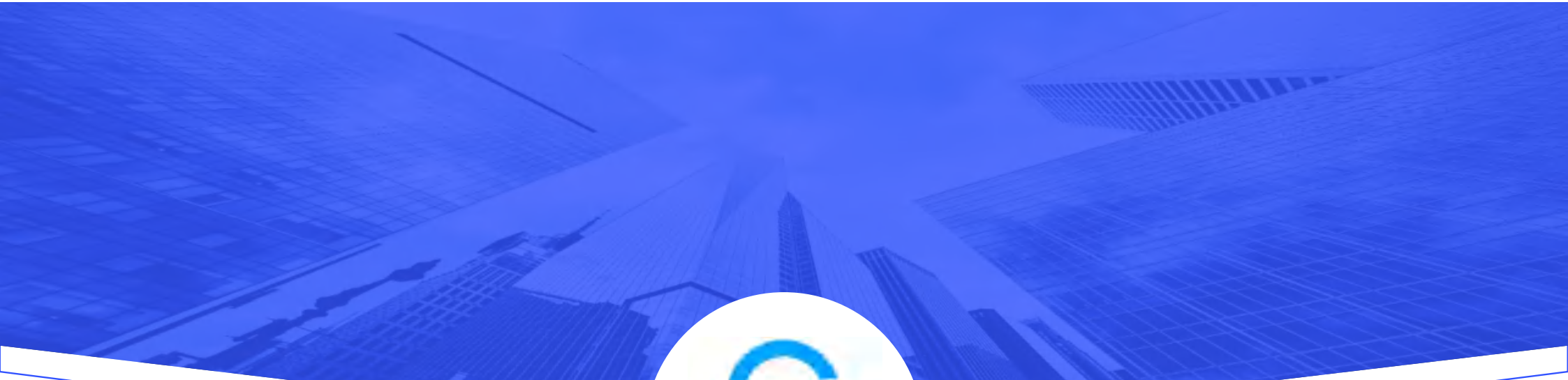


穿越智算奇点

——解锁大模型的无限可能——

(本材料所有文本、插图，均由AI自动生成)

腾讯云 秦若毅

目录 CONTENT

> 他来了！ 他来了！
> 他脚踏祥云走来了！

大模型的基础概念

你的新伙伴已经上线

2013年电影《HER（触不到的她）》

大模型有多“大”？

ChatGPT　Gemini

DALL·E 3　Midjourney

ANTHROPIC　Meta AI

stability ai　MISTRAL AI_

商汤日日新 sensenova　ChatGLM

腾讯混元　通义大模型

360智脑　文心一言

Kimi 智能助手　百川智能

- 参数规模大：大于 1,000,000,000,000个参数
- 数据需求大：大于 12个牛津图书馆
- 资源消耗大：单次训练需耗费1.5亿人民币

“你从哪里来？我的朋友”：大模型的历史脉络

1950 人工智能的起源：

- 人工智能概念最早可追溯到20世纪40年代和50年代
  1950年，阿兰·图灵提出图灵测试

1956 早期的探索：

- 1956年的达特茅斯会议，“人工智能”被正式提出
  早期的AI研究集中在基于规则的系统和专家系统

1990 神经网络的兴起与衰退：

- 20世纪60年代至80年代，神经网络开始受到关注
  由于算力和数据的限制，神经网络在90年代初遭遇低谷

2006 机器学习的复兴：

- 21世纪，机器学习开始迅速发展
  2006年，杰弗里·辛顿等人提出“深度学习”概念

2012 大模型的诞生：

- 2012年，AlexNet实现图像识别领域的突破，
  随后，大模型开始在各个领域展现出其强大的能力。

2017 Transformer（转换器）架构的革命：

- 2017年，谷歌提出Transformer架构
  Transformer架构为后续大模型的发展提供关键支撑

2022 预训练模型的兴起：

- GPT等预训练模型出现，在语言任务领域取得突破
  预训练模型通过大量文本数据训练，能够更好地理解和生成自然语言

NOW 多模态模型的探索：

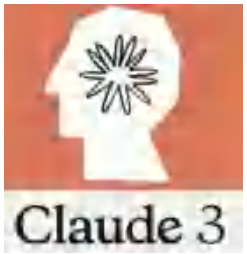
- 探索将文本、图像、声音等多种模态数据融合的大模型
  多模态模型为创意、娱乐、教育等领域提供新的可能性

“千万里，我追寻着你”：国际大模型的创新浪潮

Meta

Llama 3

- Meta于2024年4月发布
- 开源，应用非常广泛



Claude 3.5

- Anthropic于2024年6月发布
- 多模态能力、非英语语言处理

GPT-4o

- OpenAI的最新多模态模型
- 更高的参数量，当前最强大模型

Midjourney

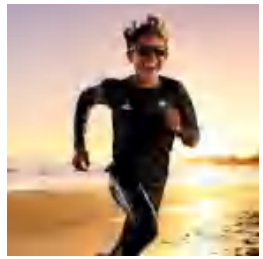
- 专注于图像生成的AI模型
- 艺术创作、设计领域

Dall-E 3

- OpenAI开发，文生图的先驱
- 图像细节、多样性表现

Stable Diffusion

- 开源的文生图大模型
- 开源，可控生成、模型多样性

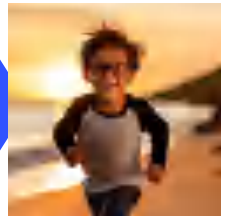


“中国欢迎你，为你开天辟地”：国内大模型的极速崛起

01

腾讯混元：多模态智能的领跑者

- 2024年5月，发布元宝。开源中文文生图模型
- 应用：游戏、社交、广告等场景



04

月之暗面Kimi：长文理解的先驱

- 2024年3月，开放支持200万字超长文本
- 应用：长文本理解、智能写作等

02

阿里通义千问：电商智能的新篇章

- 2024年6月，开源QWEN2
- 应用：电商与营销领域的智能化

05

智谱清言：分析与编程的高手

- 2024年6月，开源GLM-4-9B
- 应用：代码生成、Excel分析编写

03

百度文心一言：探索知识的深海

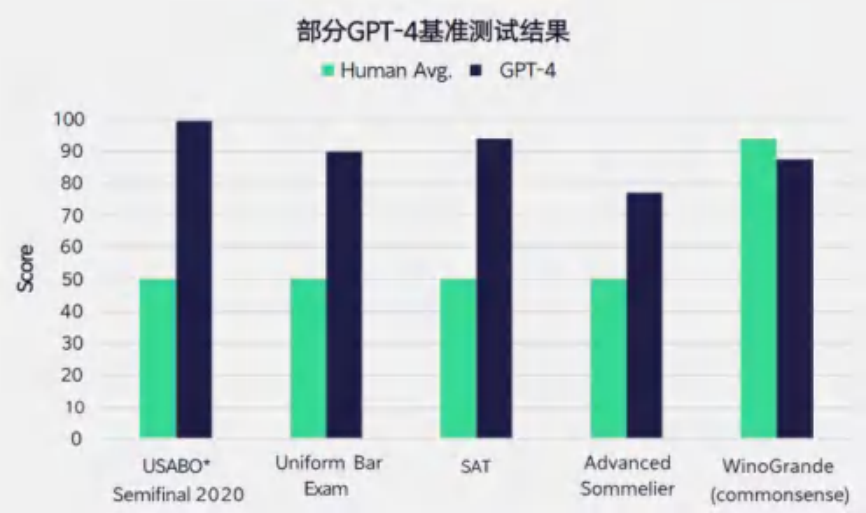
- 2023年10月，发布文心一言4.0
- 应用：在线搜索、知识问答、学术研究

01

百川百小应：会提问的小助手

- 2024年5月，发布“百小应”APP
- 应用：通过主动追问，引导用户提出需求

"你是电，你是光，你是唯一的神话"



