

# 多模态AI研究框架

## ——人工智能行业专题报告

行业评级：看好

2023年4月5日

分析师

刘雯蜀

邮箱

liuwenshu03@stocke.com.cn

证书编号

s1230523020002

### 1、AI模型从单模态向多模态演进，未来有望实现认知智能：

AI模型走向多模态必然性的三大因素：跨模态任务需求 + 跨模态数据融合 + 对人类认知能力的模拟。

### 2、多模态 AI 融合多种数据，可大幅延伸应用场景：

多模态 AI 能够实现基于文本、语音、图片、视频等多模态数据的综合处理应用，完成跨模态领域任务。

### 3、多模态 AI 五大技术环节，模态融合为核心：

多模态 AI 以模态融合为核心技术环节，围绕“表征-翻译-对齐-融合-联合学习”五大技术环节，解决实际场景下复杂问题的多模态解任务。

### 4、国内外多模态 AI 布局进展，关注 OpenAI 及谷歌动向：

国内外大厂及科研院所自2021年起相继推出跨模态 AI 模型，目前OpenAI 及谷歌（DeepMind）布局较完善，未来有望基于各任务模型，构建多模态AI生态。

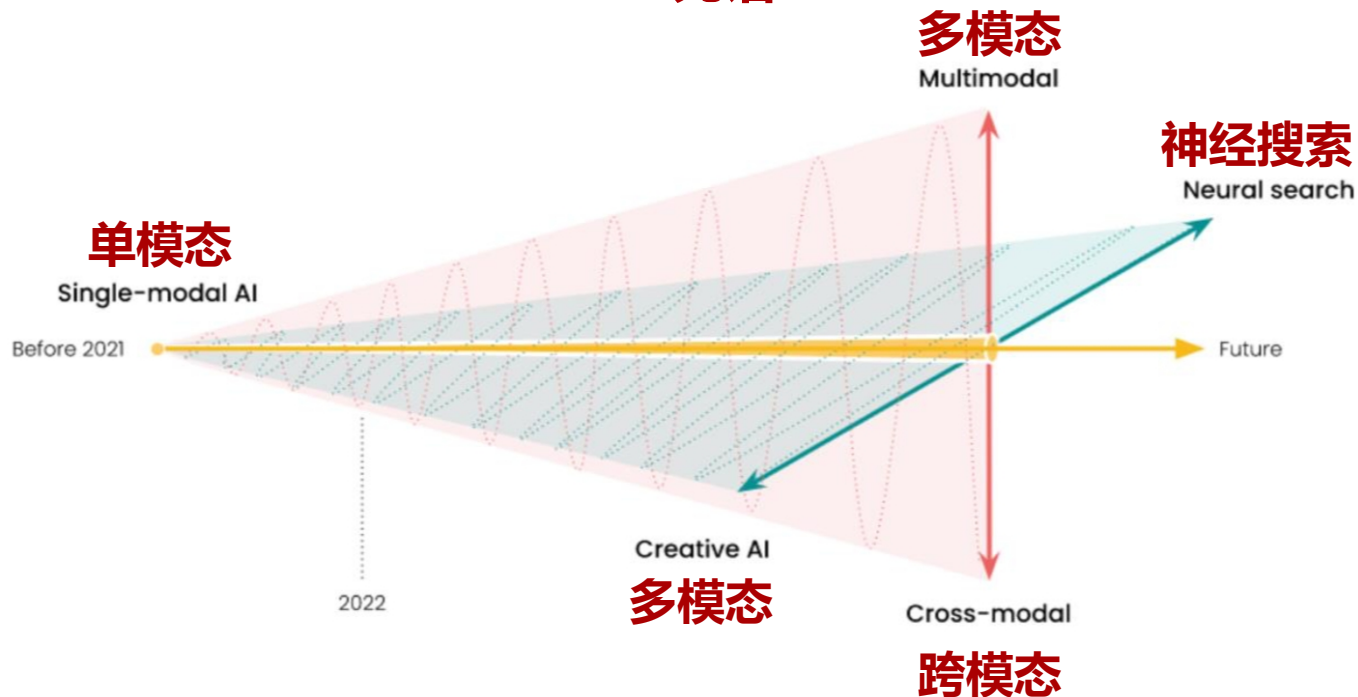
### 5、重点关注标的：

(1) **基础层**：大模型：三六零，科大讯飞； 数据服务：海天瑞声

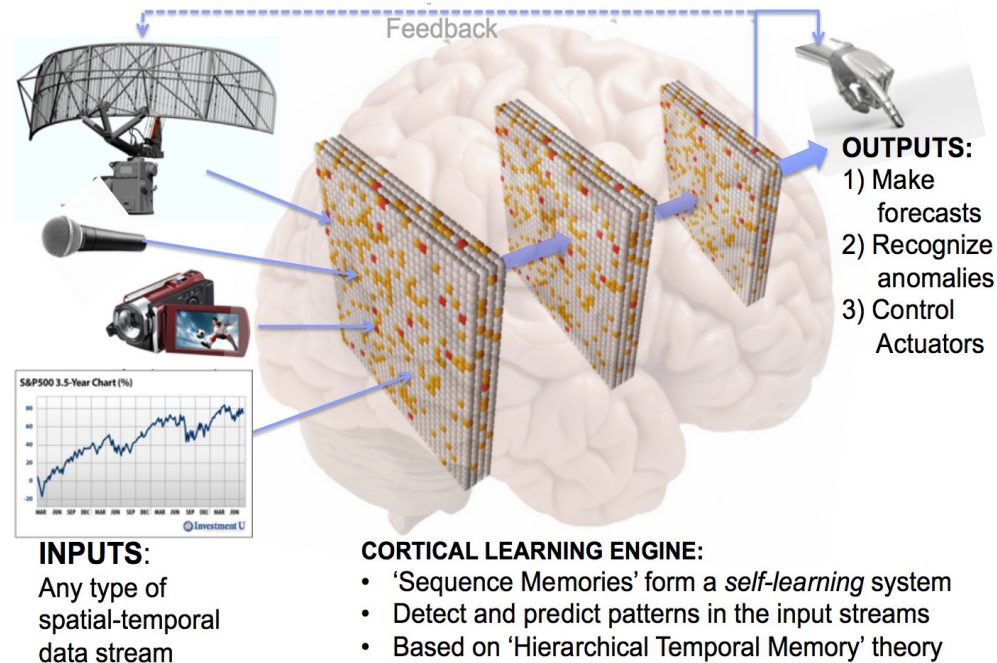
(2) **应用层**：AI+工具：金山办公； AI+建筑：广联达； AI+法律：通达海； AI+医疗：创业慧康，久远银海； AI+教育：科大讯飞； AI+网安：安恒信息、奇安信； AI+金融：同花顺； AI+交通：佳都科技； AI+政务：拓尔思； AI+遥感：航天宏图；

**风险提示**：1、AI技术发展不及预期；2、版权、伦理和监管风险；

## Jina AI光谱

AI实现  
模态融合

## 多模态AI

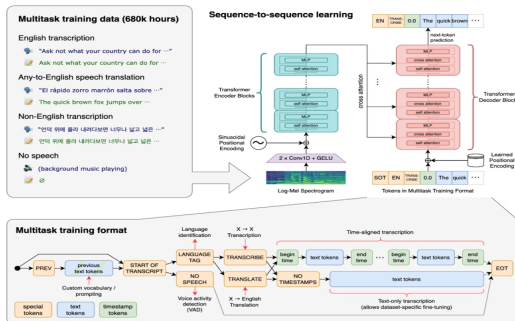
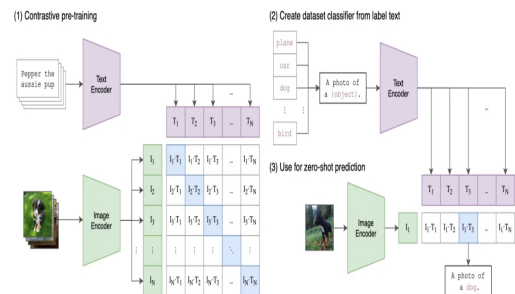
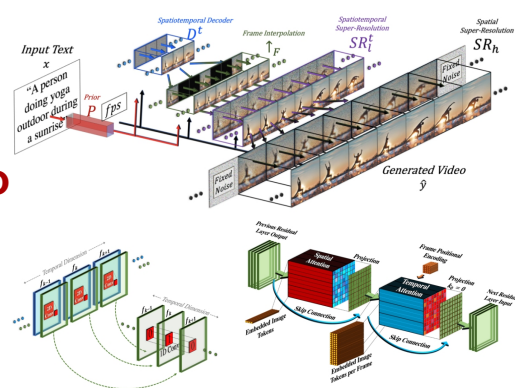


