人工智能行业点评

OpenAI 访问限流，GPT-4 算力测算

超配

◆ 公司研究·公司快评 ◆ 计算机 ◆ 投资评级：超配（维持评级）

证券分析师： 熊莉 021-61761067 xiongli1@guosen.com.cn 执证编码：S0980519030002
证券分析师： 朱松 021-60875155 zhusong@guosen.com.cn 执证编码：S0980520070001

事项：

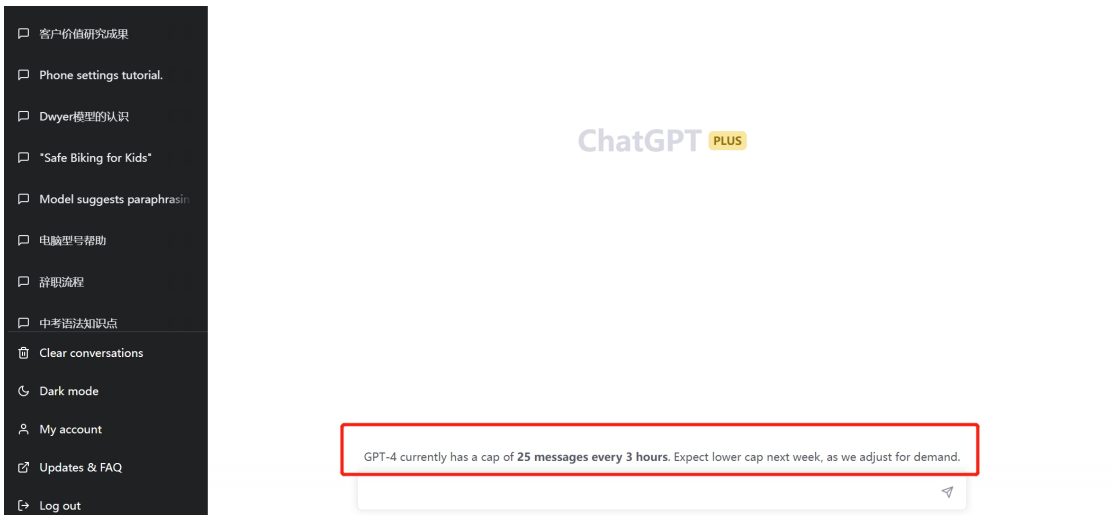
OpenAI 对于 Plus 付费用户的 GPT-4 访问连续下降阈值，GPT-4 访问限制由第一天的 150 msg/ 4 hr 到 100 msg/4 hr 到 50msg/3 hr 到最近的 25msg/3hr。在较短的时间内 GPT-4 下降了 4 次访问阈值。

评论：

◆ ChatGPT-4 的访问被持续限流

OpenAI 对于 Plus 付费用户的 GPT-4 访问连续下降阈值，GPT-4 访问限制由第一天的 150 msg/ 4 hr 到 100 msg/4 hr 到 50msg/3 hr 到最近的 25msg/3hr。在较短的时间内 GPT-4 下降了 4 次访问阈值。对于 GPT-4 的限流，OpenAI 的官方回复是“开发者可能会对使用 GPT-4 的应用程序或者服务施加类似限制，以确保合理的资源分配，避免滥用或者控制成本。通常这类限制取决于 API 访问限制或者特定的订阅服务”。官方主要从控制成本的角度对 GPT-4 访问量进行优化，也明确了限流几乎只针对 GPT-4。

图1：ChatGPT-4 访问被限流



资料来源：OpenAI 官网，国信证券经济研究所整理

◆ OpenAI 限流的背后是其日活和周活用户数的持续攀升

由于功能的强大以及回复率准确的高企，ChatGPT 的日活和周日跃用户数量屡创新高，截止 3 月 15 日，ChatGPT 日活突破 5837 万(3 月 13 日的日活用户数据是 4846 万，两天新增接近 800 万日活），截止到 3 月 9 日这周，ChatGPT 周度的活跃用户数量为 331.81 万，相比上一周活跃用户数量增加了接近 25 万。GPT 的日活和周度活跃用户数量持续攀升，一方面是 GPT4 的能力更加强大，使用的体验更加流畅、回复的答

案更加精准、上下文溯源的能力更加强大，使得越来越多的用户加入到体验和运用模型的队伍中来，另一方面是因为经过此前 GPT-3 经历后的付费用户数量越来越多。目前，OpenAI 没有提价，对于 Plus 会员的价格还是 20 美金/月，但是随着用户数量的持续增加，预计限流的动作会持续。

◆ **OpenAI 的需求持续增长将持续推高算力成本**

ChatGPT 应对终端访问，每 1000 tokens 需要算力成本 0.02 美金，3 月 3 日 OpenAI 发布 GPT-3.5-turbo ，对外开放 API 接口，宣称算力成本下降 90%，1000 tokens 算力成本下降到 0.002 美金。目前，以英伟达 DGX A100 服务器作为计算资源，采用云服务单天成本约为 460 美元，假设按照 ChatGPT 日活 1300 万人，每个人每天平均 1000 字的问题（共计 173.3 亿个 token），假设峰值为一天均值的 5 倍，那么需要 602 万台 DGX A100 服务器，每天的租用成本为 27.77 万美金，平均到每 1000 个 token 的推理成本为 0.02 美金左右（详细的测算过程欢迎参考前期报告《人工智能行业点评：ChatGPT 对算力的需求究竟如何？》）。目前，GPT-3.5-turbo 的算力成本下降了 90%达到 0.002 美金每 1000 个 tokens，我们预期模型的参数有所下调（此前测算成本按照 GPT-3.5 的模型参数 3000 亿个进行测算），此结果和用户对于 new bing 不及 Chat GPT 本身交互准确的感受相一致。

