

证券研究报告 | 传媒行业 | 2023年7月21日

传媒团队·行业深度报告

Open AI 的盈亏分析

分析师 | 杨晓峰 | 登记编号: S1220522040001

联系人 | 杨昊 | 马梓燕

1、算力的逻辑：“10%的精度 vs 50%的算力”。根据Yufan Liu等人的论文研究，模型精度每下降10%，算力可能减半。在2500万DAU的假设条件下，若每个用户10次的使用次数，GPT-3.5所需GPU数量大约为27.2万A100。在同样的假设条件下，若New Bing之后使用压缩6倍的模型（得分76.2分），所需算力约为4.5万GPU。

2、Open AI的总体盈亏分析。据data.ai数据显示，截至2023年6月19日，ChatGPT iOS端上线首月的日活付费率约为4.36%。在GPT-3.5压缩成本的情况下，如果日活月活比例达到35%，月活付费率突破12%，对于压缩后的GPT-3.5或能实现盈亏平衡。Open AI资金预测：对于压缩成本后的GPT-3.5模型和GPT-4模型，若月付费率每月提升0.25%或不能持续；若月付费率每月提升0.5%或能扭亏。

3、Open AI“单用户”盈亏分析。日活/月活比例与成本的权衡：日活跃用户与月活跃用户比例用于衡量产品的用户粘性，比率越大粘性越好。如果在统计周期内，ChatGPT用户使用产品的频率越大，日活月活比例越高，产品使用年限越长，此时面临的问题是成本也越高。GPT-4访问限制推出后，“高频用户”的使用受限，平均请求次数可能下降。

4、如何提高GPT-4付费率。在ChatGPT用户中，主要使用者为程序员；其次还包括学生和教师用户、设计师/创作者用户等；此外，还有社会科学家、游戏玩家、投资者等多类用户。GPT-4独特的第三方插件及“代码解释器”功能，推动用户付费：2023年5月，Open AI全面开放了GPT-4第三方插件功能（plugins），这些功能仅向Plus用户开放，促进了用户的付费；2023年7月11日，ChatGPT“代码解释器”（Code Interpreter）测试版正式向所有Plus用户开放，代码解释器扩展了ChatGPT的功能，为用户带来更好的交互式编程体验和强大的数据可视化功能，可能推动用户付费。

5、投资建议：

- 1) 若Open AI实现盈利，或可刺激所有海外巨头加大资本开支，建议关注算力相关公司。
- 2) 在Open AI收费的带动下，或有利于用户养成付费意识，各类AI相关应用有望百花齐放。建议关注：“AI+游戏”宝通科技、凯撒文化、恺英网络、巨人网络、神州泰岳、吉比特、完美世界；“AI+角色”奥飞娱乐、芒果超媒、掌阅科技；“AI+图像”美图；“AI+营销”三人行、蓝色光标、捷成股份等。
- 3) 若Open AI实现盈利，则其他公司在大模型的投入或将加大，“数据要素”的地位进一步凸显。建议关注：人民网、新华网。

6、风险提示：Open AI营收平衡效果不及预期，用户付费意愿不及预期，海外大厂资本开支有限。

Open AI 是整个AI大模型的领跑者，其商业模式决定了其他企业后续的AI资本开支。

1、提供“高精度模型平台”（GPT-3.5及以上，含GPT-4）的巨头会加大资本开支。

1.1 “搜索场景公司”：谷歌，百度

谷歌：谷歌搜索场景与Open AI相似，如果Open AI实现盈利，谷歌可以在Bard中加入会员付费，或加大其算力资本投入。**百度：**百度的搜索场景同样与Open AI类似，Open AI的盈利可能会影响百度的资本投入，可能在其搜索中加入AI功能。

1.2 “社交场景”：Meta，腾讯

Meta：Meta存在社交场景，如果Open AI能实现盈利，Meta或在社交场景中加入聊天机器人，如Inflection的Pi聊天机器人（情商机器人）。**腾讯：**腾讯拥有微信等社交场景，如果Open AI能够实现盈利，腾讯或也能在微信等社交场景中接入AI功能，影响其资本开支。

1.3 “云计算厂商”：微软，阿里

微软：微软主要提供API服务，如果获得微软云计算资源大力支持的Open AI能够盈利，第三方AI应用的盈利预期受到影响，未来微软或也将继续加大资本投入。**阿里：**阿里作为云计算厂商同样提供API服务，若Open AI实现盈利，第三方AI应用的盈利预期也会提高，可能增加对于API的需求，可能也会影响到阿里的资本开支。

2、接入API的第三方公司：只有Open AI实现盈亏平衡了，才能进一步加大API的投入。算力需求首先提供给自己服务，自己应用的用户是Open AI的直接用户，粘性较强，而第三方应用的用户为间接用户，粘性较弱，所以优先考虑直接用户，其次考虑第三方接入API的公司，因此当Open AI实现盈利，才能影响进一步加大API的投入。

一、算力的逻辑：“10%的精度 vs 50%的算力”

二、Open AI 的总体盈亏分析

三、Open AI “单用户” 盈亏分析

四、如何提高GPT-4付费率

五、投资建议及风险提示

一、算力的逻辑：“10%的精度 vs 50%的算力”

1、算力与精度的关系

算力与精度的关系：根据Yufan Liu等人的论文研究，模型精度每下降10%，算力可能减半。模型的大小由其参数量及其精度决定，精度通常为FP64、FP32、FP16、BF16、TF32、INT8、IN4等，精度下降使得算力承载扩大的同时，也会导致性能在一定程度上下降。根据Yufan Liu等人的研究，其模型测试精确度降低到原来的90%左右时，模型被修剪后剩余FLOPs的数量约占原始网络中FLOPs总数的50%。因此，可以通过使用更低的精度来减少GPU需求，具体关系大约为精度每下降10%，所需算力减少到原来的二分之一。