GPT4的考试成绩远超人类均值

- **美国生物奥林匹克竞赛**：超过99%考生
  美国律师资格考试：超过90%考生
- **美国高考**：1410/1600（写作+数学）
- ……

AI生成作品在人类艺术竞赛获奖

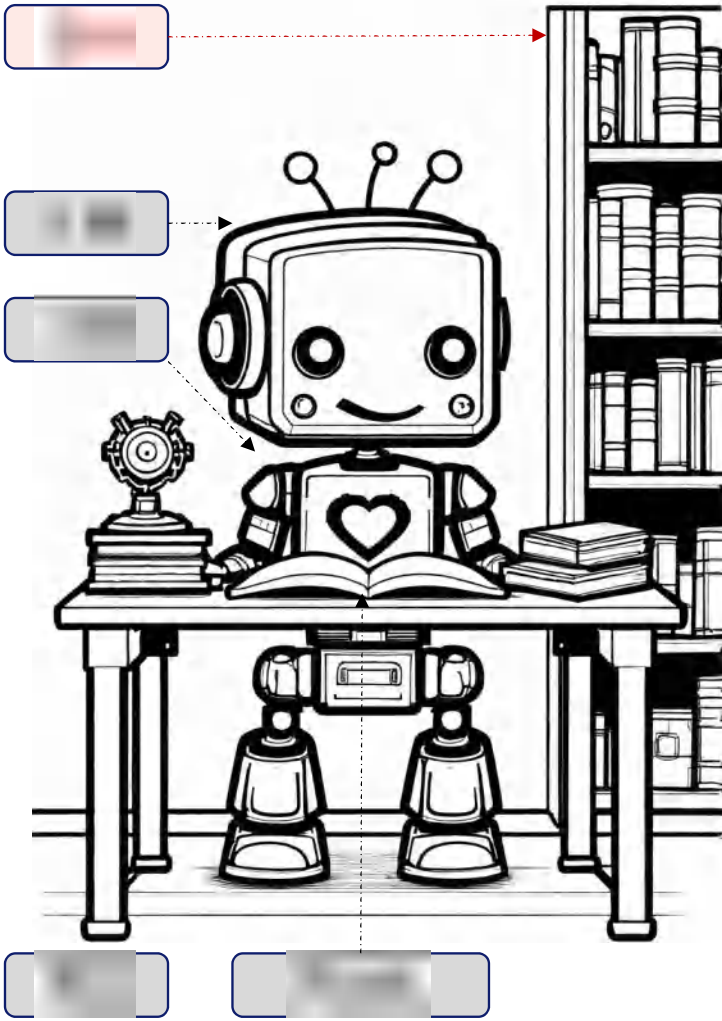
- 2022.10 美国科罗拉多州年度艺术竞赛
  Midjourney生成的作品获得蓝丝带奖