预计 GPT-4 的算力消耗量远大于 GPT-3 和 GPT-3.5-turbo。目前根据根据 GPT-4 的公开数据，在 8K 的上下文长度下，每 1000 个 token 的提问成本为 0.03 美金，每 1000 个 token 的回答完成成本为 0.06 美金；在 32K 的上下文长度下，每 1000 个 token 的提问成本为 0.06 美金，每 1000 个 token 的回答完成成本为 0.12 美金。这个算力成本相比与 GPT-3 的成本（每 1000 个 tokens 的算力成本约为 0.02 美金）上升了较高（输入成本增加 50%-200%，输出升本增加 200%-500%），相比于 GPT-3.5-turbo 的成本上升更为可观，输入成本增长了 14-29 倍，输出成本增长了 29-59 倍。

图2：ChatGPT-4 访问价格

Model	Prompt	Completion
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

资料来源：OpenAI 官网，国信证券经济研究所整理

预期在应用逐步增长的背景下，GPT 的算力成本将进一步攀升。目前 OpenAI 主要的用户为 C 端用户，未来有望逐步扩大到 B 端领域，GPT-3.5-turbo 下降精度降低算力成本以适应更多的应用场景是扩大生态中坚实的一步。随着 GPT 生态的建立、相关应用的爆发，算力的需求将持续扩大。目前，根据微软的数据，Bing 的日活首次突破 1 亿（集成搜索+聊天功能的必应预览版自推出以来总聊天次数已超过 4500 万次），目前 GPT 的日活数量接近 5800 万，Bing 日活 1 亿用户，假设 NewBing4000 万日活，其他应用 4000 万日活，合计 1.4 亿日活。假设普通 98.5%的用户使用 GPT-3.5-turbo，1.5%的用户使用 GPT-4，那么 98.5%普通用户的一天的算力成本在 28 万美金左右（对应 607 台 DGX A100 系列服务器），而 1.5%付费使用 GPT-4 的用户的算力成本为每天 840 万美金（对应 1.8 万台 DGX A100 系列服务器）。

具体测算和假设如下：1）98.5%的用户使用 GPT-3.5-turbo，每个人每天问 1000 tokens，按照每 1000 tokens 需要算力成本 0.002 美金计算，每天的成本大概为 28 万美金，对应 607 台 DGX A100 系列服务器；2）1.5%付费用户使用 GPT-4，每个人每天 1 万个 tokens，按照目前 GPT-4 给出的最低价格（每 1000 个 tokens 0.06 美金），那么算力成本为约 840 万美金，对应 1.8 万台 DGX A100 系列服务器。而根据微软在 2019 年的公告，投资 OpenAI 10 亿美金，投资 1 万个 GPU 建设 Azure AI 超算平台，可以预见的是，算力将很快被 OpenAI 的 GPT 模型消耗完，这也是 OpenAI 此前持续公布限流的原因之一。

随着 OpenAI 官方公告 6.14 号开出 GPT-4 的 API 接口，按照目前用户数量和问询数量的趋势，公司必须做

出一些应对措施，要想保留 GPT-4 强大和精准的能力，就需要继续扩大算力，或者继续降低精度以控制算力成本。在全球的 AI 浪潮下，国内自从百度文心一言大模型推出后，也有几百家的公司成为其合作伙伴，未来阿里和腾讯的大模型也有望推出，对应的算力缺口有望持续扩大，从国内的浪潮信息、中科曙光、工业富联、海光信息、紫光股份等公司下游需求来看已经有些端倪。

◆ **投资建议：**

目前，全球正处于 AI 浪潮中，海外以 OpenAI 旗下的 GPT 大模型为代表，越来越多的 C 端用户和 B 端厂商接入生态，在 AI 赋能后，产品和应用端都出现了较为明显的积极变化。微软已经率先将 GPT-4 接入了 Office 等全系列办公场景，预期后续会有更多的现象级应用出现，国内市场也受益于百度文心一言大模型的发布，未来有望在应用端迎来百花齐放。随着应用端逐渐丰富，对算力的需求提出了更多的需求，在此趋势下，预计未来的算力需求缺口将会持续扩大，建议重点关注人工智能算力领域的相关标的，如海光信息等。

◆ **风险提示：**

模型假设不合理对测算结果造成偏差，ChatGPT 商业化落地不及预期，算力芯片进展不及预期。

相关研究报告：

《人工智能行业点评-Microsoft 365 Copilot 发布，国内外 AI 应用有望加速落地》 ——2023-03-19
《国央企 ERP 专题报告：数字化转型下的新机遇》 ——2023-03-10
《计算机行业 2023 年 3 月投资策略暨年报前瞻-2022 行业业绩承压，关注 ChatGPT 引发的 AI+应用表现》——2023-03-06
《人工智能行业点评-ChatGPT 开放 API 接口，应用侧有望迎来全面爆发》 ——2023-03-05
《人工智能行业点评-2017 年以来人工智能行情复盘》 ——2023-02-17

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

类别	级别	说明
股票 投资评级	买入	股价表现优于市场指数 20%以上
	增持	股价表现优于市场指数 10%-20%之间
	中性	股价表现介于市场指数 ±10%之间
	卖出	股价表现弱于市场指数 10%以上
行业 投资评级	超配	行业指数表现优于市场指数 10%以上
	中性	行业指数表现介于市场指数 ±10%之间
	低配	行业指数表现弱于市场指数 10%以上

重要声明

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032