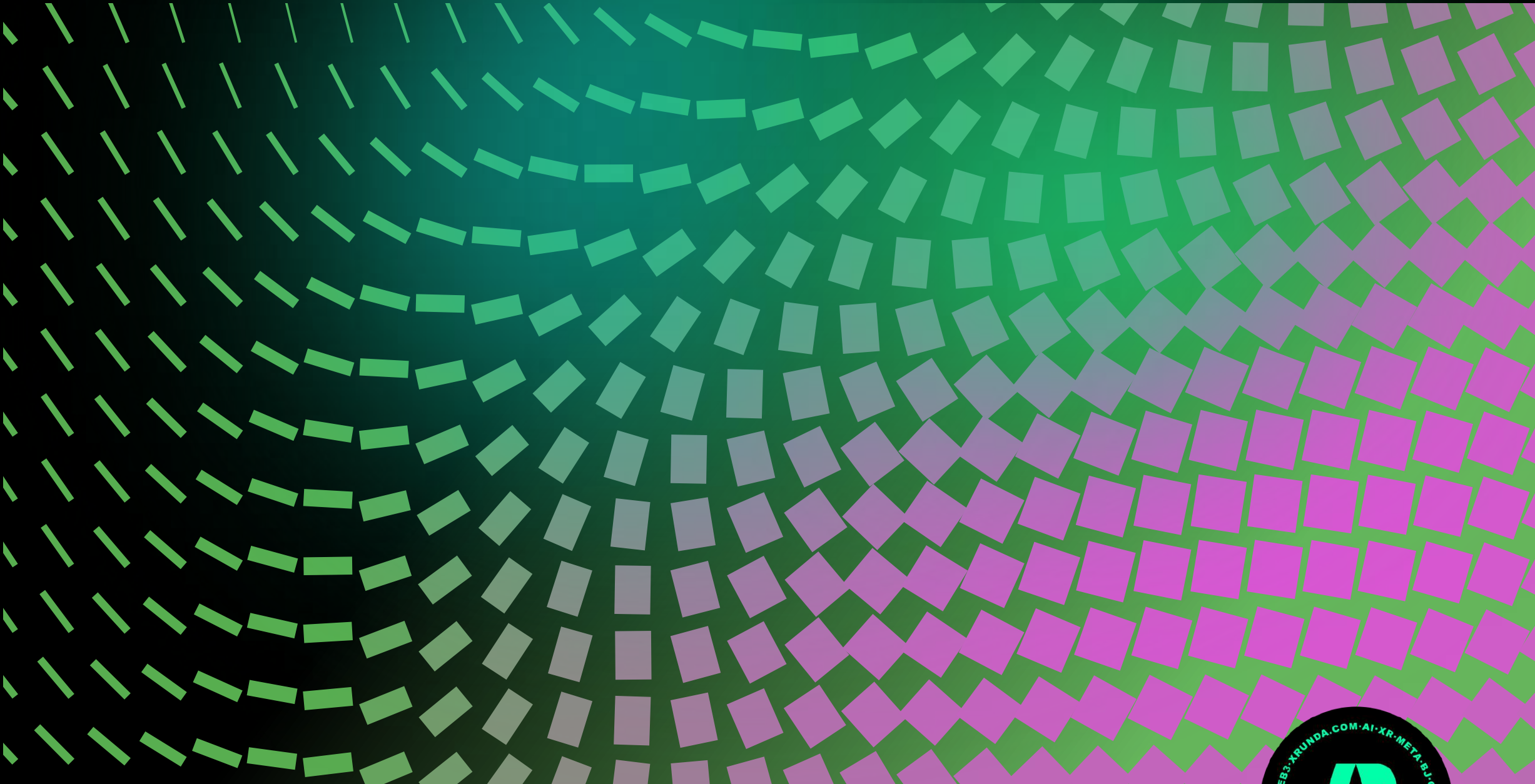


xRunda AI

20231009

TRANSFORMING THE INDUSTRY WITH INTELLIGENCE



A

公司介绍

- P04 · COMPANY
- P05 · OUR TEAM
- P06 · OUR VISION
- P07 · PARTNERSHIPS

www.xRunda.com

B

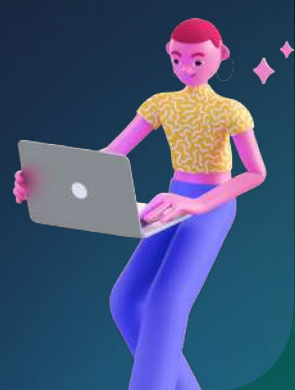
解决方案

- P09 · 解决方案框架
- P10 · 咨询顾问服务
- P13 · 原生应用开发
- P14 · 知识向量引擎
- P15 · 模型微调研发
- P16 · 企业智能化服务
- P17 · 特色能力服务

C

技术方案

- P26 · 技术栈总览
- P27 · 原生应用方案
- P33 · 知识向量引擎方案
- P39 · 模型微调方案
- P53 · 企业智能化方案
- P62 · xMaaS 集成方案



D

产品案例

- | 最新产品 | 实项展示 |
|-------------------|------------|
| P68 · 即摘 xGeekSum | P75 · 行业服务 |
| P69 · 即听 xGPTing | P76 · 孵化项目 |
| P70 · 即试 xGPTtest | P77 · 实验项目 |
| P71 · 即答 xChatDA | |
| P72 · 即调 xTune | |
| P73 · 即画 xDraw | |

E

前沿技术研究

- | | | | |
|------------------------|---------------|-------------|---------------------|
| P79 · Agent | P93 · 音频生成技术 | P102 · 算法问题 | P107 · 数据要素化 |
| P86 · Vector Embedding | P96 · 视频生成技术 | P103 · 具身智能 | P108 · 深度学习融合路线 |
| P88 · MoE | P97 · 数字人生成技术 | P104 · 端侧模型 | P109 · AI+WEB3 融合路线 |
| P89 · Knowledge Graph | P98 · 3D 生成技术 | P105 · 跨平台 | P110 · 新模型 |
| P91 · Multimodal 多模态 | P100 · 安全性 | P106 · 卓越中心 | |
| P92 · 图像生成技术 | P101 · 工程问题 | | |

A

公司介绍 Company Profile

PO4 · COMPANY

PO5 · OUR TEAM

PO6 · OUR VISION

PO7 · PARTNERSHIPS





INTEGRATOR OF INTELLIGENT HYBRID CLOUD SOLUTIONS 智能混合云集成

COMPANY

北京一润一达科技有限公司成立于 2015 年，国家高新技术企业，追求以深度的行业理解和科技洞察力，融合新一代人工智能及云服务，开创面向产业革命和智能资源新生态的解决方案中心，帮助客户和伙伴实现数智化让 AI 应用早日助力企业发展，让 AI 服务造福人类生产与生活。



