

ChatGPT: AI模型框架研究 ——AI行业深度报告

行业评级：看好

2023年3月25日

分析师	刘雯蜀
邮箱	liuwenshu03@stocke.com.cn
证书编号	s1230523020002

一、AI框架重要性日益突显，框架技术发展进入繁荣期，国内AI框架技术加速发展：

- 1、AI框架作为衔接数据和模型的重要桥梁，发展进入繁荣期，国内外框架功能及性能加速迭代；
- 2、Pytorch、Tensorflow占据AI框架市场主导地位，国内大厂加速布局AI框架技术；
- 3、AI框架技术从工具逐步走向社区，生态加速形成，未来围绕安全可信、场景落等维度呈现显著发展趋势；

二、GPT开启AI大模型时代，国内外大厂发力布局，商业化空间加速打开：

- 1、数据、算法、模型三轮驱动AI发展，大模型优势显著，成为AI主流方向；
- 2、GPT开启千亿参数级AI大模型时代，语言、视觉、科学计算等大模型快速发展；
- 3、微软加速AI商用化进程，国内大厂发力布局，看好在细分场景下的应用落地；

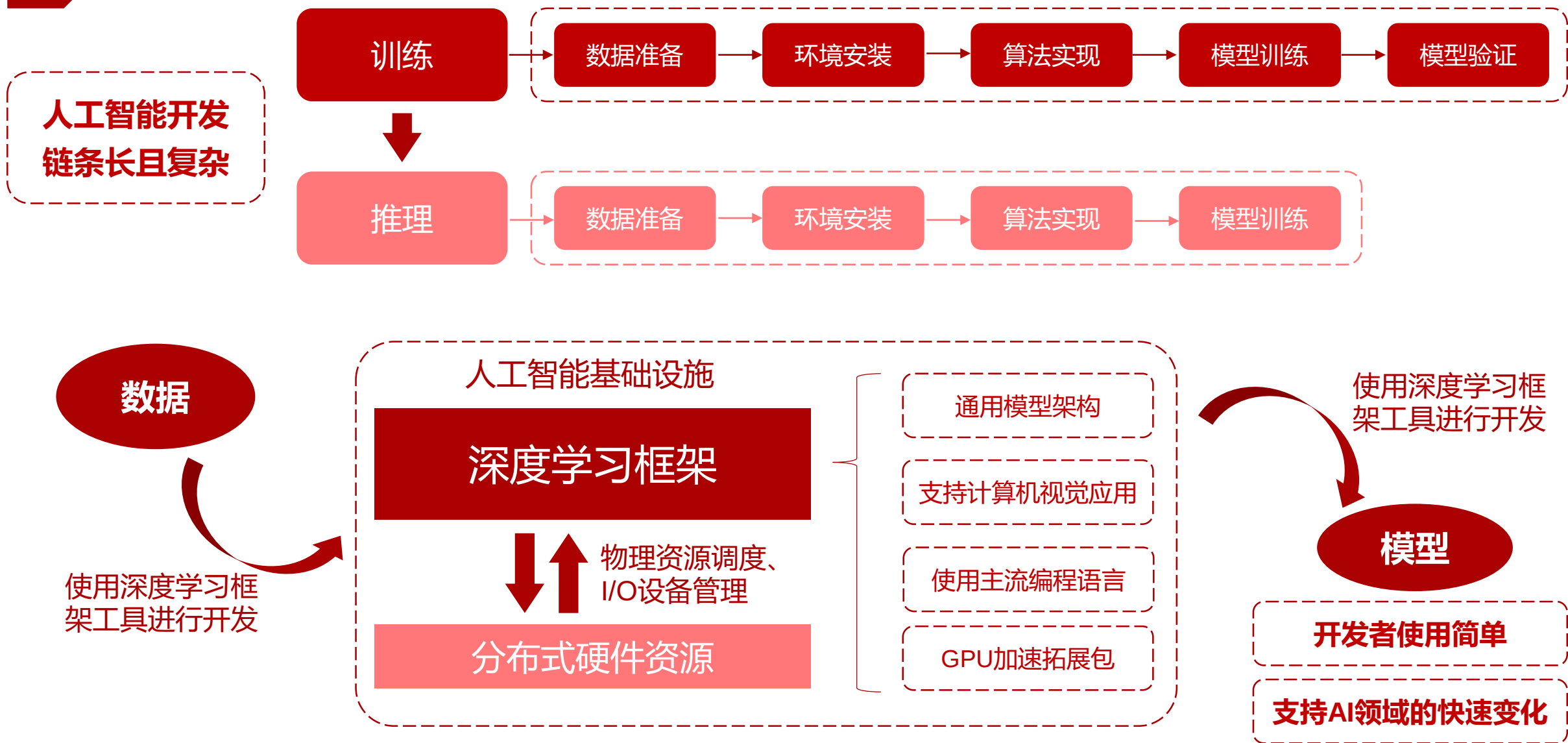
三、建议关注标的：

- 1、基础层：AI算力：中科曙光；大模型：360，科大讯飞
- 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