图表：模型算力与精度的关系

Model Architecture	Method	Training Settings		Test Accuracy		PR _{FLOPs}
		Pre-defined?	Fine-tuning?	Top-1	Top-5	
ResNet34	Baseline	N/A	N/A	73.92%	91.62%	0
	SFP [31]	✓	✗	71.83%	90.33%	41.10%
	FPGM [29]	✓	✗	72.11%	90.69%	41.10%
	ASFP [32]	✓	✗	71.72%	90.65%	41.10%
	DPEPS [47]	✗	✗	72.25%	90.80%	43.29%
	Li <i>et al.</i> [28]	✓	✓	72.17%	—	24.20%
	IE [64]	✓	✓	72.83%	—	24.20%
	DMCP [65]	✗	✓	72.96%	90.99%	39.87%
	EagleEye [54]	✓	✓	72.87%	90.94%	39.37%
	ResRep [53]	✗	✗	73.02%	91.02%	41.65%
	Logits [5]	✓	✗	72.42%	90.90%	28.38%
	AT [17]	✓	✗	72.38%	90.88%	28.38%
	IRG [19]	✓	✗	72.73%	90.99%	28.38%
	MetaDistiller [27]	✓	✗	72.58%	90.92%	28.38%
	VID [55]	✓	✗	72.56%	90.92%	28.38%
	CRD [56]	✓	✗	72.71%	90.99%	28.38%
	SAD [57]	✓	✗	72.67%	90.97%	28.38%
	DPEPS+IRG	✗	✓	72.53%	90.89%	42.56%
	Ours	✗	✗	73.26%	91.09%	49.14%

图表：在假设条件为模型每降低10%，算力减半的背景下：模型算力与精度的计算公式

$$\text{算力} = \frac{\text{原算力}}{2^{\log_{0.9} \frac{\text{精度}}{\text{原精度}}}}$$

2、单张A100芯片每日吞吐量测算

单张A100卡的每日吞吐量测算：根据英伟达官网数据，NVIDIA DGX A100 640GB的服务器，包含8个NVIDIA A100 80GB Tensor Core GPU，DGX A100具有高达640GB的总GPU显存，可将大规模训练作业的性能提升高达3倍，并将MIG实例的大小增加一倍，从而从容应对颇为复杂的大任务，以及简单轻松的小任务。据微软官方介绍，每个ND A100 v4 series虚拟机有8块A100芯片，对于Davinci 3模型（GPT-3.5）使用英伟达80G显存的NDA100芯片，若需要大约3个虚拟机，共需要约24块芯片。考虑到高并发和低并发的情况，Davinci 3模型在低并发情况下平均每秒可以处理0.28个请求；而在高并发状态下，平均每秒钟可以处理0.34个请求，此时，一张卡每天的吞吐量大约为1224次（ $0.34 \times 60 \text{ Secs} \times 60 \text{ Mins} \times 24 \text{ Hrs} / 24 \text{ GPU Cards}$ ）。但在实际使用时，若芯片全部打满容易出现崩掉的情况，因此需要考虑芯片的使用效率。

图表：模型算力支持拆解因素

模型	硬件支持	硬件布局	RPS (每秒钟请求数)	并发能力
Davinci 3	NDA100 80GB	3VMs, 每个有8GPUs 1.92 TB GPU memory 花费: \$52.5/h	0.28	低并发
			0.34	高并发



$$0.34 \times 60 \text{ Secs} \times 60 \text{ Mins} \times 24 \text{ Hrs} / 24 \text{ GPU Cards} = 1224 \text{ turns/Card*day}$$

3、精度和算力的换算

不同模型得分及精度转换：根据Tim Dettmers等人的论文，团队使用大模型GPT-4当裁判，对不同模型的回答进行打分，以GPT-3.5的成绩作为100%，最终GPT-4自己的得分是114.5%。Michal Kosinski的研究表明，GPT-3可以解决70%的心智理论任务，而GPT-3.5解决了93%的任务，以GPT-3.5的成绩作为100，GPT-3的分数约为75（70/93*100）。如果使用分数对精度进行衡量，则GPT-4精度相当于GPT-3.5的1.145倍，GPT-3精度相当于GPT-3.5的0.75倍。

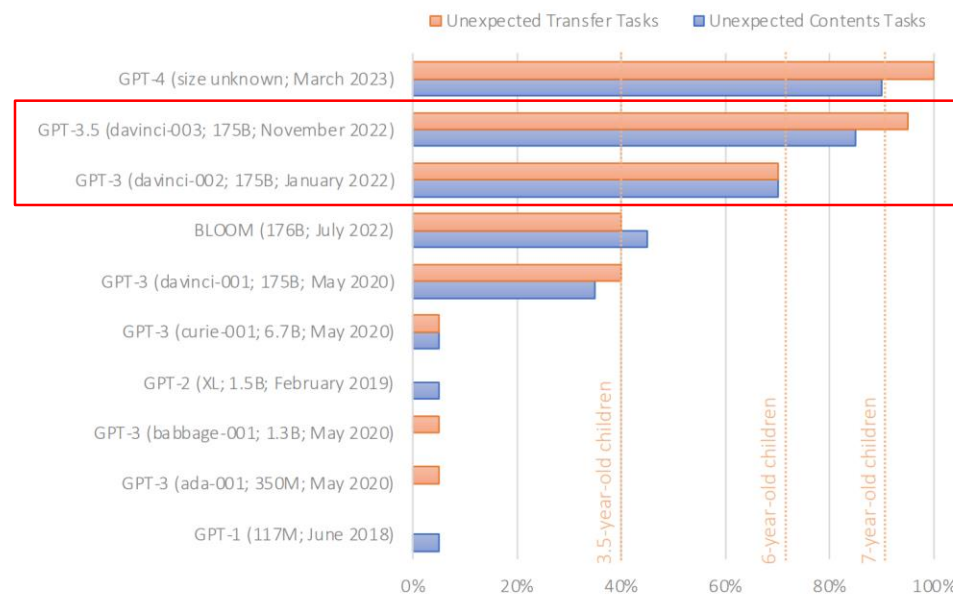
此外，如果New Bing目前使用的模型为压缩约6倍后的模型，所需算力约减少到原来的六分之一。同样，模型的质量会受到一定的影响，压缩后大约从GPT-3.5下降到GPT-3的水平，具体其打分换算过来约76.2分。

图表：不同模型以GPT-4自动评估的得分

Table 6: Zero-shot Vicuna benchmark scores as a percentage of the score obtained by ChatGPT evaluated by GPT-4. We see that OASST1 models perform close to ChatGPT despite being trained on a very small dataset and having a fraction of the memory requirement of baseline models.

