# 人工智能行业点评

## ChatGPT 对算力的需求究竟如何？

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
联系人： 黄浩峻 0755-81981812 huanghaojun@guosen.com.cn

## 事项：

2022 年 11 月 30 日人工智能实验室 OpenAI 发布全新聊天机器人模型 ChatGPT，其是一款人工智能技术驱动的自然语言处理工具。自从 ChatGPT 推出以来，受到市场广泛关注，当前每日用户超过 1000 万人。

## 评论：

**Chatgpt 成本主要可以拆分成训练和推理两个阶段。**人工智能主要分为计算机视觉和自然语言处理两大基础方向，当前自然语言处理类任务基于大型语言模型（LLM，Large Language Model）演进出了最主流的两个主要方向，BERT（基于 Transformer 的双向编码器表示技术）和 GPT（基于 Transformer 生成预训练技术），Google 属于 BERT 技术方向，微软投资的 OpenAI 属于 GPT 技术方向。从计算过程上，人工智能计算主要可以分为模型训练与推理两个阶段，针对大语言模型 LLM 更是如此，随着参数与数据规模的不断增大，将带动拉动算力需求的快速增长。

- **“token”是当前语言类模型的数据单位。**当前的自回归语言模型是根据 token 来作为单位进行数据处理和计算，分词（tokenization）就是将句子、段落、文章这类型的长文本分解为以 token 为单位的数据结构，把文本分词后每个词表示成向量进行模型计算。例如在英文语境下，“happy”可能被分解为“hap”、“-py”两个 token，中文语境下，“我很开心”可以分成“我”，“很”，“开心”三个 token。

- **以英伟达 DGX A100 服务器作为计算资源：（1）单台服务器售价 20 万美元；（2）采用云服务单天成本约为 460 美元。**根据英伟达官网数据，英伟达超算 GPU 系列从旧到新包括 P100、V100、A100、H100 等，其中，DGX A100 系列服务器为 2020 年发布，是当前主流使用的超算服务器，单机有 8 个 A100 系列 GPU，AI 算力性能为 5 PetaFLOP/s，单机最大功率 6.5kw，售价 19.9 万美元；如果租用云服务，根据亚马逊数据显示，在亚马逊 AWS 预定一年的 A100 系列 GPU，有 8 个 A100 的 AWS P4 实例的平均成本约 19.22 美元，一天的平均成本约为 461.28 美元。

- **ChatGPT 上一个 30 字的问题需要消耗计算资源 0.12 PetaFLOP/S。**最常见的 Transformer 类语言模型在推理过程中每个 token 的计算成本（以 FLOPs 为指标）约为 2N，其中 N 为模型参数数量（20 年发布的 GPT-3 拥有 1750 亿参数，22 年谷歌发布的 PaLM 拥有 5400 亿参数，由于并未公布当前 GPT3.5 的参数数量，当前假定参数数量为 3000 亿），假设模型的 FLOPS 利用率约为 20%，**粗略估计 ChatGPT 一个 30 字（假设约 40 个 token，注：在英文语境下，一般 1000 个 token=750 个单词）问题需要的算力资源为 2*40*3000 亿/20%=0.12 PetaFLOP/S。**

- **推理成本：为满足当前用户访问产生的推理成本，自建 IDC 初始投入约在 4 亿美元，租用云服务每日成本约 28 万美元。**根据 Similarweb 的数据，23 年 1 月份当前 ChatGPT 日活约 1300 万人，每人平均 1000 字左右的问题，因此合计产生约 130 亿字（173.3 亿个 token），假设 24 小时平均分配任务，需要的 A100 GPU 数量为 173.3 亿*2*3000 亿/（20%*24 小时*3600 秒）=601.75 PetaFLOP/S，由于访问流量存在峰值，假定访问峰值是一天均值的 5 倍，因此**共需要 602 台 DGX A100 服务器能够满足当前的访问量。**

(1) **自建 IDC：**服务器成本约占数据中心成本 30%左右，为满足当前日常访问需求，**前期一次性成本投入约为 602*19.9/30%=3.99 亿美元；**

(2) **云服务：**假设每天租用亚马逊 AWS 云服务，**每天成本为 461.28*602=27.77 万美元。**

- **训练成本：**训练阶段每个 Token 的训练成本约为 6N（推理成本为 2N），由于每年训练成本都在快速下降，此处引用 OneFlow 的测算结果，在公有云中训练 OPEN AI 的 GPT-3 模型需花费训练成本约 140 万美元，Google 的 PaLM 模型需花费训练成本约 1120 万美元。
- **预计在 ChatGPT 结合 Bing 搜索功能后，其对算力资源的消耗将成数倍增长。**当前 ChatGPT 模型可以理解为在一个在庞大训练数据集上训练的 LLM，它会将训练期间的知识存储到模型参数中。在推理过程中（使用模型生成输出），LLM 无法访问外部知识，仅依靠模型参数进行计算；如果将 ChatGPT 与搜索功能结合，如 Bing 等搜索引擎，其计算过程将通过搜索引擎返回多个查询结果，并通过 GPT 计算生成多个响应，在返回最高分的响应给用户，其对算力资源的消耗将成数倍增长，增长倍数取决于搜索和响应的个数。

## ◆ 投资建议：

当前处在以 ChatGPT 为主线的新一轮人工智能创新周期，ChatGPT 为人工智能产业注入新活力，有望带动 AIGC 类应用快速爆发，人工智能技术作为驱动数据经济的技术底层，有望迎来新的发展机遇。数据、算力与算法是人工智能时代的三大基石，三者相互促进带动 AI+应用快速落地，ChatGPT 为首的自然语言处理类技术及应用，有望迎来全面爆发，建议重点关注人工智能相关赛道。

## ◆ 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。

## 相关研究报告：

《信息安全深度剖析 5：密评和信创双催化，密码产业开启从 1 到 N》 ——2023-02-13
《计算机行业 2023 年 2 月投资策略-人工智能赋能产业升级，把握数字经济时代浪潮》 ——2023-02-05
《计算机行业 2023 年 1 月投资策略-紧抓高景气赛道，关注业绩高增品种》 ——2023-01-02
《计算机行业 12 月暨 2023 年投资策略-以信创和安全为基，数据要素驱动数字经济大发展》 ——2022-12-05
《大数据系列专题（2）：国产数据库百花齐放，崛起正当时》 ——2022-11-20

# 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<br>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<br>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