2023 年 03 月 06 日

计算机行业研究

行业专题研究报告

证券研究报告

计算机组

分析师：孟灿（执业 S1130522050001）

mengcan@gjzq.com.cn

分析师：王倩雯（执业 S1130522080001）

wangqianwen@gjzq.com.cn

ChatGPT 训练及多场景推理成本测算

行业观点

- 2023 年 3 月 1 日，OpenAI 宣布开发者可通过 API 将 ChatGPT 和 Wisper 模型集成到他们的应用程序和产品中。本次 ChatGPT API 接入的模型名为 GPT-3.5-turbo，只需 0.2 美分/千 Tokens。本文尝试测算训练和不同推理场景之下的实际成本；未来随着模型压缩的持续发展，推理成本可能进一步降低，也有望大幅推动生成式模型在各个场景的大规模商用。我们的成本估算思路是：计算 AI 模型在进行训练与推理时所需的浮点运算次数，与 AI 算力集群的平均算力（以每秒浮点运算次数计）做比，以此估算 AI 模型的训练成本与推理成本。

- 在通用大模型训练方面，经测算，使用云计算时 ChatGPT 的训练成本约为 170 万美元，若自建 AI 算力中心进行模型训练，训练成本有望降至约 51 万美元；在大模型推理方面，使用云计算时的 ChatGPT 每处理 1,000Tokens 信息需花费约 0.177 美分，自建 AI 算力中心有望将成本降至 0.053 美分。

- 不同应用场景中 AI 模型面临的任务复杂度有所不同，我们考虑模型缓存命中率、计算集群闲置率、模型压缩等因素，综合估算各典型场景下模型的推理成本：1）搜索引擎场景中，以新版 Bing 为例，完成一次搜索的成本约为 1.73 美分；2）办公软件融合 ChatGPT 后可支持文字生成、文字修改等功能，完成一次用户需求的成本约为 1.70 美分；3）AI 客服作为对话场景应用，有望率先实现 B 端落地。经测算，AI 客服场景解决一次用户需求的成本约为 0.08 美分。

- 未来随着模型压缩技术的持续发展，推理成本可能进一步降低，也有望大幅推动生成式模型在各个场景的大规模商用。

投资逻辑

- 我们认为有海外场景的公司有可能率先与 GPT-3.5-turbo 进行对接，建议关注福昕软件、万兴科技等海外营收占比较高的应用公司。

风险提示

海外基础软硬件使用受限；应用落地不及预期；行业竞争加剧风险

内容目录

1. AI 模型训练成本估算 3
2. 通用 AI 模型推理成本估算 4
3. 多应用场景下 ChatGPT 推理成本估算 4
3.1 搜索引擎场景 5
3.2 办公软件场景 6
3.3 AI 客服场景 6
4. 投资建议 7
5. 风险提示 7

图表目录

图表 1: AI 模型训练成本估算 3
图表 2: AI 模型训练所需浮点运算次数计算与模型参数规模正相关 3
图表 3: 前沿 AI 模型的有效算力比率 3
图表 4: 考虑有效算力比率后的训练所需浮点运算次数计算 4
图表 5: AI 模型推理成本估算 4
图表 6: AI 模型推理所需浮点运算次数计算 4
图表 7: 具体应用场景中 AI 模型完成一次任务所需成本 4
图表 8: 新版 Bing 搜索引擎的运作方式 5
图表 9: Bing 搜索引擎场景任务的所需 Tokens 5
图表 10: 办公软件场景任务的所需 Tokens 6
图表 11: AI 客服场景任务的所需 Tokens 6

2023 年 3 月 1 日，OpenAI 宣布开发者可通过 API 将 ChatGPT 和 Wisper 模型集成到他们的应用程序和产品中。本次 ChatGPT API 接入的模型名为 GPT-3.5-turbo，是许多非聊天用例的最佳模型，且只需 0.2 美分/千 Tokens。本文尝试测算训练和不同推理场景之下的实际成本；未来随着模型压缩的持续发展，推理成本可能进一步降低，也有望大幅推动生成式模型在各个场景的大规模商用。

AI 模型的成本主要由训练成本和推理成本构成：

- 训练成本：衡量从头开发一个 AI 模型的算力费用、或是对现有 AI 模型知识库进行迭代更新所需的算力费用。
- 推理成本：衡量用户使用 AI 模型时产生的算力费用。