Model / Dataset	Params	Model bits	Memory	ChatGPT vs Sys	Sys vs ChatGPT	Mean	95% CI
GPT-4	-	-	-	119.4%	110.1%	114.5%	2.6%
Bard	-	-	-	93.2%	96.4%	94.8%	4.1%
Guanaco	65B	4-bit	41 GB	96.7%	101.9%	99.3%	4.4%
Alpaca	65B	4-bit	41 GB	63.0%	77.9%	70.7%	4.3%
FLAN v2	65B	4-bit	41 GB	37.0%	59.6%	48.4%	4.6%
Guanaco	33B	4-bit	21 GB	96.5%	99.2%	97.8%	4.4%
Open Assistant	33B	16-bit	66 GB	91.2%	98.7%	94.9%	4.5%
Alpaca	33B	4-bit	21 GB	67.2%	79.7%	73.6%	4.2%
FLAN v2	33B	4-bit	21 GB	26.3%	49.7%	38.0%	3.9%
Vicuna	13B	16-bit	26 GB	91.2%	98.7%	94.9%	4.5%
Guanaco	13B	4-bit	10 GB	87.3%	93.4%	90.4%	5.2%
Alpaca	13B	4-bit	10 GB	63.8%	76.7%	69.4%	4.2%
HH-RLHF	13B	4-bit	10 GB	55.5%	69.1%	62.5%	4.7%
Unnatural Instr.	13B	4-bit	10 GB	50.6%	69.8%	60.5%	4.2%
Chip2	13B	4-bit	10 GB	49.2%	69.3%	59.5%	4.7%
Longform	13B	4-bit	10 GB	44.9%	62.0%	53.6%	5.2%
Self-Instruct	13B	4-bit	10 GB	38.0%	60.5%	49.1%	4.6%
FLAN v2	13B	4-bit	10 GB	32.4%	61.2%	47.0%	3.6%
Guanaco	7B	4-bit	5 GB	84.1%	89.8%	87.0%	5.4%
Alpaca	7B	4-bit	5 GB	57.3%	71.2%	64.4%	5.0%
FLAN v2	7B	4-bit	5 GB	33.3%	56.1%	44.8%	4.0%

图表：GPT模型测试



4、GPT模型不同精度模型下的算力需求

在2500万DAU的假设条件下，若每个用户10次的使用次数，GPT-3.5所需GPU数量大约为27.2万A100。如果2024年Bing Chat的DAU达2500万，每个用户提10次问题，若按照芯片75%的使用效率进行测算，GPT-3.5所需GPU数量大约为27.2万（ $2500 \times 10 \text{ turns} / 1224 \text{ turns} / 0.75 \text{ GPU utilization rate}$ ）。若根据精度下降10%，算力除以2的逻辑，可以推算出GPT-4在2500万日活及10次平均使用次数下条件下所需算力大约为66.4万（ $27.2 \text{ GPU} / (2^{\log_{0.9}(114.5/100)})$ ）。

在同样的假设条件下，若New Bing之后使用压缩6倍的模型（得分76.2分），所需算力约为4.5万GPU。如果微软对模型进行压缩，压缩后的模型使用4块芯片，其算力需求约为4.5万。同样地，在相同假设前提下，GPT-3所需的算力约为4.1万（ $27.2 \text{ GPU} / (2^{\log_{0.9}(75/100)})$ ）。根据机器之心，GPT-4的推理成本是Davinci模型（GPT-3.5为Davinci3）的3倍，与我们测算结果相似。

图表：不同模型的精度及所需算力

模型	分数	精度	GPU（万）
GPT-3	75	0.75	4.1
GPT-3.5	100	1	27.2
GPT-4	114.5	1.145	66.4
New Bing目前模型	76.2	0.76	4.5

$$\frac{27.2}{2^{\log_{0.9} \frac{114.5}{100}}} = 66.4$$

二、Open AI 的总体盈亏分析

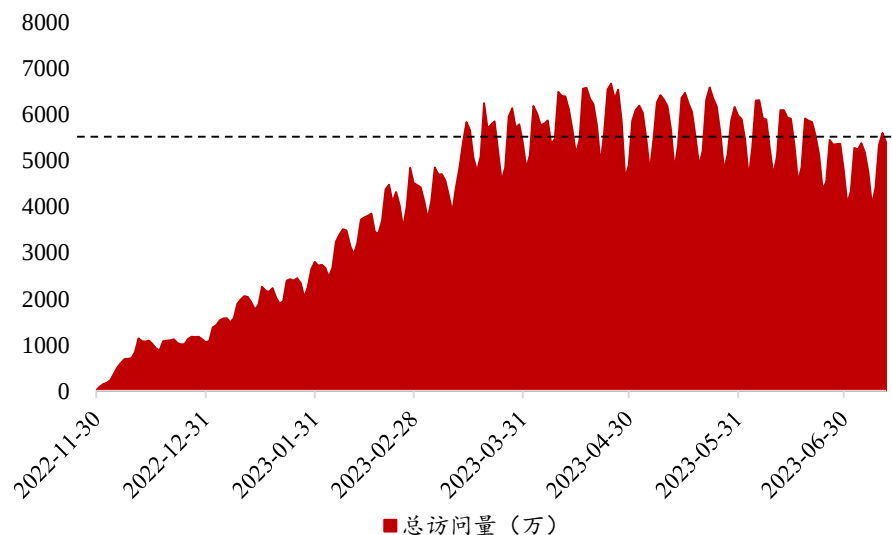
- 1、据data.ai数据显示，截至2023年6月19日，ChatGPT iOS端上线首月的日活付费率约为4.36%。截至2023年6月19日，ChatGPT iOS端5月21日-6月19日美国地区平均日活用户量约94.64万人，累计付费用户数约为4.13万人，因此日活付费率（月付费用户数/日活跃用户）约为4.36%。
- 2、GPT-3.5的成本分析：如果GPT-3.5的精度下降为原来的95.7%，成本大约下降25%。目前，由于大量GPT-3.5用户没有付费，Open AI可能选择降低其成本。2023年6月13日，Open AI宣布降低GPT-3.5等模型成本，具体为gpt-3.5-turbo每1K输入tokens降低25%至0.0015美元、每1K输出tokens降低至0.002美元。
- 3、GPT-4的付费分析：ChatGPT使用用户画像：在ChatGPT用户中，程序员所占比例最多；学生和教师用户次之；设计师、创作者用户所占比例排名第三；此外，还有社会科学家、游戏玩家、投资者等多类用户。为什么为GPT-4付费：GPT-4具有独特的第三方插件及“代码解释器”功能。2023年5月，Open AI向Plus用户全面开放了GPT-4第三方插件功能（plugins）；2023年7月11日，ChatGPT“代码解释器”（Code Interpreter）测试版正式向所有Plus用户开放，具体来说包括帮助分析数据、创建图表、编辑文件、执行数学运算等功能。

