应的 token 数**。

（1）字数：单次问答所包含的字数或多或少会受到大模型上下文窗口（Context Window）的限制。随着上下文窗口瓶颈的快速突破，**长文本趋势成为主流。**以 OpenAI 为例，从 GPT-3.5 升级至 GPT-3.5-Turbo，上下文窗口从 4k 升级为 16k；而 GPT-4 版本时隔一年后升级为 GPT-4-Turbo，也将上下文窗口从 32k 提升至 128k。而以长文本能力成为“顶流”的 Kimi，2023 年 10 月上线时支持无损上下文长度最多为 20 万汉字，24 年 3 月已支持 200 万字超长无损上下文，长文本能力提高 10 倍。按照 AI 领域的计算标准，200 万汉字的长度大约为 400 万 token，在全球范围内也属于领先的标准。

（2）每个字对应的 token 数：不同的大模型均有各自的分词器设计，以 OpenAI 为例，1000 个 token 通常代表 750 个英文单词或 500 个汉字。以其官网 Tokenizer 计算工具结果来看，基本与此结论相契合。

图表 13：OpenAI Platform Tokenizer



来源：OpenAI Platform，华福证券研究所
注：以上文字为测试语句，来自该网页下方原文

3 文本大模型云端 AI 推理对 GPU 的需求量如何求解？

本文按第一章所示“文本大模型云端 AI 推理算力供给需求公式”，逐步拆解计算过程。首先，测算 AI 大模型推理所需计算量，随后通过对单 GPU 算力供给能力的假设，逐步倒推得到 GPU 需求数量。需要说明的是，根据图 9-12 中 Similarweb 统计数据，我们发现当前 OpenAI 在全球范围内的访问量仍然断层领先，由此我们推断 OpenAI 在全球推理算力需求中占据较大比重，因此本文测算以 OpenAI 为例。

一、云端 AI 推理算力需求

1、每参数每 token 所需计算量：根据 OpenAI《Scaling Laws for Neural Language Models》，并结合我们对于推理算力的理解，我们拆解出云端 AI 推理算力需求≈2×模型参数量×数据规模×峰值倍数，即每参数每 token 推理所需计算量为 2 Flops。

2、大模型参数量：根据我们此前外发报告《如何测算文本大模型 AI 训练端算力需求？》，我们认为 Scaling Law 仍将持续存在，大模型或将持续通过提升参数量、预训练数据规模（token 数）带动计算量提升，进而提升大模型性能，我们按过往提升速度大致推断未来增长情况。

3、数据规模（tokens）：从大模型访问量来看，我们以 Similarweb 统计的访问量数据为测算基础，通过趋势的判断，以及对 AI 推理能力提升带动 AI 应用渗透趋势的信心，我们预计 24-26 年 OpenAI 访问量有望同比增长 60%/50%/30%。需要说明的是，由于 OpenAI 在 Similarweb 统计的访问量数据中以断层优势处于领先地位，我们认为 OpenAI 在全球 AI 推理需求中占据相当大的比重。**从单次访问产生的数据规模来看，**运算公式可以拆解为单次提问的问题与答案所包含的 token 数总和乘以单次访问提出的问题数，其中单次问答所包含的 token 数取决于**字数&每个字对应的 token 数**。（1）首先，我们假设 23 年大模型单次访问的问答次数为 5 次，随着大模型的使用频率提高，用户粘性增强，该次数有望稳步提升。（2）我们假设单次提问一般对应 30tokens 左右，另外我们取 23 年一次问答实验中 12 次回答的平均字数（523 个汉字）作为假设，基于 1:2 的换算比例，得到 23 年单次问答产生的 token 数为 1077 字节。我们假设 24-26 年单次问答产生的 token 数有望跟随大模型上下文窗口的长文本化趋势，而呈现爆发式增长。

4、峰值倍数：我们假设 1）推理需求在一日之内存在峰谷，算力储备比实际需求高，2）随算力应用的进一步泛化，峰值倍数有望逐渐下降。

二、云端 AI 推理算力供给

1、GPU 计算性能：根据我们此前外发报告《如何测算文本大模型 AI 训练端算力需求？》，我们假设未来英伟达新产品的 FP16 算力在 Blackwell 架构的基础上延续过往倍增趋势。此外，我们假设训练卡供不应求，由此推断推理需求的实现相较于训练需求的实现或延迟半代，AI 推理侧或以英伟达新一代及次新一代 GPU 为主，我