我们的成本估算思路是：计算 AI 模型在进行训练与推理时所需的浮点运算次数，与 AI 算力集群的平均算力（以每秒浮点运算次数计）做比，以此估算 AI 模型的训练成本与推理成本。

1. AI 模型训练成本估算

我们以 ChatGPT 为例，采用以下公式估算 AI 模型的训练成本：

图表1：AI 模型训练成本估算

$$\text{训练成本} = \frac{\text{AI模型训练所需的浮点运算次数}}{\text{AI算力集群单位时间有效浮点运算次数}} \times \text{AI算力集群单位时间价格}$$

来源：量子位公众号，国金证券研究所

- 训练所需的浮点运算次数：根据 OpenAI 于 2020 年发表的相关研究，对于 GPT-3 等采用 Decoder 结构 Transformer 骨干网络的大型语言模型而言，其训练所需的浮点运算次数可以遵照下述公式：

图表2：AI 模型训练所需浮点运算次数计算与模型参数规模正相关

$$\text{AI模型训练所需的浮点运算次数} = 6 \times \text{模型参数规模} \times \text{训练集大小}$$

来源：《Scaling Laws for Neural Language Models》（Jared Kaplan 等，2020），国金证券研究所

ChatGPT 采用 Decoder 结构 Transformer 骨干网络，虽然他的参数量包含 1.3B、6B、175B 这几种，但目前开放 API 接口的 GPT-3.5-turbo 与 Instruct Davinci 相近，因而采用 1,750 亿参数规模进行计算。

在训练集大小上，我们假设 ChatGPT 采用的训练集大小为 4,000 亿 Tokens（数据单位，1,000 Tokens 约对应 750 个英文单词、500 个中文字符），由此得出 ChatGPT 训练所需的浮点运算次数。

- AI 算力集群单位时间有效浮点运算次数：在利用 GPU 进行 AI 模型训练时，GPU 算力除了用以训练模型，还被用以通信、训练数据读写等，因此有效浮点运算次数才能衡量 AI 算力集群的算力水平。

图表3：前沿 AI 模型的有效算力比率

模型名称	时间	使用硬件	有效算力比率
GPT-3	2020 年 5 月	NVIDIA V100	21.3%
MT-NLG	2021 年 10 月	NVIDIA A100	30.2%
PaLM	2022 年 4 月	Google TPU	46.2%

来源：《PaLM: Scaling Language Modeling with Pathways》（Aakanksha Chowdhery 等，2022），国金证券研究所

ChatGPT 模型在底层算法结构层面与 MT-NLG 模型类似，并且同样采用 NVIDIA A100 进行训练，我们保守估计 ChatGPT 有效算力比率为 30%。

图表4：考虑有效算力比率后的训练所需浮点运算次数计算

AI 算力集群单位时间有效浮点运算次数 = 算力集群峰值算力 × 有效算力比率

来源：机器学习算法与自然语言处理公众号，国金证券研究所

OpenAI 的模型训练算力由 Microsoft Azure 通过云服务提供，由于 ChatGPT 采用 NVIDIA A100 训练，我们选取了 Microsoft Azure 的 ND A100 v4 系列作为对标的 AI 计算集群，该 AI 计算集群由 8 块 NVIDIA A100 组成，NVIDIA A100 的峰值算力为 312PFlops，使用价格为每小时 10.9 美元。

通过上述计算，我们得出在使用云计算时 ChatGPT 的训练成本约为 170 万美元。根据微软 2022 年财报，微软云的毛利率为 70%，因此如果自建 AI 算力中心进行模型训练，ChatGPT 的训练成本将降至约 51 万美元。

2. 通用 AI 模型推理成本估算

由于AI模型的应用场景多元化，在不同应用场景中AI模型会进行复杂度不同的推理运算，为便于后续 AI 模型场景成本估算，我们将估算 AI 模型每处理 1,000 Tokens 信息所需的推理成本。

我们同样以 ChatGPT 为例，采用如下公式估算 AI 模型的推理成本：

图表5：AI 模型推理成本估算

$$\text{推理成本} = \frac{\text{AI模型推理所需的浮点运算次数}}{\text{AI算力集群单位时间有效浮点运算次数}} \times \text{AI算力集群单位时间价格}$$