**数据：**将多模态数据融合处理

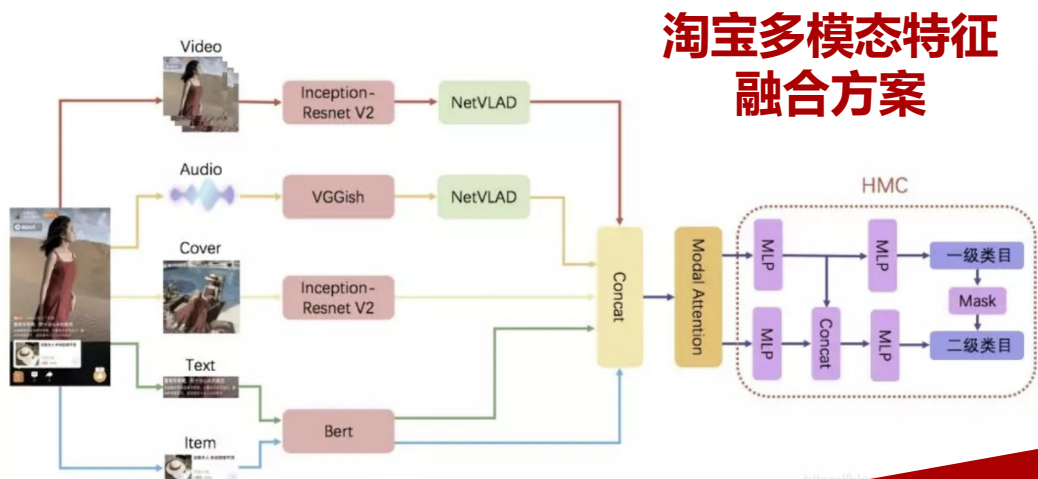
**模型：**更多融合人类脑神经机制，提升性能

**功能：**实现智能决策、跨模态任务、运动控制、智能预测反馈.....

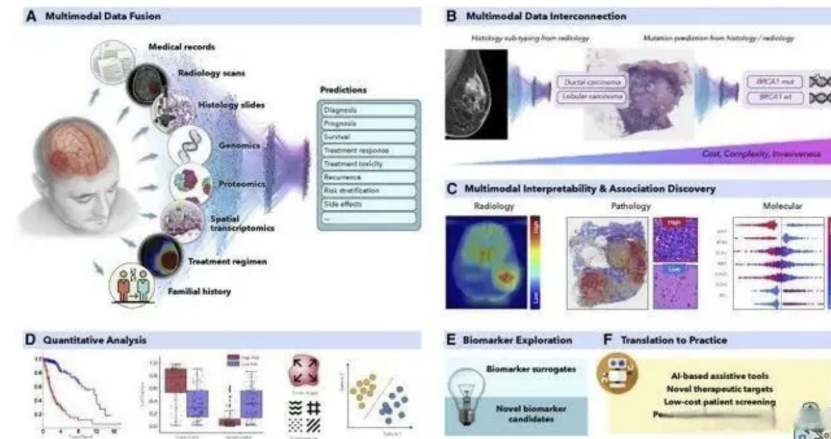
## 多模态AI模型应用场景（按架构分）

**Whisper**  
 语音-文本

**CLIP**  
 文本-图像

**Make-a-Video**  
 文本-视频

**模态融合**

架构	应用领域	融合内容
联合架构	视频分类	语音、视频、文本
	事件检测	语音、视频、文本
	情绪分析	语音、视频、文本
	视觉问答	图像、文本
	情感分析	语音、视频、文本
	语音识别	语音、视频
协同架构	跨模态搜索	图像、文本
	图像标注	图像、文本
	跨模态嵌入	图像、视频、文本
	转移学习	图像、文本
编解码器架构	图像标注	图像、文本
	视频解码	视频、文本
	图像合成	图像、文本



