

计算机

报告日期：2023年03月15日

# GPT-4 发布，多模态处理+复杂推理能力有望打开应用空间

## ——行业点评报告

### 投资要点

- 北京时间3月15日凌晨，OpenAI官方发布了新一代AI大模型GPT-4，模型在多模态任务、复杂问题推理领域取得重大突破，在多项人类测试中表现优异。OpenAI已在ChatGPT Plus上提供API接口供开发者构建应用和服务，标志着AI多模态化进程持续推进，商业化空间有望加速打开。
- GPT-4实现多模态任务处理能力，并在数学、物理等学科领域实现重大进步
  - 1、GPT-4可接受图像、文本信息作为输入并生成说明文字、分类和分析，允许长文内容创建、扩展对话以及文档搜索和分析等功能；
  - 2、GPT-4在复杂问题的推理能力上大幅超越GPT-3.5，在SAT、Bar等大多数专业测试以及相关学术基准评测中，GPT-4的分数高于GPT-3.5并超越人类平均水平；
  - 3、GPT-4目前已对ChatGPT Plus用户开放API，收费为每1000 prompt tokens 0.12美元，而每1000 completion tokens收费为0.06美元。同时OpenAI还开源了Evals框架，以自动评估AI模型性能，允许用户报告模型中的缺点，以帮助其改进。
  - 4、GPT-4引入了更多人类反馈数据进行训练，以改进GPT-4的行为，不断吸取现实世界使用的经验教训，并基于模型的高级推理和指令遵循能力加快模型安全性的迭代。
- GPT-4已投入实际应用，多模态趋势下商用空间有望打开
  - 1、微软已宣布新必应（Bing）搜索引擎已接入GPT-4，未来有望将GPT-4技术嵌入到业务生态中的更多应用；
  - 2、Duolingo、Strip、冰岛政府等企业、教育机构、政府已接入GPT-4，借助其能力实现不同的需求和功能；
  - 3、多模态化将成为AI大模型发展的核心趋势，看好未来GPT模型融合音频、视频等模态数据，实现更多跨模态任务，进一步打开应用空间；
- 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 《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

1 OpenAI 发布 GPT4 大模型，多模态为核心亮点 .....	4
1.1 GPT 4 大模型发布，多模态任务+复杂推理实现重大突破 .....	4
1.2 GPT- 4 在多项任务环节实现重要突破 .....	5
1.3 新 Bing 已使用 GPT-4 模型，多模态趋势下商用空间有望打开 .....	6
2 风险提示 .....	7

## 图表目录

图 1: GPT-4 可基于图片输入生成对应文字描述 .....	4
图 2: GPT-4 解决复杂推理问题 .....	4
图 3: GPT-4 在多项人类测试上的表现实现大幅进步 .....	5
图 4: GPT-4 在传统机器学习模型测试上表现优异 .....	5
图 5: GPT-4 在生成内容的合理性和真实性方面进步显著 .....	5
图 6: GPT-4 在生成内容的合理性和真实性方面进步显著 .....	6
图 7: AI Jina 光谱——未来多模态模型将成为核心发展方向 .....	7
表 1: 多家企业、政府及教育机构借助 GPT-4 实现功能 .....	7

# 1 OpenAI 发布 GPT4 大模型，多模态为核心亮点

## 1.1 GPT 4 大模型发布，多模态任务+复杂推理实现重大突破

北京时间 2023 年 3 月 15 日凌晨，OpenAI 官方发布了 GPT-4 大模型，与 GPT-3、GPT-3.5 模型相比，GPT-4 具备了多模态功能，支持接受图像和文本输入并输出文本，我们认为随着模型的持续迭代，GPT 有望在图像生成、音视频处理等跨模态任务领域上持续拓展，应用空间非常广阔。

**GPT-4 距离人类认知水平仍有差距，但在多项指标表现上实现重大迭代。**OpenAI 官方表示，GPT-4 在许多现实世界场景中的能力仍不如人类，但在各种专业和学术基准上的表现已经超越了人类的平均水平。在模拟 bar (美国律师执照) 考试中，GPT-4 测试成绩排名前 10%，而此前发布的 GPT-3.5 仅排在后 10%，除此之外，在多项不同领域的专业测试中，GPT-4 能够排到非常考前的位置，在复杂问题推理上的能力表现出了显著进步。

**GPT-4 在图像-文字、复杂推理任务领域展现优秀能力。**GPT-4 可根据图像信息识别图像中包含的信息，并生成文字反馈。同时在论文理解方面，GPT-4 可以智能识别论文截图中包含的文字以及图表内容并进行汇总分析，实现论文的总结归纳功能。在复杂问题推理上，GPT-4 可准确解答力学题，并根据题目要求输出推理计算过程，相比于 GPT-3 和 GPT-3.5 进步显著。

图1： GPT-4 可基于图片输入生成对应文字描述

User Below is part of the InstructGPT paper. Could you read and summarize it to me?



