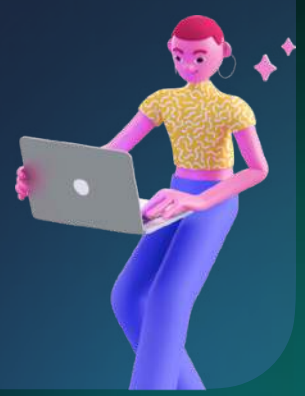
智 能 混 合 云 集 成 商

xRunda.AI

xRunda AI Solution Architecture

北京一润一达科技有限公司



A

公司介绍

Company Profile

CLICK TO CONTINUE

一润一达 | 智能混合云集成商

XRUNDA

INTEGRATOR OF INTELLIGENT HYBRID CLOUD SOLUTIONS

智 能 混 合 云 集 成 商

COMPANY

成立于2015年，国家高新企业，追求以深度的行业理解和用户洞察力，
结合新灵智慧的技术服务为每个企业客户定制最匹配的项目解决方案，
并以有感染力的交互创造令人愉悦的体验。

28+
软著认证

50+
跨领域专家团队

500+
产品创新项目

OUR TEAM 北京一润一达科技有限公司

极客工作室团队

Our studio is delicate yet powerful. As a learning organization, we are empowered by self-drive, judgment, analytical abilities, learning capabilities, decision-making, creativity, and execution. We deeply understand emerging markets and technologies, integrating our technical expertise with our rich experience in enterprise services.

我们的工作室精致而强大。作为一个学习型组织，我们拥有自我驱动、判断、分析能力、学习能力、决策、创造力和执行力。我们深入了解新兴市场和技术，将我们的技术专长与我们在企业服务方面的丰富经验相结合。

前沿技术实验室

案例制造机

智能向共创模式

大模型养成

超级进击实验室 xRunda AI Lab

OUR VISION

智变未来 TRANSFORMING THE FUTURE WITH INTELLIGENCE

公司使命
Company Mission

助力企业数智化升级，以科技驱动革新，
让商业变得更智慧。

We are Transforming the Industry with Artistic Intelligence.

团队目标
Team Goals

成为企业数智化转型伙伴

Strategic Partners for Enterprise Digital & Intelligent Transformation.

价值观念
Corporate Values

客户为本

Customer-centric

Prioritizing our customers always.

合作共赢

Collaborative Win-Win

Pursuing partnerships for mutual success.

洞察创新

Insightful Innovation

Creating value through innovative insights.

科技向善

Technology for Good

Steering technology to benefit humanity.

时代机遇
Opportunity

人工智能在许多任务上超越了人类，并且在新任务上超越人类的速度正在加快

State-of-the-art AI performance on benchmarks, relative to human performance

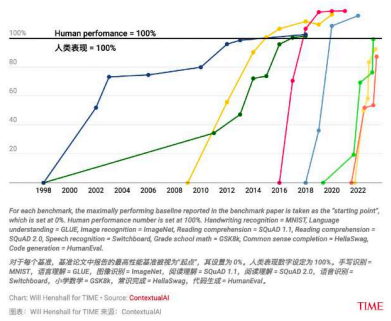
相对于人类表现，在基准测试中具有最先进的人工智能性能

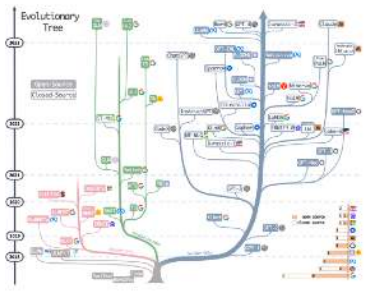
Handwriting recognition · Speech recognition · Image recognition · Reading comprehension

手写识别 语音识别 图像识别 阅读理解

Language understanding · Common sense completion · Grade school math · Code generation

语言理解 常识完成 小学数学 代码生成





PARTNERSHIPS

生态伙伴共成长 GROWING TOGETHER WITH ECOSYSTEM PARTNERS

汽车之家 autohome.com.cn
MINI
抖音支付 更快 更省
东方卫视
快手
友邦保险
Tencent 腾讯
PICC
中国工商银行 INDUSTRIAL AND COMMERCIAL BANK OF CHINA
西瓜视频
万达集团 WANDA GROUP
北京师范大学 BEIJING NORMAL UNIVERSITY
中国邮政 CHINA POST
中软 CS&S
喜百年装饰
JUSHU 聚朗生物
HUAWEI
Haier
adidas
平安银行
搜狐 SOHU.com
yes auto
Microsoft
China unicom中国联通
高德地图 amap.com
smart
Coca-Cola
牛咖斯NCARS 爱车信任牛咖斯
Sciverse.

B

解决方案

Solution

CLICK TO CONTINUE

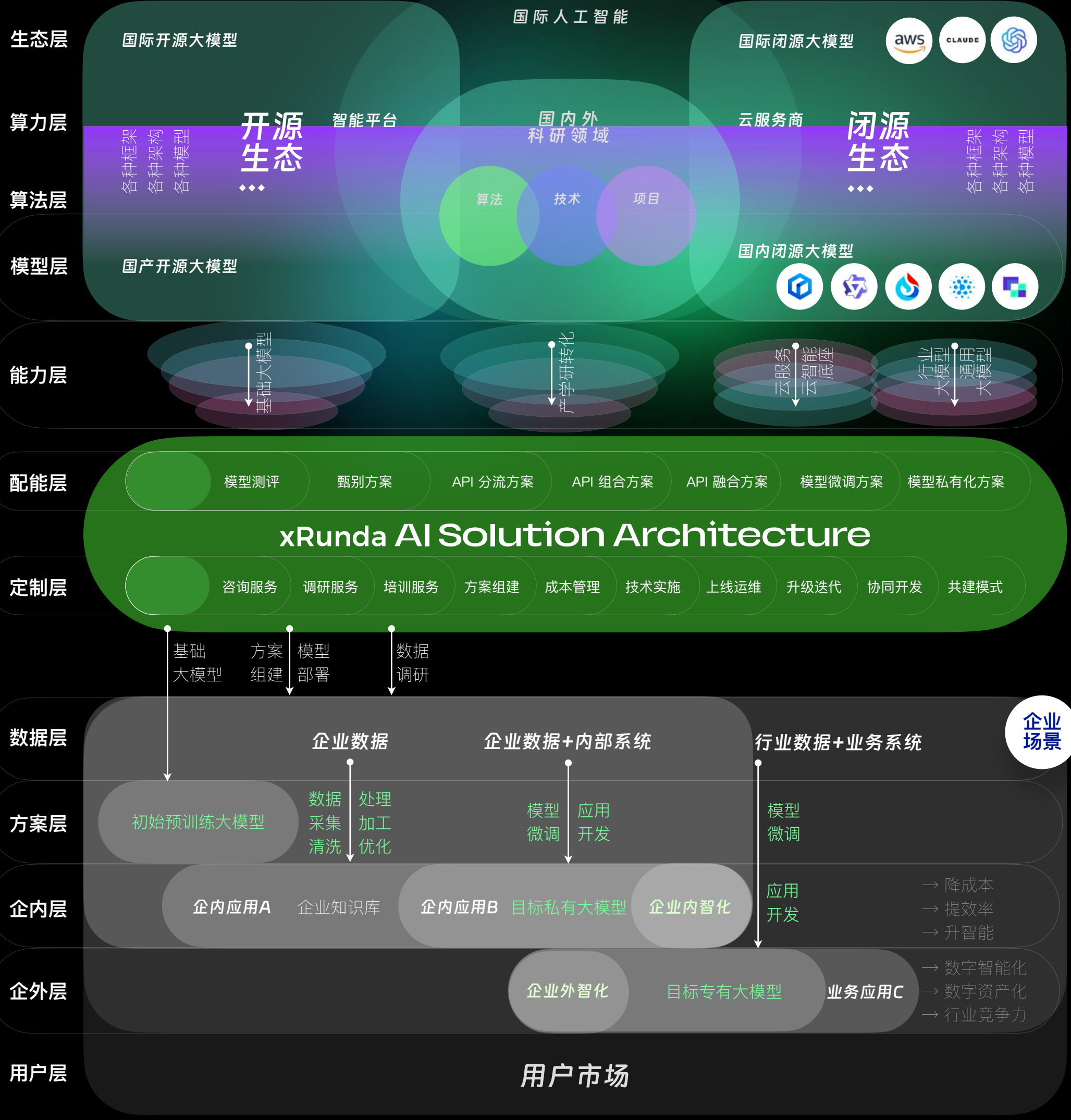
B

Solution 解决方案

新智能资源生态系统

MaaS IaaES Industry Ecosystem

模型即服务 智能即能源



CLICK TO CONTINUE

xRunda AI Solution Architecture

xRundaAI解决方案框架

咨询顾问服务

Consultant

	模型场景	办公	金融	生活	娱乐	自动驾驶	智慧城市	商业	医疗	工业	教育
通用类	ERNIE 3.0	语言理解与生成	保险合同条款智能解析、金融风控、智能对话	智能对话			市长热线工单分类、投诉工单信息抽取		医学病例抽取		知识挖掘、词条管理、版权保护
	M6	文案生成、服务搜索	金融信息问答和推荐	语义对话、事件点评		汽车外观设计		商品图片生成、内容封面设计、图文商品检索			
	紫东太初				影视创作	智能驾驶			手术机器人	质量检测	
	华为云盘古		金融OCR识别				智慧物流		药物开发	煤矿安全、异物监控、铁路轨道机车缺陷、电力巡检	
	文心一言	内容创作、数据分析	贷前审核及尽职调查、贷后风险监测与预警、生态内企业风险管理		智能对话		突发事件预警监控				
	通义千问	写报告,SWOT分析、商品描述生成		对话机器人、写菜谱、写作文	彩虹屁专家、写情书、为你写诗						
	悟道2.0	图像翻译、手语播报		基于图片的开放域问答、用语言操作数字可视化、可控人设的开放域问答	生成式问答、诗词创作						
	星火	会议纪要整理、稿件撰写、数字员工,AI虚拟人				人机交互、多模感知					作文批改、语言学习
	书生（INTERN）	以文生图		居家机器人		自动驾驶、图像分类					
	Mengzi(孟子)	内容创作、以文生图	金融					营销文案生成、文化创意、舆情分析			
	HunYuan（混元）	内容创作			3D虚拟场景			广告内容理解			

专业类										
ProteinLM	MEGA-Protein	鹏程神农	空天.灵眸	秦岭.西电遥感脑	PanGu-Coder	CodeGeeX	盘古·气象	东方御风	天佑	Huatuo
生物制药			遥感		代码生成/编辑		气象	流体仿真	轨道交通	医学知识问答

咨询服务 | 底座甄选服务

Large Language Model Selection

主流模型研究

百度	阿里	讯飞	智谱	百川	商汤	StabilityAI
文心一言	通义大模型	星火认知大模型	GLM	百川大模型	日日新大模型 SenseNova	Stable Diffusion
• 知识增强大语言模型 • 插件市场开放	• 局部开源 • Qwen-7B • 通义大模型架构	• 多模态 • 行业方案	• 局部开源 • ChatGLM2-6B	• 局部开源 • Baichuan-7B/13B	• 商量 SenseChat • 秒画 SenseMirage • 如影 SenseAvatar • 琼宇 SenseSpace	• 开源 • Stable Diffusion XL

xRunda.com 模型研究
2023年8月版

其他模型研究

国际跟踪

OpenAI	• GPT
Anthropic	• Claude
Meta	• LLaMA2
Google	• PaLM2 / Bard
Microsoft	• Bing
BigScience	• BLOOM
Falcon	• The Technology Innovation Institute

国产跟进

360	• 智脑
智源	• 悟道·飞鹰 Aquila
OpenBMB	• 面壁 CPM-Bee
元象	• XVERSE

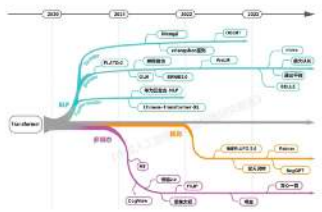
其他探索

Code	• StableCode / StarCoder WizardCoder-15B / AquilaCode / PanGu-Coder2
Embedding	• GTE / BGE / M3E
OpenAI	• Shap-E / Whisper
OpenMMLab	• Open Models
Raven	• RWKV 7B
	• AudioGen
	• WizardMath

xRunda.com 模型研究
2023年8月版

模型利用策略

底座预研



中国人工智能大模型地图研究报告

开源大模型

浅研探索	私有化
6B+ • 自然语言交互轻应用 • 终端设备内置	10B+ • 小微功能应用 • 传统业务功能升级 30B+ • 轻度智能应用 60B+ • 行业智能化

闭源大模型

API 调用	云端私有化	本地私有化
100B+ • 原生应用 • 知识库方案	60B+ • 企业信息系统升级 100B+ • 大型企业智能化改造	10B+ • 小微功能应用 • 敏感业务功能升级 30B+ • 轻度智能应用 • 企内局部功能升级

xRunda.com 模型研究
2023年8月版

Solution 解决方案

咨询服务 | 定制方案服务

xRunda.com Solution Customization

前期咨询 Consultation → 立项策划 Project Planning → 项目调研 Project Research

预算评估 Budget Assessment ← 方案定制 xRunda - Solution Customization

PoC / 验证 ↔ 研发实施 Proof of Concept / Validation ↔ Product Development Stage → 成果交付

xRunda.com 定制方案研究
2023年8月版

10B ⩽
适合展示及微调探索

⩾ 60B
适合API级快速PoC验证

⩾ 100B
适合落地应用及展望

CLICK TO CONTINUE

原生应用开发

大模型原生应用方案

原生开发工作流

需求调研 → 概念验证 → 方案确认 → 应用研发 → 应用上架 → 应用运维

xRunda.com

Prompt Engineering
提示工程

LLM Plugins
模型插件

Function Calling
函数调用

Code Interpreter
代码解释器

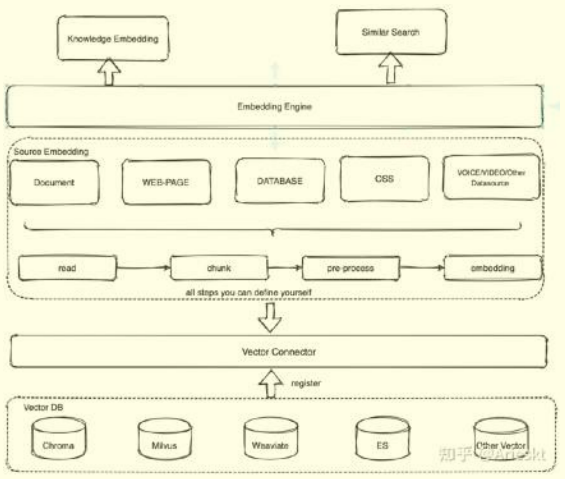
B
Solution 解决方案

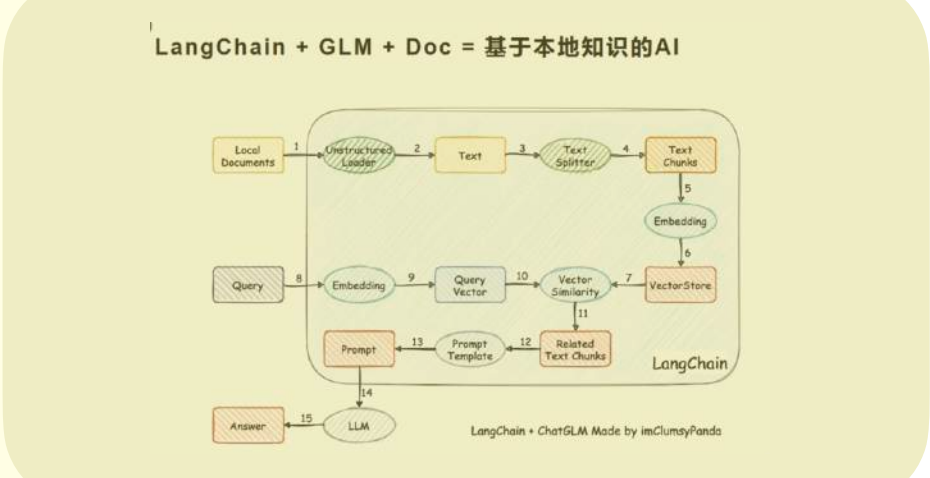
嵌入产品研发

大模型嵌入产品方案

嵌入方案工作流

数据调研 → 概念验证 → 方案确认 → 数据处理 → 应用研发（数据要素化） → 应用上线 → 应用运维





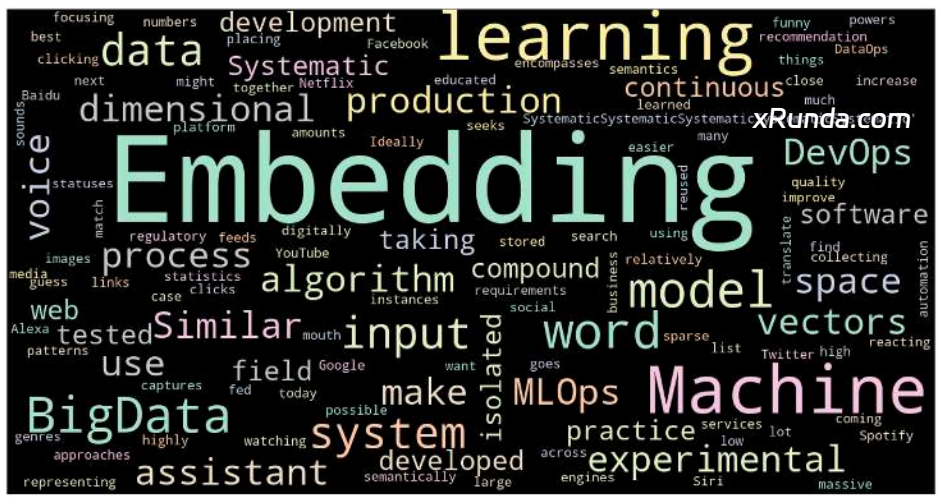
Chunk
内容处理

Embedding
向量嵌入

Recall
召回

Knowledge Base
知识库

Text-to-SQL
数据库交互

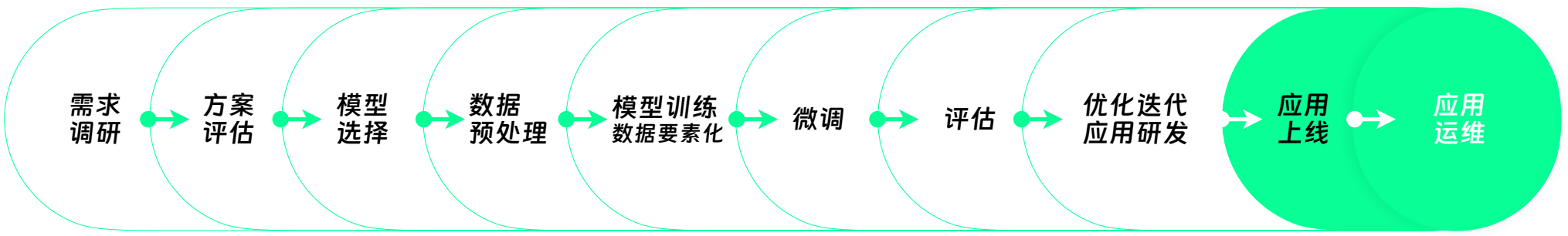


CLICK TO CONTINUE

模型微调研发

模型微调技术方案

微调研发工作流

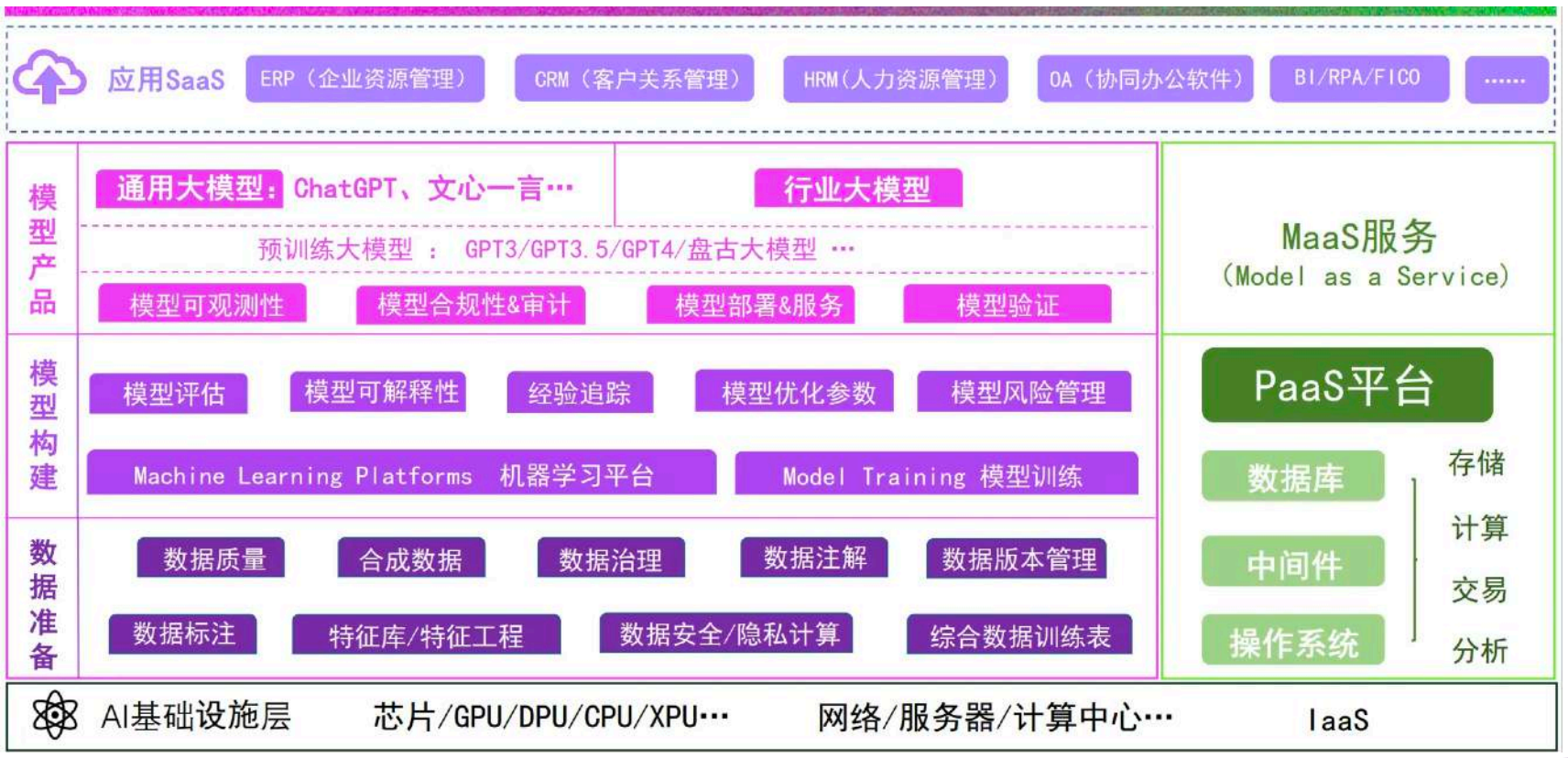


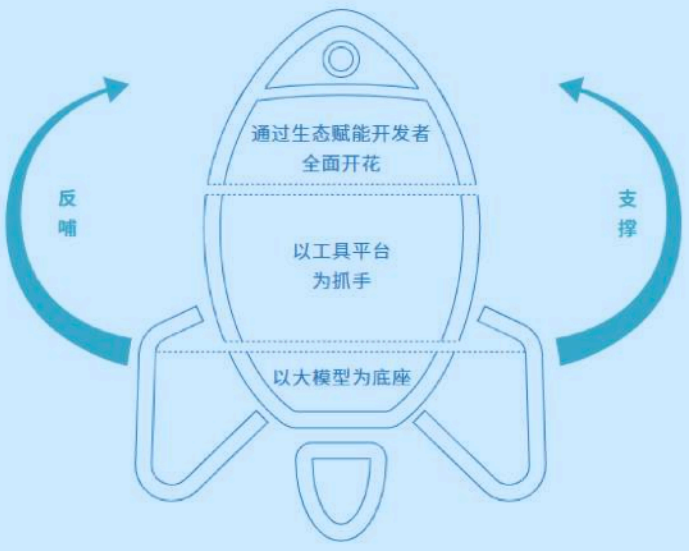
- Adapter Fusion Tuning
- Prefix Tuning
- Prompt Tuning
- Instruction Tuning 指令微调
- SFT, Supervised FineTune 监督微调
- PEFT, Parameter-Efficient Fine-Tuning 参数高效的微调方法
- PT, P-Tuning v2 xRunda.com
- Freeze 监督微调 xRunda.com
- LoRA, Low-Rank Adaptation of Large Language Models 大语言模型的低阶自适应
- QLoRa: Quantized LLMs with Low-Rank Adapters 使用低秩适配器的量化大语言模型
- Knowledge Distillation 知识蒸馏
- RL, Reinforcement learning 强化学习方式微调
- Pre-training 预训练

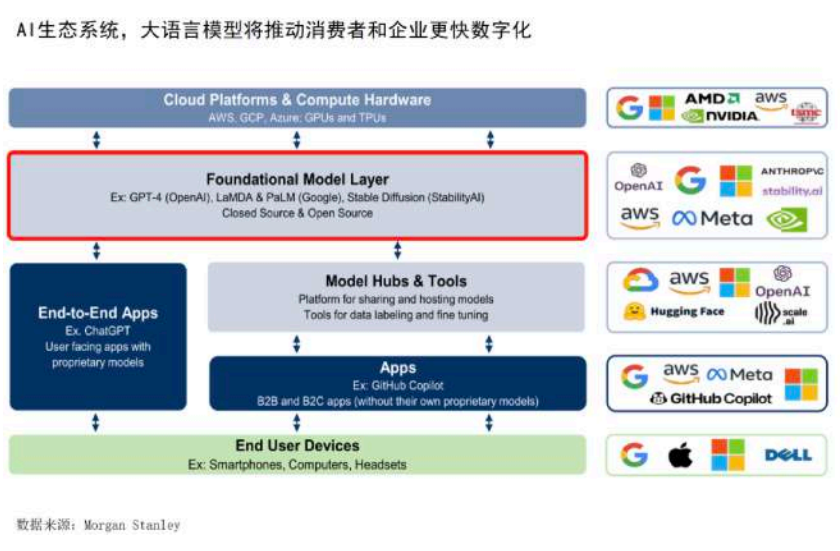
企业智化服务

企业数智化方案

自象限 AI Infra 架构图







ERP | CRM | sCRM | HRM | OA | BI | RPA

Distinctive Capabilities and Services

特色能力服务

特色能力服务

Copilot X

Efficient Development Intelligent Teams

特色能力服务

LangChain

PoC 快速概念验证

LangChain 提供的功能或支持的集成

数据预处理	UnstructuredIO、Airbyte...
数据索引	GPT-Index...
Doc&Text Splitter	Generic Recursive Text Splitter、Markdown Splitter、Python Code Splitter...
向量数据库与检索	FAISS、Pinecone、Weaviate、Elastic...
图数据库	Chroma...
外部知识或操作	SerpApi、Searx、Wikipedia API、Wolfram Alpha、Zapier Natural Language Actions API...
LLM API	OpenAI、Hugging Face、Cohere、Anthropic、PaLM、GooseAI、Cerebrium AI、Forefront AI、Petals...
Embedding 引擎	OpenAI、Hugging Face、Cohere...
可观测性	Helicone、Prompt Layer、Weights&Biases...
应用部署	Streamlit、Hugging Face(Gradio)、Steamship、Kookaburra...
模型评估数据集	Hugging Face (truthful qa)、LangchainDatasets...
模型回复结构化及验证	kor、guardrails...

特色能力服务 AI x Lab

Frontier Technology Lab

前沿技术实验室

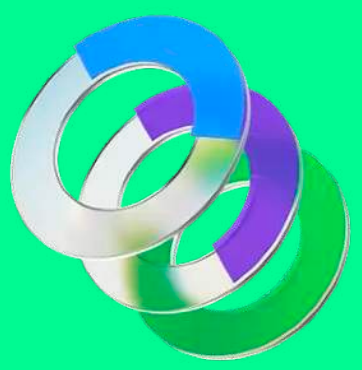
xRunda.com

Big Model Middleware

大模型中间件

Studio Mode

工作室模式



Co-Creation Model

共创模式

特色能力服务 xLLMda 一问多答方案

多模型API同步调用

搜索辅助

数据管理

xRunda.com

大模型调度增强

服务托管

特色能力服务 | xTune 一通多调方案

数据处理

多模型同步微调训练

利用LLM实现多微调同步验证

xRunda.com

服务托管

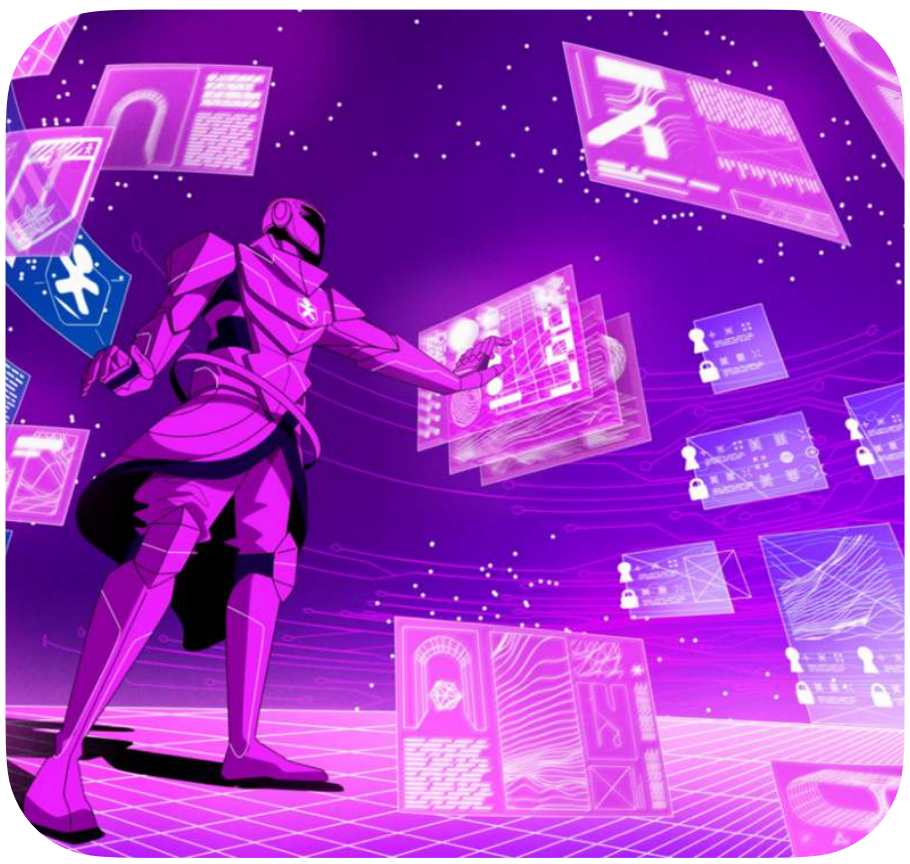
共建模式

特色能力服务 | xBenchmark 评测服务

中文通用大模型测评基准

- 定期在系列国内外代表性大模型上使用多个维度能力考题进行测试
- 借鉴国际较为知名的预训练大语言模型评测集，汇集经验，增加亮点，构建适合普适商用评估的新评测框架
- 为项目定制策划测评题库，推出针对性测评
- 提升模型研发验收能力

xRunda.com

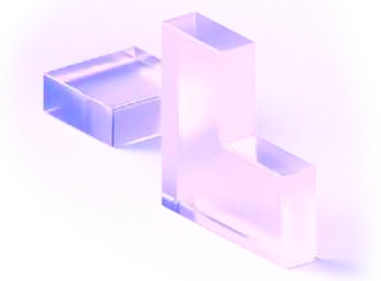


MELT

Multi-task Evaluation in Language and Thinking

多任务语言与思维评测

一个多任务的评测，覆盖了语言和思维两个方面



AGILE

Agent-guided and Integrated Language Evaluation

Agent引导综合语言评测

突出了Agent导向，并且强调了这是一个综合的语言评测

Solution 解决方案

特色能力服务

WEB3 服务

C

技术方案

Technical Solutions

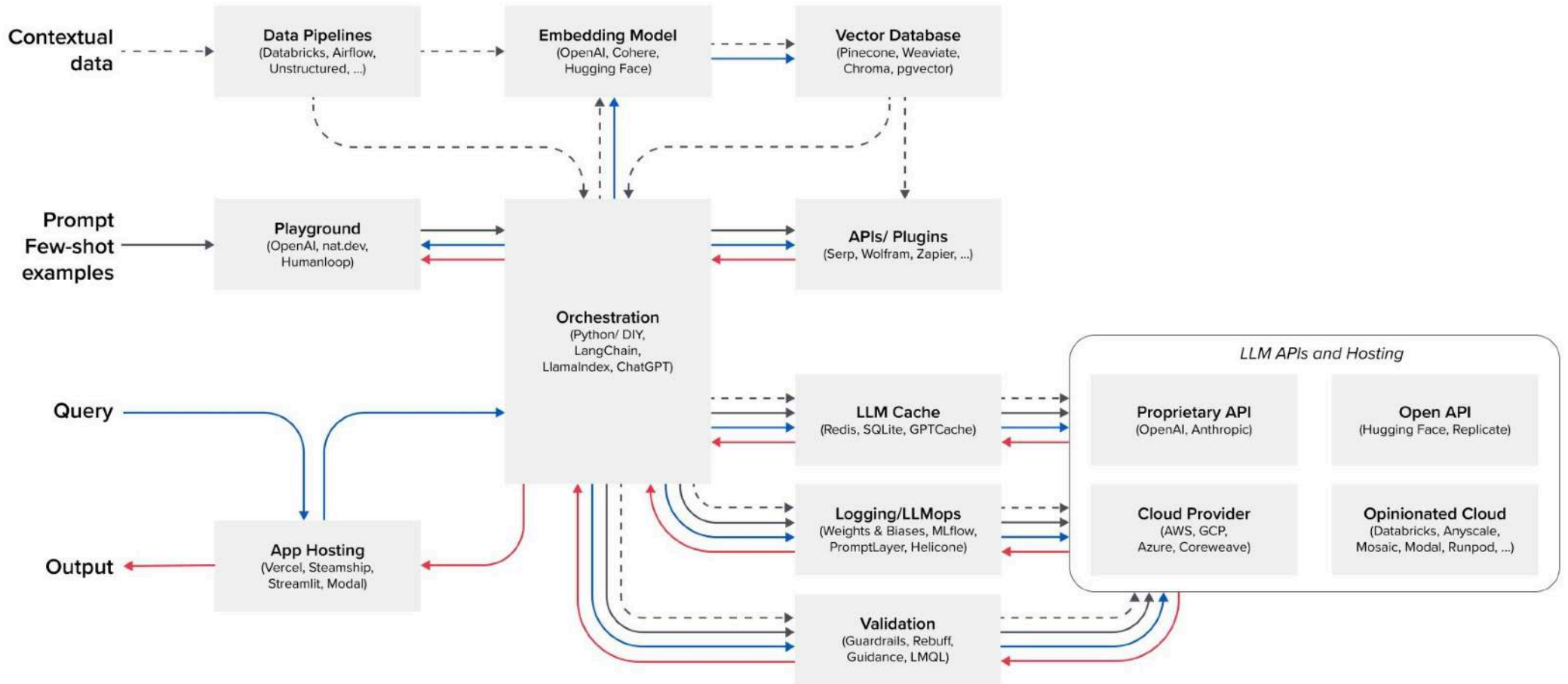
C

技术栈总览
Technology Stack

Web2 工具箱			Web3 工具箱			AI 工具箱		
类别	技术/工具	适用场景	类别	技术/工具	适用场景	类别	技术/工具	适用场景
前端开发	React	网站、APP的用户界面开发	**区块链开发**	Solidity	Ethereum智能合约开发	**基于LLM的开发**	GPT系列	语言生成、理解
	Vue.js	网站、APP的动态界面设计		Vyper	安全的智能合约开发		BERT	文本分析、分类
	Angular	单页面应用开发		Ethereum	分散式应用开发		Transformer	序列转换、翻译
样式设计	CSS/SASS/LESS	网页样式设计		Binance Smart Chain	高性能区块链开发		Hugging Face Transformers	模型训练和微调
	Bootstrap	响应式设计		Polkadot	跨链技术		OpenNLP	语言处理
	Tailwind CSS	快速原型设计	**NFT**	OpenZeppelin	NFT标准和合约开发	**多模态应用开发**	OpenCV	图像处理
后端开发	Java	企业级应用、API开发		Truffle	区块链开发环境		TensorFlow Image	图像识别
	Python	数据处理、API开发	**DAO**	Aragon	分散式组织开发		Speech-to-Text	语音转文本
	Node.js	实时应用、轻量级服务端开发		DAOstack	DAO协议和应用开发		Text-to-Speech	文本转语音
	Ruby	快速开发	**隐私计算**	Multi-Party Computation	数据隐私保护	**知识库**	RDF	知识图谱构建
移动开发	Android (Kotlin, Java)	Android端原生开发	**DID**	Decentralized Identifiers	身份验证和管理		SPARQL	图谱查询
	iOS (Swift, Objective-C)	iOS端原生开发				**Langchain**	自定义链上语言处理	链上语言分析
跨平台开发	React Native	iOS和Android跨平台开发				**Prompt工程**	Custom Prompt Design	定制化问题解决方案
	Flutter	iOS和Android高性能跨平台开发						
小程序开发	WeChat Mini Program	微信小程序开发						
	Alipay Mini Program	支付宝小程序开发						
More			More			More		