1、Open AI为什么用户放缓：扭亏是个关键

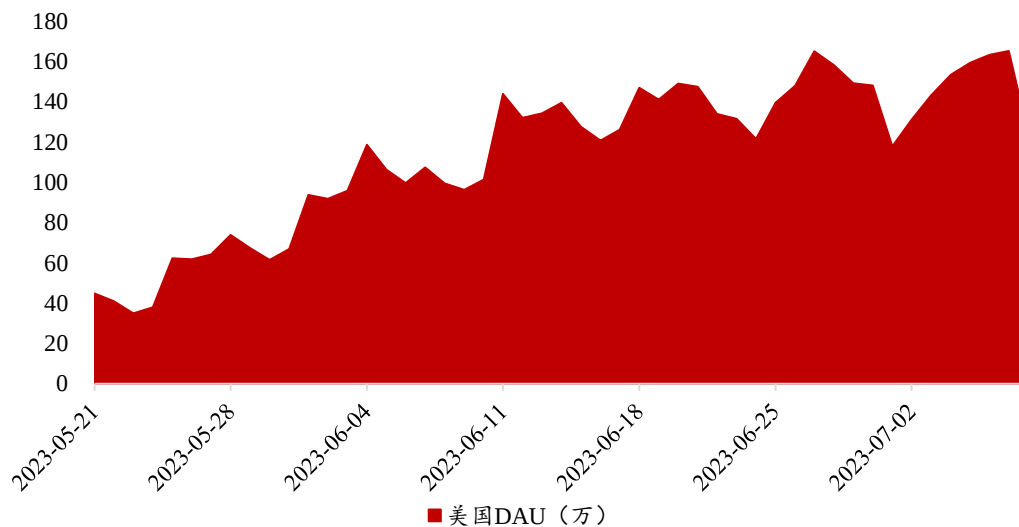
根据Similarweb的数据，截至2023年7月12日，ChatGPT网页端日访问量基本持平。2022年11月30日发布以来，ChatGPT用户数量持续上涨。据官方数据显示，ChatGPT上线2个月活跃用户突破一亿人次。而根据Similarweb网页访问数据，截至2023年7月12日，ChatGPT网页日访问量已基本持平，维持在五千多万。

据data.ai数据显示，ChatGPT iOS端加速扩张，每日活跃用户量呈增长趋势。2023年5月18日，Open AI正式发布ChatGPT iOS版本。根据App Annie的数据，截至2023年6月19日，Open AI ChatGPT iOS端美国地区前30日平均日活跃用户94.6万人。

图表：目前ChatGPT网页端日访问量基本持平



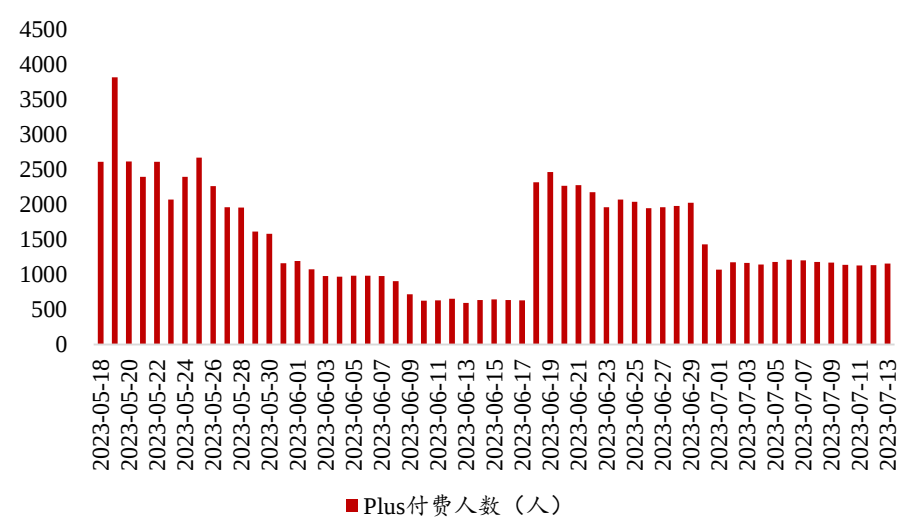
图表：美国地区ChatGPT iOS端日活增长趋势



2、提高收入：提高付费率

据data.ai数据显示，截至2023年6月19日，ChatGPT iOS端上线首月的日活付费率约为4.36%。根据data.ai的数据，截至2023年6月19日，ChatGPT iOS端美国地区累计下载约264.8万次，累计收入约100万美元。当前，GPT-4 PLUS的付费价格为20美元/月。根据data.ai，截至2023年6月19日，ChatGPT iOS端5月21日-6月19日美国地区平均日活用户量约94.64万人，累计付费用户数约为4.13万人。因此日活付费率（月付费用户数/日活跃用户）约为4.36%（4.13÷94.64）。据Questmobile数据，百度APP日活月活用户比例为37%左右，因此如果ChatGPT日活月活比例为37%时，月活跃人数约为255.6万（94.64÷37%），此时月活付费率（月付费用户数/月活跃用户）约为1.6%（4.13÷255.8）。