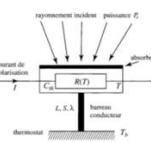
GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

图2： GPT-4 解决复杂推理问题

Un barreau, conducteur thermique, homogène, de longueur  $L$ , de section  $S$  et de conductivité thermique  $\lambda$  et sans échange thermique latéral, relie le bâtonnage à un thermostat de température  $T_0$  fixe.



I.1. Dans cette première partie, on étudie le comportement des barreaux en régime stationnaire, à la température de fourrétement  $T = T_0$ .

I.1.a) Exprimer la température  $T(x)$  en tout point du barreau conducteur rejoint par son absorbeur  $x$ ; on prendra  $T(0) = T_0$  et  $T(L) = T_b$ .

GPT-4

To answer question I.1.a, we need to express the temperature  $T(x)$  in every point of the conductive bar, which is characterized by the abscissa  $x$ . We are given that  $T(0) = T_0$  and  $T(L) = T_b$ .

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:  

$$\frac{d^2T}{dx^2} = 0$$

Integrating once, we get:  

$$\frac{dT}{dx} = C_1$$

Integrating a second time:  

$$T(x) = C_1x + C_2$$

To determine the constants  $C_1$  and  $C_2$ , we use the boundary conditions:  
 $T(0) = T_0$ , which gives  $C_2 = T_0$   
 $T(L) = T_b$ , which gives  $C_1 = (T_b - T_0)/L$

Finally, the expression for  $T(x)$  is:  

$$T(x) = (T_b - T_0) * (x/L) + T_0$$
  
 This represents a linear temperature profile along the conductive bar, with a slope of  $(T_b - T_0)/L$ .

资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

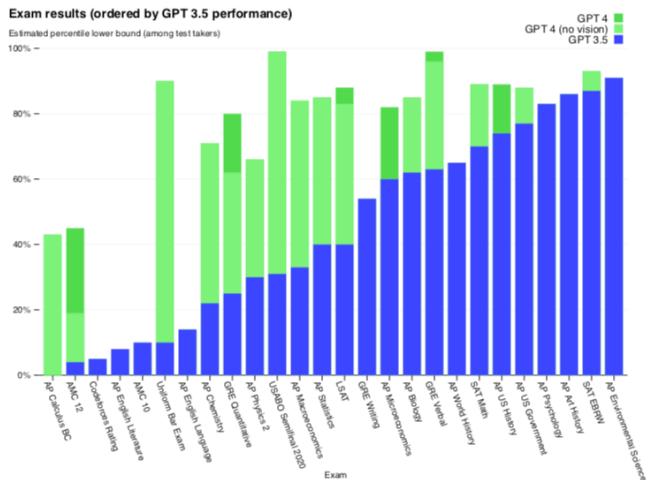
**OpenAI 开放 GPT-4 API 申请渠道并开源 OpenAI Evals 项目，GPT-4 用户生态有望快速构建。**OpenAI 官方宣布面向开发者开放 GPT-4 API 的申请通道，用户可提交申请进入 waitlist 等待通过，而 ChatGPT Plus 订阅会员可直接获得 GPT-4 的试用权限。与 GPT-3.5 类似，GPT-4 模型 API 根据用户调用 tokens 数量进行收费，定价为每 1000 prompt tokens 0.06 美元，而每 1000 completion tokens 收费 0.12 美元。另外，OpenAI 开源 Evals 项目，可用于自动评估 AI 模型性能，以便进一步优化与改进模型。

## 1.2 GPT-4 在多项任务环节实现重要突破，思维能力接近人类平均水平

OpenAI 通过各类 benchmark 进行测试，GPT-4 在多项测试上表现优于 GPT-3.5。相比于 GPT-3.5，GPT-4 在物理、生物、数学等理科测试中的表现明显优于 GPT-3.5，模型展现了处理复杂推理任务能力的大幅迭代。同时，GPT-4 在多数测试中的成绩能够超越 80% 的人类应试者，反映出模型在多项人类测试上的能力已超越平均水平。