原生应用技术方案

Emerging LLM APP Stack
新兴大模型应用程序堆栈



LEGEND

Gray boxes show key components of the stack, with leading tools/systems listed

Arrows show the flow of data through the stack

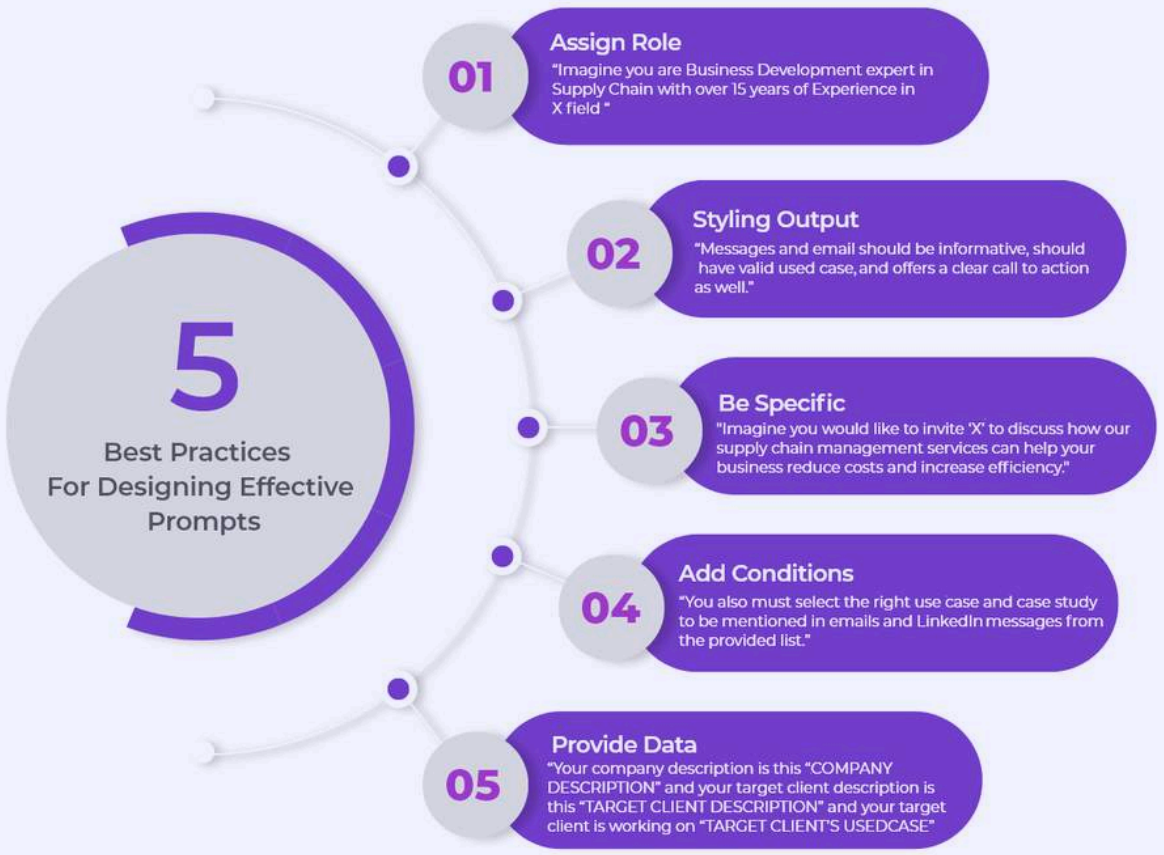
- - - → Contextual data provided by app developers to condition LLM outputs
——→ Prompts and few-shot examples that are sent to the LLM
——→ Queries submitted by users
——→ Output returned to users

a16z Enterprise

原生应用技术

Prompt Engineering 提示工程

最佳实践



针对 Prompt 进行结构、内容等维度进行优化的 AI 技术，把大模型的输入限定在一个特定的范围，进而更好地控制模型的输出

角色生成器等

垂类激发器等

业务化PT模板

多重技术策略

LLM Plugins 模型插件

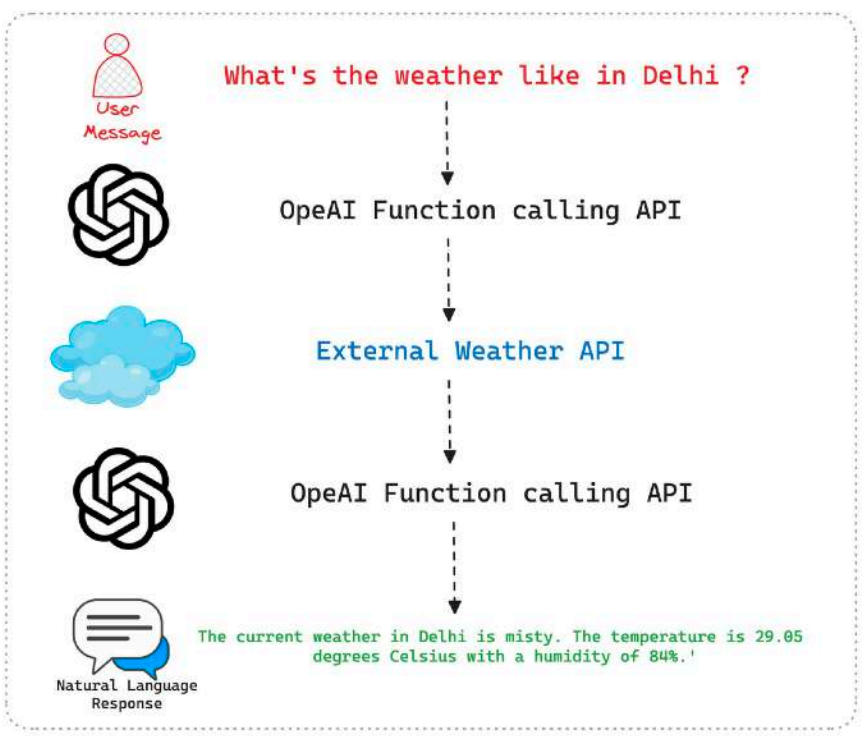
Plugins are tools designed specifically for language models with safety as a core principle, and help LLMs access up-to-date information, run computations, or use third-party services.
插件是专为以安全为核心原则的语言模型而设计的工具，可帮助大语言模型访问最新信息、运行计算或使用第三方服务。

检索实时信息

检索知识库信息

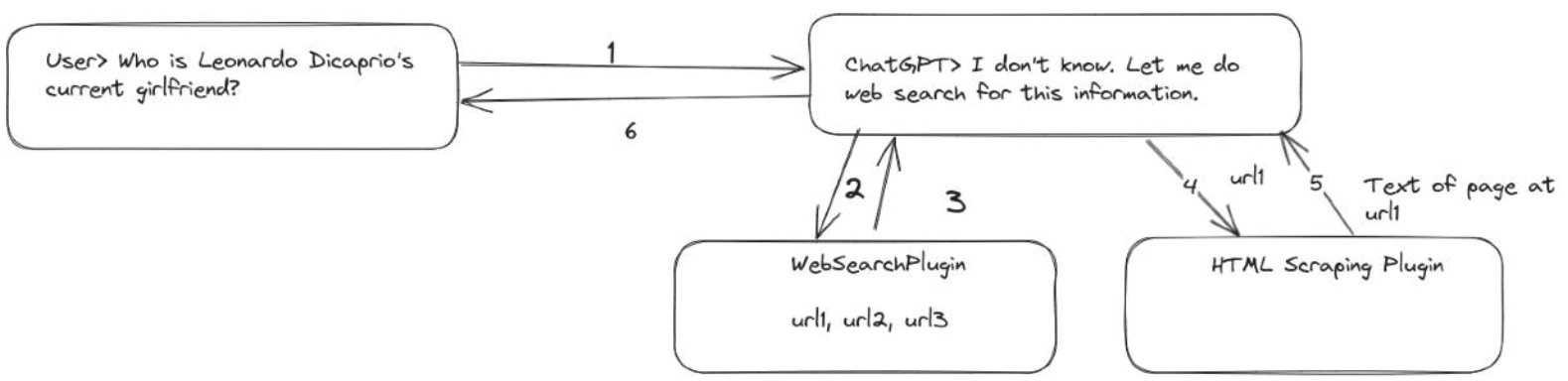
代表用户执行操作

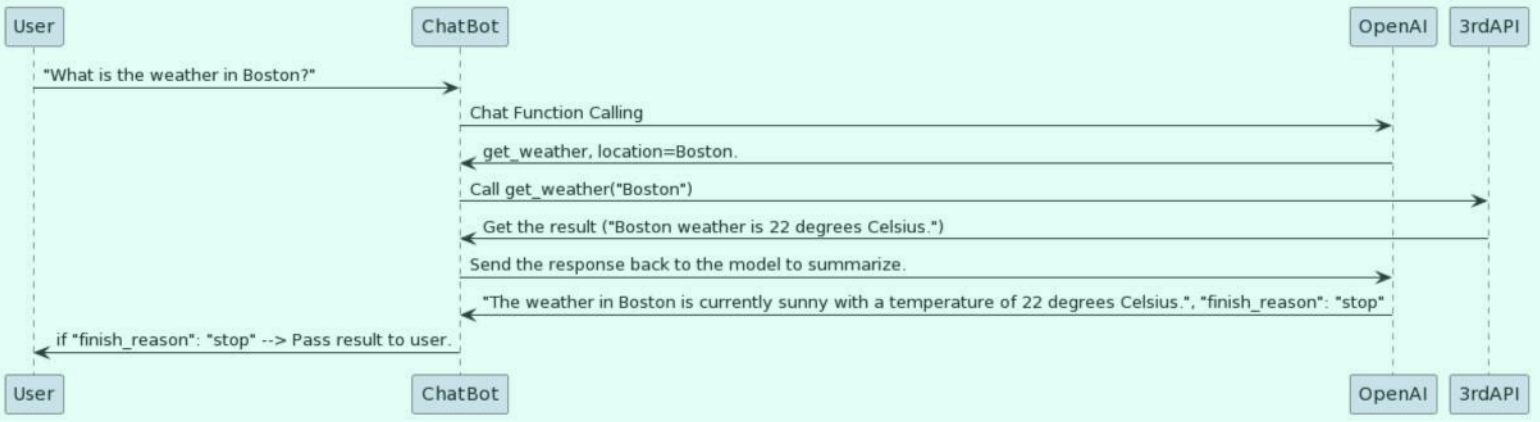
Function Calling 函数调用



函数回调可以让模型输出一个请求调用函数的消息，其中包含所需调用的函数信息、以及调用函数时所携带的参数信息。

这是一种将GPT能力与外部工具/API连接起来的新方式。

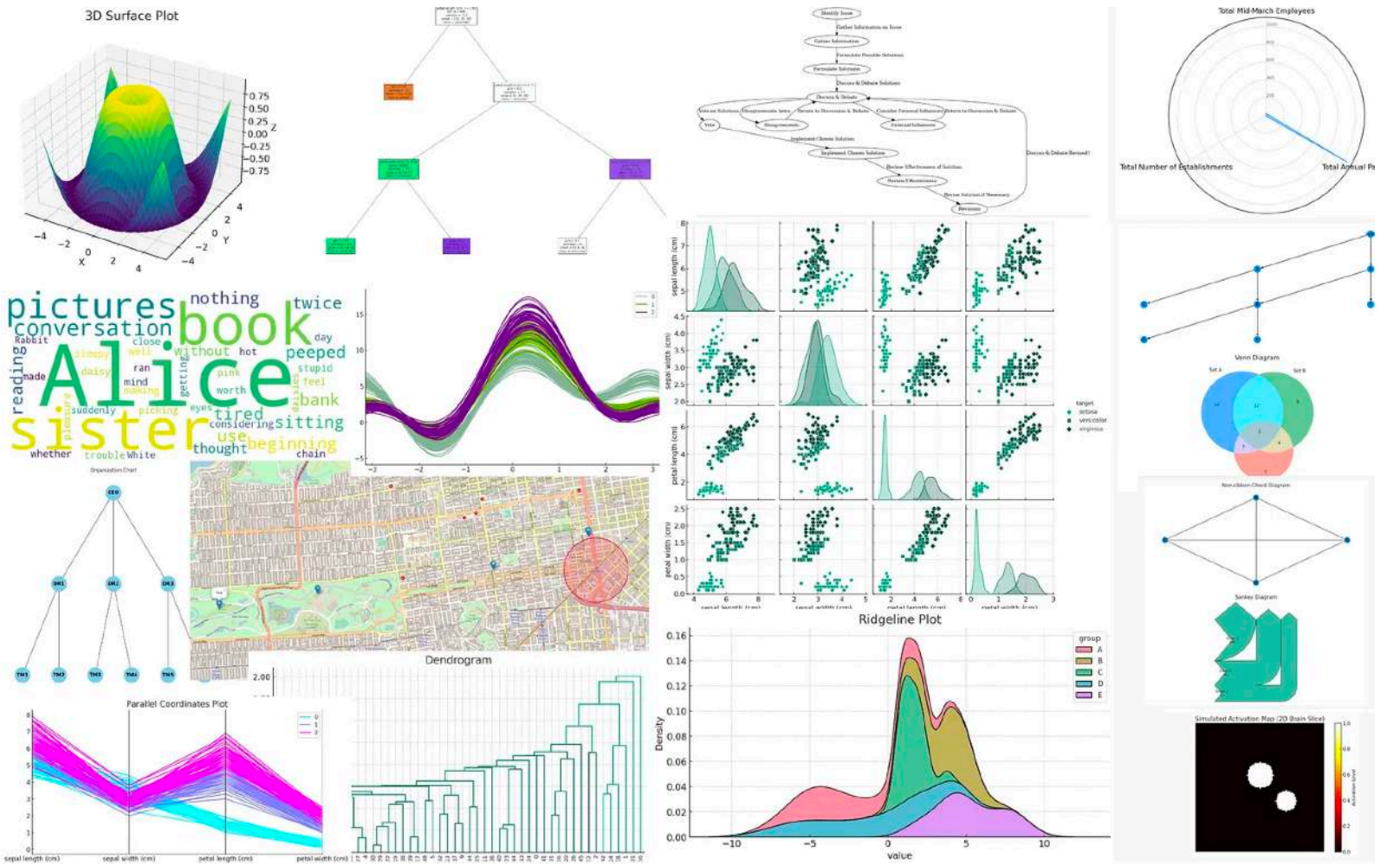




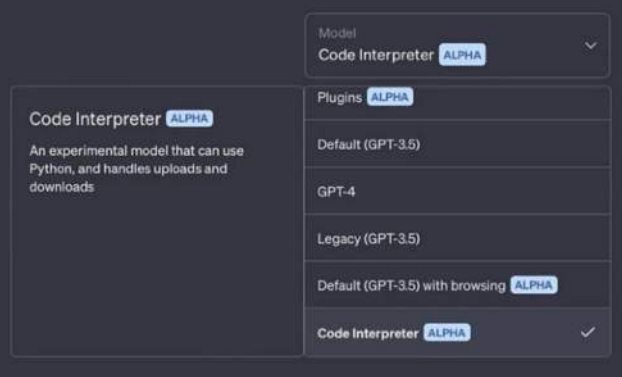
原生应用技术

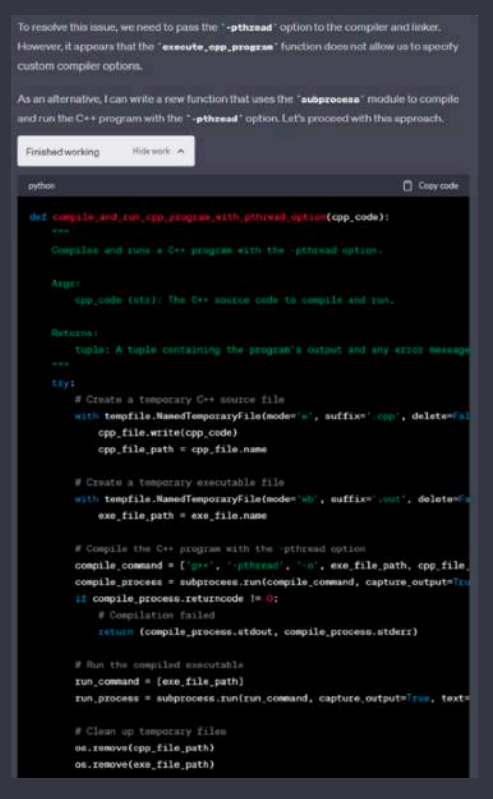
Code Interpreter 代码解释器

专业级别的代码解释器，专门用于执行Python脚本，允许用户处理各种类型数据的工具



- Solving mathematical problems, both quantitative and qualitative
  解决数学问题，包括定量和定性问题。
- Doing data analysis and visualization 进行数据分析和可视化
- Converting files between formats 在不同格式之间转换文件





嵌入产品技术方案

Chunk 内容分块

NLTK
句子分词器

提供语言上连贯的块，但大小差异很大

spaCy
分句器

提供语言上连贯的块，但大小差异很大

LangChain
文本分词器

此方法提供一致的块长度，从而产生均匀分布。该方法对文本的特定语言结构不太敏感

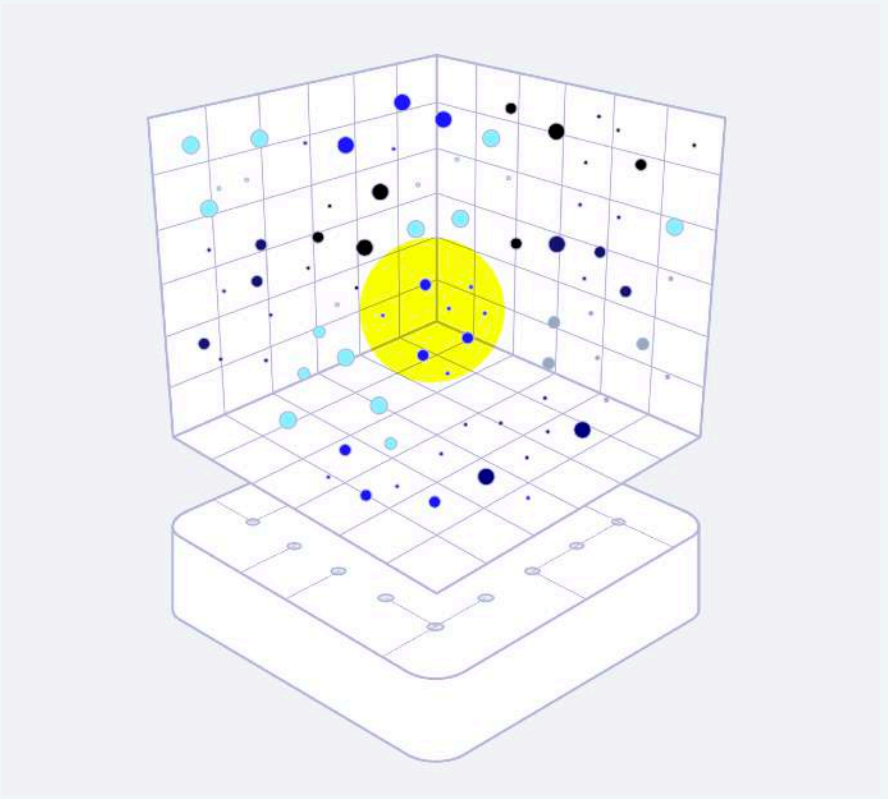
KMeans
聚类

基于语义相似度对句子进行分组的技术。通过使用句子嵌入和 K-means 等聚类算法，可以实现句子聚类。

Adjacent Sentences 聚类

克服 KMeans 聚类的一些限制，特别是句子顺序的丢失。基本前提是在文本中连续出现的两个句子比相隔较远的两个句子更有可能在语义上相关

Embedding 向量嵌入



Word2Vec 基于 Seq2Seq 的神经网络结构

Glove 词共现矩阵

Item2Vec 推荐中的双塔模型

FastText 浅层神经网络

ELMo 独立训练双向，Stacked Bi-LSTM 架构

GPT 从左到右的单向 Transformer

BERT, Bidirectional Encoder Representations of Transformers 双向 Transformer 的 Encoder，Attention 联合上下文双向训练。生成文本嵌入变换器模型

Baidu Ernie-3.0-base-zh

https://github.com/PaddlePaddle/PaddleNLP

https://zhuanlan.zhihu.com/p/523727481

模型	模型大小	（中文）数据量	训练方法
ERNIE1.0	参考bert base(110M)	Wiki，baike，news，tieba	pretraining + finetuning
ERNIE2.0	参考bert base(110M)， bert large（340M）	wiki，news，dialogue，IR，discourse relation	pretraining + finetuning
ERNIE3.0	10B	4TB（ERNIE2.0，search,web,QA-long，QA-short，novel，poetry&couplet，medical，law，financial,KG）	progressive training + finetuning / zero-shot

Text2vec 文本表征及相似度计算

Text2vec-large-chinese（LERT，升级版）

约占用显存3GB，可修改为CPU中运行。基于CoSENT方法训练，将MacBERT替换为LERT，其他训练条件不变

Base（CoSENT方法训练，MacBERT）

属于余弦句子模型，基于CoSENT方法训练，使用MacBERT，它将句子映射到768维的密集向量空间，可用于句子嵌入、文本匹配或语义搜索等任务

Azure OpenAI Embeddings

相似性嵌入

擅长捕获两个或更多文本片段之间的语义相似性

文本搜索嵌入

可帮助度量长文档是否与简短查询相关

代码搜索嵌入

可用于嵌入代码片段和嵌入自然语言搜索查询

M3E, Moka Massive Mixed Embedding

SOTA

评测 BenchMark 使用 MTEB-zh，通过千万级 (2200w+) 的中文句对数据集进行训练

Recall 召回

向量召回

- 嵌入捕获向量空间中的语义相似性，从而能更轻松地对表示字词的大型输入进行机器学习
- 使用嵌入来确定两个文本区块在语义上是否相关或相似，并提供一个分数来评估相似性
- Cosine distance余弦相似度

Cosine distance
余弦相似度

L2-Squared distance
欧氏距离

Dot Product distance
点积距离

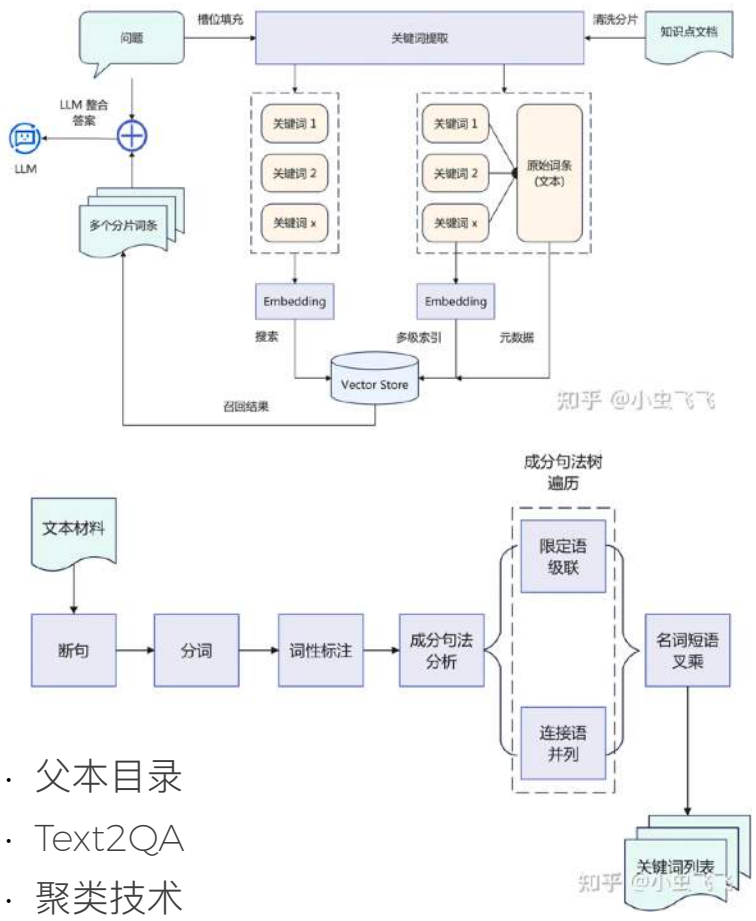
Hamming distance
汉明距离

生成结果方案

Stuff　Refine　Map Reduce　Map Rerank

辅助召回

- 关键词提取

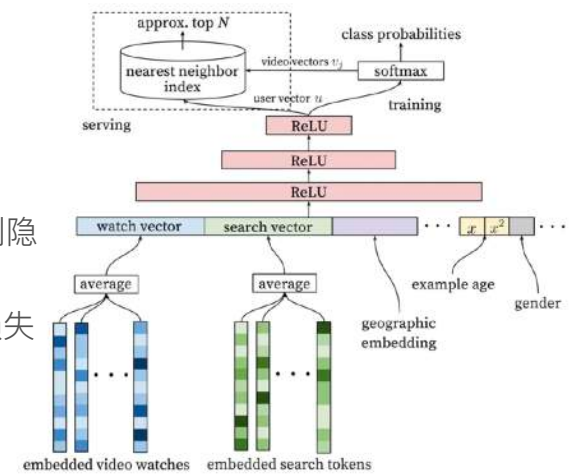


- 父本目录
- Text2QA
- 聚类技术
- 项目表征任务

xRunda.com

推荐系统等算法召回

- 深度候选生成模型架构，展示了嵌入式稀疏特征与密集特征的连接
- 在连接前对嵌入进行平均处理，将可变大小的稀疏 ID 包转换为适合输入到隐藏层的固定宽度向量。所有隐藏层全连接
- 在训练过程通过在采样 Softmax 的输出上进行梯度下降，最小化交叉熵损失
- 在服务过程中执行近似的最近邻查找，生成数百个候选推荐