## 基于多模态融合的AI疾病诊断

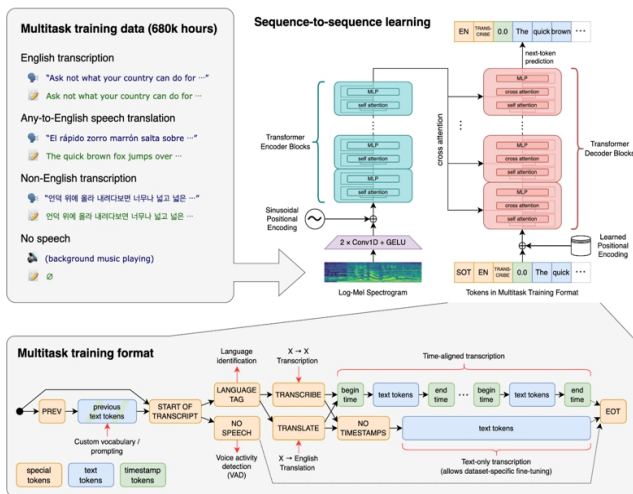


## 多模态AI的5大核心技术环节

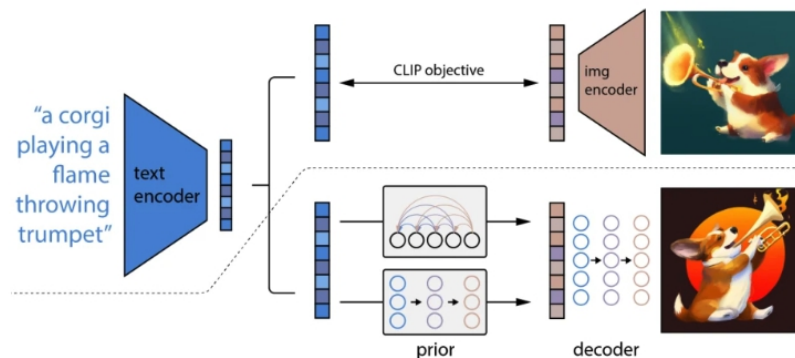
	表征 (Representation)	翻译 (Translation)	对齐 (Alignment)	融合 (Fusion)	联合学习 (Co-learning)
目标	实现模态互补 剔除模态冗余	模态映射	模态子成分 关联分析	信息整合	模态知识填充
技术路线	联合表示	Example-based	无监督方法	早期/晚期融合	Parallel learning
	协同表示	Decoder-Encoder	监督方法	多核学习	Zero Shot
			图像模型方法	图像模型	Hybrid
			神经网络方法	神经网络	

时间	提出者	模型名称	功能	意义
2021年1月	OpenAI	CLIP-DALL·E	以文搜图，按照文字描述生成对应图片	CLIP的zero-shot learning技术在各种数据集上的表现都很好
2021年5月	Google	MUM	多功能统一模型	可从 75 种不同语言中挖掘出的上下文信息对用户搜索结果进行优先排序
2021年9月	百度	DocVQA	跨模态文档理解	登顶DocVQA榜首
2021年11月	NVIDIA	GauGAN2	根据输入的文本/简笔画生成对应逼真的风景图、输入图像并编辑部分内容	可用文字和图画混合创造逼真的艺术
2021年11月	Microsoft & 北大	NvWa女娲	实现文本/草图转图像、图像补全、文字指示修改图像视频、文字/草图转视频、视频预测等	在8种图像和视频处理的视觉任务上具有出色的合成效果
2021年12月	NVIDIA	PoE GAN	文字描述、图像分割、草图都可以转化为图片，还可同时接受以上几种输入模态的任意两种组合	可以在单模态、多模态输入甚至无输入时生成图片。
2022年1月	百度	ERNIE-ViLG	图文双向生成	刷新文本生成图像、图像描述等多个跨模态生成任务最好效果
2022年1月	Meta	Au-HuBERT	通过输入语音音频和唇语视频内容，输出对应文本	在嘈杂的环境下，通过读唇可以将语言识别的准确性最高提升6倍。
2022年7月	Meta	Make-a-Scene	文本生成图像，并允许文本输入进行有针对性创作	用户获得更丰富的个人理念定制，从而生成更加具有针对性的画作
2022年9月	OpenAI	Whisper	语音生成文本，支持语音转录和翻译两项功能并接受各种语音格式	多模态AI模型有望进入商用时代
2022年11月	Meta	Make-a-Video	文本、图片生成短视频，根据输入的自然语言文本生成一段5秒钟左右的短视频。	AIGC进入视频创作领域
2022年11月	NVIDIA	Magic3D	根据文字描述生成 3D 模型，可将低分辨率生成的粗略模型优化为高分辨率的精细模型	3D建模效率更高，且成本更低