在传统的机器学习测试上，GPT-4 在 MMLU（大规模多任务语言理解）、HellaSwag（常识基准）、GSM-8K（数学测试）等指标上相较 GPT-3.5 以及之前的模型最佳结果，均有不同程度的提升。

图3： GPT-4 在多项人类测试上的表现实现大幅进步



资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

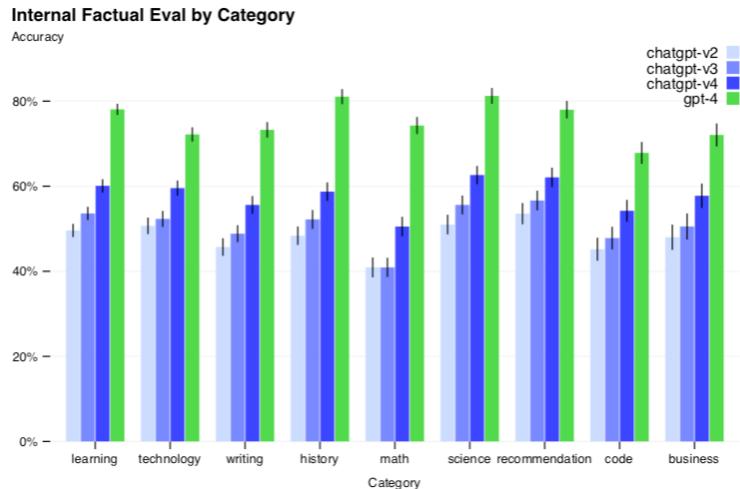
图4： GPT-4 在传统机器学习模型测试上表现优异

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [43] Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [44]	75.2% 5-shot Flan-PaLM [45]
HellaSwag [46] Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [47]
AI2 Reasoning Challenge (ARC) [48] Grade-school multiple choice science questions. Challenge set.	<b>96.3%</b> 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [49]	86.5% ST-MOE [18]
WinoGrande [50] Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [37] Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [51]
DROP [52] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	<b>88.4</b> QDGAT [53]
GSM-8K [54] Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [55]	87.3% Chinchilla + SFT+ORM-RL, ORM ranking [56]

资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

针对 GPT-3 以及 ChatGPT 会生成与事实不符信息的问题，GPT-4 实现显著优化。与 GPT-3.5 相比，GPT-4 在历史、数学、写作等各交互场景下上生成内容拥有更好的合理性与真实性。OpenAI 研究人员在论文中指出，在包含 9 大类话题的内容事实性评估测试中，GPT-4 的准确度较 GPT-3.5 平均提高了 19 个百分点。

图5： GPT-4 在生成内容的合理性和真实性方面进步显著



资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

GPT-4 基于人类反馈强化实现改进，但在复杂推理问题上已接近人类水平导致优化效果并不显著。OpenAI 透露，GPT-4 引入了更多人类反馈数据进行训练，以提升模型的安全性和对齐性。同时，研究人员在各项人类测试中将 GPT-4 基础模型（Base Model）和引入 RLHF 机制的模型（RLHF model）进行对比，结果显示引入 RLHF 后的 GPT-4 平均得分率为 74.0%，仅比未引入的 73.7% 高出 0.3%，也反映出 GPT-4 在专业性测试中的表现已经接近甚至超越人类。

