

OpenAI 发布 Whisper API，再添新收费产品

——行业点评报告

事件

北京时间 3 月 2 日，OpenAI 宣布开放 ChatGPT API，价格为每 1000 tokens 0.002 美元，较 GPT-3.5 模型价格降低 90%。同时，OpenAI 还推出 Whisper 官方 API，Whisper 模型可实现语音-文本（Speech-to-Text）的跨模态任务，使用者可实现语音的转录和翻译，费用为每分钟 0.006 美元。

投资要点

□ 微软宣布开放 Whisper API 接口，AI 多模态进程加速

- 1、Whisper 支持语音转录和翻译两项功能并接受各种语音格式，模型中、英、法、德、意、日等主流语言上取得 85% 以上的准确率，完全符合工业准确率标准，未来有望打开商业化空间；
- 2、Whisper 模型根据参数量和语言不同，共有 9 种版本，可适应不同使用者的需求。在中文语料测试下，模型在语音识别、语气识别、自动断句等方面表现出色，可满足各类使用场景需求；
- 3、多模态将成为 AI 大模型发展的重要趋势，Whisper 模型在跨模态任务上的出色表现，有望为 AI 大模型多模态化发展奠定重要基础。

□ Whisper 模型拆解，性能优异应用场景丰富

- 1、Whisper 模型的编码/解码器架构仍基于 Transformer，通过不同大小和类型的数据输入，实现模型的缩放性能，以及各项语音-文本任务的泛化性和鲁棒性；
- 2、Whisper 模型较之前的语音识别模型有多项改进，在识别准确率、断句、漏音等维度上均有显著优化。目前已有 Speak 为代表的产品使用 Whisper API 提升服务质量，未来看好 Whisper 在更多商业场景中的深度应用。

□ 投资建议：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（ChatGPT，Whisper），Meta（OPT 模型），百度（“文心”模型），腾讯；

□ 风险提示

- 1、AI 技术迭代不及预期的风险；
- 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 《人工智能行业深度：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

2 《拓尔思（300229）：语义智能领军者，大数据+AI 双轮驱动打开公司广阔成长空间》

2022.11.29

正文目录

1 微软宣布开放 Whisper API 接口，AI 多模态进程加速	4
1.1 OpenAI 开放 Whisper API 端口，AI 商用空间持续打开	4
1.2 Whisper 模型灵活多样可满足不同需求，中文识别表现出色	5
1.3 AI 模型有望从单模态走向多模态，应用场景持续延伸	6
2 Whisper 模型拆解，性能优异应用场景丰富	7
2.1 Whisper 模型简约不简单，Transformer 架构发挥重大作用	7
2.2 Whisper 模型优势显著，未来应用场景丰富	8
3 投资建议	9
4 风险提示	9

图表目录

图 1: Whisper 模型在全球主流语种的语音识别错误率	4
图 2: 2022 年中国智能语音产品成熟度曲线分布	5
图 3: Whisper 在 small 模型下的识别结果	6
图 4: Whisper 在 medium 模型下的识别结果	6
图 5: 未来人工智能应用中的 Jina AI 光谱	6
图 6: 多模态 AI 将融合各类型数据之间的界限, 提升 AI 解决任务的能力和范围	7
图 7: Whisper 模型架构	7
图 8: Whisper 采用基于 Transformer 的编码器/解码器架构	7
图 9: Whisper 相比 wav2vec 在不同数据集的词错误率上表现更加出色	8
图 10: Speak 基于 AI 模型辅助用户英语学习	9
表 1: Whisper 模型支持多种参数数量和语种版本	5

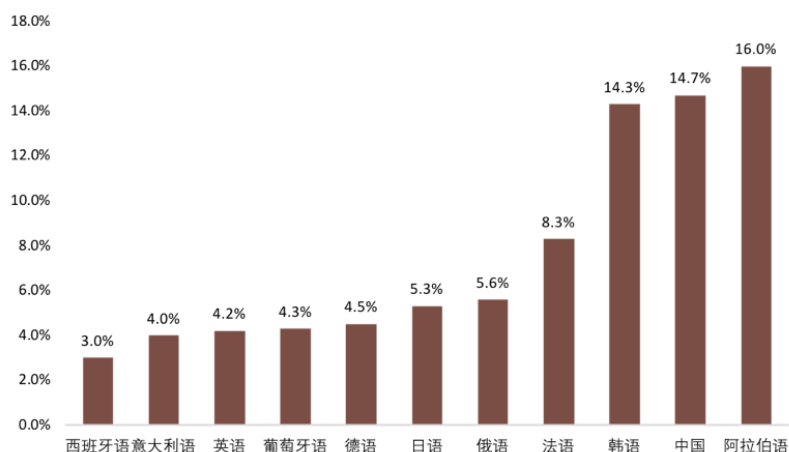
1 微软宣布开放 Whisper API 接口，AI 多模态进程加速

1.1 OpenAI 开放 Whisper API 端口，AI 商用空间持续打开

OpenAI 开放 Whisper 模型 API，AI 商业化进程加速推进。北京时间 2023 年 3 月 2 日，OpenAI 宣布将允许第三方开发者通过 API 将 ChatGPT 集成到他们的应用程序和服务中，并开放 API 允许用户使用语音到文本的 AI 模型 Whisper，对其开发人员服务条款进行了一些修改。

OpenAI 在 2022 年 9 月开源了 Whisper 语音到文本（Speech-to-Text）模型，经过一系列优化后，向使用者开放 API 接口以调用模型，定价为 0.006 美元/分钟。Whisper 支持语音转录和翻译两项功能，并接受各种格式（m4a、mp3、mp4、mpga 等）。模型在全球 98 种语言进行语料训练，并在中、英、法、德、意、日等 55 种语言上取得了低于 50% 的错误率（WER，word error rate），达到语音转文字的工业准确率标准。

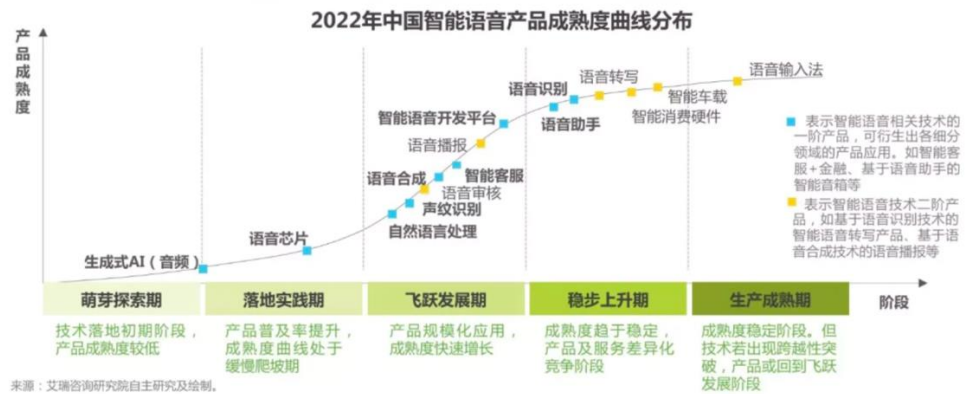
图1： Whisper 模型在全球主流语种的语音识别错误率



资料来源：Github，浙商证券研究所

Whisper 模型标志着语音智能走向成熟，未来在实际场景中的应用渗透率有望进一步加深。根据艾瑞咨询绘制的 2022 年中国智能语音产品成熟度曲线，国内语音智能在语音转写、语音输入等领域已处于较为成熟的阶段，已有规模化的产品应用；而在语音合成、自然语言处理和声纹识别等领域处于成熟度快速上升的阶段。我们认为此次 Whisper 模型的大规模推广，有望凭借其出色的性能和较低的服务成本，对智能语音市场带来变革性的影响，同时也看好 OpenAI 在 Whisper、ChatGPT 等 AI 模型的基础上，推出多模态的 AI 大模型，拓宽 AI 落地场景，打开商业化空间。

图2： 2022 年中国智能语音产品成熟度曲线分布



资料来源：艾瑞咨询，浙商证券研究所

1.2 Whisper 模型灵活多样可满足不同需求，中文识别表现出色

Whisper 模型基于海量数据训练，模型选择灵活多样可满足不同需求。Whisper 模型在 68 万小时标记音频数据上进行训练，其中包括 11.7 万小时 96 种不同语言的演讲和 12.5 万小时“从任意语言到英语”的翻译数据。训练数据集还包括从 YouTube 视频中提取的短语音片段的集合，并根据视频标题和描述的语言进行标记，带有额外的步骤以去除误报。

目前 Whisper 有 9 种模型（分为纯英文和多语言），其中四种只有英文版本，开发者可以根据需求在速度和准确性之间进行权衡，根据模型的大小、内存要求和速度需求选择合适的模型版本，最大的 large 模型版本拥有 1.55 亿参数量，最小的 tiny 版本仅有 3900 万参数量，但运行速度可达到 large 版本的 32 倍。

表1： Whisper 模型支持多种参数量和语种版本

模型大小	参数量	纯英文模型	多语种模型	显存要求	模型相对速度
tiny	39M	tiny.en	tiny	~1GB	~32x
base	74M	base.en	base	~1GB	~16x
small	244M	small.en	small	~2GB	~6x
medium	769M	medium.en	medium	~5GB	~2x
large	1550M		large	~10GB	1x

资料来源：Github，浙商证券研究所

Whisper 模型可较好地实现中文语料的转写，并根据语义进行断句。随着模型参数量的增加，Whisper 模型对中文语料的识别理解能力不断提高。在 small 版本下，Whisper 识别中文的错别字以及语法错误很少，而在 medium 版本下，模型可以出色地完成断句，并且能够判断说话者的语气。我们认为随着模型在识别准确度、转写实时性等维度上的性能进一步优化，未来有望充分打开在各类语音-文本交互场景下的 AI 应用空间。

图3: Whisper 在 small 模型下的识别结果

```
[00:47.800 → 00:50.400] 中国人学习中国史
[00:50.400 → 00:53.000] 不是纯粹的知识
[00:53.000 → 00:56.480] 而是一种带着身体和血液的精力
[00:56.480 → 00:59.480] 因此历史的汇暗与光明
[00:59.480 → 01:02.040] 对我们并非比喻
[01:02.040 → 01:04.800] 而是实际的光与暗
[01:04.800 → 01:06.360] 乐于苦
[01:06.360 → 01:10.240] 但身体需要百般感触
[01:10.240 → 01:13.880] 失去风雨植物已无法伸张
[01:13.880 → 01:19.000] 因此沐浴历史中的光与暗
[01:19.000 → 01:21.480] 才让我们今日的苦乐
[01:21.480 → 01:24.280] 得以接入巨大的历史身体
```

资料来源: 少数派, 浙商证券研究所

图4: Whisper 在 medium 模型下的识别结果

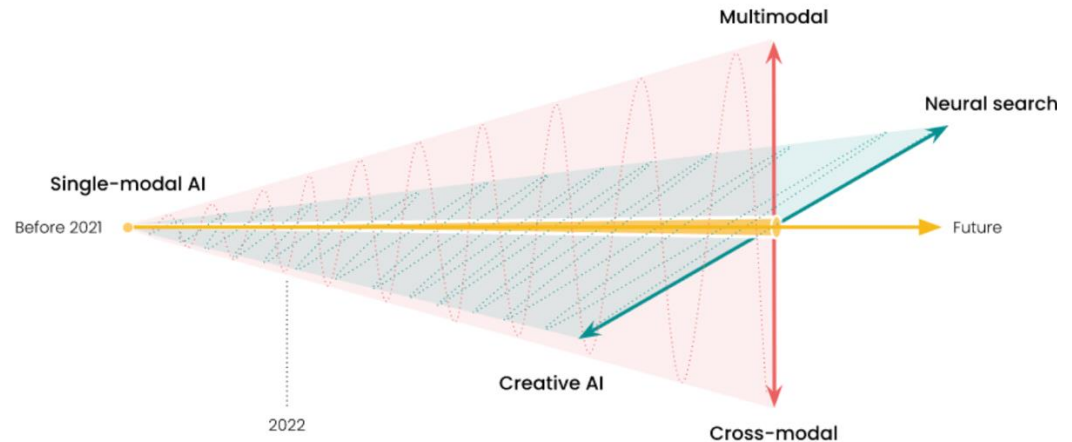
```
[00:00.000 → 00:14.000] 想想你现在的生活中啊,用于我们解释和阐释人世间的现象,不管是物理学的现象还是人间现象,有没有什么大权理论啊?
[00:14.000 → 00:17.000] 其实现在在大权理论还是挺多的啊。
[00:17.000 → 00:24.000] 在我们的传统之中呢,有一个非常重要的大权理论来解释一切,就是这个阴阳五行理论。
[00:24.000 → 00:36.000] 所以这期呢,我们就回到阴阳五行理论被发明的历史现场,来看看这个大权理论是怎么弄出来,又很可能会对历史产生什么样的影响。
[00:36.000 → 00:56.000] 中国人学习中国史,不是纯粹的知识,而是一种带着身体和血液的精力。
[00:56.000 → 01:07.000] 因此历史的晦暗与光明,对我们并非比喻,而是实际的光与暗,乐与苦。
[01:07.000 → 01:15.000] 但身体需要百般感触,失去风雨,植物亦无法生长。
[01:15.000 → 01:27.000] 因此,沐浴历史中的光与暗,才让我们今日的苦乐得以接入巨大的历史身体,流入历史的血液。
[01:27.000 → 01:35.000] 这会给我们全新的世界,和全新的理解,与全新的道路。
[01:35.000 → 01:45.000] 饭店2.0第三章,中国历史与思想史景观,接入历史的身体。
```

资料来源: 少数派, 浙商证券研究所

1.3 AI 模型有望从单模态走向多模态, 应用场景持续延伸

AI 模型将从单模态走向跨模态、多模态, 未来 AI 应用空间将呈现爆炸式扩展。开源神经搜索公司 Jina AI 曾描述了未来 AI 产品及应用的四大特征, 即跨模态 (cross-modal)、多模态 (Multi-modal)、神经搜索 (neural search) 和生成式 AI (Creative AI), 相比于单模态 AI 模型, 多模态 AI 的应用空间将大幅延伸, 能够将文本、语音、图像以及音视频等数据信息进行综合处理分析, 完成跨模态的应用, 从而胜任真实世界中的各类复杂任务。

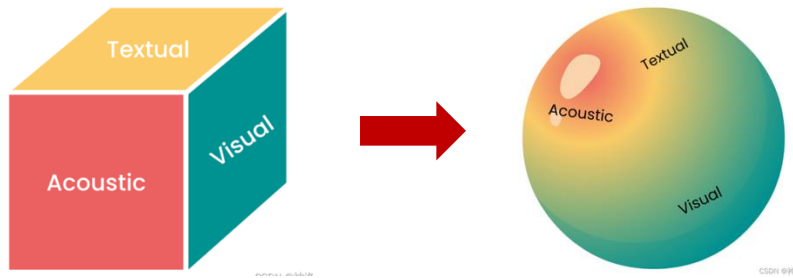
图5: 未来人工智能应用中的 Jina AI 光谱



资料来源: CSDN, 浙商证券研究所

Whisper 模型的应用标志着 AI 在跨模态任务上的突破性进展, 为多模态 AI 大模型的发展奠定良好基础。目前单模态 AI 算法的两大弊端在于: (1) 任务只针对一种模态 (文本、图像、音频等); (2) 知识只能从一种模态学习, 并应用在这一种模态中 (视觉算法智能从图像中学习并应用于图像中)。而 Whisper 模型能够出色完成语音到文本的跨模态任务, 突破了不同模态数据之间的界限, 随着模型不断发展, 我们有望看到 AI 任务在多模态之间的共享传输, 知识也能够从多模式中学习并应用到多种模式中去, AI 大模型朝多模态方向发展的进程有望加速。

图6：多模态 AI 将融合各类型数据之间的界限，提升 AI 解决任务的能力和范围



资料来源：CSDN，浙商证券研究所

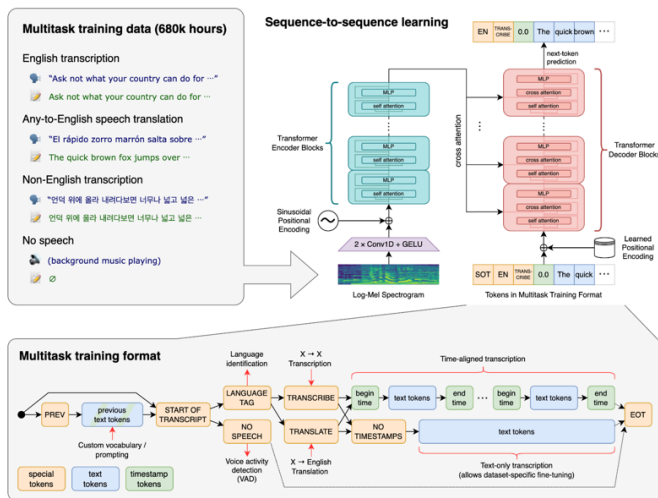
2 Whisper 模型拆解，性能优异应用场景丰富

2.1 Whisper 模型简约不简单，Transformer 架构发挥重大作用

Whisper 模型使用 Transformer 架构的 seq2seq 模型针对各种语音处理任务进行训练。Whisper 模型可完成多语言语音识别、语音翻译、口语识别和语音活动检测，所有任务都被联合表示为解码器预测的令牌序列，允许使用单个模型替换传统语音处理管道的许多不同阶段，多任务训练格式则使用一组特殊标记作为任务说明符或分类目标。

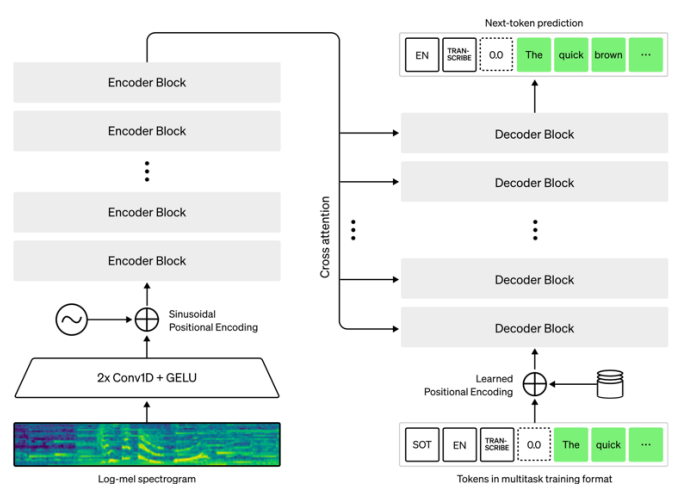
Whisper 模型工作中，输入的音频被分割成 30 秒的小段、转换为 log-Mel 频谱图，然后传递到编码器。解码器经过训练以预测相应的文字说明，并与特殊的标记进行混合，这些标记指导单一模型执行诸如语言识别、短语级别的时间戳、多语言语音转录和语音翻译等任务。

图7：Whisper 模型架构



资料来源：Github，浙商证券研究所

图8：Whisper 采用基于 Transformer 的编码器/解码器架构



资料来源：OSCHINA，浙商证券研究所

Whisper 模型在编码过程中，编码器首先使用一个包含两个卷积层的词干处理输入表示，并使用 GELU 激活函数。第二个卷积层的步幅为 2。然后将正弦位置嵌入添加到词干的输出中，然后应用编码器 Transformer 块。而在解码过程中，Whisper 模型使用了学习位置嵌入和绑定输入输出标记表示。

Whisper 模型为改进缩放属性，在不同的输入大小上进行了训练，依靠数据集难解的多样性来促进泛化性和鲁棒性，Whisper 采用的训练数据中有约 43.82 万小时（占比约 65%）的是英语音频和匹配的英语文本，12.57 万小时（占比约 18%）非英语音频和英语文本，11.71 万小时（占比约 17%）非英语音频和相应的文本，非英语部分共包含 98 种不同语言。使用这样一个大而多样的数据集可以提高对口音、背景噪音和技术语言的鲁棒性。这也使得模型能够变化出大小不同的版本，并在基础性能上保持优秀的表现。

2.2 Whisper 模型优势显著，未来应用场景丰富

Whisper 模型相较前几代语音识别模型优势明显，看好未来在语音转录与翻译场景下的应用。wav2vec2、Conformer 和 Hubert 等模型的发展推动了语音识别领域的发展，这些模型可从原始音频中进行学习，能够有效地使用未标记语音的大型数据集，但 Whisper 模型以监督方式跨多个数据集和领域预训练的模型，表现出更好的鲁棒性和对持有数据集的泛化。

除了以上优势之外，Whisper 模型能够检测清音区域并应用 NLP 技术在转录本中正确进行标点符号的输入，同时模型是可扩展的，允许从音频信号中提取转录本，而无需将视频分成块或批次，从而降低了漏音的风险。而最直观的是，Whisper 模型在各类数据集上取得了更高的准确率，其相比于 wav2vec 模型取得了目前最低的此错误率。

图9： Whisper 相比 wav2vec 在不同数据集的词错误率上表现更加出色

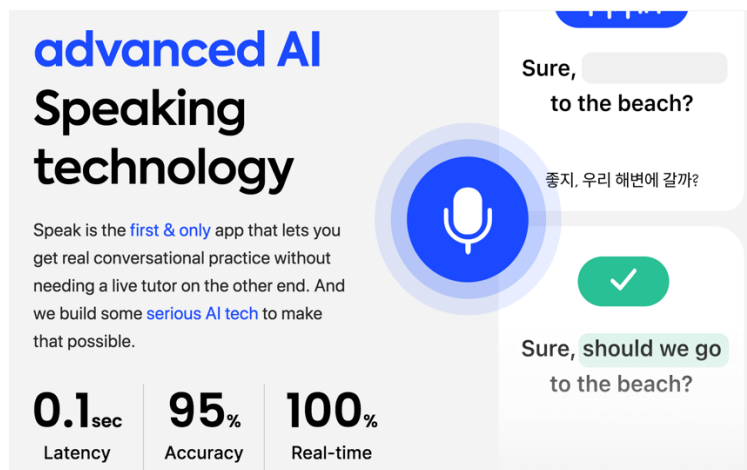
Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	2.7	2.7	0.0
Artie	24.5	6.2	74.7
Common Voice	29.9	9.0	69.9
Fleurs En	14.6	4.4	69.9
Tedlium	10.5	4.0	61.9
CHiME6	65.8	25.5	61.2
VoxPopuli En	17.9	7.3	59.2
CORAAL	35.6	16.2	54.5
AMI IHM	37.0	16.9	54.3
Switchboard	28.3	13.8	51.2
CallHome	34.8	17.6	49.4
WSJ	7.7	3.9	49.4
AMI SDMI	67.6	36.4	46.2
LibriSpeech Other	6.2	5.2	16.1
Average	29.3	12.2	54.8

资料来源：阿里云社区，浙商证券研究所

Whisper 可应用于教育、办公、医疗等多种场景，市场前景广阔，目前已有 Speak 使用 Whisper 模型为其产品服务赋能。Speak 是一款由人工智能驱动的语言学习应用程序，专注于构建口语流畅性的最佳途径。它们是韩国增长最快的英语应用程序，并且已经在使用 Whisper API 为新的人工智能口语伴侣产品提供动力，并迅速将其带到全球其他地区。

Whisper 对每个级别的语言学习者的人级准确性解锁了真正的开放式对话练习和高度准确的反馈。

图10: Speak 基于 AI 模型辅助用户英语学习



资料来源: Speak 官网, 浙商证券研究所

3 投资建议

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;

4 风险提示

- 1、AI 技术迭代不及预期的风险;
- 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

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。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。

本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、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的意见或建议不一致的投资决策。

本报告版权均归本公司所有，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、发布、传播本报告的全部或部分内容。经授权刊载、转发本报告或者摘要的，应当注明本报告发布人和发布日期，并提示使用本报告的风险。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

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>