28+

软著认证

50+

跨领域专家团队

500+

产品创新项目



OUR TEAM

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。

赛博极客工作室

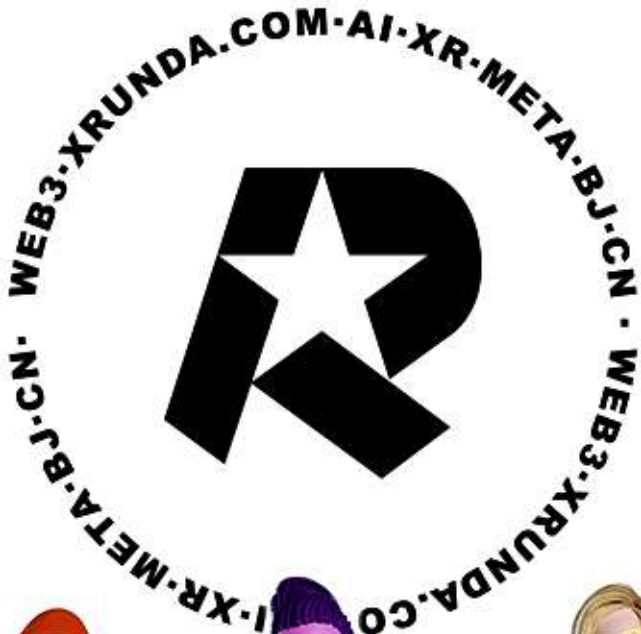
案例制造机

前沿技术实验室

产品探索器

浪尖上共创社群

项目反应炉



项目经理

产品经理

算法工程师

UI/UX设计师

数据运营师

前端工程师

后端工程师

AI工程师

运维工程师

OUR VISION

TRANSFORMING THE FUTURE WITH INTELLIGENCE 智變未來

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.

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

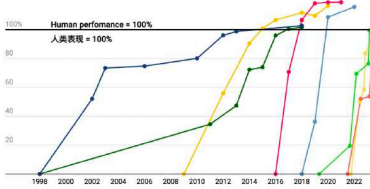
相对于人类表现，在基准测试中具有最先进的人工智能性能

● Handwriting recognition ● Speech recognition ● Image recognition ● Reading comprehension

手写识别 语音识别 图像识别 阅读理解

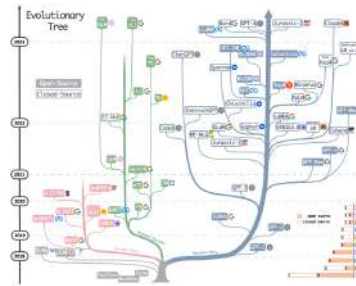
● Language understanding ● Common sense completion ● Grade school math ● Code generation

语言理解 常识完成 小学数学 代码生成



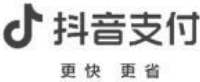
For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point" which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = G5K8K, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI
图表: Will Henshall for TIME 来源: ContextualAI



PARTNERSHIPS

GROWING TOGETHER WITH ECOSYSTEM PARTNERS 生态伙伴共成长



B

解决方案

Solution

P09

解决方案框架

- AI+增长飞轮
- 智能资源生态系统

P10

咨询顾问服务

- 底座甄选服务
- 定制方案服务

P13

原生应用开发

- 大模型原生应用方案

P14

知识向量引擎

- Embedding/RAG方案

P15

模型微调研发

- 模型微调技术方案

P16

企业智化服务

- 企业数智化方案

P17

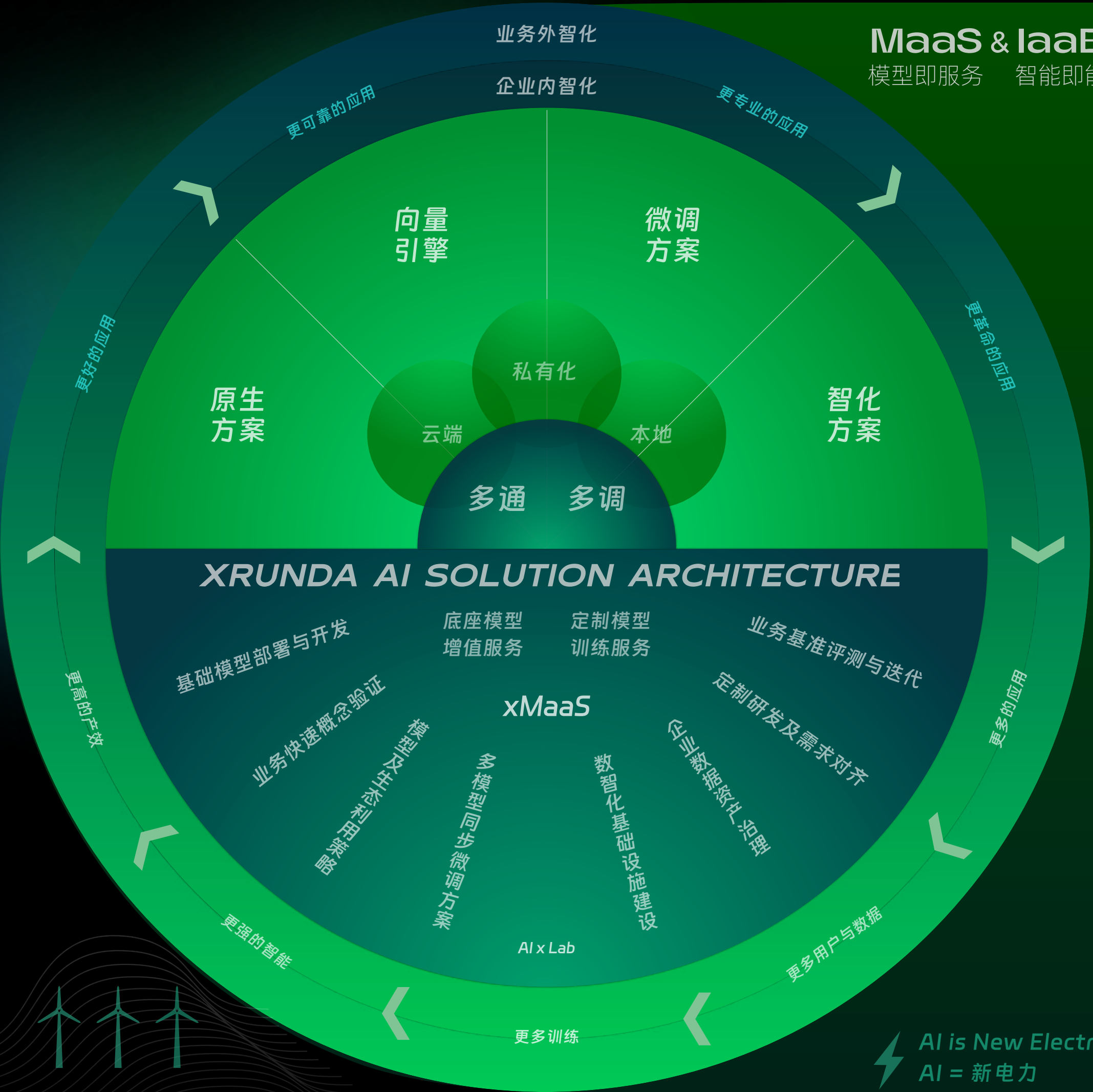
特色能力服务

-  Copilot X
-  xLLMDA 一问多答方案
-  WEB3 服务
-  LangChain
-  xTune 一通多调方案
-  AI x Lab
-  xBenchmark 评测服务

xRunda AI Solution Architecture

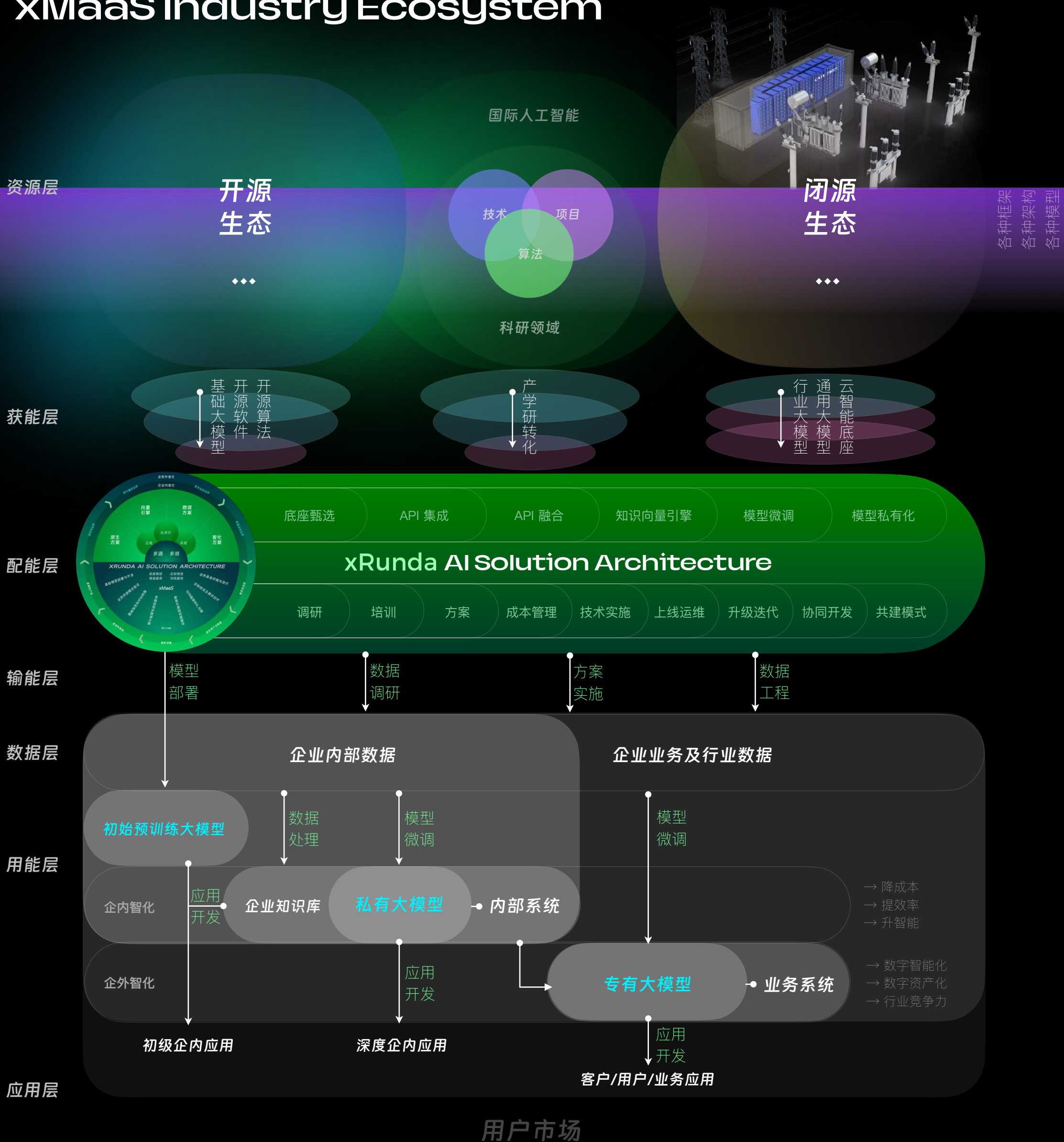
xRundaAI解决方案框架

MaaS & IaaS
模型即服务 智能即能源

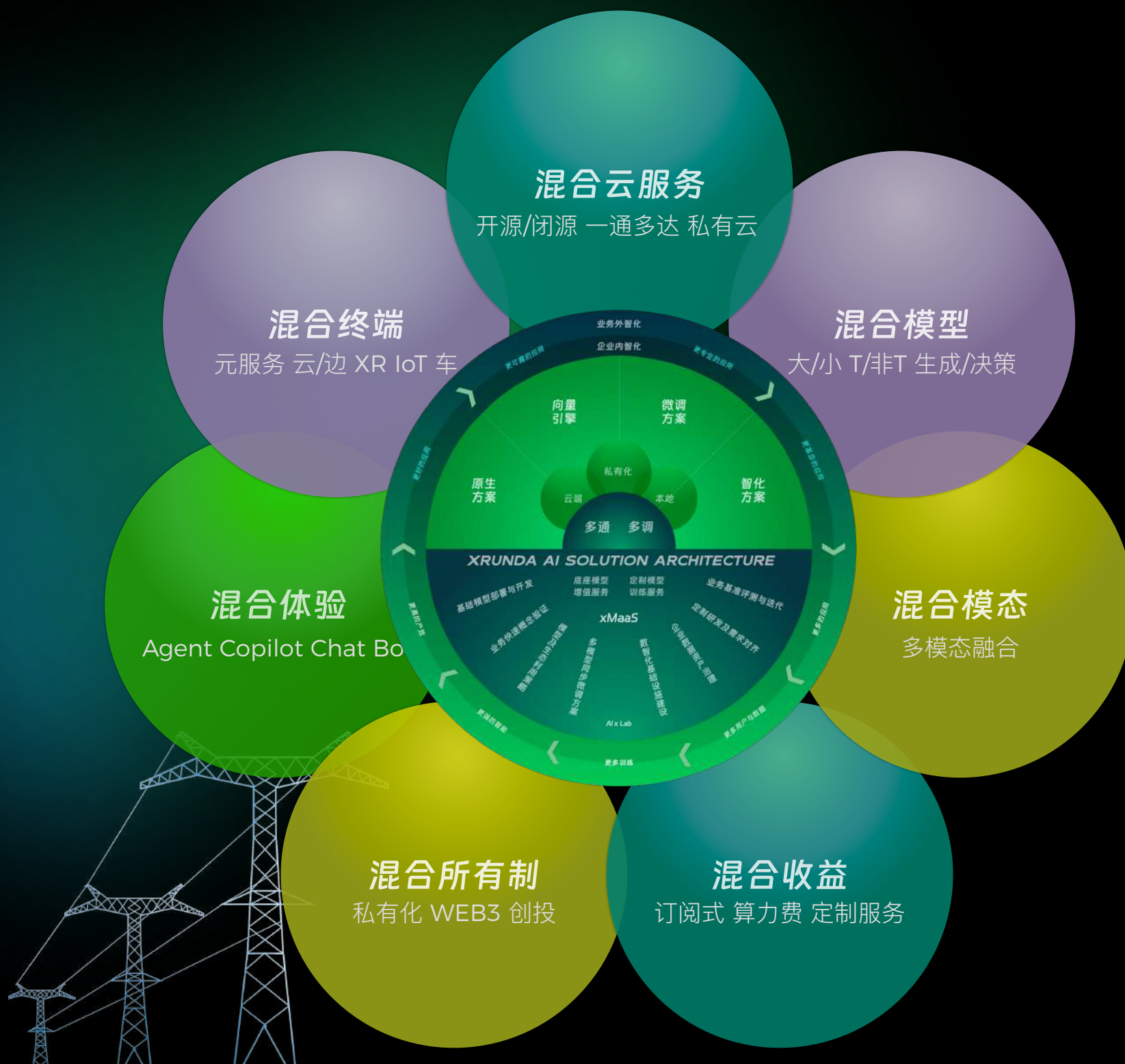


AI is New Electricity
AI = 新电力

国际
国内



混合 MaaS 解决方案



MaaS & IaaS

模型即服务 智能即能源

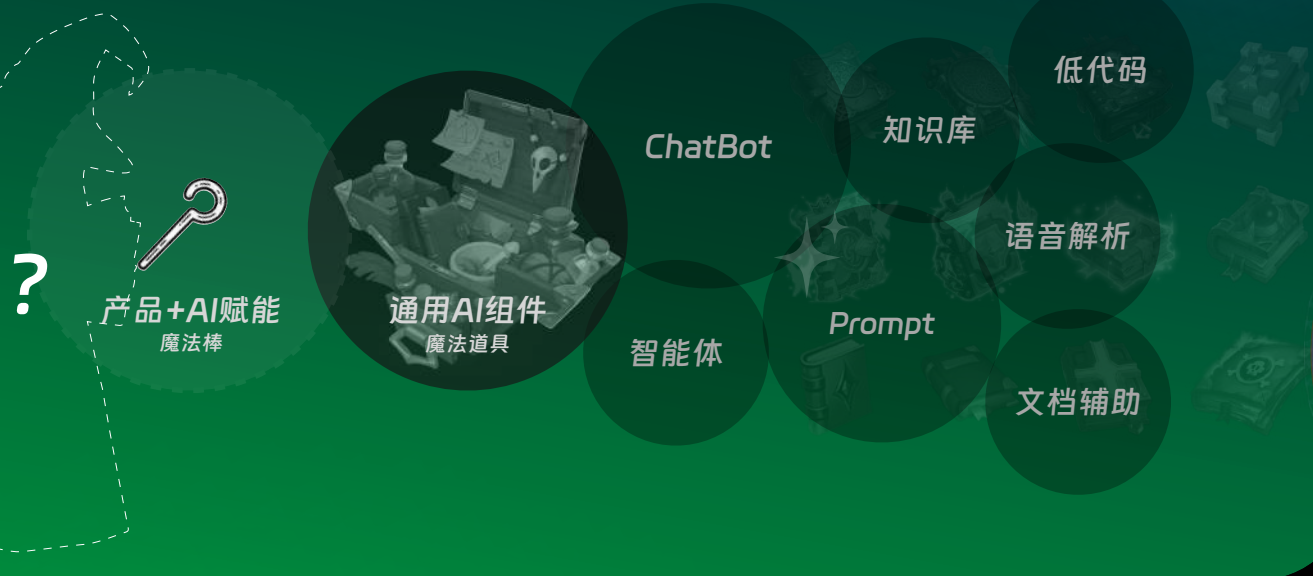
xRunda AI xSaaS Solution

混合 SaaS 解决方案

企业信息化系统+AI

钉钉 / SaaS / SCRM / ERP / SCM / OMS / PLM / BI / OA / ...

AI Magic Store



数智化入门企业

xRunda 定制服务企业数智化

钉钉 / SaaS / SCRM / ERP / SCM / OMS / PLM / BI / OA / ...

AI Magic Story



数智化深化企业

调研 | PoC | 应用开发 | 系统融合 | 资源整合 | 交付 | 运维 | 迭代

企业信息系统数智化定制研发服务
SaaS+AI 的数智化魔法世界



CLICK TO CONTINUE

咨询顾问 服务

Consultant



• 定制方案服务 – P11

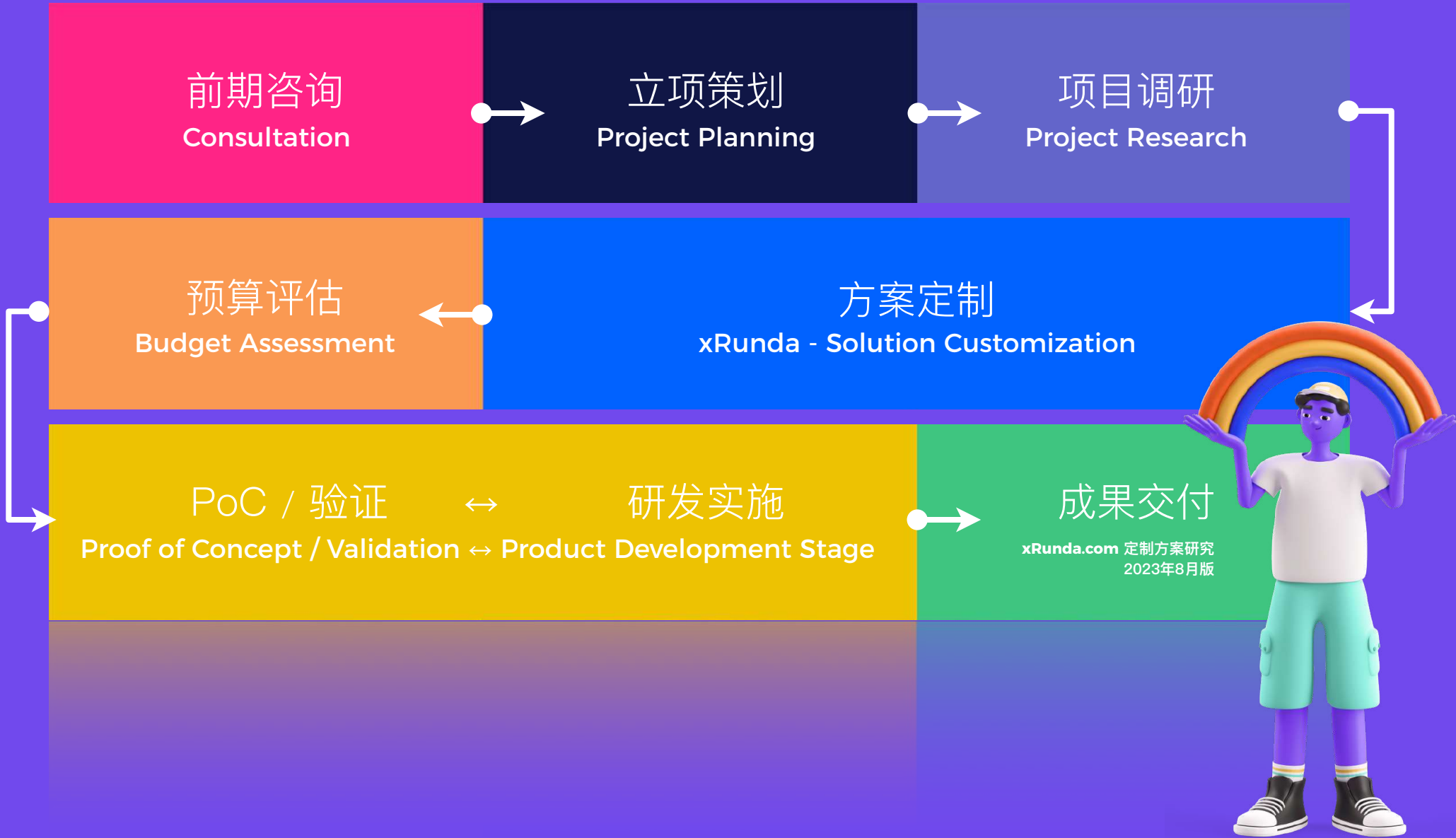
- 定制方案 workflow
- 模型参数规模影响

• 底座甄选服务 – P12

- 主流可选模型
- 其他可选模型
- 国际模型
- 国产模型
- 更多探索
- 模型利用策略

定制方案服务

xRunda Solution Customization



主流模型

百度



文心一言

- 知识增强大语言模型
- 插件市场开放

阿里



通义大模型

- 局部开源
- Qwen-7B
- 通义大模型架构

讯飞



星火认知大模型

- 多模态
- 行业方案

智谱



GLM

- 局部开源
- ChatGLM2-6B

百川



百川大模型

- 局部开源
- Baichuan-7B/13B

商汤



日日新大模型 SenseNova

- 商量 SenseChat
- 秒画 SenseMirage
- 如影 SenseAvatar
- 琼宇 SenseSpace

StabilityAI



Stable Diffusion

- 开源
- Stable Diffusion XL

xRunda.com 模型研究
2023年8月版

其他模型

国际模型

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

国产模型

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



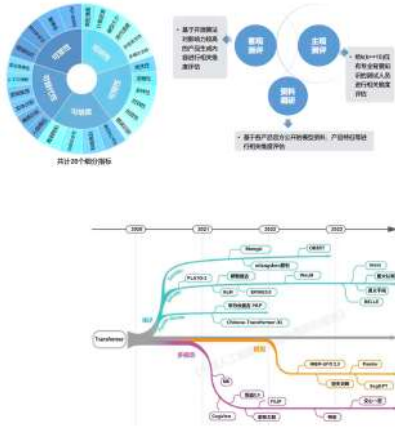
更多探索

- 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

开源大模型

浅研探索

6B+

- 自然语言交互轻应用
- 终端设备内置

私有化

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

100B+

- 原生应用
- 知识库方案

云端私有化

60B+

- 企业信息
系统升级

100B+

- 大型企业
智能化改造

本地私有化

10B+

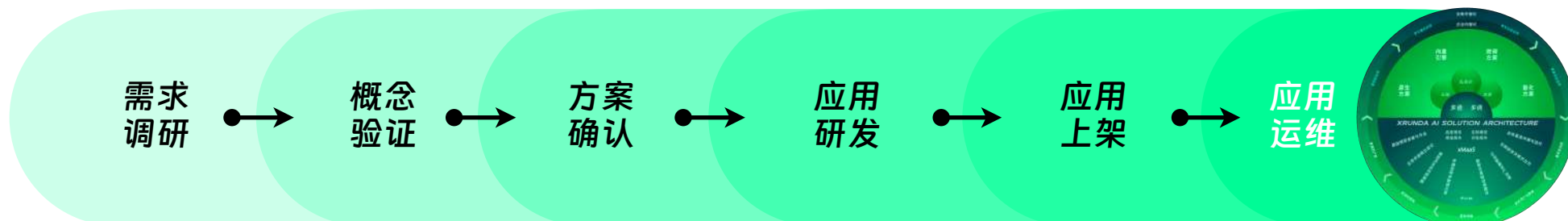
- 小微功能应用
- 敏感业务功能
升级

30B+

- 轻度智能应用
- 企内局部功能
升级

Native Application 方案

原生方案工作流

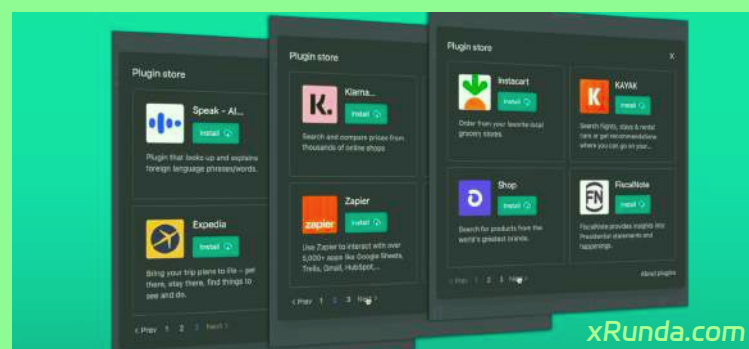


Prompt Engineering

提示工程



LLM Plugins



Function Calling 函数调用

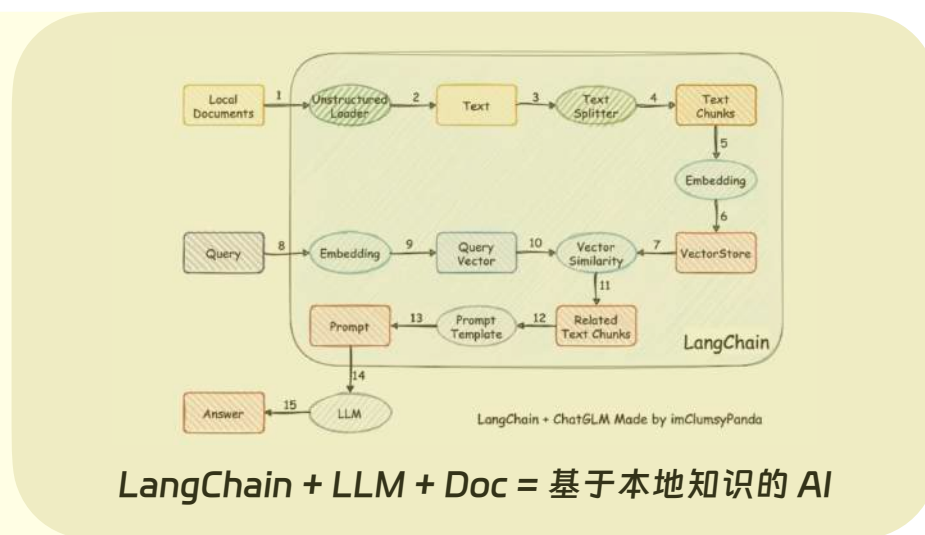
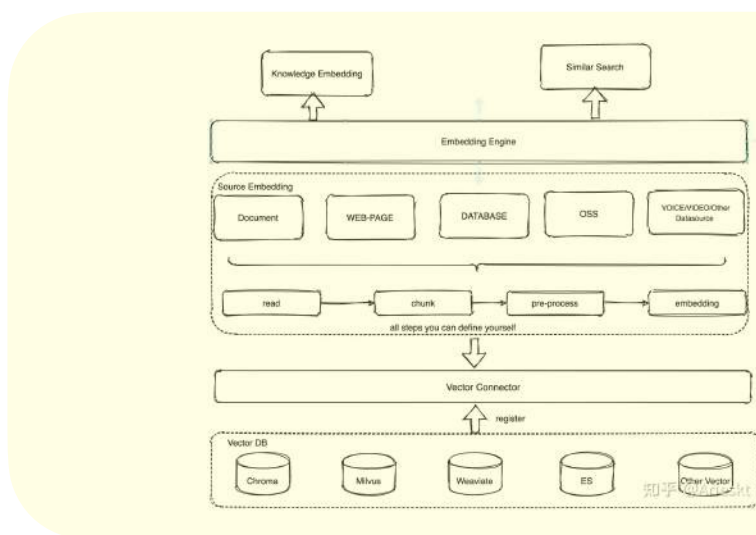
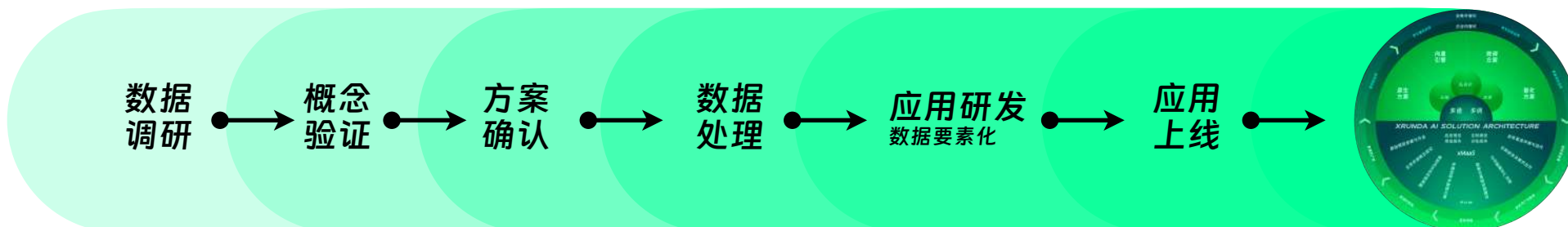


Advanced Data Analysis

高级数据分析



向量引擎方案 workflow



Chunk

内容処理



Embedding

Recall

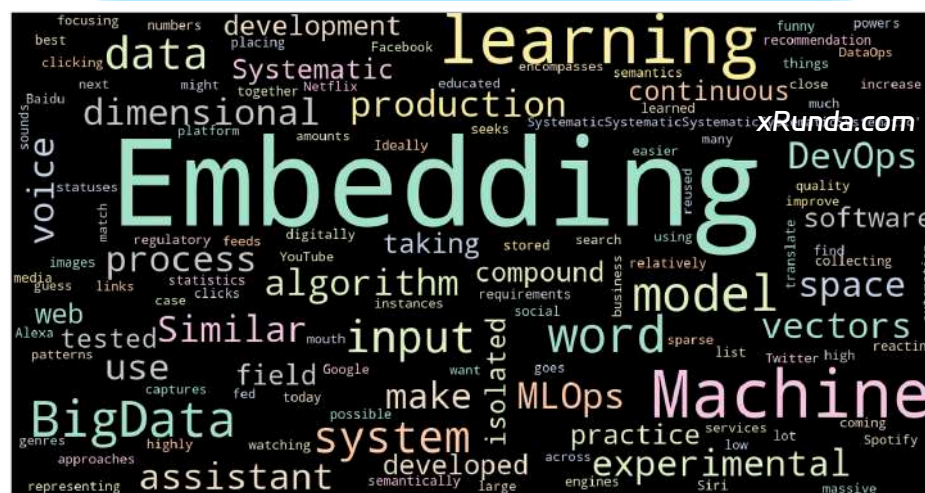


Knowledge Base



Text-to-SQL

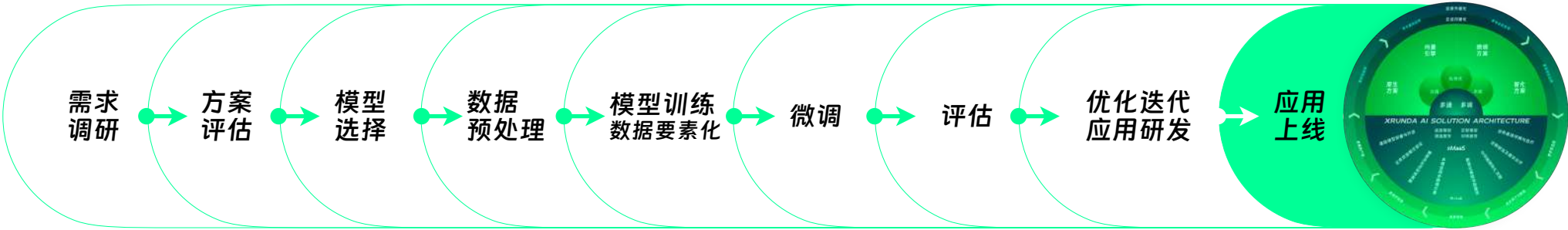
数据库交互



模型微调研发

Fine-Tuning 方案

微调方案 workflow



- 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
- 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



