计算机

报告日期：2023 年 03 月 13 日

AI 应用成本快速下降，MaaS 模式下商用空间有望打开

——行业专题报告

投资要点

- **我们认为 OpenAI 作为 AI 模型提供商，未来有望在 MaaS 模式下打开商业空间。MaaS 模式下 AI 驱动搜索业务的成本已满足商用要求（对标 Google 搜索），服务价格有望具备市场竞争优势（对标 AWS），并且技术迭代可满足 C 端市场需求。建议关注 OpenAI、百度等国内外大厂主导的 AI 生态加速建设过程，并围绕 AI 商业化三大主线挖掘投资标的。**

- **MaaS 有望成为 AI 大模型提供商的核心商业模式**

  1、ChatGPT 单次搜索成本接近商业化要求，商用前景广阔。基于 OneFlow，经我们测算，ChatGPT 驱动的单次搜索成本约为 0.0066 美元，远低于 Google 搜索引擎单次检索收入（0.055 美元），商用条件已经满足。

  2、对标 AWS Comprehend 服务，经我们测算，OpenAI 基于 MaaS 模式提供服务价格更低，未来有望在下游市场具备竞争优势。

  3、模型迭代提升 AI 服务能力，Meta MultiRay 平台提供 AI 大模型服务，基于此支撑的搜索引擎服务每日可支持 8000 亿次查询服务，满足 C 端市场需求。

- **AI 商业化进程加速，看好 MaaS 模式下商业空间打开**

  1、微软加速 OpenAI 技术与业务生态融合，包含搜索引擎、浏览器、移动办公等，经测算 ChatGPT 技术及相关产品可为公司提供显著营收增量。

  2、百度深耕 AI 大模型多年，已构建完善产品技术生态，国内细分行业企业加速布局 AI 应用并宣布接入百度 AI 生态，看好国内生态加速完善，推进 AI 商业化落地进程。

- **投资建议：围绕 AI 商业化三大主线挖掘投资标的**

  1、关注具备底层算法模型核心技术优势的厂商：

  （1）推荐标的：拓尔思（中文 NLP 龙头厂商），科大讯飞（智能语音处理及合成）；

  （2）建议关注：谷歌（DeepMind），微软（OpenAI，ChatGPT），Meta（OPT 模型），百度（“文心”模型），腾讯，浪潮信息（“源”大模型）；

  2、关注各细分赛道下兼具场景理解与 AI 布局优势的垂类厂商：

  （1）推荐标的：海康威视（智能物联领域龙头）；

  （2）建议关注：云从科技（智慧城市），格灵深瞳（智慧金融），金山办公（办公），万兴科技（AI 绘画），商汤（智慧安防）；

  3、围绕 AI 数据、算力等基础设施选择优质投资标的：

  （1）推荐标的：海天瑞声（国内 AI 训练数据龙头提供商）；

  （2）英伟达（GPU），寒武纪（AI 芯片）；

- **风险提示**

  1、AI 技术迭不及预期的风险； 2、AI 商业化产品发布不及预期；3、政策不确定性带来的风险；4、下游市场不确定性带来的风险；

行业评级：看好(维持)

分析师：刘雯蜀
执业证书号：s1230523020002
liuwenshu03@stocke.com.cn

相关报告

1 《计算机行业点评报告：OpenAI 发布 Whisper API，再添新收费产品》 2023.03.04

2 《计算机行业深度报告：潮起潮落，拐点已过，AIGC 有望引领人工智能商业化浪潮》 2023.02.12

正文目录

图表目录

1 模型成本大幅下降，MaaS 模式下 AI 商用空间加速打开

我们认为 ChatGPT 在文本类交互场景下的应用有两条路径，一是将知识储存在大模型中（ChatGPT），单次使用直接得到信息输出，即 ChatGPT Equivalent 模式；另外一种方法是基于现有搜索引擎，AI 通过对实时信息的智能检索和分析，经过加工后再生成信息反馈，即 2-Stage Search Summarizer 模式。考虑未来 ChatGPT 技术在实际应用中的准确性和实时性要求，我们认为后一种有望成为主流方案。

我们以信息检索为核心应用场景，以上述两种模式对 ChatGPT API 模型调用成本进行测算。基于 Oneflow，经我们测算，我们认为对标谷歌搜索引擎及相关业务，ChatGPT 技术成本已大幅下降，满足商用化要求，未来有望加速打开应用空间。

OpenAI 持续完善产品服务，有望加速构建 AI 应用生态。OpenAI 除提供模型 API 之外，还布局模型推理和训练的租用服务，根据推理单元和模型版本进行收费，用户可根据自己实际需求进行定制化开发。对标 Amazon Comprehend，经测算 OpenAI 在价格维度具备显著优势。

1.1 ChatGPT 单次搜索成本接近商业化要求，商用前景广阔

以 ChatGPT 为例，测算 AI 大模型驱动搜索引擎成本，对标谷歌搜索引擎业务，我们认为 MaaS 模式有望成为 AI 模型提供商的核心商业模式。根据 OneFLow，我们基于以下两个假设，测算 ChatGPT 驱动搜索引擎的经济成本：