图表： ChatGPT iOS端付费人数



图表： 不同日活月活比例下的月活付费率

日活月活比例	月活付费率
20%	0.87%
25%	1.09%
30%	1.31%
35%	1.53%
40%	1.75%
45%	1.96%
50%	2.18%

2.1 Open AI总体营收分析：12%的付费率或为盈亏线

在GPT-3.5精度没有下降的情况下，假设在8次平均使用次数条件下，如果日活月活比例达到30%，月活付费率突破14%，对于未压缩情况下的GPT-3.5或能实现盈亏平衡。在每用户平均使用8次的条件下，Open AI全年整体算力成本约为65.5亿美元，此时处于亏损状态。而若日活月活比例为30%，付费率突破14%时，年收入可能达到67.2亿美元/年，或将实现盈亏平衡。

在GPT-3.5精度下降的情况下，如果日活月活比例达到35%，月活付费率突破12%，对于压缩后的GPT-3.5或能实现盈亏平衡。如果GPT-3.5的精度下降约5%，成本大约下降25%，在6000万DAU，每用户平均8次请求次数的条件下，Open AI整体算力成本约为47.4亿美元。如果日活月活比例达到35%，月活付费率突破12%，对于压缩后的GPT-3.5或能实现盈亏平衡。

图表： 不同付费率及日活月活比例条件下总体盈亏情况

付费率	总体用户收入（亿美元/年）				
	20%	25%	30%	35%	40%
2%	14.4	11.5	9.6	8.2	7.2
4%	28.8	23.0	19.2	16.5	14.4
6%	43.2	34.6	28.8	24.7	21.6
8%	57.6	46.1	38.4	32.9	28.8
10%	72.0	57.6	48.0	41.1	36.0
12%	86.4	69.1	57.6	49.4	43.2
14%	100.8	80.6	67.2	57.6	50.4

3、Open AI总算力成本：GPT-3.5是主要成本来源

假设每用户平均使用8次/天，未压缩的GPT-3.5/GPT-4在2500万DAU及高并发条件下分别需21.8/53.1万GPU_s。计算GPT-3.5的算力时，我们的假设条件为用户平均请求次数为10次，但根据实际的用户使用情况，平均每个活跃用户每天发送7~8次请求。若每用户平均请求8次，GPT-3.5和GPT-4模型在2500万日活用户及高并发条件下分别需要21.8万、53.1万GPU_s（27.2/10*8）。

假设ChatGPT共有6000万DAU，每用户平均8次请求次数的条件下，未压缩的GPT-3.5/GPT-4全年整体算力成本约为65.5亿美元。据财经网科技，截止3月15日，ChatGPT日活突破5837万。假设按照Open AI整体DAU6000万计算，根据月活付费率，ChatGPT的用户中，大多数用户仍为GPT-3.5的免费用户，只有不到2%为GPT-4的付费用户。若按单价每张卡每小时1.4美元计算，未压缩的GPT-3.5及GPT-4模型一年的总算力成本约为65.5亿美元
 （（21.8*98.5%+53.1*1.5%）*\$1.4*24hours*365days/2500*6000/10000）。

图表：每用户8次使用次数时不同模型所需的算力

平均请求次数（次）	所需GPU（万）	
	GPT-3.5	GPT-4
8	21.8	53.1
10	27.2	66.4

图表：整体算力成本与人均使用次数的敏感性分析

平均请求次数（次）	一年整体算力成本（亿美元）		
	GPT-3.5	GPT-4	总算力成本
6	47.4	1.8	49.1
7	55.3	2.1	57.3
8	63.2	2.3	65.5
9	71.1	2.6	73.7
10	79.0	2.9	81.9
11	86.8	3.2	90.1

3.1 GPT-3.5总成本：通过精度可控制算力成本

如果GPT-3.5的精度下降为原来的95.7%，成本大约下降25%，在6000万DAU，每用户平均8次请求次数的条件下， GPT-3.5整体算力成本约为45.8亿美元。根据Open AI官网消息，2023年6月13日起GPT-3.5成本减少25%，具体包括：text-embedding-ada-002嵌入模型的每1K tokens降低至0.0001美元，相当于原来价格的25%；gpt-3.5-turbo每1K输入tokens降低25%至0.0015美元、每1K输出tokens降低至0.002美元；gpt-3.5-turbo-16k新价格为每1K输入tokens价格为0.003美元，每1K输出tokens价格为0.004美元。此时，模型精度下降约5%，假设在6000万DAU，每用户平均8次请求次数的条件下， GPT-3.5整体算力成本约为45.8亿美元。

图表： GPT-3.5降低精度总成本减少

平均请求次数（次）	GPT-3.5总算力成本（亿美元/年）		
	精度下降1%	精度下降5%	精度下降10%
6	45.0	34.3	24.0
7	52.5	40.0	28.1
8	60.0	45.8	32.1
9	67.5	51.5	36.1
10	75.0	57.2	40.1
11	82.5	62.9	44.1

4、Open AI 资金预测：月付费率每月提升0.25%或不能持续

在Open AI的100亿美元资金的情况下，敏感性假设分析：如果在6000万DAU、付费率逐月提升0.25%的条件下，压缩成本后的GPT-3.5模型和GPT-4总体算力成本100亿美金或许可以亏到2026年1月。2019年，微软向ChatGPT的所有者Open AI投下10亿美元，2023年1月，微软再投100亿美元。根据Similarweb访问量数据，如果按压缩后的GPT-3.5模型，并以单价每张卡每小时1.4美元计算，2023年上半年Open AI 总算力成本大约可达到23.9亿美元。从2023年7月开始，若日活跃用户为6000万，以平均每用户使用8次，付费的GPT-4所占的比例为1.5%计算，压缩后的GPT-3.5和GPT-4模型每月整体的算力成本约为4.09亿美元。

如果按GPT-4 Plus用户付费率约2%以及6000万日活跃用户计算，成本压缩25%的GPT-3.5及GPT-4所需算力成本约4.13亿美元/月。若OpenAI的月活付费率每月提高0.25%，其100亿美金或许可以亏到2026年1月。