刷题不刷万亿遍，哪敢自称学霸？

大模型的关键技术

学霸的书架：海量训练数据

GPT4共学习13万亿token的训练数据，类似于读完5200万本《唐诗三百首》



学霸的大脑：基底大模型（Foundation Model）

基底大模型的定义：智能的基石

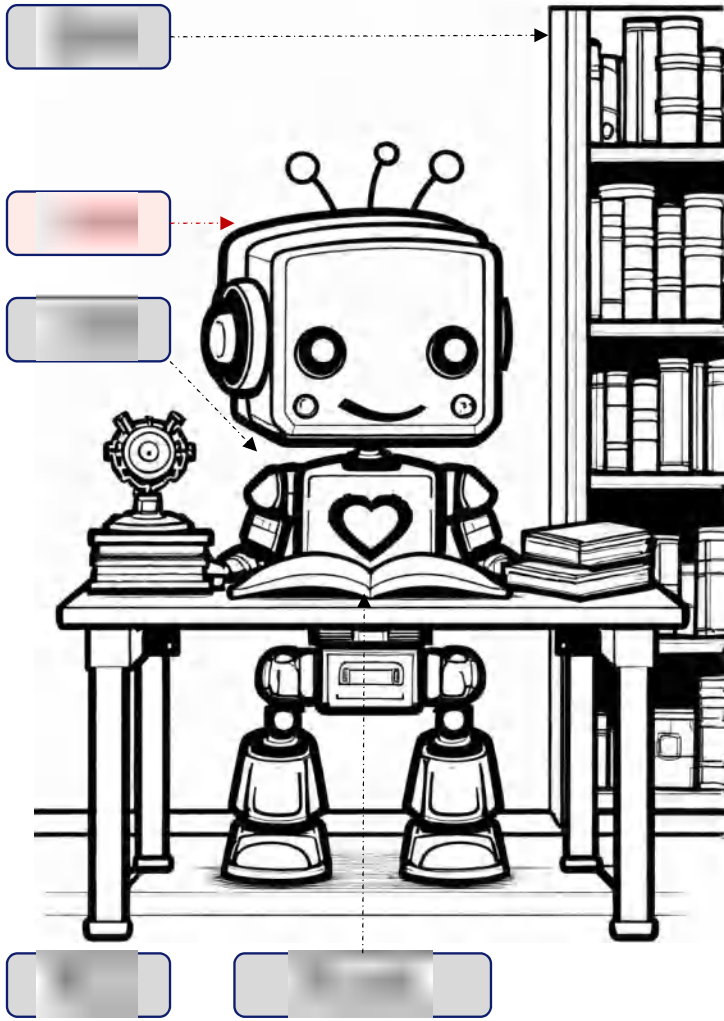
- 通常具有数十亿甚至数万亿的参数
- 基底大模型为智能系统提供基础能力

模型规模与性能：规模的力量

- 基底大模型的性能与规模密切相关
- 当前最大规模：GPT4，1.8万亿参数

模型压缩：效率与性能的平衡

- 压缩模型，以减少模型的算力需求
- 技术：知识蒸馏、迁移、剪枝



学霸的心脏：智算底座

高性能集群

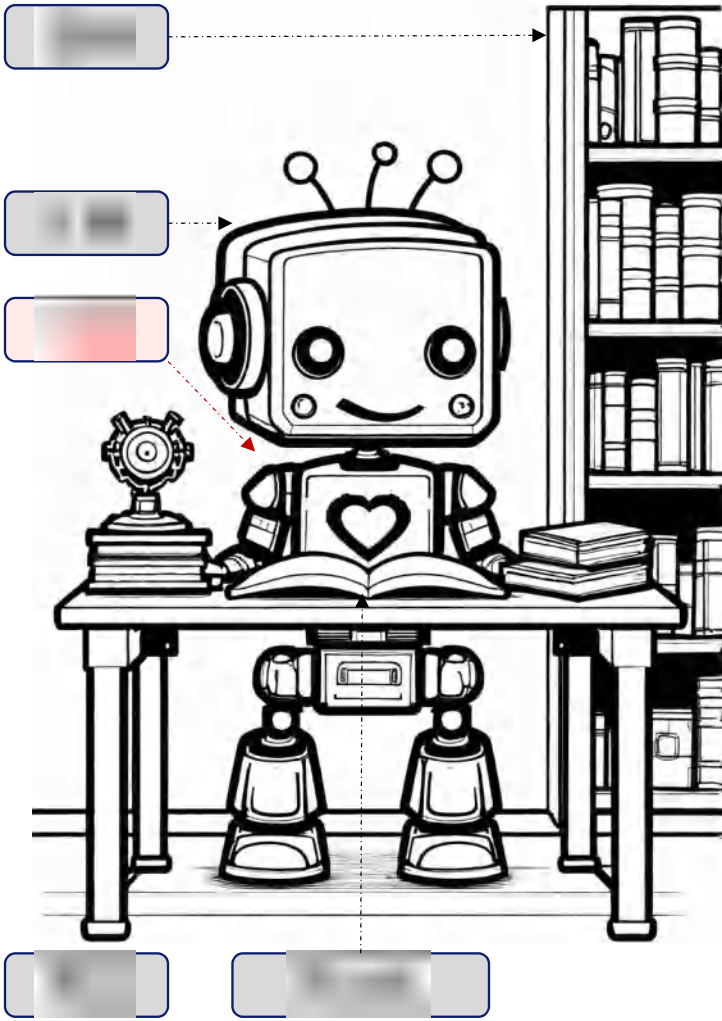
- GPU组成的大规模并行计算集群
- 要点：并行处理、高速互联、资源调度

高性能网络

- 0.1%的掉包，会损失50%的算力
- 要点：低时延、高可靠

高性能存储

- 读写数据时，GPU资源处于等待状态
- 要点：快速读写，高吞吐、冷热分层



学霸的书桌：训练平台

资源优化与训练效率提升

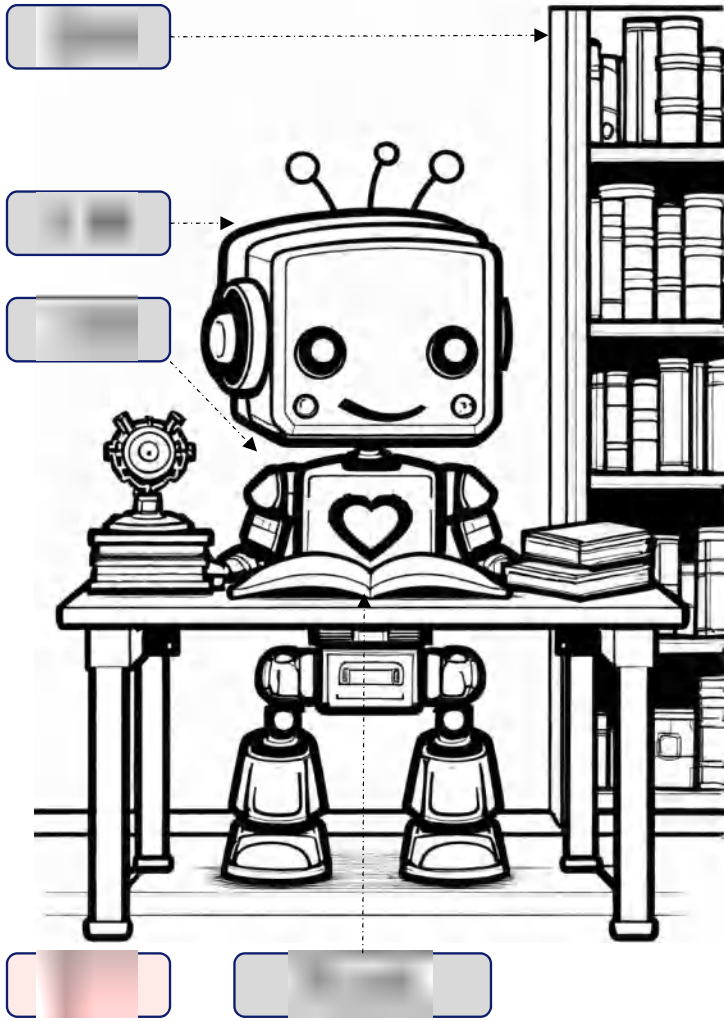
- 简单堆机器，会导致分布式性能反降
- 优化算法，提高资源利用率，提升推理/训练速度

一站式服务的训练平台

- 功能：模型选型、训练共建、部署应用
- 训练方法：监督训练、人工反馈、强化学习

数据管理与模型版本控制

- 训练数据的准备、知识挖掘与增强、数据清洗与切块
- 向量库建设与维护、模型版本控制



学霸的作业本：模型优化

模型精调：领域专家的塑造者

- 技术：针对特定任务或数据集进一步训练
- 目的：学习特定领域知识，提升专业任务表现

  应用举例：医疗行业，更好地理解电子健康记录

提示工程：智能引导的艺术

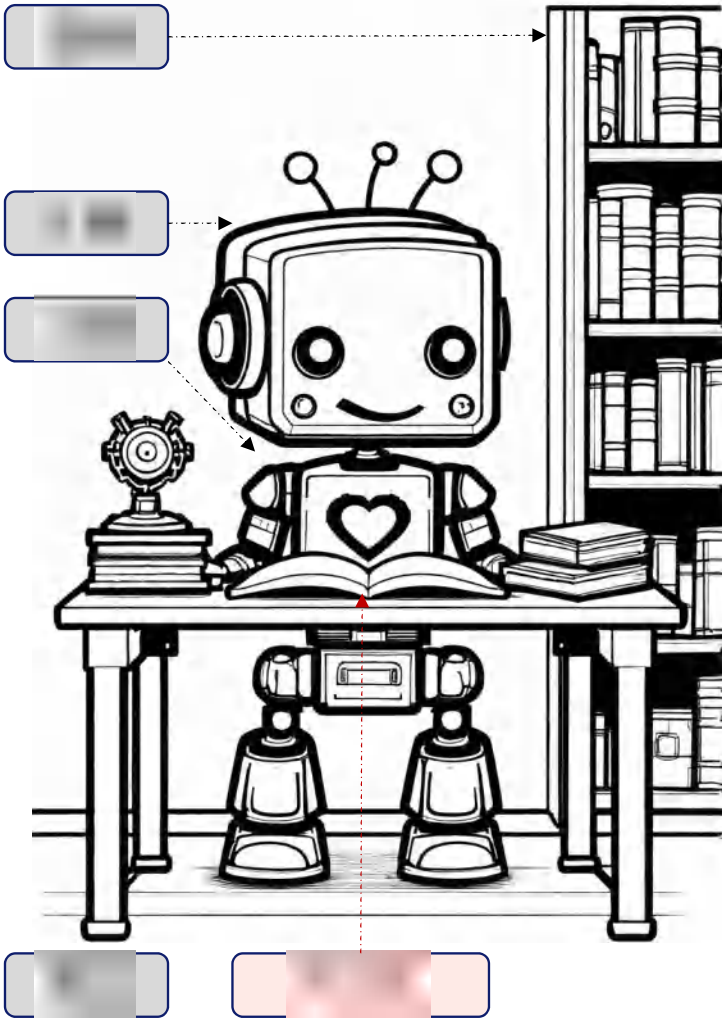
- 技术：设计精确的输入提示来引导模型
- 目的：在没有大量训练数据的情况下解决特定任务

  应用举例：教育行业，快速生成定制化学习材料

检索增强：获得实时的海量信息

- 技术：检索外部信息源，提供辅助信息
- 目的：提供更加精准，更加及时的信息

  应用举例：金融分析领域的智能投研助手



你是我的阿拉丁神灯吗？

大模型的典型应用

通才或者专才：通用大模型与行业大模型

通用大模型

- 概念：捕捉通用特征，处理多种通用任务
- 构建方式：侧重于广泛的数据覆盖

  训练目标：良好的泛化能力
- 场景举例：搜索引擎、语言翻译、内容推荐

VS

行业大模型

- 概念：针对特定行业定制，基于通用大模型精调
- 构建方式：侧重于特定领域数据的深度和质量

  训练目标：特定任务的性能
- 场景举例：医疗诊断、法律分析、金融风险评估

个性化智能：面向个人的大模型应用

辅助创作

- 文案创作
- 音乐创作
- 图像生成

个性化教育

- 语言学习
- 编程学习
- 知识问答

虚拟助理

- 日程管理
- 信息查询
- 智能家居控制

娱乐互动

- 虚拟陪伴
- 游戏角色
- 虚拟偶像

个性化智能：全球C端大模型应用概览

	大模型赋能	大模型原生	市场已有产品数量 >2300
搜索	Google Search, Bai百度, New Bing	Perplexity	>150
会议	Microsoft Teams, 腾讯会议		>20
文档	腾讯文档, Office, WPS		>450
陪伴	DISCORD	character.ai, Inflection Pi, Replika, MINIMAX	>150
游戏	王者荣耀, 网易 逆水寒, 米哈游 崩坏3	AI DUNGEON, inworld	>50
个人助理	Windows Copilot	Rewind, ADEPT	>1000
学习	科大讯飞 IFLYTEK, duolingo		>150

来源：腾讯研究院

浅窥智能生活：腾讯的C端大模型应用案例

腾讯会议：自动会议纪要

腾讯云智能 AI写真

腾讯文档：智能起草

天涯明月刀：NPC

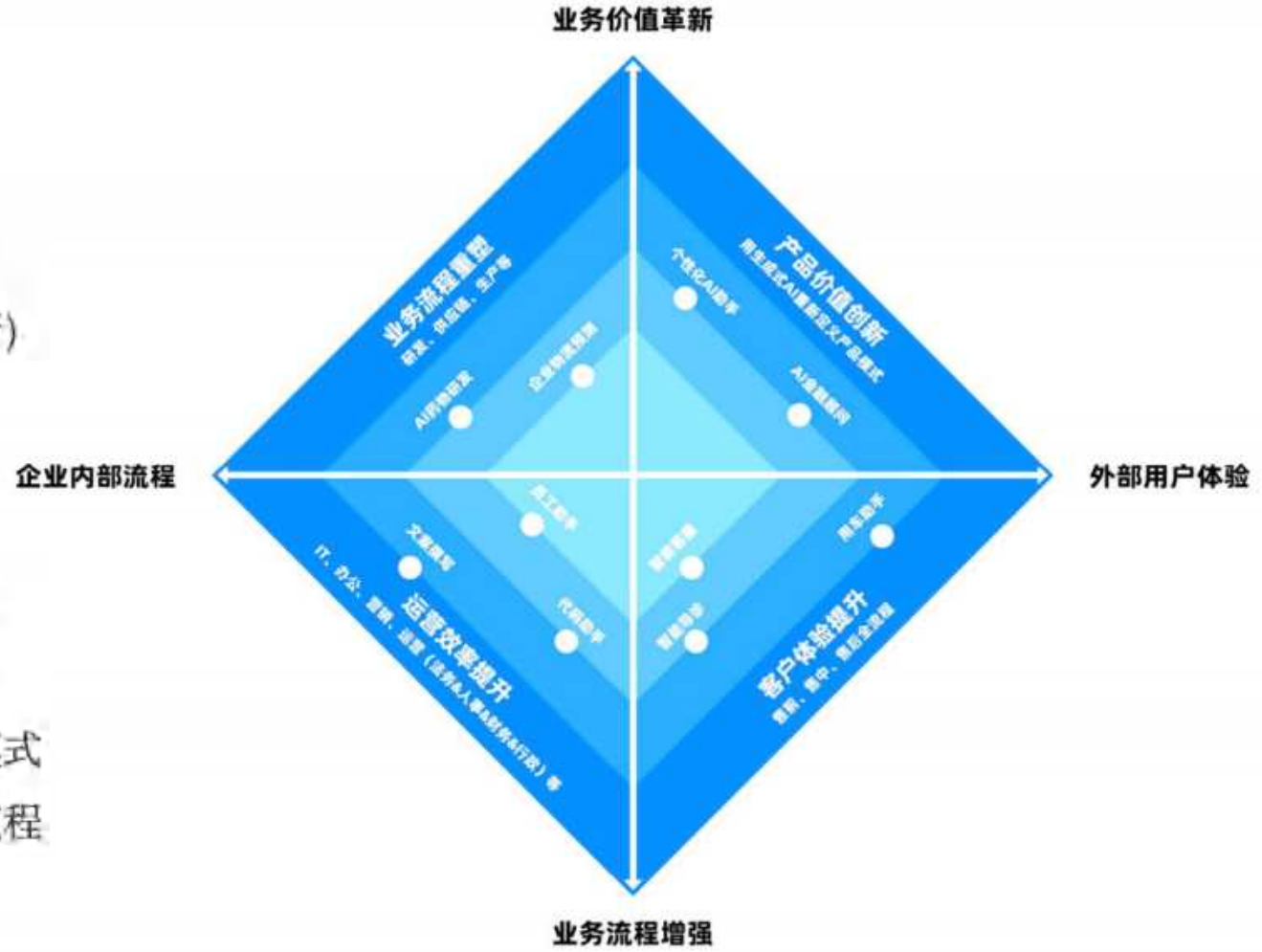
企业智能化：面向企业的大模型应用

生成式AI应用场景矩阵

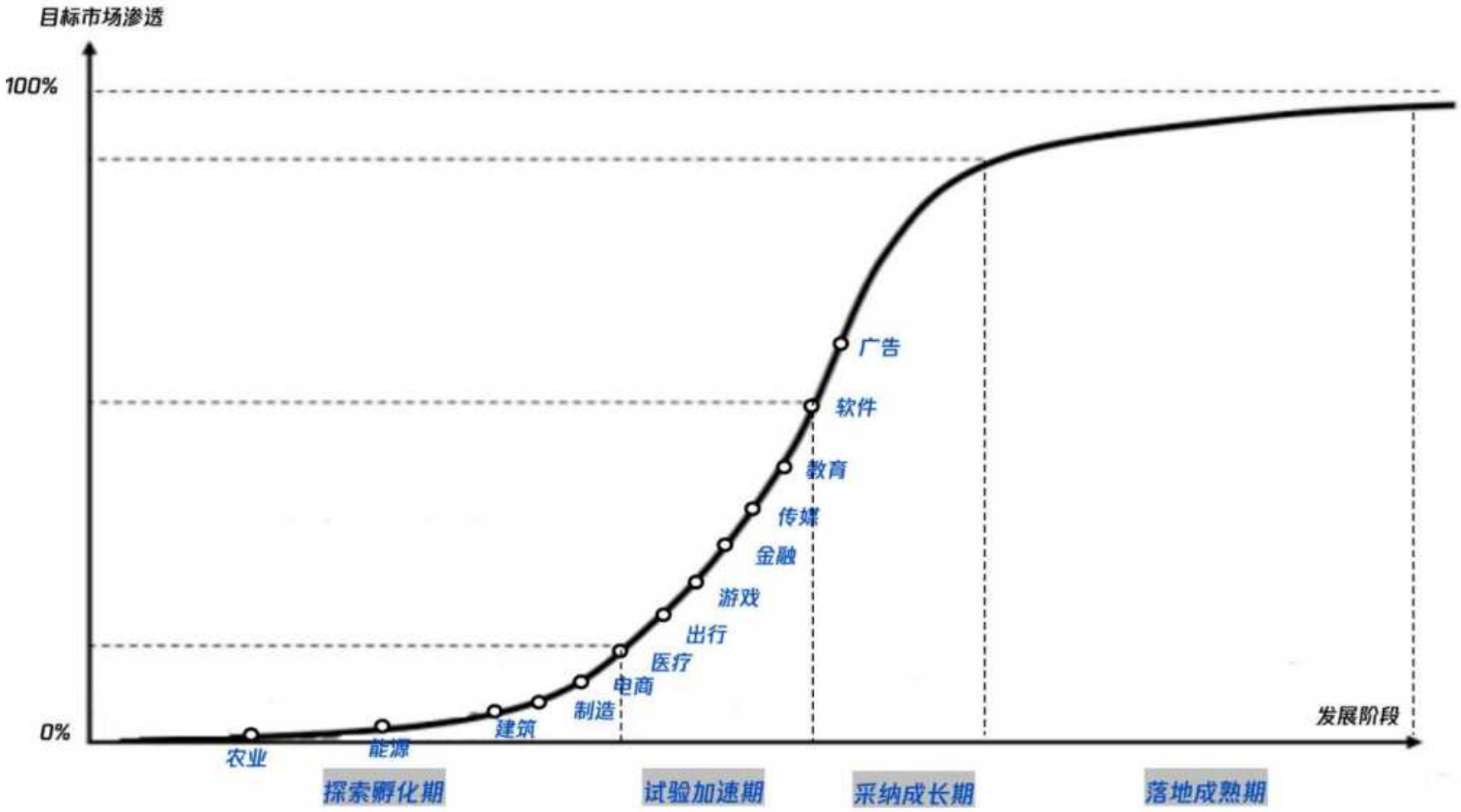
- **横轴**：企业业务流程（内部<=>外部）
  纵轴：业务改造程度分类（改善 <=> 创新）

四大场景价值象限详解

- 运营效率提升象限：内部用户/降本增效
  客户体验提升象限：外部用户/提升体验
  产品价值创新象限：外部用户/颠覆服务模式
  业务流程重塑象限：内部用户/重构内部流程



企业智能化：大模型技术的行业渗透情况



来源：腾讯研究院，2024年初调研

企业智能化：AI百大应用场景

来源：腾讯云《生成式AI产业落地路径研究报告》

运营效率提升：以金融行业落地案例为例

生成式AI应用场景渗透金融业务流程

企业内部流程	生成式AI应用场景
IT	代码助手
办公	办公管理
营销	个性化营销
运营(法务&人事&财务&行政)	数据分析

IT业务流程优化

- 高信息化程度的金融机构
- 生成式AI辅助代码补全
- 加速APP研发迭代速度

办公业务流程简化

- 会议纪要的自动生成
- 理解并准确生成专业术语
- 从流程中释放员工

营销业务流程赋能

- 客户经理面对的产品知识库
- 生成式AI提供快速问答
- 节省学习理解产品的时间

运营业务流程自动化

- 定期向监管机构报送数据
- 辅助获取分析数据与生成报告
- 对话式交互快速定位异常原因

客户体验提升：以医疗行业落地案例为例：

生成式AI应用场景渗透医疗服务流程

外部用户体验	生成式AI应用场景
诊前（售前）	智能导诊
诊中（售中）	检查报告自动生成
诊后（售后）	智能随访复诊

诊前环节优化

- 解决挂错诊室、医生不匹配等问题
  生成式AI机器人协助患者了解就诊科室
  完成预约挂号与精准医生推荐

诊中环节辅助

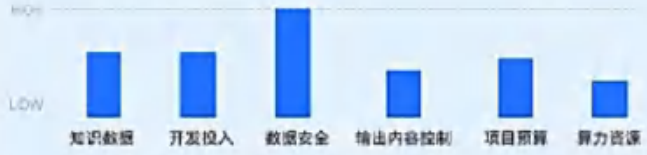
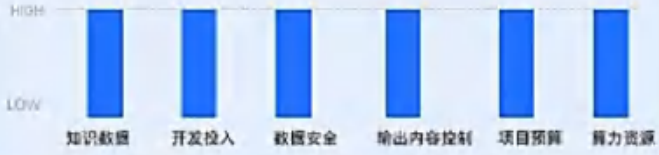
- 患者面对检查报告的困惑
- 生成式AI提供医疗辅助诊断报告
- 提升医生检查效率

诊后环节关怀

- 弥补诊后与医生的交互空缺
- 生成式AI随时解答患者用药疑惑
- 提醒用药，持续关心患者健康

大模型应用落地的三大路线

购买 → 自建

	1. 标准软件	2. 标准模型能力增强	3. 定制化模型精调训练
	基于生成式AI模型能力增强的开箱即用的应用软件	基于标准生成式AI模型，通过RAG等提示工程增强模型能力	加入企业专属数据精调训练，自建企业专属大模型，沉淀模型资产
评估维度			
合作分工	服务商托管：应用开发；模型精调训练；提示工程与检索增强生成；基础模型构建	企业参与：应用开发；提示工程与检索增强生成 服务商托管：模型精调训练；基础模型构建	企业参与：应用开发；提示工程与检索增强生成；模型精调训练 服务商托管：基础模型构建
腾讯云产品方案	• 智能增强软件 • GenAI助手	• 模型服务 • 数据平台 • 应用研发 • AI开发平台 • 云原生平台	• 数据平台 • AI开发平台 • 硬件创新 • 云原生平台 • 应用研发 • 云原生平台 • 软硬加速 • 分布式AI算力

混元大模型、行业大模型、开源模型

全栈安全防护

标准软件：腾讯云AI代码助手

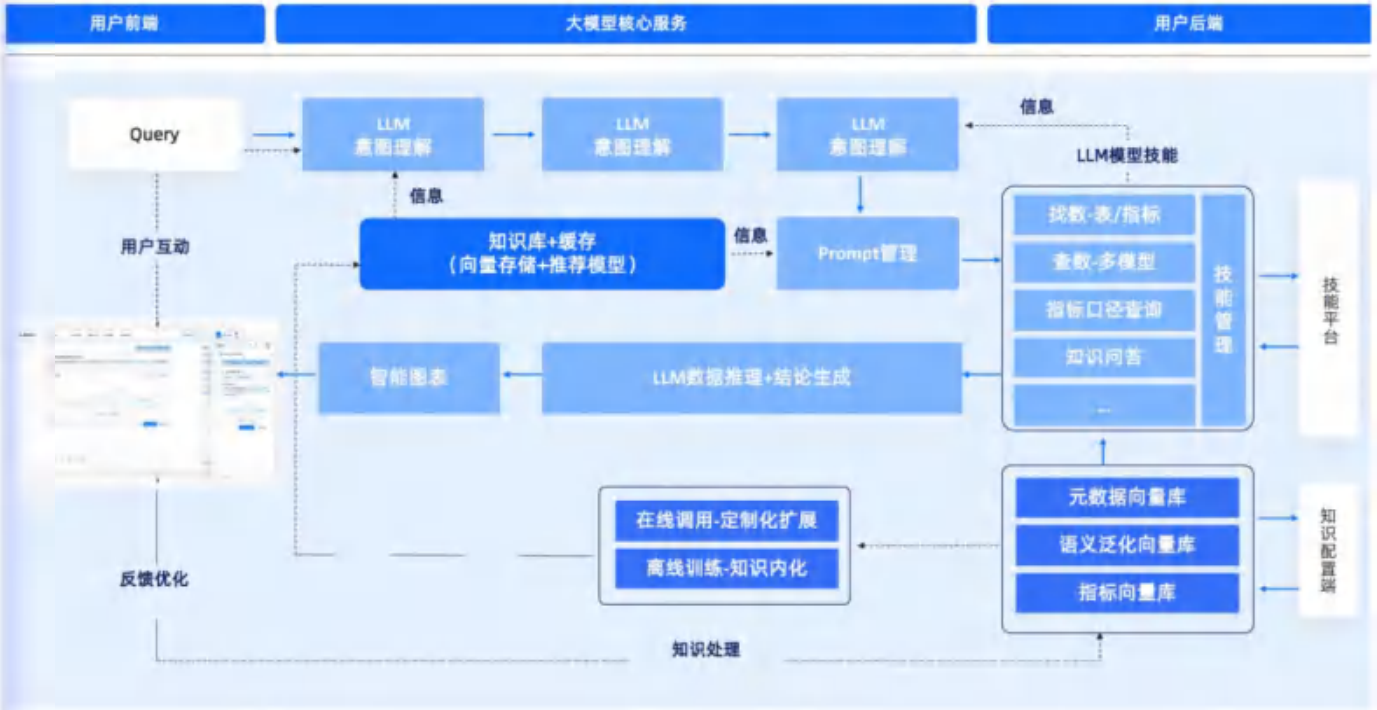
标准软件：腾讯企点营销云AI助手

场景挑战

- 业务与数据分析技能的脱节
- 海量数据处理效率低下
- 数据洞察依赖主观经验，缺乏严谨性

方案价值

- 专业：行业知识微调的大模型
  高效：智能归因自动识别指标异动
  易用：对话式交互，业务人员轻松上手
  可信：复杂问题的模型推理步骤可视化



标准软件：腾讯乐享AI助手

场景挑战

- 高门槛的内容生产与低效的内容流转
- 长链路的内容获取与低利用率
- 知识管理成本高昂，效果有限

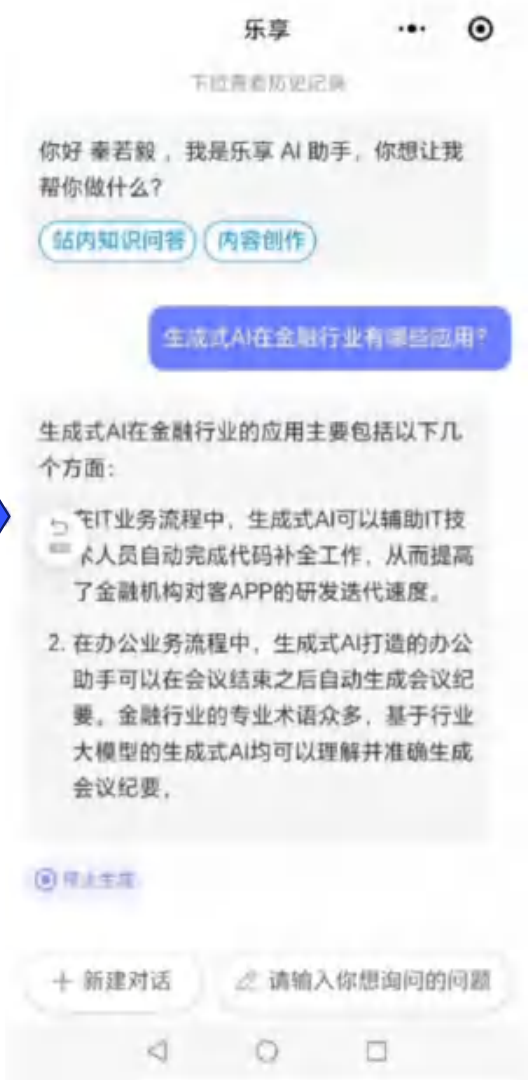
方案价值

- 智能问答：理解并自然语言回复
- 智能写作：提供适配企业业务的写作大模型
- 智能生成：理解业务知识并按规则出题

小彩蛋——送给听众的礼物

专门解答智算与大模型问题的AI助手

(扫码可获得使用权限)



标准模型能力增强：腾讯云大模型知识引擎

方案价值

- 多模态复杂知识处理：表格问答，复杂公式，图文关联
- 解决复杂文档问题：文档解析，切分，检索，推理，生成
- 多模态交互模式：支持多种输入，输出格式

大模型客服机器人的优势

- 低使用门槛：开箱可用的应用模板，可集成的原子能力API
- 完善的工具链：模型选择，角色设定，提示词优化，知识库管理
- 一站式流程：对话测试，修正，发布，反馈增强的一站式处理