来源：量子位公众号，国金证券研究所

- 推理所需的浮点运算次数：根据 OpenAI 于 2020 年发表的相关研究，对于 GPT-3 等采用 Decoder 结构 Transformer 骨干网络的大型语言模型而言，其推理所需的浮点运算次数可以按照如下公式计算：

图表6：AI 模型推理所需浮点运算次数计算

$$\text{AI模型推理所需的浮点运算次数} = 2 \times \text{模型参数规模} \times \text{训练集大小}$$

来源：《Scaling Laws for Neural Language Models》（Jared Kaplan 等，2020），国金证券研究所

推理成本计算中的其余参数与训练成本计算中相同。根据计算，我们得出在使用云计算时 ChatGPT 每处理 1,000 Tokens 信息需要花费约 0.177 美分，如果自建 AI 算力中心为模型推理提供算力支持，成本将进一步降低至 0.053 美分。

3. 多应用场景下 ChatGPT 推理成本估算

由于在不同的应用场景中，AI 模型所面临的任务复杂度有所不同，因此 AI 推理成本的估算必须基于应用场景。我们采用如下公式估算各应用场景中 ChatGPT 的推理成本：

图表7：具体应用场景中 AI 模型完成一次任务所需成本

$$\text{场景成本} = \text{场景任务所需Tokens} \times \text{推理成本} \times (1 - \text{缓存命中率}) \times \frac{1}{1 - \text{计算集群闲置率}} \times \text{模型压缩因子}$$

来源：新钛云服公众号，算法邦公众号，机器之心公众号，国金证券研究所

其中，各参数的意义如下：

- 场景任务所需 Tokens：其指在具体应用场景中，为了完成给定的场景任务，如搜索信息、文书修改撰写等，AI 模型需要处理的 Tokens 数量。这既包括向 AI 模型输入的信息，也包括 AI 模型自主生成的信息。
- 推理成本：我们在前文计算得到在使用云计算时 ChatGPT 每处理 1,000 Tokens 信息

需要花费约 0.177 美分，如果自建 AI 算力中心，推理成本将降低至 0.053 美分/每 1,000Tokens。

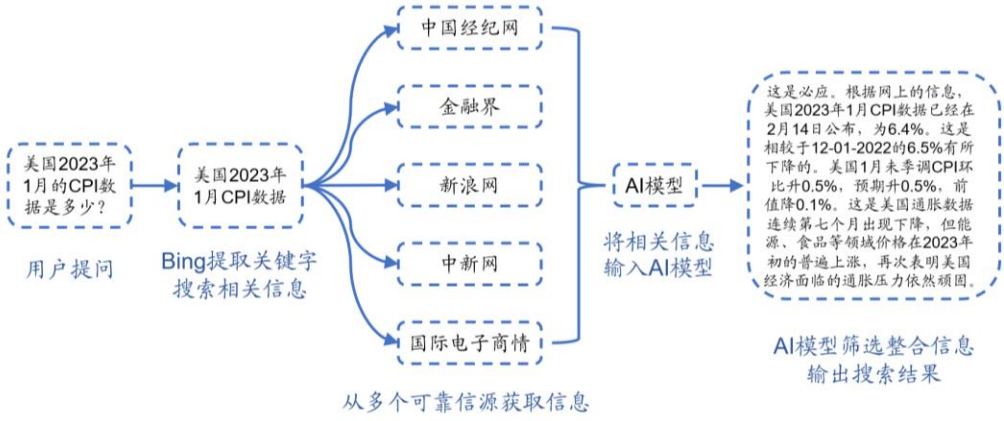
- 缓存命中率：在 AI 模型应用至具体场景时，用户的任务需求存在重叠性。当一位用户向 AI 模型提出一个任务需求时，该任务需求可能在此之前已为 AI 模型所解决，因此无需 AI 模型进行推理，仅输出缓存信息即可完成场景任务。
- 计算集群闲置率：当 AI 模型应用至具体场景时，为保证能够及时响应用户需求，AI 算力集群需采用冗余配置方式，这会导致算力闲置。
- 模型压缩因子：用以衡量当 AI 模型应用至具体场景时，是否可以通过模型压缩技术来压缩 AI 模型的参数规模，从而降低场景下的推理成本。