图表：月活付费率每月提升0.25%

时间	付费率	Open AI算力成本 (亿美元/月)	Open AI营收 (亿美元/月)
2023年7月	2.00%	4.13	0.69
2023年8月	2.25%	4.16	0.77
2023年9月	2.50%	4.18	0.86
2023年10月	2.75%	4.20	0.94
2023年11月	3.00%	4.22	1.03
2023年12月	3.25%	4.24	1.11
2024年1月	3.50%	4.27	1.20
2024年2月	3.75%	4.29	1.29
2024年3月	4.00%	4.31	1.37
2024年4月	4.25%	4.33	1.46
2024年5月	4.50%	4.36	1.54
2024年6月	4.75%	4.38	1.63
2024年7月	5.00%	4.40	1.71
2024年8月	5.25%	4.42	1.80
2024年9月	5.50%	4.44	1.89
2024年10月	5.75%	4.47	1.97
2024年11月	6.00%	4.49	2.06
2024年12月	6.25%	4.51	2.14
2025年1月	6.50%	4.53	2.23
2025年2月	6.75%	4.56	2.31
2025年3月	7.00%	4.58	2.40
2025年4月	7.25%	4.60	2.49
2025年5月	7.50%	4.62	2.57
2025年6月	7.75%	4.64	2.66
2025年7月	8.00%	4.67	2.74
2025年8月	8.25%	4.69	2.83
2025年9月	8.50%	4.71	2.91
2025年10月	8.75%	4.73	3.00
2025年11月	9.00%	4.76	3.09
2025年12月	9.25%	4.78	3.17
2026年1月	9.50%	4.80	3.26
总计		162.3	63.52

4.1 Open AI 资金预测：月付费率每月提升0.5%或能扭亏

在Open AI的100亿美元资金的情况下，敏感性假设分析：如果在6000万DAU、付费率逐月提升0.5%的条件下，压缩成本后的GPT-3.5模型和GPT-4总体可能在2025年11月转亏为盈。根据Similarweb访问量数据，如果按压缩后的GPT-3.5模型，单价每张卡每小时1.4美元，从2023年7月开始，如果按GPT-4 Plus用户付费率约2%以及6000万日活跃用户计算，若OpenAI的月活付费率每月提高0.5%，OpenAI每月营收到2025年11月大约可以达到5.49亿美元/月，成本压缩25%的GPT-3.5及GPT-4所需算力的成本为5.38美元/月，月营收超过成本，实现扭亏。

图表：月活付费率每月提升0.5%

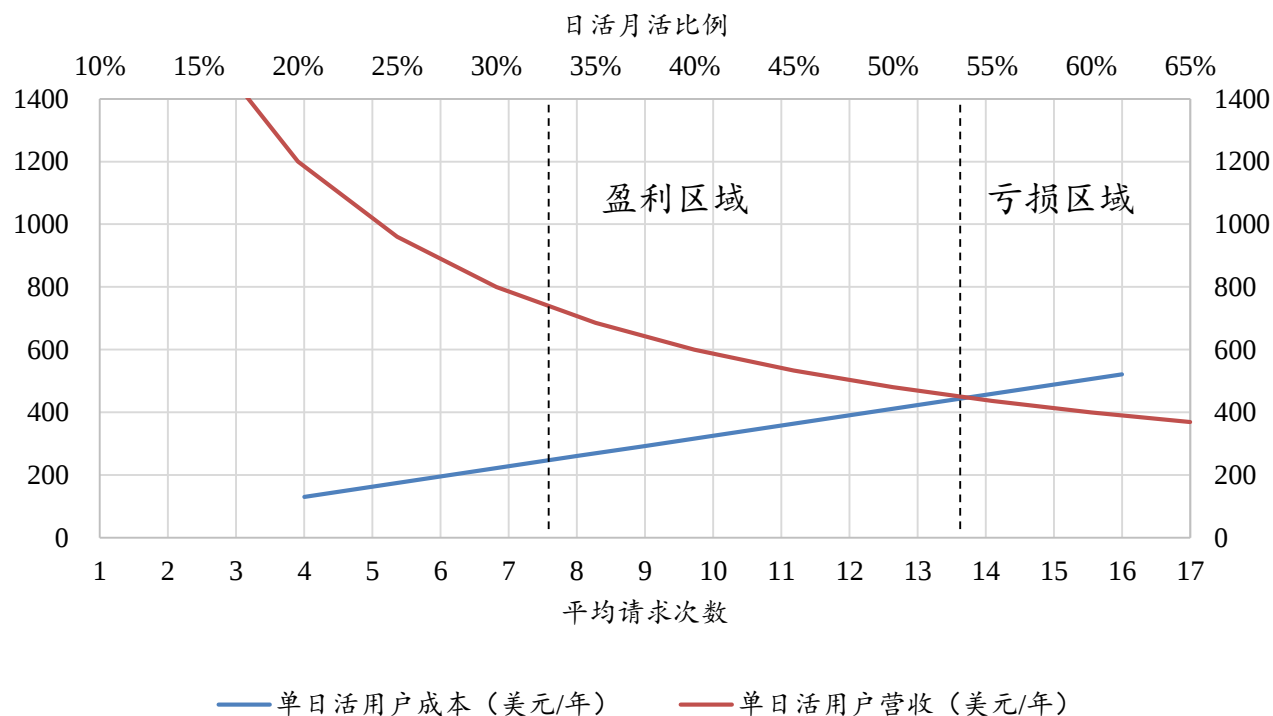
时间	付费率	Open AI算力成本 (亿美元/月)	Open AI营收 (亿美元/月)
2023年7月	2.00%	4.13	0.69
2023年8月	2.50%	4.18	0.86
2023年9月	3.00%	4.22	1.03
2023年10月	3.50%	4.27	1.20
2023年11月	4.00%	4.31	1.37
2023年12月	4.50%	4.36	1.54
2024年1月	5.00%	4.40	1.71
2024年2月	5.50%	4.44	1.89
2024年3月	6.00%	4.49	2.06
2024年4月	6.50%	4.53	2.23
2024年5月	7.00%	4.58	2.40
2024年6月	7.50%	4.62	2.57
2024年7月	8.00%	4.67	2.74
2024年8月	8.50%	4.71	2.91
2024年9月	9.00%	4.76	3.09
2024年10月	9.50%	4.80	3.26
2024年11月	10.00%	4.84	3.43
2024年12月	10.50%	4.89	3.60
2025年1月	11.00%	4.93	3.77
2025年2月	11.50%	4.98	3.94
2025年3月	12.00%	5.02	4.11
2025年4月	12.50%	5.07	4.29
2025年5月	13.00%	5.11	4.46
2025年6月	13.50%	5.16	4.63
2025年7月	14.00%	5.20	4.80
2025年8月	14.50%	5.24	4.97
2025年9月	15.00%	5.29	5.14
2025年10月	15.50%	5.33	5.31
2025年11月	16.00%	5.38	5.49
2025年12月	16.50%	5.42	5.66
2026年1月	17.00%	5.47	5.83