图6：GPT-4 在生成内容的合理性和真实性方面进步显著

Exam	Base model	RLHF model
LSAT (MCQ)	67.0 %	72.0 %
SAT EBRW – Reading Portion	92.3 %	90.4 %
SAT EBRW – Writing Portion	90.9 %	84.1 %
SAT Math (MCQ)	91.4 %	86.2 %
Graduate Record Examination (GRE) Quantitative	57.5 %	67.5 %
Graduate Record Examination (GRE) Verbal	87.5 %	90.0 %
USNCO Local Section Exam 2022	51.7 %	63.3 %
AP Art History (MCQ)	72.5 %	66.2 %
AP Biology (MCQ)	98.3 %	96.7 %
AP Calculus BC (MCQ)	66.7 %	57.8 %
AP Chemistry (MCQ)	58.3 %	71.7 %
AP English Language and Composition (MCQ)	55.6 %	51.1 %
AP English Literature and Composition (MCQ)	63.6 %	69.1 %
AP Environmental Science (MCQ)	72.5 %	67.5 %
AP Macroeconomics (MCQ)	83.3 %	76.7 %
AP Microeconomics (MCQ)	90.0 %	76.7 %
AP Physics 2 (MCQ)	62.2 %	71.1 %
AP Psychology (MCQ)	98.0 %	96.0 %
AP Statistics (MCQ)	60.0 %	62.5 %
AP US Government (MCQ)	85.5 %	83.6 %
AP US History (MCQ)	89.1 %	87.3 %
AP World History (MCQ)	94.5 %	98.2 %
MKSAP Questions (MCQ)	77.9 %	74.7 %
AMC 10	28.0 %	24.0 %
AMC 12	20.0 %	32.0 %
Introductory Sommelier (theory knowledge)	90.5 %	92.2 %
Certified Sommelier (theory knowledge)	83.2 %	86.2 %
Advanced Sommelier (theory knowledge)	74.8 %	77.1 %
Average	73.7 %	74.0 %

资料来源：OpenAI《GPT-4 Technical Report》，浙商证券研究所

### 1.3 新 Bing 已使用 GPT-4 模型，多模态趋势下商用空间有望打开

GPT-4 性能大幅提升，海外多家企业已相继布局 GPT-4 应用。GPT-4 在回答准确性、识图能力、歌词以及文本生成等任务能力的大幅提升，使其具备更高的商业价值和更多的应用场景。3月15日，微软副总裁兼消费者首席营销官 Yusuf Mehdi 发文确认了新 Bing（必应）搜索引擎已运行 GPT-4，而包含摩根士丹利、duolingo 在内的政府及企业已布局 GPT-4 技术以实现多样化的功能。

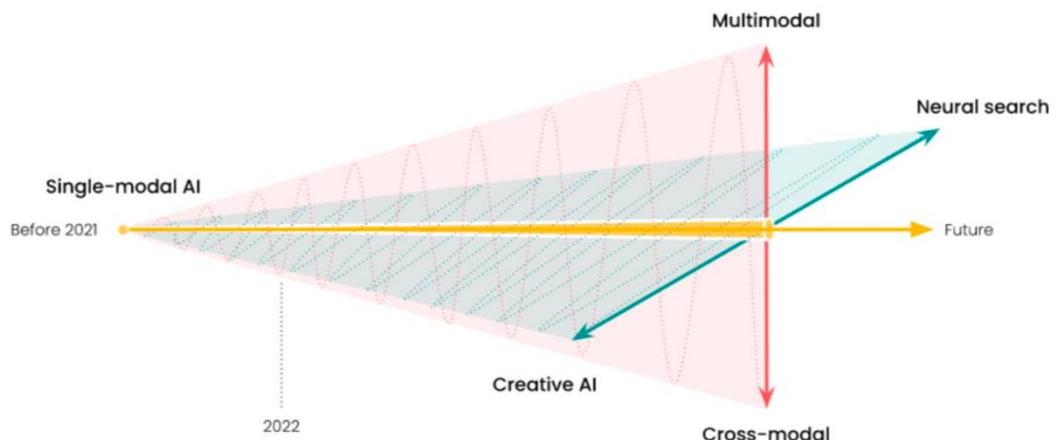
表1：多家企业、政府及教育机构借助 GPT-4 实现功能

企业/政府/机构	应用场景	基于 GPT-4 实现功能
Duolingo	语言教育	实现更加智能的语言沟通交互
Be My Eyes	视觉残障辅助	帮助盲人识别图像、文本信息
Stripe	互联网支付	识别和预防欺诈
摩根士丹利	金融	财富管理部门利用 GPT-4 组织庞大的知识库
可汗学院	教育	为学生成个性化的学习计划
冰岛政府	文化	利用 GPT-4 保护本国语言文化

资料来源：OpenAI 官网，浙商证券研究所

我们认为，GPT-4 实现了从单模态大模型到多模态大模型的重要突破，未来有望随着模型算法的加速迭代，实现对音频、视频等多模态数据的融合，完成各类跨模态任务，充分打开 AI 商业化应用空间。

图7：AI Jina 光谱——未来多模态模型将成为核心发展方向



资料来源：CSDN，浙商证券研究所

## 2 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>