Distinctive
Capabilities and Services
特色能力服务

特色能力
服务

xCopilot



智能开发副驾



Efficient Development Intelligent Teams



Copilot X
The revolution
of AI-support for
developers

xRunda

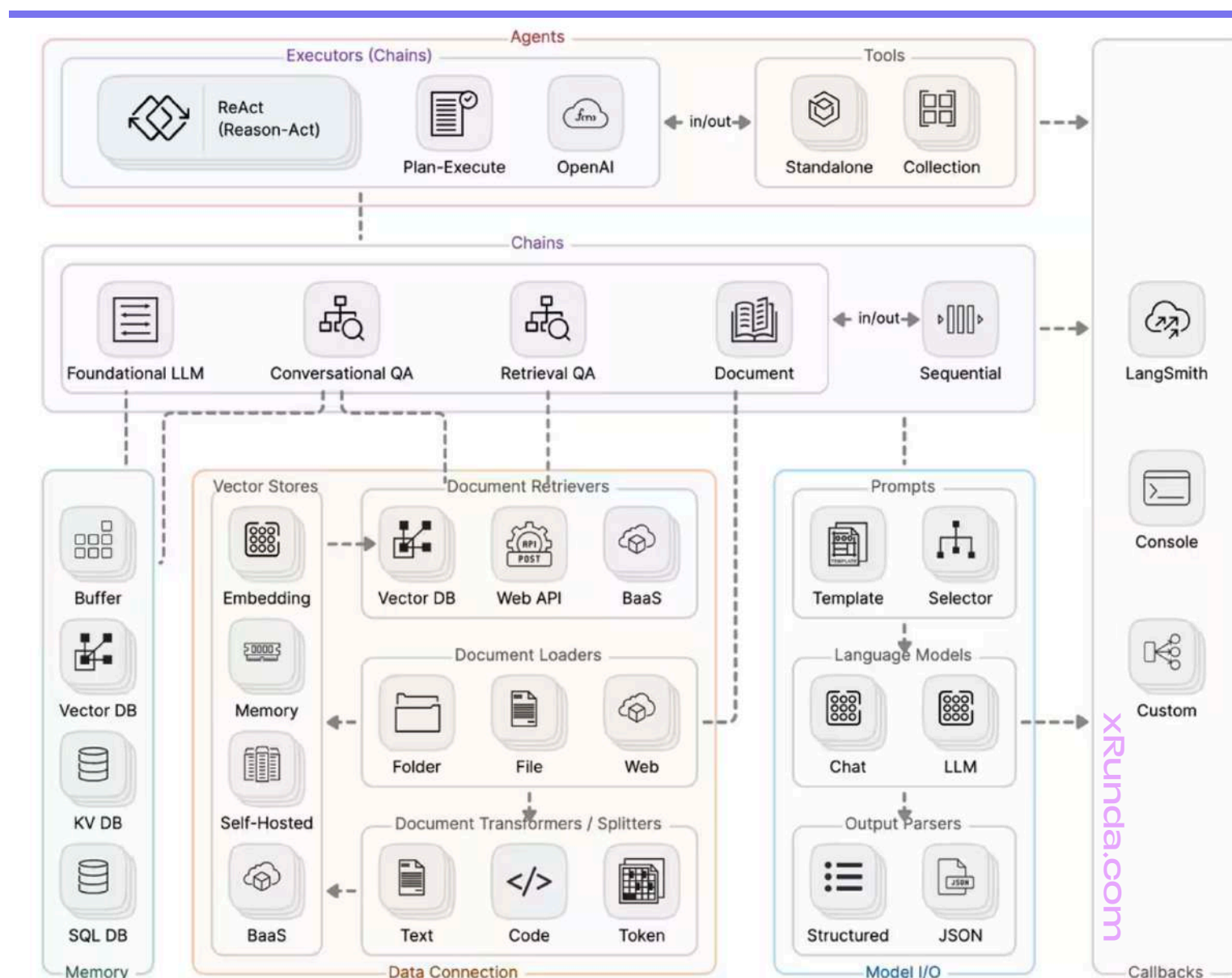


高效率开发 智能化团队



LangChain

LangChain 赋能及集成模块助力 xPoC



特色能力服务

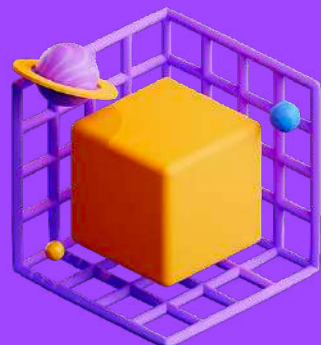


AI x Lab



Frontier Technology Lab

前沿技术实验室



xRunda.com

Large 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



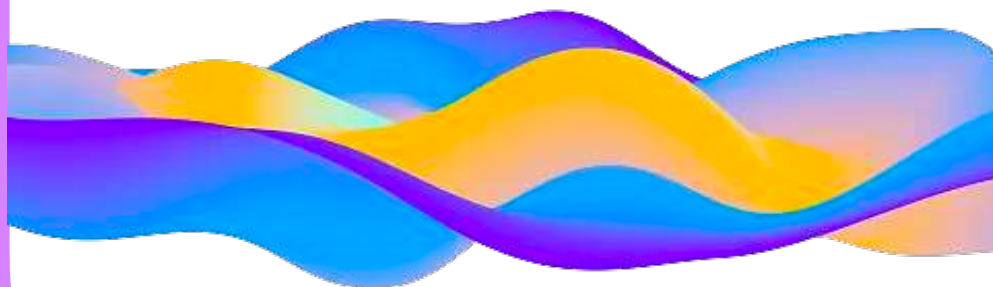
多模型API同步调用

搜索辅助



数据管理

xRunda.com



大模型调度增强



服务托管





xTune 一通多调方案



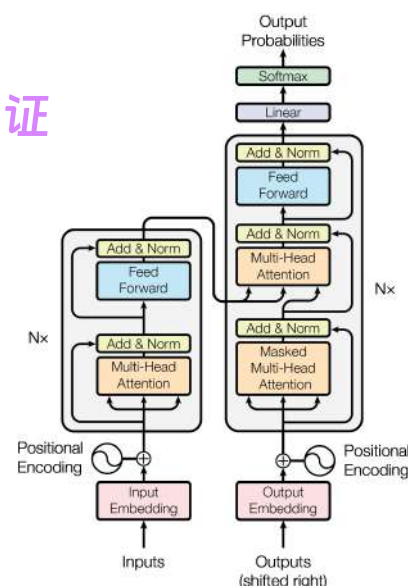
数据处理



多模型同步微调



多微调同步验证



xRunda.com

服务托管



共建模式





中文通用大模型基准测评

- ❖ 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

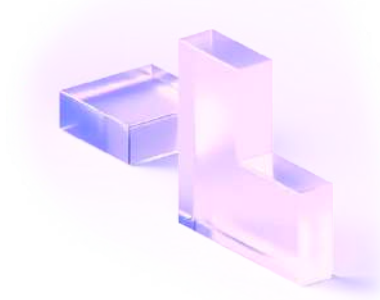


MELT

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

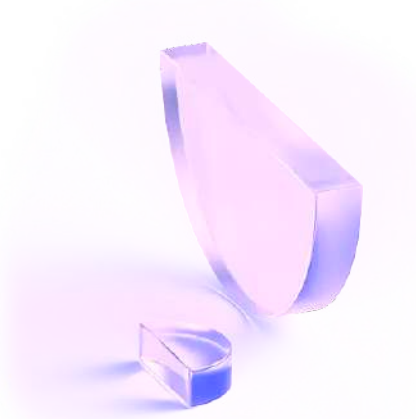


AGILE

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



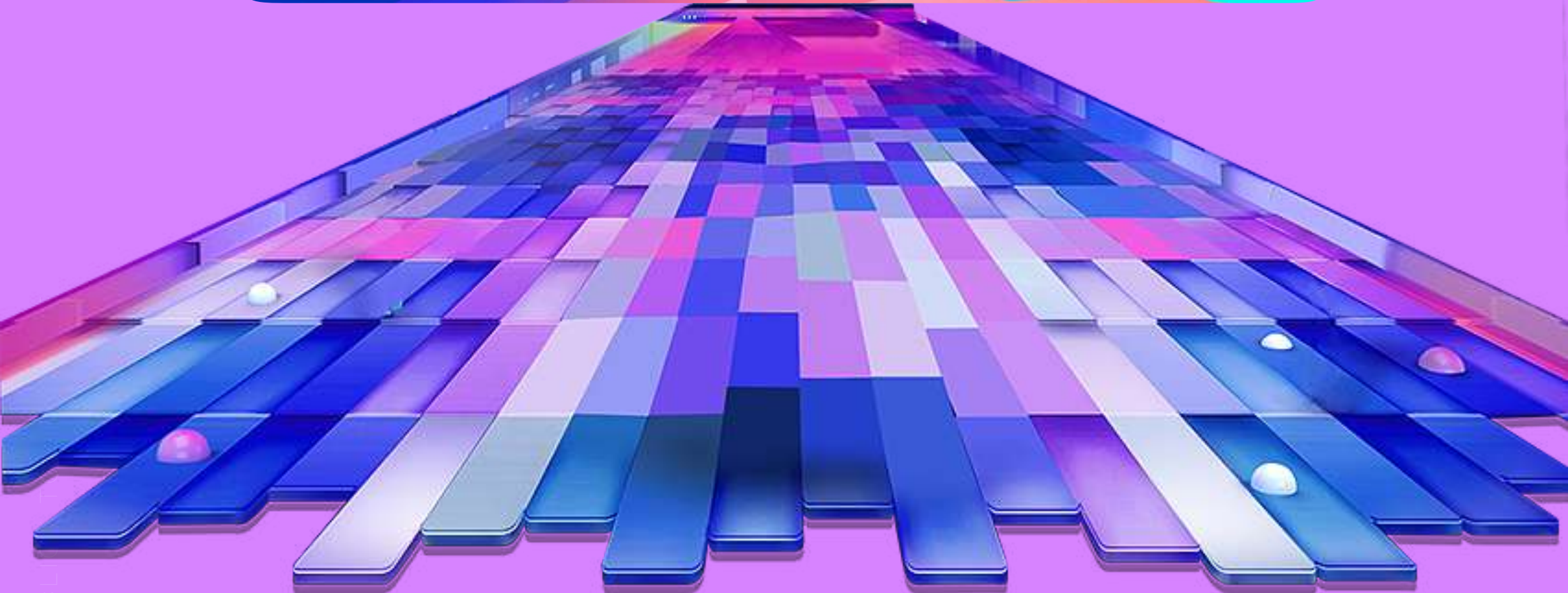
B

solution
解决方案

特色能力
服务



WEB3 融合服务





技术方案 Technical Solutions

P27

原生应用技术方案

- 新兴 LLM 应用堆栈
- Prompt Engineering
- LLM Plugins
- Function Calling
- Code Interpreter

P33

知识向量引擎方案

- Chunk
- Embedding
- Recall
- Knowledge Base
- RAG

P39

模型微调技术方案