三、Open AI “单用户” 盈亏分析

1、单用户盈亏平衡分析

如果按GPT-4单用户每日吞吐量约7-8次，日活月活比例大约为30%-35%，处于盈利状态。
如果日活月活比例达到50%-55%，则当用户使用次数大于14次时，可能会出现亏损。

图表：Open AI PC端GPT-4单用户盈亏平衡分析



2、GPT-4日活月活比例：使用与成本的权衡

行业日活月活比例情况：根据Questmobile的数据，百度搜索引擎的日活跃用户与月活跃用户比例大约在37%左右，以此作为参考，行业日活月活比例应在这个数值附近。如果以日活月活比例37%为参考，那么GPT-4单日活用户营收大约每年649美元（ $20 \div 37\% \times 12$ ）。

日活/月活比例与成本的权衡：日活跃用户与月活跃用户比例用于衡量产品的用户粘性，比率越大粘性越好。如果在统计周期内，ChatGPT用户使用产品的频率越大，日活月活比例越高，产品使用年限越长，此时面临的问题是成本也越高。

图表： 移动端不同日活月活比例下的单日活用户营收

日活月活比例 (移动端)	单日活用户营收 (美元/天)	单日活用户营收 (美元/年)
35%	1.88	685.7
36%	1.83	666.7
37%	1.78	648.6
38%	1.73	631.6
39%	1.69	615.4
40%	1.64	600.0

3、GPT-4平均次数：通过控制“高频用户”限制平均次数

GPT-4访问限制推出后，“高频用户”的使用受限，平均请求次数可能下降。

由于GPT-4所需的算力消耗远高于GPT-3.5，为了控制算力成本或避免滥用，OpenAI对于高频次数的GPT-4用户的使用进行了限制。据消息，OpenAI对GPT-4的限制从每4小时150条消息到每4小时100条消息、每3小时50条消息到最新的每3小时25次，一个月内连续下降了4次访问阈值。通过对“高频用户”的使用限制，整体平均请求次数可能下降。

图表： GPT-4加大使用限制后用户平均使用次数下降

GPT-4使用限制	限制平均使用次数/小时
无限制	/
150次/4小时	38
100次/4小时	25
50次/3小时	17
25次/3小时	8

四、如何提高GPT-4付费率

1、ChatGPT用户构成：程序员、学生/教师、设计师

1、ChatGPT使用用户画像：在ChatGPT用户中，主要使用者为程序员；其次还包括学生和教师用户、设计师/创作者用户等；此外，还有社会科学家、游戏玩家、投资者等多类用户。对于程序员来说，ChatGPT可以帮助他们调试、生成代码、结构设计、修复代码错误；对于学生和教师群体来说，GPT-4可以切换不同“身份”进行智能对话、撰写文章、回答问题等，为学生使用群体提供了方便；对于设计创作者来说，ChatGPT可以用于生成创意和灵感，协助设计师进行创意设计。

2、GPT-4的多模态功能和Code Interpreter为用户提供工作帮助，促进用户付费。2023年7月11日，Open AI推出代码解释器（Code interpreter）功能，无论是代码编写还是数据分析、图表生成、执行数学运算，甚至是生成视频、分析股票市场，都能得到高质量的结果。从编程的角度来看，即使不是程序员，只需要用自然语言向ChatGPT下达指令，也可以完成需要复杂编程技术的任务。从设计领域来看，GPT-4的多模态功能可以输入图片，Code Interpreter可以进行GIF 动画渲染以及基本视频编辑，促进了创作型用户对GPT-4的付费。

2、GPT-3.5与GPT-4对比：不同用户视角分析

GPT-4在多项测试中表现优于GPT-3.5。Open AI通过各类benchmark进行测试，GPT-4在多项测试上表现优于GPT-3.5。相比于GPT-3.5，GPT-4在物理、生物、数学等理科测试中的表现明显优于GPT-3.5，模型展现了处理复杂推理任务能力的大幅迭代。

程序员视角分析：在编码和软件开发方面，GPT-4可以使用自然语言生成代码，自动化编程过程，提高编码效率，其生成效率明显高于GPT-3.5；此外，GPT-4可以帮助程序员解决代码问题，通过指出问题并提供修改建议来提高代码质量，帮助程序员避免常见错误。

学生/教师视角分析：GPT-4可以切换不同“身份”进行智能对话、撰写文章、回答问题等，GPT-4对于有大量搜索及提问需求的学生或教师群体来说，其作用要大于GPT-3.5。

设计师/创作者视角分析：ChatGPT可以用于生成创意和灵感，协助设计师进行创意设计，GPT-4的多模态功能可以输入图片，较GPT-3.5更为便利，GPT-4新功能Code Interpreter可以进行GIF 动画渲染以及基本视频编辑，为创作者提供帮助。