（1）ChatGPT Equivalent：即 AI 大模型经过庞大训练数据集的训练，将获得的知识储存在模型参数中，用户在调用模型进行推理时直接获取来自模型的信息输出；

（2）2-Stage Search Summarizer：即在前一种模式的基础上，AI 大模型在推理时会访问传统的搜索引擎（Google、Bing 等），通过搜索引擎运行查询以检索前 K 个结果。在第二阶段，通过 LLM 运行每个结果以生成 K 个响应，该模型再将得分最高的响应返回给用户。

OpenAI 提供四种收费模式，取 GPT 3.5 turbo 模型收费进行测算。Davinci API 由 GPT-3 的 1750 亿参数版本提供支持，据 OpenAI 官网显示，该模型进行推理的价格为 0.02 美元/750 词（1000token 约等于 750 个单词），而最新的 GPT 3.5 Turbo 模型 API 接口调用价格仅为 Divinci 的 10%，即每 1000 tokens 0.002 美元，用于计算定价的单词总数包括输入和输出。**我们基于最新价格对单次搜索的成本进行测算，认为 ChatGPT 模型成本已满足商业化条件。**

图1： OpenAI Instruct GPT 模型四种收费模式

资料来源：OpenAI 官网，浙商证券研究所

在 ChatGPT Equivalent 假设下，单词检索服务收费约为 0.0055 美元。假设在平均 50 个词的信息输入条件下，为获得更高质量的效果，每次检索将返回 5 个结果响应（并从中选择最佳响应），且每个响应包含 400 词的内容，在该假设下，每次检索服务 OpenAI 收费为 0.0055 美元。

图2： 基于 ChatGPT Equivalent 假设下的单次搜索服务费用

$$\frac{\text{OpenAI revenue}}{\text{query}} = \underbrace{\left(\frac{\text{50 query prompt words}}{\text{query}}\right.}_{\text{input words}} + \underbrace{\left.\frac{\text{400 generated words}}{\text{sampled response}} \times \frac{\text{5 sampled responses}}{\text{query}}\right)}_{\text{output words}} \times \underbrace{\frac{\text{1000 tokens}}{\text{750 words}}}_{\text{words to tokens conversion}} \times \underbrace{\frac{\$0.002}{\text{1000 tokens}}}_{\text{token pricing}}$$

$$= \$0.0055/\text{query}$$

资料来源：OneFlow，网易，浙商证券研究所

在 2-Stage Search Summarizer 假设下，单词检索服务收费约为 0.0375 美元。在该情况下，我们假定模型会检索传统搜索引擎中的相关内容，并基于每条搜索内容生成结果响应，并将模型认为最好的答案输出给用户。我们假定在 50 词的信息输入条件下，每次基于传统搜索引擎生成 10 个响应，每个响应中有平均 1000 词，并最终生成 400 词的搜索结果，则单次检索服务费用为 0.0375 美元。

图3： 基于 2-Stage Search Summarizer 假设下的单次搜索服务费用

$$\frac{\text{OpenAI revenue}}{\text{query}} = \underbrace{\left(\frac{\text{50 query prompt words}}{\text{query}} + \left(\frac{\text{1000 search result prompt words}}{\text{link}} + \frac{\text{400 generated words}}{\text{link}}\right) \times \frac{\text{10 links}}{\text{query}}\right)}_{\text{input and output words}} \times \underbrace{\frac{\text{1000 tokens}}{\text{750 words}}}_{\text{words to tokens conversion}} \times \underbrace{\frac{\$0.002}{\text{1000 tokens}}}_{\text{token pricing}}$$

$$= \$0.0375/\text{query}$$

资料来源：OneFlow，网易，浙商证券研究所

考虑缓存命中率和毛利率，经测算 ChatGPT equivalent 和 2-Stage Search Summarizer 两种情况下检索服务成本分别为 0.001 美元/次和 0.0066 美元/次。考虑搜索效果的准确度，参考谷歌搜索引擎的缓存命中率在 30%~60%，我们取保守估计，ChatGPT 嵌入搜索引擎服务的初期缓存命中率较低，取 30%作为估计值；考虑 OpenAI 业务毛利率，我们参考 SaaS 类公司的毛利率，取 75%作为参考，经测算，ChatGPT equivalent 假设下单次检索服务，模型提供商的成本为 0.001 美元，而 2-Stage Search Summarizer 情况下单次成本为 0.0066 美元。

测算成本与 OpenAI CEO 透露数字接近，ChatGPT 商用模式采用 2-Stage Search Summarizer 可能性更大。OpenAI 首席执行官 Sam Altman 曾在社交平台上透露 ChatGPT 单次使用的成本为几美分（single-digits cents），与测算结果接近。考虑 ChatGPT 落地应用在实时性、准确性等方面的要求，我们认为未来基于 2-Stage Search Summarizer 模式进行部署的可能性更大。