们假设二者各占 50%，以其平均 FP16 算力作为计算基准。

2、训练时间&算力利用率：我们假设推理应用需求在全年 365 天是持续存在的，由此假设全年推理时间为 365 天。根据我们此前外发报告《如何测算文本大模型 AI 训练端算力需求？》，我们假设算力利用率在 30-42%区间，逐年提升。

三、结论：若以英伟达当代&前代 GPU 卡供给各占 50%计算，我们认为 2024-2026 年 OpenAI 云端 AI 推理 GPU 合计需求量为 148/559/1341 万张。

图表 14：OpenAI 云端 AI 推理算力需求-供给测算

市场空间测算 | 需求侧

假设

1、根据Scaling Law，我们给出OpenAI大模型参数量预测
其中：23年由于经历从GPT-3.5迭代至GPT-4的过程，我们取二者参数量平均值
2、根据Similarweb数据测算OpenAI全年访问量，我们保守估计访问量增速或将逐年下降
3、假设23年大模型单次访问的问答次数为5次，随着大模型的使用频率提高，用户粘性增强，该次数有望稳步提升
4、单次问答产生的token数，为提问+回答产生的token数之和。
其中：1）23年：我们假设单次提问一般对应30tokens左右
我们取23年一次问答实验中12次回答的平均字数对应的token数，作为单次回答产生的token数假设
2）24-26年：我们假设单次问答产生的token数有望跟随大模型上下文窗口的长文本化趋势，而呈现爆发式增长
5、假设：1）推理需求在一日之内存在峰谷，算力储备比实际需求高，2）随算力应用的进一步泛化，峰值倍数有望逐渐下降
基于此，我们假设23-26年峰值倍数约为5.0、4.5、4.0、3.0

			2023	2024E	2025E	2026E
OpenAI推理算力需求	**ZFlops**	**10^21Flops**	**807520**	**17080845**	**221418356**	**1554356856**
每参数每token所需计算量			2	2	2	2
参数量		万亿	1.0	1.8	3.5	7.0
token数	TeraByte	万亿字节	82	1054	7908	37008
全年访问量		亿次	167	269	404	525
单次访问产生的token数		字节	5384	43073	215367	775320
单次访问的问答次数		次	5	8	10	12
单次问答产生的token数		字节	1077	5384	21537	64610
峰值倍数			5.0	4.5	4.0	3.0

市场空间测算 | 供给侧

假设

1、假设推理应用需求在全年365天是持续存在的
2、假设训练卡供不应求，由此推断推理需求的实现以英伟达新一代及次新一代GPU为主，我们以平均FP16算力作为计算基准
3、假设随着AI大模型行业发展，AI训练侧算力利用率逐年提升

			2023	2024E	2025E	2026E
OpenAI推理GPU需求		**万张**	**39**	**148**	**559**	**1341**
每秒平均算力需求	ZFlops	10^21Flops	0.1	1.7	19.5	117.4
算力利用率			30%	32%	36%	42%
GPU平均FP16性能	TFlops		219	1146	3490	8750
新一代GPU			Ampere	Hopper	Blackwell	下一代产品
FP16性能			312	1979	5000	12500
用量占比			50%	50%	50%	50%
次新一代GPU			Volta	Ampere	Hopper	Blackwell
FP16性能			125	312	1979	5000
用量占比			50%	50%	50%	50%

来源：英伟达，Similarweb，OpenAI Platform，OpenAI《Language Models are Few-Shot Learners》，OpenAI《Scaling Laws for Neural Language Models》，NVIDIA&Stanford University&Microsoft Research《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》，腾讯科技，量子位，手机光明网，第一财经，内蒙古自治区通信学会等，华福证券研究所测算

4 风险提示

AI 需求不及预期风险、Scaling Law 失效风险、长文本趋势发展不及预期风险、GPU 技术升级不及预期的风险、测算模型假设存在偏差风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，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。

一般声明

华福证券有限责任公司（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。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。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所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn