

GPT-4-Turbo专题

多模态能力提升，应用生态加速

西南证券研究发展中心
海外研究团队 王湘杰
2023年11月

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

核心观点

- **当前变化：GPT4-Turbo模型优化，GPTs生态加速繁荣。** 2023年11月7日，OpenAI通过开发者大会推出新产品：**1) GPT4-Turbo**：该模型通过增加上下文窗口以支持更长的工作流，同时具备视觉和语音等多模态能力，输入和输出的tokens价格大幅下降，从而帮助开发者以更低的价格获取更高的模型性能。**2) GPTs**：用户只需输入指令并提供外设的知识库即可创建专属GPT，应用开发门槛大幅降低，未来在收益计划的助力下，AI应用有望迎来大爆发，形成全新的AI Agent生态。**3) Assistant API**：开发者可以通过函数调用、知识检索、代码解释器简化应用开发流程、实现效率提升。
- **未来影响：算力&存力亟需扩容，AI应用潜力无限，垂类数据或成关键。****1) 算力端**：OpenAI一系列新产品的推出成功打开新流量入口，更多的用户和开发者希望参与其中，巨大的流量对算力底座提出更高的要求；叠加图片等多模态生成所要求的tokens计算量远高于文本模态（根据OpenAI官网信息推算，在GPT-4-Turbo的Vision pricing calculator高保真度模式下，1张图片所产生的tokens数大约是1个单词的570或830倍），算力供给亟需扩容。**2) 存力端**：在算力提效到达一定瓶颈的情况下，AI芯片未来将逐步通过堆叠HBM的方式来提升性能，扩大单位算力的存储能力，HBM等存力需求将迎来暴增。**3) 应用端**：类比移动互联网时代，AI时代的应用市场有望如同移动互联网时代具备无限潜力，GPTs数量将呈现非线性高速增长；此外，OpenAI的GPTs通过提供API，使得开发者只需喂给大模型更多的垂类数据即可打造垂类AI应用，同时使满足更多长尾需求成为可能。**4) 数据端**：从OpenAI GPTs的Knowledge功能来看，专业知识与大模型通用能力的结合将成为未来的重中之重，私域数据库和专业数据库方向将会不断产生新热点、新需求，因此，如何在合规前提下留存垂类数据并构建体系化数据库、以及保证知识产权的确权或成为未来的重要议题。
- **投资建议：1) 算力**：把握云侧芯片龙头、端侧格局变化和大厂自研趋势；**2) 存力**：建议关注高性能存储器芯片制造商；**3) 应用**：建议关注以OpenAI和微软为代表的领军企业；**4) 数据**：建议关注具备垂类优势和优质产权的数据提供商。相关标的：英伟达(NVDA.O)、超威半导体(AMD.O)、英特尔(INTC.O)、高通(QCOM.O)、微软(MSFT.O)等。
- **风险提示**：相关理论和技术发展不及预期风险；行业竞争加剧风险；应用生态发展不及预期风险等。

目 录

1 技术原理：图像理解+视觉生成持续发展，助力多模态能力提升

1.1 图像理解能力提升：三大视觉学习方法

1.2 视觉生成：多模态内容理解和生成的闭环

2 当前变化：GPT4-Turbo模型优化，GPTs生态加速繁荣

2.1 GPT-4 Turbo：多模态能力增强，使用成本降低

2.2 GPTs：加入Agents战场，与开发者共享收益

2.3 Assistant API：解决API开发者痛点，拓展OpenAI收入来源

3 未来影响：算力&存力亟需扩容，AI应用潜力无限，垂类数据或成关键

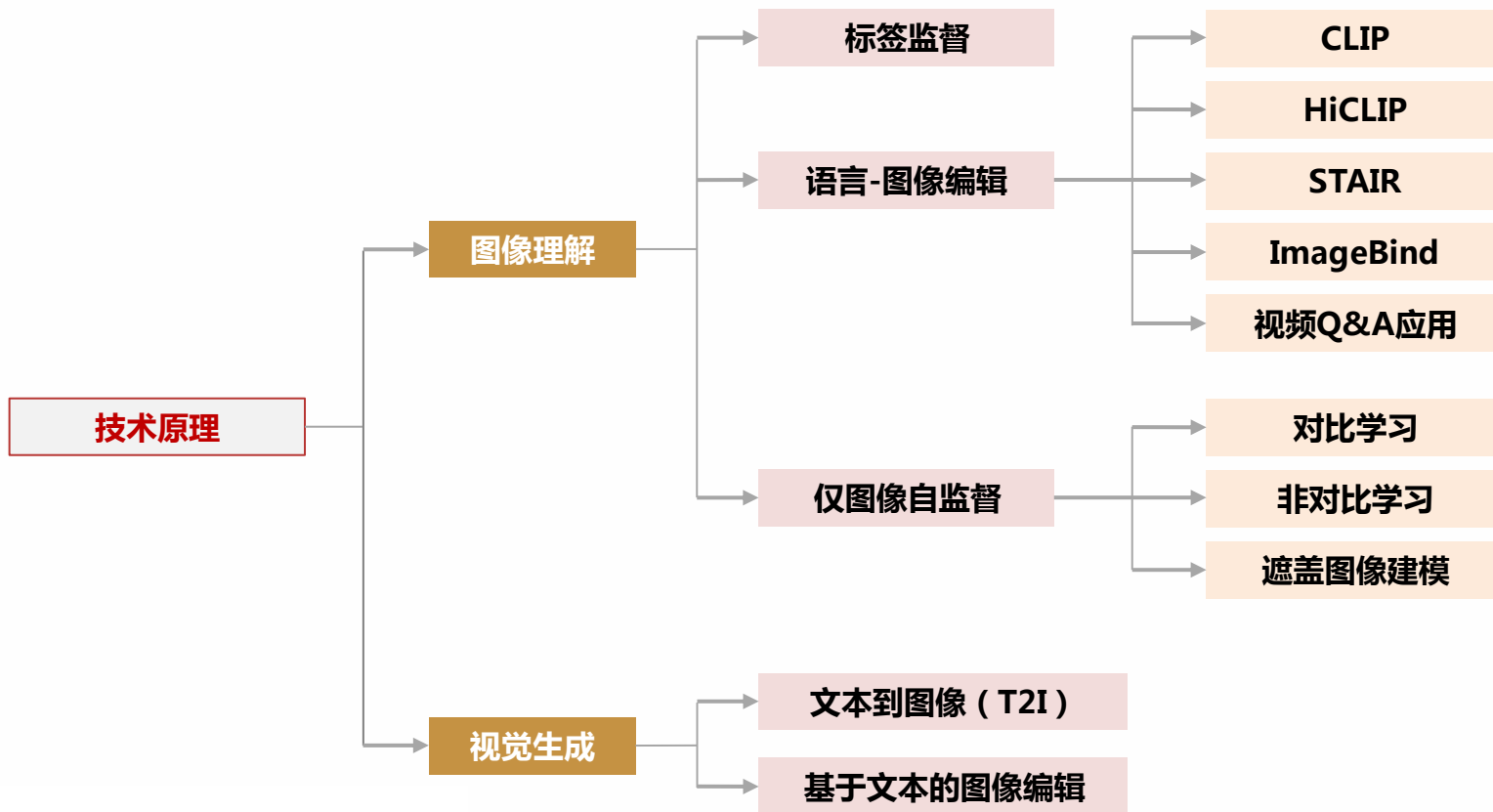
3.1 算力&存力端：流量激增+多模态生成拉大算力缺口，同时带动存力需求

3.2 应用端：强化垂类属性&满足长尾需求，数量有望实现快速增长

3.3 数据端：专业数据或成关键，重视垂类数据留存及知识产权价值

1 技术原理：图像理解+视觉生成持续发展，助力多模态能力提升

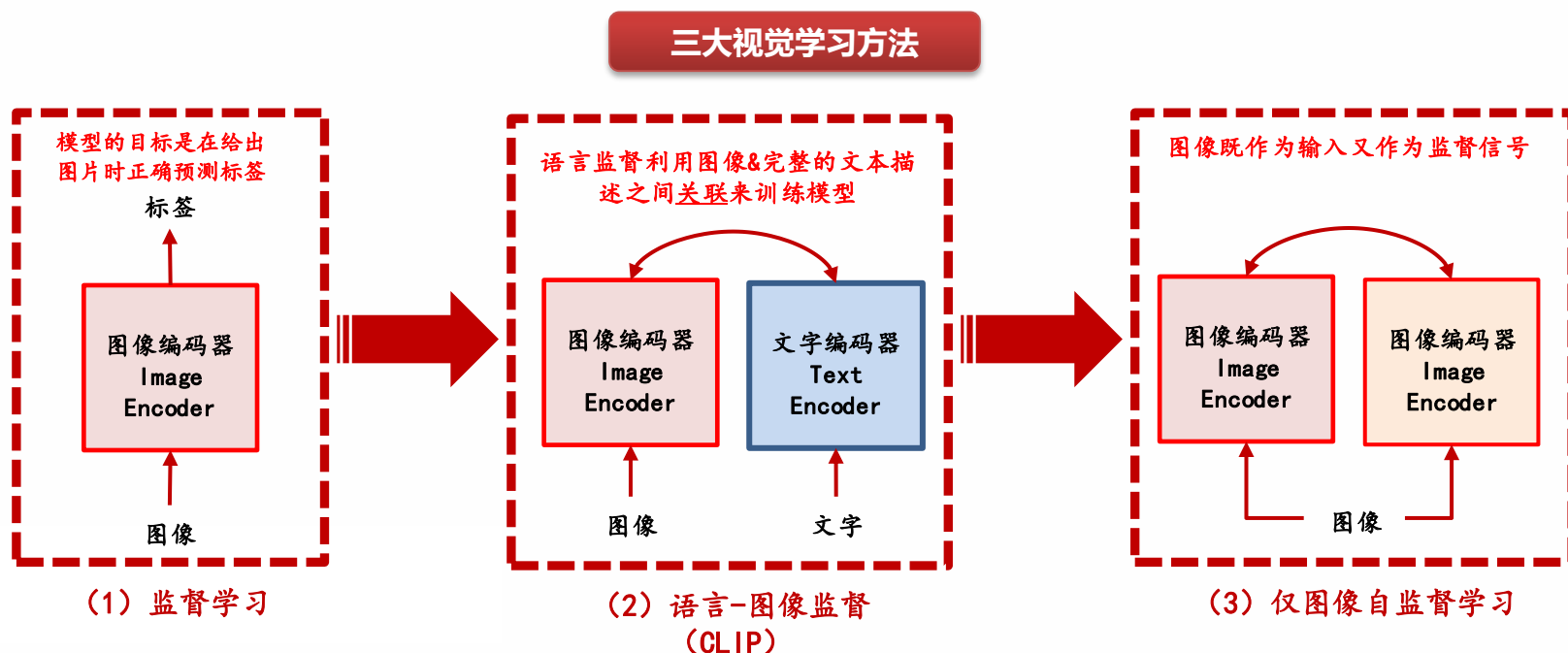
章节小结



资料来源：西南证券

1.1 图像理解能力提升：三大视觉学习方法

- ❑ 为打造视觉大模型，建立高效视觉预训练体系以提升图像理解能力至关重要，这影响到从图像整体（如分类、图文检索、标注）到部分区域（如物体识别、短语定位）及到像素细节（如语义、实例、全景分割）的各项任务。
- ❑ 通用视觉预训练方法主要可归纳为三大类。**1) 标签监督**：此方法在每张图片都配有对应标签的数据集上进行训练，如图像分类中，一张狗的照片会对应“狗”的标签，模型的核心任务是准确预测此标签。**2) 语言-图像监督**：利用完整的文本描述来引导模型学习，使模型能够深入挖掘图像内容与文本语义间的关联。**3) 仅图像自监督**：利用图像本身固有的结构和信息来学习有意义的表示，而不依赖于显式的人工注释标签。



1.1.1 图像理解能力提升：标签监督

❑ **监督式预训练已成为机器学习和计算机视觉领域的核心技术。**在这种策略中，模型首先在大规模标注数据集上进行预训练，随后针对特定任务进行精细调整。这种方法充分利用了如ImageNet这样的大型人工标注数据集，为模型赋予了高度可迁移的视觉特征。

❑ **其核心原理是将图像与预设的标签相对应，这些标签往往代表某一视觉物体。**多年来，这种策略在各类视觉基础架构，如AlexNet、ResNet以及ViT的发展中都发挥了不可或缺的作用。监督式预训练为计算机视觉领域带来了革命性的进步，从基础的图像分类和物体检测，到更为高级的视觉问答和图像标注任务。**受限于人工标注的高成本，这些模型所学习到的特征会受到预训练数据集的规模和多样性的制约。**

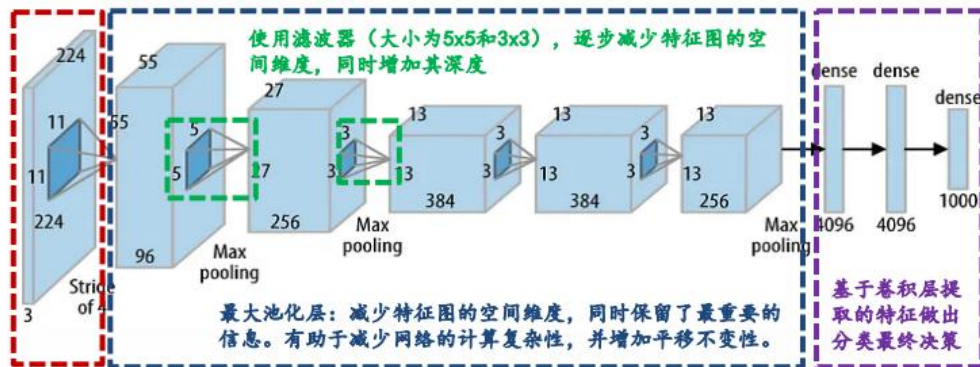
ImageNet数据集



超过1,000万个手工标注的高分辨率图像，涵盖2万多个类别
每个图像都被详细地标注为某一特定的类别，如“狗”、“猫”或“汽车”

AlexNet 深度神经网络

接受一个大小为224x224的输入图像



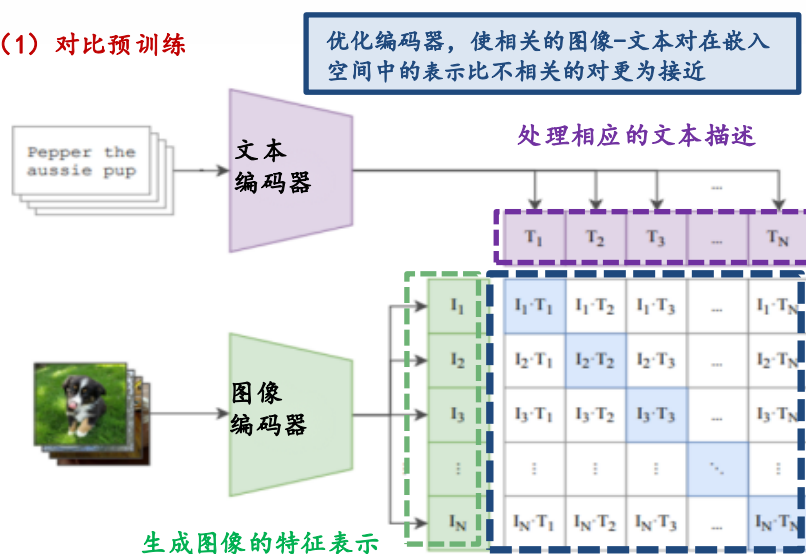
标签监督预训练允许AlexNet等深度神经网络学习多层次的特征表示—从基本的边缘和纹理到更高级的对象部件和结构。这种层次化的表示对于图像分类任务至关重要。

1.1.2 图像理解能力提升：语言-图像监督预训练——CLIP

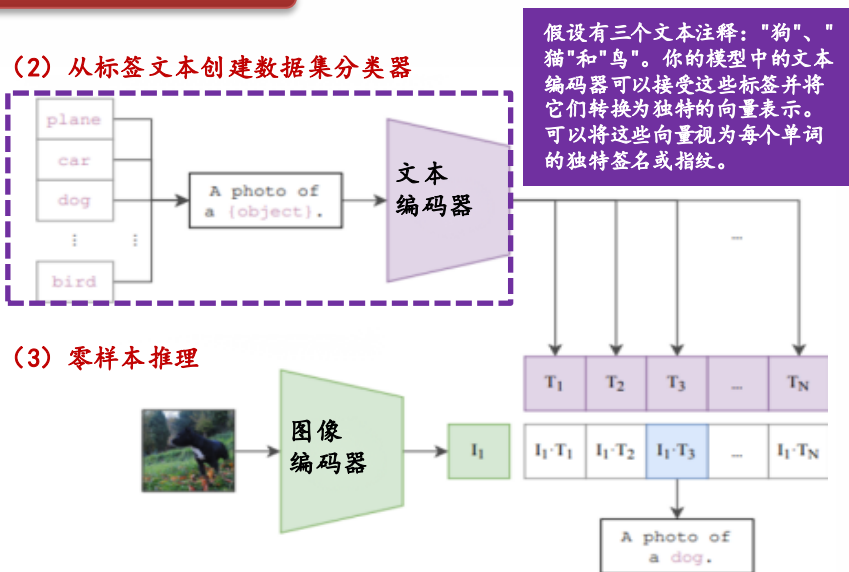
- **CLIP (对比性语言-图像预训练)** 利用图像的“alt-text”（一种为图像提供文字描述的方法，常用于网页以增强无障碍访问）来进行训练。此阶段的训练涵盖了大量图像与其相关文本描述的匹配对，确保模型精确捕捉到文本与图像间的关联性。如果图片和文字有对应关系，那么在高维特征空间中代表两者的点就会非常接近。经过无数图像-文本配对训练的文本编码器可以接受任何给定的文本标签并产生一个独特的向量表示。当提供一个图像时，其对应的向量表示将与这些文本向量进行比较，以分配最合适的类别。
- **CLIP在零样任务中展现出卓越的泛化性能和对于领域偏移的高度鲁棒性。** CLIP不仅能够应对视觉识别和图像-文本的联合检索任务，更在接受适当的提示后表现出强大的视觉-语言推断能力。**由于CLIP基于简单的跨模态交互，它在推理效率上超越了基于交叉注意力的视觉-语言模型。**

CLIP (对比性语言-图像预训练)

(1) 对比预训练



(2) 从标签文本创建数据集分类器

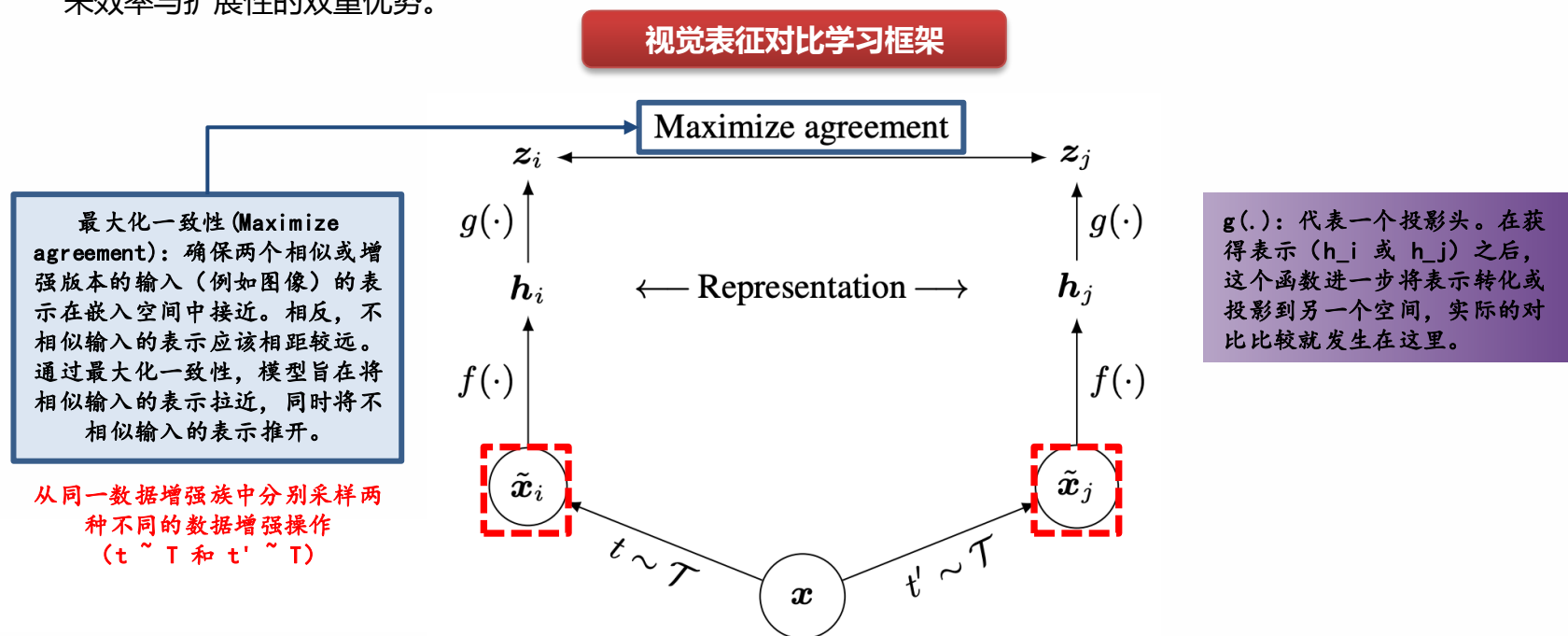


(3) 零样本推理

给定一个新的动物图像，图像编码器为这个图像产生一个向量表示。为了对这个图像进行分类，模型将比较图像的向量与文本注释的向量。假设图像的向量最接近“狗”的向量。那么，模型将把图像分类为“狗”。

1.1.3 图像理解能力提升：图像自监督预训练——对比/非对比学习

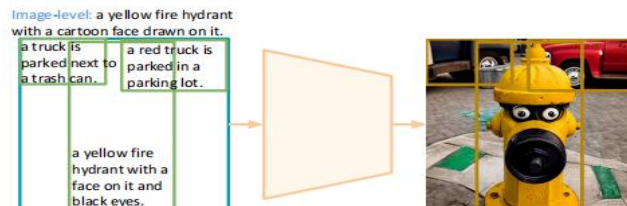
- 在没有标签的无监督学习场景中，对比学习通过将正例（相似的样本）和负例（不相似的样本）对比，强化模型对于相似性和差异性的判别能力。通过这种方法，即便在没有明确的类别标签的情况下，模型也能够被训练成可识别出哪些特征更有可能来自同一分布或类别。对比学习的成功归功于其能够捕捉到数据的内部结构和模式。在图像处理中，对比学习可以用来学习视觉表征，这些表征能够区分不同的物体和场景。一旦模型学习有区分性的特征，它将被用于多种下游任务，如分类、检索和聚类，并展现出强大的泛化能力。
- 近年来，非对比学习逐渐崭露头角，不再重视明确的样本对比，而是致力于挖掘数据的内在结构和不变特性。例如基于Siamese架构的模型，可以有效处理多个数据视图，而不必过多地依赖样本之间的对比性质，这为模型训练带来效率与扩展性的双重优势。



1.2 视觉生成：多模态内容理解和生成的闭环

- ❑ 在多模态大型模型的发展过程中，视觉生成技术是整合视觉内容与文本信息的关键手段。拥有理解和生成视觉内容的能力使模型能够参与到更深层次和细致的任务中，如图像标注、视觉叙事以及复杂的设计任务。由于人类的理解和沟通本质上是多模态的，通过整合视觉生成功能，模型能够以更接近人类认知的方式处理和生成信息。此外，多模态融合也为虚拟现实、增强现实以及交互式数字平台等领域奠定基础，未来将实现更为自然、无缝且高效的用户界面。
- ❑ “人类行为对齐”旨在融合人类的认知过程和AI驱动的视觉内容生成。传统视觉生成模型仅依赖数据，导致输出内容往往缺乏人类的逻辑思维和行动模式。例如，在生成‘繁忙的街道’图像时，传统模型只机械地排列车辆、行人和商店，无法捕捉人类对‘繁忙’的深刻理解，例如行人的匆忙步伐、车辆的密集流动和商店的热闹场景。而通过引入人类行为对齐，模型可确保生成的视觉内容不仅准确，更与人类的感知和期望相符合，推动模型向更以用户为中心的方向转变，使系统在‘思考’和‘感知’方面更符合人类需求。在遵循人类意图合成所需视觉内容的图像生成模型方面，主要涉及四个方向：
1) 空间可控的T2I生成；2) 基于文本的图像编辑；3) 更好地遵循文本提示；4) 在T2I生成中实现物体定制化。

(1) 空间可控的T2I生成



(2) 基于文本的图像编辑



(3) 更好地遵循文本提示



(4) 在T2I生成中实现物体定制化

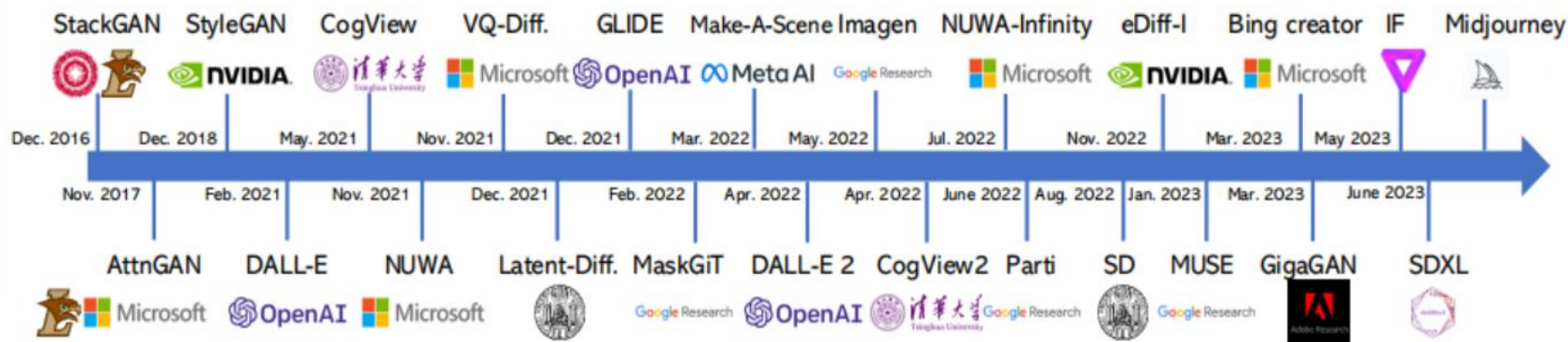


1.2.1 视觉生成：文本到图像（T2I）

□ 在人工智能领域，文本到图像（T2I）生成技术的目标是根据输入的文本语义生成视觉质量高的图像。目前主流的图像生成技术包括生成对抗网络（GAN）、变分自编码器（VAE）、离散图像标记预测和扩散模型。

- ① **GAN**：包括生成器和判别器两个关键组件。生成器从随机噪声输入创建合成图像，并根据文本条件调整噪声输入以生成与文本语义相关的图像。判别器的任务是区分合成生成的图像和真实图像。
- ② **VAE**：通过编码器和解码器网络模块来生成图像。编码器优化将图像编码成潜在表示，解码器则将采样的潜在表示转换回新图像。VQ-VAE通过离散的潜在空间和向量量化进一步改进VAE。
- ③ **离散图像标记预测方法**：核心是图像标记器和去标记器的配对。VQ-GAN将连续的视觉信号转换成离散标记，进而转化为离散标记预测任务。自回归Transformer逐行生成视觉标记，最终通过去标记化得到图像预测结果。
- ④ **扩散模型**：通过随机微分方程将随机噪声变为图像，经过多次迭代的去噪过程，基于输入文本的条件精炼图像。

T2I生成技术发展历程



1.2.2 视觉生成：基于文本的图像编辑

- ❑ **基于文本的图像编辑技术**依赖于已有的图片和文本描述来合成新图像。其核心目的是维持图像的主体视觉内容，对特定区域进行微调，如局部物体的调整或全局的风格改变，以更准确地满足用户的意图。
- ❑ **基于文本的编辑领域有3种主流功能。**1) **针对图像的特定区域进行修改**：如物体的增减或属性更改。用户所提供的区域蒙版与图像生成中的空间潜在变量操作结合。2) **针对图像的特定区域进行修改2.0**：用户简单描述所需区域外观，即可作为明确的编辑指示，例如指导“将图像中的A物体替换为B物体”。3) **专业模块融合**：相较于单一T2I模型的编辑扩展，某些编辑系统已开始融合多种专业模块，包括图像分割模型和大型语言处理模型。

基于文本的图像编辑示例



指令：“将房间变成彩色的”
大模型自动调整权重给“彩色的”指令

目 录

1 技术原理：图像理解+视觉生成持续发展，助力多模态能力提升

1.1 图像理解能力提升：三大视觉学习方法

1.2 视觉生成：多模态内容理解和生成的闭环

2 当前变化：GPT4-Turbo模型优化，GPTs生态加速繁荣

2.1 GPT-4 Turbo：多模态能力增强，使用成本降低

2.2 GPTs：加入Agents战场，与开发者共享收益

2.3 Assistant API：解决API开发者痛点，拓展OpenAI收入来源

3 未来影响：算力&存力亟需扩容，AI应用潜力无限，垂类数据或成关键

3.1 算力&存力端：流量激增+多模态生成拉大算力缺口，同时带动存力需求

3.2 应用端：强化垂类属性&满足长尾需求，数量有望实现快速增长

3.3 数据端：专业数据或成关键，重视垂类数据留存及知识产权价值

2 GPT4-Turbo模型优化，GPTs生态加速繁荣

章节小结

当前变化

GPT-4-Turbo

- | | | |
|--------------|--------|------------------|
| 1) 更长的上下文 | -----> | 可以支持更长的工作流 |
| 2) 更丰富的世界知识 | -----> | 知识储备能够实现迅速更新 |
| 3) 更多的控制 | -----> | 开发者能更准确地控制函数和API |
| 4) 多模态支持 | -----> | 图片/语音理解能力和创作效率提升 |
| 5) 定制化 | -----> | 特定领域的训练和应用进一步发展 |
| 6) 更高的频率限制 | -----> | 支持开发者完成更多工作 |
| 7) 大幅降低开发者成本 | -----> | 以更低的成本实现更强的性能 |

GPTs

- | | | |
|--|--------|------------------------------|
| 1) 什么是GPTs | -----> | 用户定制化ChatGPT，实现私人化和场景化 |
| 2) 如何使用 | -----> | 语言指令方便简捷，程序集成赋予多模态能力 |
| 3) 如何创建 | -----> | 无需编程即可构建，大幅降低创作门槛 |
| 4) 如何分发 | -----> | GPT Store推出在即，OpenAI与开发者收益共享 |
| 5) OpenAI Agent生态 VS Open source Agent生态 | -----> | iOS VS 安卓 |

Assistant API

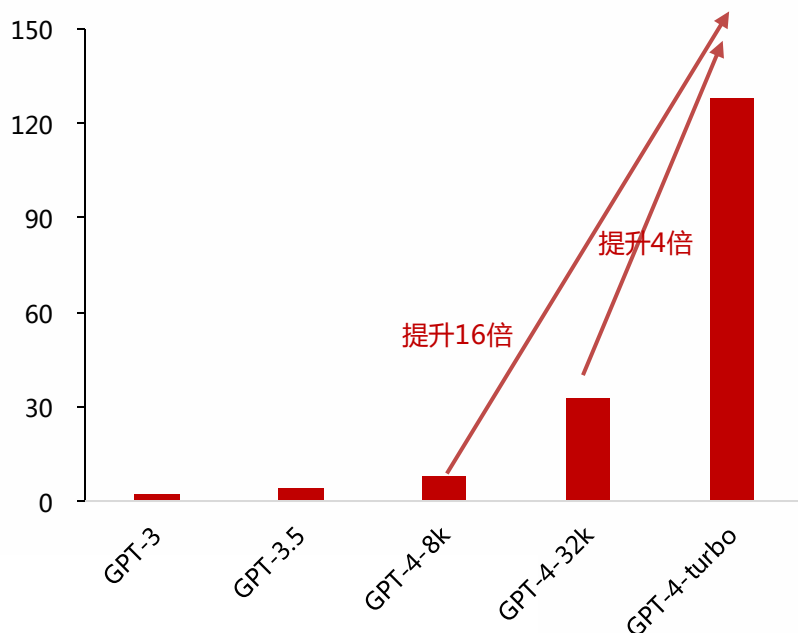
- | | | |
|----------|--------|------------------------|
| 1) 函数调用 | -----> | 减轻开发者调用函数工作，助力AI赋能UI |
| 2) 检索功能 | -----> | 助手获取更多知识，简化环节推动使用效率提升 |
| 3) 代码解释器 | -----> | 赋予AI写作能力、及时执行代码、甚至生成文件 |

资料来源：西南证券

2.1.1 GPT-4 Turbo : 文本理解能力再次提升，知识储备迅速更新

- ❑ **1) 更长的上下文 (Context Length)** : GPT-4-turbo支持的上下文窗口 (128k) 相较于GPT-4-8k提升16倍, 相较于GPT-4-32k提升4倍, 即GPT-4-turbo可在单个prompt中处理超过300页的文本, 且GPT-4-turbo模型在较长的上下文中更加准确。我们认为GPT-4-turbo更长且更准确的上下文处理能力将支持更长的工作流, 在B端有能力承担更多的工作负载, 提升用户体验。
- ❑ **2) 更丰富的世界知识 (Better Knowledge)** : 外部文档和数据库的截止更新日期从21年9月更新至23年4月, 意味着OpenAI的大模型在半年内已学习互联网一年半的知识, 学习速度极快。

GPT-4-turbo支持更长上下文窗口



资料来源: OpenAI官网, 西南证券整理

GPT-4-turbo更新知识截点至23年4月



资料来源: OpenAI开发者大会, 西南证券整理

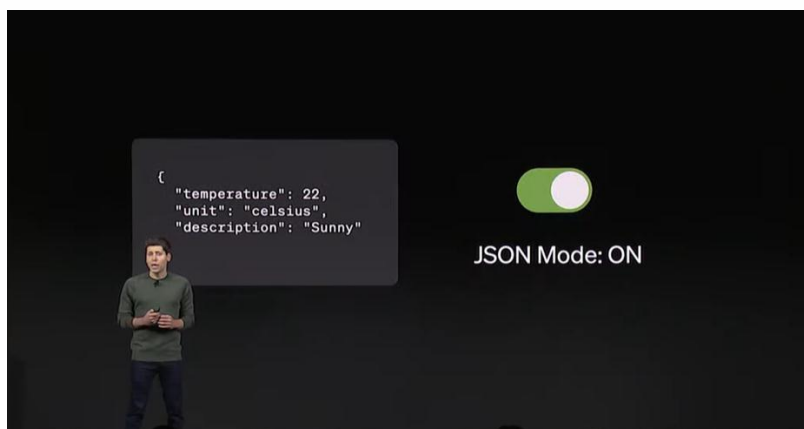
2.1.2 GPT-4 Turbo : 多工具帮助调用函数 , 开发者控制能力进一步加强

❑ 3) 更多的控制 (More Control) :

- 支持JSON Mode新模式 : 在JSON模式下 , 新的API参数response_format使模型能够约束模型输出 , 以生成语法正确的JSON对象。
 - 更好的Function calling : 用户可以在发送一条消息时请求多个操作 , 且函数调用的准确性提升。
 - Reproduceable outputs : 通过 seed 参数控制输出可重复的内容。
- ❑ → 我们认为 , 开发人员通过运用GPT-4-turbo模型将对函数和API有更准确、更高的控制能力。

GPT-4 Turbo支持JSON Mode新模式

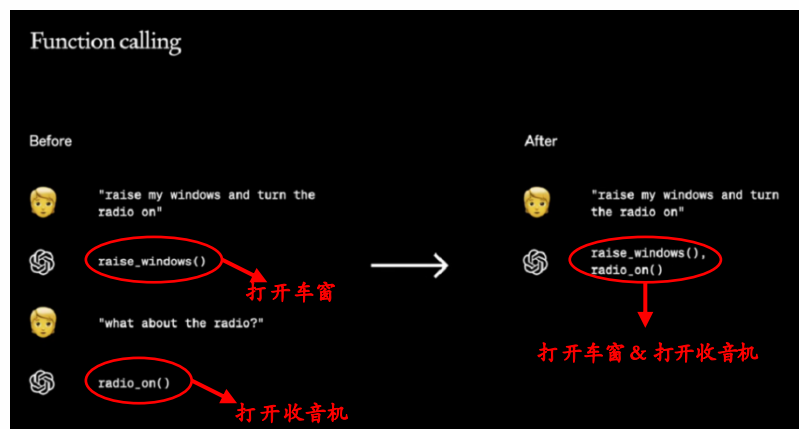
Json Mode 下开发者可以更方便的调用 API



资料来源 : OpenAI开发者大会 , 西南证券整理

GPT-4 Turbo支持更好的function calling

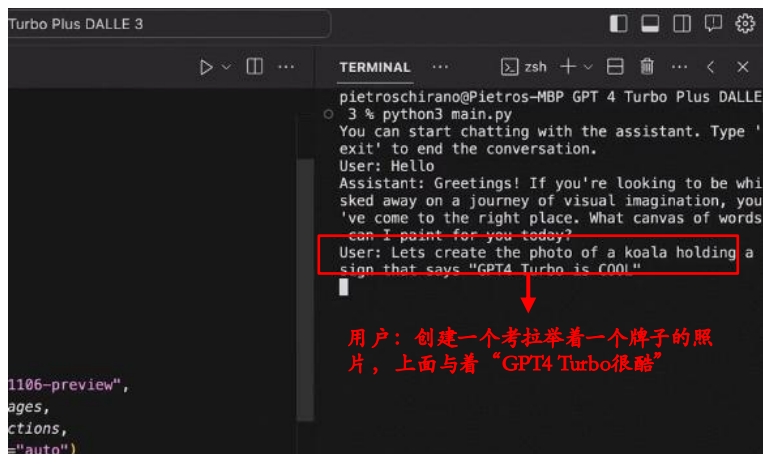
- **Before** : GPT-4在一条消息请求中只能调用其中一个函数、完成一个操作 , 完成多个操作需要进一步请求 ;
- **After** : GPT-4-turbo在一条信息请求中可以调用多个函数并完成多个操作 , 且函数调用的准确性得到提升。



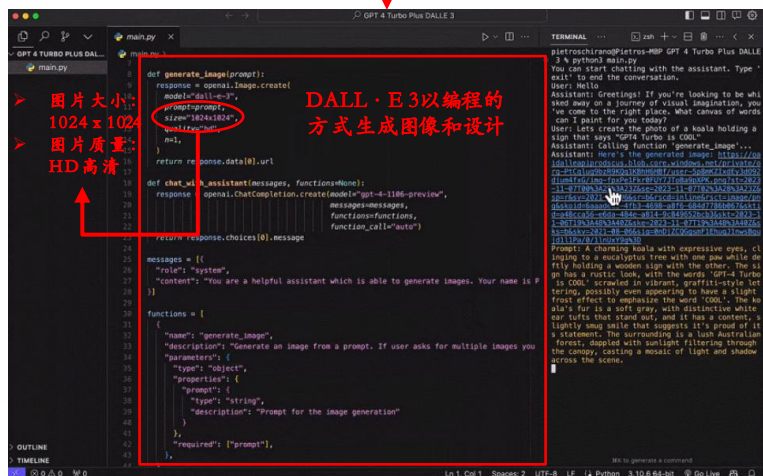
资料来源 : OpenAI开发者大会 , 西南证券整理

2.1.3 GPT-4 Turbo : 推出多个API , 多模态能力实现跨越

- ❑ 4) 【多模态能力—文生图】GPT-4 Turbo + DALL·E 3 : 开发者只需指定DALL·E 3作为模型, 即可以通过Images API将DALL·E 3直接集成到开发者的应用程序和产品中。根据大会演示, Snap、可口可乐和Shutterstock等公司已经使用DALL·E 3为客户生成图像和设计。当前API已内置适度的审核功能, 帮助开发者保护其应用程序免受滥用。
- ❑ 价格: 每张图像0.04美元起, 根据不同的格式和质量而定。

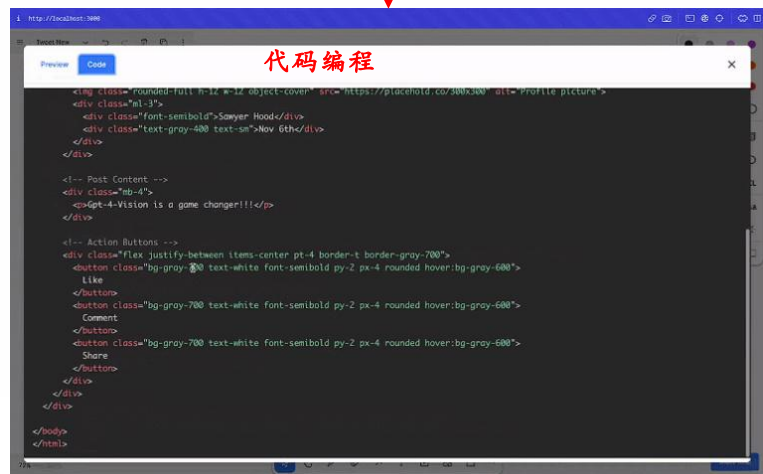


用户: 创建一个考拉举着一个牌子的照片, 上面写着 "GPT4 Turbo很酷"

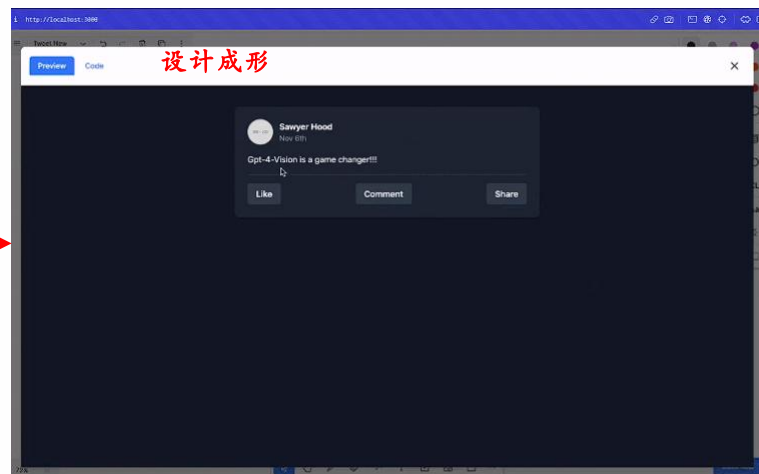


资料来源: OpenAI, 机器之心, 西南证券整理

2.1.3 GPT-4 Turbo : 推出多个API , 多模态能力实现跨越



- ❑ 4) 【多模态能力—图生文or图生图】GPT-4 Turbo + GPT-4V : OpenAI于2023年9月25日发布视觉多模态大模型GPT-4V, GPT-4V既能像原版 ChatGPT一样通过文字聊天, 也能读懂用户在聊天中给到的图像, 通过开放GPT-4V API, 开发者可以用OpenAI最新的GPT-4 Turbo (视觉版) 来开发新应用。
- ❑ 图像理解能力提升&创作大幅提效: 以设计UI界面为例, 几个小时的工作变成几分钟。



资料来源: OpenAI, 机器之心, 西南证券整理

2.1.3 GPT-4 Turbo : 推出多个API , 多模态能力实现跨越

- ❑ 4) 【多模态能力—文字生成语音】 GPT-4 Turbo + TTS : 开发者可以通过文本转语音API生成高质量的语音。新TTS模型提供六种预设的语音选择, 以及两种模型变体 (TTS-1和TTS-1-HD)。TTS针对实时应用进行优化, 而TTS-1-HD针对质量进行优化。
- ❑ 价格 : 定价从每1,000个字符的输入开始, 价格为0.015美元。

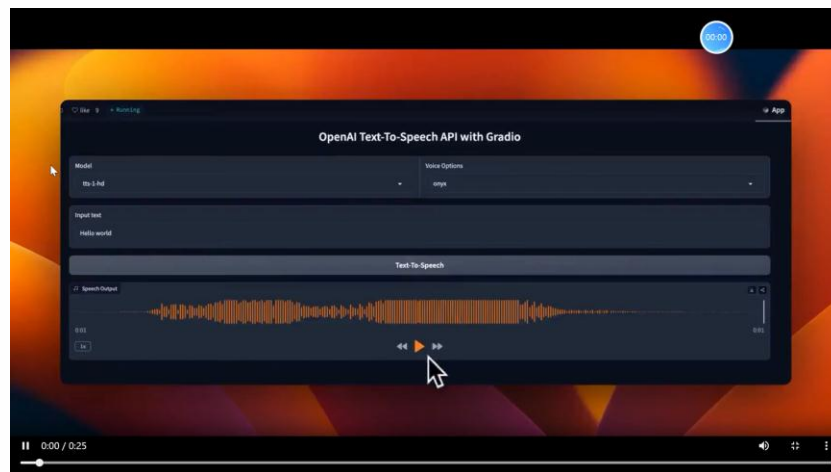
GPT-4V+TTS : 完成足球解说视频

- ① 提取视频帧创建结构化提示;
 - ② 发送GPT请求;
 - ③ 制作语音解说提示
 - ④ 生成语音解说脚本
 - ⑤ TTS API发送请求将脚本转换为音频
 - ⑥ 将音频和视频相结合
- 视频共1131帧, 每10帧选1张图
 - 足球比赛视频AI解说共花费30\$



TTS可根据标点符号生成不同语气的语音

- “hello world (无标点符号)” => 语气平缓
- “hello world ! ” => 语调略高
- “hello world ? ? ? ” => 略带疑问语气
- “hello world … ” => 语气低沉



资料来源 : OpenAI , 新智元 , 西南证券整理

资料来源 : OpenAI , 新智元 , 西南证券整理

2.1.4 GPT-4 Turbo：微调推动定制化发展，速率翻倍加速任务运行

- ❑ **5) 定制化 (Customization)**：开发者可以针对GPT-4和GPT-3.5进行微调，并面向大公司推出深度定制的Custom Models，**定制化模型意味着针对特定领域的训练和应用将进一步发展。**
- ❑ **6) 更高的频率限制 (Higher rata limits)**：GPT-4付费客户的每分钟令牌限制将翻倍，并且可以在API设置中申请更高的频率限制。通常，速率限制有五种衡量方式：RPM（每分钟请求）、RPD（每天请求）、TPM（每分钟令牌）、TPD（每天令牌）和IPM（每分钟图像）。**更高的频率限制将支持开发者完成更多工作。**

OpenAI部分老模型的速率限制

Free tier rate limits

This is a high level summary and there are per-model exceptions to these limits (e.g. some legacy models or models with larger context windows have different rate limits). To view the exact rate limits per model for your account, visit the [limits](#) section of your account settings.

MODEL	RPM	RPD	TPM	TPD
gpt-3.5-turbo	3	200	20,000	
text-embedding-ada-002	3	200	150,000	
whisper-1	3	200		
tts-1	3	200		
da11-e-2	5 img/min			
da11-e-3	1 img/min			

* The models `gpt-4-1106-preview` and `gpt-4-vision-preview` are currently under preview with restrictive rate limits that make them suitable for testing and evaluations, but not for production usage. We plan to increase these limits gradually in the coming weeks with an intention to match current gpt-4 rate limits once the models graduate from preview. As these models are adopted for production workloads we expect latency to increase modestly compared to this preview phase.

每分钟可请求得的tokens数翻番

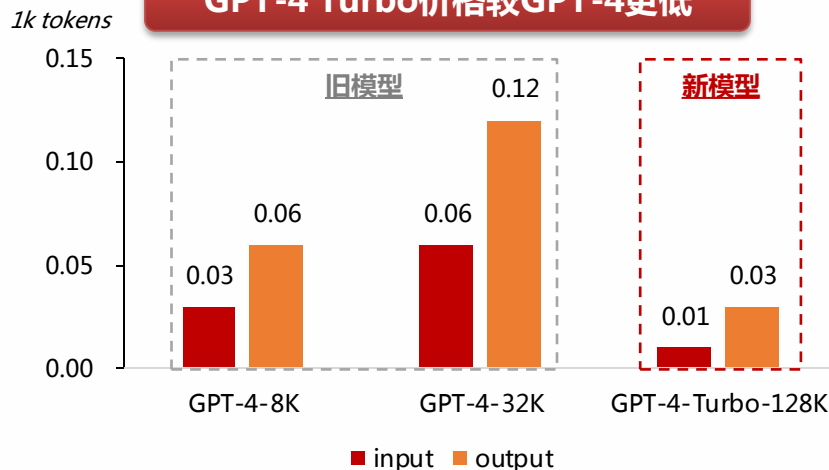


资料来源：OpenAI官网，西南证券整理

资料来源：OpenAI开发者大会，西南证券整理

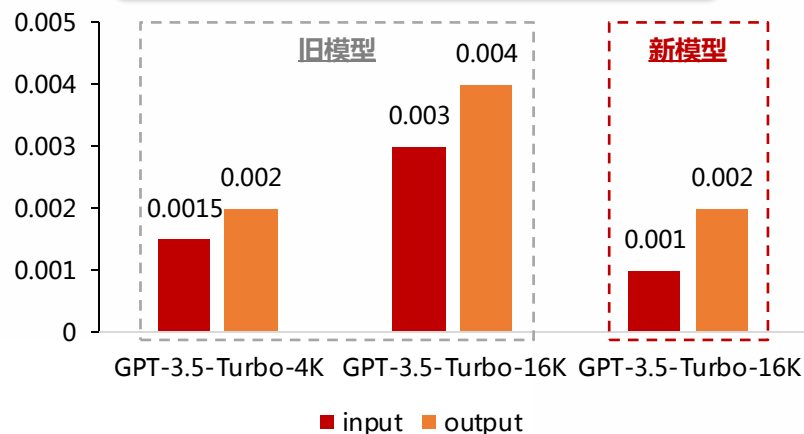
2.1.5 GPT-4 Turbo : 更强的性能，更低的成本

GPT-4 Turbo价格较GPT-4更低



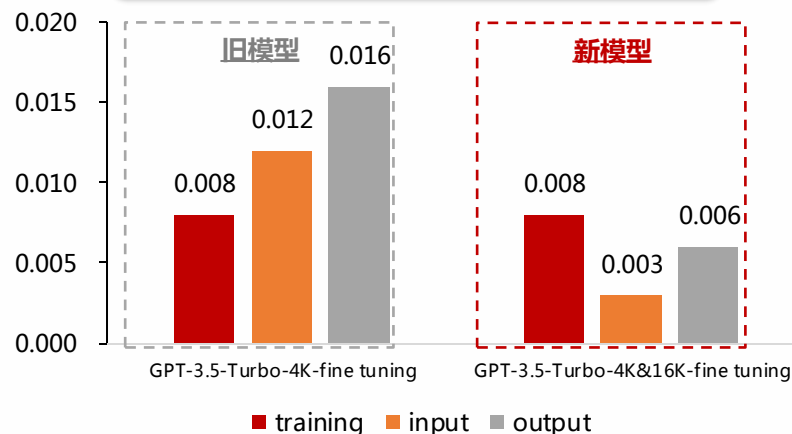
资料来源：OpenAI官网，西南证券整理

GPT-3.5 Turbo新模型价格更低



资料来源：OpenAI官网，西南证券整理

GPT-3.5 Turbo微调新模型价格更低



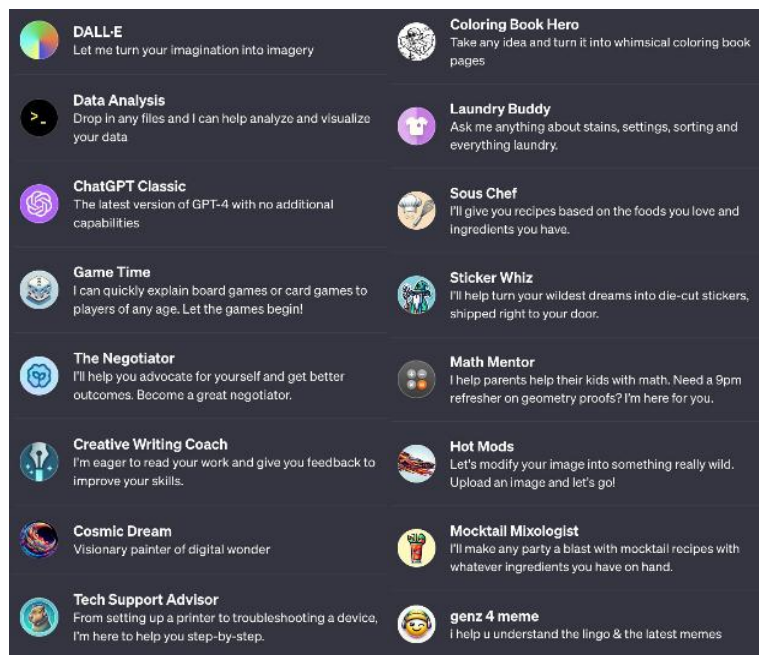
资料来源：OpenAI官网，西南证券整理

- 7) 大幅降低开发者成本。① GPT-4-Turbo的输入令牌价格比GPT-4-8K便宜3倍，价格为\$0.01，输出价格则便宜2倍，价格为\$0.03。② GPT-3.5-Turbo-16K的输入价格比之前的16K模型便宜3倍，价格为\$0.001，输出标记便宜2倍，价格为\$0.002。③ 通过微调，新的GPT-3.5-Turbo-fine tuning模型支持与4K相同价格的16K上下文，且GPT-3.5-Turbo-4K-fine tuning模型的输入价格减少4倍为\$0.003，输出价格降低2.7倍为\$0.006。

2.2.1 什么是GPTs：用户定制化ChatGPT，实现私人化和场景化

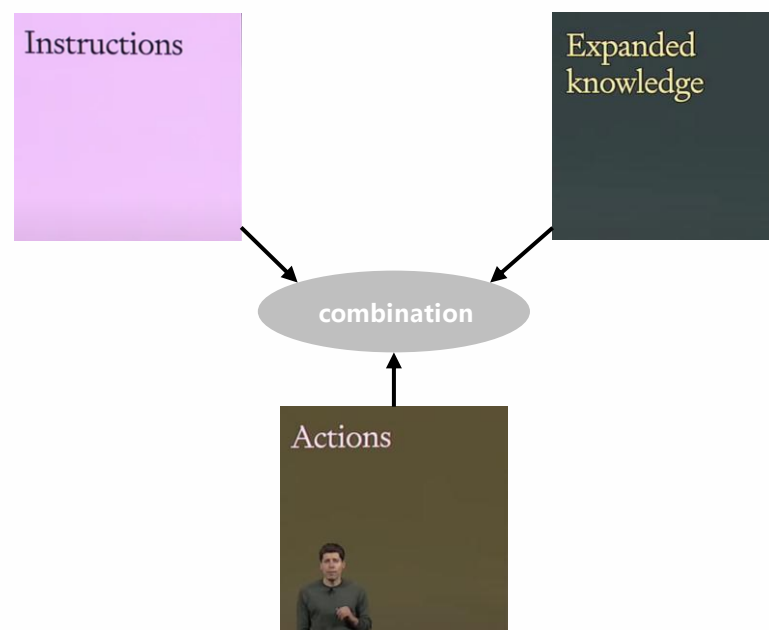
- ❑ **什么是GPTs**：GPTs是指“针对特定目的定制的ChatGPT”，用户可以通过自定义行为创建一个定制版的ChatGPT，定制版的ChatGPT具备带有任何功能的可能性（在保证隐私和安全的情况下）。不论是开发者还是不会写代码的普通人，都可以拥有自定义版本的GPT。
- ❑ → **更加个性化、私人化、场景化，每个人都可以拥有自己的AI Agent**。GPTs通过结合①说明/Instructions+②扩展的知识/Expand knowledge+③操作/Actions，能够在很多情况下更好地工作，并且为用户提供更好地控制，用以帮助用户轻松完成各种任务、或者获得更多乐趣。

OpenAI首批上线的16个AI Agent



资料来源：OpenAI官网，西南证券整理

GPTs结合说明、扩展的知识、及操作

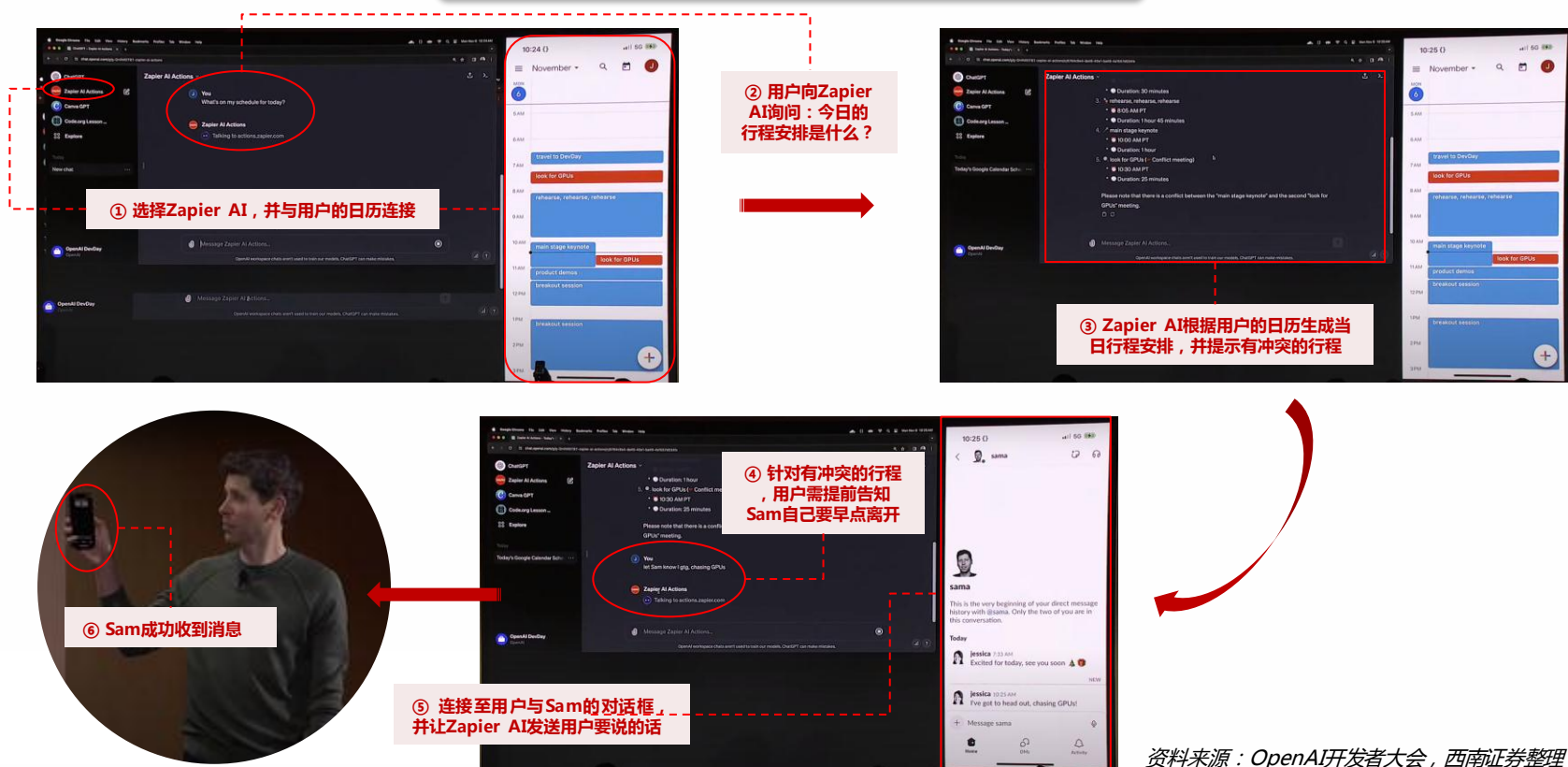


资料来源：OpenAI开发者大会，西南证券整理

2.2.2 如何使用GPTs：语言指令方便简捷，程序集成赋予多模态能力

- ❑ 如何使用GPTs：在ChatGPT中直接调用任意GPTs，用户提出要求，并可以通过继续聊天使GPTs生成的答案不断迭代或完成一系列指令。目前仅向ChatGPT Plus和Enterprise用户开放。
- ❑ → 灵活集成外部程序，提升多模态能力。例如ChatGPT可通过集成Code.org帮助教师规划课程、通过集成Zapier AI实现日程管理，以及可以通过集成Canva设计并微调海报等。

OpenAI演示如何使用Zapier AI进行日程管理

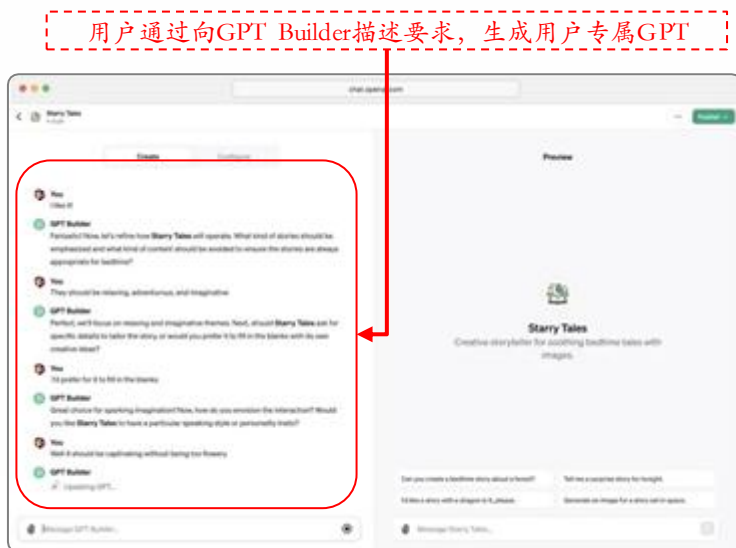


资料来源：OpenAI开发者大会，西南证券整理

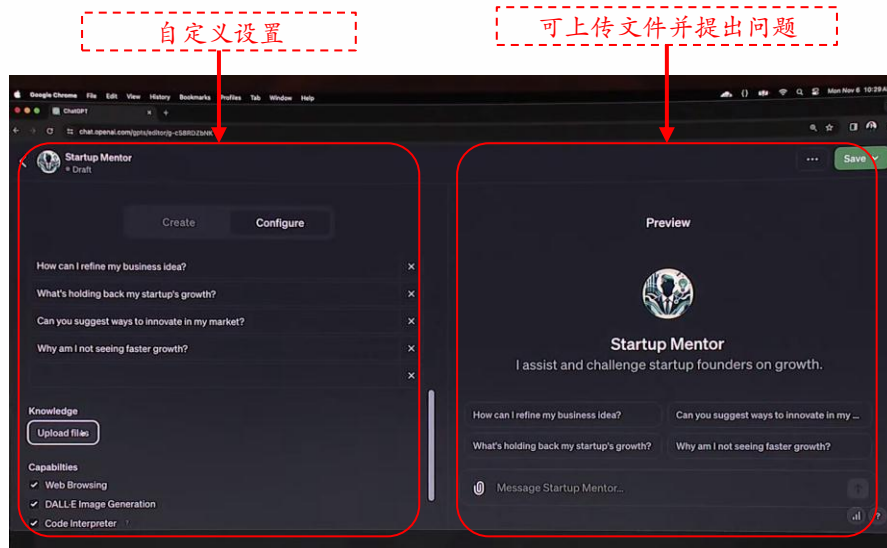
2.2.3 如何创建GPTs：无需编程即可构建，大幅降低创作门槛

- ❑ **如何创建GPTs**：OpenAI推出GPTs制作工具——GPT Builder。ChatGPT Plus和Enterprise用户只需要与GPT Builder进行语言交谈就可以进行定制化构建，用户向GPT Builder阐述需求，描述想要构建的GPT，GPT Builder即能生成专属GPT。
- ❑ **→大幅降低开发门槛，推动社区活跃与繁荣**。GPTs将由广泛的社区构建，用户无论是何种职业、无论是否知道如何制作工具或编程，均可以通过聊天对话打造一个专属于自己的个性化GPT，用户只需要提供基本信息、上传额外文件、明确要求即可。我们认为，GPTs不要求用户掌握编程，仅通过自然语言聊天即可打造专属GPT，将大大降低创作门槛、推动用户积极实践、AI Agent有望实现百花齐放，从而对市场上的AI Agent应用产生一定的威胁。

GPT Builder界面



Sam Altman演示构建“创业导师”



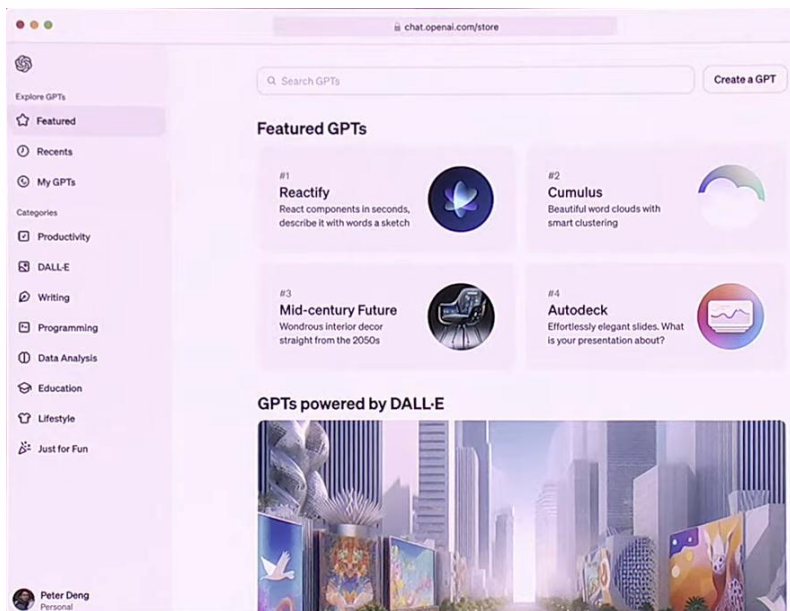
资料来源：OpenAI开发者大会，西南证券整理

资料来源：OpenAI开发者大会，西南证券整理

2.2.4 如何分发GPTs：GPT Store推出在即，OpenAI与开发者收益共享

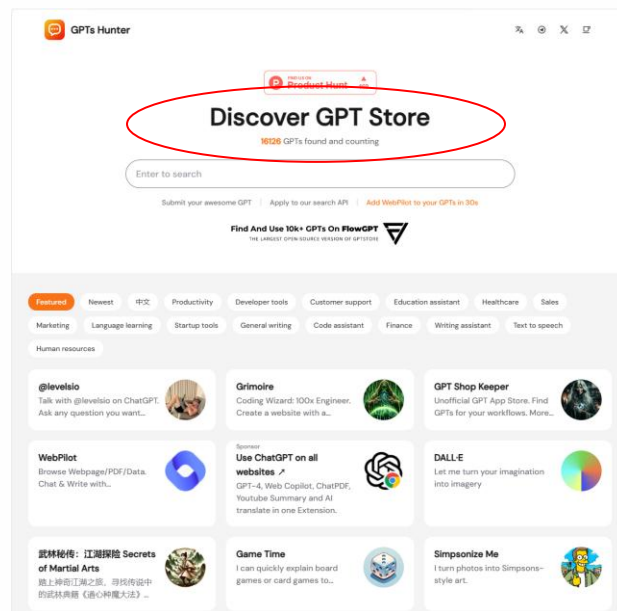
- ❑ **如何分发GPTs**：OpenAI计划于11月末推出GPT Store，用于收录并共享GPTs。未来，用户在创建个性化GPT之后，可以选择公开并共享，那么该GPT将**进入商店、且变得可搜索**。随着各种GPT在GPT Store汇集，ChatGPT Plus和Enterprise用户可以直接获取**开箱即用**的定制版ChatBot或AI Agent。以此同时，也将产生最受欢迎的GPT，并可能在**排行榜**上不断攀升。
- ❑ **→与开发者共享收益，构建完整生态**。未来，OpenAI将进一步推出收益计划，GPTs的构建者在提供定制化GPT后可以赚取一定的收益，并与OpenAI共享。该收益机制将大力推动创作社区的积极性，构建GPTs的盈利模式，完善Agent生态系统。

OpenAI GPT Store预期示意图



资料来源：OpenAI开发者大会，西南证券整理

GPTs Hunter已收录16000+ GPTs

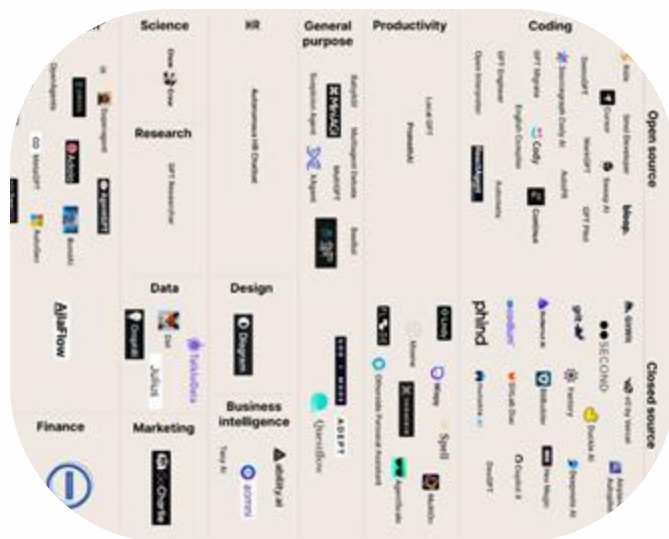


资料来源：GPTs Hunter（截至中国时间2023.11.20 9:00），西南证券整理

2.2.5 OpenAI Agent生态 VS Open source Agent生态



OpenAI Agent生态——GPTs Store



Open source Agent生态

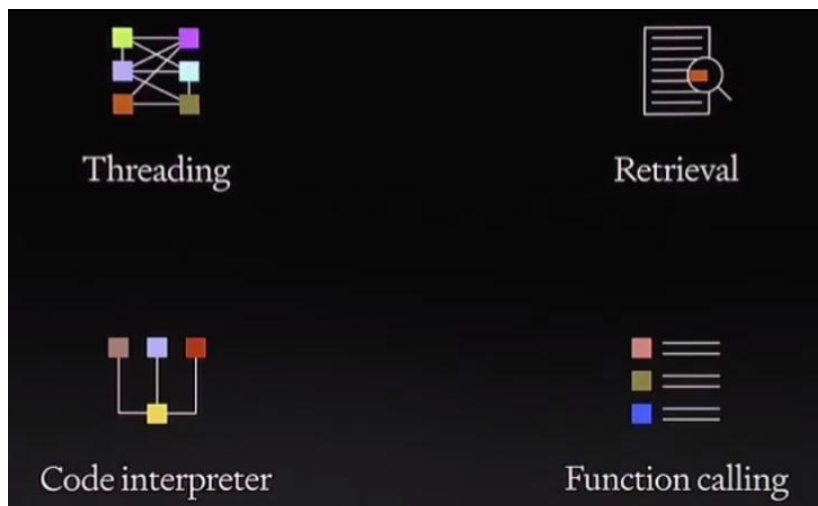
- **GPTs store有望如同Apple APP store**，通过鼓励开发者创建应用并在商店中展示，**进行收入分成**。以抖音充值斗币为例，苹果手机抖音APP中1人民币可购买7斗币；而安卓手机的抖音APP中1人民币可购买10斗币；苹果抽取约3成。
- **规模优势：预计OpenAI生态的AI Agent将在数量上具备绝对优势**。根据GPTs Hunter数据，截至中国时间11月13日21:30，GPTs Hunter已收录八千多个GPTs。

- **以Meta和Huggingface为代表，如同手机时代的安卓生态。**
- **破局之道：**
 - ① **价格/成本优势：**GPT-4 Turbo的输入token比GPT-4便宜3倍，为0.01\$/1000 tokens，输出token便宜2倍，为0.03\$/1000 tokens，尽管价格已显著下降，但相较于开源生态的大模型和Agent工具，价格仍然高出几倍。
 - ② **专业任务方面的优势：**开源AI Agent可以在特定任务方面打造高质量和轻量化优势。

2.3 Assistant API：解决API开发者痛点，拓展OpenAI收入来源

- ❑ **针对开发者在开发API中的痛点，OpenAI推出Assistant API，致力于为开发者赋能。**根据此前市场上推出的各种API，我们可以发现API通过接入各种程序和应用，有助于帮助应用实现特定功能。例如，Shopify的Sidekick允许用户在平台上进行操作；Discord的Clyde允许discord版主设置自定义人格；Snap my AI作为定制聊天机器人工具，可以添加至群聊中并提出建议。但以上API的构建可能需要开发者耗费几个月的时间、并由数十名工程师搭建，而目前Assistant API的推出将使其变得容易实现。
- ❑ **定价：除常规的tokens计费外，部分组件还需收取额外费用。**其中，代码解释器/Code interpreter定价为单次0.003美元；检索/Retrieval定价为0.20美元/GB/助理/天。

OpenAI Assistant API 核心组件



资料来源：OpenAI开发者大会，西南证券整理

OpenAI Assistant API 收费标准

Assistants API and tools (retrieval, code interpreter) make it easy for developers to build AI assistants within their own applications. Each assistant incurs its own retrieval file storage fee based on the files passed to that assistant. The retrieval tool chunks and indexes your files content in our vector database.

[Learn more](#)

The tokens used for the Assistant API are billed at the chosen language model's per-token input / output rates and the assistant intelligently chooses which context from the thread to include when calling the model

[Learn about Assistants API](#)

Tool	Input
Code interpreter	\$0.03 / session (free until 11/17/2023)
Retrieval	\$0.20 / GB / assistant / day (free until 11/17/2023)

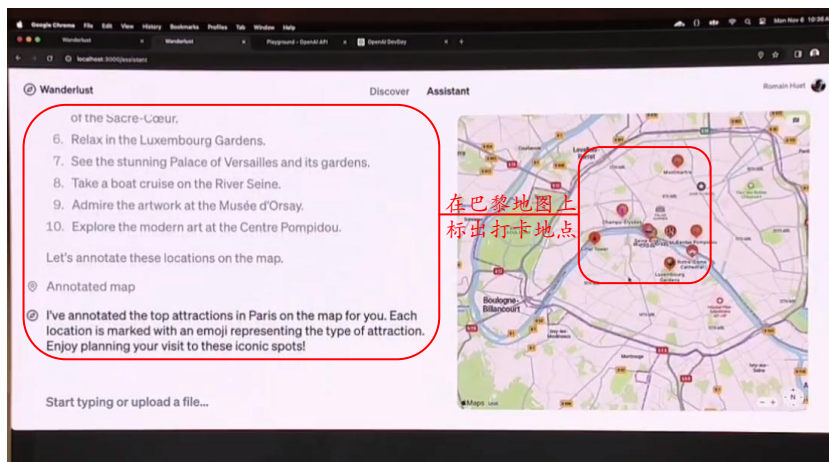
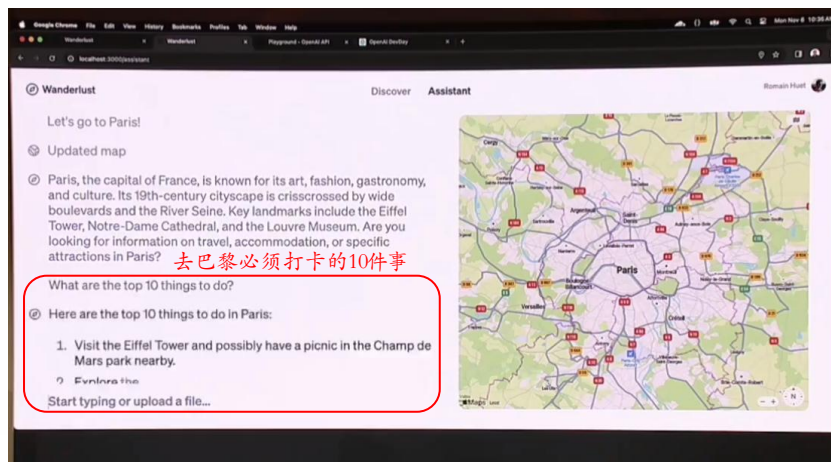
资料来源：OpenAI官网，西南证券整理

2.3.1 Assistant API-函数调用

- ❑ **1) 函数调用/Function calling**：在一次API调用/call中，用户可以描述函数/functions，让模型输出JSON对象来调用一个或多个函数。GPT-4-turbo经过训练，既可以检测何时应该调用函数（取决于输入），也可以保证JSON输出不会有延迟。
- ❑ **→减轻开发者调用函数工作，助力AI赋能UI**：该集成允许开发者通过自然语言就能与各种应用程序的组件和功能进行流畅的交互，实现AI与UI的更好融合。与此同时，开发者可以将自身更多的精力放在想法的创新，无需自己分析并调用函数，将脏活累活交给助手。

OpenAI开发者大会演示：在旅行应用中使用“函数调用”

① 用户提问“去巴黎必做的10件事” → ② 助手回答用户问题 → ③ 助手调用函数以在地图上进行标注



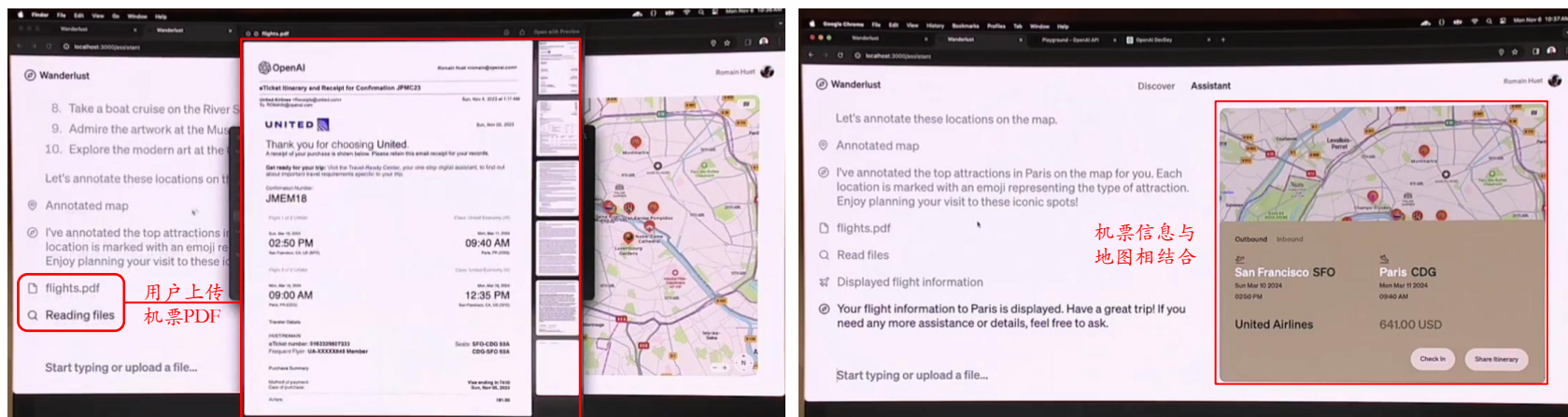
资料来源：OpenAI开发者大会，西南证券整理

2.3.2 Assistant API-检索功能

- ❑ **2) 知识库检索/Retrieval**：用户首先将文件上传至助手/Assistant，然后文件将在后台被自动分块、索引、存储和嵌入，助手再实施向量搜索以检索相关内容，最终结合用户要求回答问题。
- ❑ **→ 助手可以获取更多知识**：助手通过获取模型以外的文件信息，增强助手的专业化或个性化功能。
- ❑ **→ 用户无需自己解析文件**：助手具备解析长格式文档的能力，从普通文本到特定格式。
- ❑ **→ 省去部分环节&提升效率**：过去，开发者需要计算embeddings、设置分块算法、并在每次调用API时重新发送整个历史对话记录，例如设置键值存储（key value store）以处理上下文窗口和序列消息；而现在，助手可以帮助开发者省去以上环节，进一步提升效率。
- ❑ **→ 去黑盒化**：开发者可以在Dashboard的线程中看到历史步骤，如调用的函数和上传的文件等。

OpenAI开发者大会演示：在旅行应用中使用“检索”功能

① 用户上传已经订好的机票PDF → ② 后台“检索”读取文件 → ③ 机票信息将呈现在屏幕上，并与地图界面相结合



资料来源：OpenAI开发者大会，西南证券整理

2.3.3 Assistant API-代码解释器

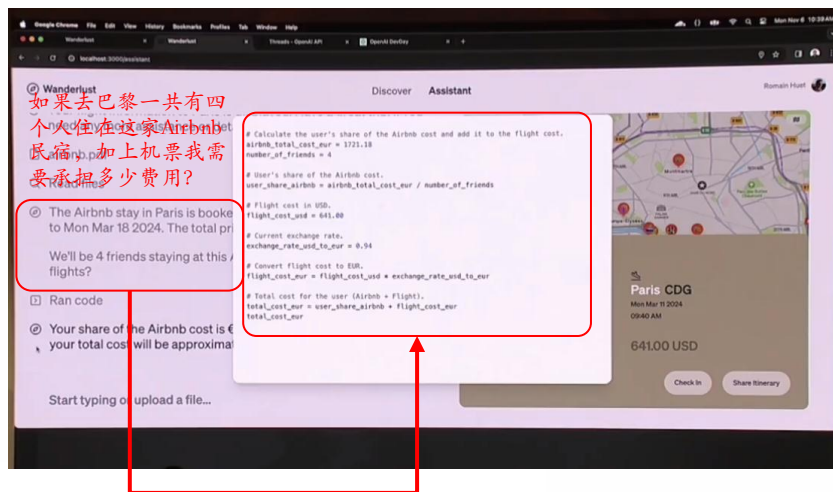
- ❑ **3) 代码解释器/Code interpreter**：代码解释器允许Assistant API在沙盒执行环境中编写和运行Python代码。该工具可以处理具有不同数据和格式的文件，并生成具有数据和图形图像的文件。代码解释器允许用户的助手迭代运行代码，以解决具有挑战性的代码和数学问题。当助手编写的代码无法运行时，它可以通过尝试运行不同的代码来迭代此代码，直到代码执行成功。
- ❑ → **代码解释器赋予人工智能写作的能力，并及时执行代码、甚至生成文件。**

OpenAI开发者大会演示：在旅行应用中使用“代码解释器”

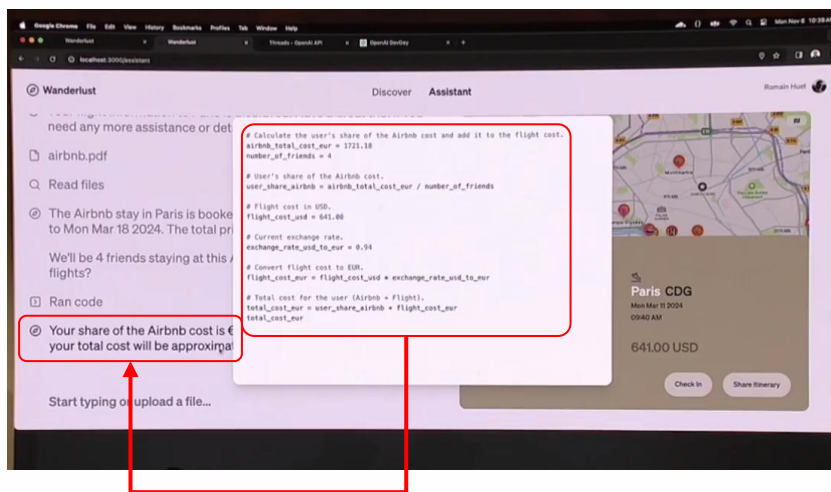
① 用户提出需要进行计算或编写代码的问题

② 代码解释器意识到需要编写代码来回答问题

③ 代码解释器即时编写代码进行运算并得出答案



代码解释器根据问题编写代码



代码解释器通过计算在巴黎的旅游天数、人员数量、汇率等，向用户给出答案

资料来源：OpenAI开发者大会，西南证券整理

目 录

1 技术原理：图像理解+视觉生成持续发展，助力多模态能力提升

1.1 图像理解能力提升：三大视觉学习方法

1.2 视觉生成：多模态内容理解和生成的闭环

2 当前变化：GPT4-Turbo模型优化，GPTs生态加速繁荣

2.1 GPT-4 Turbo：多模态能力增强，使用成本降低

2.2 GPTs：加入Agents战场，与开发者共享收益

2.3 Assistant API：解决API开发者痛点，拓展OpenAI收入来源

3 未来影响：算力&存力亟需扩容，AI应用潜力无限，垂类数据或成关键

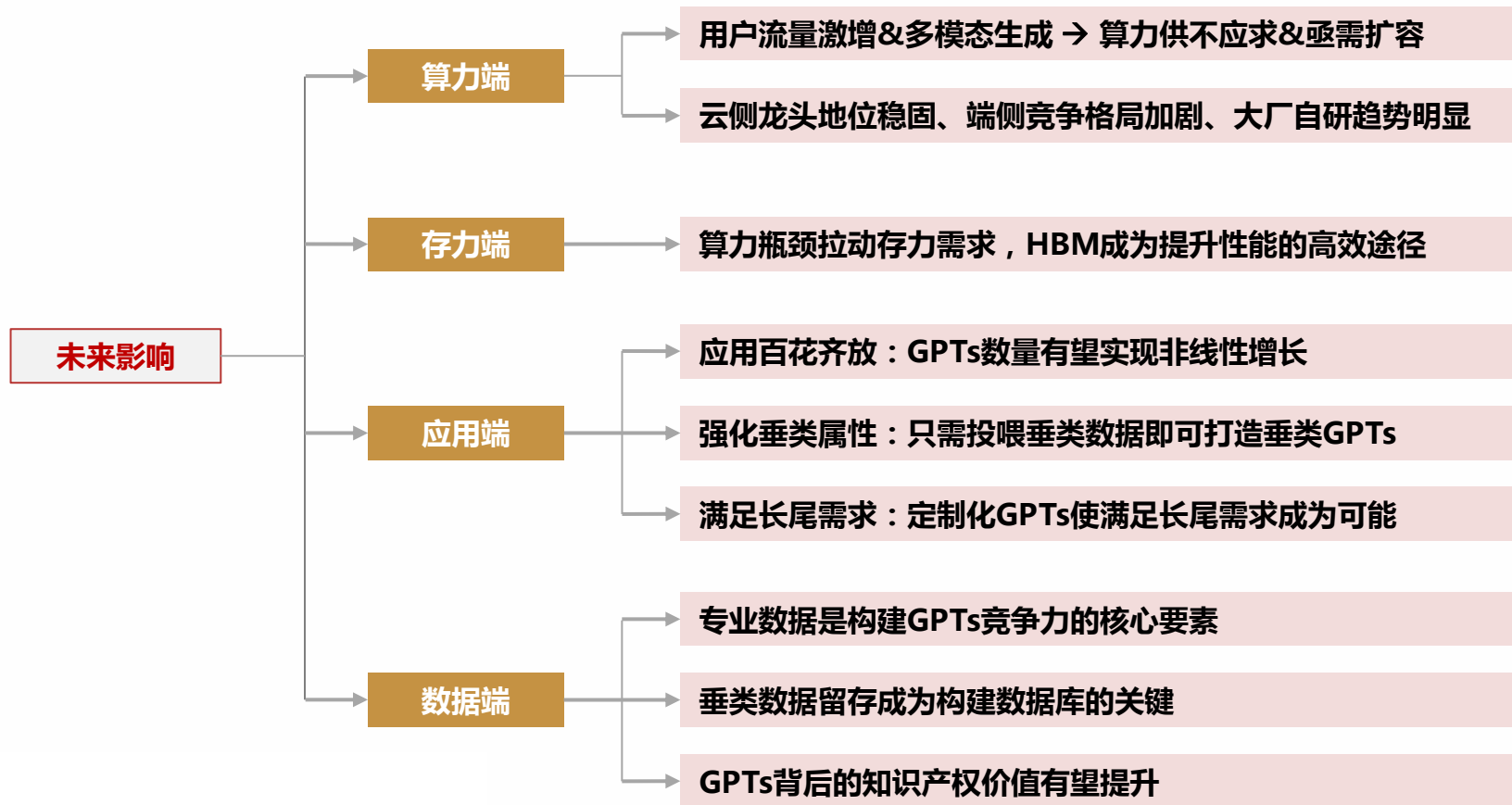
3.1 算力&存力端：流量激增+多模态生成拉大算力缺口，同时带动存力需求

3.2 应用端：强化垂类属性&满足长尾需求，数量有望实现快速增长

3.3 数据端：专业数据或成关键，重视垂类数据留存及知识产权价值

3 未来影响：算力存力亟需扩容，AI应用潜力无限，垂类数据或成关键

章节小结



资料来源：西南证券

3.1.1 算力端：用户流量激增+多模态生成，算力亟需扩容

- ❑ 1) 用户流量激增导致算力供不应求：OpenAI旗下GPT-4-Turbo、GPTs和Assistant API等产品的推出已打开新流量入口，更多的用户和开发者希望参与其中，巨大的流量对算力底座提出更高的要求。2023年11月7-9日，ChatGPT及API出现多次宕机，面对开发者大会后的巨大用户流量，OpenAI的算力基础设施频繁遇到容量瓶颈；11月15日，创始人Sam Altman在推特上表示，在开发者大会召开后，ChatGPT及一系列新产品的使用量出现激增，且远超内部预期，其算力负载短期难以为继，OpenAI为确保现有用户体验，决定停止ChatGPT Plus新用户注册，近期OpenAI服务器的不稳定以及对用户数量的限制已经表明全球算力亟需扩容。

11月7日ChatGPT&API宕机

Partial outage across ChatGPT and the API

Incident Report for OpenAI

Resolved	This incident has been resolved. Posted 7 days ago. Nov 07, 2023 - 22:21 PST
Monitoring	A fix has been implemented and we are monitoring the results. Posted 7 days ago. Nov 07, 2023 - 21:59 PST
Identified	API and ChatGPT has degraded once again – we're continuing the investigation. Posted 7 days ago. Nov 07, 2023 - 21:02 PST
Monitoring	A fix has been implemented and we're seeing services back to normal, but we are still monitoring to ensure we're fully restored. Posted 7 days ago. Nov 07, 2023 - 20:47 PST
Update	API and ChatGPT continue to have periodic outages – we have implemented a couple mitigations but are still investigating the issue. Posted 7 days ago. Nov 07, 2023 - 20:33 PST
Investigating	We are currently investigating this issue. Posted 7 days ago. Nov 07, 2023 - 19:52 PST

11月8日ChatGPT&API宕机

Major Outage across ChatGPT and API

Incident Report for OpenAI

Resolved	Between 5:42AM - 7:16AM PT we saw errors impacting all services. We identified the problem and implemented a fix. We are now seeing normal responses from our services. Posted 7 days ago. Nov 08, 2023 - 07:46 PST
Monitoring	A fix has been implemented and we are gradually seeing the services recover. We are currently monitoring the situation. Posted 7 days ago. Nov 08, 2023 - 07:33 PST
Identified	We've identified an issue resulting in high error rates across the API and ChatGPT, and we are working on remediation. Posted 7 days ago. Nov 08, 2023 - 06:50 PST
Update	We are seeing high error rates across the API and ChatGPT and are actively investigating possible causes. Posted 7 days ago. Nov 08, 2023 - 06:49 PST
Update	We are continuing to investigate this issue. Posted 7 days ago. Nov 08, 2023 - 06:02 PST
Investigating	We are currently investigating this issue. Posted 7 days ago. Nov 08, 2023 - 05:54 PST

11月9日ChatGPT&API宕机

Periodic outages across ChatGPT and API

Incident Report for OpenAI

Resolved	The incident has been resolved and status of our services have returned to normal. Posted 5 days ago. Nov 09, 2023 - 13:21 PST
Update	We are dealing with periodic outages due to an abnormal traffic pattern reflective of a DDoS attack. We are continuing work to mitigate this. Posted 6 days ago. Nov 08, 2023 - 19:49 PST
Update	We are seeing periodic outages across ChatGPT and the API still. Posted 6 days ago. Nov 08, 2023 - 17:23 PST
Update	We are continuing to monitor for any further issues. Posted 7 days ago. Nov 08, 2023 - 12:56 PST
Monitoring	A fix has been implemented and we are monitoring the results. Posted 7 days ago. Nov 08, 2023 - 12:41 PST
Investigating	We are currently investigating this issue. Posted 7 days ago. Nov 08, 2023 - 12:03 PST

资料来源：OpenAI官网，西南证券整理

3.1.1 算力端：用户流量激增+多模态生成，算力亟需扩容

- ❑ 2) 多模态生成大幅提高算力需求：图片等多模态生成所要求的tokens计算量将远高于文本模式，从而大幅提升模型训练和推理的算力需求。根据OpenAI官网信息，1000个tokens大约为750个words，因此在文本方面，1个单词对应为1.33个tokens。而多模态方面，在GPT-4-Turbo的Vision pricing calculator高保真度模式下，1张图片则需要765或1105个tokens，对应来看，1张图片所产生的tokens数大约是1个单词的570或830倍。

高保真度模式下1024x1024图像tokens数

Vision pricing calculator

1024 px by 1024 px = \$0.00765 ①

☐ Low resolution

Price per 1K tokens (fixed)	\$0.01
Resized width	768
Resized height	768
512 × 512 tiles	2 × 2
Total tiles	4
Base tokens	85
Tile tokens	170 × 4 = 680
Total tokens	765
Total price	\$0.00765

◆ 高保真度模式下的1024 x 1024图像：

- 由于1024小于2048，因此不用调整初始大小；
- 由于最短边是1024，因此系统将图像缩小到768 x 768；
- 因此需要4个512px的方形图块表示图像；
- 因此最终的tokens数为4 x 170 + 85 = 765 tokens。

高保真度模式下2048x4096图像tokens数

Vision pricing calculator

2048 px by 4096 px = \$0.01105 ①

☐ Low resolution

Price per 1K tokens (fixed)	\$0.01
Resized width	768
Resized height	1536
512 × 512 tiles	2 × 3
Total tiles	6
Base tokens	85
Tile tokens	170 × 6 = 1020
Total tokens	1105
Total price	\$0.01105

◆ 高保真度模式下的1024 x 1024图像：

- 由于1024小于2048，因此不用调整初始大小；
- 由于最短边是1024，因此系统将图像缩小到768 x 768；
- 因此需要4个512px的方形图块表示图像；
- 因此最终的tokens数为4 x 170 + 85 = 765 tokens。

- 图像输入以token计量和收费。图像token成本由两个因素决定：① 图像的大小；② 图片的保真度要求（low or high resolution）。
- 在低保真度模式下：图像每张花费85个token。
- 在高保真度下：模型首先将图像缩放至适合2048 x 2048的正方形，并保持其纵横比；然后对图片进行缩放，使图像最短边为768像素；最后计算图像应该由多少个512px的正方形组成，每个512px的方格花费170个token，且另外85个token始终会添加到最终总数中。

3.1.2 算力端：云侧龙头地位稳固、端侧竞争格局加剧、大厂自研趋势明显

❑ 1) 云侧芯片龙头：英伟达龙头地位稳固，AMD加速发展。

- ① **英伟达**：2023年11月13日，英伟达推出H200，内存方面首次采用HBM3e，容量高达141GB，带宽实现大幅提升；性能方面着重强化推理能力和HPC性能，可将Llama2模型的推理速度提高近一倍，相较于H100可降低50%的TCO和能耗成本。2023年以来，英伟达AI芯片已发布多个产品，在云侧算力芯片领域龙头优势明显。
- ② **AMD**：2023年6月，AMD正式发布MI300系列，MI300将CPU、GPU和内存封装为一体，大幅缩短DDR内存行程和CPU-GPU PCIe行程，提高性能和效率；MI300采用Chiplet设计，拥有13个基于3D堆叠的小芯片（5nm: 3个CPU，6个GPU；4个6nm芯片），包括24个Zen4 CPU内核，同时融合CDNA 3和8个HBM3显存堆栈，集成5nm和6nm IP，总共包含128GB HBM3显存和1460亿个晶体管。对比MI250加速卡，MI300可带来8倍AI性能和5倍每瓦性能的提升(FP8)，使ChatGPT和DALL-E等超大AI模型的训练时间可以从几个月缩短到几周。

英伟达主要AI芯片和AMD MI系列芯片参数情况

类别	英伟达主要AI芯片			AMD MI系列芯片		
产品名称	H200 SXM	H100 SXM	A100 SXM	AMD MI250	AMD MI250X	AMD MI300
发布时间	2023H2	2022/3/22	2020/5/14	2021/11/8	2021/11/8	2023H2
GPU架构	Hopper	Hopper	Ampere	CDNA2	CDNA2	CDNA3
专用显存	141 GB HBM3e	80 GB HBM3	80 GB HBM2e	128GB HBM2e	128GB HBM2e	128GB HBM3
总版卡功耗（TBP）	Up to 700W (configurable)	Up to 700W (configurable)	400W (for standard configuration)	560W Peak	560W Peak	600W Peak
峰值单精度（FP32）性能	67 TFLOPS	67 TFLOPS	19.5 TFLOPS	45.3 TFLOPS	47.9 TFLOPS	/
峰值双精度（FP64）性能	34 TFLOPS	34 TFLOPS	9.7 TFLOPS	45.3 TFLOPS	47.9 TFLOPS	/
互联	PCIe Gen5: 128 GB/s; NVLink: 900GB/s	PCIe Gen5: 128 GB/s; NVLink: 900GB/s	PCIe Gen4: 64 GB/s; NVLink: 600GB/s	PCIe® 4.0 x16; Fabric™ Link: 100GB/s	PCIe® 4.0 x16; Fabric™ Link: 100GB/s	/

资料来源：英伟达官网，AMD官网，西南证券整理

3.1.2 算力端：云侧龙头地位稳固、端侧竞争格局加剧、大厂自研趋势明显

❑ 2) 端侧芯片格局：对于AI PC端侧适用芯片，英特尔进展保持领先，高通有望实现从0到1。

- ① **英特尔**：PC主芯片包括CPU和GPU。GPU方面，根据Statista数据，英特尔在22Q4全球PC GPU市场中占据71%的市场份额；CPU方面，根据Counterpoint Research数据，英特尔在2022年全球Notebook CPU/SoC市场中占据69.6%的市场份额，当前英特尔在PC主芯片市场中处于主导地位。23H2，英特尔推出Meteor Lake处理器，助力PC效能提升，专为AI任务设计，在AI PC主芯片市场中具备先发优势。
- ② **高通**：2023年10月，高通于骁龙峰会上推出骁龙X Elite芯片，可支持130亿参数大模型，为ows-on-Arm笔记本设计，预计在2024年中期发布。算力方面，AI PC对PC芯片的算力要求更高，高通在端侧AI推理能力优于英特尔。与此同时，生态方面，2022年至今ows开始支持高通，已发布多轮支持Arm架构芯片的操作系统。未来，高通在以骁龙X Elite为代表的AI PC芯片的助力下，将在PC领域实现重要突破，逐步抢占市场份额。

英特尔和高通在AI PC芯片上的布局

厂商	芯片	CPU	GPU	NPU	其他
英特尔	Meteor Lake处理器 	1) 采用Intel 4制程工艺； 2) 具有快速响应能力，适合需要快速决策和低延迟的轻量级AI任务；	1) 采用台积电N5工艺构建的Arc GPU系统； 2) 具有性能并行性和高吞吐量； 3) 擅长处理与媒体、3D应用程序和图形渲染有关的任务，能够同时运行处理大量AI任务；	1) Intel首款集成式NPU； 2) 专门为AI任务设计； 3) 节能、可持续运行和处理AI任务；	采用 Foveros 封装技术，在芯片内实现极低功耗和高密度的晶片连接
高通	骁龙X Elite 	1) 采用Oryon CPU内核，由台积电4nm工艺代工； 2) CPU性能表现远超英特尔13代酷睿i7处理器及苹果M2处理器； 3) 功耗较低；	1) 采用自研Adreno GPU； 2) 是英特尔酷睿i7-13800H中集成的GPU性能2倍，峰值功耗降低74%； 3) 比AMD Ryzen 9 7940HS中的Radeon 780M GPU快80%，峰值功耗降低80%；	1) 采用自研Hexagon NPU； 2) 最高可提供 45 TOP(每秒万亿次运算) 算力； 3) 可在设备上以“惊人的速度”运行超过13B参数的生成式AI LLM；	1) 内存带宽136GB/s； 2) 缓存总数42MB； 3) 节能效果突出，可实现长达数天的电池续航； 4) 通过5G和Wi-Fi7连接，可快速下载文件、流式传输和同步；

资料来源：英特尔ON技术创新峰会，2023高通峰会，西南证券整理

3.1.2 算力端：云侧龙头地位稳固、端侧竞争格局加剧、大厂自研趋势明显

- ❑ 3) 大厂自研趋势：科技巨头加速自研，优先服务于自身云服务业务及AI条线。微软于23年11月16日Ignite技术大会上发布两款自研芯片——Azure Maia 100和Azure Cobalt 100，分别用于大语言模型的训练推理和通用云服务的支持。近年来，各大科技厂商纷纷自研芯片，一是为了降低自身对第三方芯片和外部供应链的依赖；二是自研芯片可帮助各大厂商克服一定的通用芯片局限，通过CPU+GPU+DPU+定制芯片等结合方案，提升全系统整合效率、实现业务赋能；三是提高计算能效、减少长期硬件成本。

各大厂自研AI芯片和CPU产品情况

AI芯片布局						
厂商	Google	Amazon		Meta	Tesla	Microsoft
芯片	Cloud TPU v5e	AWS Inferentia 2	AWS Trainium	MTIA v1	Dojo D1	Azure Maia 100
推出时间	2023年8月29日	2019年	2022年	2023年5月18日	2021年	2023年11月16日
代际	第五代	第二代	第一代	第一代	第一代	第一代
工艺&制程	最多允许256个芯片互连，总带宽超过400 Tb/s，INT8性能达到100 petaOps	5nm	5 nanometer	7nm；内部内存可以从128MB扩展到128GB	7nm；500 亿个晶体管；400W TDP	5nm；1050亿个晶体管
用途	专用于大中型训练与推理；支持Google cloud和聊天机器人Bard等应用产品	推理芯片	训练芯片	推理芯片	训练芯片；用于搭建Tesla Dojo超算平台，以支持自动驾驶和机器人业务	专门用于云端训练和推理；支持Microsoft Azure OpenAI服务和Microsoft Copilot(Bing Chat)等应用产品

CPU芯片布局				
厂商	Google	Amazon	Microsoft	
芯片名称	Cypress	Maple	Graviton 3	Azure Cobalt 100
推出时间	预计2025年部署上线	预计2025年部署上线	2021年	2023年11月16日
代际	第一代	第一代	第三代	第一代
工艺&制程	5nm；基于Arm	5nm；基于Arm	5nm；550亿个晶体管	5nm；基于Arm；128核
用途	自研用于运营数据中心的 Plan A 处理器	自研用于运营数据中心的 Plan B 处理器	云原生通用处理器	旨在用于执行常规计算任务，如为微软Teams提供动力；暂时没有销售计划，更倾向于供内部使用

资料来源：The information，各公司官网，西南证券整理

3.1.3 算力端：算力瓶颈拉动存力需求，HBM成为提升性能的高效途径

□ **单位算力存储需求提升，HBM实现功耗降低&带宽提升。** HBM3e是一种基于3D堆叠工艺的DRAM存储芯片，AI服务器对其需求强烈。从英伟达H200芯片参数来看，H200首次搭载HBM3e，拥有141GB的内存及4.8TB/秒的带宽，HBM容量提升76%，而算力保持不变。此外，根据官网披露数据，相较于H100，H200可以使Llama-70B的推理性能几乎翻倍，运行GPT3-175B的效率可以提升60%。可以看出，在算力到达一定瓶颈的情况下，AI芯片将逐步通过堆叠HBM以提升性能，未来芯片将通过存力升级进行优化迭代，实现单位算力存储能力的大幅提升，因此，HBM等存力需求也将迎来暴增，从而刺激存储产品进一步涨价。

英伟达H100和H200算力和存力对比

存力大幅提升：+76% ←

产品名称	H200 SXM	H100 SXM
GPU Architecture	Hopper	Hopper
GPU Memory	141 GB HBM3e	80 GB HBM3
GPU Memory Bandwidth	4.8 TB/s	3.35 TB/s
INT8 FP8 Tensor Core	3958 TFLOPS	3958 TFLOPS
BF16 FP16 Tensor Core*	1979 TFLOPS	1979 TFLOPS
TF32 Tensor Core*	989 TFLOPS	989 TFLOPS
FP32	67 TFLOPS	67 TFLOPS
FP64	34 TFLOPS	34 TFLOPS
TDP	Up to 700W (configurable)	Up to 700W (configurable)

算力保持不变 ←

海力士HBM实现带宽提升&功耗下降



资料来源：英伟达官网，西南证券整理

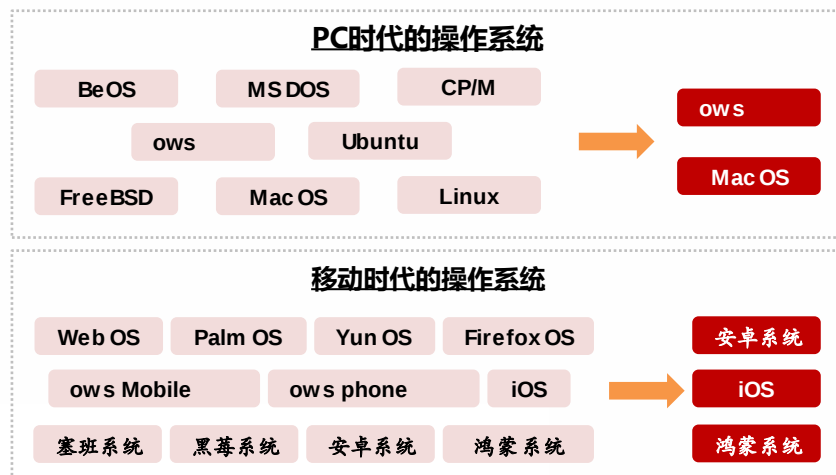
资料来源：海力士官网，西南证券整理

3.2.1 应用端：AI应用市场潜力无限，数量有望实现非线性增长

❑ 大模型将由少数大厂垄断，应用有望百花齐放。

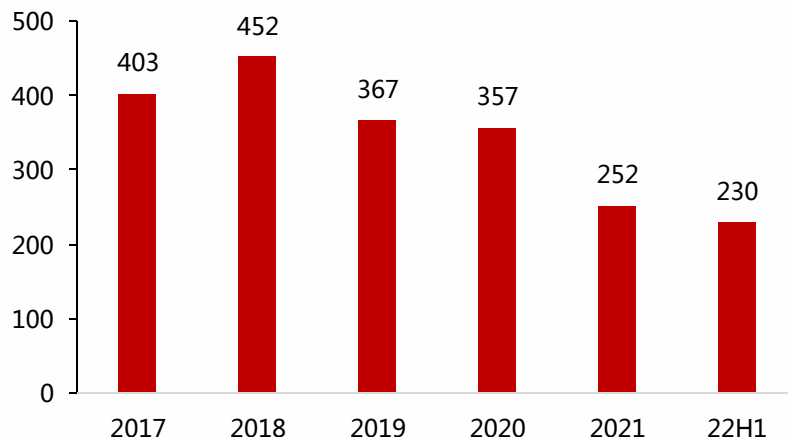
- **底层架构方面**：类比互联网时代，PC操作系统和移动操作系统在发展初期面临激烈竞争，但PC操作系统最终仅由ows和Mac OS系统主导、移动操作系统由iOS、安卓系统、鸿蒙系统主导，我们认为AI时代的底层架构，即大模型，也将由当前百模大战的局面向寡头垄断的方向发展。
- **应用市场方面**：类比移动互联网时代，无论是iOS系统还是安卓系统，其中的应用数量均达数百万款。根据工信部数据，截至22H1，我国国内市场上监测到的APP数量多达232万款，APP应用生态繁荣。我们认为AI时代的应用市场也将如同移动互联网时代具备无限潜力，且GPTs的创作门槛比APP的开发门槛更低，与百模大战相比，未来AI应用的数量将达到更高的量级，并在发展初期有望实现非线性的高速增长，而原生AI应用的繁荣，则是人类真正进入AI时代的标志。

底层架构大浪淘沙：操作系统最终由寡头垄断



资料来源：洞见数据研究院，西南证券整理

应用市场百花齐放：中国移动应用在架数量



资料来源：工信部，前瞻产业研究院，西南证券整理

3.2.2 应用端：GPTs垂类属性加强，有望覆盖更多长尾空间

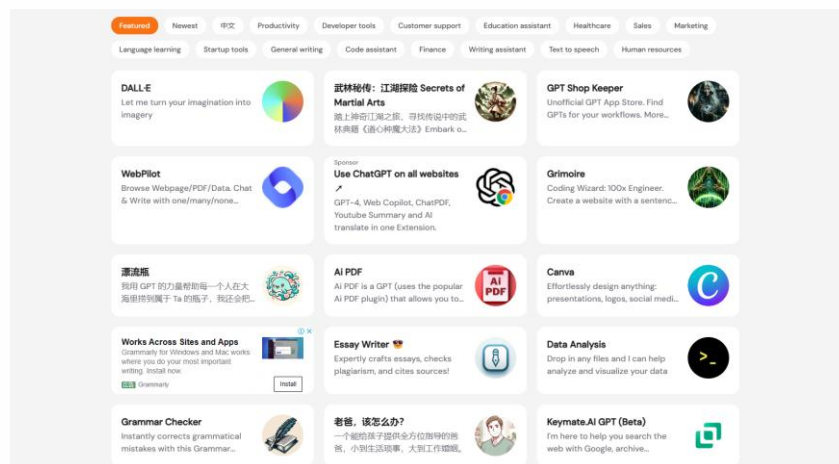
- ❑ **强化垂类属性**：根据GPTs Hunter，当前GPTs应用的赛道布局主要包括生产力工具、开发者工具、客服助手、教育助手、医疗健康助手、营销工具、语言学习助手、创业工具、代码&写作助手、文字转语音和人力资源工具等。**OpenAI的GPTs通过提供API，开发者只需喂给大模型更多的垂类数据，即可打造垂类AI应用**，对此前开源生态中的垂类AI应用以及传统软件都将产生一定威胁。
- ❑ **满足长尾需求**：与AI应用相比，传统软件孤岛特征明显，应用开发相对受限，难以满足长尾需求。而**OpenAI的GPTs使满足任何人的长尾需求成为可能**，开发者和用户可以通过上传自身的个性化数据，链接模型和特征数据，生成定制化助手。此外，个人助手还可以通过中间层的一次性代码调用其他二级Agent，完成其他任务，实现只需唯一助手即可满足复杂和长尾的个人需求。

GenAI产品月访问量Top50

Top 50 GenAI Web Products, By Monthly Visits				
1. ChatGPT	11. YOU	21. NightCafe	31. GPTGPT.ai	41. FLIKI
2. character.ai	12. leonardo.	22. lIeflicate	32. runway	42. pornpen.ai
3. Bard	13. PIXLR	23. Speechify	33. Playground	43. KAPWING
4. Poe	14. VEED.IO	24. IIElevenLabs	34. Kaiber	44. Gamma
5. QuillBot	15. Tome	25. Lexica	35. Hotpot	45. Looka
6. PhotoRoom	16. AI-Novel	26. VocalRemover	36. Stable Diffusion	46. humanornot
7. CIBITAI	17. cutout.pro	27. Writesonic	37. copy.ai	47. PIXAI
8. Midjourney	18. ForefrontAI	28. CHATPDF	38. ZeroGPT	48. WRITER
9. Hugging Face	19. Clipchamp	29. D-ID	39. Smodin	49. NovelAI
10. Perplexity	20. TheB.AI	30. Chubai	40. ZMO.AI	50. DeepSwap

资料来源：Consumer，西南证券整理

GPTs Hunter特色GPTs榜单

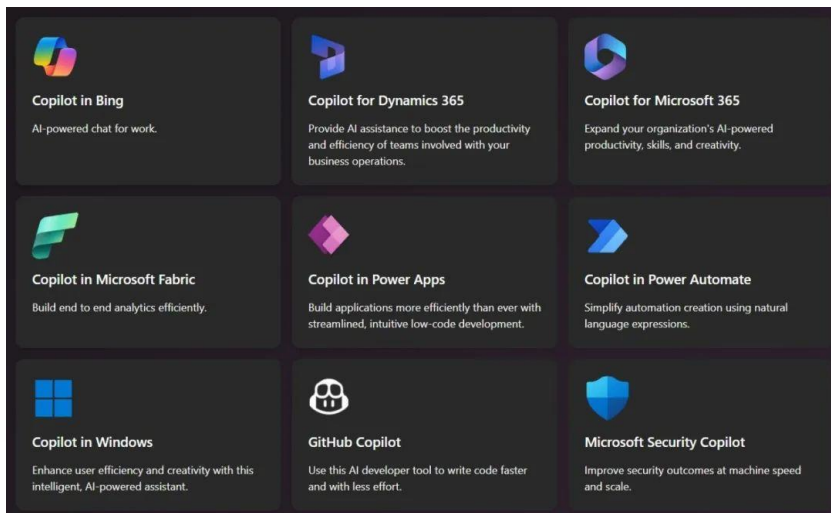


资料来源：GPTs Hunter官网，西南证券整理

3.2.3 应用端：微软Copilot赋能旗下产品，先发优势&头部地位突出

- ❑ 2023年11月16日，微软在Ignite技术大会上宣布开启Copilot时代：
- 1) **Bing Chat更名为Copilot（可类比为OpenAI的ChatGPT）**：Bing Chat和Bing Chat for Enterprise版本均更名为Copilot，后续不仅可以通过浏览器网页进行访问，还会上线移动设备。可以免费使用GPT-4、DALL·E 3功能，并且拥有独立网站，Copilot将嵌入微软旗下所有应用中。
- 2) **推出低代码工具Microsoft Copilot Studio（可类比为OpenAI的GPTs）**：Copilot Studio允许用户通过低代码开发方式在网页中构建、部署、分析和管理内容，通过拖放等简单易懂的方式，直接构建和发布插件。Copilot Studio可集成OpenAI的GPTs和第三方应用数据源，用户可通过提供关键业务数据，定制开发Copilot功能，生成自定义应用。个性化程度有望日益丰富。

微软将Copilot嵌入旗下所有应用中



微软推出Copilot Studio开发助手



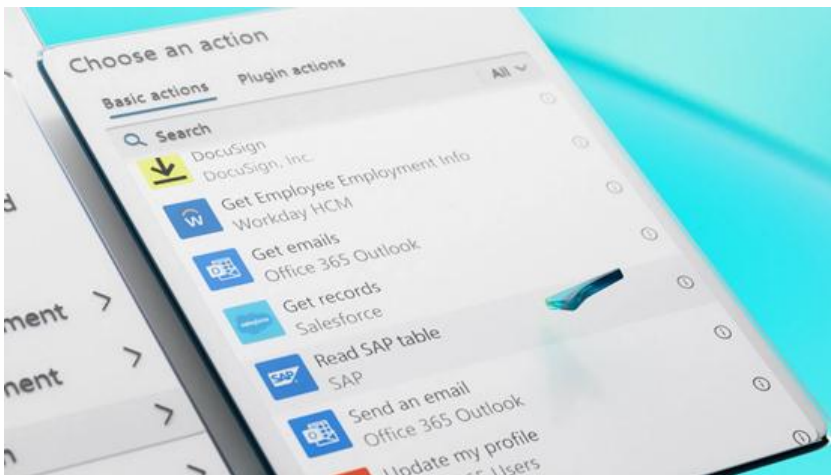
资料来源：微软 Ignite 2023，西南证券整理

资料来源：微软 Ignite 2023，西南证券整理

3.2.3 应用端：微软Copilot赋能旗下产品，先发优势&头部地位突出

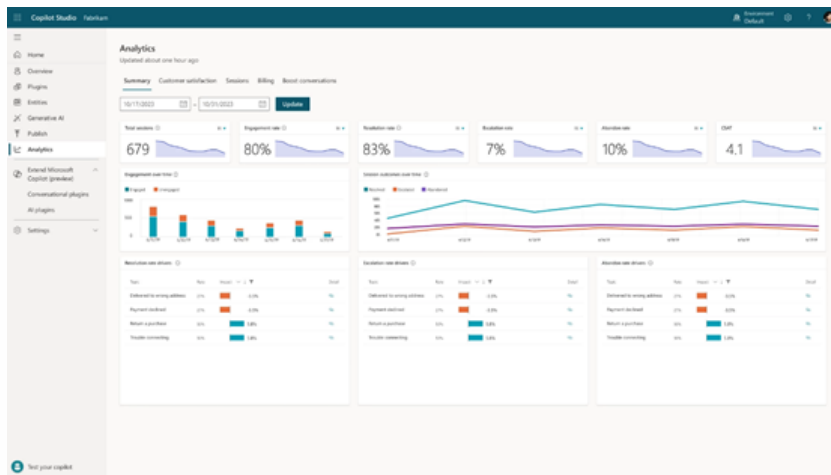
- ❑ 微软Copilot有望率先成为AI颠覆性应用。GPTs浪潮将推动未来涌现众多的垂类应用开发者和创业者。而在垂类领域之外，微软Copilot有望率先成为大的颠覆者之一。
- ❑ 相较于OpenAI的GPTs，微软Copilot Studio功能更强、应用更广：1) 与企业办公软件深度集成：用户可以通过自己的数据、文件（包含Word、PPT、PDF等15种类型）、SharePoint站点来开发助手，开发后的助手可以通过SAP、Workday和ServiceNow等1100多个预制连接器无缝集成在CRM、ERP、OA等日常办公系统中，并与微软的Power Bi、Power Virtual Agents、Microsoft Teams等产品实现深度集成，提供可视化数据分析等商业服务，帮助企业生产提效。2) 提供带有安全管理和控制的可视化后台界面：管理员可实时监控自定义助手的使用情况并进行数据分析，同时通过后台控制来选定谁可以使用自定义助手，从而进一步加强数据安全管理。

Copilot Studio可集成于微软产品系统中



资料来源：微软 Ignite 2023，西南证券整理

Copilot Studio提供后台实时可视化界面



资料来源：微软 Ignite 2023，西南证券整理

3.2.4 应用端：热门AI应用流量周度跟踪

- **周度日均访问量TOP20总览**：11.05-11.11全球AI应用日均访问量前20排名与上周相同，环比各有升降，前三分别为 ChatGPT、New Bing 以及 Canva Text to Image，环比分别 +3.79%-0.86%/+7.64%，其他AI应用日均访问量均处于1000万次以下，其中AI+生产力应用Gamma获得本周前20榜单中最高增速，环比+9.58%。

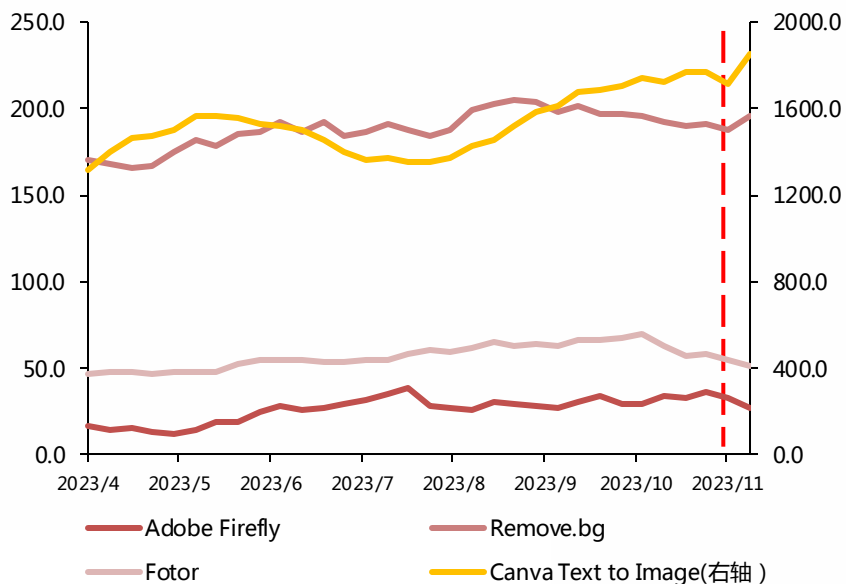
应用领域	名称	11.05-11.11 日均访问量（万次）	环比	名次较上周变化	11.05-11.11 日均访问时长（秒）	环比
AI+聊天机器人	ChatGPT	5654.4	3.79%	0	508	9.96%
AI+搜索引擎	New Bing	4246.4	-0.86%	0	423	-0.24%
AI+图像	Canva Text to Image	1849.7	7.64%	0	624	0.16%
AI+聊天机器人	Bard	884.5	7.11%	0	355	-9.44%
AI+写作	Deepl	837.6	2.42%	0	546	-1.80%
AI+教育	Q-chat	506.3	-0.02%	0	420	-1.64%
AI+办公	Notion AI	499.7	-0.82%	0	465	-0.43%
AI+办公	Slack	396.5	4.57%	0	778	1.97%
AI+生产力	Jambot	275.9	2.95%	0	950	3.83%
AI+教育	Duolingo	254.5	0.51%	0	610	-0.65%
AI+写作	Grammarly	216.8	-1.38%	0	200	-4.76%
AI+写作	Quillbot	214.9	2.45%	0	363	4.01%
AI+聊天机器人	Poe	207.6	1.20%	0	426	-7.59%
AI+图像	Remove.bg	195.7	3.87%	0	343	-0.29%
AI+搜索引擎	Perplexity.ai	174.2	-0.26%	0	603	8.45%
AI+教育	Khancademy	156.0	-1.14%	0	495	3.34%
AI+搜索引擎	Feedly	106.2	-3.05%	0	508	2.83%
AI+聊天机器人	Claude	88.7	-1.08%	0	297	-3.26%
AI+生产力	Gamma	87.6	9.58%	0	199	5.85%
AI+编程	Replit	73.2	0.14%	0	420	-5.19%

数据来源：Similarweb，西南证券整理

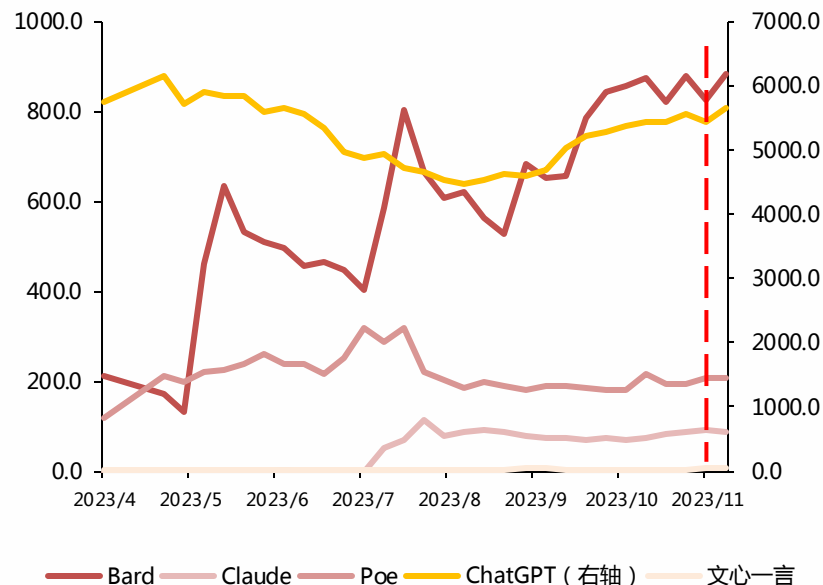
3.2.4 应用端：热门AI应用流量周度跟踪

- ❑ **细分赛道1—AI+图像**：Canva text to image在AI+图像应用中日均访问量最高，11.05-11.11日均访问量达1849.7万次，环比+7.64%；平均访问时长624秒，环比+0.16%。该应用是一款在线图片编辑工具，可以帮助用户将文字转化为图片、轻松地创建高质量的图像。用户可以在此应用上自由选择字体、颜色、背景等元素，并添加各种图形、效果和滤镜，以丰富其图片设计。
- ❑ **细分赛道2—AI+聊天机器人**：ChatGPT是目前AI+聊天机器人应用中日均访问量最高的应用，同时也是所有AI应用中流量最高的，11.05-11.11日均访问量达到5654.4万次，环比+3.79%；平均访问时长508秒，环比-9.96%。

AI+图像周度日均访问量（万次）



AI+聊天机器人周度日均访问量（万次）

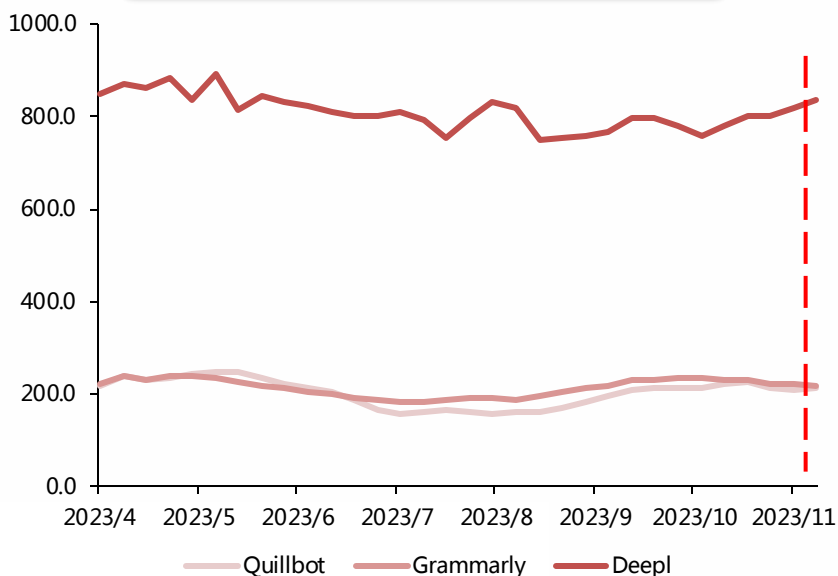


数据来源：Similarweb，西南证券整理

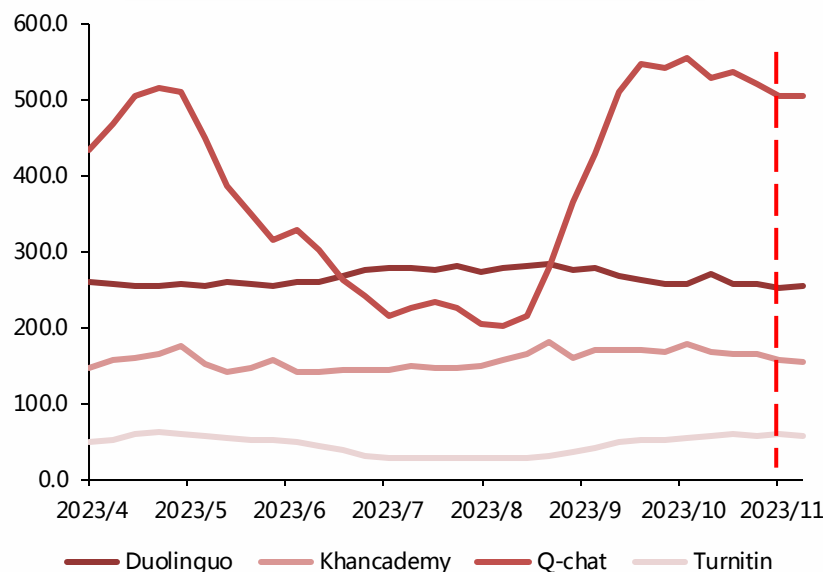
3.2.4 应用端：热门AI应用流量周度跟踪

- ❑ **细分赛道3—AI+写作**：DeepL是目前AI+写作应用中日均访问量最高的应用，11.05-11.11日均访问量达到837.6万次，环比+2.42%；平均访问时长546秒，环比-1.80%。DeepL是一家德国的AI翻译公司，成立于2009年，前身为Linguee，致力于将人工智能技术赋能写作翻译领域。
- ❑ **细分赛道4—AI+教育**：Q-chat是目前AI+教育应用中日均访问量最高的应用，11.05-11.11日均访问量达到506.3万次，环比-0.02%；平均访问时长420秒，环比-1.64%。Q-chat是Quizlet推出的AI应用，定位于在线学习平台，可以帮助学生轻松学习各种内容，从学术到职业，覆盖各个领域。Quizlet可以创建和记忆各种内容，包括词汇、概念、图片、视频、音频等，让学习变得更加高效和有趣。Quizlet还提供各种测验和游戏，帮助用户巩固所学知识，提高学习效果。

AI+写作周度日均访问量（万次）



AI+教育周度日均访问量（万次）

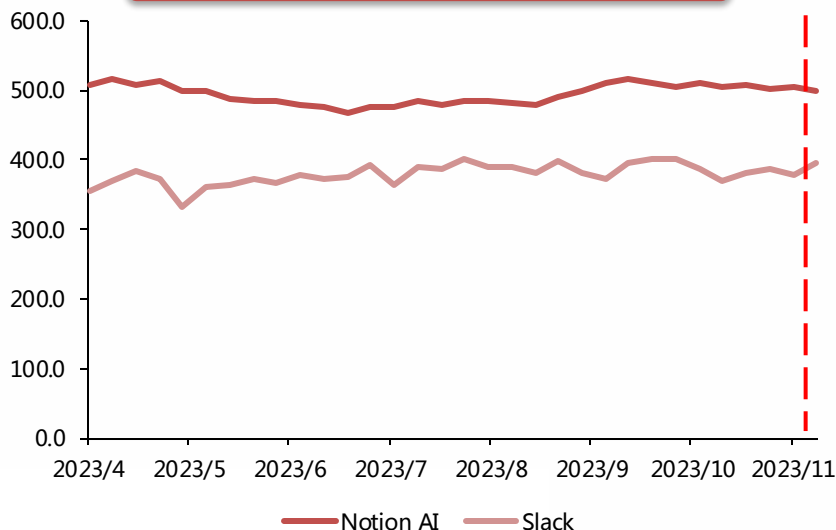


数据来源：Similarweb，西南证券整理

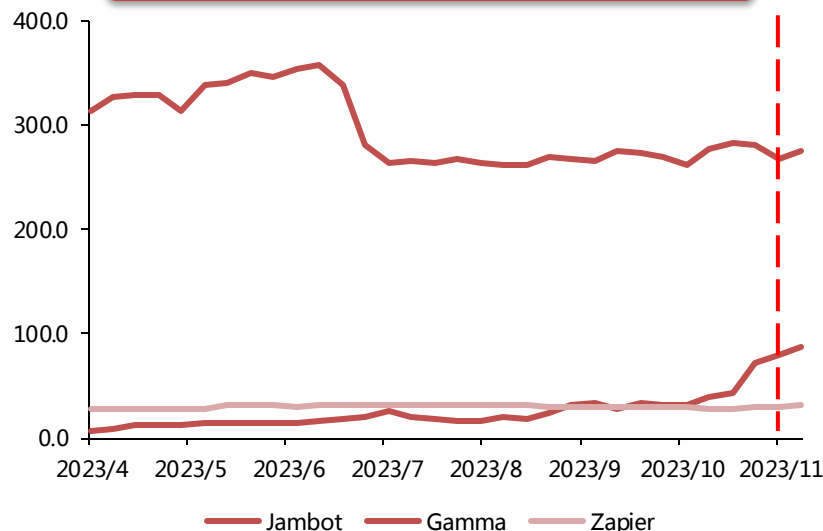
3.2.4 应用端：热门AI应用流量周度跟踪

- ❑ **细分赛道5—AI+办公**：Notion AI是目前AI+办公应用中日均访问量最高的应用，11.05-11.11日均访问量达到499.7万次，环比-0.82%；平均访问时长465秒，环比-0.43%。Notion是一款功能强大的工作协作应用程序，用户可以使用它创建和组织笔记、任务清单、数据库、文档等多种类型的内容。Notion AI是Notion应用程序中内置的人工智能功能，通过使用机器学习技术，以帮助用户更高效地组织和管理他们的内容。
- ❑ **细分赛道6—AI+生产力**：Jambot是目前AI+生产力应用中日均访问量最高的应用，11.05-11.11日均访问量达到275.9万次，环比+2.95%；平均访问时长950秒，环比+3.83%。Jambot是一个在线设计协作平台Figma为其白板工具FigJam最新推出的AI插件，借助ChatGPT的AI能力可帮助设计人员或团队激发灵感快速进行头脑风暴和生成思维导图。

AI+办公周度日均访问量（万次）



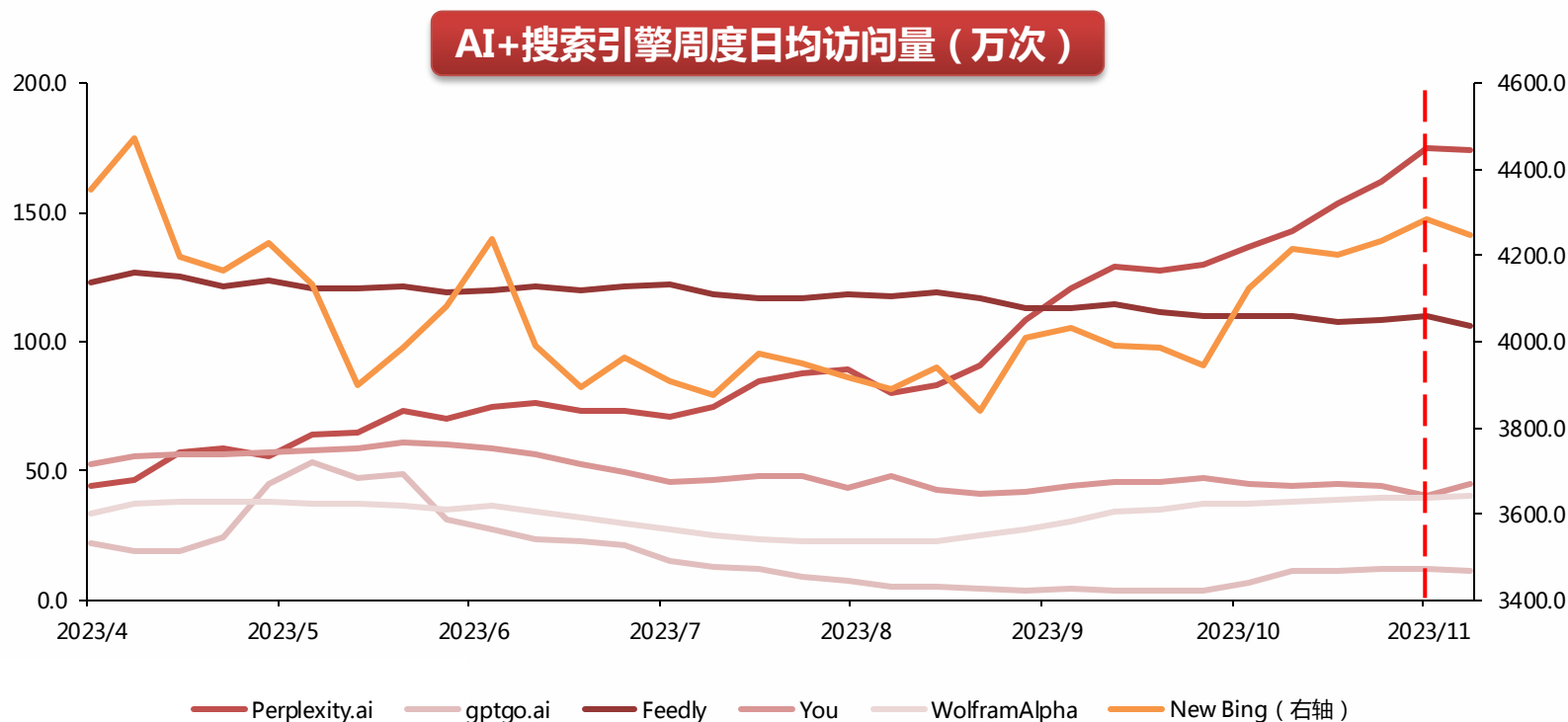
AI+生产力周度日均访问量（万次）



数据来源：Similarweb，西南证券整理

3.2.4 应用端：热门AI应用流量周度跟踪

- **细分赛道7—AI+搜索引擎**：New Bing是目前AI+搜索引擎应用中日均访问量最高的应用，也是AI应用全品类中仅次于ChatGPT的应用，11.05-11.11日均访问量达到4246.4万次，环比-0.86%；平均访问时长423秒，环比-0.24%。New Bing是微软在2023年推出的一款搜索引擎，它基于GPT-4技术，可以提供更加智能和强大的搜索和语言交互功能。与传统的搜索引擎相比，New Bing不仅可以提供网页链接和简单的文本信息，还可以生成更加丰富和多样化的回答，并且能够根据用户的意图和上下文进行更加智能的回答。



数据来源：Similarweb，西南证券整理

3.3 数据端：专业数据或成关键，重视垂类数据留存及知识产权价值

- ❑ **专业数据是构建GPTs竞争力的核心要素之一**：OpenAI发布的GPTs包含Knowledge的功能，该功支持用户上传个人或专业数据，有助于提升AI应用的定制化使用效果和用户体验。我们认为，在GPTs模式下，专业知识与大模型通用能力的结合将成为未来的重中之重，私域数据库或专业数据库方向将会不断产生新需求、新热点。
- ❑ **垂类数据留存成为构建数据库的关键**：当前垂类数据分布相对分散，难以形成系统的数据集，GPTs模式有望收集更多垂类和专业数据，而如何在合规前提下留存以上数据并构建体系化数据库或成为关键。此外，面对业务连续性的要求，OpenAI和微软等AI应用和技术提供商需要在硬件端构建完善的数据存储系统，即使出现设备宕机，也应保障计算任务实现连续运行，避免因服务器故障或者网络故障导致数据丢失等问题。
- ❑ **GPTs背后的知识产权价值有望提升**：在GPTs模式下，开发者和用户在制作专属GPTs时，可能需要调用个人、专业或垂类数据库的知识和信息，未来随着GPTs大规模的爆发和应用，知识产权的确权和保护议题将变得愈发重要，知识产权的价值有望大幅提升。

风险提示

- ❑ 相关理论和技术发展不及预期风险；
- ❑ AI行业竞争加剧风险；
- ❑ 应用生态发展不及预期风险。

分析师：王湘杰
执业证号：S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司 评级

买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下

行业 评级

强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会核准的证券投资咨询业务资格。

本公司与作者在自身所知范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本公司签约客户使用，若您并非本公司签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司也不会因接收人收到、阅读或关注自媒体推送本报告中的内容而视其为客户。本公司或关联机构可能会持有报告中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或财务顾问服务。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告，本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。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。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，本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。

本报告及附录版权为西南证券所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用须注明出处为“西南证券”，且不得对本报告及附录进行有悖原意的删节和修改。

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	陈阳阳	销售经理	17863111858	cyyf@swsc.com.cn
	崔露文	销售经理	15642960315	clw@swsc.com.cn	李煜	销售经理	18801732511	yfliyu@swsc.com.cn
	谭世泽	销售经理	13122900886	tsz@swsc.com.cn	卞黎旻	销售经理	13262983309	bly@swsc.com.cn
	薛世宇	销售经理	18502146429	xsy@swsc.com.cn	龙思宇	销售经理	18062608256	lsyu@swsc.com.cn
	岑宇婷	销售经理	18616243268	cyryf@swsc.com.cn	田婧雯	销售经理	18817337408	tjw@swsc.com.cn
	汪艺	销售经理	13127920536	wyf@swsc.com.cn	阚钰	销售经理	17275202601	kyu@swsc.com.cn
	张玉梅	销售经理	18957157330	zmyf@swsc.com.cn	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	胡青璇	销售经理	18800123955	hqx@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	张鑫	销售经理	15981953220	zhxin@swsc.com.cn
	杨薇	高级销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	销售经理	18500981866	wangyuf@swsc.com.cn
	王一菲	销售经理	18040060359	wyf@swsc.com.cn	路漫天	销售经理	18610741553	lmtyf@swsc.com.cn
	姚航	销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
广深	郑龔	广深销售负责人	18825189744	zhengyan@swsc.com.cn	张文锋	销售经理	13642639789	zwf@swsc.com.cn
	杨新意	销售经理	17628609919	yxy@swsc.com.cn	陈紫琳	销售经理	13266723634	chzlyf@swsc.com.cn
	龚之涵	销售经理	15808001926	gongzh@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	丁凡	销售经理						