嵌入产品技术 Knowledge Base 知识库

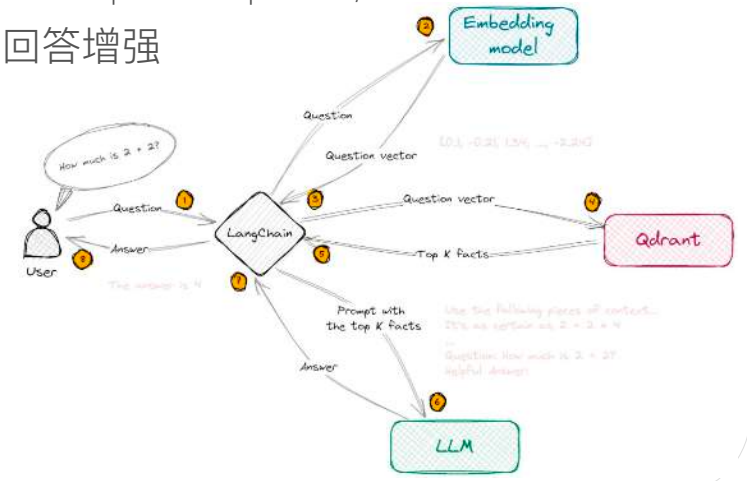
数据预处理

- PDF / MARKDOWN /
- WORD / PPT / HTML /
- CSV / FROM

- 多模态技术
- 多模态嵌入
- 多模态理解

LLM 接入

Prompt Engineering /
Prompt Template /
回答增强



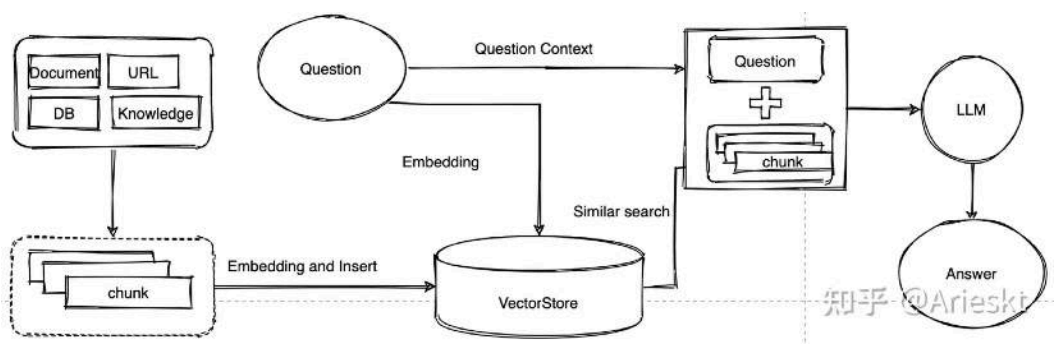
数据接入

- 系统改造 / 数据改造 ⟶ 迭代优化

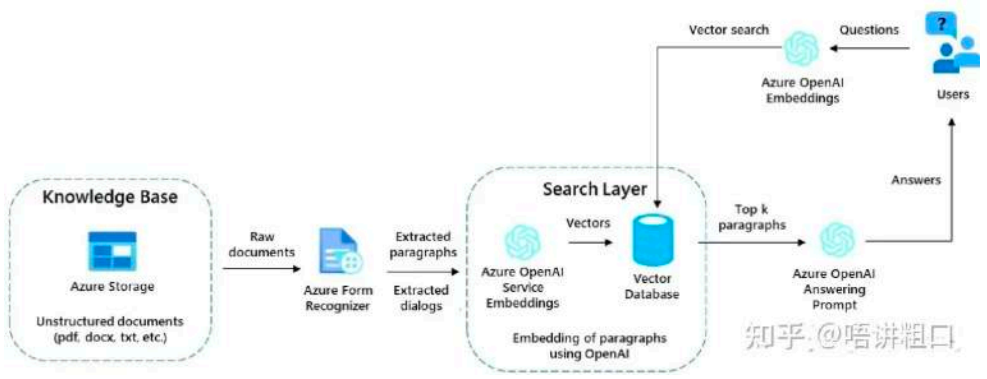
方案参考

DB-GPT

https://github.com/eosphoros-ai/DB-GPT



Azure OpenAI



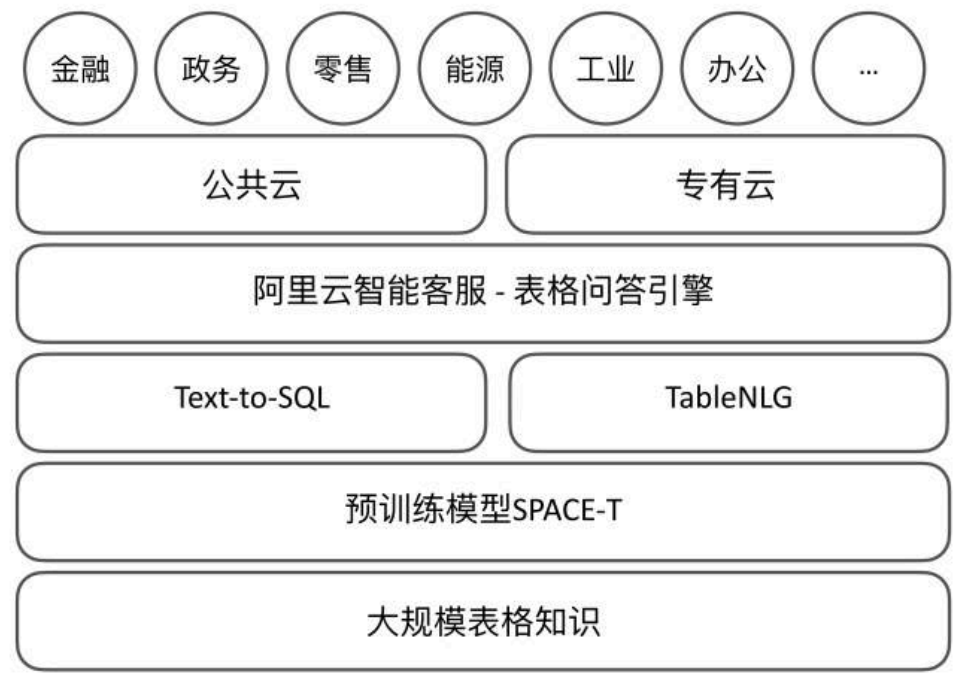
TableQA / Text2SQL 表库联合

TableQA

利用模型将自然语言转换为 SQL 查询语言，允许用户使用自然语言与表格知识直接交互并返回直观、流畅、忠实的结果

Text2SQL

将一个自然语言问题转换为相应的可执行结构语句（SQL）



多领域支撑：SPACE-T及表格问答引擎已经支撑了各领域客户，包括金融、政务、零售、能源等；

多平台输出：SPACE-T及表格问答引擎已经在公共云和专有云同时输出满足不同客户的需求；

多项目落地：SPACE-T系列模型已经实现规模化落地应用，帮助企业基于二维表格结构化数据快速构建机器人；

模型微调技术方案

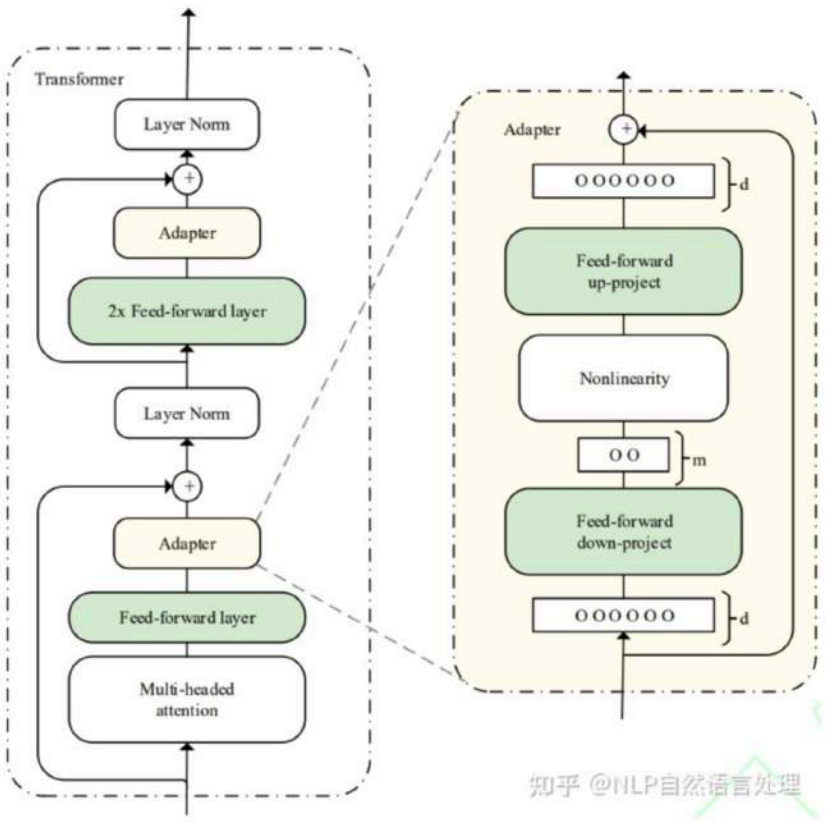
更多访问 www.xRunda.com

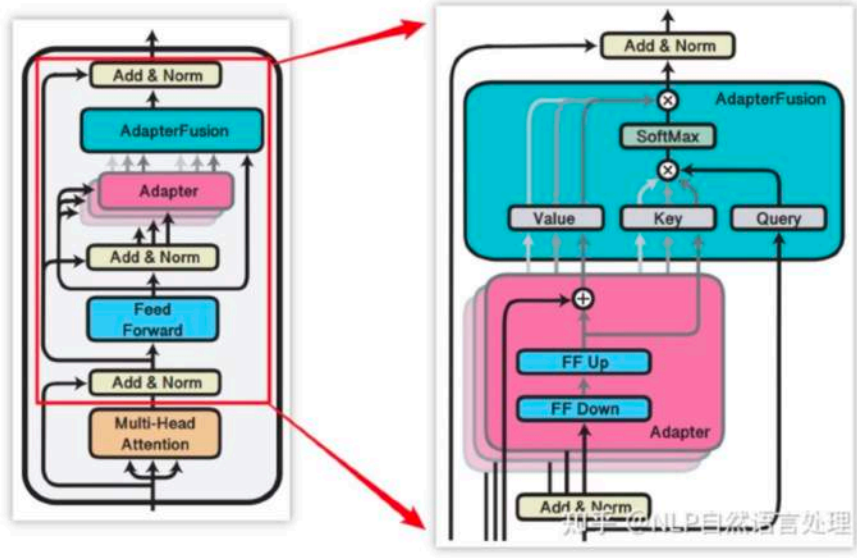
模型微调技术

Adapter Fusion Tuning

在预训练语言模型的每一层中间插入一个小型的可训练模块来调整模型的输出

Adapter Fusion 算法改进 —— 用以实现多个 Adapter 模块间的最大化任务迁移





知识提取阶段

训练 Adapter 模块学习下游任务的特定知识，
将知识封装在 Adapter 模块参数中

知识组合阶段

将预训练模型参数与特定于任务的 Adapter 参数固定，
引入新参数学习组合多个 Adapter 中的知识，提高模型
在目标任务中的表现

优势

· 解决了灾难性遗忘、任务间干扰和训练不稳定的问题

· 大多数情况下性能优于全模型微调和 Adapter

局限

· 增加了模型层数，引入了额外的推理延迟

Adapter

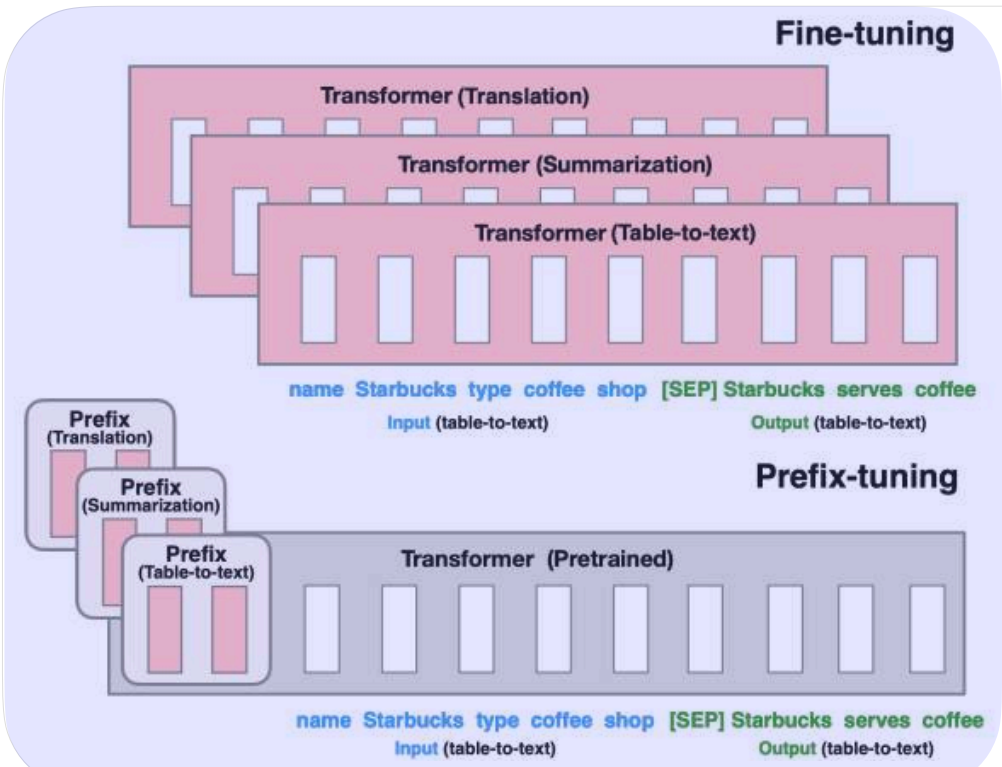
Paper

Parameter-Efficient Transfer Learning for NLP https://arxiv.org/pdf/1902.00751.pdf

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer https://arxiv.org/pdf/2005.00052.pdf

Prefix Tuning

在预训练语言模型的输入端添加一个可训练的前缀网络来控制模型的生成行为



局限性

难训练，效果不如直接 Finetune

Paper

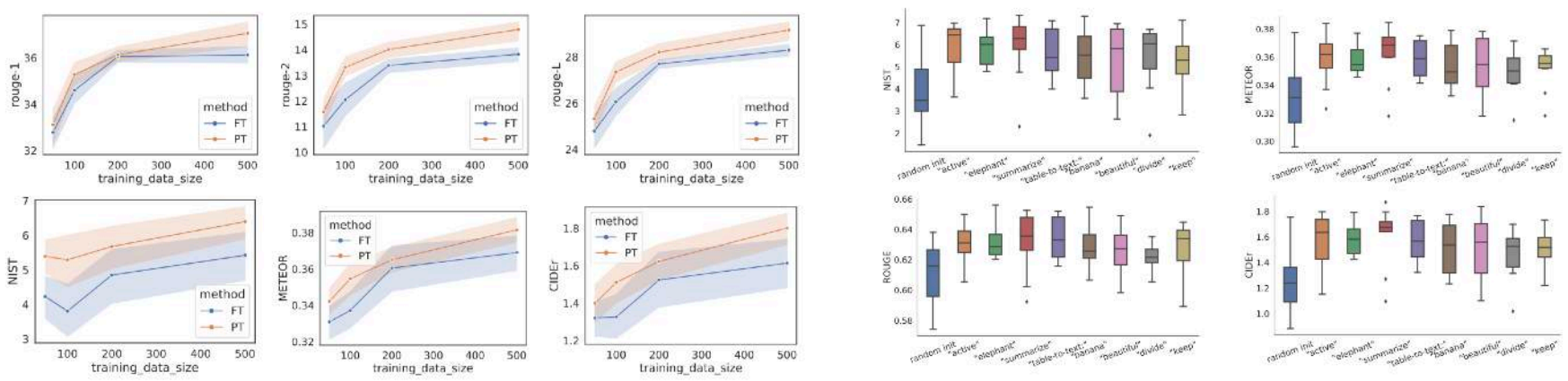
Prefix-Tuning: Optimizing Continuous Prompts for Generation

将一个连续的特定于任务的向量序列添加到输入，称之为前缀

与 Prompt 方法不同的是，前缀完全由自由参数组成，与真正的 Token 不对应

相比于传统的微调，前缀微调只优化了前缀

只需存储一个大型 Transformer 和已知任务特定前缀的副本，对每个额外任务产生非常小的开销



https://arxiv.org/pdf/2101.00190.pdf

模型微调技术

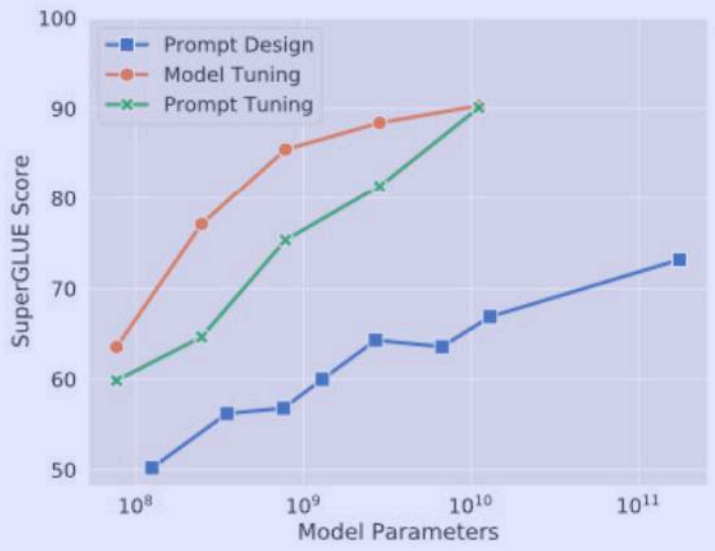
Prompt Tuning

在预训练语言模型的输入端添加一个可训练的前缀网络，来控制模型的生成行为

Prompt-based Methods

在预训练语言模型的输入端添加一个固定的文本片段 Prompt 来引导模型完成特定任务

给每个任务定义专属 Prompt，拼接到数据上作为输入，
同时 Freeze 预训练模型进行训练，
在没有加额外层的情况下，可以看到随着模型体积增大，
效果越来越好，最终追上精调的效果

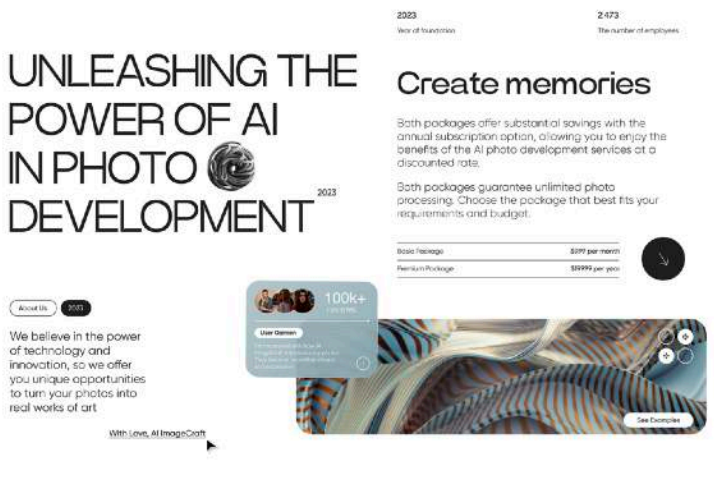


构建模板（Template Construction）
标签词映射（Label Word Verbalizer）

Prompt-ensembling

在一个 Batch 里同时训练同一个任务的不同 Prompt

相当于训练了不同「模型」，比模型集成的成本小很多



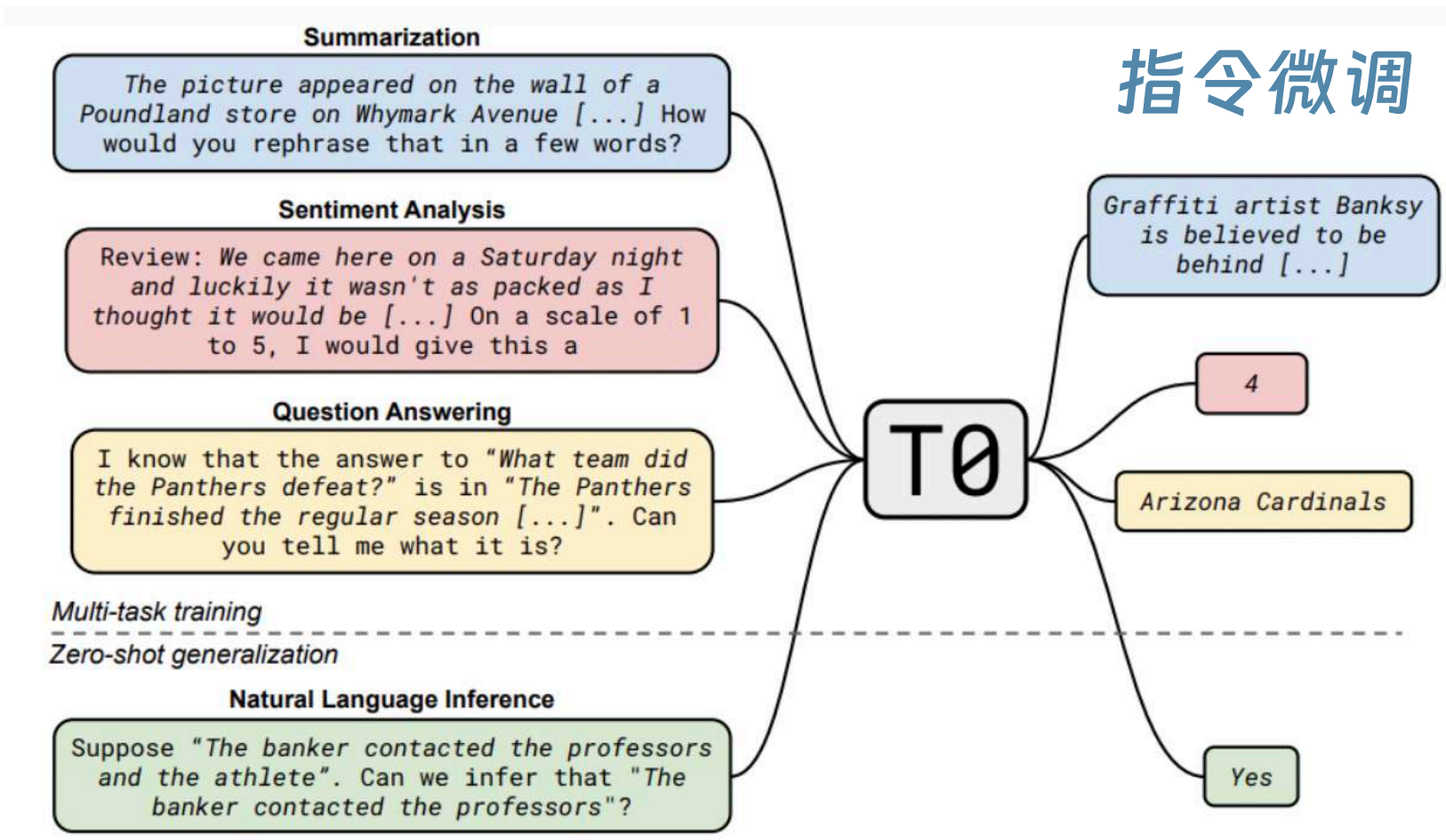
典型案例

GPT3

Language Models are Few-Shot Learners（NIPS2020）

模型微调技术

Instruction Tuning



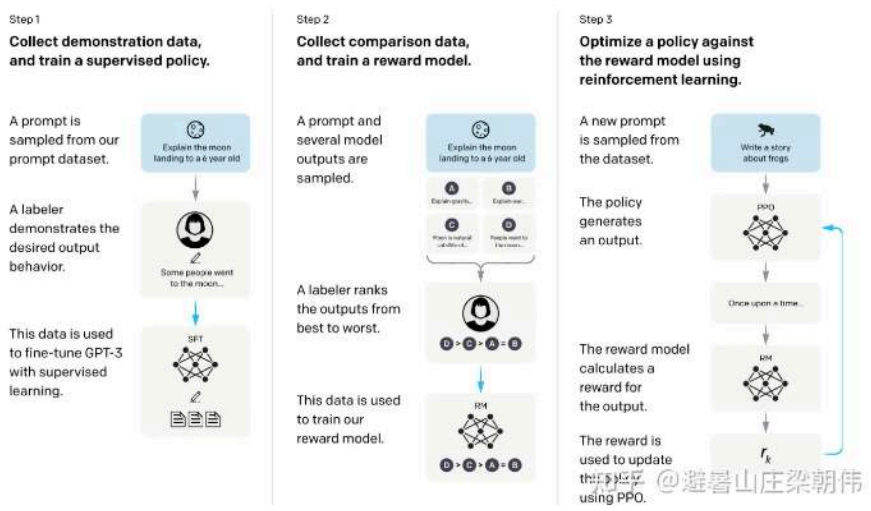
针对每个任务，单独生成 Instruction（hard token），通过在若干个 Full-shot 任务上进行微调，然后在具体的任务上进行评估泛化能力（zero shot），其中，预训练模型参数是 Unfreeze 的

Instruction Tuning 与 Prompt Tuning 的区别是激发语言模型的理解能力，通过给出更明显的指令，让模型去理解并做出正确的反馈

典型案例

InstructGPT

Language Models are Few-Shot Learners（NIPS2020）



模型微调技术 Supervised FineTune

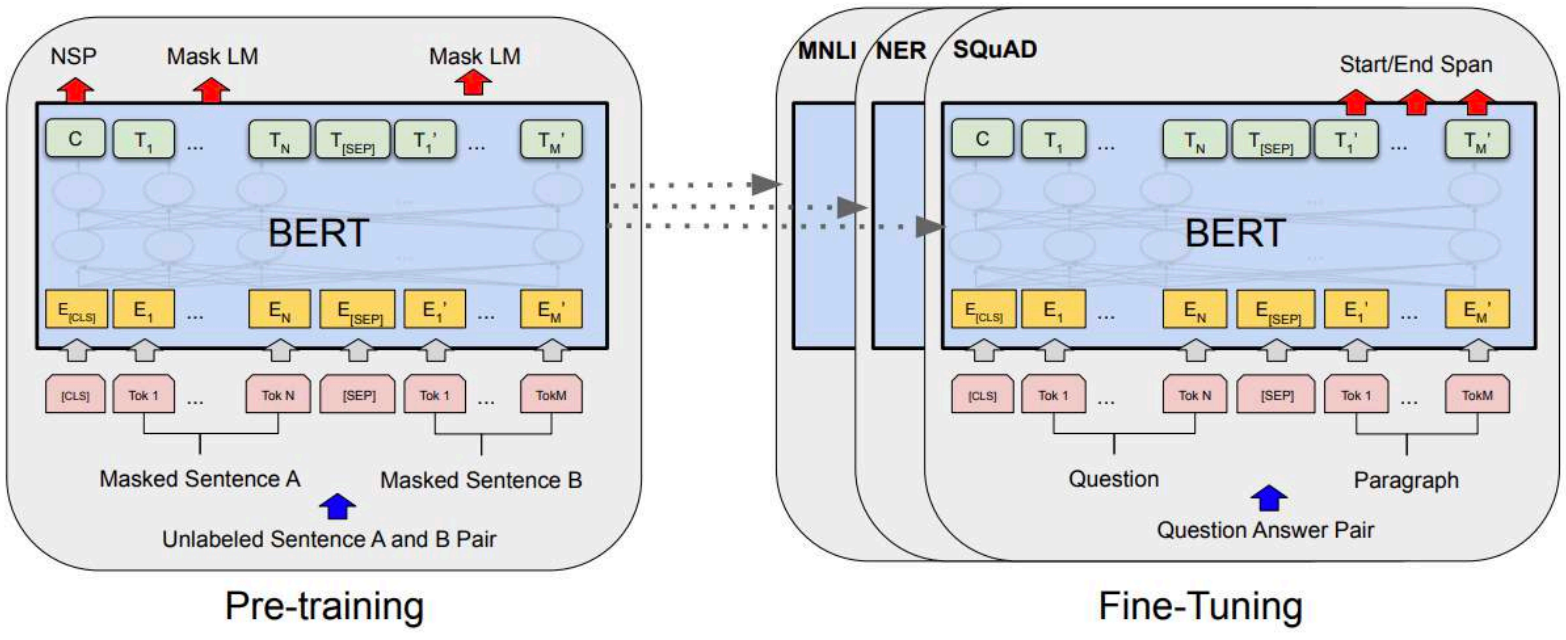
监督微调 SFT

1. 在源数据集上预训练一个神经网络模型，即源模型。
2. 然后创建一个新的神经网络模型，即目标模型。
3. 目标模型复制了源模型上除了输出层外的所有模型设计及其参数。这些模型参数包含了源数据集上学习到的知识，且这些知识同样适用于目标数据集。
4. 源模型的输出层与源数据集的标签紧密相关，因此在目标模型中不予采用。
5. 为目标模型添加一个输出大小为目标数据集类别个数的输出层，并随机初始化该层的参数
6. 在目标数据集上训练目标模型时，将从头训练到输出层，其余层的参数都基于源模型的参数微调得到。

典型案例

Bert

Language Models are Few-Shot Learners（NIPS2020）



模型微调技术 Parameter-Efficient Fine-Tuning

PEFT

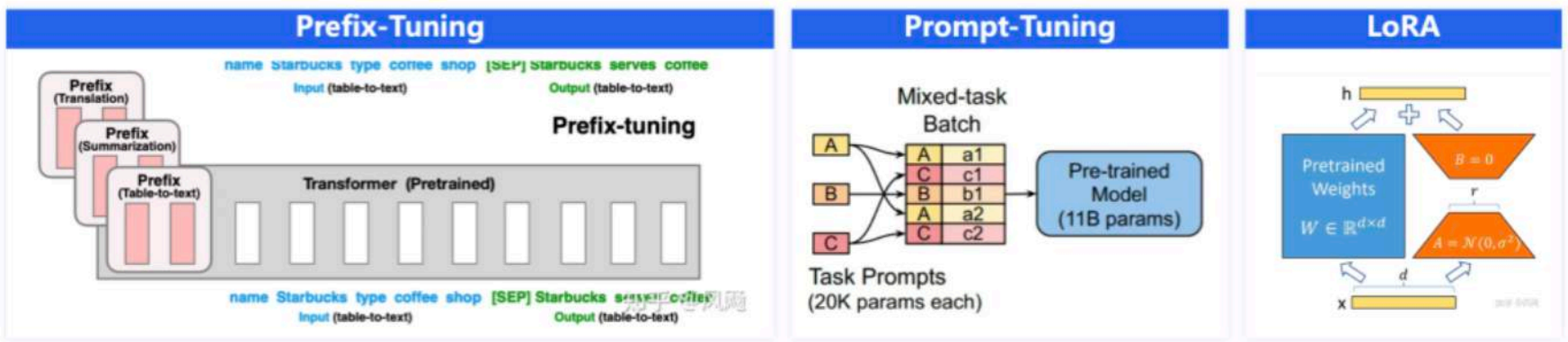
参数高效的微调方法

I. 全量微调对大型预训练语言模型的训练成本很高

II. 简单的微调接近下游任务的最后几层参数，难以达到较好的效果

III. 在面对特定的下游任务时，如果进行全量微调，太过低效

PEFT 中常用的 P-Tuning/LoRA 方法 都是在合适的位置增加需要微调的参数



Prefix tuning/Prompt tuning特点

- 可以复用基础模型的预测服务，外挂多个p tuning模型
- 减少了模型的可用序列长度
- 效果差于full-finetuning

LoRA特点

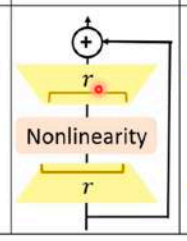
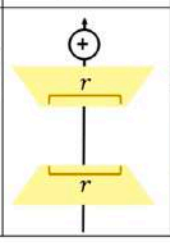
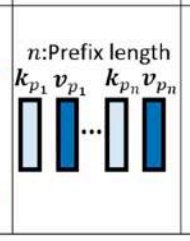
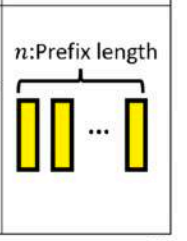
- 推理过程与Full-finetune一样，没有额外的计算量
- 不会减少模型的可用序列长度
- 训练效果损失较少

Parameter-Efficient Fine-Tuning

更多访问 www.xRunda.com

https://blog.csdn.net/weixin_43154149/article/details/124370319

• Benefit 1: Drastically decreases the task-specific parameters

	Adapter	LoRA	Prefix Tuning	Soft Prompt
Task-specific parameters*	$\Theta(d_{model}rL)$	$\Theta(d_{model}rL)$	$\Theta(d_{model}nL)$	$\Theta(d_{model}n)$
Percent Trainable	<5%	<0.1%	<0.1%	<0.05%
Illustration				

*not including the classifier head

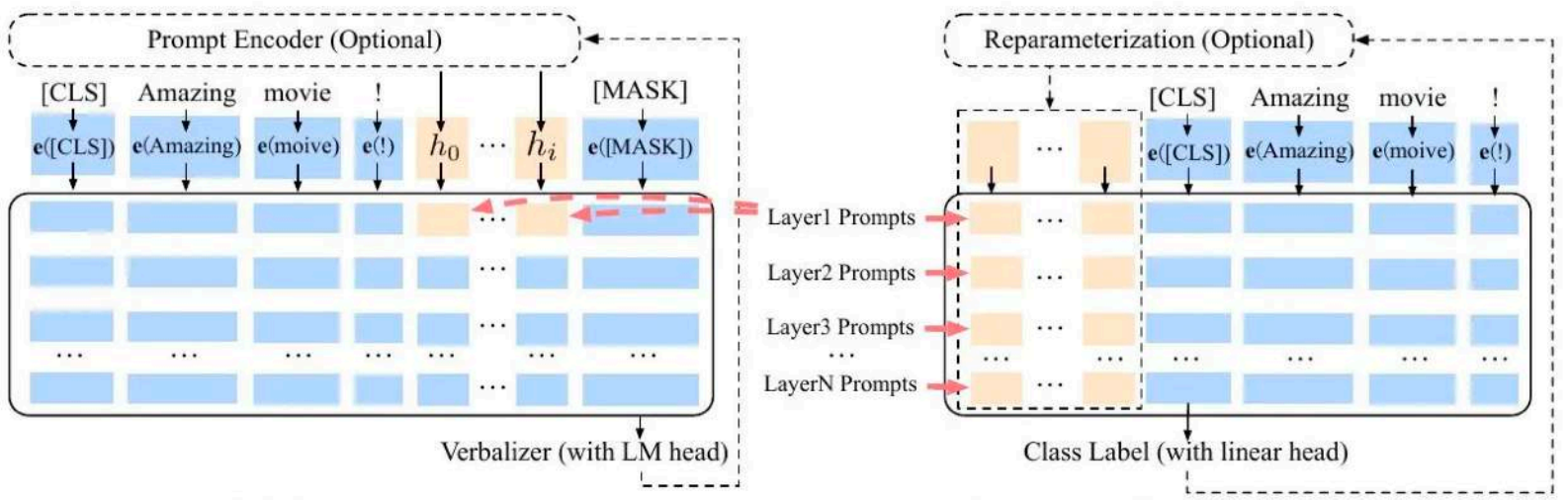
P-Tuning v2

PT 一种针对于大模型的 Soft-prompt 方法

PTv1 VS PTv2

P-Tuning 仅对大模型的Embedding加入新的参数

P-Tuning-V2，将大模型的Embedding和每一层前都加上新的参数



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks) (b) P-tuning v2 (Frozen, most scales, most tasks)

Figure 2: From Lester et al. (2021) & P-tuning to P-tuning v2. Orange tokens (include h_0, h_i) refer to prompt embeddings we add; blue tokens are embeddings stored or computed by frozen pre-trained language models. Compared to Lester et al. (2021), P-tuning v2 adds trainable continuous prompts to inputs of every transformer layer independently (as prefix-tuning (Li and Liang, 2021) does). Additionally, P-tuning v2 removes verbalizers with LM head and returns to the traditional class labels with ordinary linear head to allow its task-universality.

Paper

[2103.10385\] GPT Understands, Too

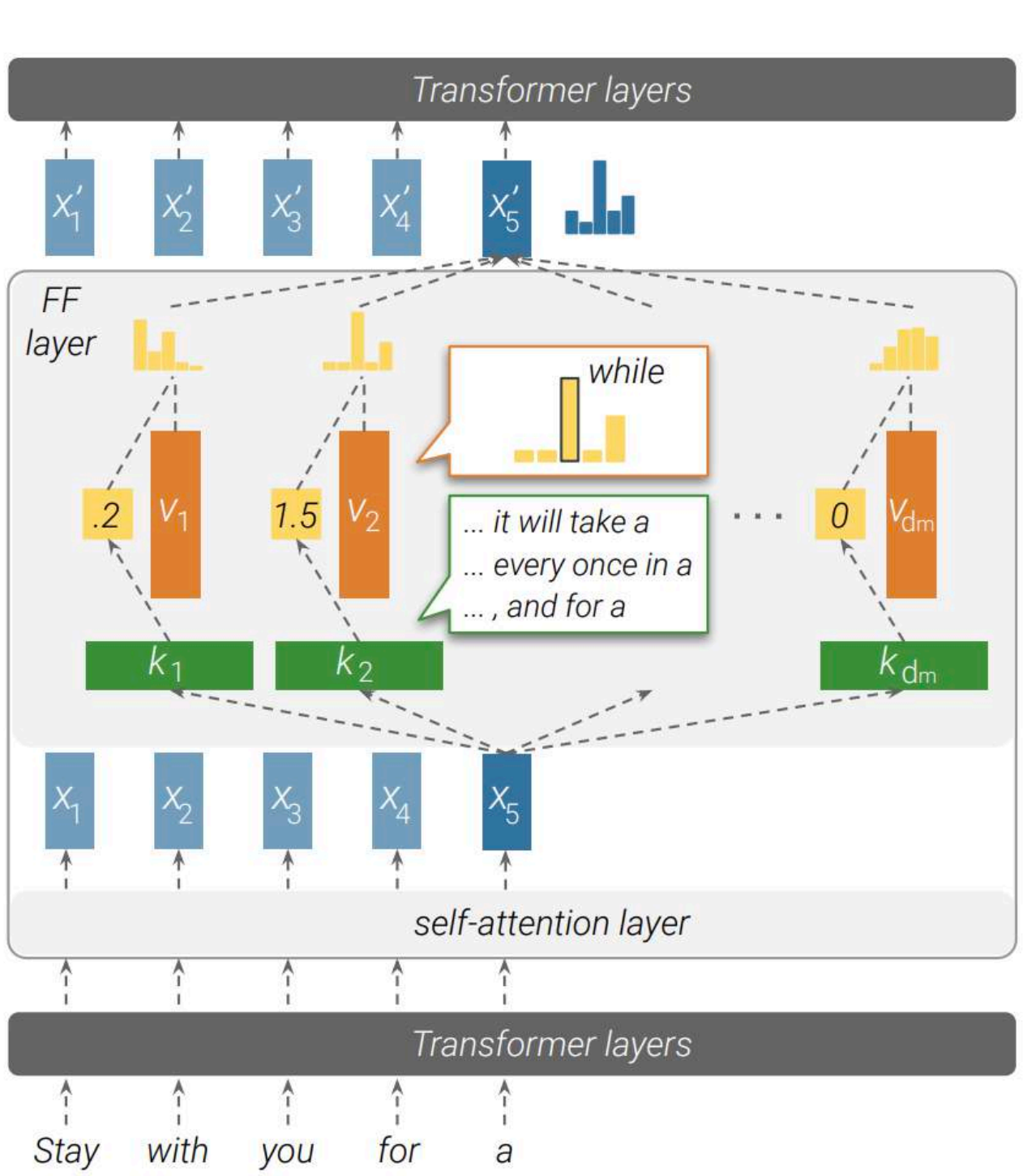
P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks

Xiao Liu[1,2*], **Kaixuan Ji**[1*], **Yicheng Fu**[1*], **Zhengxiao Du**[1,2], **Zhilin Yang**[1,2†], **Jie Tang**[1,2†]
[1]Tsinghua University, Beijing, China
[2]Beijing Academy of Artificial Intelligence (BAAI), Beijing, China

模型微调技术 Freeze

参数冻结监督微调

对原始模型部分参数进行冻结操作，仅训练部分参数，以达到在单卡或多卡，不进行TP或PP操作就可以对大模型进行训练



Freeze 微调方法原理

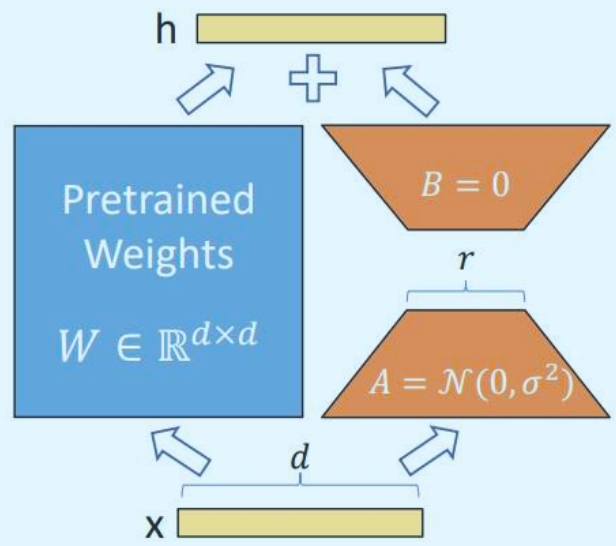
Low-Rank Adaptation of Large Language Models

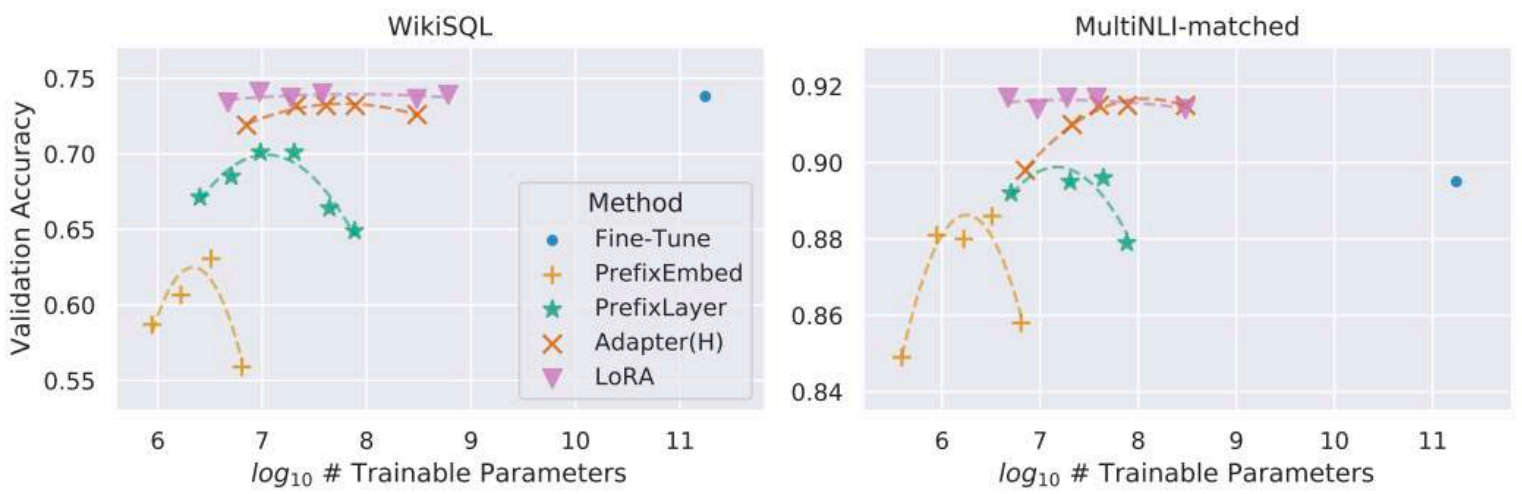
LoRA

大语言模型的低阶自适应

在 LLM 上对指定参数（权重矩阵）并行增加额外的低秩矩阵，并在模型训练过程中，仅训练额外增加的并行低秩矩阵的参数。当“秩值”远小于原始参数维度时，新增的低秩矩阵参数量也就很小。在下游任务 Tuning 时，仅须训练很小的参数，但能获取较好的表现结果

1. 在原始预训练语言模型（PLM）旁边增加一个旁路，做一个降维再升维的操作，来模拟所谓的内在秩。
2. 训练的时候固定 PLM 的参数，只训练降维矩阵 A 与升维矩阵 B。
3. 模型的输入输出维度不变，输出时将 BA 与 PLM 的参数叠加。
4. 用随机高斯分布初始化 A，用 0 矩阵初始化 B，保证训练的开始此旁路矩阵依然是 0 矩阵。





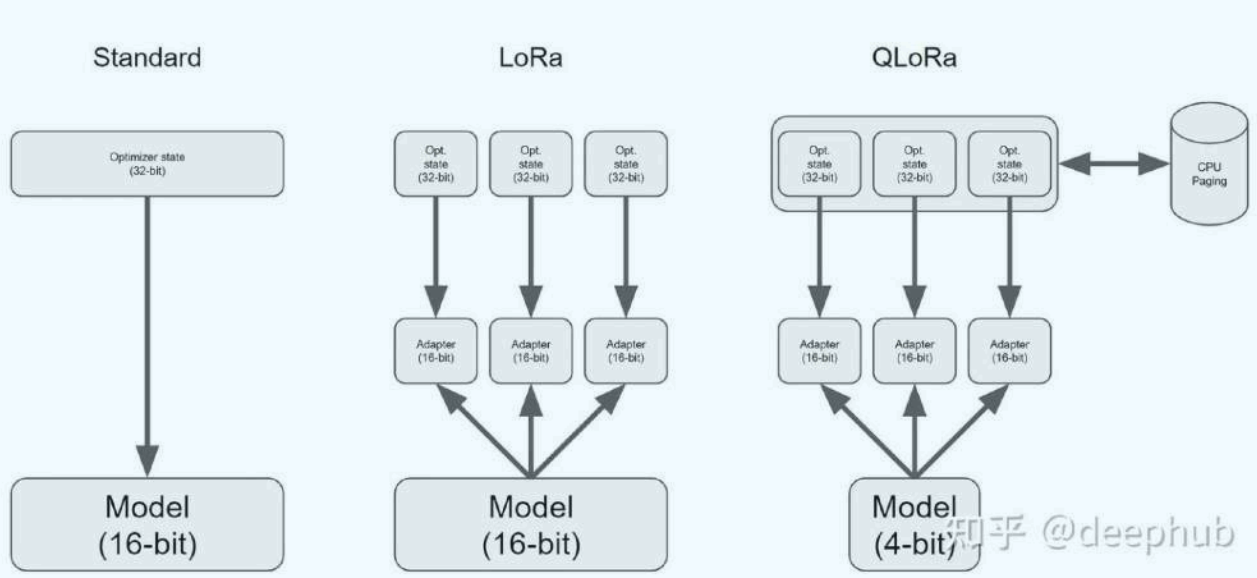
https://arxiv.org/pdf/2106.09685.pdf

https://github.com/microsoft/LoRA

Quantized LLMs with Low-Rank Adapters

QLoRa

Efficient Fine-tuning of Quantized LLMs



4 位 NormalFloat 量化

一种改进量化的方法。确保每个量化仓中有相同数量的值。这避免了计算问题和异常值的错误。

双量化

对量化常量再次量化以节省额外内存的过程

统一内存分页

它依赖于NVIDIA统一内存管理，自动处理CPU和GPU之间的页到页传输。它可以保证GPU处理无错，特别是在GPU可能耗尽内存的情况下。

使用低秩适配器的量化LLM

通过降低内存使用，实现在单个GPU上对大型语言模型进行微调，并取得了先进的性能结果

https://github.com/artidoro/qlora

https://arxiv.org/abs/2305.14314

更多访问 www.xRunda.com

模型微调技术

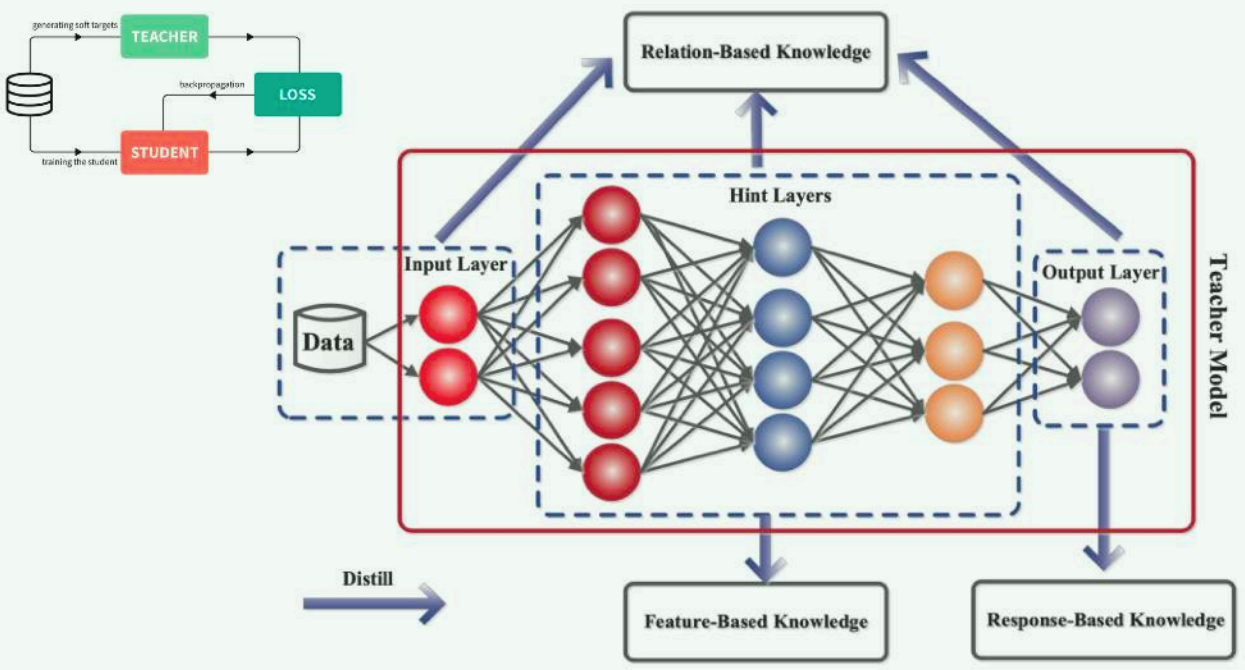
Knowledge Distillation

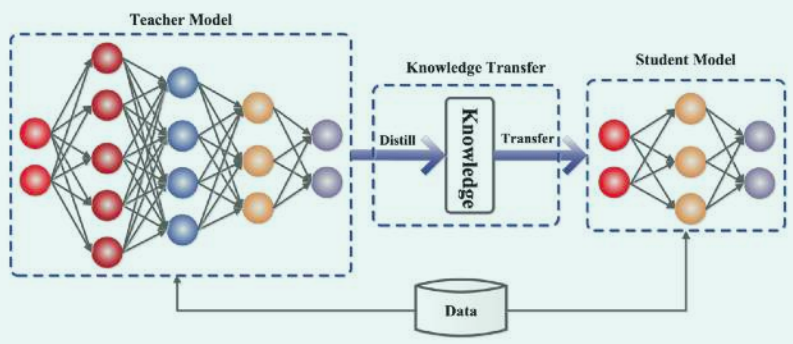
知识蒸馏

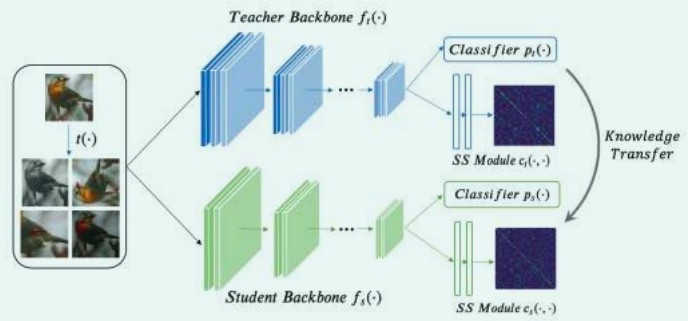
模型压缩技术

将教师模型的知识转移到学生模型中的方法

使用一个已经过大量训练的较大模型作为教师模型，将其知识转移到较小的学生模型中







https://blog.51cto.com/u_15668366/6148835

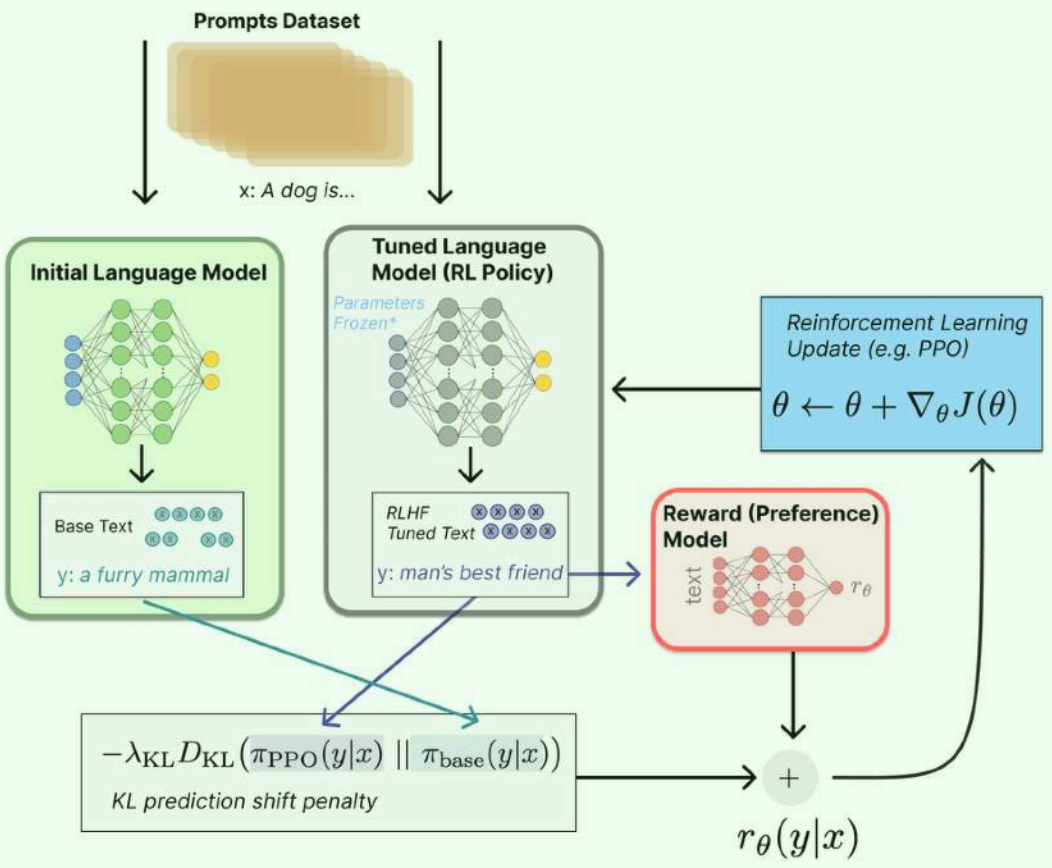
模型微调技术 | Reinforcement Learning

人在回路 强化学习方式微调

RL

RLHF, Reinforcement Learning from Human Feedback

基于人类反馈的强化学习



1. 对同一输入的多个生成结果进行人工排序，获得包含人类偏好反馈的标注数据，从而训练出一个奖励模型（Reward Model）
2. 在强化学习的过程中，奖励模型将对大语言模型的多个生成结果的排序进行判定。
3. 最终，强化学习通过更新大模型的参数，使得输出结果符合奖励模型的判定要求。

→ 暂无OpenAI之外的成功经验

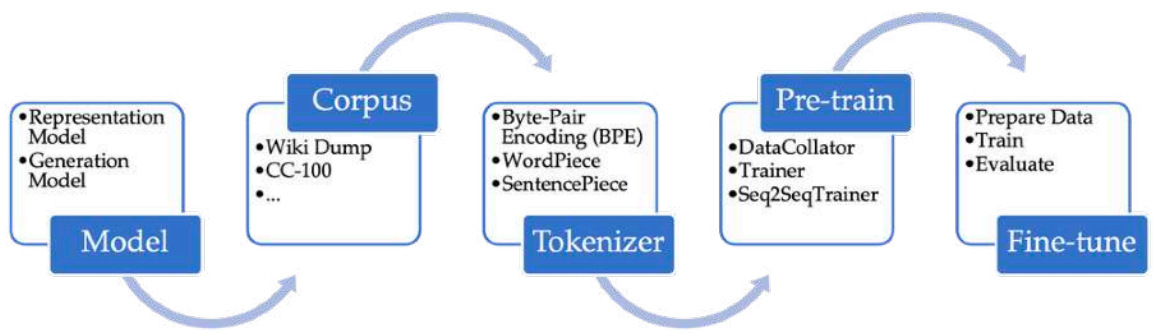
ReinforceFineTune

一种基于强化学习的微调方法，可以通过人类反馈来优化预训练模型的目标函数，并生成更自然和更有趣的对话

Pre-training

Pre-trained
在大规模无标注文本上训练语言模型的过程或结果
以捕捉语言的通用知识和规律

预训练

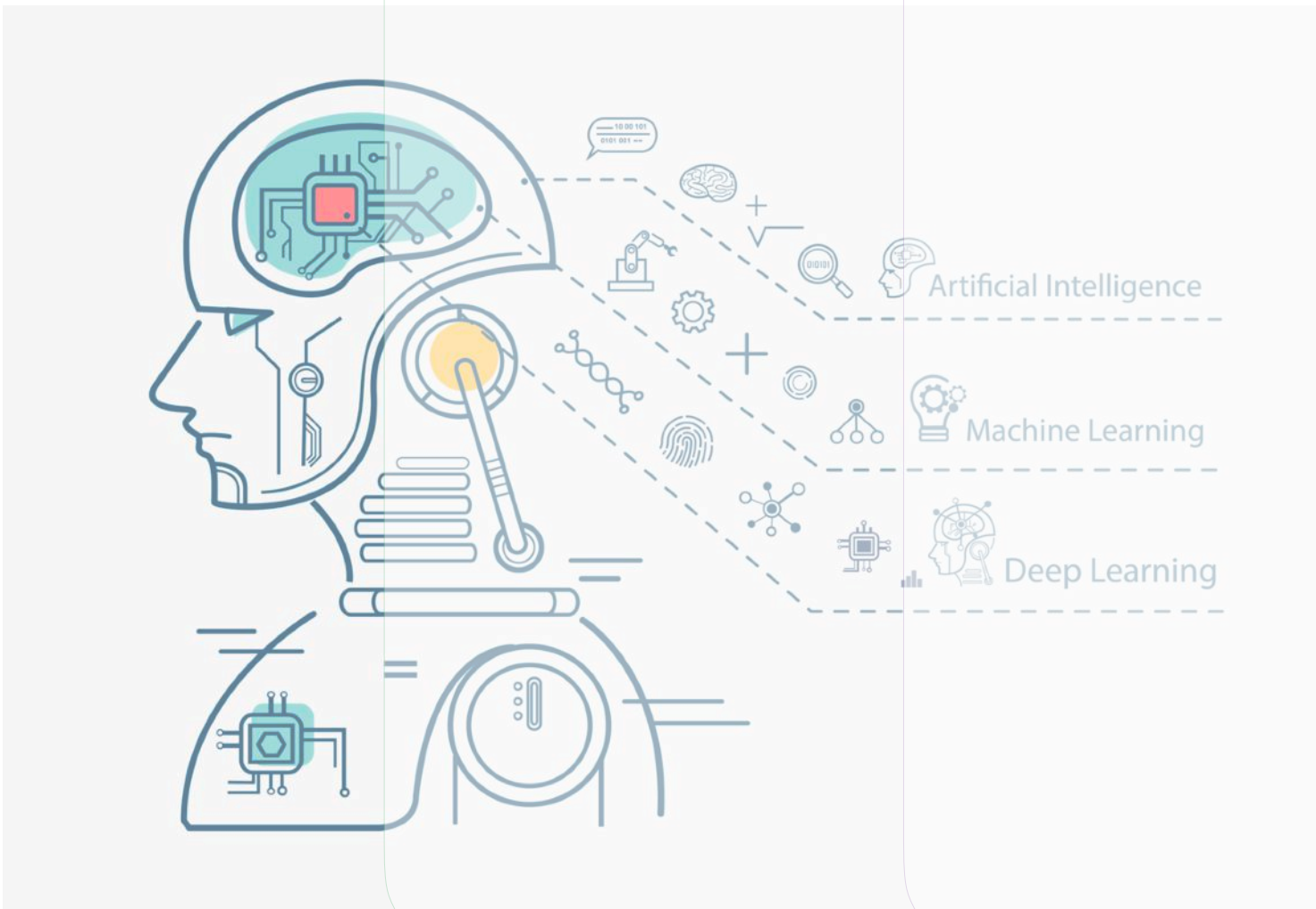


预训练（共建）行业大模型

大规模行业知识

海量无结构数据

行业数据挖掘



企业智化方案

企业智化 | 通用企业方案 工业企业

安全生产智能应用平台

产品功能台

解决方案

解决方案

企业智化 | 平台支持资源

阿里企业大模型

华为盘古大模型

https://www.huaweicloud.com/product/pangu.html

企业智化 | 法律行业方案

LaWGPT

基于中文法律知识的大语言模型

该系列模型在通用中文基座模型（如 Chinese-LLaMA、ChatGLM 等）的基础上扩充法律领域专有词表、大规模中文法律语料预训练，增强了大模型在法律领域的基础语义理解能力。在此基础上，构造法律领域对话问答数据集、中国司法考试数据集进行指令精调，提升了模型对法律内容的理解和执行能力。

LaWGPT project directory structure

```
LaWGPT
├── assets        # Static resources
├── resources     # Project resources
├── models        # Base models and Lora weights
│   ├── base_models
│   └── lora_weights
├── outputs       # Fine-tuned instruction outputs
├── data          # Experimental data
├── scripts       # Script directory
│   ├── finetune.sh    # Instruction fine-tuning script
│   └── webui.sh       # Service startup script
├── templates     # Prompt templates
├── tools         # Toolkits
├── utils
├── train_clm.py      # Secondary training
├── finetune.py       # Instruction fine-tuning
├── webui.py          # Service startup
├── README.md
└── requirements.txt
```

https://github.com/pengxiao-song/LaWGPT

ChatLaw

ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases

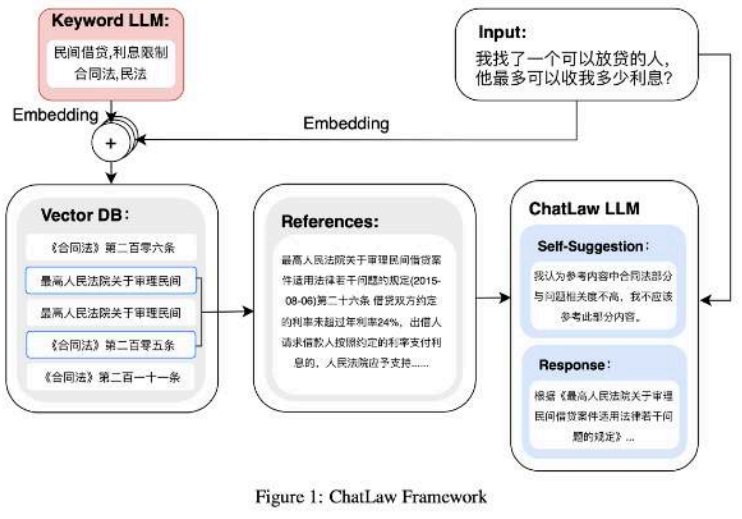


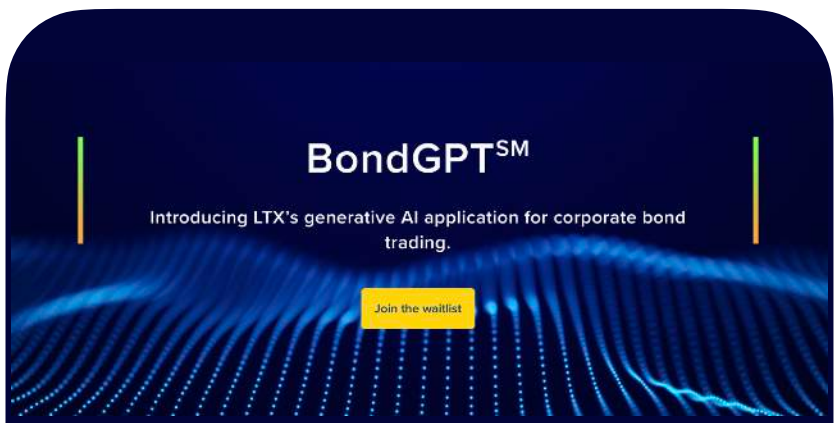
Figure 1: ChatLaw Framework

ChatLaw法律大模型目前开源的仅供学术参考的版本底座为姜子牙-13B、Anima-33B，我们使用大量法律新闻、法律论坛、法条、司法解释、法律咨询、法考题、判决文书等原始文本来构造对话数据。

由北京大学深圳信息工程学院完成，指导教师为袁粒

https://arxiv.org/pdf/2306.16092.pdf
https://github.com/PKU-YuanGroup/ChatLaw

企业智化 | 金融行业方案



BondGPT

Broadridge推出BondGPT，服务债券市场

度小满

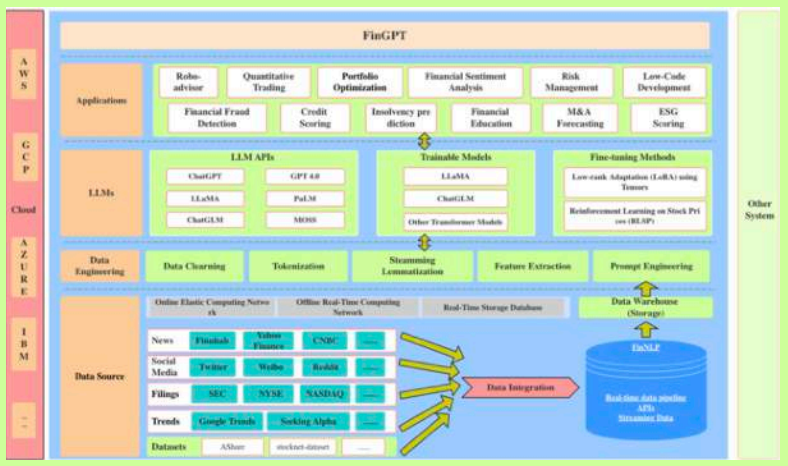
轩辕（BLOOM）

彭博社

BloombergGPT

- 500亿参数、基于BLOOM模型的LLM，过程中采用了一种兼具通用能力和特定领域的方法
- 基于Bloomberg 40年来积累的数据构造了目前最大的金融领域数据集
- 在金融领域取得好效果的同时，并没有以牺牲模型通用能力为代价

- 论文解读：
- https://zhuanlan.zhihu.com/p/619444812



FinGPT

FinGPT: Open-Source Financial Large Language Models
https://arxiv.org/abs/2306.06031

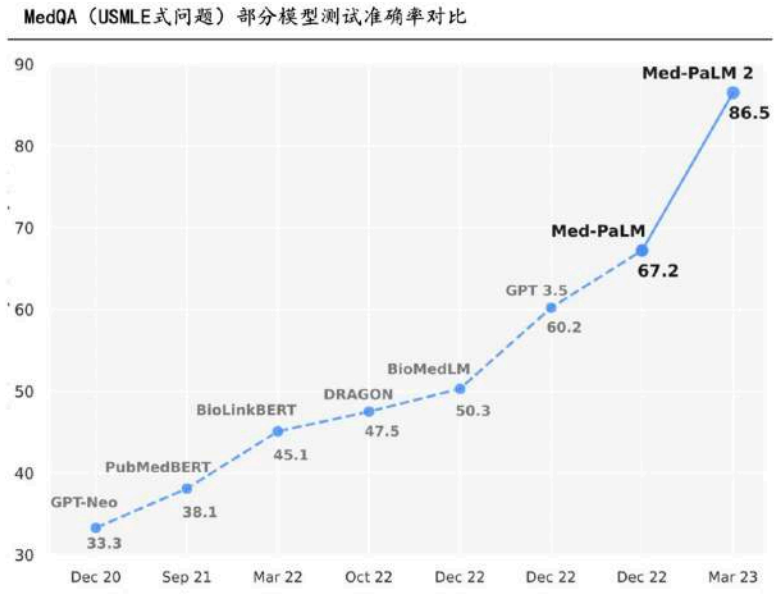
企业智化 | 医药行业方案

医联 / MedGPT
为治愈而生

MedGPT

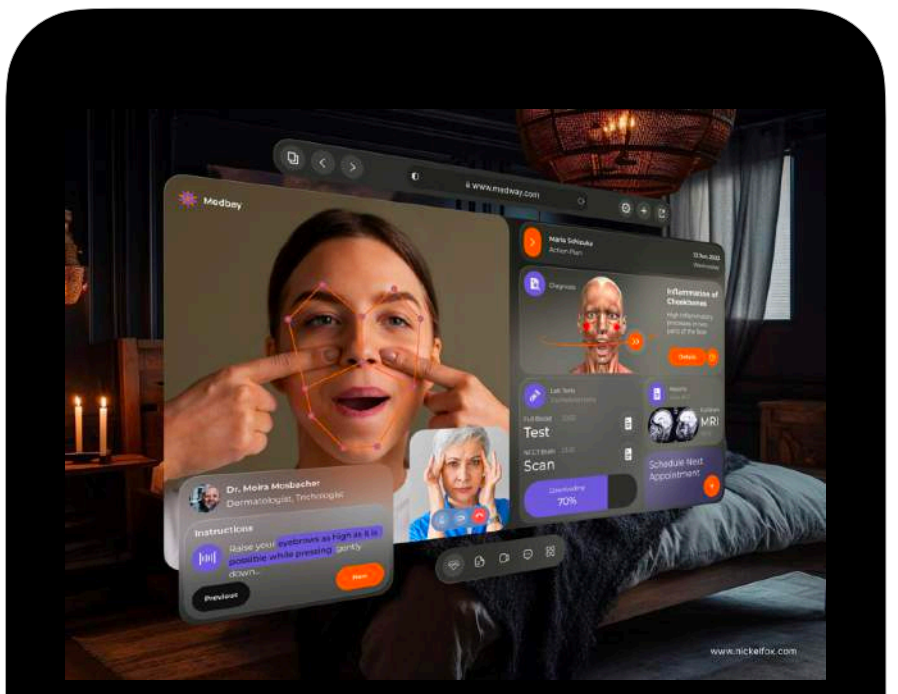
医联发布国内首款医疗大语言模型MedGPT

- 参数100B规模，预训练阶段使用了超过20亿的医学文本数据，微调训练阶段使用了800万条结构化临床诊疗数据基于Bloomberg 40年来积累的数据构造了目前最大的金融领域数据集
- 通过收集足够信息并做出符合医学的决策，以“治愈”为目的而进行人机交互。将大模型技术与工程调优技术以及医学一致性校验技术相结合
- 模型微调训练阶段采用100+真实医生参与的RLHF（Reinforcement Learning from Human Feedback) 监督微调
- 通过多轮问诊引导患者收集足够的诊断决策因子之后再进入到诊断环节，从而保证准确性
- 建立了基于专家评议的AI诊疗准确性与真实世界医生对标测试机制，不断将AI与真实诊疗场景对齐，最终实现准确诊断