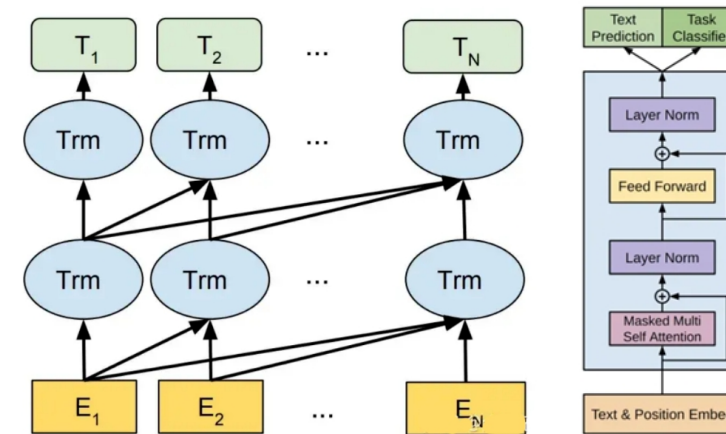
## Whisper 语音-文本模型



## DALL-E2 文本-图像模型



## ChatGPT



## 多样的模型调用接口

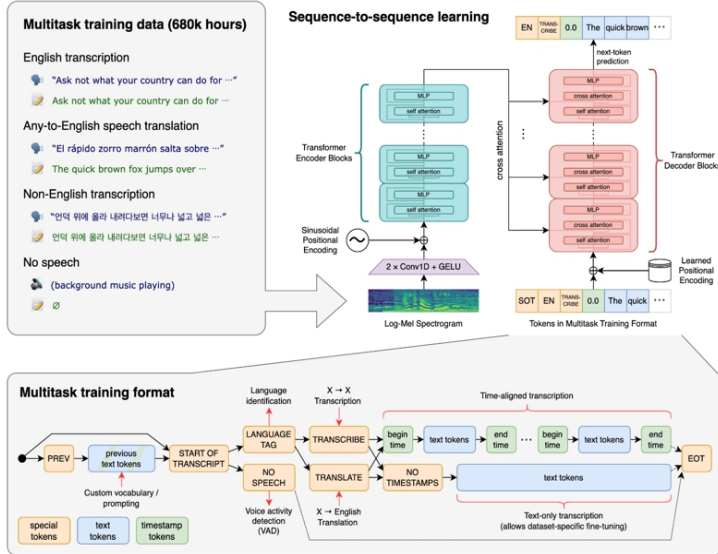
## GPT模型迭代

	GPT-1	GPT-2	GPT-3	Instruct GPT
论文年份	2018	2019	2020	2022
Transformer层数	12	48	96	—
参数量	1.2亿	15.8亿	1750亿	13亿
预训练数据量	5GB	40GB	45TB	—