3.1 搜索引擎场景

2023 年 2 月 7 日，微软宣布推出新版 Bing 搜索引擎，将 ChatGPT 技术引入搜索引擎，有望为用户提供全新的搜索体验。搜索引擎场景是 ChatGPT 的第一个落地场景，我们将以新版 Bing 搜索引擎为例，估算搜索引擎场景下的 AI 推理成本。

新版 Bing 搜索引擎在执行搜索任务时，采用下图中的方式获取信息并加以整合，最终输出搜索结果。

图表8：新版 Bing 搜索引擎的运作方式



来源：Bing，智东西公众号，国金证券研究所

我们将搜索引擎场景任务定义为完成一次搜索，即与 Bing 进行一次问答。根据新版 Bing 搜索引擎的运作方式，我们采用如下公式估算搜索引擎场景任务的所需 Tokens：

图表9：Bing 搜索引擎场景任务的所需 Tokens

$$\text{场景任务所需Tokens} = \text{用户输入} + (\text{模型从每个信源中读取的Tokens} \times \text{每次搜索的信源数量}) + \text{模型输出Tokens}$$

来源：智能试听研究院，OneFlow 公众号，国金证券研究所

我们对相关参数进行如下估算：

- 用户输入：我们估计用户进行一次搜索时输入中文 25 字，对应 50 Tokens。
- 模型从每个信源中读取的信息：Bing 在对信源进行信息读取时，会首先读取标题、摘要和文章的前两段，再根据读取信息与关键词是否匹配，决定是否读取更多的信息。Bing 从每个信源中读取的文字量从 300 字到 1,000 字不等，我们估计模型平均从每个信源中读取的信息为 1,300 Tokens。
- 每次搜索的信源数量：在绝大多数情况，Bing 针对用户问题会对 6 个信源进行搜索，并从中挑选 4-6 个信源来生成最终的搜索结果，我们估计每次搜索的信源数量为 6。
- 模型输出 Tokens：Bing 生成的搜索结果力求简洁精炼，输出内容通常在 200-400 字间，我们保守估计模型每次的输出为 600 Tokens。
- 缓存命中率：我们参照了 Google 搜索引擎的缓存命中率，Google 不同时期的缓存命中率在 30%-60%不等，我们保守估计新版 Bing 的缓存命中率为 30%。

- 计算集群闲置率：为使得用户搜索得到及时详细的结果，需对算力集群实行冗余布置，我们估计计算集群闲置率为 20%。
- 模型压缩因子：由于在搜索引擎场景，AI 模型需要处理信息来源极为广泛，要求 AI 模具备各领域的知识，因此在当前技术水平下，无法通过模型压缩技术来降低搜索引擎场景的推理算力。我们将模型压缩因子设定为 1。

由此，我们计算得出在采用云计算的情况下新版 Bing 搜索引擎完成一次搜索的成本为 1.73 美分；如果自建 AI 算力中心，成本将下降至 0.52 美分。

3.2 办公软件场景

2023 年 2 月 7 日，微软宣布最快或于 3 月将 ChatGPT 整合至 Microsoft Office 系列软件，并开放相关测试。办公软件场景或将成为 ChatGPT 的第二个落地场景。

我们将办公软件场景任务定义为完成一次用户需求，包括文字生成、文字修改等。我们采用如下公式估算办公软件场景任务的所需 Tokens。

图表10：办公软件场景任务的所需 Tokens

场景任务所需Tokens＝（用户输入＋模型输出Tokens）×满足用户需求的平均响应轮次

来源：国金证券研究所

我们对相关参数进行如下估算：

- 用户输入：在进行 AI 辅助文章撰写、文章修改时，需要用户提供相关信息，包含大纲、写作要求、追加信息、修改意见等。我们保守估计用户每次输入中文 200 字，对应 400Tokens。
- 模型输出 Tokens：由于办公软件场景涉及文章撰写、修改、翻译等工作，模型每次输出的 Tokens 较大，我们估计模型每次输出中文 1,000 字，对应 2,000Tokens。
- 满足用户需求的平均响应轮次：我们认为现有 ChatGPT 的智能水平有限，同时存在用户对需求描述不清的现象，因此想要满足用户需求需要进行多轮次响应。我们估算满足用户需求的平均响应轮次为 3。
- 缓存命中率：办公软件场景任务多为创意类工作，无法通过缓存机制降低推理算力消耗，我们估算缓存命中率为 0。
- 计算集群闲置率：为使得用户的办公需求得到及时详细的响应，同样需对算力集群实行冗余布置，我们假设计算集群闲置率为 20%。
- 模型压缩因子：办公软件场景任务多为创意类工作，要求 AI 模型具备多领域知识，因此在当前技术水平下，无法通过模型压缩技术来降低搜索引擎场景的推理算力。我们将模型压缩因子设定为 1。