Med-PaLM 2

https://www.nature.com/articles/s41586-023-06291-2

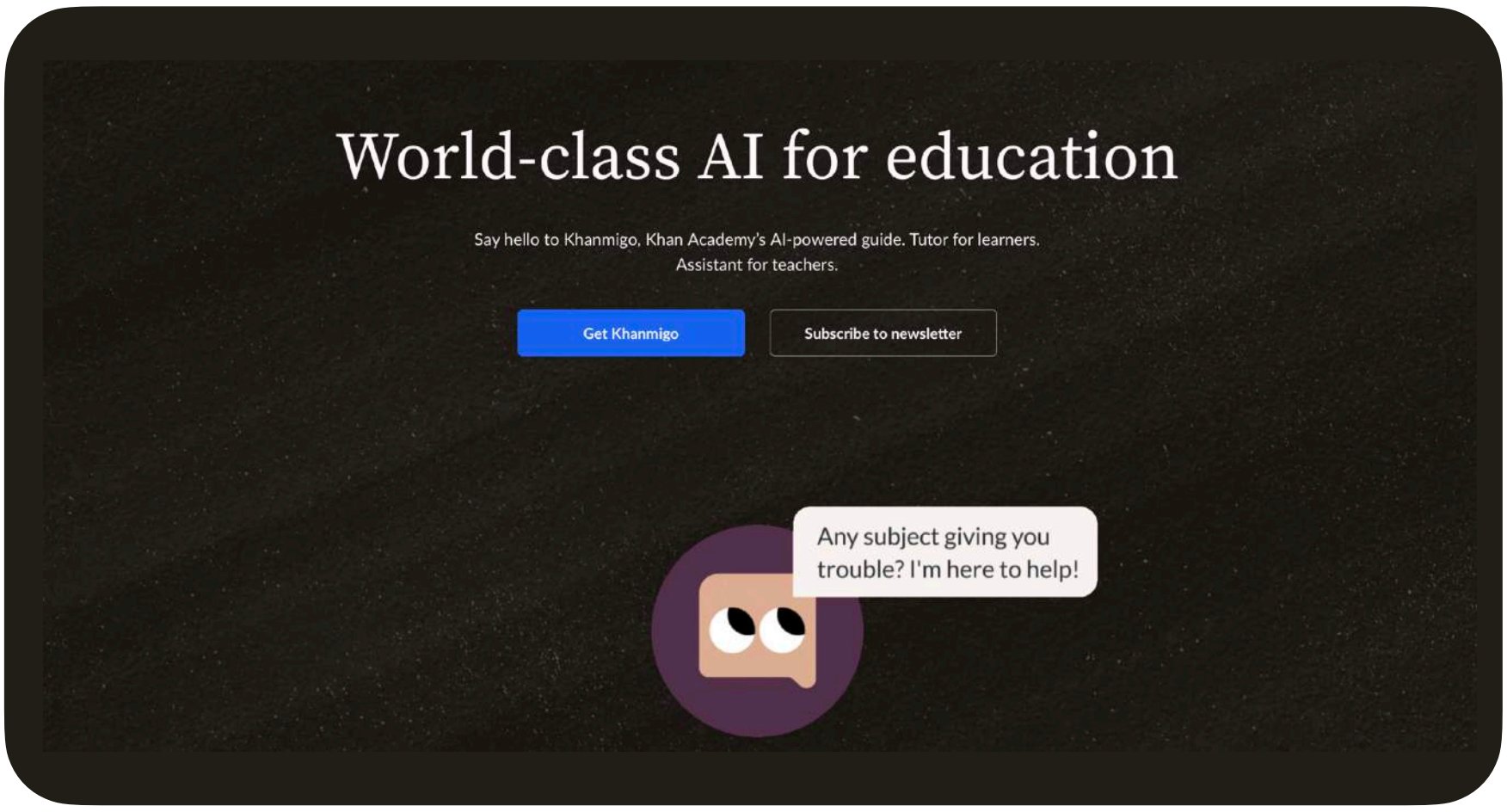


AI+医疗研究报告

https://pdf.dfcfw.com/pdf/H301_AP202306281591795603_1.pdf

企业智化

教育行业方案



可汗学院

Khan Labs: World-class AI for education
https://www.khanacademy.org/khan-labs

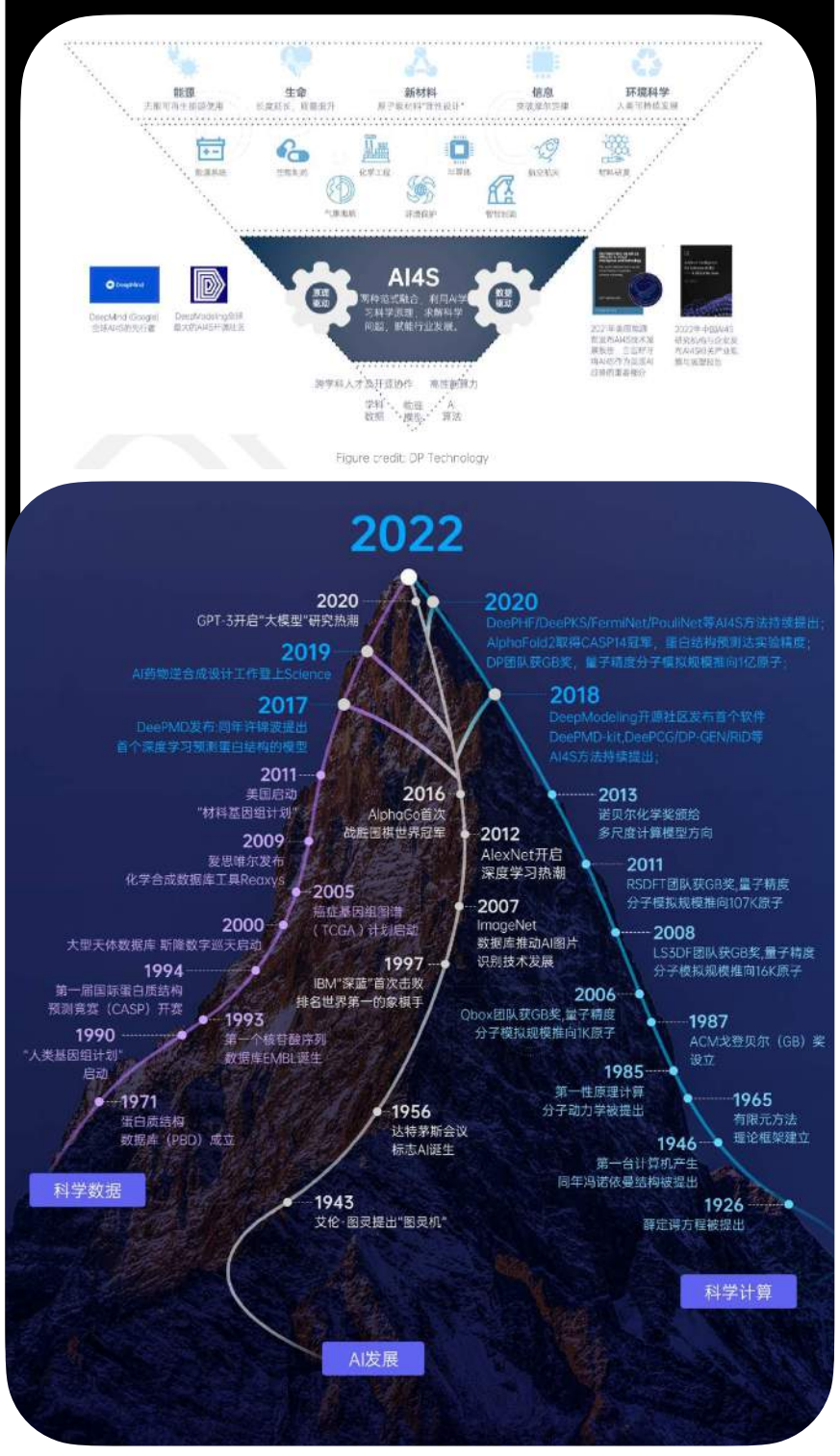
企业智化 | 汽车行业方案

理想汽车 MindGPT　多模态感知与交互　知识增强　模型和API控制器　Memory

吉利+百度　知识增强的汽车行业大模型

企业智化 | 科学领域方案 AI For Science

络绎科学



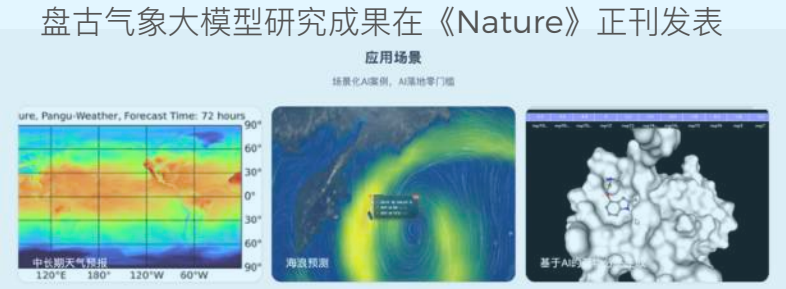
Google DeepMind

- 2010年创立，使用机器学习解决传统计算机难以处理的问题，如在围棋和蛋白质折叠等领域超越人类。
- 2014年被收购成为谷歌旗下子公司。
- 2023年4月更名 Google DeepMind，宣布推出通用 AI GATO 模型
- https://www.deepmind.com/
- https://www.livescience.com/what-is-deepmind.
- https://zh.wikipedia.org/wiki/DeepMind.
- 技术已应用于现实世界场/ 景医疗、能源和教育等

盘古科学计算大模型

盘古气象大模型

盘古气象大模型研究成果在《Nature》正刊发表

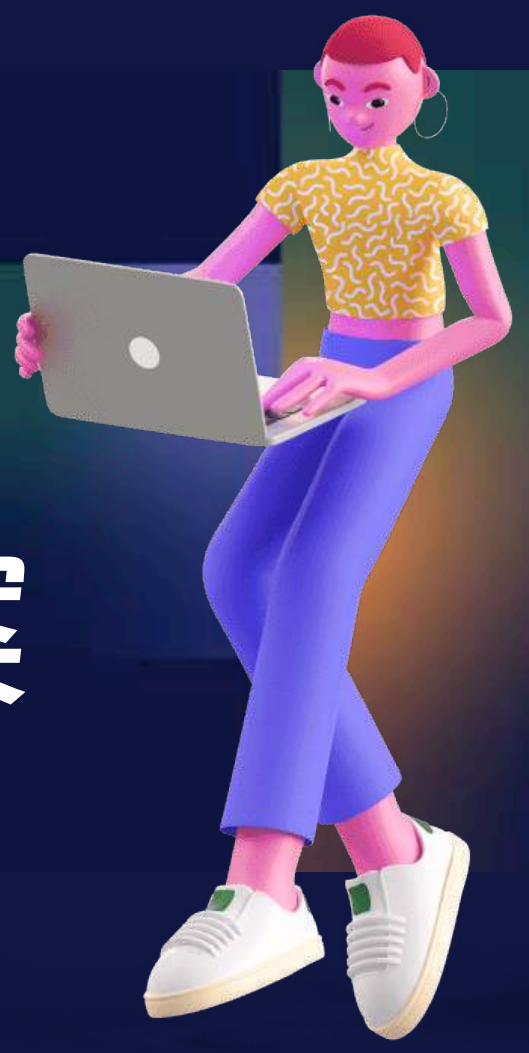


盘古药物分子大模型

赋能药物研发的全链条任务，旨在帮助医药企业机构显著提升药物研发的效率。囊括了大规模药物虚拟筛选、分子动力学模拟等传统CADD药物研发软件，基于AIDD的蛋白质结构预测、分子属性预测等服务。助力新靶标药物的发现，让医药公司搭乘AI辅助药物研发的快车

医疗智能体 EIHealth

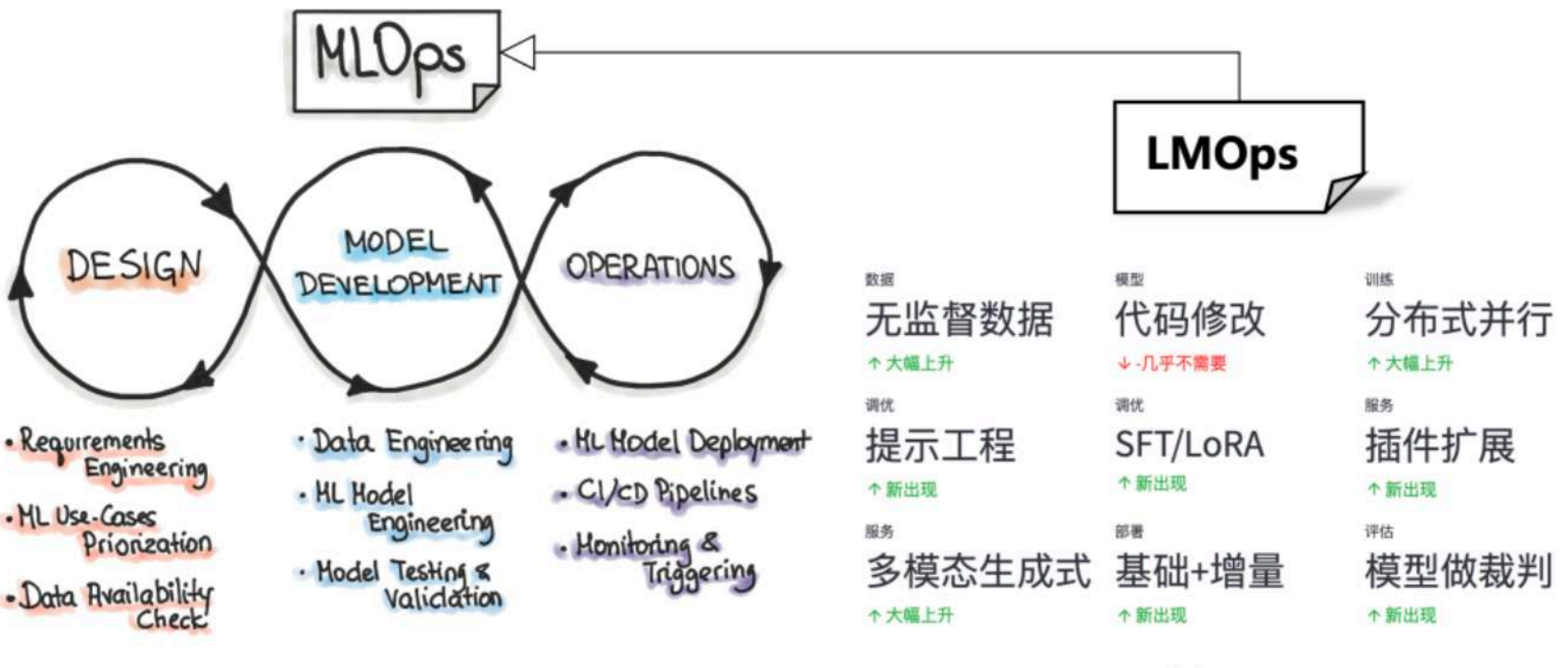
应用场景



xMaaS 集成方案

xMaaS
集成方案

LMOps 方案



方案框架

继承了 MLOps 整体的框架和机器学习的全生命周期等主要环节，并且针对大模型的变化进行了微调适配

更多访问 www.xRunda.com

Landscape

xMaaS 集成方案 | xMaaS 服务

数据存储 1
向量化和存储

数据准备 2
对大规模未标注的数据进行加工和配比

Cleaning
首先是对特殊字符的一些清除，如特殊标点清除等，替换部分异常文本

Filtering
删除低质量文档。建立低质文档的统计指标，超过某个阈值就进行删除。或通过定制的分类模型对文档质量进行自动分类。

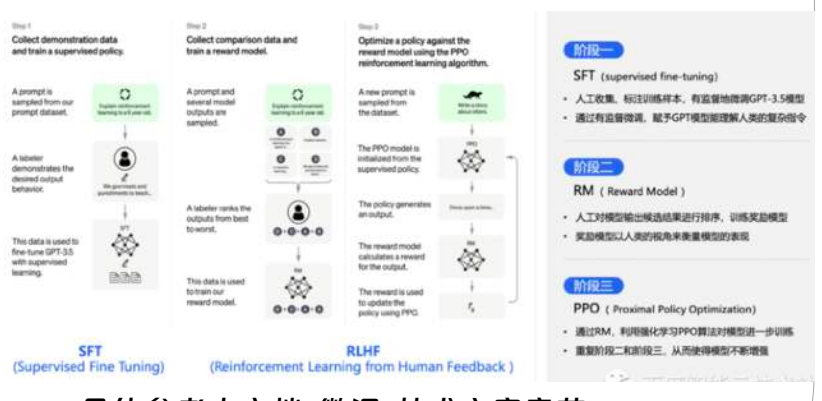
Deduplication
文档去重。针对文档中的句子和段落等进行文档内的内容去重；针对两个内容重复阈值较高的相似文档，可以进行跨文档去重。

Privacy Reduction
去除隐私数据。使用基于规则的正则表达式的方法检测个人信息，来进行隐私数据的脱敏。

Tokenization
建立词表（Tokenize 的过程）。常用 Sentence Piece 等方法。当将原始语料加工成 token，并建立 token_id 后，再喂给大模型进行训练或者推理。

更多访问 www.xRunda.com

模型训练 LMOps 训练环节 3



具体参考本文档 微调 技术方案章节

模型评估 LMOps 评估环节 4

评估标准 多维度 | 评估数据集 验证数据集 | 评估任务模板 Prompt Engineering | 评估工具

模型运维 6

安全环节 | 监控 | 管理

模型部署 5

推理环节

部署环节

xMaaS 集成方案 MaaS 平台

阿里魔搭社区

ModelScope 魔搭社区

阿里巴巴 MaaS

百度文心千帆

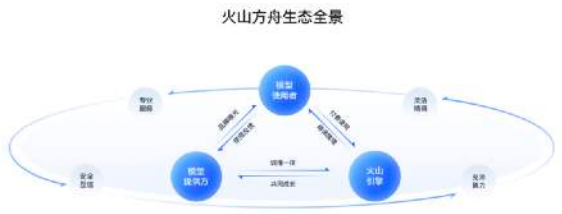
https://cloud.baidu.com/product/wenxinworkshop

腾讯云TI平台

腾讯云 MaaS 全景图

https://cloud.tencent.com/product/ti

字节跳动火山方舟



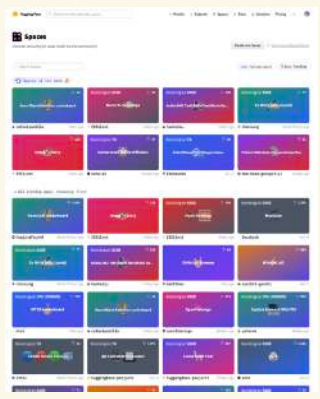
火山方舟生态全景 https://www.volcengine.com/product/ark

商汤大装置 SenseCore

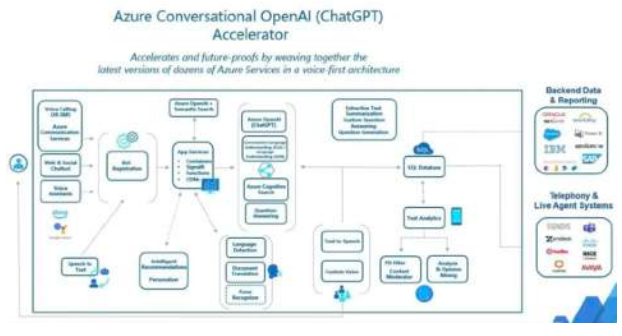
AIaaS https://www.sensecore.cn/about

HuggingFace

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard



Microsoft Azure



More

华为云盘古大模型
https://www.huaweicloud.com/product/pangu.html
科大讯飞星火认知大模型
https://xinghuo.xfyun.cn/
360智脑大模型
https://ai.360.cn/
智谱 AI GLM

更多访问 www.xRunda.com

D 产品案例 Product Case

最新产品

Recent Products

最新产品 | 即摘 xGeekSum ✨

即摘 GeekSum →

TO C / B / G

高效阅读助手
提效 阅读 智能 助手

xRunda AI Lab 推出基于大模型技术的创新产品，即摘 GeekSum，旨在提升阅读效率，为您打造高效的信息订阅工具。

产品亮点

智能快摘
通过 AI 可快速生成文章摘要

广泛支持
支持大部分主流媒体文章内容，包括微信公众号、知乎、即刻、少数派、虎嗅、雪球等

多端智读
小程序+H5 多端触达

操作便捷
粘贴链接+微信公众号文章一键摘要

展望 增值服务/未来展望/持续迭代

智慧收藏夹
保存历史记录，随时回顾

音频生成
文章内容随时听取

AI 问答
巩固理解，深度探究

阅读笔记
捕捉灵感，记录思考

智能分类
高效管理文章

专属分享
与他人分享阅读心得

最新产品

即听 xGPTing

Generative Podcast Transforming

即听 xGPTing →

TO C / B / G

开启智能化播客工作流体验

用一种很 COOL 的方式打开文档，通过 AI 实现高效地“读”文档，自由地“问”文档，便捷地“听”文档，轻松的“分享”文档。

产品亮点

智能快摘
通过 AI 可快速生成文章摘要

广泛支持
支持大部分主流媒体文章内容，包括微信公众号、知乎、即刻、少数派、虎嗅、雪球等

多端智读
小程序+H5 多端触达

操作便捷
粘贴链接+微信公众号文章一键摘要

展望

增值服务/未来展望/持续迭代

智慧收藏夹
保存历史记录，随时回顾

音频生成
文章内容随时听取

AI 问答
巩固理解，深度探究

阅读笔记
捕捉灵感，记录思考

智能分类
高效管理文章

专属分享
与他人分享阅读心得

最新产品

即试 xGPTest

Generative Product Test

即试 xGPTest →

TO C / B / G

Prompt 集市 – 追求极智的 Prompt 创作体验

用即试 xGPTest 创建和管理您的 Prompt，积累沉淀每个不经意的优雅创意。

产品亮点

写试一体	在即试中轻松储存、编辑和运行您的提示词。
数据安全	轻松组织您的提示，每个项目独立管理。
创作笔记	记录Prompt 创作过程中的收获与反思，亮点与不足。
版本追溯	每次运行提示词都有版本记录，方便您随时回退到以前版本。
动态数据	即试 xGPTest 支持插入多个变量到您的提示词中。
多模型	支持不同类型的模型，包括 GLM、讯飞星火、OpenAI、Claude、Azure OpenAI。

最新产品

即答 xChatDA

Chat Distribute Agent

即答 xChatDA →

TO C / B / G

你的智能中枢
xRunda 一问多答解决方案

即答是一款集服务托管、数据管理、隐私保护、高效调度、多源比较、内容检索与聚合于一体的智能问答平台。

产品亮点

AI多元交互
提供与多种AI模型的交互，包括聊天、知识问答、创意生成等，满足不同的需求。

高效大模型托管
无论是企业还是个人用户，均可便捷地托管大型AI模型，享受无与伦比的性能和稳定性。

数据管理
基于上传数据集检索历史及返回内容等打造个性化知识库。

隐私保护
采用先进的安全技术，确保用户和企业的数据隐私得到全面保护。

算力优化与调度
采用独特的算法，使用户在多个应用场景下都能获得最高性价比的算力消耗。

智能内容检索与聚合
提供高效的内容检索，聚合多个来源内容，并通过先进的算法提炼综述、去伪存真。

一问多答比较系统
每个用户提问均返回多个LLM结果，方便用户比较参考，并自动提供综述，增强决策支持。

多端支持
多端接入，使用户能够随时随地访问服务。

投资与市场前景

即答汇聚了行业最佳实践和创新技术，具备强大的市场竞争力。我们的商业模式旨在满足日益增长的信息处理和决策支持需求，预计将在企业和个人用户中迅速获得广泛认可。

结语

让我们一同踏上这场智能问答的探索之旅，即答将成为您的得力伙伴，共同开启未来的崭新篇章

最新产品 | 即调 xTune

即调 xTune →

TO C / B / G

大模型微调的极智新选
xRunda 一通多调解决方案

前沿的大模型微调平台，将高效微调技术与一站式服务结合，为开发者和企业提供便捷与效能。
从数据管理到自动化模型定制，再到云端部署，我们致力于让大模型的微调更加智能、灵活和高效。

适用人群　AI开发者　企业与机构

主要功能

多模型支持
支持多种预训练的大型模型

一站式微调服务
提供数据管理、自动化模型定制、云端部署的一体化解决方案，助您轻松掌握微调全流程。

交互式在线平台
支持多种语言和领域的微调选项，多样的输出格式和参数设置，让微调更加直观和便捷。

批量微调与比较
基于相同数据集和微调目标，一次性微调多个底层模型，横向比较效果，直观查看成本与性能，进一步提供调整建议。

产品亮点

批量微调与横向比较
这是我们的核心竞争力所在，用户可以在同一平台上对多个模型进行横向对比和批量微调，节省时间，提高效率，并能得到更加精确的调整策略。

灵活与易用
无论是初学者还是专家，我们的平台都能为您提供友好的用户体验，使微调变得触手可及。

投资与市场前景

随着AI的发展，大模型微调已经成为了关键技术。作为市场上首款集成了这些功能的微调平台，具有巨大的市场潜力和增长空间。

结语

即调是您微调大模型的最佳选择，是未来智能化微调技术的领航者。

最新产品 | 即画 xDraw

即画 xDraw →

TO C / B / G

喜AI图梦想服务器
重见想象的AI绘画平台

即画 xDraw 是一款前沿的AI绘画平台，将先进的人工智能技术与艺术创造力完美融合。通过精准的算法解读，您的文字描述将被转化为高质量的图片作品。无论是一些关键词还是一段精彩的场景描述，"即画"能在瞬息之间为您呈现出独一无二的视觉艺术。

主要功能

文字生图
只需输入关键词或描述文字，几秒内即可生成属于您自己的图片，版权完全归属于您。

丰富的绘画风格选择
提供丰富的绘画风格，从古典到现代，从抽象到写实，满足您多样化的审美需求。

艺术家参考灵感
多位世界著名艺术家的作品参考，让您的作品充满艺术气息，创造出无与伦比的视觉效果。

批量生成与多尺寸选择
支持批量生成多图，提供多种尺寸和高分辨率的选择，让每一幅作品都适合应用场景。

快速尝试与调整
提供的灵活工具使您可以快速尝试和调整图片，无论专业设计师还是新手用户都能轻松获得满意的作品。

产品亮点

批量微调与横向比较：
这是我们的核心竞争力所在，用户可以在同一平台上对多个模型进行横向对比和批量微调，节省时间，提高效率，并能得到更加精确的调整策略。

灵活与易用：
无论是初学者还是专家，我们的平台都能为您提供友好的用户体验，使微调变得触手可及。

创新与领先

即画 xDraw 不仅是一款工具，更是一场想象与科技的无限碰撞。我们的AI算法不仅将文字描述转化为图像，还能捕捉文字背后的情感和氛围，为您的创作赋予生命和灵魂

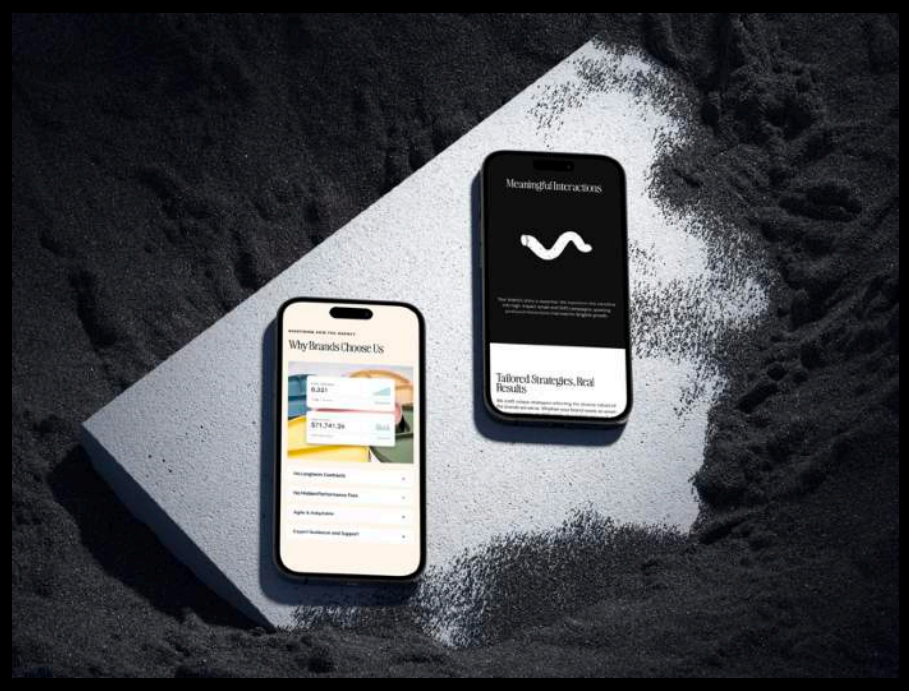
结语

在"即画 xDraw"的世界里，每一个文字都可以绽放成视觉的花朵，每一个想法都可以化身为绚丽的画卷。我们诚邀您加入这场视觉与文字的盛宴，用"即画 xDraw"描绘您心中的世界！