图表： GPT-3.5与GPT-4模型对比

名称	模型类型	模式	可接收文字输入长度
GPT-3.5	自然语言处理模型	仅文本	3000字
GPT-4	多模态模型	文本和图像	25000字

3、ChatGPT推出代码解释器功能，推动付费用户增加

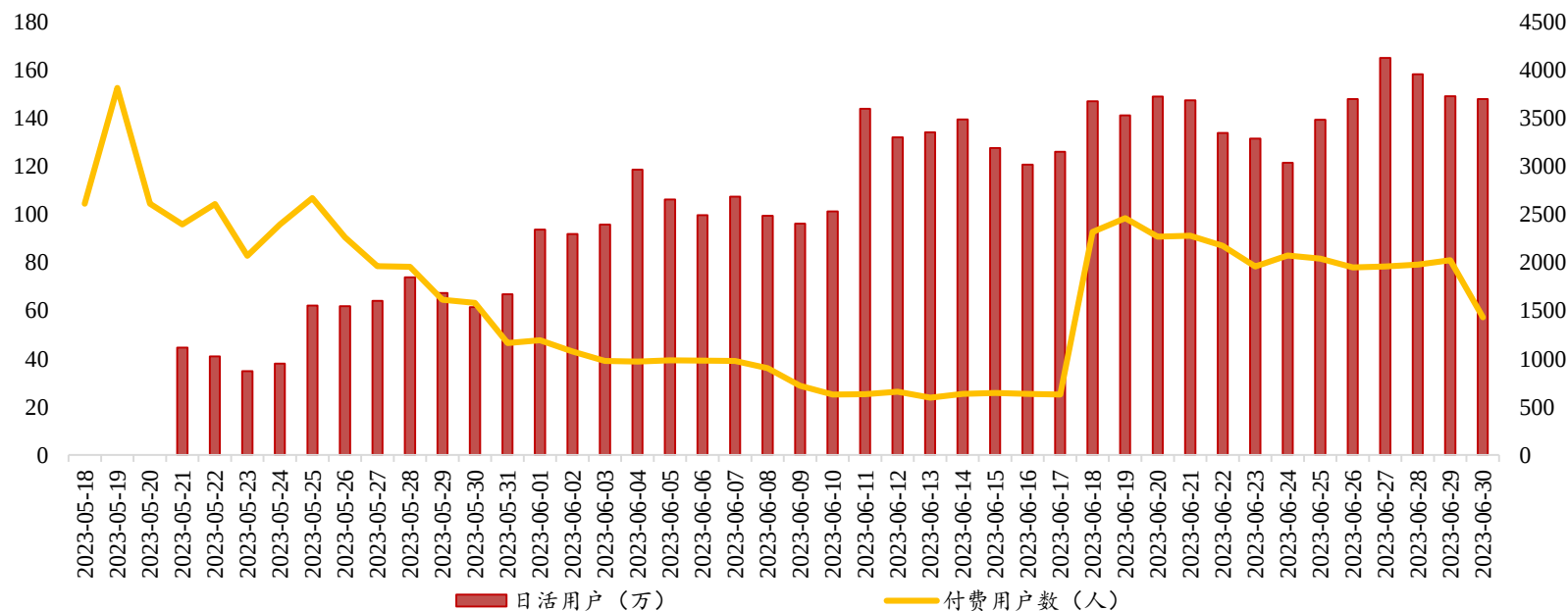
GPT-4独特的第三方插件及“代码解释器”功能，推动用户付费。

2023年5月，Open AI全面开放了GPT-4第三方插件功能（plugins），这些功能仅向Plus用户开放，促进了用户的付费；

2023年7月11日，ChatGPT“代码解释器”（Code Interpreter）测试版正式向所有Plus用户开放，具体来说包括帮助分析数据、创建图表、编辑文件、执行数学运算等功能，代码解释器扩展了ChatGPT的功能，为用户带来更好的交互式编程体验和强大的数据可视化功能，可能推动用户付费。

根据data.ai数据，截至2023年6月17日，ChatGPT iOS端前30天的平均付费用户数约1469人，平均每日活跃用户91万人。

图表： ChatGPT iOS端日活及付费用户



五、投资建议及风险提示

投资建议：

- 1) 若Open AI实现盈利，或可刺激所有海外巨头加大资本开支，建议关注算力相关公司。
- 2) 在Open AI收费的带动下，或有利于用户养成付费意识，各类AI相关应用有望百花齐放。建议关注：“AI+游戏”宝通科技、凯撒文化、恺英网络、巨人网络、神州泰岳、吉比特、完美世界；“AI+角色”奥飞娱乐、芒果超媒、掌阅科技；“AI+图像”美图；“AI+营销”三人行、蓝色光标、捷成股份等。
- 3) 若Open AI实现盈利，则其他公司在大模型的投入或将加大，“数据要素”的地位进一步凸显。建议关注：人民网、新华网。

风险提示：

Open AI营收平衡效果不及预期，用户付费意愿不及预期，海外大厂资本开支有限。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，保证报告所采用的数据和信息均来自公开合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法存在局限性。特此声明。

免责声明

本研究报告由方正证券制作及在中国（香港和澳门特别行政区、台湾省除外）发布。根据《证券期货投资者适当性管理办法》，本报告内容仅供我公司适当性评级为C3及以上等级的投资者使用，本公司不会因接收人收到本报告而视其为本公司的当然客户。若您并非前述等级的投资者，为保证服务质量、控制风险，请勿订阅本报告中的信息，本资料难以设置访问权限，若给您造成不便，敬请谅解。

在任何情况下，本报告的内容不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求，方正证券不对任何人因使用本报告所载任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

本报告版权仅为方正证券所有，本公司对本报告保留一切法律权利。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。如需引用、刊发或转载本报告，需注明出处且不得进行任何有悖原意的引用、删节和修改。

类别	评级	说明
公司评级	强烈推荐	分析师预测未来12个月内相对同期基准指数有20%以上的涨幅。
	推荐	分析师预测未来12个月内相对同期基准指数有10%以上的涨幅。
	中性	分析师预测未来12个月内相对同期基准指数在-10%和10%之间波动。
	减持	分析师预测未来12个月内相对同期基准指数有10%以上的跌幅。
行业评级	推荐	分析师预测未来12个月内行业表现强于同期基准指数。
	中性	分析师预测未来12个月内行业表现与同期基准指数持平。
	减持	分析师预测未来12个月内行业表现弱于同期基准指数。
基准指数说明		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指数为基。



方正证券研究所

专注 · 专心 · 专业

北京市西城区展览路48号新联写字楼6层
长沙市天心区湘江中路二段36号华远国际中心37层

上海市静安区延平路71号延平大厦2楼
深圳市福田区竹子林紫竹七道光大银行大厦31层

广州市天河区兴盛路12号楼隼峰院2期3层方正证券