表1：两种假设下的 OpenAI 检索服务测算

	ChatGPT equivalent	2-Stage Search Summarizer
Query Prompt Words / Query	50	
Generated Words/ Sampled Response	400	——
Sampled Response / Query	5	——
Search Result Prompt Words / Link	——	1000
Generated Words / Link	——	400
No. of Links / Query	——	10
No. of Words / 1000 Tokens	750	
OpenAI Revenue / 1000 Tokens	$0.02	
Estimated OpenAI Revenue / Query	**$0.0055**	**$0.0375**
Cache Hit Rate	30%	
OpenAI Gross Margin	75%	
Estimated Cost / Query	**$0.0010**	**$0.0066**

资料来源：OneFlow，网易，浙商证券研究所

对比谷歌搜索引擎业务，基于 ChatGPT 的 AI 技术应用接近商用要求。根据 Alphabet 公司财报披露，2021 年 Google 搜索引擎相关收入约为 1489.51 亿美元。据 OBERLO 报道，Google 搜索引擎 2022 年使用次数超过三万亿次。我们取 2.7 万亿次作为假设，则 2021 年平均单次使用搜索引擎为 Google 创造的收入约为 0.055 美元，而基于 ChatGPT 技术的检索成本为 0.0066 美元，占收入的 12%。**我们认为在 MaaS 商业模式下，随着模型的优化带来的成本进一步下降，商业化空间有望充分打开。**

模型优化有望。我们考虑随着模型算法的优化，2-Stage Search Summarizer 模式下单次搜索成本有望持续下降。我们认为随着 AI 模型的优化，未来在搜索引擎中单次检索调用的信息数量以及文本内容均有望实现减少，基于此进行敏感性测算，在单次调用 5 个链接，并且每个链接下传统搜索引擎返回 700 词，则单次查询访问的成本可下降至 0.0026 美元，仅占 0.055 美元收入的 4.73%。

表2：单次检索成本的敏感性测算

		单次检索生成 links				
		7	6	5	4	3
传统搜索引擎返回词数	900	0.0043	0.0037	0.0031	0.0025	0.0018
	800	0.0039	0.0034	0.0028	0.0023	0.0017
	700	0.0036	0.0031	**0.0026**	0.0021	0.0016
	600	0.0033	0.0028	0.0024	0.0019	0.0014
	500	0.0030	0.0025	0.0021	0.0017	0.0013

资料来源：OneFlow，网易，浙商证券研究所

基于以上测算，我们认为 AI 商业化已经走过了拐点，未来随着模型准确度的提升、数据调用需求下降以及算力成本的优化等途径，有望实现基于 AI 大模型的服务成本持续下降，模型提供商采用 MaaS 模式有望打开商业空间。

1.2 OpenAI 新业务支持个性化模型部署，对标 AWS 产品更具竞争力

OpenAI 布局 Model Instance 新业务，支持用户在 OpenAI 模型基础上通过微调训练个性化模型。OpenAI 在推出 GPT 3.5 Turbo 和 Whisper API 的同时，还推出了新业务模式 Dedicated Instances，支持大规模推理，用户可以基于 OpenAI 的大模型自由控制模型配置和性能设定，以满足自己的使用需求。

Model Instance 采用按月付费模式，每个推理单元月租价为 260 美元。根据 OpenAI 官方披露信息，该模式下主要会有三种版本，分别有 100、300 和 600 个训练单元，最多可支持每日超过 4.5 亿个 token 的模型推理训练工作。

表3： OpenAI Foundry 收费模式

Model Instance	Units/ Instance	3-month commit		1-year commit	
		Monthly cost	Total commit	Monthly cost	Total commit
GPT-3.5 Turbo	100	$26,000	$78,000	$22,000	$264,000
DV (8K max context)	300	$78,000	$234,000	$66,000	$792,000
DV (32K max context)	600	$156,000	$468,000	$132,000	$1,584,000

资料来源：Neoteric，twitter，浙商证券研究所

对标 AWS Comprehend 自然语言处理服务，OpenAI 收费模式或更具竞争力。Amazon Comprehend 提供自然语言处理、个人身份信息 (PII) 检测和修订、自定义分类和实体检测以及主题建模，以支持可分析原始文本的广泛应用程序，并且还使用一些 API 提供 PDF 和 Word 之类的文档格式。AWS Comprehend 收费主要由推理模型调用、模型训练和模型存储这三部分构成。

表4： Amazon Comprehend 主要功能及收费模式

类型	主要功能	收费模式
自然语言处理	适用于实体识别、情绪分析、语法分析、关键短语提取和语言检测的 Amazon Comprehend API 可用于从自然语言文本中提取见解。	以 100 个字符为单位（1 个单位=100 个字符）进行计算，每个请求最低按 3 个单位收费
个人身份信息（PII）	可以查找文档中选定的个人识别信息实体的位置，并可用于创建修订版文档。包含 PII API 会告知文档是否包含选定的 PII。	以 100 个字符为单位（1 个单位=100 个字符）进行计算，每个请求最低按 3 个单位收费
自定义 Comprehend	自定义分类和实体 API 可以训练自定义 NLP 模型以对文本进行分类并提取自定义实体。	异步推理以 100 个字符为单位进行计算，每个请求最低按 3 个单位收费。另外还有模型训练支付费用，每小时 3 USD（按秒计费），以及模型管理费用，每月 0.50 USD。
主题建模	主题建模可从存储在 Amazon S3 的文档集合中识别相关术语或主题。它会识别集合中最常见的主题，并按组整理，然后将文档映射到相应主题。	基于每个作业处理的文档总大小支付费用。前 100 MB 按统一费率收费。超过 100MB，按 MB 收费。

资料来源：AWS Comprehend 官网，浙商证券研究所

经我们测算在假设情形下，OpenAI 费用低于 Amazon Comprehend。我们取自定义 Comprehend 收费模式，考虑 AWS 每个推理单位的承载能力为 100 字符每秒，假设每秒钟用户产生 150 字符的推理需求，每月用户模型训练用时为 10 小时。在 Amazon Comprehend 方案下，用户总计需支付 894.5 美元，其中模型推理费用为 864 美元，模型训练和管理费

用分别为 30 美元和 0.5 美元。而在 OpenAI GPT 3.5 Turbo 模式下，我们假定用户需要租用 2 个推理单元，则每月费用为 584.8 美元，明显低于 Amazon 收费方案。**我们认为，未来 OpenAI 等模型提供商基于 MaaS 模式提供服务将更具市场竞争力。**

表5： Amazon Comprehend 与 OpenAI 两种收费模式下的成本对比

	Amazon Comprehend	OpenAI Davinci
用户推理需求假设	150 字符/秒	
每日工作时长	8h	
租用推理单元数量	2	
每月（30 日）租用时长	240h	
推理单元租用单价	$0.0005/s	
每月推理总字符数		129,600,000
每月推理总 token 数		32,400,000
OpenAI 调用单价		$0.002/1000 token
模型推理费用（单月）	**$864**	**$64.8**
模型训练时长（单月）	10h	
模型训练费用	**$30**	**$520**
模型管理费用	**$0.5**	
总费用	**$894.5**	**$584.8**

资料来源：AWS Comprehend 官网，OpenAI 官网，浙商证券研究所

1.3 模型迭代实现成本大幅降低，多样化模型选择加速商用进程

基于前文的测算逻辑我们认为，实现模型成本大幅降低的主要途径有减少单次模型调用的参数和数据量，以及提高语言模型准确率，而相比于 GPT3 Davinci，GPT 3.5 Turbo 在各项指标上均实现大幅优化，实现 90%的成本缩减。

GPT 3.5 Turbo 响应速度显著提升。根据 36 氪报道，进行对比测试后 GPT 3.5 Turbo 的响应速度相比前版本平均快 1.44 倍，我们假设响应速度与单次调用数据和参数量成反比，我们认为在 GPT 3.5 Turbo 模型下单次调用数据量可减少 30%以上，显著节约成本。

GPT 3.5 Turbo 模型准确度显著提升。在文本分类、情感分析、数学等领域，GPT 3.5 Turbo 在准确度上相对于 Davinci 模型有显著优化，尤其在数学领域，GPT 3.5 Turbo 完成数学测试的准确率达到 75%，并且在更深入的对于数学测试题的解释任务中，准确度提升更显著（68% vs 44%）。

表6： GPT 3.5 Turbo 与 GPT3 Davinci 003 在各项测试上的准确度对比

	GPT 3.5 Turbo	Davinci 003
0-shot classification	**91.72%**	82.02%
k-shot classification	88.37%	**89.47%**
Sentiment analysis	**84.26%**	78.57%
Math	**75%**	61.11%
Math Explanation	**68%**	44%

资料来源：Scale AI，浙商证券研究所

单次文本生成量增长，显著优化用户使用体验。根据 Scale AI 对 GPT 3.5 Turbo 和 Davinci 003 的测试，开发者设计了包含 30 个问题的测试集，使用两个模型智能生成回答。结果显示 GPT 3.5 Turbo 的回答更长，且包含更多的细节，平均每个问题的回答长度约为 156 个单词（约 208 个 token），而 Davinci 003 为 83 个单词（约 111 个 token）。

OpenAI 提供多种版本的模型接口，满足用户定制化需求。OpenAI 目前已上线多款不同的模型，根据用户在代码生成、文本分析、内容过滤等需求提供针对性的模型端口，可支持用户以较低成本和较高效率实现 AI 功能，未来随着模型在实际场景中的渗透率提升和功能延伸，有望推出更多的模型接口，加速商用进程。

表7：OpenAI 提供灵活多样的模型接口

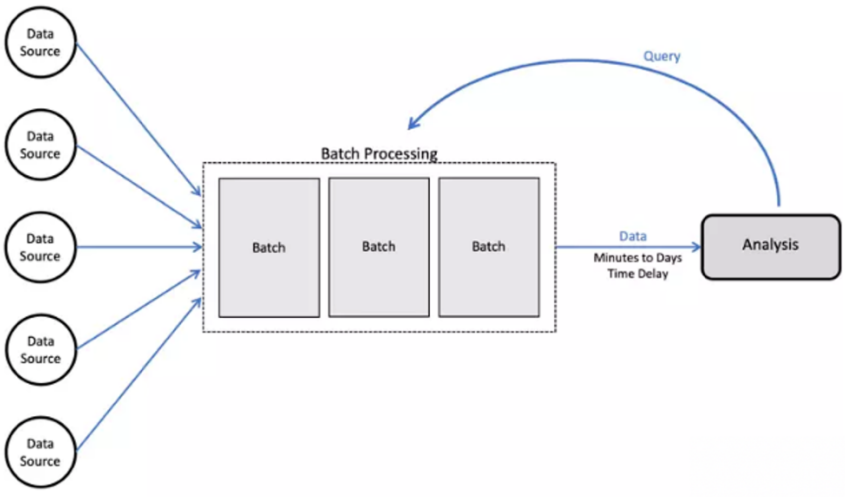
类别	功能描述	模型名	参数量
基础 GPT-3 模型	——	Davinci	175B
		Curie	6.7B
		Babbage	1B
CodeX models	代码生成	Code-cushman-001	12B
Similarity embeddings models	文本相似度分析	Text-similarity-davinci-001	175B
		Text-similarity-curie-001	6B

资料来源：OpenAI 官网，浙商证券研究所

1.4 技术迭代提升 AI 平台负荷，有望拓展模型供应商应用场景

Meta 推出人工智能平台 MultiRay，可运行大规模 AI 模型并大幅提升运行效率。Meta 开发 MultiRay 平台来优化 AI 模型运行，降低整体的模型运算成本。平台集中化大型模型使用以共同分担大部分的处理成本，用户除了能够有效降低成本之外，获得的模型服务也比团队自己发展的更好。

图4：Meta MultiRay 平台通过并行处理提升运算效率



资料来源：新浪财经，浙商证券研究所

MultiRay 平台聚合各类高质量 AI 模型，可广泛应用在各种任务中。MultiRay 于 2020 年发布 TextRay 模型，支持文本理解应用程序，可检测虚假内容并改善用户搜索体验；而 PostRay 模型可以将文本和图像理解集成到同一个模型中，用户可直接通过平台调用该模型，无需再重复开发文本和图像理解模型，大幅提升使用效率。

MultiRay 实现大规模 AI 基础模型访问成本的大幅降低，平台支持高并发查询服务，未来商业化应用空间已经打开。目前，MultiRay 在 Meta 中支持超过 125 个用例，每秒支持高达 2000 万个查询（QPS 达 2000 万），每天可服务 8000 亿次查询，可满足 C 端市场应用的并发要求，未来完全有望投入 C 端产品应用中，实现商业价值的深度变现。

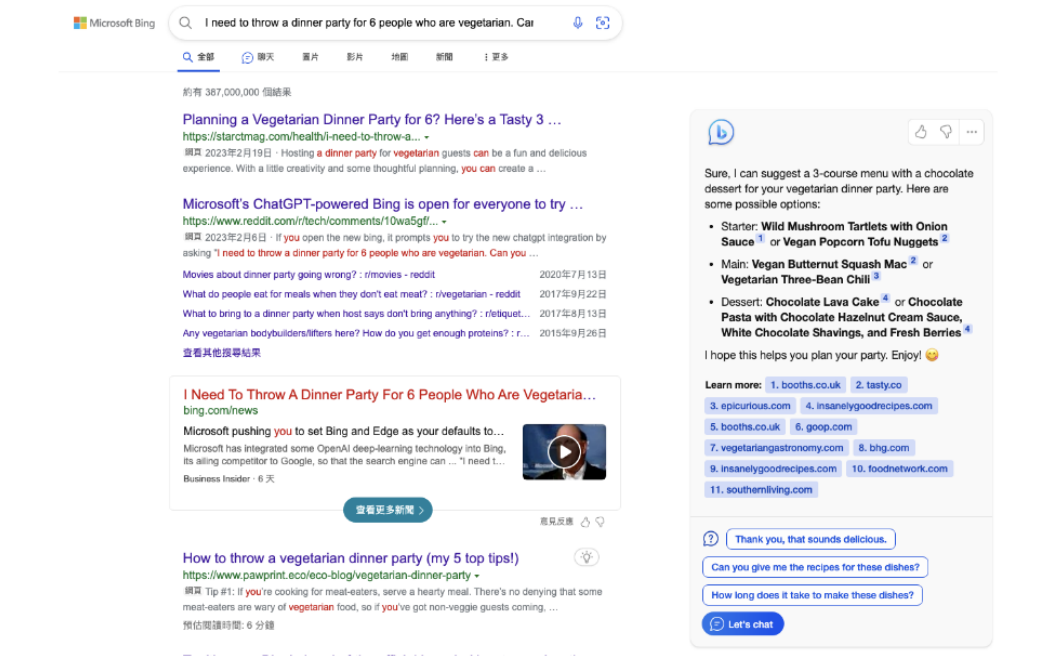
2 AI 商业化进程加速，看好 MaaS 模式下商业空间打开

2.1 微软加速 OpenAI 技术与业务生态融合，市场空间广阔

微软计划推出一系列 AI 应用服务，ChatGPT 有望在实际应用中持续成长。2023 年 2 月 2 日，OpenAI 公司宣布推出付费试点订阅计划 ChatGPT Plus，定价每月 20 美元。付费版功能包括高峰时段免排队、快速响应以及优先获得新功能和改进等。同时，OpenAI 方面仍将提供对 ChatGPT 的免费访问权限。以 ChatGPT 目前约 1 亿的月活用户量为基础，假设 10%的用户有付费意愿，则 ChatGPT 该订阅计划的潜在收入空间为 24 亿美元，市场空间广阔。

新版搜索引擎发布，融合 OpenAI 核心技术。根据钛媒体报道，2 月 8 日微软宣布推出集成 ChatGPT 的全新 Bing 搜索服务，以及集成 AI 辅助的 Edge 浏览器。新版 Bing 带有一个扩展的聊天框，它现在可以做的不仅仅是回答事实问题和为你提供各种链接，在 ChatGPT 的帮助下，它还能够为你即时生成各种个性化的规划、建议、分析等，解决更复杂的搜索问题。

图5：微软新版搜索引擎可根据问题内容提供完整解答及方案建议



资料来源：CSDN，浙商证券研究所

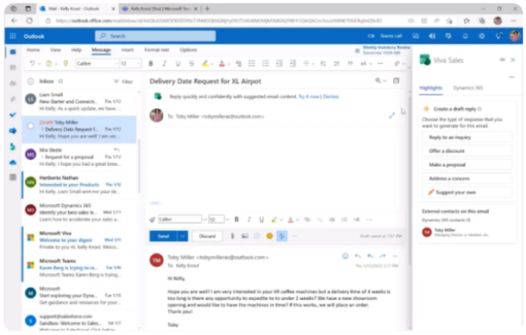
微软有望将 OpenAI 核心技术嵌入 Office 应用中，使其可以根据提示自动生成文本填充文档，并根据用户的需要自动编写电子邮件。微软在其客户关系管理软件 Viva Sales 中集成 OpenAI 技术，AI 程序从客户记录和 Office 电子邮件软件中提取数据，将它们用于生成个性化文本、定价细节和促销信息的电子邮件。

微软 Teams 高级版利用 OpenAI 技术，提升用户办公效率。据澎湃网报道，Teams 高级版新增的人工智能功能（例如智能回顾会议内容）由 OpenAI 的 GPT-3.5 系列大型语言模型提供，Teams 高级版人工智能功能可以通过微软的 Azure OpenAI 服务 API 提供给开发者。

微软为 Teams 高级版订阅提供每月 7 美元的首次折扣价格。首次折扣价适用于初始订阅的整个期限（除了一些按月付费的订阅和一些三年期按年付费的订阅），该优惠于 2023 年 6 月 30 日结束，届时将恢复到每个用户每月 10 美元的标准价格。Teams 高级版人工智能生成的章节将 PPT 实时会议记录分为几个部分。Teams 本身的智能复述则是根据会议记录做这项工作。另外用户可根据参会者加入或离开时间的标记帮助这些人补上错过的会议记录。个性化时间轴还会最终标记出用户的名字在何时被提及、何时屏幕被共享、谁在会议上发过言以及用户在会议中何时发言。

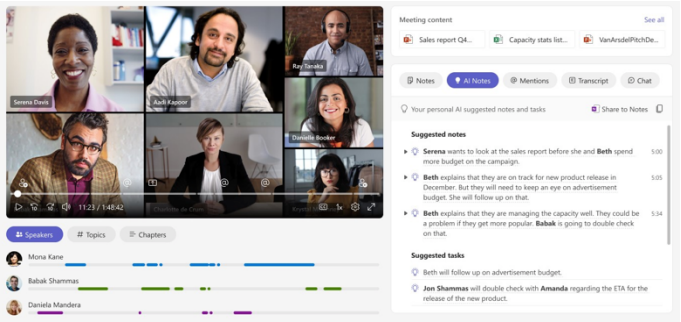
据澎湃网统计，目前 Teams 月活用户数约为 2.8 亿，**我们以每月 7 美元的订阅费计，假设 10%的用户有付费意愿，则该项服务每年可为微软提供约 23.52 亿美元收入。**