实项展示

Project Presentation

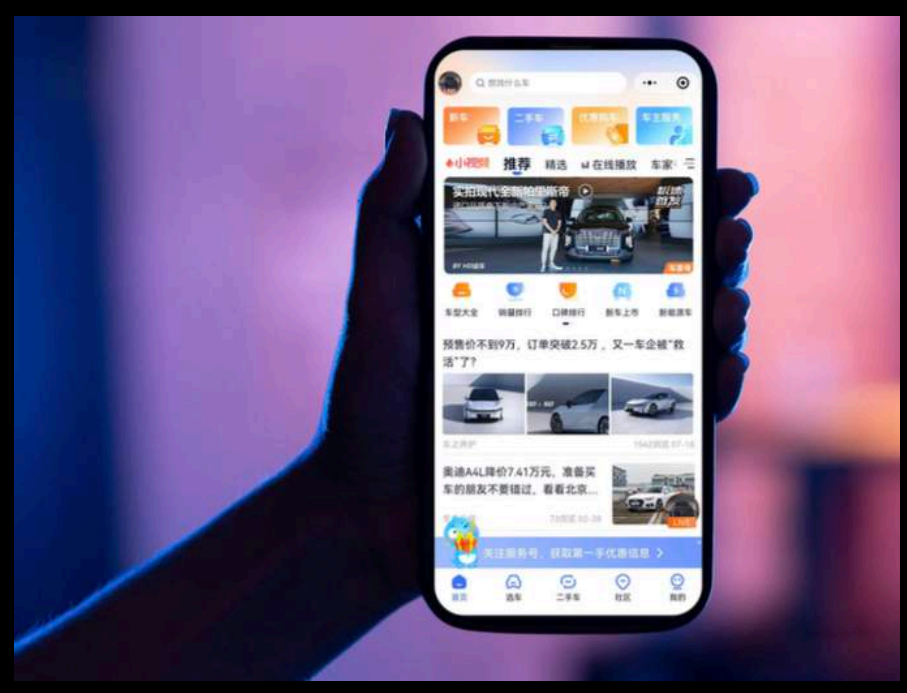
实项展示 | 行业服务



保险公司海报协作平台

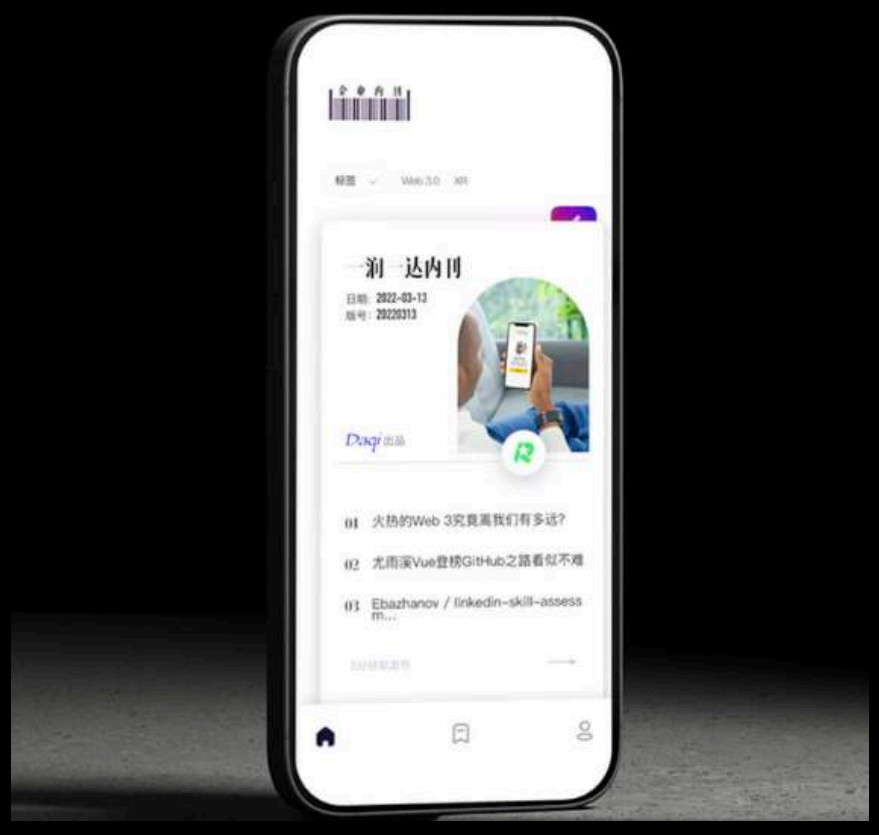
快手2022AIGC

络绎科学 AI4S

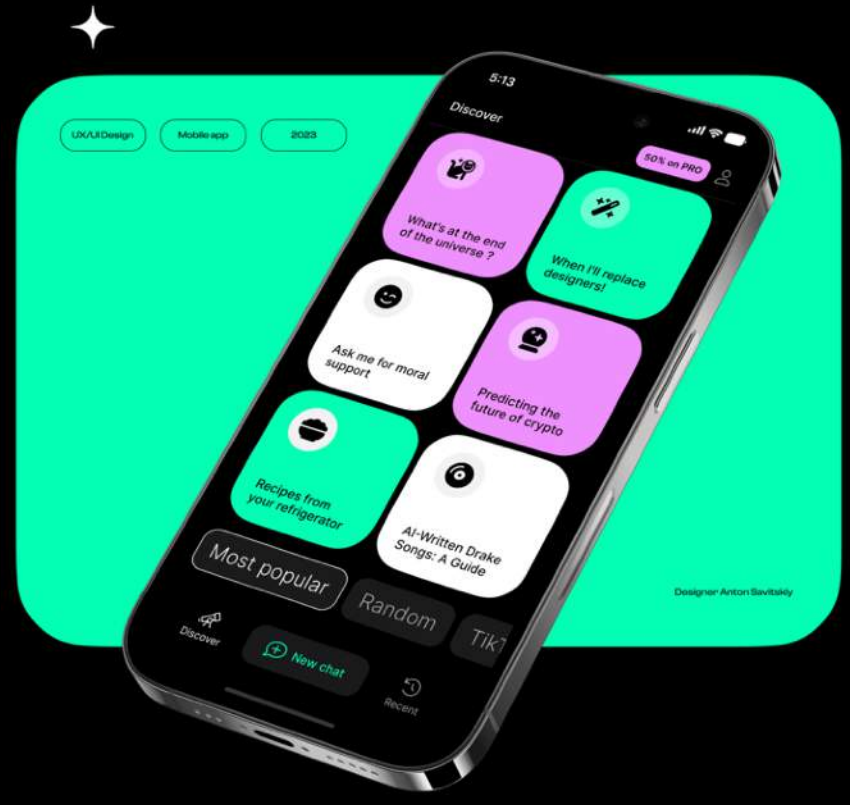


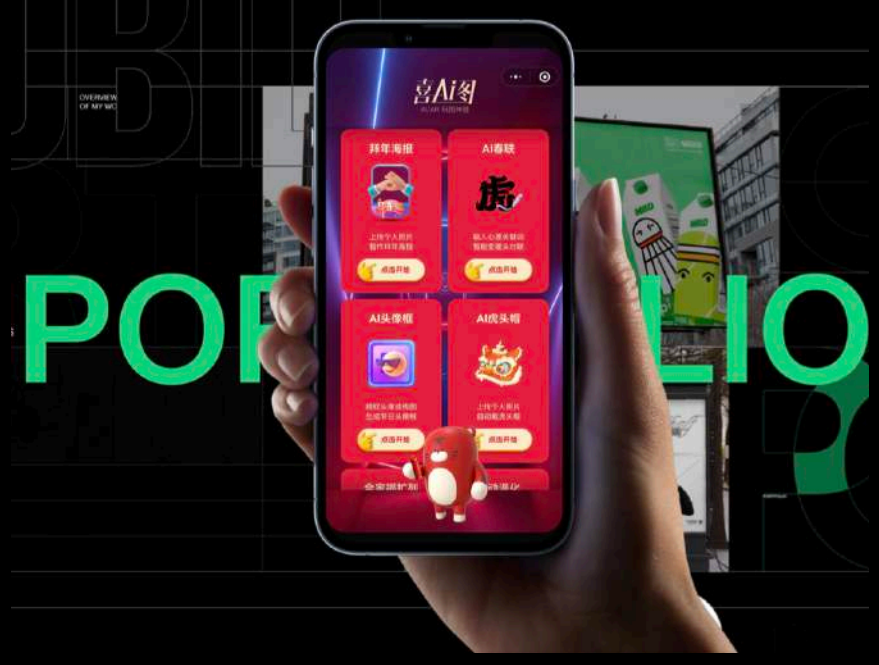
汽车之家

实项展示 | 孵化项目



企内刊



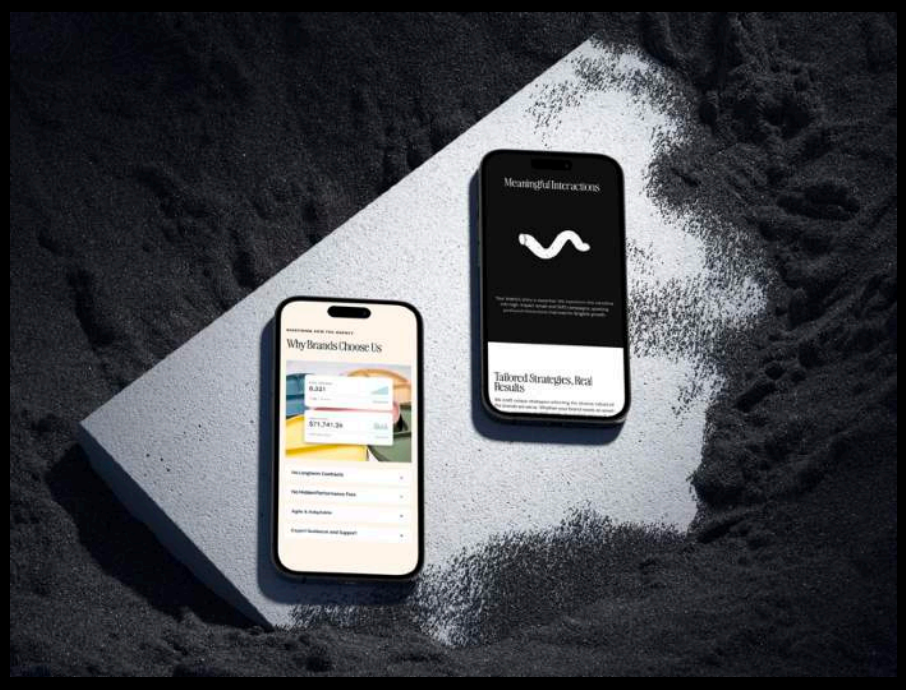


喜AI图

AIGA Town

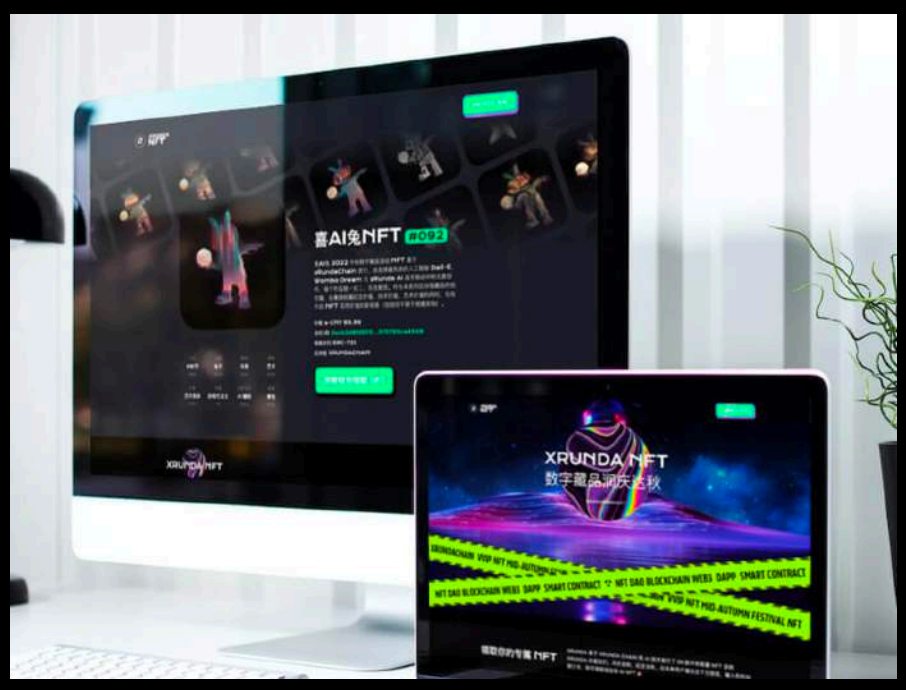
实项展示

实验项目

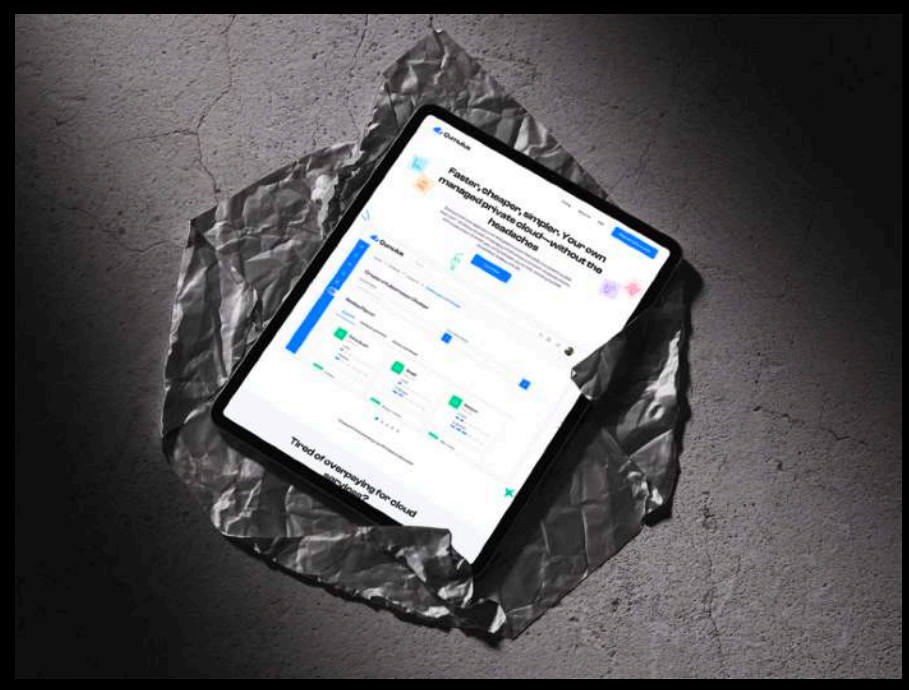


车型口碑 ChatBot

AIGC多模态少儿绘本生产线



LoRA微调画家风格模型



个人声音定制模型训练程序

www.xRunda.com

E

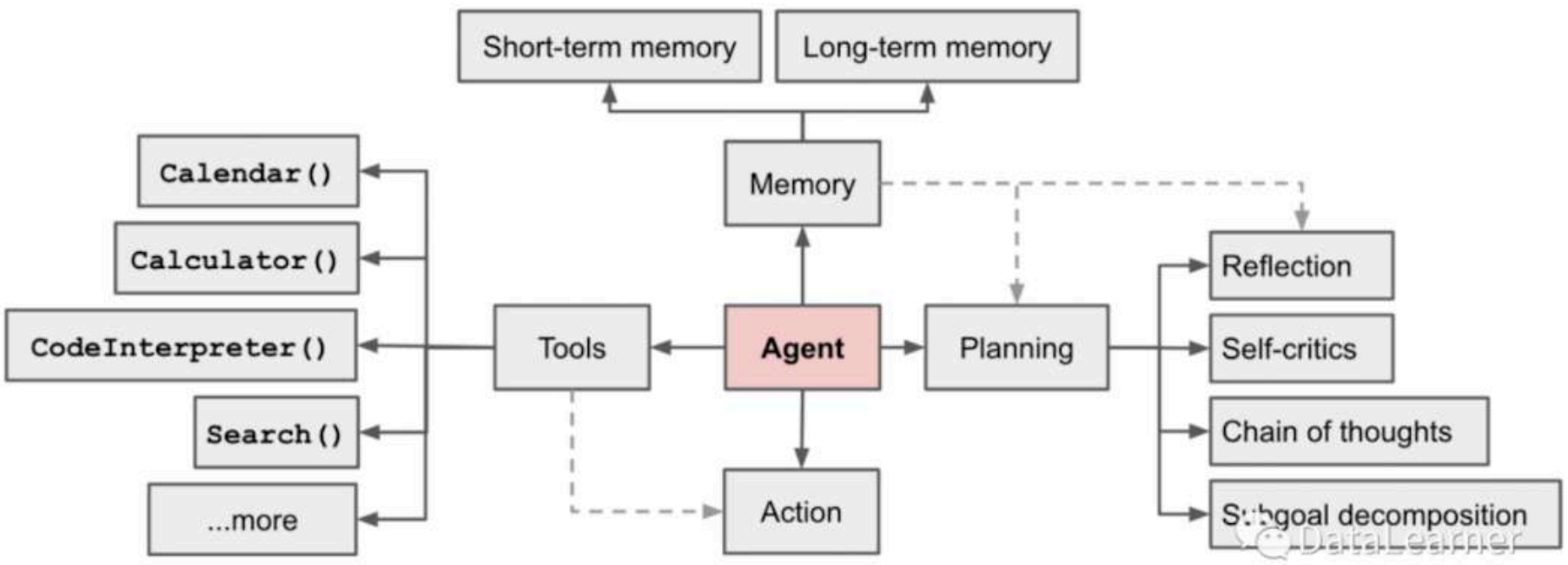
前沿研究
Frontier Research

Agent

Agent

Agent 概述

Agent: LLM 赋能的自主智能体系统概览



关键组件

规划 Planning

子目标和分解：
智能体将大型任务分解为更小、可管理的子目标，从而高效处理复杂的任务；

反思和完善：
智能体可以对过去的行为展开自我批评和自我反思，从错误中吸取教训，并针对未来的步骤进行完善，提高最终结果的质量。

记忆 Memory

短期记忆：
所有的上下文学习（参见提示工程）都是利用模型的短期记忆来学习。

长期记忆：
为智能体提供了长时间保留和回忆信息的能力，通常利用外部向量存储和快速检索实现。

工具使用 Tool Use

调用外部 API：
获取模型权重中缺失的额外信息
（通常在预训练后很难更改）

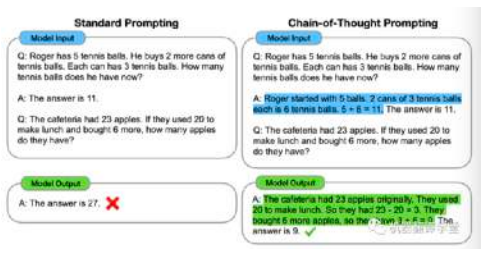
当前信息　　代码执行能力　　对专有信息源的访问等

Agent 规划 Planning

CoT

Chain of Thought, 思维链

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models



CoT 将大型任务转化为多个可管理的小任务，并解释清楚模型的思维过程

ToT

Tree of Thoughts, 思维树

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

1. 通过在每一步探索多种推理可能性来扩展 CoT
2. 首先将问题分解为多个思考步骤
3. 并在每个步骤中生成多个思考
4. 创建一种树结构
5. A 搜索过程 广度优先搜索（BFS） 深度优先搜索（DFS）
6. B 状态评估 分类器（通过提示） 多数 Vote

LLM+P

外部经典规划器

LLM+P: Empowering Large Language Models with Optimal Planning Proficiency

1、该方法利用规划领域定义语言（PDDL）作为描述规划问题的中间接口

(1) 将问题转化为「Problem PDDL」 → (2) 请求经典规划器基于现有的「Domain PDDL」生成 PDDL 规划 → (3) 将 PDDL 规划转换回自然语言

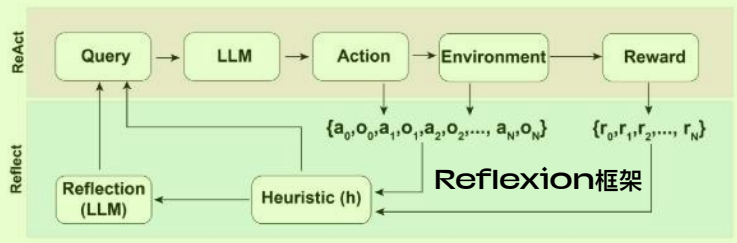
2、规划步骤被外包给了外部工具，并假设特定领域的 PDDL 和合适的规划器可用

3、在某些机器人设置中很常见，而在许多其他领域并不常见

Self-reflection

自我反思 ReAct 将「动作空间」扩展为一个任务特定的组合，将推理和动作集成在 LLM 中

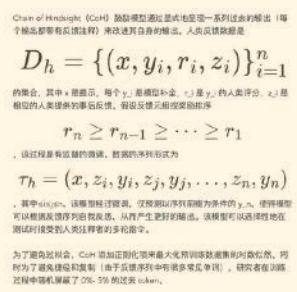
允许自主智能体通过完善以往行动决策和纠正以往错误来迭代改进，因而会在出现试错的现实世界任务中发挥至关重要的作用。



CoH

Chain of Hindsight

鼓励模型通过显式地呈现一系列过去的输出（每个输出都带有反馈注释）来改进其自身的输出



Chain of Hindsight (CoH) 鼓励模型通过显式地呈现一系列过去的输出（每个输出都带有反馈注释）来改进其自身的输出。人类反馈数据是

$$D_h = \{(x, y_i, r_i, z_i)\}_{i=1}^{n}$$

的集合，其中 x 是提示，每个 y_i 是模型补全，r_i 是 y_i 的人类评分，z_i 是相应的人类提供的事后反馈。假设反馈元组按奖励排序

$$r_n \geq r_{n-1} \geq \cdots \geq r_1$$

，该过程是有监督的微调，数据的序列形式为

$$\tau_h = (x, z_i, y_i, z_j, y_j, \ldots, z_n, y_n)$$

，其中 $i \leq j \leq n$。该模型经过微调，使得模型可以根据反馈序列进行反思，从而产生更好的输出。该模型可以选择性地在测试时接受到人类注释者的多轮指令。

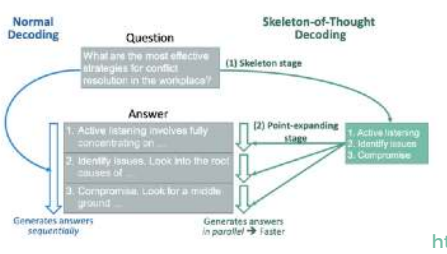
为了避免过拟合，CoH 添加正则化项来最大化预训练数据集的对数似然。同时为了避免捷径和复制（由于反馈序列中有很多常见单词），研究者在训练过程中随机屏蔽了 0%- 5% 的过去 token。

https://arxiv.org/abs/2302.02676

SoT

Skeleton of Thought, 思维骨架

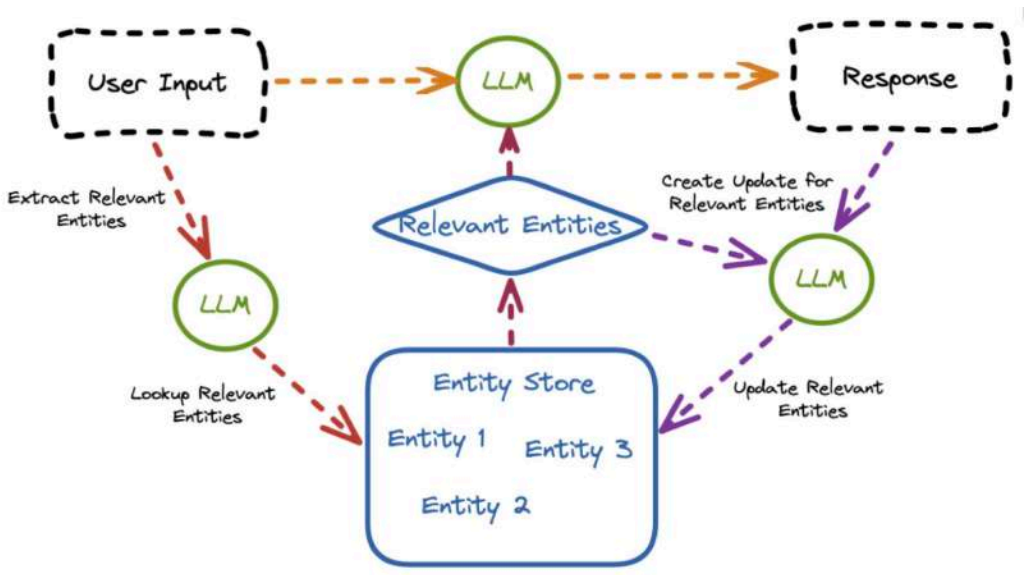
鼓励模型通过显式地呈现一系列过去的输出（每个输出都带有反馈注释）来改进其自身的输出



SoT 引导 LLM → 先生成答案骨架 → 并行 API 调用或分批解码 → 并行完成每个骨架点的内容

https://arxiv.org/pdf/2307.15337.pdf

Agent | 记忆 Memory



分类与映射

感知记忆 **Sensory Memory**	**记忆的早期阶段** Iconic Memory (Visual) 图像记忆 Echoic Memory (Auditory) 回声记忆（听觉） Haptic Memory (Touch) 触摸记忆（触感）	**多模态学习嵌入表示**
短期记忆或工作记忆 **Short-term memory (Working memory)**	**短期记忆**	**上下文学习**
长期记忆 **Long-term memory**	**显式、陈述性记忆** Explicit / Declarative memory (Conscious) Episodic memory (Life events) 情景记忆（事件和经过） Semantic memory (Facts, Concepts) 语义记忆（事实和概念） **隐式、程序性记忆** Implicit / Procedural memory (Unconscious, skills)	**外部向量存储** 查询 | 检索 | 访问

Agent | 使用工具 Tool Use

CoA

Chain of Action, 行为链

TALM

Parisi et al. 2022

MRKL 架构

Parisi et al. 2022

模块化推理（Modular Reasoning）、知识（Knowledge）和语言（Language）

专家模块

神经:深度学习模型
符号:数学计算器　货币转换器　天气 API

LLM 路由器

ToolFormer

Schick et al. 2023

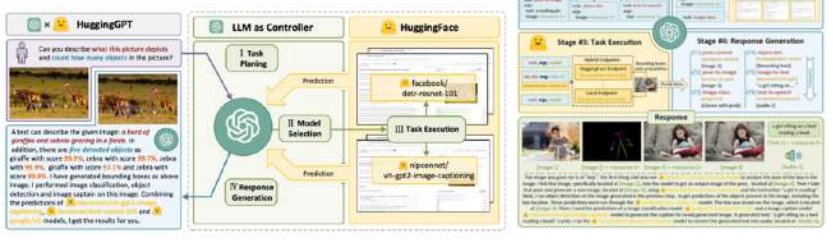
Figure 2: Key steps in our approach, illustrated for a *question answering* tool: Given an input text $\mathbf{x}$, we first sample a position i and corresponding API call candidates $c_i^1, c_i^2, \ldots, c_i^k$. We then execute these API calls and filter out all calls which do not reduce the loss L_i over the next tokens. All remaining API calls are interleaved with the original text, resulting in a new text $\mathbf{x}^*$.

TaskFormer

HuggingGPT

工作原理示意图

使用 ChatGPT 作为任务规划器的框架，根据模型描述选择 HuggingFace 平台中可用的模型，并根据执行结果归纳总结出响应



(1) 任务规划：LLM 作为大脑，将用户请求解析为多个任务。每个任务有四个关联属性：任务类型、任务 ID、依赖项和参数。研究团队使用少量例子来指导 LLM 进行任务解析和规划。 → (2) 模型选择：LLM 会从一个模型列表中选择模型，将任务分配给专家模型。由于上下文长度有限，需要进行基于任务类型的过滤。 → (3) 任务执行：专家模型执行具体任务，并记录执行结果。

(4) 响应生成：LLM 接收执行结果，并向用户提供总体结果。

API-Bank

Li et al. 2023

评估工具增强型 LLM 性能的基准:
53 个常用的 API 工具
工具增强型 LLM 工作流
涉及 568 个 API 调用的 264 个已注释的对话

LLM 首先通过 API 搜索引擎找到合适的 API 调用:
使用相关文档调用 API
可选择多样化 API
搜索引擎　日历查询　计算器　智能家居控制　日程管理等

Agent 探索项目

RPA, Robotic Process Automation
机器人流程自动化

AutoGPT

MetaGPT

Transformers Agents

ToolLLM

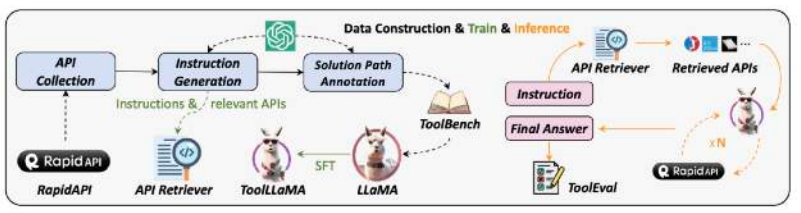


Figure 1: Three phases of constructing ToolBench and how we train our API retriever and ToolLLaMA. During inference of an instruction, the API retriever recommends relevant APIs to ToolLLaMA, which performs multiple rounds of API calls to derive the final answer. The whole reasoning process is evaluated by ToolEval.

Tool Learning with Foundation Models

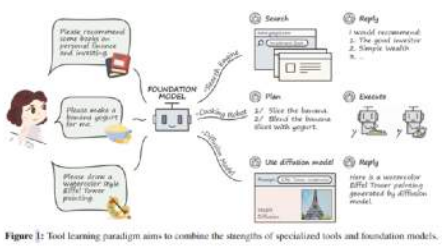
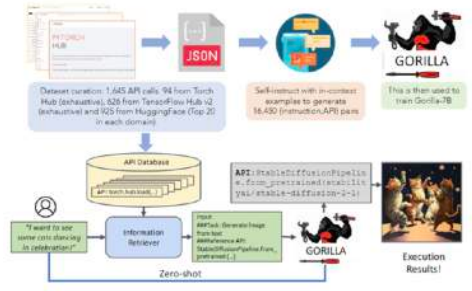


Figure 1: Tool learning paradigm aims to combine the strengths of specialized tools and foundation models.

Gorilla: Large Language Model Connected with Massive APIs



Multi-Agents

CAMEL: Communicative Agents for “Mind” Exploration of Large Scale Language Model Society

Generative Agents: Interactive Simulacra of Human Behavior

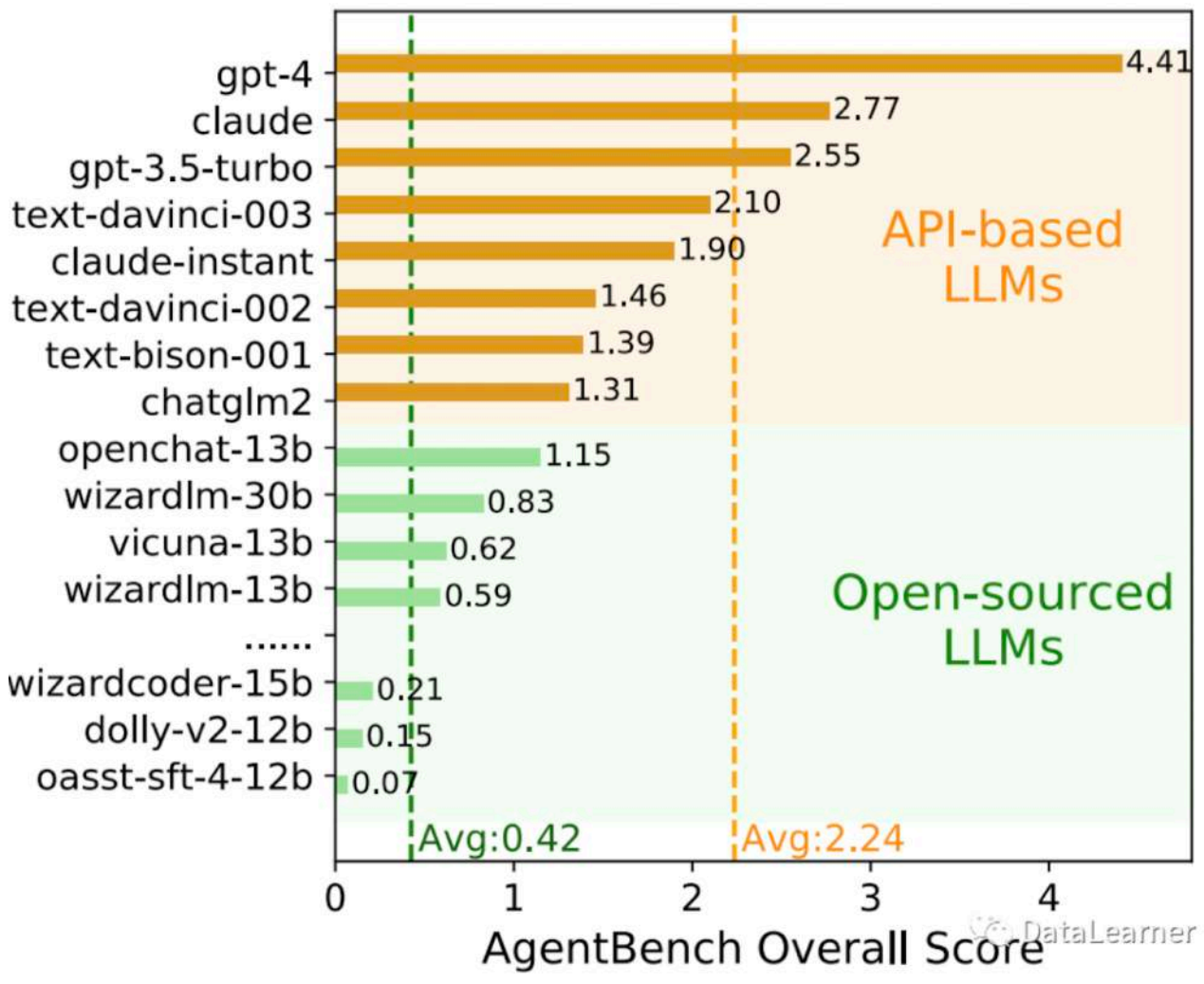
Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory

METAGPT: META PROGRAMMING FOR MULTI-AGENTCOLLABORATIVE FRAMEWORK

Paradot

斯坦福小镇

Agent | AgentBench



商业顶级模型展现出在复杂环境中完成代理任务的强大能力，能够理解指令并进行多轮交互。这显示了LLM作为代理的潜力。

目前开源模型与商业模型之间还存在显著的差距，开源模型在AgentBench上普遍表现较弱。

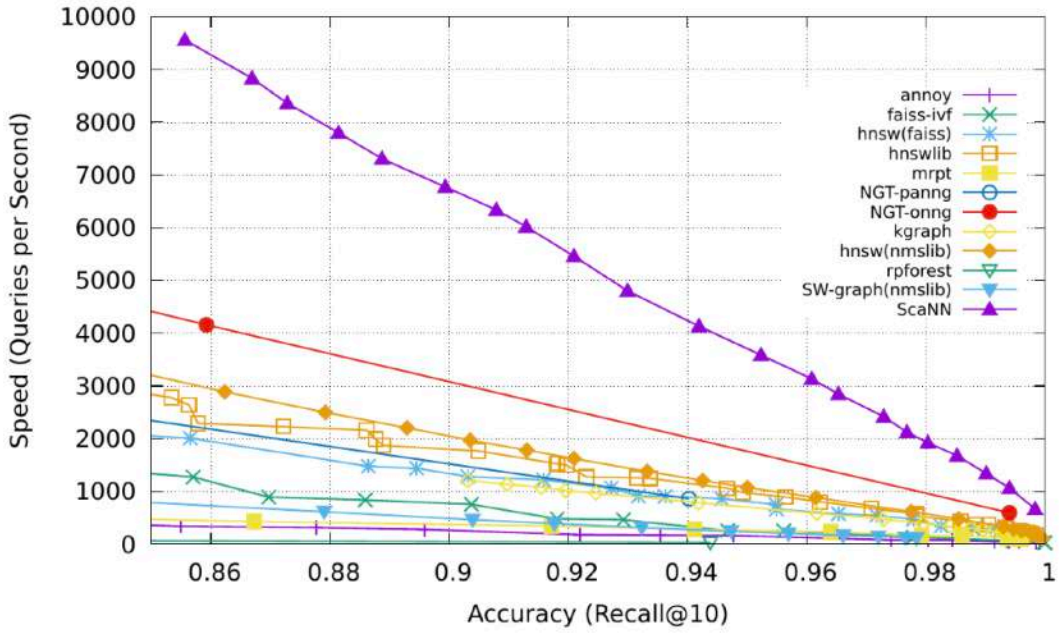
不同环境有不同的挑战，如操作系统和数据库考察编码能力，知识图谱需要复杂推理，网页浏览需要处理庞大inputs。不同模型之间也存在明显的优劣。

Vector Embedding

向量存储技术

将信息的嵌入表示保存到向量存储数据库

- MIPS 最大内积搜索
- MIPS 算法比较



近似最近邻算法 ANN, Approximate nearest neighbors

局部敏感哈希 LSH

引入了一个哈希函数，使得相似的输入项以高概率映射到相同的 buckets 中

其中 buckets 的数量远远小于输入的数量

分层可导小世界 HNSW, Hierarchical Navigable Small World

受小世界网络（small world networks，一种图结构）的启发。大多数节点可以在很少的步骤内与其他节点相连

构建了小世界图的层次结构，底层包含实际的数据点，中间层创建了快捷方式以加速搜索
在执行搜索时从顶层的一个随机节点开始，并向目标节点导航，当无法再靠近目标时，它向下移动到下一层，直到达到底层

近似最近邻 ANNOY

核心数据结构：随机投影树（Random Projection Trees）

一组二叉树，其中每个非叶节点表示一个超平面，将输入空间分割为两部分，而每个叶节点则存储一个数据点。树是独立且随机构建的，因此在某种程度上类似于哈希函数。
这个想法与 KD 树（一种将空间中点分开存储的树状数据结构）密切相关，但扩展性更强。

FAISS

在高维空间中，节点之间的距离遵循高斯分布，因此应该存在数据点的聚类。

FAISS 通过将向量空间分割成聚类并在聚类内进行量化来应用向量量化。

可扩展最近邻 ScaNN

各向异性矢量量化（Anisotropic Vector Quantization，AVQ）

减少了数据点之间的距离误差

Vector Embedding

向量存储方案

单纯向量数据库

Milvus zilliz Pinecone vespa marqo drant Weaviate Chroma Vald LanceDB

API OpenAI Pinecone Chroma LanceDB Marqo Weaviate Milvus/ Zilliz Qdrant Vald Vespa

全文检索数据库

elasticsearch APACHE LUCENE OpenSearch Apache Solr

API Elastic / Lucene OpenSearch Solr

支持向量的NoSQL数据库

DataStax Astra cassandra Azure Cosmos DB MongoDB neo4j redis ROCKSET

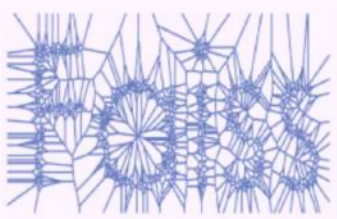
API MongoDB Cassandra / DataStax Astra CosmosDB Rockset 键值数据库 其他特殊用途的数据库

支持向量的SQL数据库

SingleStore ClickHouse kinetica

API SingleStoreDB PostgreSQL Clickhouse Kinetica Pgvector Supabase Vector

开源向量库



非商用权重

Faiss Annoy Hnswlib M3E

商用权重

BGE

www.xRunda.com

MoE Mixture-of-Experts 专家混合模型

DeepSpeed-MoE

多个专家 + 门控网络

输入不同方面数据→专家
A 共享参数
B 自有参数

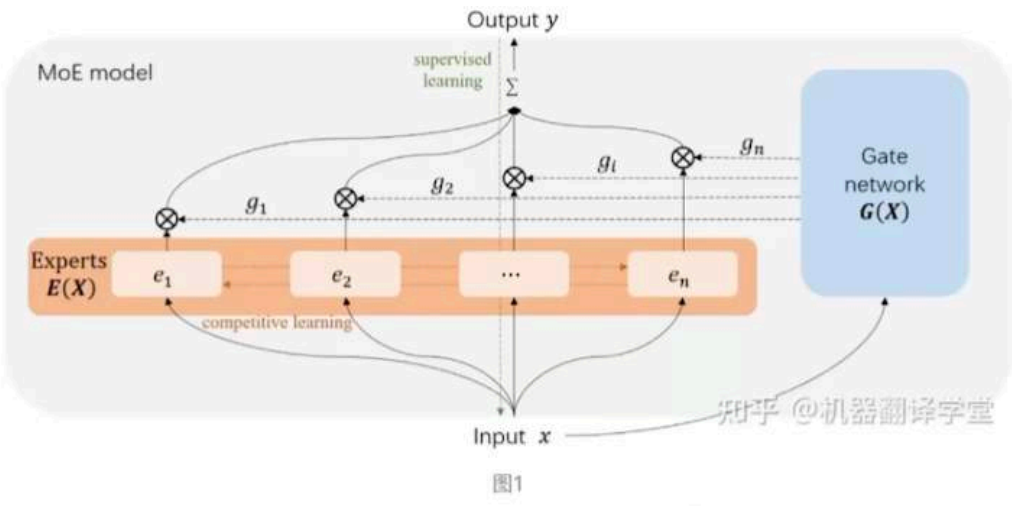


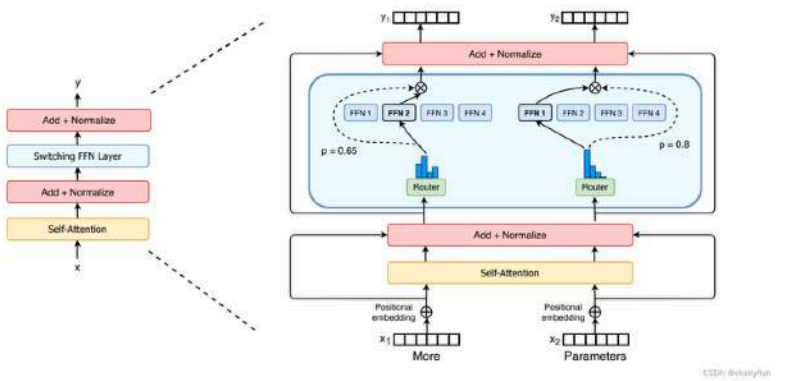
图1

Switch Transformer

模型的每一层都是一个专家网络的集合，输入数据会被动态地路由到不同的专家进行处理

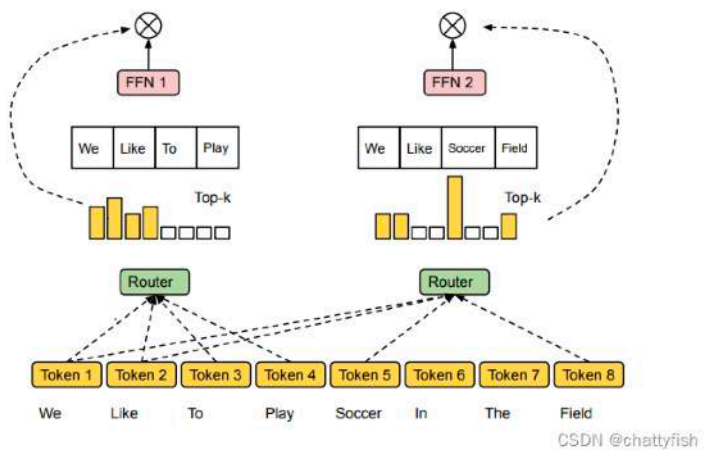
用一个稀疏的 Switch 前馈网络 FFN 层（浅蓝色）替换 Transformer 中存在的密集 FFN 层。该层独立地对序列中的标记进行操作，然后路由到多个 FFN 专家中