风险提示： 1、AI技术发展不及预期； 2、版权、伦理和监管风险；

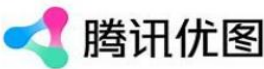
AI框架

深度学习框架：人工智能时代的操作系统

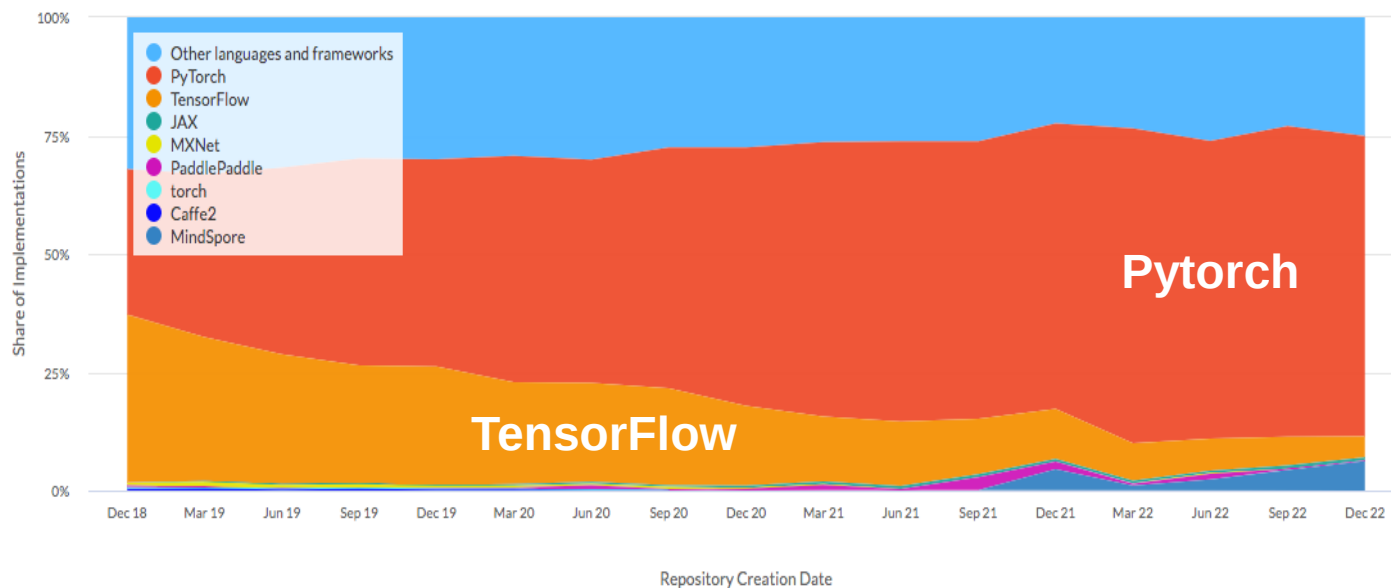




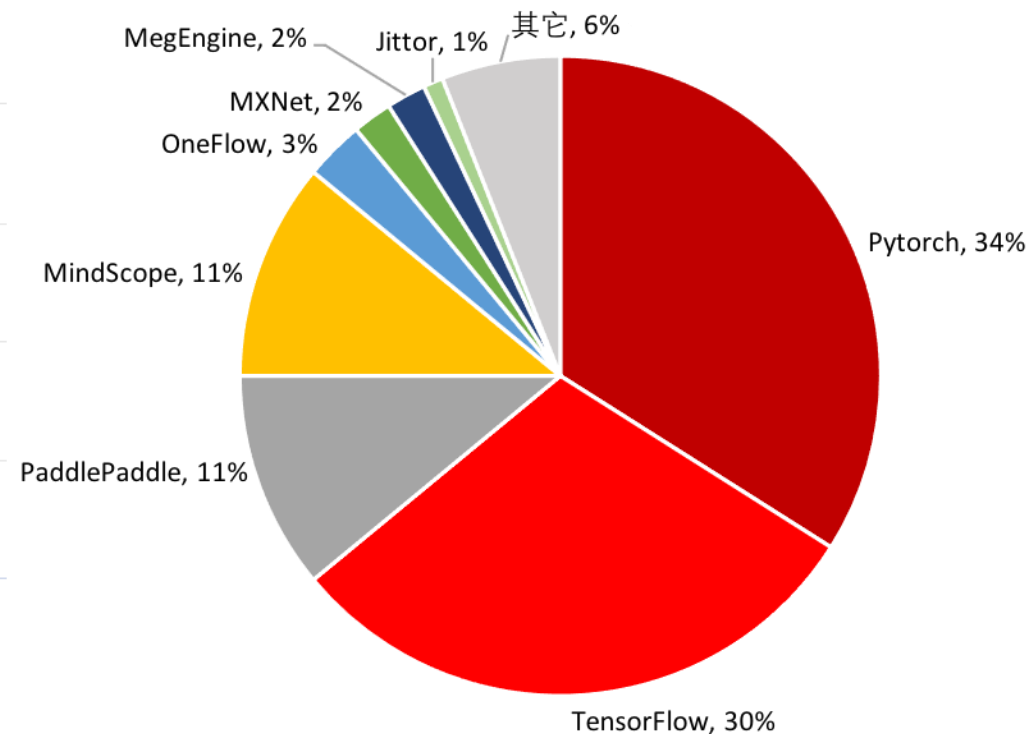
国内外深度学习框架

发布时间	开发公司	深度学习框架	语言	是否开源	计算图	是否是分布式框架	特点/优点
2013		Caffe	Python	√	静态	×	速度快、使用方便、社区好
2014		CNTK	Lua, Python (new)	√	静态	√	性能高、适合做语音任务
2015			C++	√	动态	√	高效灵活、易用
2016			Python	√	动静兼容	√	容易上手
2017			C++	√	静态	√	简单清晰
2020			Lua, Python (new)	√			移动端高性能、通用轻便
2020			Python	√	基于源码转换自动微分，不依赖计算图	√	高效灵活、易用
202x			C++、CUDA、Python	√	动静合一		灵活高效

2018-2022年全球论文发表数量（按使用框架分）



2022年中国开发者人工智能框架使用率



Pytorch版本平均每3~4个月更新一次，功能服务持续扩充



多维优势支持Pytorch实现对TensorFlow的反超

门槛低

只需要Numpy和基本深度学习概念

代码简洁灵活

基于动态图机制，网络搭建更方便

文档规范

官方社区可查看各版本文档

资源丰富

arXiv新算法大多基于Pytorch实现

开发者多

Github上贡献者1100+

大厂支撑

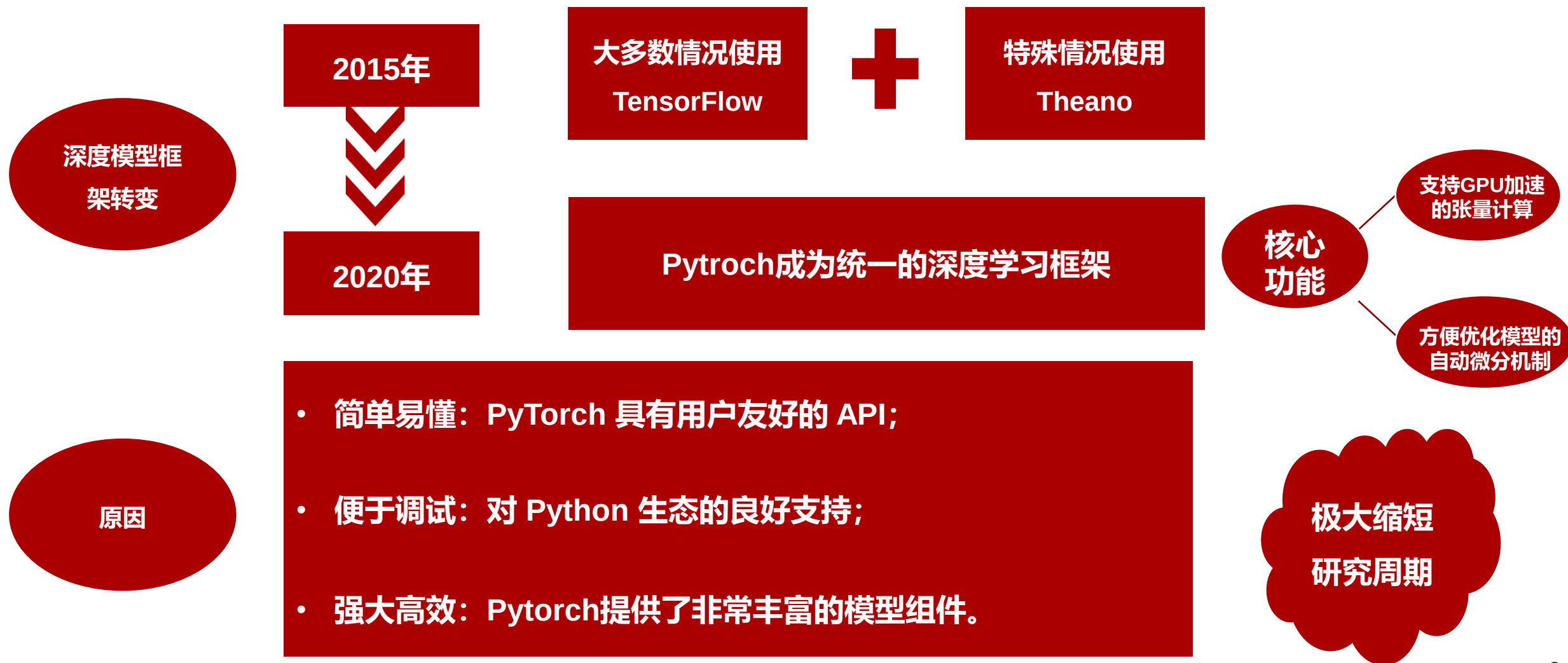
Meta维护开发

适用人群广泛

深度学习初学者: 快速实现模型算法，加深深度学习概念认识；
机器学习爱好者: 快速实现人脸识别、目标检测、图像生成等AI功能及实验；
算法研究员: 最新arXiv论文算法快速复现及开发；



01 Open AI: 从多种框架的使用到专注于Pytorch

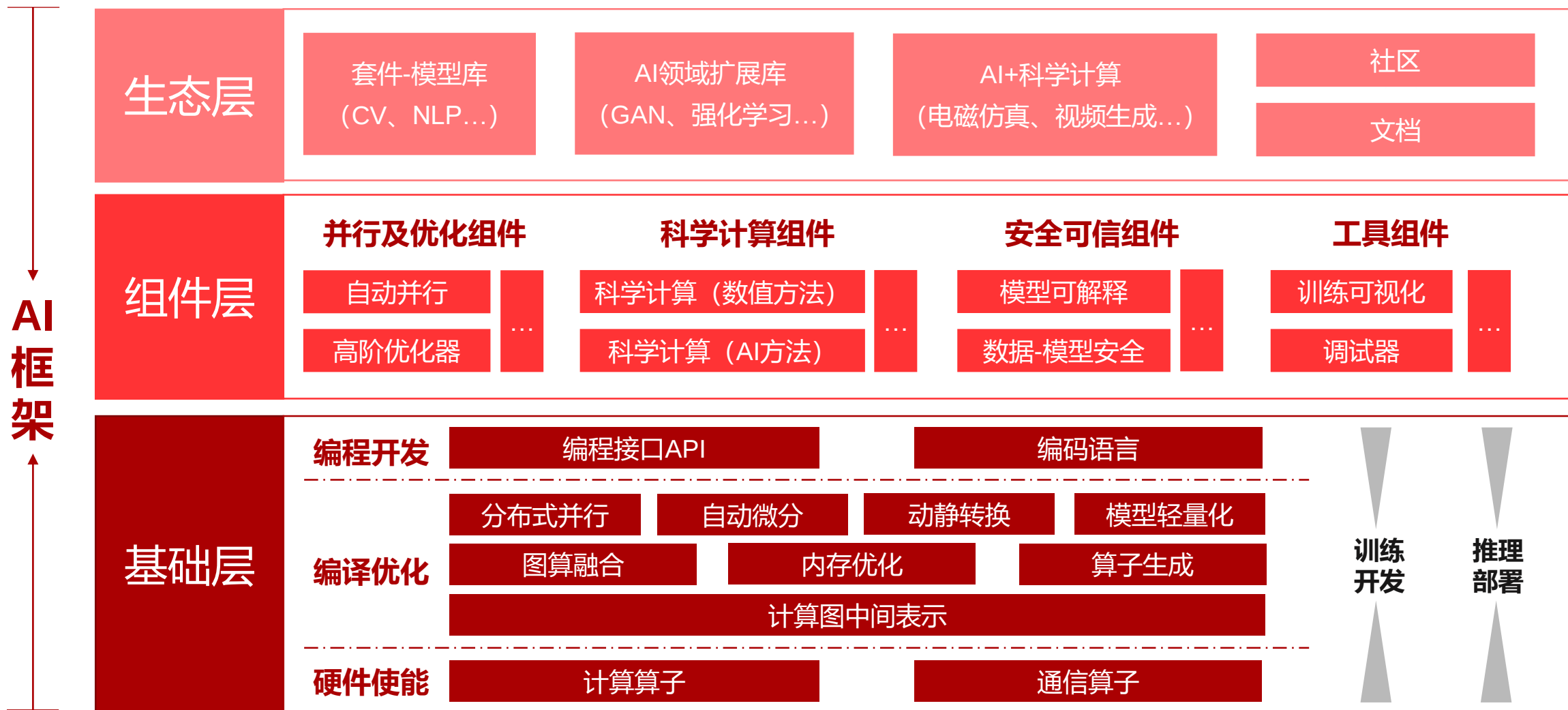


01 Tensorflow: 谷歌开源的向更加易用发展的主流学习框架

Tensorflow从0.1到2.0的发展历程



AI框架技术呈现三层次结构，从工具走向社区生态



飞桨企业版

零门槛AI开发平台

全功能AI开发平台

飞桨产业级深度学习开源开放平台

工具与
组件自动化
深度学习

强化学习

联邦学习

图学习

科学计算

量子机器学习

生物计算

低代码开发工具

预训练模型应用工具

可视化分析工具

安全与隐私工具

资源管理与调度工具

云上部署编排工具

端到端
开发套件

语义理解

文字识别

图像分类

目标检测

图像分割

图像生成

大模型训推一体

基础
模型库

自然语言处理

计算机视觉

语音

推荐

时间序列

文心大模型

核心
框架

开发

动态图

静态图

训练

大规模
分布式训练产业级
数据处理

推理部署

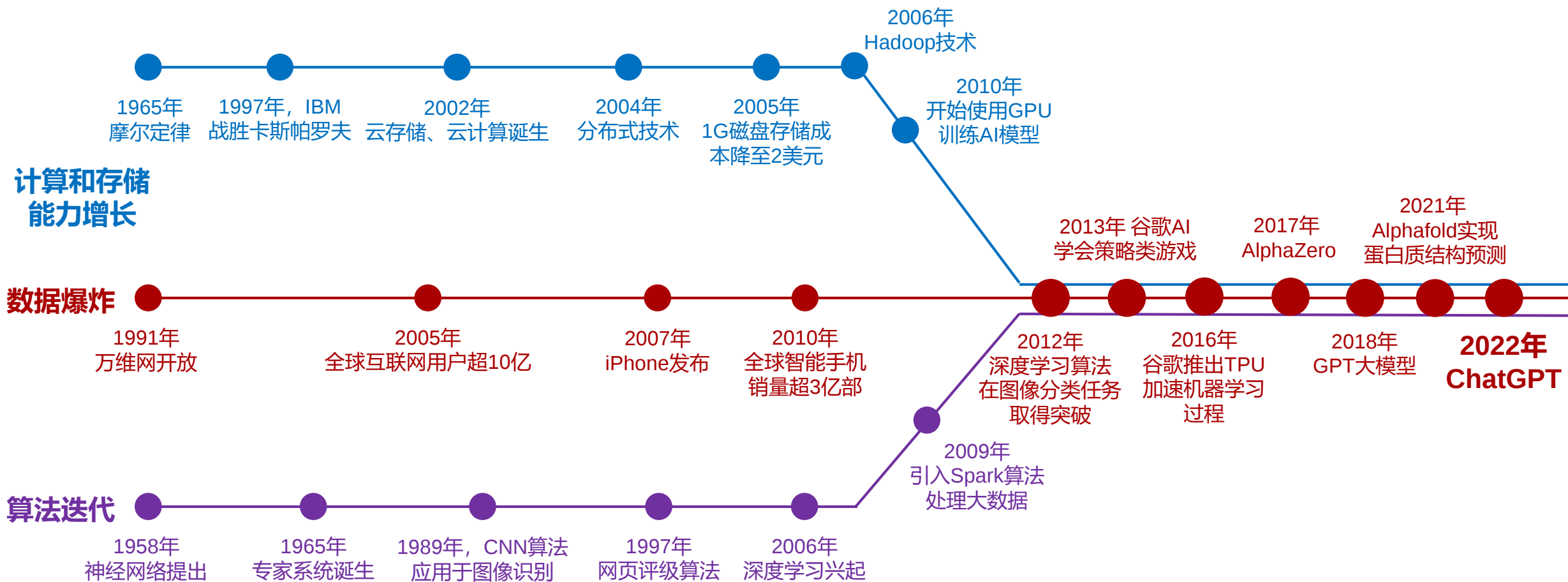
模型
压缩服务器
推理引擎边缘与移动
端推理引擎前端
推理引擎服务化
部署全场景
统一部署学习
与实训
社区

发展趋势

前景展望

泛开发	前端便捷化	后端高效化	多种开发语言无缝衔接	+	动静图转换能力提升 后端运行效率
全场景	全场景标准化互通		AI框架与硬件平台解耦，通过标准接口实现跨设备平台快速部署		
超大规模	混合并行	分布式处理	突破五堵墙：内存墙+算力墙+通信墙+调优墙+部署墙		
科学计算	自动微分	统一加速引擎	丰富编程接口	+	内置专业领域科学计算套件
安全可信	鲁棒性检测	模型可解释	提供丰富的 AI 鲁棒性检测工具		
工程化	模型自适应	框架精细化	AI 模型的压缩和端侧推理框架的轻量化		