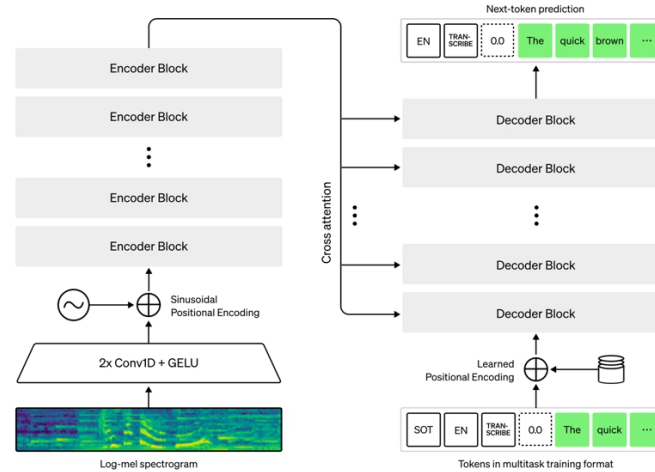
类别	名称	参数量
基础版本	Davinci	1750亿
	Curie	67亿
	Babbage	10亿
代码生成	Code-Cushman-001	120亿
关联分析	Text-similarity-davinci-001	1750亿
	Text-similarity-curie-001	60亿

## Whisper 语音-文本模型



架构简单

## 编码器/解码器基于Transformer 与其它单模态模型类似



畅想  
GPT X.0  
ChatGPT  
+  
Whisper  
+  
DALL-E2  
+  
...

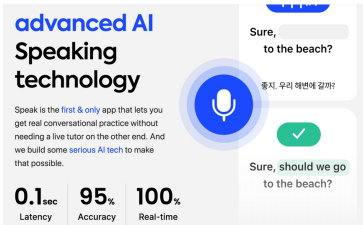
场景广泛

物美价廉

价格：0.006美元/分钟

模型选择灵活多样

语言学习



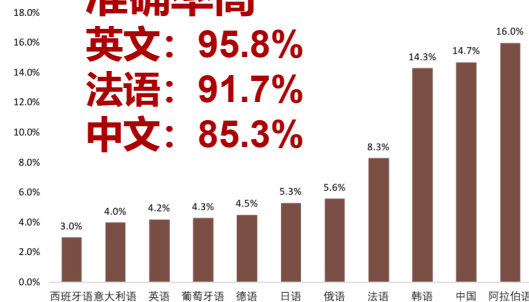
残障交流

智能翻译

电子病历

准确率高

英文：95.8%  
法语：91.7%  
中文：85.3%



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



## GPT-4

## 多模态

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

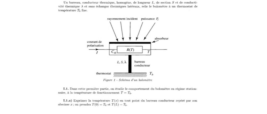


GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply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

## 复杂推理



GPT-4 To answer question 11.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  $T(L) = T_1$ .

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

$$\frac{dT}{dx} = 0$$

Integrating once, we get:

$$dT = C_1 dx \Rightarrow T = C_1 x + C_2$$

Integrating a second time:

$$T(x) = C_1 x + C_2$$

To determine the constants  $C_1$  and  $C_2$ , we use the boundary conditions:

$$T(0) = T_0 \Rightarrow C_2 = T_0$$

$$T(L) = T_1 \Rightarrow C_1 L + T_0 = T_1 \Rightarrow C_1 = \frac{T_1 - T_0}{L}$$

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

$$T(x) = \frac{T_1 - T_0}{L} x + T_0 = T_0 + \frac{T_1 - T_0}{L} x$$

The opposite of a linear temperature profile along the conductive bar, with a slope of  $\frac{T_1 - T_0}{L}$ .

## 视觉功能

User: What is funny about this image? Describe it panel by panel.



GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

## 深度语言理解

Language	Example
English >1B speakers	Why is the sky blue? A) Because the molecules that compose the Earth's atmosphere have a blue-ish color. B) Because the sky reflects the color of the Earth's oceans. C) Because the atmosphere preferentially scatters short wavelengths. D) Because the Earth's atmosphere preferentially absorbs all other colors. आकाश निरंज का रंग?
Marathi 90M speakers	A) कारण पृथ्वीच्या वातावरणाची रचना कार्बोनायड रेणूंचा रंग निरंज असतो B) कारण आकाशातून पृथ्वीच्या महासागरांचा रंग प्रतिबिंबित होतो C) कारण वातावरण श्रमूच्याने लहान तरंगलांबी विखते D) कारण पृथ्वीचे वातावरण इतर सर्व रंगांना श्रमूच्याने शोषून घेते Kāpēc debēis ir zilas?
Latvian 2M speakers	A) Jo molekūlām, kas veido Zemes atmosfēru, ir zilgana krāsa. B) Jo debesis atspoguļo Zemes okeānu krāsu. C) Jo atmosfēra galvenokārt izkliedē īsus viļņu garumus. D) Jo Zemes atmosfēra galvenokārt absorbē visas citādas krāsas.