定制化模型精调： 腾讯云行业大模型

腾讯云行业大模型概览

- 多个行业大模型：金融、医学、汽车、教育、文旅等
- 超50个行业大模型解决方案
- 助力客户构建专属大模型及智能应用

算法层面的创新

- 基于混元技术能力，结合垂直行业业务场景
- 加入大量垂直行业专业数据进行精调训练
- 数据安全和生成内容控制能力得到提升

定制化模型精调： 腾讯云行业大模型——医学大模型

定制化模型精调：腾讯云TI平台

业务挑战

- 大规模并行训练的高稳定性要求
- 计算资源的稀缺性和高效调度需求

方案价值

- 完善的大规模训练工具平台
  算力、存储、计算多重优化加速

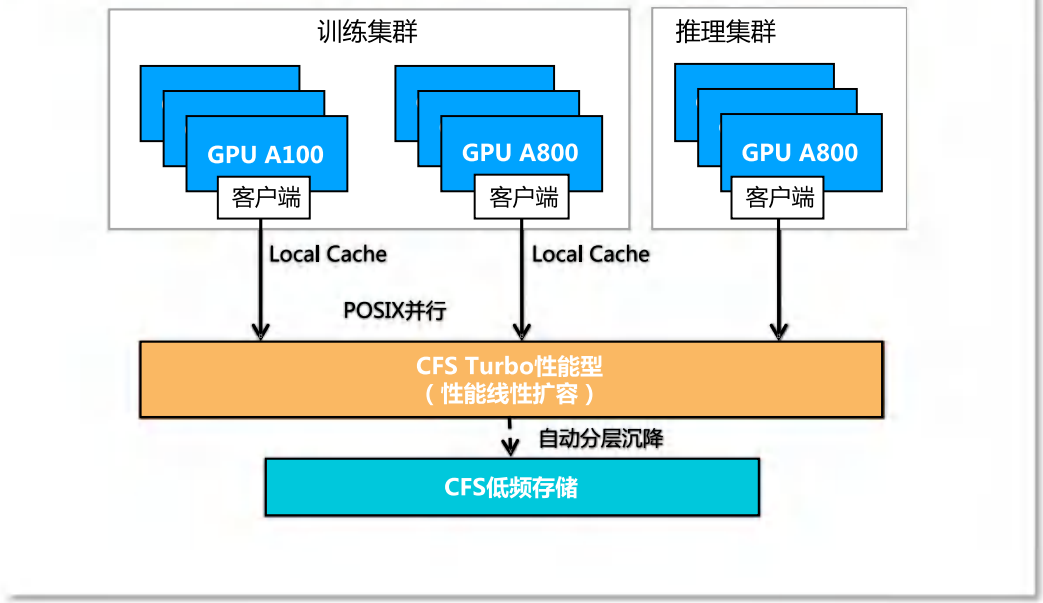
定制化模型精调：腾讯云高性能存储

快：超大带宽存储
省：自动分层存储

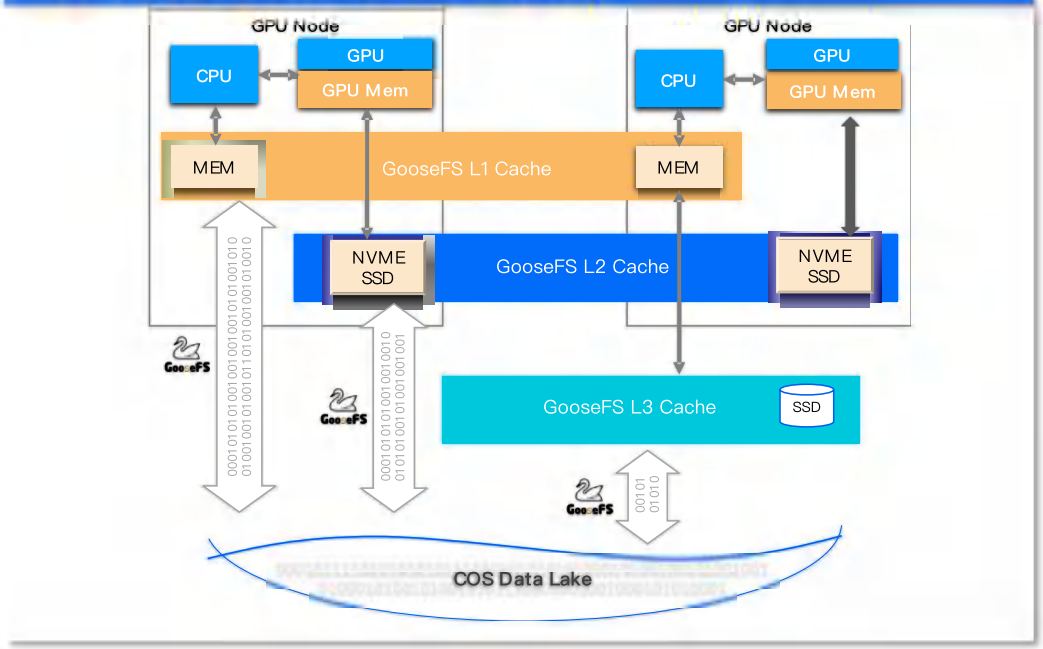
数据集相对固定

数据集频繁更新

腾讯云Turbo CFS文件存储方案



腾讯云COS+GooseFS缓存加速方案

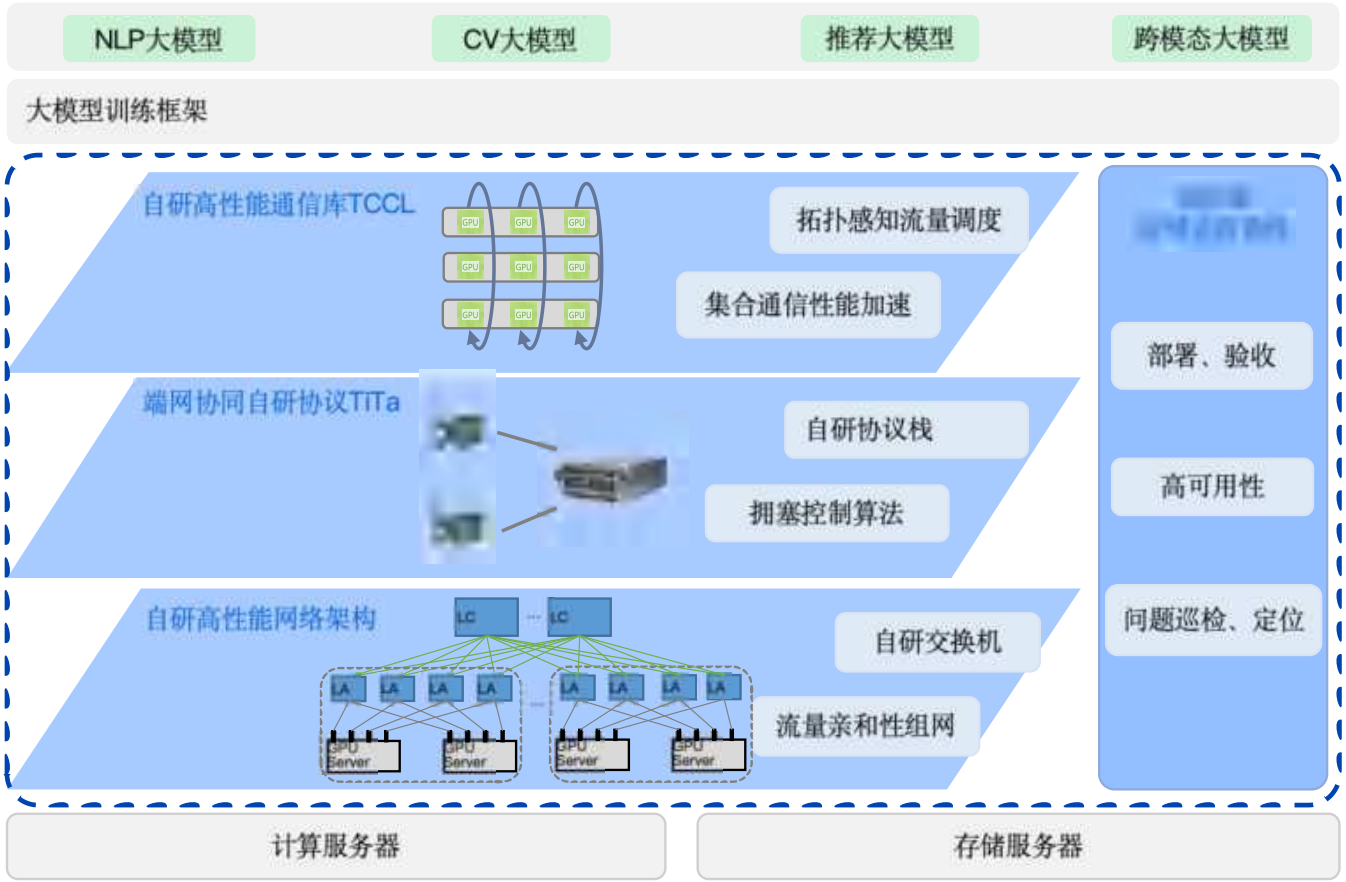


定制化模型精调： 腾讯云星脉RDMA网络

自研高性能通信库TCCL
——业务流量的“导航仪”

自研端网协同协议TiTa
——以太网车道的“交警”

自研高性能网络架构
——3.2Tb的“超宽车道”