AI大模型

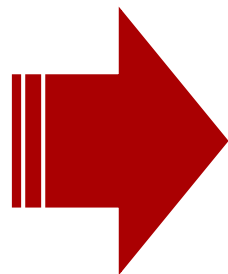


AI大模型优势

泛化性+通用性



开发门槛低



大模型意义

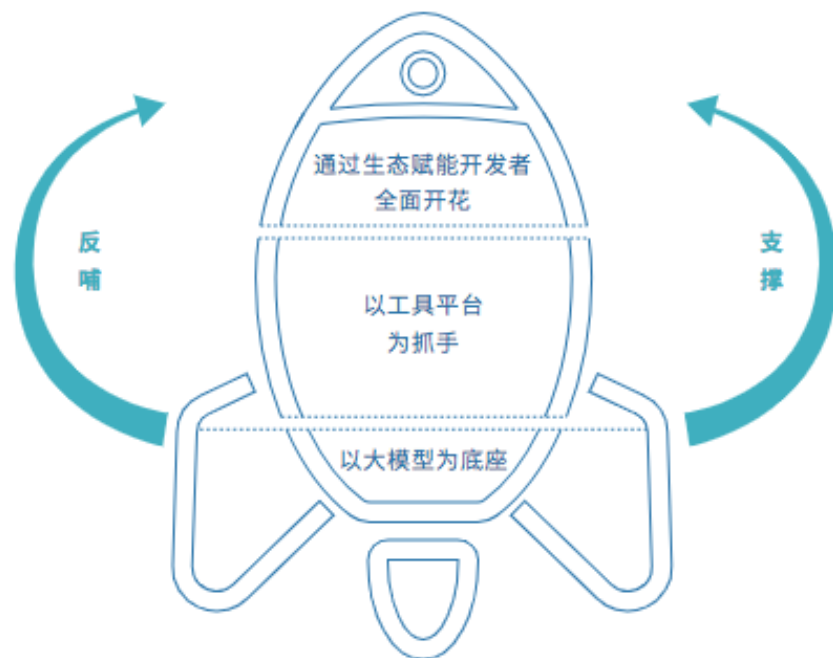
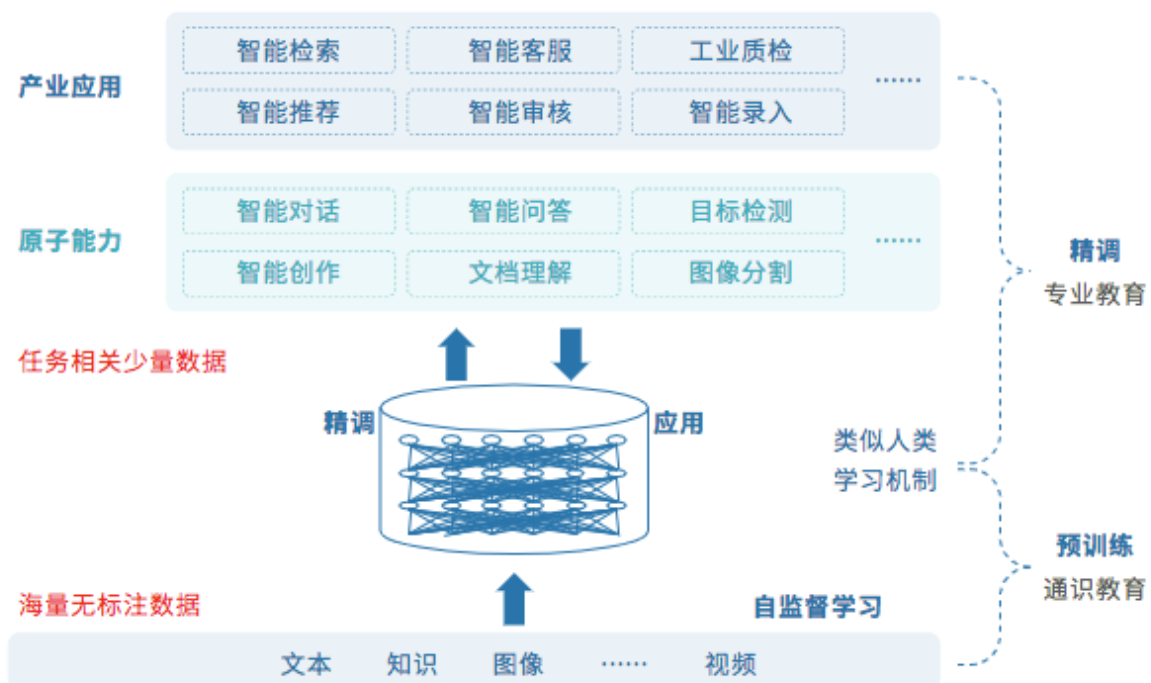
AI应用通用化



AI开发工程化



项目建设集约化





02

AI迈入大模型时代，参数量过千亿

	GPT-1	GPT-2	GPT-3	GPT-4
推出年份	2018	2019	2020	2023
Transformer 层数	12	48	96	-
参数量	1.2亿	15.8亿	1750亿	-
预训练数据量	5GB	40GB	45TB	-

OpenAi GPT-3 (1758)

Google BigBird (1750)

Microsoft Truning-NLG (172)

Facebook M2m-100 (150)

Google T5 (110)

NVIDIA Megatron-LM (83)

OpenAi GPT-2 (15.8)

Facebook RobertTa (3.35)

Google ALBERT (0.31)

Facebook BART

Google ELECTRA (1.02)

Google BERT-Large (3.4)

Facebook XLM

OpenAi GPT-1 (1.2)

百度 ERINE2.0

Google BERT-base (1.1)

百度 ERINE1.0

Google LaMDA (2800)

Google Gopher (2800)

百度 ERNIE 3.0 Titan (2600)

Naver Corp HyperCLOVA (2040)

Google FLAN (1370)

GLM (1300)

百度 ERINE3.0 (100)

Eleuther AI GPT-j (60)

Google PaLM (5400)

微软和英伟达 Megatron-Turing NLG (5300)

BigScience BLOOM (1760)

Meta AI OPT (1750)

EleutherAI GPT-NeoX (200)

OpenAi InstructGBT (13)

2018

2019

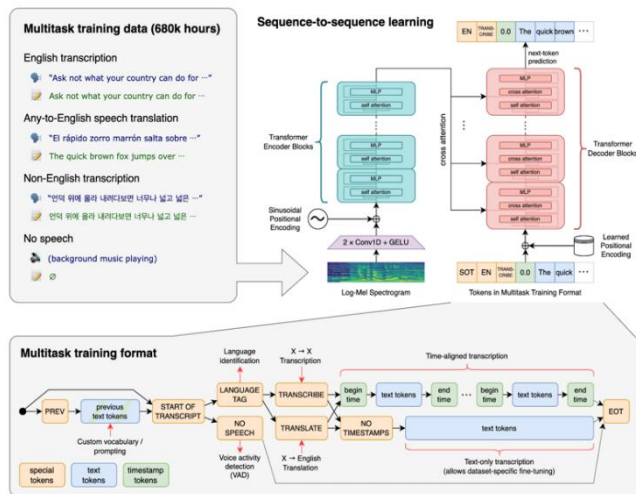
2020

2021

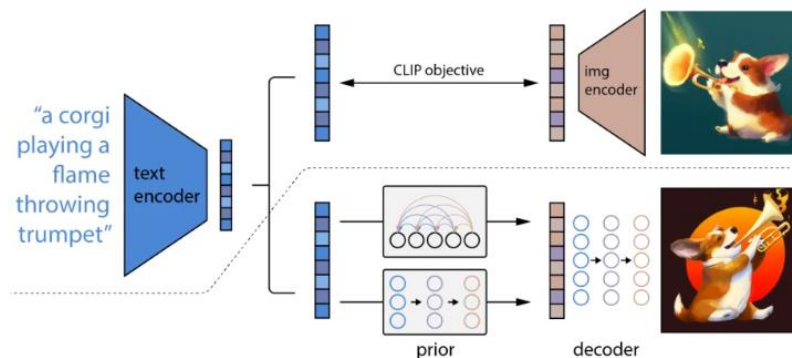
2022

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	NVIDA	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	NVIDA	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9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	NVIDA	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

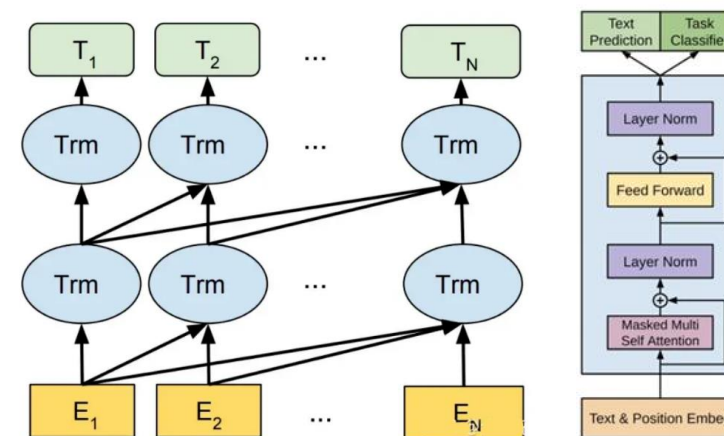
Whisper 语音-文本模型



DALL-E2 文本-图像模型



ChatGPT



多样的模型调用接口

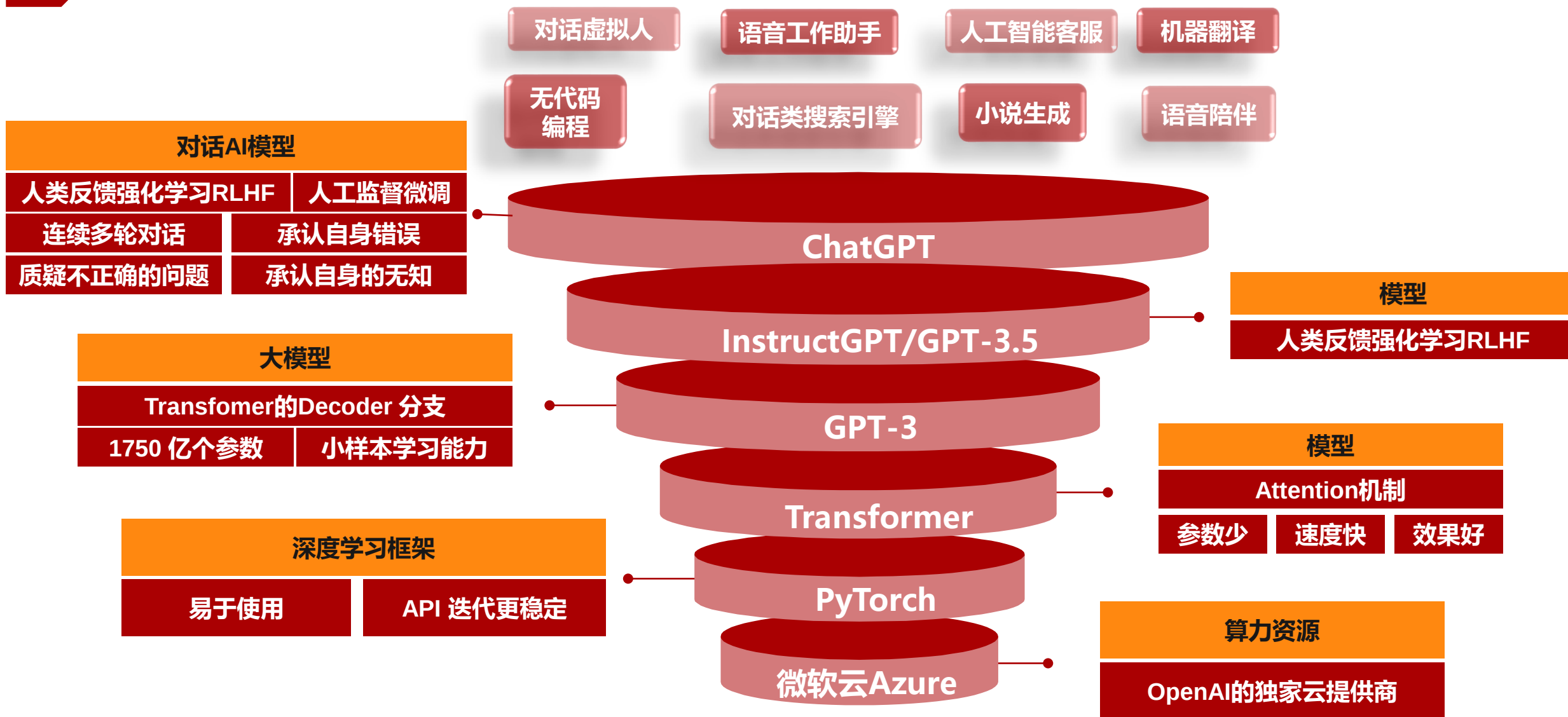
GPT模型迭代

	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

02 ChatGPT实现路径：算力与框架支持，应用百花齐放



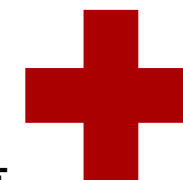
Chat
GPTInstruct
GPT

GPT-3



- ✓ 增加Chat属性
- ✓ 网页公众测试入口

➤ 略微降低参数量



- ✓ 代码训练
- ✓ 指令微调 (instruction tuning)
- ✓ 基于人类反馈的强化学习 (RLHF)

➤ 参数数量降低了100倍
(1750亿->13亿)

A prompt is
sampled from our
prompt dataset.

从问题库中抽取
问题

A labeler
demonstrates the
desired output
behavior.

标记者书写期待
的回复

This data is used to
fine-tune GPT-3.5
with supervised
learning.

被标记的数据用来调
优GPT-3.5



A prompt and
several model
outputs are
sampled.

采样问题，并列出
所有模型和标记者
的回复

A labeler ranks the
outputs from best
to worst.

标记者排序所有
标记答案

This data is used
to train our
reward model.

用排序答案训练
奖励模型



A new prompt is
sampled from the
dataset.

The PPO model is
initialized from the
supervised policy.

The policy generates
an output.

The reward model
calculates a reward
for the output.

The reward is used
to update the policy
using PPO.



通过模型生
成初步答案

输入奖励模
型得到分数
和优化参数



将ChatGPT整合进Bing和Edge

新版Bing搜索引擎
四大技术突破

Bing在OpenAI的下一代LLM模型上运行，该模型专门为搜索定制，比ChatGPT更强大

普罗米修斯（Prometheus）模型：可以提高搜索结果相关性，并对答案进行注释

通过将人工智能模型应用于核心搜索算法，改进了核心搜索指数，使得搜索结果相关性实现飞跃

搜索与聊天相结合，除了传统的搜索结果外，还提供了聊天界面

搜索模型

搜索性能

答案相关

用户体验

能动的提供解决方案：
创建菜谱、制定旅行计划、诗歌创作等

新增聊天窗口

Sure, I can help you with that. Based on the web search results, here is a possible workout plan for your arms and abs with no sit-ups and no gym equipment. It should only take 30 minutes. Please consult your doctor before starting any new exercise routine and stop if you feel any pain or discomfort.

- Warm up for 5 minutes with some light cardio, such as jogging, skipping, or jumping jacks.
- Do 3 sets of 10 repetitions of each of the following exercises, resting for 30 seconds between sets:

You're already on the waitlist

Access the new Bing faster

Learn more See another example

传统信息搜索框

新版Bing功能展示

Create a 3-course menu.

Help plan my special anniversary trip.

What art ideas can I do with my kid?

Can you help me get fit?

Write a rhyming poem.

I just went fishing in Bighorn in Montana and I would like to go fishing in the Florida Keys in the Spring. What do I need to do differently to prepare?

Can you help me write a story?

I need some help with my coding.

I need a big fast car.

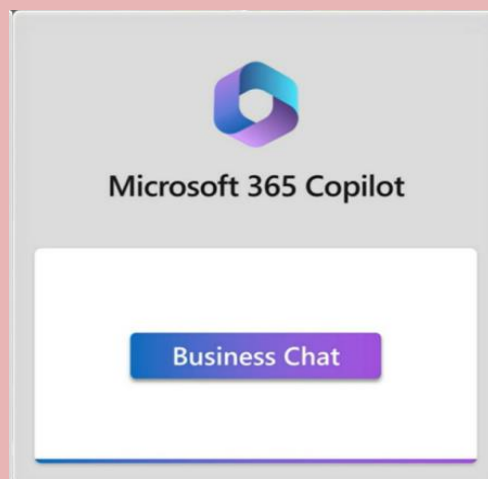
Microsoft 365 Copilot

- Copilot旨在协助用户生成文档、电子邮件、演示文稿和更多内容
- Copilot主要由OpenAI的GPT-4驱动，会与微软365应用程序一起，作为聊天机器人的模式，出现在侧边栏

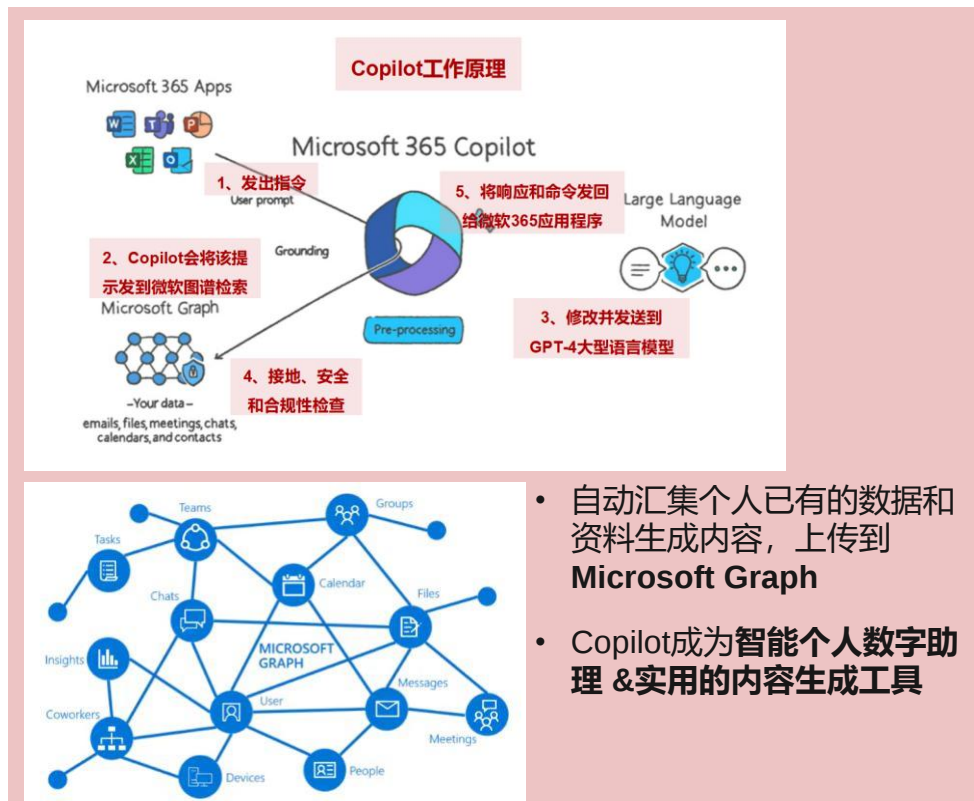


Copilot工作方式

- Copilot嵌入到人们每天使用的Microsoft 365 应用中
- **商务聊天 Business Chat.** Business Chat 将汇总电子邮件、文件、文档、会议、聊天记录、日历等资料，并归纳总结



Copilot工作原理





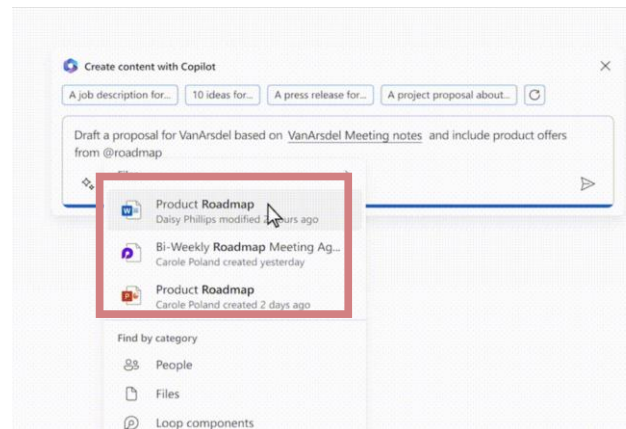
应用场景

Copilot in PowerPoint



- Copilot 可以跨应用程序生成内容。例如，根据Word文档，可以生成一个10张幻灯片的PPT
- 提升演讲效果，增加字体大小和间距，在演讲稿中添加演讲提醒
- 一键压缩冗长的演示文稿，调整布局、重新格式化文本和完美的时间动画。

Copilot in Word

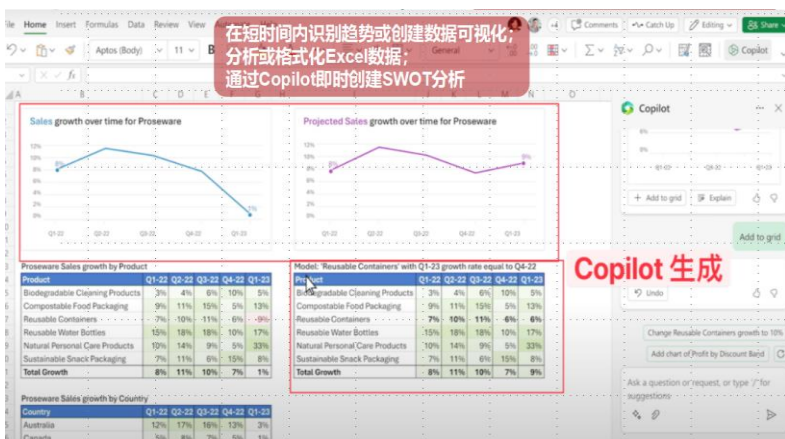


- Copilot可以根据需求创建初稿
- 对文本内容进行提炼、改写、简化，查漏补缺
- 用户还可以根据需求调整AI的语气，包括严肃、热情、感谢等



应用场景

Copilot in Excel



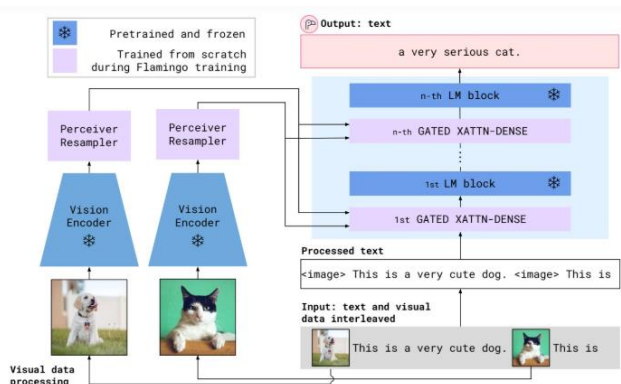
- 在短时间内识别趋势或创建数据可视化
- 数据归纳处理，分析或格式化Excel数据，生成直观图像
- Excel用户可以通过Copilot即时创建SWOT分析或基于数据的PivotTable

Copilot in Teams

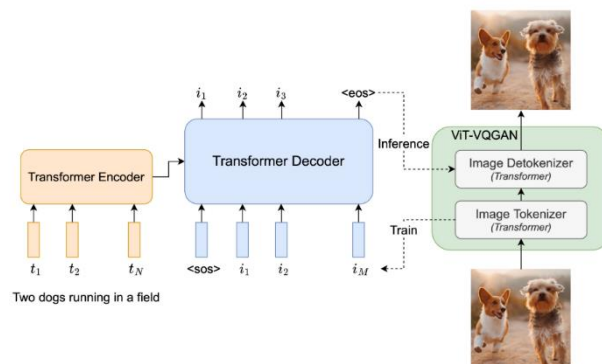


- 在对话上下文中提供实时摘要和操作项，进行会议内容总结，提醒可能错过的东西
- 如果参加会议时间较晚，copilot会提供一份错过的内容摘要，从而提高会议效率

Flamingo 图像-文本



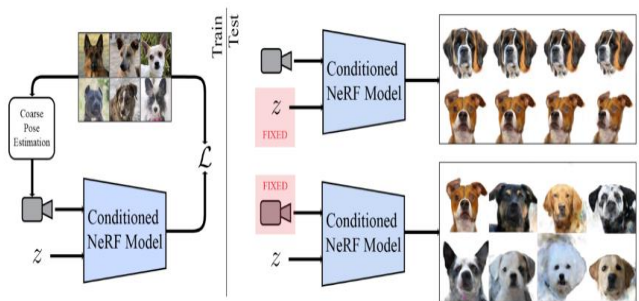
Parti 文本-图像



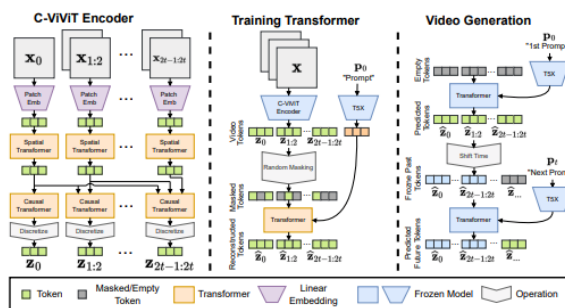
官方开源多个多模态模型

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
多模式模型	DeViSE	视觉语义嵌入
	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
音频生成	VDTTS	视觉驱动文本到语音
	AudioLM	基于语言建模的音频生成

LOLNerf 2D图像-3D图像



Phenaki 文本-视频



■ 大模型的主要玩家有科技大厂、高校和新型研发机构，形成了四种合作模式

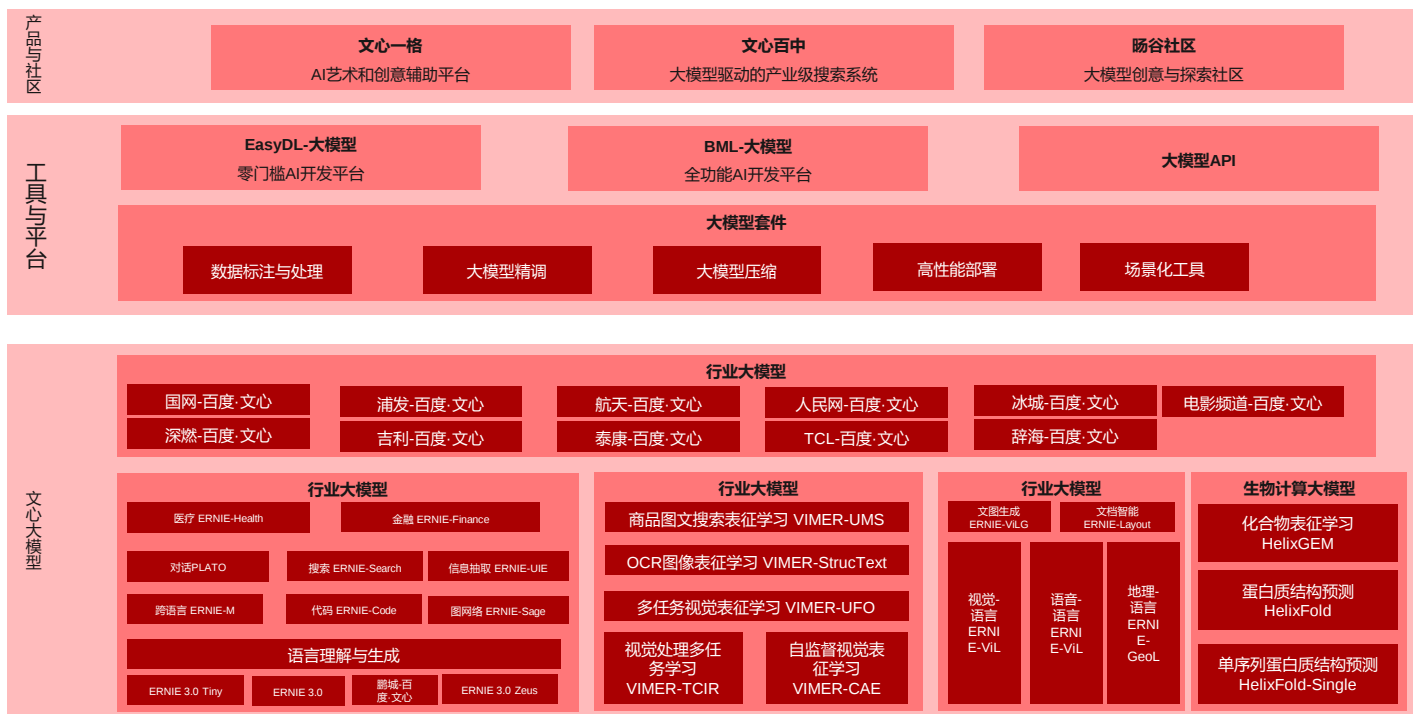
- (1) 大厂独立完成 (2) 机构+高校 (3) 大厂+高校 (4) 大厂+机构+高校。大厂通过资金优势、数据优势往往可以独立完成或主导合作。机构凭借行业领袖的团队和政府的资金支持，可以主导合作。而高校凭借行业领袖的团队提供科研能力支持。
- 过去来看，由于大厂受到商业任务限制，资金和数据优势未能充分发挥。而未来，在ChatGPT之后，经过验证的模式铺平商业决策之路，将逐步成为未来大模型的主导力量。



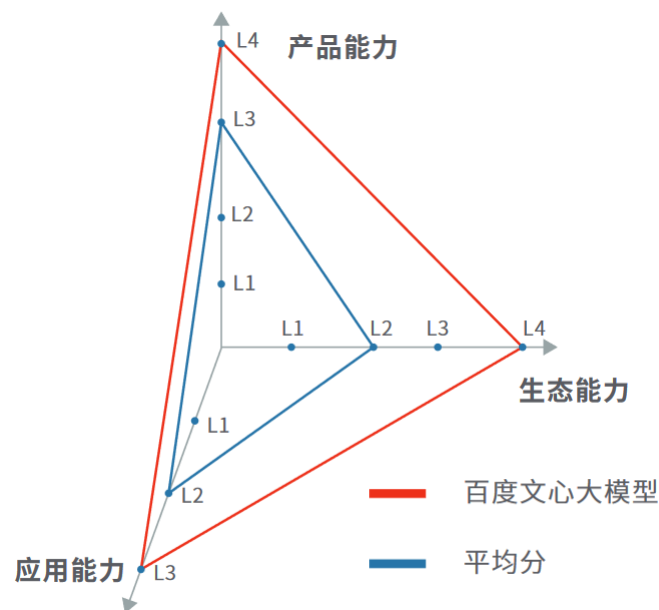
坐拥大模型+训练框架+数据+社区多重优势，百度有望成为AIGC领域率先实现商业化的领头羊。

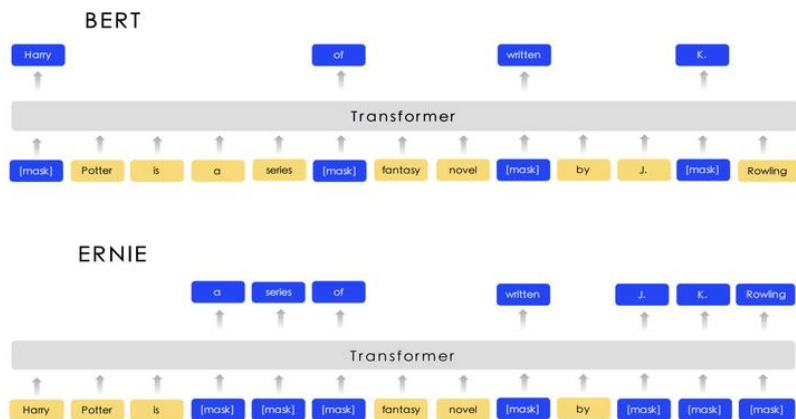
- 自2019年发布ERNIE 1.0，百度持续投入大模型的技术创新与产业应用，布局了NLP、CV、跨模态等大模型，率先提出行业大模型，成了支撑大模型产业落地的关键路径，构建文心大模型层、工具平台层、产品与社区三层体系。
- 根据IDC的大模型评分，在产品能力、生态能力和应用能力三个维度上百度均位于第一梯队，且在生态维度远高于平均水平，这得益于百度的大模型框架“飞桨”、旻谷社区。
- 百度于2023年3月发布“文心一言”，成为首款中文生成式对话大模型产品。

文心大模型与产品框架

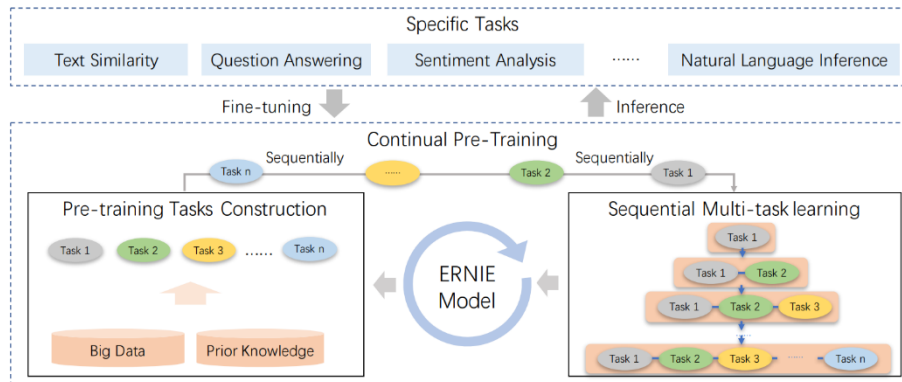


文心大模型评分

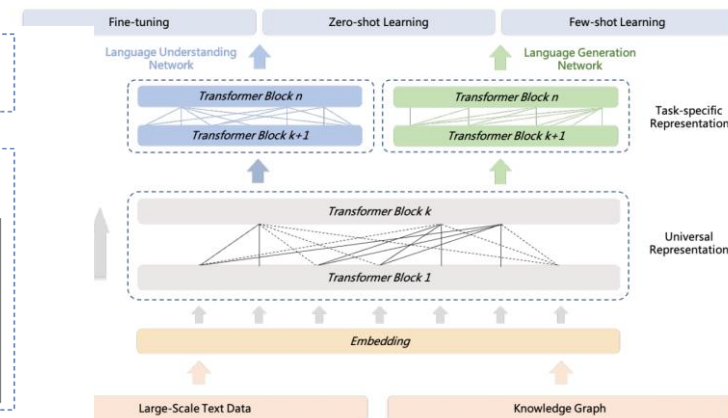




ERNIE 1.0架构：改进了MLM任务



ERNIE 2.0：+持续学习框架



ERNIE 3.0、3.0TITAN：+参数量

ERNIE版本	1.0	2.0	3.0	3.0 TITAN
推出年份	2019	2020	2021	2022
参数量	参考bert base(1.1亿)	参考bert base(1.1亿), bert large (3.4亿)	100亿	2608亿
预训练数据量	Wiki, baike, news, tieba	wiki, news, dialogue, IR, discourse relation	4TB	-

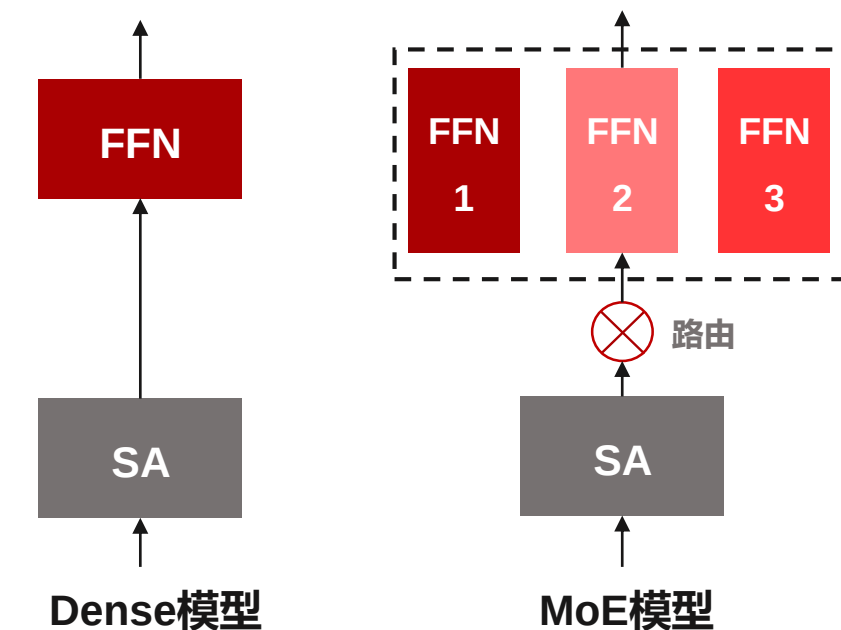
■ 阿里巴巴率先构建大模型统一底座、通过训练策略大幅提升稀疏参数大模型框架训练效率，在大模型框架上具备领先地位。

- 阿里巴巴2021年3月发布M6，成为国内最早提出千亿模型的厂商，同年发布十万亿模型M6-10T，通过expert prototyping训练策略成功实施MoE稀疏参数模型，使模型达到10万亿参数级别。
- 2022年9月发布通义大模型，通过统一学习范式M6-OFA和模块化的设计，提升大模型跨模态能力和效率。
- 2023年报电话会上，集团CEO张勇表示针对生成式AI趋势，将全力构建预训练大模型。

阿里通义大模型架构

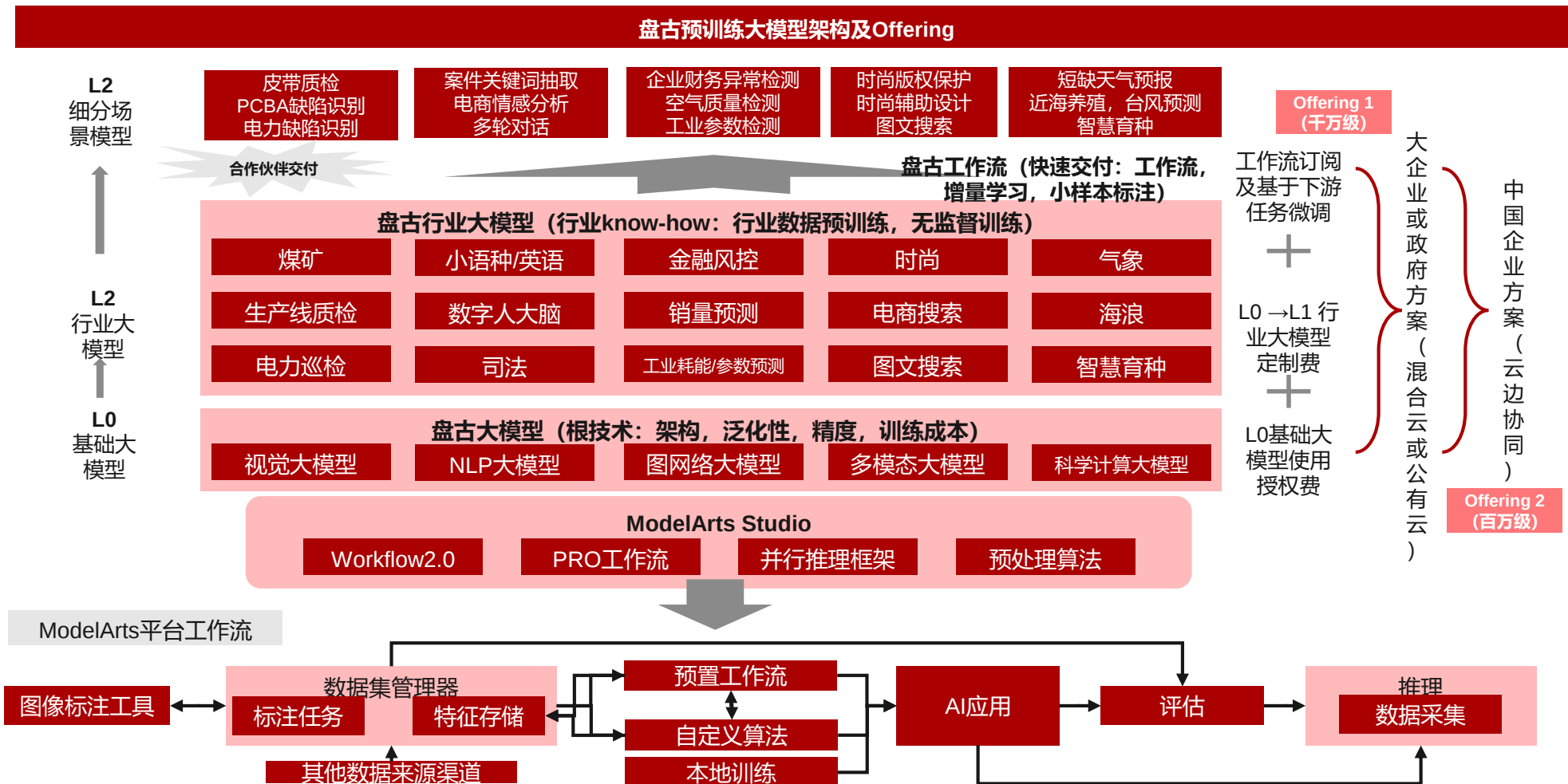


Dense模型与MoE模型



■ 华为盘古大模型深耕实业，拥有更广泛的行业大模型，具备更强的落地能力。

- 基于ModelArts AI工作平台的盘古大模型2021年4月发布，目前已应用于10+行业的100+应用场景。
- 根据信通院模型开发和模型能力两方面测评，均为优异水平。



- **基础层：**

AI算力：中科曙光

大模型：360，科大讯飞

- **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

- 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>