Switch FFN 层返回所选 FFN 的输出，然后乘以路由器阈值，然后进行合并

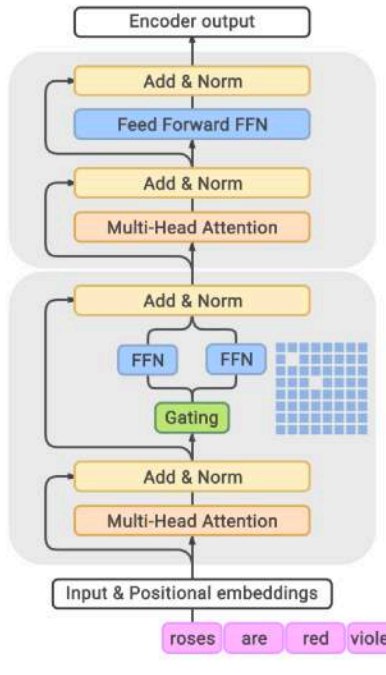


Expert Choice

设置一组具有预定缓冲区容量的专家，给专家分配给前 k 个令牌，产生一个令牌到专家的得分矩阵，然后用该矩阵做出路由决策



Generalist Language Model



使用稀疏激活的混合专家架构来扩大模型容量，同时与密集型变体相比，其训练成本也大大降低

在 Transformer 层之间加一个 MoE 层。对于每个输入标记，门控模块会动态从 64 个专家中选择两个最相关专家。

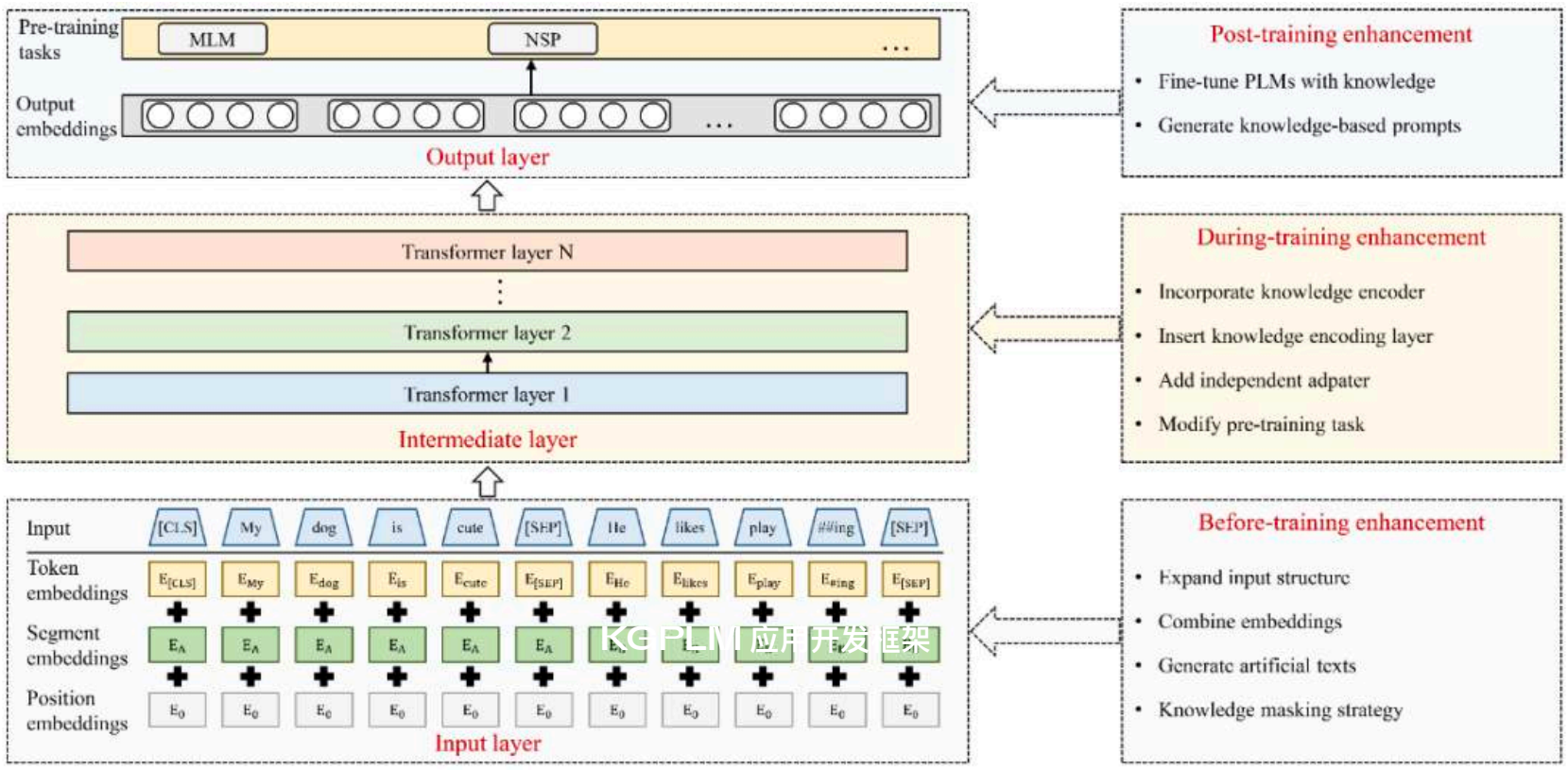
这两个专家的输出的加权平均值将然后传递给上面的 Transformer 层。对于输入序列中的下一个标记，将选择两个不同的专家来达到平衡。

Knowledge Graph

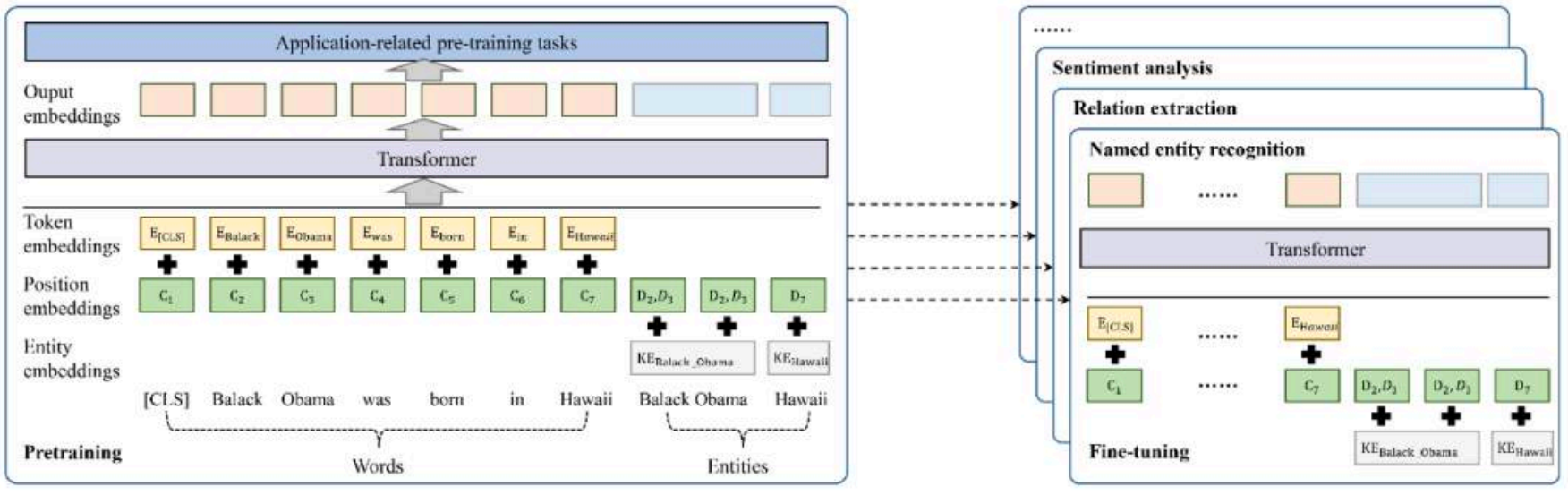
知识图谱增强预训练模型

Knowledge Graph-enhanced Pre-trained Language Model, KGPLM

整体架构

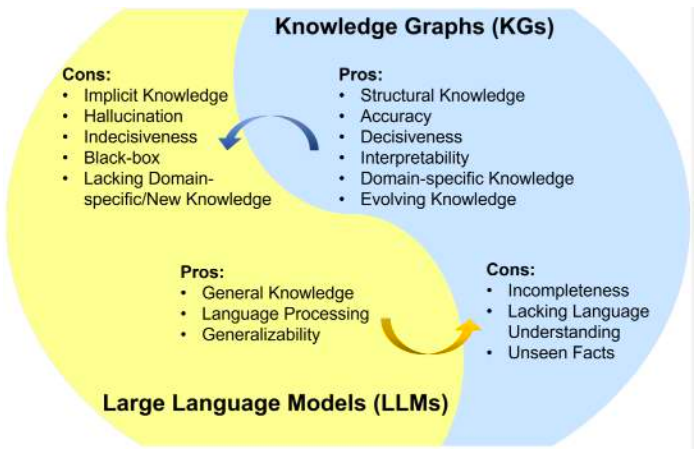


KGPLM 应用开发框架

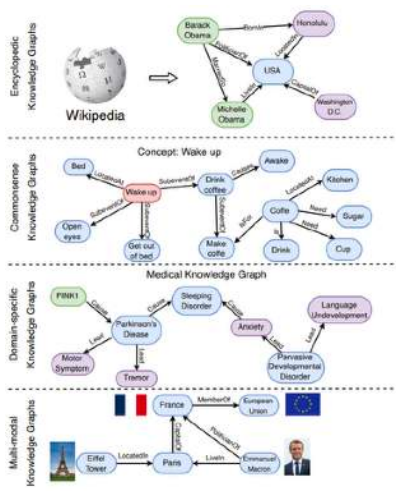


Knowledge Graph

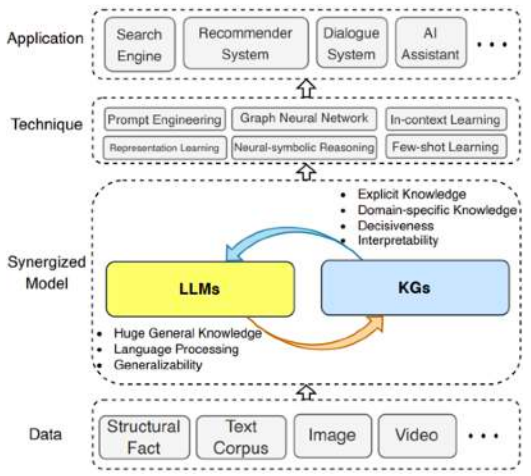
LLM与知识图谱协同研究



LLM 和知识图谱的优缺点



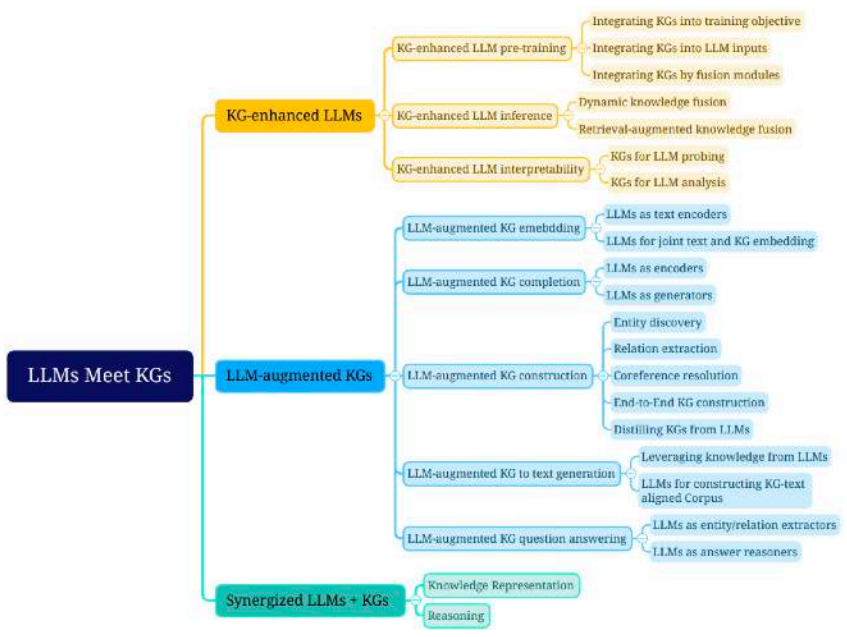
不同类别知识图谱示例



LLM 与知识图谱协同的一般框架

数据层　协同模型　技术层　应用层

联合 LLM 与知识图谱的相关研究分类

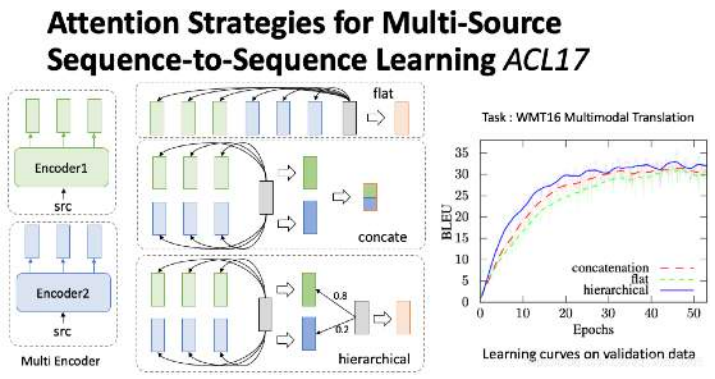


未来研究方向

- 将知识图谱用于检测 LLM 的幻觉
- 将知识图谱用于编辑 LLM 中的知识
- 将知识图谱用于黑箱 LLM 知识注入
- 将多模态 LLM 用于知识图谱
- 将 LLM 用于理解知识图谱的结构
- 将 LLM 和知识图谱协同用于双向推理

Multimodal
多模态技术研究

多模态表示 Representation

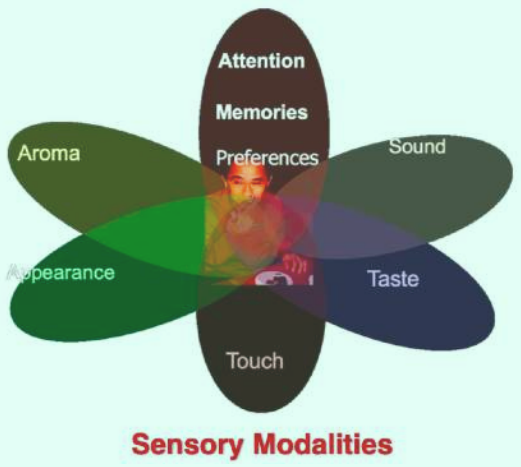


多模态中的注意力机制

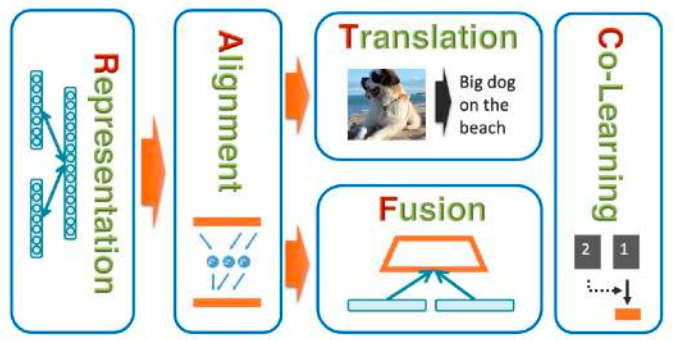
多模态转化

- 文生图
- 文生视频
- 文生音
- 文生3D

多模态融合



多模态学习技术挑战



模态联合学习

Multimodal Compact Bilinear Pooling

Cross-Modal Retrieval

跨模态学习

Deep Cross-Modal Projection Learning

Cross-Modal Transfer Learning

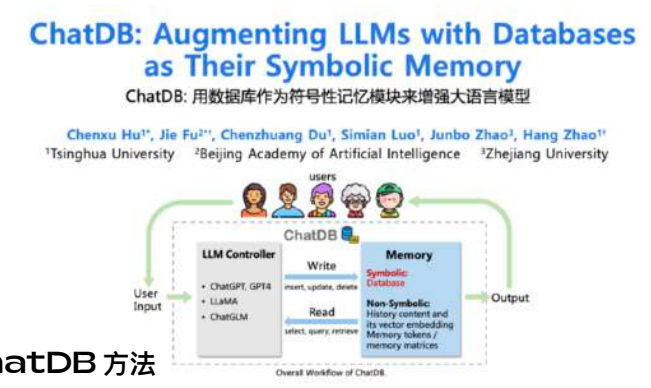
多模态自监督学习

Joint Audio-Visual Self-Supervised Learning

SimCLR-MultiTask

Text-to-SQL数据库交互技术

Chain-of-Memory（CoM，记忆链）



ChatDB 方法

Augmenting LLMs with Databases as Their Symbolic Memory

LLM 通过生成 SQL 指令来操纵数据库，从而实现对记忆模块中历史信息精确的增删改查，并在需要时为大语言模型提供信息，以帮助其回应用户的输入

图像生成技术 图像生成先进模型

MidJourney		https://www.midjourney.com	 Jason Allen 的作品《Théâtre D'opéra Spatial》獲得了科羅拉多州博覽會的「數位藝術類」獎項
StabilityAI	Stable Diffusion	https://stablediffusionweb.com	
Meta	CM3leon		
OpenAI	DALL-E2	https://openai.com/product/dall-e-2	
	GPT3 + DALL-E		
	CLIPDraw	基于CLIP模型和可微分渲染器，可以生成具有细节和几何结构的图像，例如“一个笑着的蓝色八边形“	
NVIDIA	Muse		
Google	Text2Scene	基于场景图表示和神经渲染器，可以生成具有场景结构和透视关系的图像，例如“一个在桌子上放着苹果和香蕉的房间”	
	Imagen	基于变分自编码器和对比学习	
Adobe	Firefly		
Baidu	文心一格		
西湖心辰	造梦日记	https://www.printidea.art	
北京智源	Flag Studio	https://flagstudio.baai.ac.cn	
Civitai	LoRA Model	https://civitai.com/	

音频生成技术

音频生成主要类型

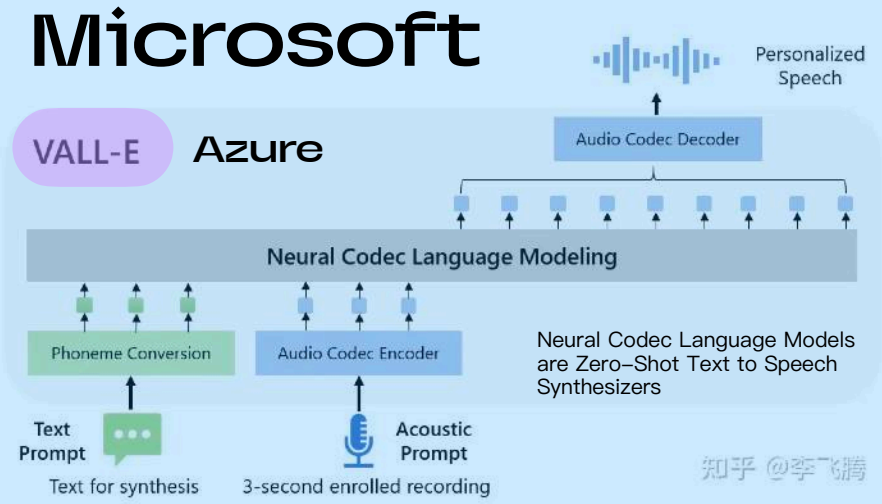
输入数据类型	定义	典型应用
文字信息	提取文字信息特征并合成语音信息	信息播报、人机交互服务
音频信息	根据给定的语音片段进行编辑，或将一种语言转换为另一种语言的语音信息	语音编辑、语音翻译、音乐制作
肌肉震动	对喉部、面部等肌肉运动情况进行感知并合成语音	医疗可穿戴设备
视觉内容	对图像、视频等视觉内容进行识别和理解，并生成与口型对应的语音信息	数字人

国内外典型音频模型

模型	机构	是否开源	简介
Tacotron2	Google	已在GitHub开源	首先提出端到端语音合成模型，作为多个语音系统解决方案的基础架构
Transformer-TTS	Google	已在GitHub开源	基于Tacotron2和Transformer的结合，目前主流的端到端语音合成框架
AudioLM	Google	未开源	基于Transformer的音频生成模型，支持根据音频片段生成语音和音乐
Whisper	OpenAI	已在GitHub开源	自动语音识别模型，通过大规模和多样化数据集提升语音识别能力，并支持语音转录、语音翻译等
WavLM	微软亚洲研究院&Azure	已在GitHub开源	基于Transformer架构的通用语音预训练模型，使用超过94000小时英文语料的大规模数据集训练提升模型鲁棒性和泛化能力，在语音识别、语音增强、语音翻译等任务中取得了很好的效果
FastSpeech2	微软&浙江大学	已在GitHub开源	基于Transformer-TTS模型的端到端语音合成模型，针对对Fastspeech的缺点进行了改进，语音生成速率快，对语音长短和韵律的控制较好
Make-an-audio	浙江大学、北京大学、火山语音	未开源	基于扩散模型的语音生成模型，提出了Distill-then-Reprogram文本增强策略，支持将文本、音频、图像、视频等多模态作为输入生成语音，是业界首次尝试在用户定义的输入模态下生成高质量音频
DeepVoice3	百度	未开源	全卷积序列到序列语音合成模型，通过扩展语音合成模型训练数据集，能够提升多人语音合成效果
文心ERNIE-SAT	百度	已在GitHub部分开源，包括语音编辑、个性化语音合成、跨语言语音合成	采用语音-文本联合训练方式的跨模态预训练大模型，融合跨语言音素知识，能够提升多种语音合成任务效果
SMART-TTS	科大讯飞	未开源	工业级中文语音预训练模型，支持多模态语音识别、情感识别、声纹识别等任务

音频生成技术 音频生成先进模型

Microsoft



OpenAI

Whisper

Adobe

Podcast https://podcast.adobe.com

阿里巴巴

通义听悟

科大讯飞

智能音频

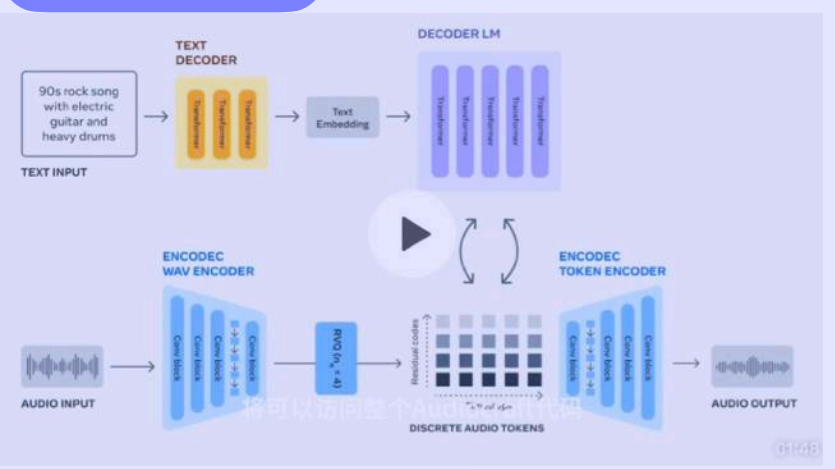
灵动音科技

DeepMusic

- AI音乐生成引擎[和弦派]
- 像游戏一样创作AI音乐[口袋乐队]
- 提示词生成音乐工具[BGM猫]
- AI歌词生成器[LYRICA]

Meta

AudioCraft



MusicGen

文本生成音乐

基于文本输入生成音乐，使用Meta拥有和专门授权的音乐进行训练

AudioGen

文本生成音效

基于文本输入生成音效（雨点声、狗吠、警笛等），使用公共音效训练

EnCodec

高质量音频压缩

Neural Audio Codec，神经音频编解码器，压缩音频，以较低的音损保持音频的高质量

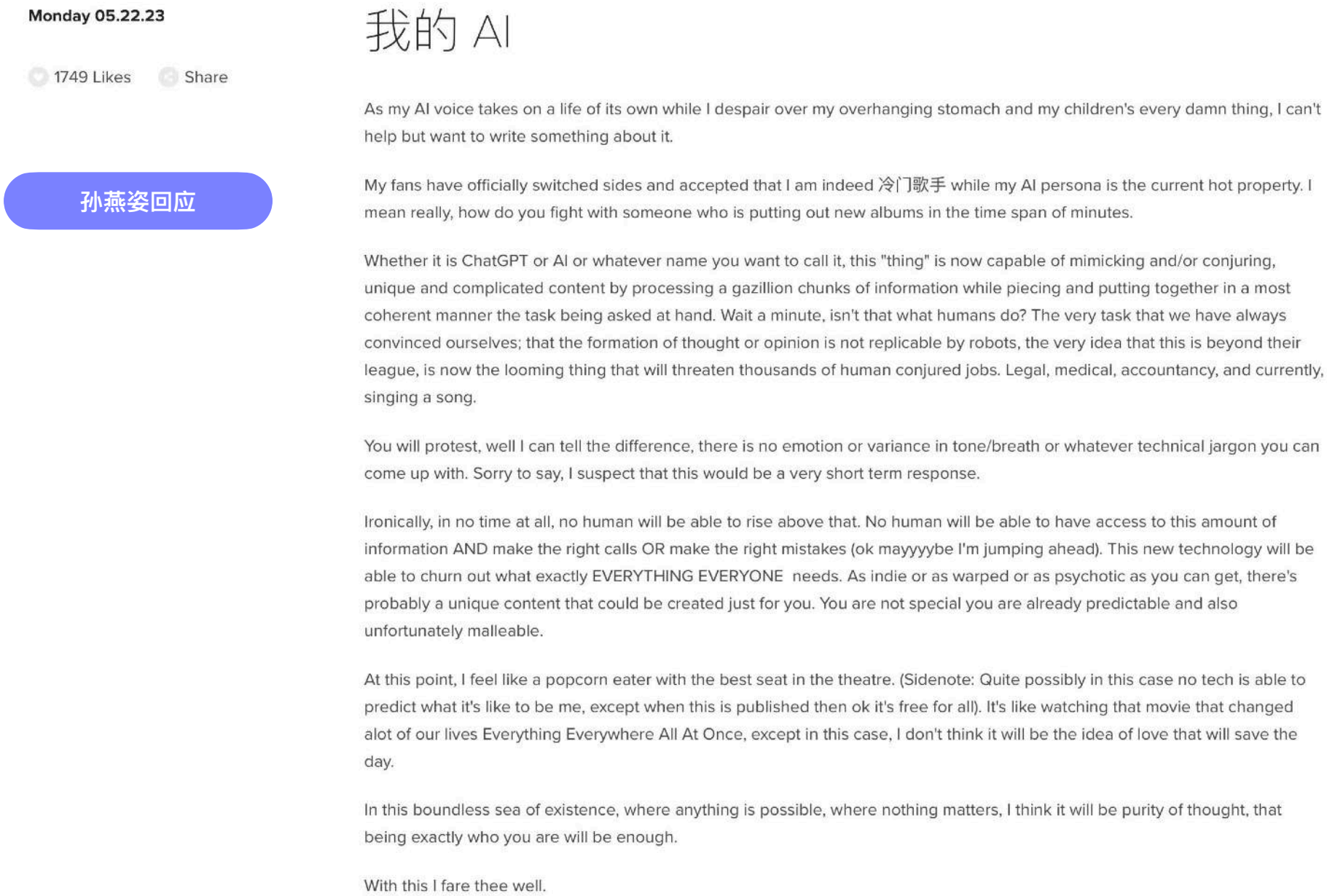
已开源

https://github.com/facebookresearch/audiocraft

Paper

https://arxiv.org/abs/2306.05284

音频生成技术 案例 AI 孙燕姿



Sovits 4.0 模型

基于 so-vits-svc 的开源项目

变分自动编码器（Variational Autoencoder, VAE）的架构

结合了条件生成对抗网络（Conditional Generative Adversarial Network, CGAN）

视频生成技术

视频生成进展与展望

- 视频生成技术
- 图像到视频的转换
- 视频编辑与合成
- 语义分割与物体识别
- 三维建模与渲染
- 风格迁移与内容生成
- 动作捕捉与人物动画
- 音频与视频的同步
- 视频中的人物、场景、物体等元素更逼真
- 实时视频生成与编辑
- 多模态信息整合
- 强化学习与交互视频
- 对视频场景智能分割、合成、渲染
- 个性化视频生成
- 虚拟现实与增强现实整合
- 解决伦理与安全问题

动画生成技术

Runway

提供多种模式和功能，可以根据文本、图片或音频生成视频，也可以分享和编辑视频

最新的文本生成视频（Text-to-video）AI模型，可以根据简单的文本提示生成4秒的视频片段

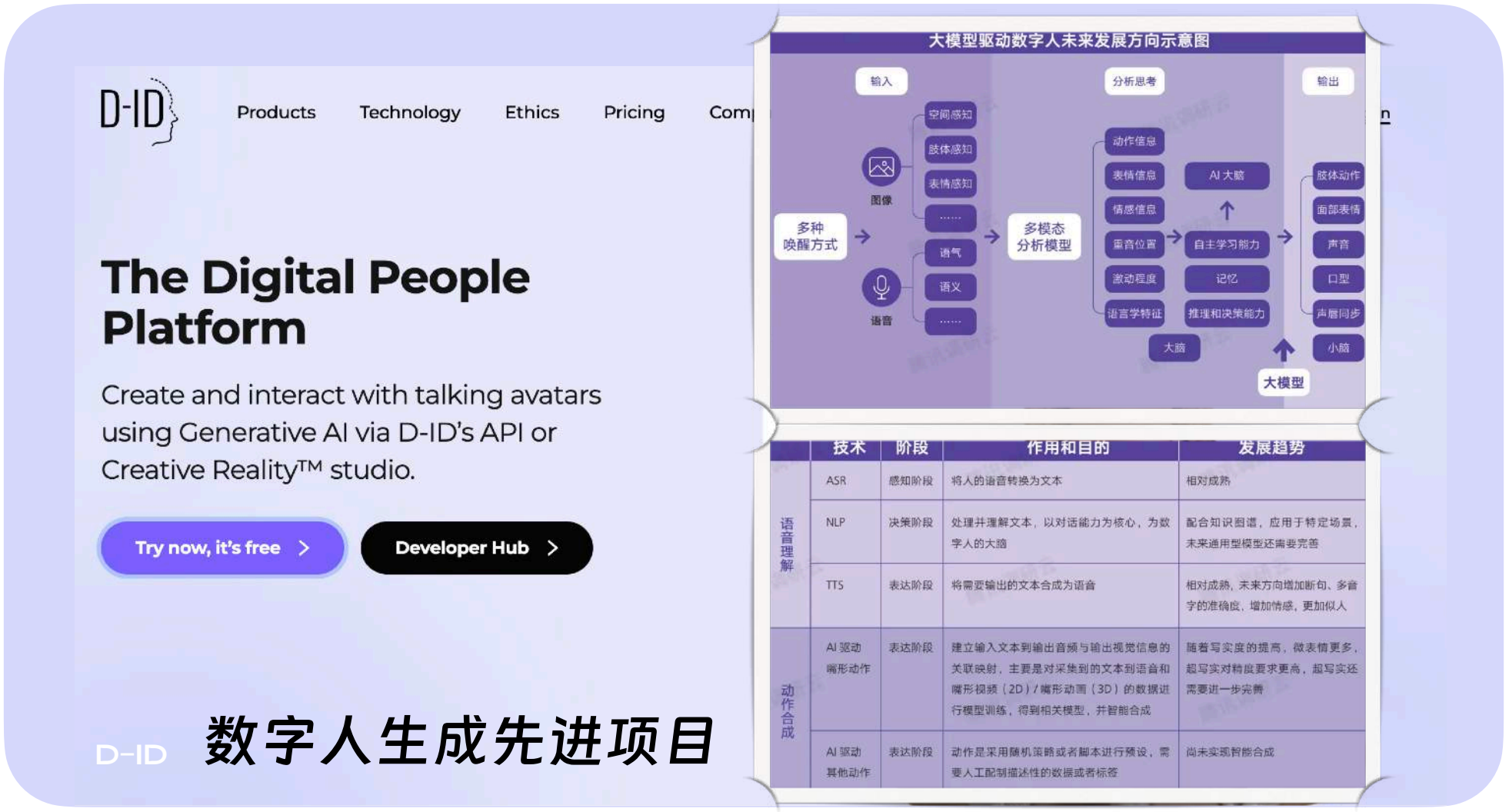
Pictory

免费的 AI 视频生成器，可以从文本或 URL 转换为视频，也可以编辑现有的视频

FlexClip

免费的 AI 视频生成器，可以提供文本到视频、预制视频模板和庞大的媒体库

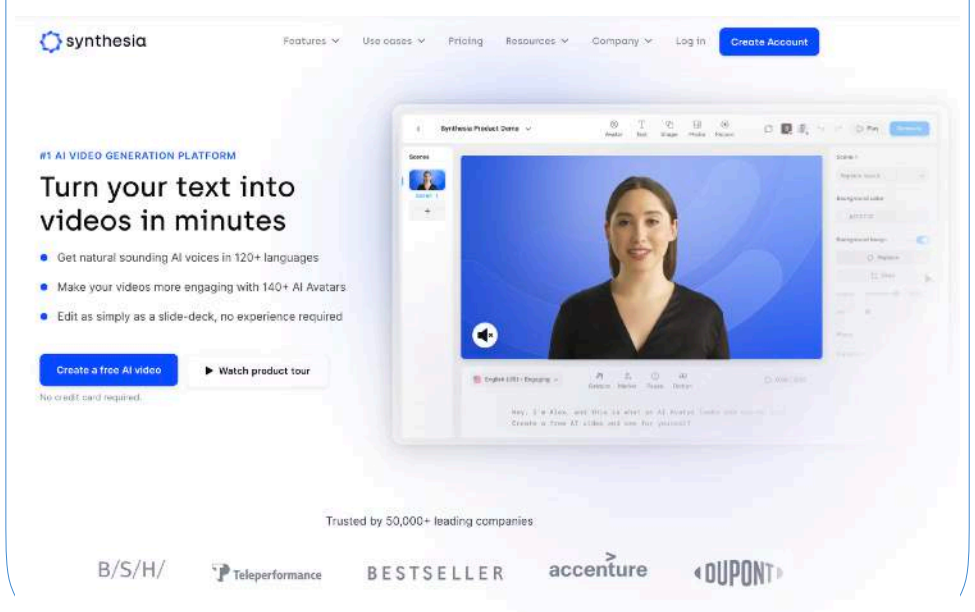
数字人生成技术



	技术	阶段	作用和目的	发展趋势
语音理解	ASR	感知阶段	将人的语音转换为文本	相对成熟
	NLP	决策阶段	处理并理解文本，以对话能力为核心，为数字人的大脑	配合知识图谱，应用于特定场景，未来通用型模型还需要完善
	TTS	表达阶段	将需要输出的文本合成为语音	相对成熟，未来方向增加断句、多音字的准确度，增加情感，更加似人
动作合成	AI 驱动嘴形动作	表达阶段	建立输入文本到输出音频与输出视觉信息的关联映射，主要是对采集到的文本到语音和嘴形视频（2D）/嘴形动画（3D）的数据进行模型训练，得到相关模型，并智能合成	随着写实度的提高，微表情更多，超写实对精度要求更高，超写实还需要进一步完善
	AI 驱动其他动作	表达阶段	动作是采用随机策略或者脚本进行预设，需要人工配制描述性的数据或者标签	尚未实现智能合成