由此，我们计算得出在采用云计算的情况下 ChatGPT 在办公场景完成一次用户需求的成本为 1.70 美分，如果采用自建 AI 算力中心，成本将下降至 0.51 美分。

3.3 AI 客服场景

ChatGPT 大幅提高了对话类人工智能的应用表现，AI 客服作为对话场景应用，有望率先实现 B 端落地，助力企业实现降本增效。

我们将 AI 客服场景任务定义为解决一次用户需求。我们采用如下公式估算 AI 客服场景任务的所需 Tokens。

图表11：AI 客服场景任务的所需 Tokens

场景任务所需Tokens＝（用户输入＋模型输出Tokens）×解决用户需求的平均对话轮次

来源：国金证券研究所

我们对相关参数进行如下估算：

- 用户输入：在用户对 AI 客服进行咨询时，需要用户提供相关信息。我们保守估计用

户在进行咨询时，平均每次输入中文 50 字，对应 100 Tokens。

- 模型输出 Tokens：为解决用户需求，我们估计模型每次输出中文 300 字，对应 600 Tokens。
- 满足用户需求的平均响应轮次：在客服场景中经常存在用户对问题描述不清的现象，这就需要模型通过与用户进行多轮对话，引导用户描述问题。因此想要满足用户需求需要进行多轮次对话。我们假设满足用户需求的平均对话轮次为 5 轮。
- 缓存命中率：由于场景相较搜索引擎较为单一，用户需求会有较高的重叠度，我们假设 AI 客服场景任务的缓存命中率为 35%。
- 计算集群闲置率：为使得用户的问题得到及时解决，同样需对算力集群实行冗余布置，我们估算计算集群闲置率为 20%。
- 模型压缩因子：AI 客服仅需针对单一场景，为解决客户问题，AI 模型仅具备单一领域知识，可以通过模型压缩技术削减模型中的无用参数，以此降低搜索引擎场景的推理算力。我们将模型压缩因子设定为 0.5。

由此，我们计算得出在采用云计算的情况下 ChatGPT 在 AI 客服场景解决一次用户需求的成本为 0.08 美分，如果采用自建 AI 算力中心，成本将下降至 0.02 美分。

4. 投资建议

从近期推进节奏来看，GPT-3.5-turbo 模型可为高需用户提供专用实例（Dedicated instances），或将进一步加速场景应用落地。开发者可上传数据，由 OpenAI 为其进行云端微调，以获取大模型在特定场景下的能力。据官方口径，若开发者每日运营超过 4.5 亿 Tokens，Dedicated instances 可能是更经济的选择。未来随着模型压缩技术的持续发展，推理成本可能进一步降低，也有望大幅推动生成式模型在各个场景的大规模商用。

我们认为有海外场景的公司有可能率先与 GPT-3.5-turbo 进行对接，建议关注福昕软件、万兴科技等海外营收占比较高的应用公司。

5. 风险提示

- 海外基础软硬件使用受限

若因国际关系等原因，高算力 GPU 等基础硬件或计算框架等基础软件使用受限，可能会对国内人工智能算法应用产生影响。

- 应用落地不及预期

若相关应用公司不能找到人工智能算法较好的商业应用落地场景，或相关场景客户没有较强的付费意愿，可能算法应用落地会不及预期。

- 行业竞争加剧风险

若相关企业加快技术迭代和应用布局，整体行业竞争程度加剧，将会对行业内已有企业的业绩增长产生威胁。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；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，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-60753903
传真：021-61038200
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号
紫竹国际大厦 7 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号
新闻大厦 8 层南侧

深圳
电话：0755-83831378
传真：0755-83830558
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：中国深圳市福田区中心四路 1-1 号
嘉里建设广场 T3-2402