## 应用场景

## New Bing



## Duolingo

## 语言教育

## Be my Eyes

## 图文转换

## Morgan Stanley

## 金融知识库

## Stripe

## 欺诈识别

## 可汗学院

## 学习规划

## 冰岛政府

## 语言文化保护

## 性能提升

## 人类测试表现优异

模拟Bar: 前10%  
生物奥赛: 前1%  
其它测试: 前20%

## ML测试进步显著

多项超越  
GPT3.5、SOTA

## 长文本处理

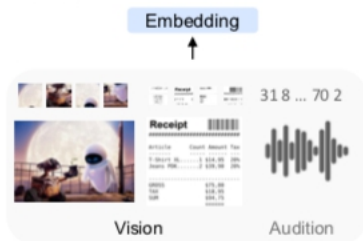

处理25,000+词

## 内容与安全

生成不被允许响应-82%  
符合事实的响应+40%

## 模型亮点：将视觉信息与大规模语言模型对齐

Kosmos-1 can perceive both language and    , learn in context , reason, and generate

What's in this picture?

Looks like a duck.


That's not a duck. Then what's it?

Looks more like a bunny.

Why?

It has bunny ears.

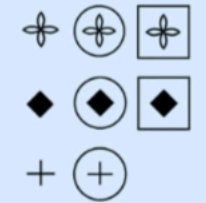
Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown.  
Description of downy woodpecker: It has white spots on its black wings and some red on its crown.



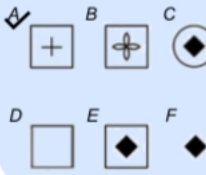
Question: what is the name of the woodpecker in the picture?

Downy

Here are eight images:



The following image is:



A

B

C

D

E

F

### 图像解释



Question: Explain why this photo is funny?  
Answer:



Question: Why did the little boy cry?  
Answer:

### 图像问答

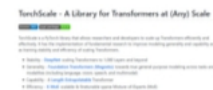


Question: What is the hairstyle of the blond called?  
Answer:



Question: When will the movie be released?  
Answer:

### 网页内容问答



What is TorchScale?

### 数学计算

$$5 + 4$$

Question: The result is?  
Answer:

### 内容识别



Question: What is the heart rate in this picture?  
Answer:



The time now is

语言任务

跨模态迁移

视觉任务

非语言推理

感知—语言 任务

### 图文交互



What's in this picture?

sausage roll.

How to cook it?

Soak sausage roll in ketchup. Bake in oven for 15 min. Cut and enjoy.

Can I put cheese in the dish?

Sure. But make sure it is melted.



What kind of animal is in the picture?

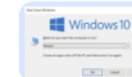
It's a zebra.

Where do they live in?

In Africa.

Tell me a joke about zebras.

Zebras get in trouble for crossing the road.



What's in this picture?

A screenshot of the Windows 10 shutdown window

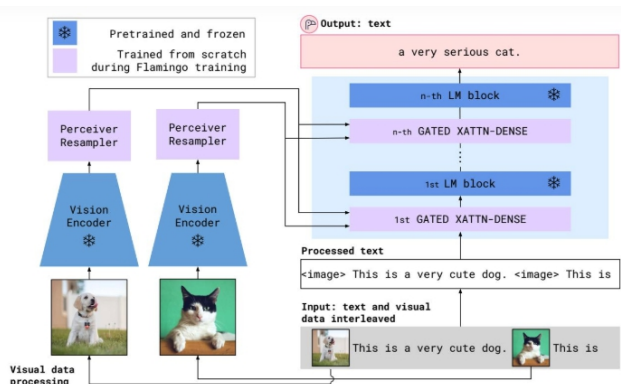
I would like to restart my computer. Which button should I click?