D-ID 数字人生成先进项目

Synthesia



HeyGen

AI Video Generator

3D 生成技术 3D 生成先进技术

NeRF, Neural Radiance Fields 神经辐射场

最新进展

英伟达提出了一种新技术，可以在单张 RTX 3090 上实时渲染 NeRF 模型，并且训练时间最快只需 5 秒

清华大学和鉴智机器人提出了 DFRF，一种快速小样本生成高真实感、自然的讲话头的方法，可以用于数字人等

最新趋势

高质量动态建模

与大模型结合

更丰富的信息嵌入

应用到其他领域

基于 Diffusion 的 NeRF

最新研究

CVPR 2022 有多篇与 NeRF 相关的论文

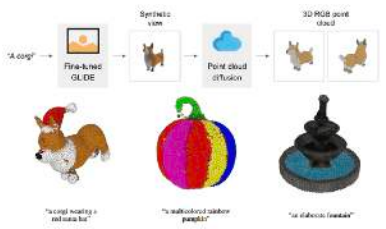
涉及到 Mip-NeRF、Point-NeRF、Human-NeRF、Urban-NeRF、Block-NeRF、Raw-NeRF 等不同的变体和应用

CVPR 2023 论文 120 多篇。

Point-E

最新进展

开源源代码和预训练模型



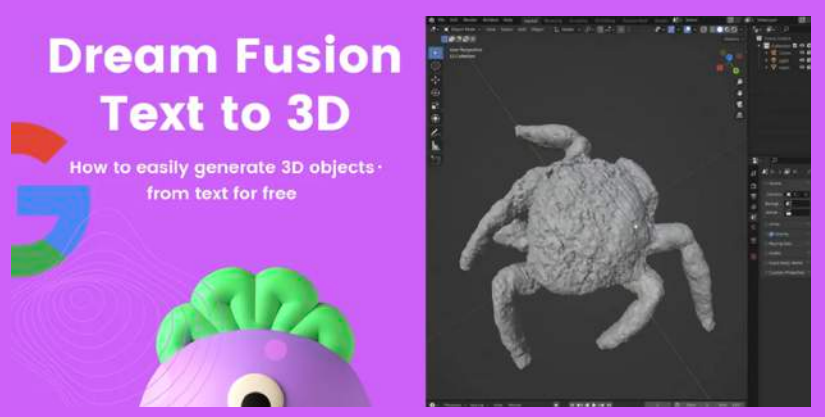
基础信息

OpenAI 开发的文本到 3D 模型的生成模型，它可以根据用户输入的文本描述，自动创建出高质量的 3D 点云模型，并且可以从不同的角度观看

技术原理是先用一个预训练的文本到图像模型 DALL-E 根据文本生成一张 2D 图像，然后用一个基于变分自编码器和正则化自回归流的模型，将 2D 图像转换为 3D 点云模型

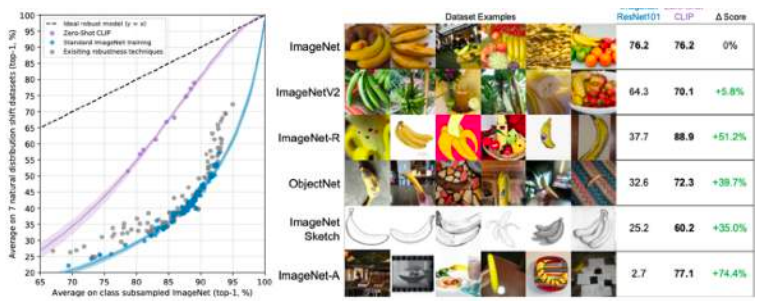
特点是它不需要任何 3D 数据进行训练，只需要大量的 2D 图像-文本对，而且它可以实现极快的 3D 生成，只需要一到两分钟就可以在单块 GPU 上生成 3D 模型

Dream Fusion



谷歌和 UC Berkeley 的研究者开发的文本到 3D 模型的生成模型

SOTA : CLIP, Contrastive Language-Image Pre-training



基于对比学习的多模态预训练 OpenAI 的一篇NLP和CV结合的多模态的工作

应用案例

Contrastive Language-Image Forensic Search

使用CLIP完成视频检索，看一个视频里面有没有出现过一个人或者一些场景，通过直接输入文本的这种形式进行检索

https://github.com/johanmodin/clifs

3D 生成技术 更多相关技术研究

3D 生成表示 / 编码方式

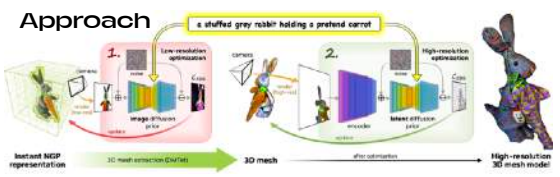
- 三维网格（Mesh）
- 八叉树（Octree）
- 三维体元（Voxels，也称体素）
- 隐函数（Implicit Function）
- 点云（Point Cloud）
- 神经场（Neural Field）
- 三平面（Tri-plane）

更多生成算法 / 网络结构

- 生成对抗网络（GAN）
- 变分自编码器（VAE）
- 扩散模型（Diffusion Model）
- Transformer模型
- 参数化（Parameterization）

其他领先项目

Magic 3D



基础信息

英伟达的新技术，可以根据文本描述生成高质量的 3D 模型，无需任何建模技能。Magic 3D 利用了深度学习和图形学的结合，可以从大量的 3D 数据中学习出一个通用的 3D 表示，然后根据用户的输入生成相应的 3D 网格模型。

最新进展

Magic 3D 可以与图像条件技术和提示编辑方法结合，提供用户更多的控制 3D 合成的方式，例如可以根据一张图片生成一个类似的 3D 模型，或者根据一些关键词修改一个已有的 3D 模型2。
Magic 3D 可以生成各种类型的 3D 模型，包括动物、人物、建筑、食物、家具等，而且可以处理复杂的细节和纹理

MCC

基础信息

- Meta FAIR Lab 开发的文本到 3D 模型的生成模型，它可以根据用户输入的文本描述，自动创建出高质量的 3D 点云模型，并且可以从不同的角度观看
- 技术原理是先用一个预训练的文本到图像模型 DALL-E 根据文本生成一张 2D 图像，然后用一个基于变分自编码器和正则化自回归流的模型，将 2D 图像转换为 3D 点云模型

最新进展

开源源代码和预训练模型

GauDi

- 苹果 AI 团队 开发的文本到 3D 场景的生成模型，它可以根据用户输入的文本描述，自动创建出沉浸式的 3D 室内场景，并且可以从不同的角度观看
- 技术原理是先用一个基于变换器的文本编码器，将文本转换为语义特征，然后用一个基于 NeRFs 的神经渲染器，将语义特征转换为 3D 场景
- 特点是它不需要任何 3D 数据进行训练，只需要大量的 2D 图像-文本对，而且它可以实现多样化和可控制的 3D 生成，可以根据用户的偏好和需求调整场景的布局、颜色、光照等

最新进展

开源源代码和预训练模型

安全性

- 合规机制
- 双新备案
- 第三方过滤
- 关键词拦截
- 鲁棒性 Robustness
- 准确性 Accuracy
- 校准误差 Calibration error
- 恶意性 Toxicitiy

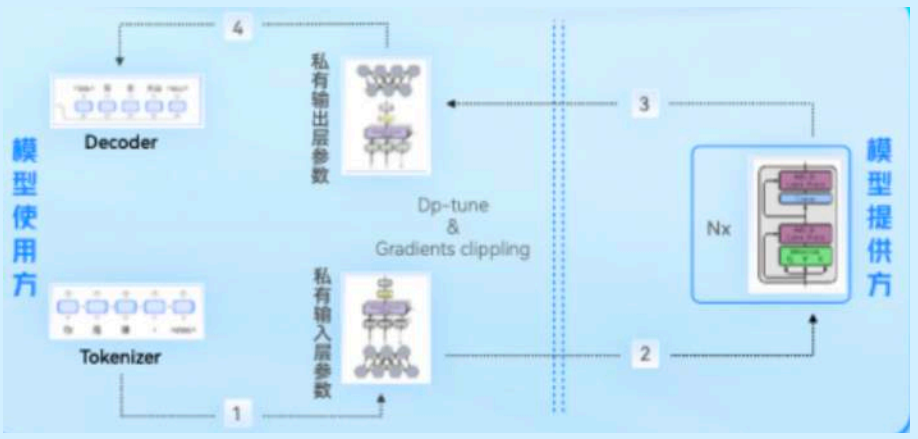
工程问题

算力运维技术

例 飞桨大模型分布式训练技术
https://xie.infoq.cn/article/1ab7515dafab97c7223c88272

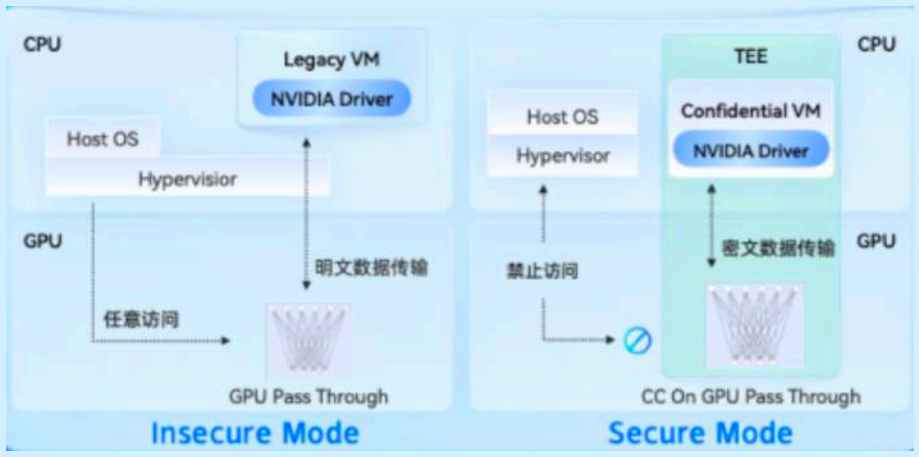
安全沙箱

联邦学习



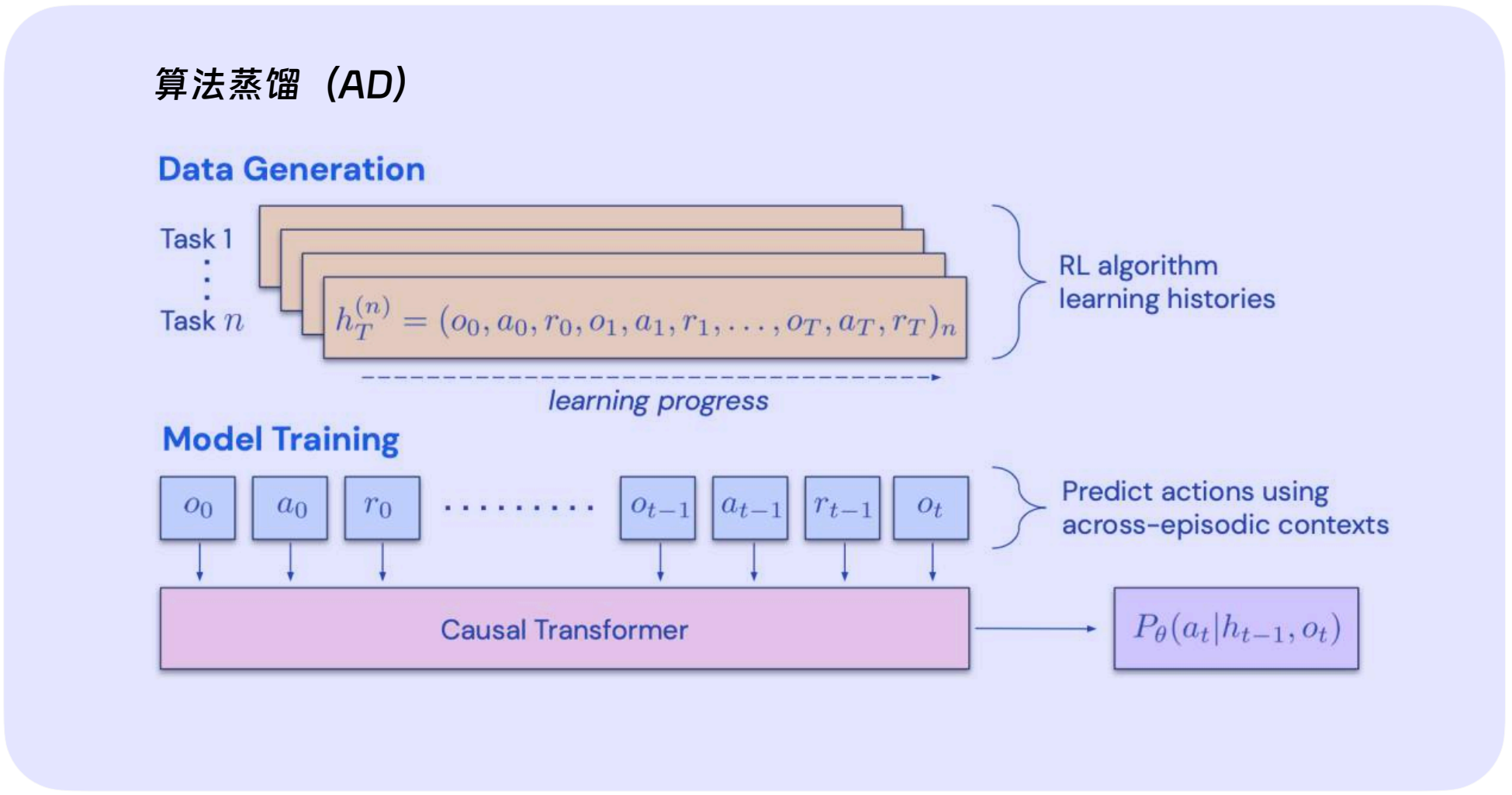
联邦学习 – 数据资产分离

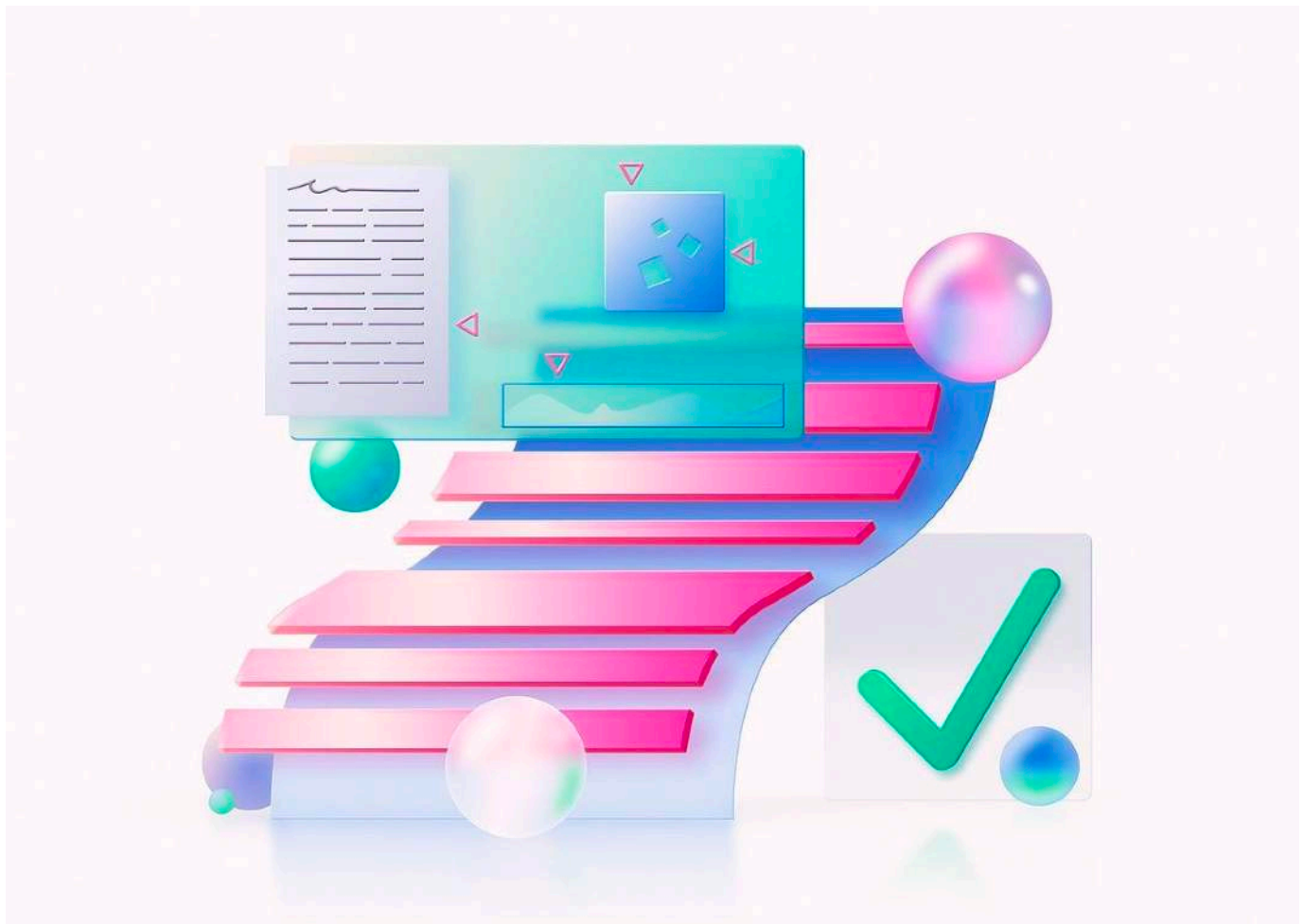
可信计算



硬件支持 – 可信计算环境

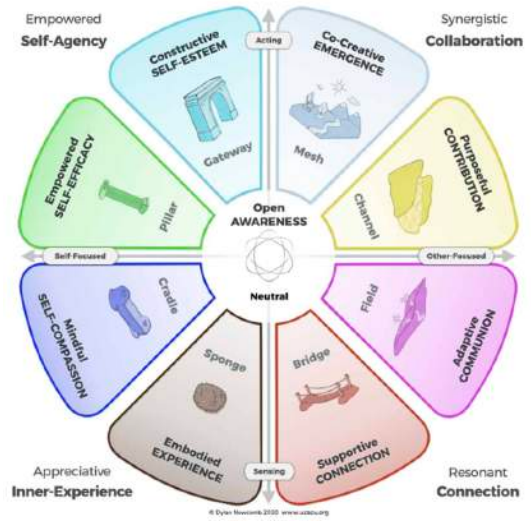
算法问题

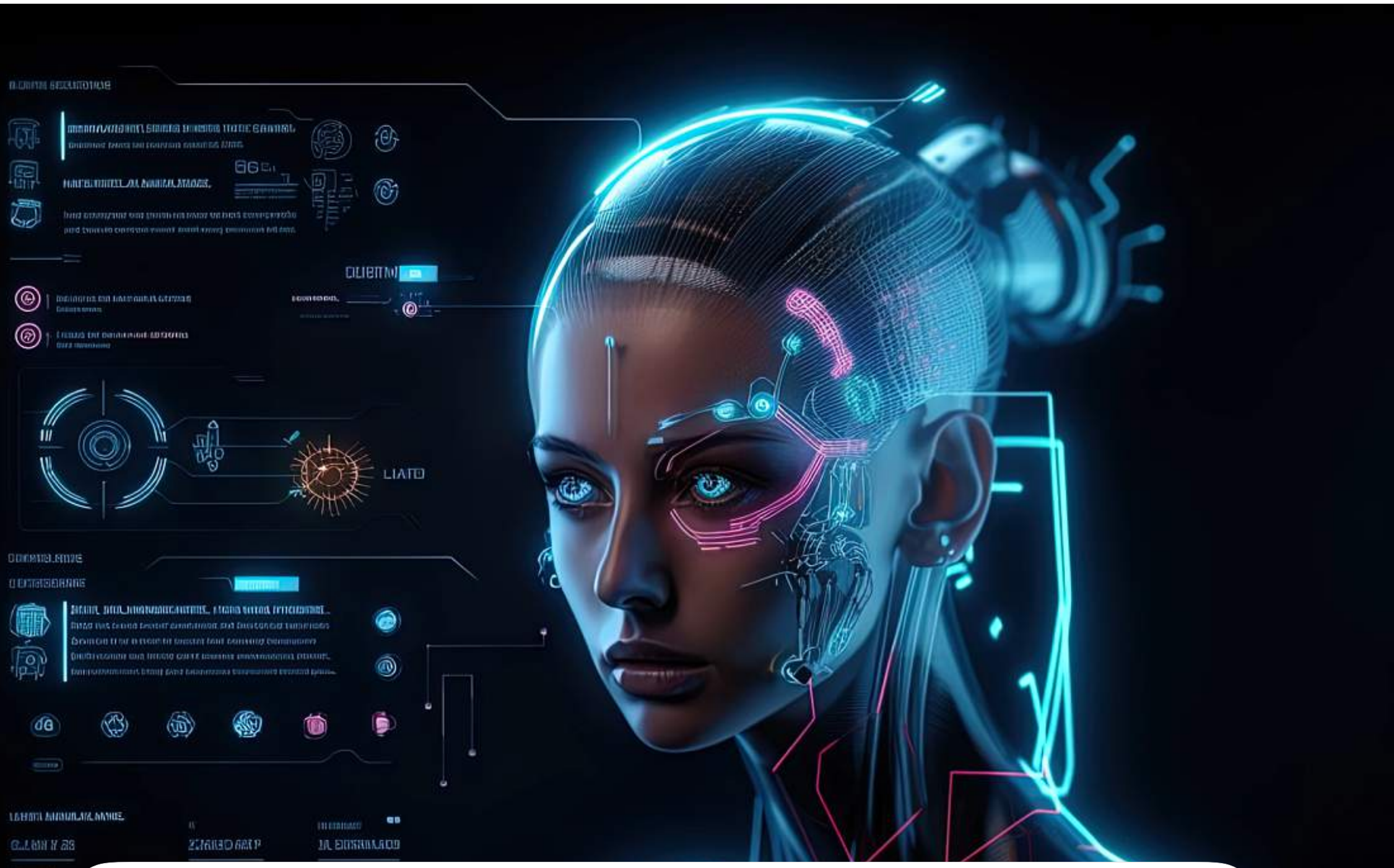




https://arxiv.org/abs/2210.14215

具身智能

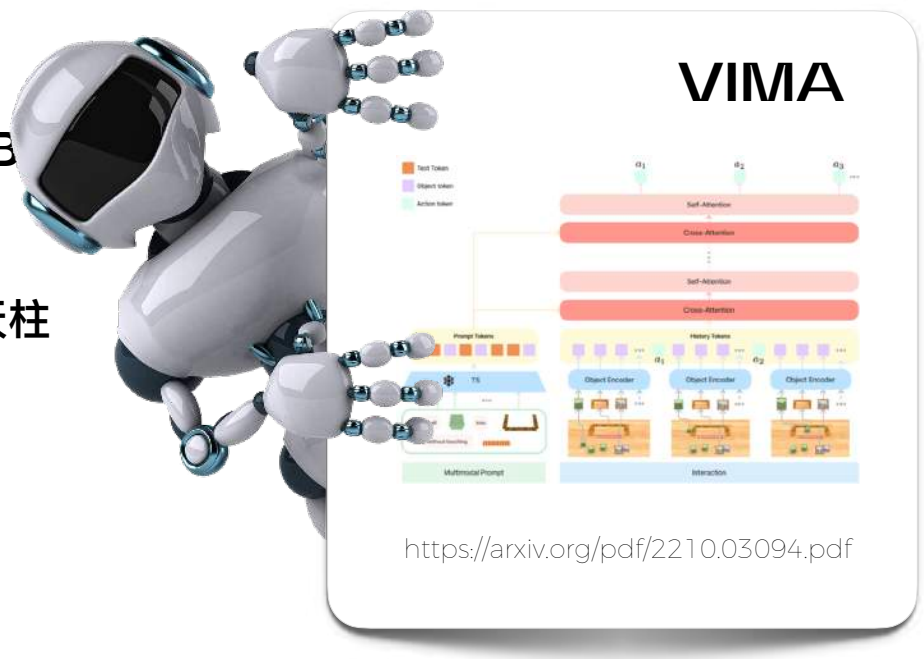




智元 AGIB

特斯拉 擎天柱

小米



端侧模型

智能终端 和 穿戴设备

监控设备

车载设备

案例研究

高通《混合AI是AI的未来》

混合 AI 架构在云端和边缘终端之间分配

并协调 AI 工作负载

全栈 AI 优化

通过跨层的模型、硬件和软件创新，加速AI应用

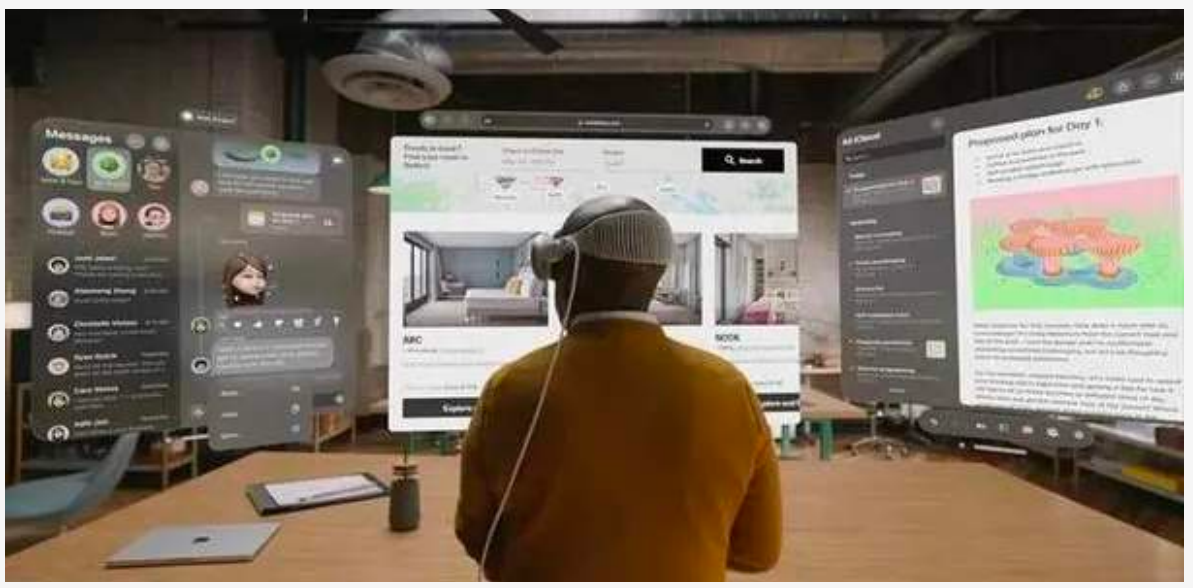
早期研发和技术发明对于引领生态系统发展至关重要

将技术转让给商业团队，并通过部署过程中的收获来影响未来的研究

跨平台应用

鸿蒙元服务

Apple Vision



安卓轻应用

CoE Center of Excellence 卓越中心

类似于 CoA, Center of Automation，自动化中心的组织结构，

但更侧重于 **推动高级、AI驱动的技术的交付和创新** 。

https://learn.microsoft.com/zh-cn/power-platform/guidance/coe/overview

针对某焦点领域提供领导力、最佳实践、研究、支持与或培训的一个团队、共享设施或实体

焦点领域可能是技术（如 Java）、商业概念（如 BPM）、技能（例如谈判）或更广泛研究（例如健康）的领域

在组织内 CoE 可以指一群人 / 一个部门 / 一个共享设施

能力中心 Competence Center 或 Capability Center)

1. figure: grouping of centers by types of knowledge and time-focus

Source: (Bryan, 2011)

https://zhuanlan.zhihu.com/p/566816384

数据要素化


智慧交通
智慧产业
智慧农业
智慧金融
智慧文旅
智慧医疗
安全的数据价值应用
交易内容安全检测
可疑线索分析研判
编排与响应
数据交易市场安全
数据可信计算
数据白名单防护
数据追踪溯源
数据可信流通安全
业务访问安全
运营访问安全
数据访问安全
数据要素供给安全
基础网络设施安全
全流程安全监管
多主体协同管理
多视角安全监管
实名制安全检测
场景化分析建模
信任技术保障
身份管理
身份认证
授权管理
责任认定

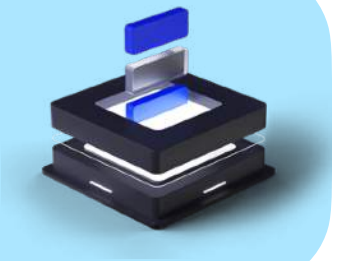
深度学习融合路线

CNN 卷积

RNN 循环

GAN 生成

DQN 强化

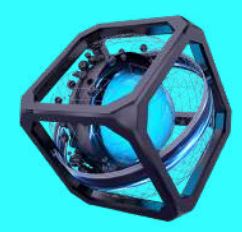


GNN 图神经

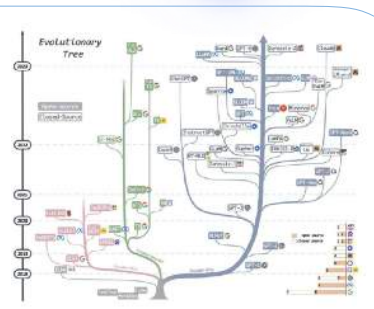
LSTM 长短期

DBN 深度置信

RWKV



Transformer

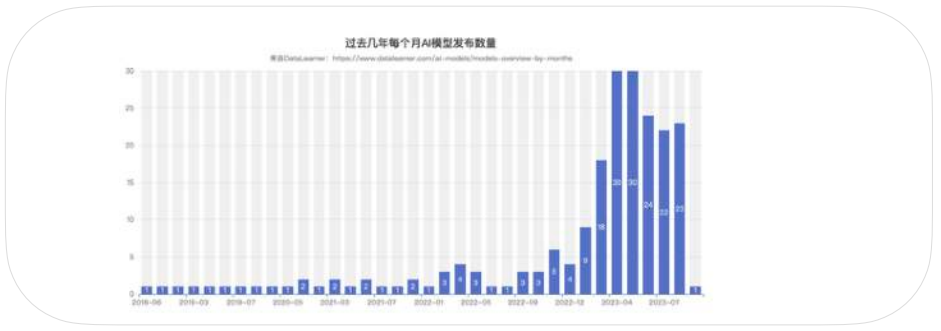


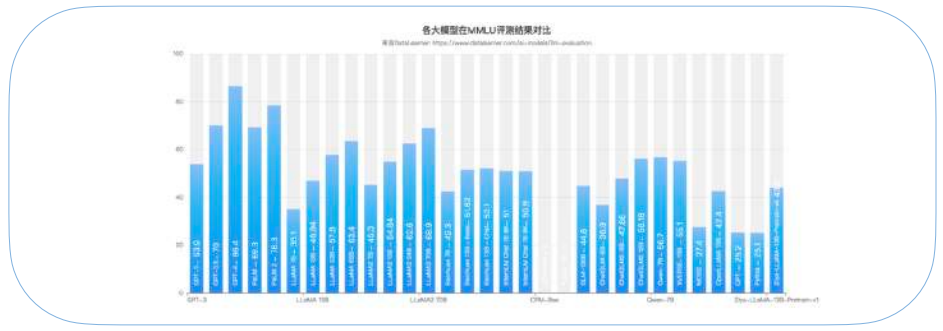
Reward Modeling 奖励模型

Reinforcement Learning

xRunda.com

新模型 新能力





表格能力

TableGPT
Space-T

数学能力

MathGPT
Microsoft *WizardMath*
科大讯飞 *星火认知大模型*
阿里云 *MindOpt Copilot*

图像能力

Segment
Meta SAM, Segment Anything Model
IDEA 研究院 Grounded-SAM
北京智源 SegGPT

3D能力

LumaAI FlyThroughs *iPhone即可录制创建专业的3D立体飞跃视频*

Benchmark

PandaLM	比较大模型回复质量
智源 FlagEval	能力-任务-指标 三维评测体系
SciBench	
L-Eval	对长上下文进行标准化评估
MMLU	目前最著名的大模型语义理解测评之一，由UC Berkeley大学的研究人员在2020年9月推出
C-Eval	全面的中文基础模型评估套件。由上海交通大学、清华大学和匹兹堡大学研究人员在2023年5月份联合推出
AGI Eval	微软在2023年4月推出，主要评测大模型在人类认知和解决问题的一般能力
GSM8K	OpenAI发布的大模型数学推理能力评测基准

声音能力

清华+火山 *SALMONN, Speech Audio Language Music Open Neural Network*

个人管理

Rewind
Notion
Dot

企业案例

麦肯锡 *Lilli*

视频能力

CoDeF, the Content Deformation Field 内容形变场

AI+WEB3融合路线

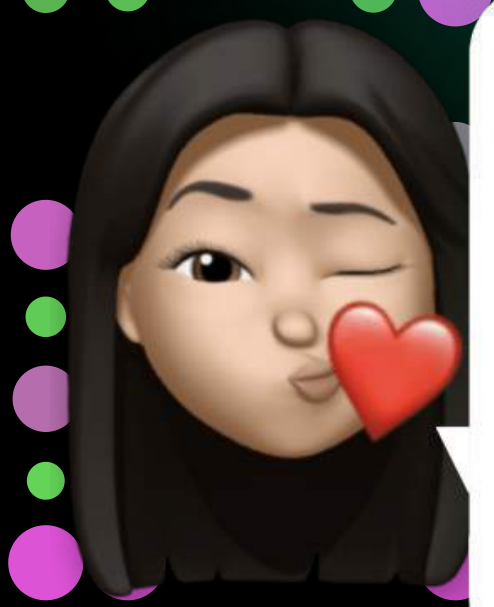
Open Challenges in LLM Research

Reduce and measure hallucinations	减少并测量幻觉
Optimize context length and context construction	优化上下文长度和构建
Incorporate other data modalities	融合其他数据模式
Make LLMs faster and cheaper	大模型降本提速
Design a new model architecture	设计新的模型架构
Develop GPU alternatives	开发GPU替代品
Make Agents usable	提升智能体可用性
Improve learning from human preference	根据人类偏好改进
Build LLMs for non-English languages	构建非英语大模型

更多访问 www.xRunda.com

TOGETHER

TRANSFORMING THE FUTURE WITH INTELLIGENCE



微信扫一扫

WWW.XRUNDA.COM

AOKVEN@XRUNDA.COM

I 8 6 I I I 7 5 0 I I