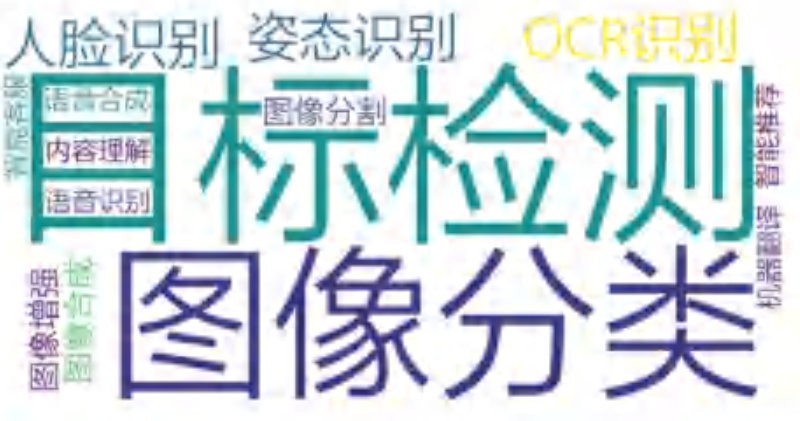


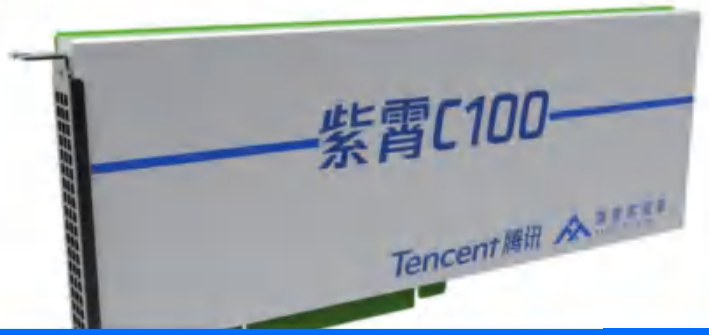
定制化模型精调： 面向视觉场景的紫霄推理卡

紫霄V1：高性价比的国产推理卡

- 设计算力规格与A10相当
- 显存带宽比A10高30%
- 在深度优化场景，最高可以做到A10的1.5-2倍性能。

适合多种视觉大模型推理场景



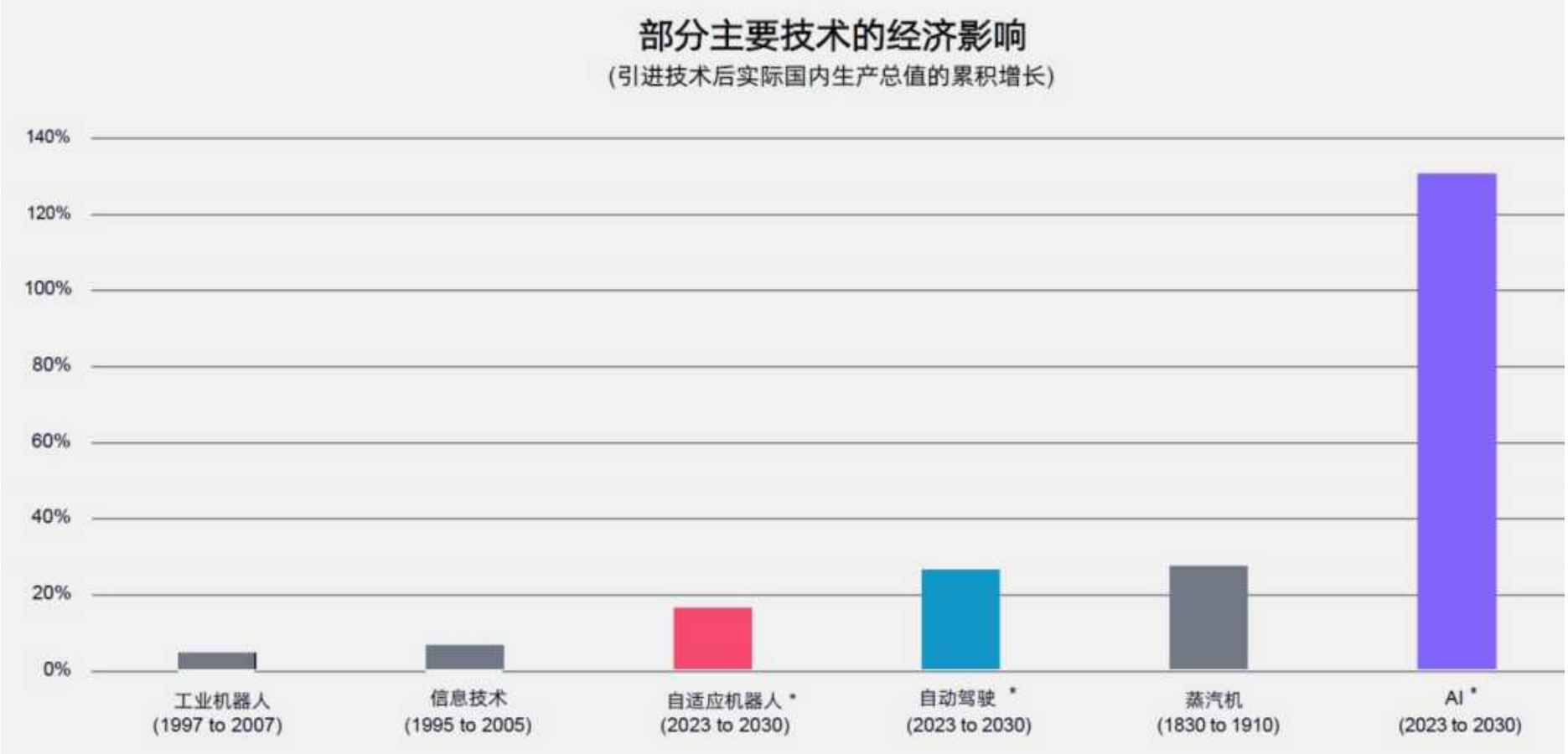


场景分类		典型模型	场景核心需求	推荐用卡	
				国外方案	国产方案
大模型	小模型 （7B以下）	SD 1.5	性价比	A10	**紫霄 v1**
	中等模型 （ 7B ~ 65B）	llama-7b llama-13b llama-30b llama-65b	较高显存 高算力 性价比	L40S	**紫霄V2**
	大模型 （65B以上）	gpt-89b bloomz-176b	高显存 高算力 性价比	A800 H800	**紫霄V2**
传统 AI	计算机视觉（CV）	resnet50	性价比	A10	**紫霄 v1**
	自然语言处理（NLP）	BERT			
	光学字符识别/语音识别 （OCR/ASR）	Deep Speech 2			
搜索/广告/推荐系统		wide&deep	大数据量 大内存 高内存带宽 高卡间互联带宽	L20	**紫霄 v2**

大模型和我，或许都有美好的未来

大模型的未来发展

大模型对未来经济的影响，可能会是颠覆性的



来源：ARK基金《BIG IDEA 2024》

国内大模型未来发展的挑战

01

人才挑战：AI人才集中在美国

- 全球TOP 2000顶级学者，美国占57%，谷歌/微软/Meta占17%
  全球60%的AI人才在美国工作，其中三分之二来自国外，主要是中印

02

算力挑战：智算需求爆发式增长

- GPT-5算力需求预计是GPT-4的9倍以上
- 美国禁令对国内企业的算力需求产生较大影响
  目前国产训练芯片仅有昇腾，且受限于产能和软件生态

03

数据挑战：缺乏高质量数据集

- 中文语料数据集占比低于5%，多元化不足，缺乏专用领域开源数据
  数据荒正在到来，全球高质量语言数据预计在三年内耗尽

强大的AI智能体：大模型+工作流

“基于GPT-3.5构建的智能体工作流在实际应用中表现优于GPT-4。”
“AI智能体将在2024年推动人工智能领域的大规模进步。”

——AI权威学者 吴恩达

AI智能体是什么？

- AI智能体是能够感知环境、做出决策并采取行动的系统
- AI智能体由感知系统、决策系统、行动系统和学习系统构成

AI智能体平台是什么？

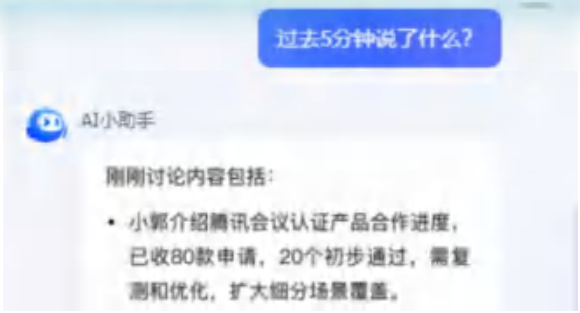
- 提供完整的工具和框架，用于便捷开发和管理AI智能体

字节跳动的"扣子"、腾讯的"元器"都是AI智能体开放平台

实例：如何快速开发一个"儿童采访教师"AI智能体？

1）儿童通过微信语音向智能体提问（感知）。
2）智能体将语音转成文字，并理解问题（行动）。
3）智能体通过网络搜索和知识库获取相关知识（学习）。
4）智能体理解知识，并以生动的儿童化文字输出知识（决策）
5）智能体将文字转化为语音，回答儿童的问题（行动）。

案例：腾讯会议AI小助手



影像创造的革新：文生视频大模型

- OpenAI发布Sora，可生成一分钟视频
  中国：Vidu，OpenSora，可灵，YiSu，……

- 将文本描述转化为视频内容
  关键技术：文本理解、场景构建、视觉生成、时间序列预测、多模态融合

- 传媒和广告行业：自动化和个性化的广告
  娱乐和动画产业：快速生成电影与角色

Sora大模型的文生视频样例

物理世界的具身智能：大模型+机器人

腾讯用生成式模型训练的机器狗

斯坦福大学的做饭机器人

生成式模型训练机器人

- 摒弃规则，完全从数据中学习

动捕系统采集动作轨迹，作为训练语料

从"自动化"到"智能化"

- 自动化：重复做预编程的动作

智能化：根据环境调整、规划行动

超越今日的边界：通用人工智能（AGI）

通用人工智能（AGI）的定义

- AGI具有类似人类的思维
  AGI能够理解、学习和应用知识于广泛领域
- AGI具有自我意识、情感智能和道德判断

人类应该如何迎接AGI的到来？（机械公敌，2004）

腾讯云WeTele公众号

欢迎探讨 期待交流

腾讯云 秦若毅

2024.6