图6： OutLook 基于 AI 智能生成邮件



资料来源：The Verge，浙商证券研究所

图7： 微软 Teams 基于 AI 技术提供更多服务



资料来源：微软官网，浙商证券研究所

2.2 百度即将推出 AI 大模型应用，垂直赛道玩家陆续加入生态

百度深度布局 AIGC 多年，围绕自有业务生态形成核心竞争力，在中文 AI 领域优势显著。2022 年 12 月，百度智能云发布国内首个全栈自研的 AI 基础设施“AI 大底座”，具备标准化输出 AI 的底层能力。百度围绕各场景，在 NLP、CV、跨模态、生物计算等领域形成大模型。百度过去十年累积投入研发资金超 1000 亿元，并且连续四年在 AI 专利申请量和授权量上保持国内第一，技术优势明显。

图8：百度“文心”AI 大模型

资料来源：百度文心大模型官网，浙商证券研究所整理

表8：百度基于文心大模型构建完整业务生态

类别	产品名称	主要特点
大模型	NLP 大模型	面向语言理解、语言生成等 NLP 场景，具备超强语言理解能力以及对话生成、文学创作等能力。创新性地将大数据预训练与多源丰富知识相结合，通过持续学习技术，不断吸收海量文本数据中词汇、结构、语义等方面的新知识，实现模型效果不断进化。
	CV 大模型	基于领先的视觉技术，利用海量的图像、视频等数据，为企业和开发者提供强大的视觉基础模型，以及一整套视觉任务定制与应用能力。
	跨模态大模型	基于知识增强的跨模态语义理解关键技术，可实现跨模态检索、图文生成、图片文档的信息抽取等应用的快速搭建，落实产业智能化转型的 AI 助力。
	生物计算大模型	融合自监督和多任务学习，并将生物领域研究对象的特性融入模型。构建面向化合物分子、蛋白分子的生物计算领域预训练模型，赋能生物医药行业。
	行业大模型	文心大模型与各行业企业联手，在通用大模型的基础上学习行业特色数据与知识，建设行业 AI 基础设施，行业覆盖能源、金融、航天、制造、传媒等。
开放 API	ERNIE 3.0 文本理解与创作	提供多种参数量级的、具备超强的语言理解能力和文本创作能力的 API 服务。
	ERNIE ViLG AI 作画	全球规模最大的中文跨模态生成模型，通过自然语言实现图像生成与编辑。无需编程在体验专区探索大模型服务能力、找到应用场景，并可通过 API 集成服务能力。
	文心 PLATO 会话生成	全球首个基于隐空间的大规模生成式开放域对话模型，具备接近真人水平的多轮聊天能力；
工具与平台	大规模套件	ERNIEKit：基于最新一代预训练范式的 NLP 算法定制开发工具集； PaddleFleetX：旨在打造一套简单易用、性能领先且功能强大的端到端大模型工具库，覆盖大模型环境部署、数据处理、预训练、微调、模型压缩、推理部署全流程，并支持语言、视觉、多模态等多个领域的前沿大模型算法。
	零门槛 AI 开发平台 EasyDL	EasyDL 文本：基于大模型为企业/开发者提供一整套 NLP 定制与应用能力； EasyDL 图像：零算法基础定制高精度图像应用 AI 模型，提供端云多种灵活部署方案 EasyDL 跨模态：基于大模型，提供领先的视觉、文本跨模态理解能力，根据业务需求轻松定制图文匹配模型；
	全功能 AI 开发平台 BML	支持一站式 AI 开发，集成飞桨全流程开发套件和丰富的产业场景使用范例；
产品	文心百中	依托大模型，以极简的策略和方案替代传统搜索引擎复杂的特征，低成本接入各类企业和开发者应用，凭借数据驱动的优化模式可实现极致的行业优化效率及应用效果；
	文心一格	AI 艺术和创意辅助平台，依托飞桨、文心大模型的技术创新推出的“AI 作画”产品；

资料来源：百度文心大模型官网，浙商证券研究所

对标 ChatGPT，百度计划近期完成产品内测，有望加速 AI 产品商用化进程。近日，百度官方宣布，将在 3 月份完成其 ChatGPT 产品的内测并面向公众开放，该项目名字确定为文心一言，英文名 ERNIE Bot。公司产业级知识增强文心大模型 ERNIE 具备跨模态、跨语言的深度语义理解与生成能力，对标微软对 OpenAI 核心技术的布局，我们认为百度有望将对标产品应用到业务矩阵下的消费级和企业级应用中，加速 AI 商业化进程。

细分赛道内多玩家宣布接入百度 AI 生态，布局垂直业务场景与 AI 的深度融合。百度文心已累计发布 11 个行业大模型，涵盖电力、燃气、金融、传媒、城市、制造等领域，加速推动行业的智能化转型升级。近期各细分赛道企业宣布接入百度“文心·一言”模型生态，借助百度在 AI 大模型领域的技术优势，深度赋能各细分业务场景的效率提升，**我们认为随着以百度、微软等科技大厂为核心的 AI 生态加速构建，商业化进程有望加速推进，而百度、微软等 MaaS 提供商也将受益 AI 大模型在业务场景中的加速渗透，为企业提供重要助力。**

表9：各细分赛道公司加入百度“文心”AI 模型生态

行业	企业	合作方向
汽车	吉利	基于文心 NLP 大模型，结合了吉利汽车专业领域行业数据（汽车领域媒体知识、客服工单、法律法规以及汽车售后维修手册）进行预训练，充分理解掌握汽车行业知识，得到性能更强、稳定性更高的汽车领域大模型。未来双方还将以吉利-百度·文心大模型为通用底座，进一步支撑吉利在智能车机、知识资产管理和用户运营与智能营销等场景实现智能化升级。
金融	宇信科技	双方也将携手共创“文心一言”在金融业务场景的率先应用，并围绕技术创新、场景孵化、生态建设等多方面展开更深入的合作，助力金融行业的智能化转型升级。
金融	度小满	度小满将基于自身金融场景积累的海量对话及解决方案数据，融合“文心一言”的全面能力，打造全新的智能客服、智能营销、智能风控服务。
传媒	凤凰网	凤凰网将把百度领先的智能对话技术成果应用在新媒体领域，探索基于 AIGC 核心技术，通过相关产品和技术的应用，提高新闻写作内容生产的效率；并进一步丰富资讯类图文和视频内容，提升用户体验。
营销	蓝色光标	蓝标传媒正式成为百度文心一言首批生态合作伙伴，并将在近期全面体验并接入文心一言的能力，快速推进全场景人工智能营销服务体系搭建工作。
办公	金蝶软件	金蝶云·苍穹将把百度领先的智能对话技术成果应用在 ERP 领域，实现更高效的信息获取、信息整合、决策分析、数据洞察，为企业提供更可靠、更敏捷、更智能、更开放的服务。
IT 运维	新炬网络	新炬网络 ZnAiops 智能运维平台与 ZnBot 数字员工产品线将通过百度智能云全面体验并接入文心一言的能力，将把百度领先的智能对话技术成果应用在企业级智能运维与数字员工智能服务领域。
应用软件	航天宏图	航天宏图将把百度领先的智能对话技术成果应用在卫星遥感领域，优先获得领先 AI 技术的加持，也标志着对话式语言模型技术在 PIE-Engine 时空遥感云平台及各行业卫星应用场景中的首次着陆。

资料来源：新浪财经，第一财经，澎湃网，凤凰网，百度文心官网，浙商证券研究所

3 投资建议：围绕 AI 商业化三大主线挖掘投资标的

1、关注具备底层算法模型核心技术优势的厂商：

（1）推荐标的：拓尔思（中文 NLP 龙头厂商），科大讯飞（智能语音处理及合成）；

（2）建议关注：谷歌（DeepMind），微软（OpenAI，ChatGPT），Meta（OPT 模型），百度（“文心”模型），腾讯，浪潮信息（“源”大模型）；

2、关注各细分赛道下兼具场景理解与 AI 布局优势的垂类厂商：

（1）推荐标的：海康威视（智能物联领域龙头）；

（2）建议关注：云从科技（智慧城市），格灵深瞳（智慧金融），金山办公（办公），万兴科技（AI 绘画），商汤（智慧安防）；

3、围绕 AI 数据、算力等基础设施选择优质投资标的：

（1）推荐标的：海天瑞声（国内 AI 训练数据龙头提供商）；

（2）英伟达（GPU），寒武纪（AI 芯片）；

4 风险提示

1、AI 技术迭不及预期的风险；

2、AI 商业化产品发布不及预期；

3、政策不确定性带来的风险；

4、下游市场不确定性带来的风险；

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.买　入 ：相对于沪深 300 指数表现＋20%以上；
2.增　持 ：相对于沪深 300 指数表现＋10%～＋20%；
3.中　性 ：相对于沪深 300 指数表现－10%～＋10%之间波动；
4.减　持 ：相对于沪深 300 指数表现－10%以下。

行业的投资评级：

以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.看　好 ：行业指数相对于沪深 300 指数表现＋10%以上；
2.中　性 ：行业指数相对于沪深 300 指数表现－10%～＋10%以上；
3.看　淡 ：行业指数相对于沪深 300 指数表现－10%以下。
我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。
建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论。

法律声明及风险提示

浙商证券研究所

上海总部地址：杨高南路 729 号陆家嘴世纪金融广场 1 号楼 25 层
北京地址：北京市东城区朝阳门北大街 8 号富华大厦 E 座 4 层
深圳地址：广东省深圳市福田区广电金融中心 33 层
上海总部邮政编码：200127
上海总部电话：(8621) 80108518
上海总部传真：(8621) 80106010
浙商证券研究所：https://www.stocke.com.cn