Press OK.

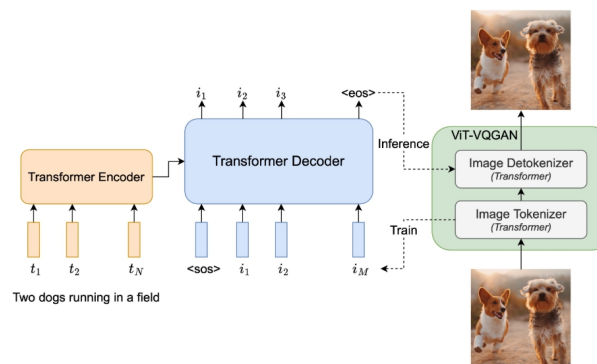
Now I would not like to restart. What can I do?

Click Cancel.

## Flamingo 图像-文本



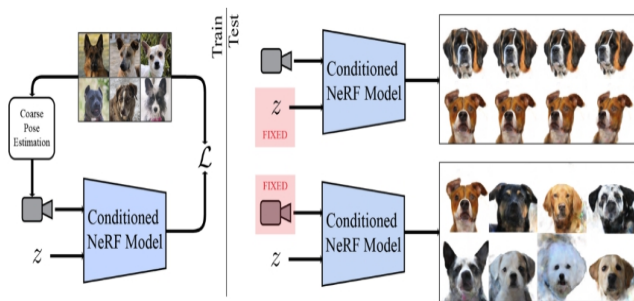
## Parti 文本-图像



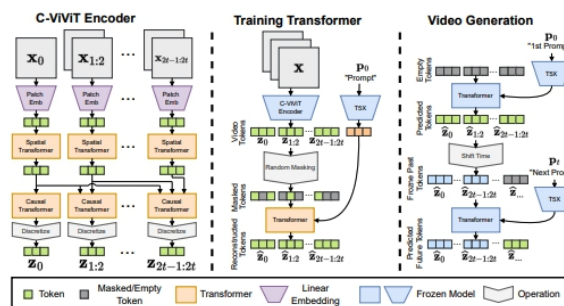
## 官方开源多个多模态模型

类别	模型	功能
计算机视觉	Pix2Seq	用于对象检测的语言建模框架
	DeViSE	视觉语义嵌入
多模式模型	LiT	将语义理解添加到图像模型
	PaLI	多语种语言图像学习
	FindIt	基于自然语言的通用对象定位
	VDTTS	视觉驱动文本到语音
音频生成	AudioLM	基于语言建模的音频生成

## LOLNerf 2D图像-3D图像



## Phenaki 文本-视频



## 文心一言

## 五大能力

文学创作



事实性问答，作品内容理解，小说续写

商业文案创作



企业命名，Slogan创作，新闻稿写作

数学逻辑推理



解决鸡兔同笼问题

中文理解



成语解释，经济学理论解释，诗词创作

多模态生成



图像、音频、视频生成

## 模型表现

## 文心千帆

平台服务

百度全套文心大模型

开发工具链

第三方开源大模型

具体功能

PPT创作

数字人直播带货

机票、酒店预订

持续优化

收费模式

0.012元/1000 tokens  
按调用输入输出总字数付费

## AI开发框架——飞桨



百度完整业务生态数据

## AI模型生态——文心



650+ 合作伙伴构建丰富应用场景

- **基础层：**

大模型：三六零，科大讯飞

数据服务：海天瑞声

- **应用层：**

AI+工具：金山办公； AI+建筑：广联达；

AI+法律：通达海； AI+医疗：创业慧康，久远银海；

AI+教育：科大讯飞； AI+网安：安恒信息、奇安信；

AI+金融：同花顺； AI+交通：佳都科技；

AI+政务：拓尔思； AI+遥感：航天宏图；

- 1、AI技术发展不及预期：**当前以ChatGPT为代表的AI模型以及其他多模态AI模型发展仍不成熟，存在一定缺陷；
- 2、版权、伦理和监管风险：**AIGC生成的内容依赖现有版权素材，另外不当使用或模型自身问题可能导致不良后果；

## 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

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：<http://research.stocke.com.cn>