- Pre-training
- Prompt Tuning
- Prefix Tuning
- Adapter Fusion Tuning
- Instruction Tuning
- SFT
- PEFT
- P-Tuning
- Freeze
- LoRA
- QLoRA
- Knowledge Distillation

P53

企业智化方案

- 行业企业数智化
- 行业大模型

P62

xMaaS 集成方案

- LMOps 方案
- xMaaS 服务
- MaaS 平台



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	NFT	Binance Smart Chain	高性能区块链开发	多模态应用开发	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		OpenZeppelin	NFT标准和合约开发		OpenCV	图像处理
后端开发	Java	企业级应用、API开发	DAO	Truffle	区块链开发环境	知识库	TensorFlow Image	图像识别
	Python	数据处理、API开发		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		Multi-Party Computation	数据隐私保护		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	Langchain	SPARQL	图谱查询
	iOS (Swift, Objective-C)	iOS端原生开发		More...	More...		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	More...	More...	More...	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				More...	More...	More...
小程序开发	WeChat Mini Program	微信小程序开发	More...	More...	More...	More...	More...	More...
	Alipay Mini Program	支付宝小程序开发						
More	More...	More...	More...	More...	More...	More...	More...	More...





原生应用方案



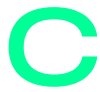
P28 新兴LLM应用程序堆栈

P31 Function Calling

P29 Prompt Engineering

P32 Advanced Data Analysis

P30 LLM Plugins

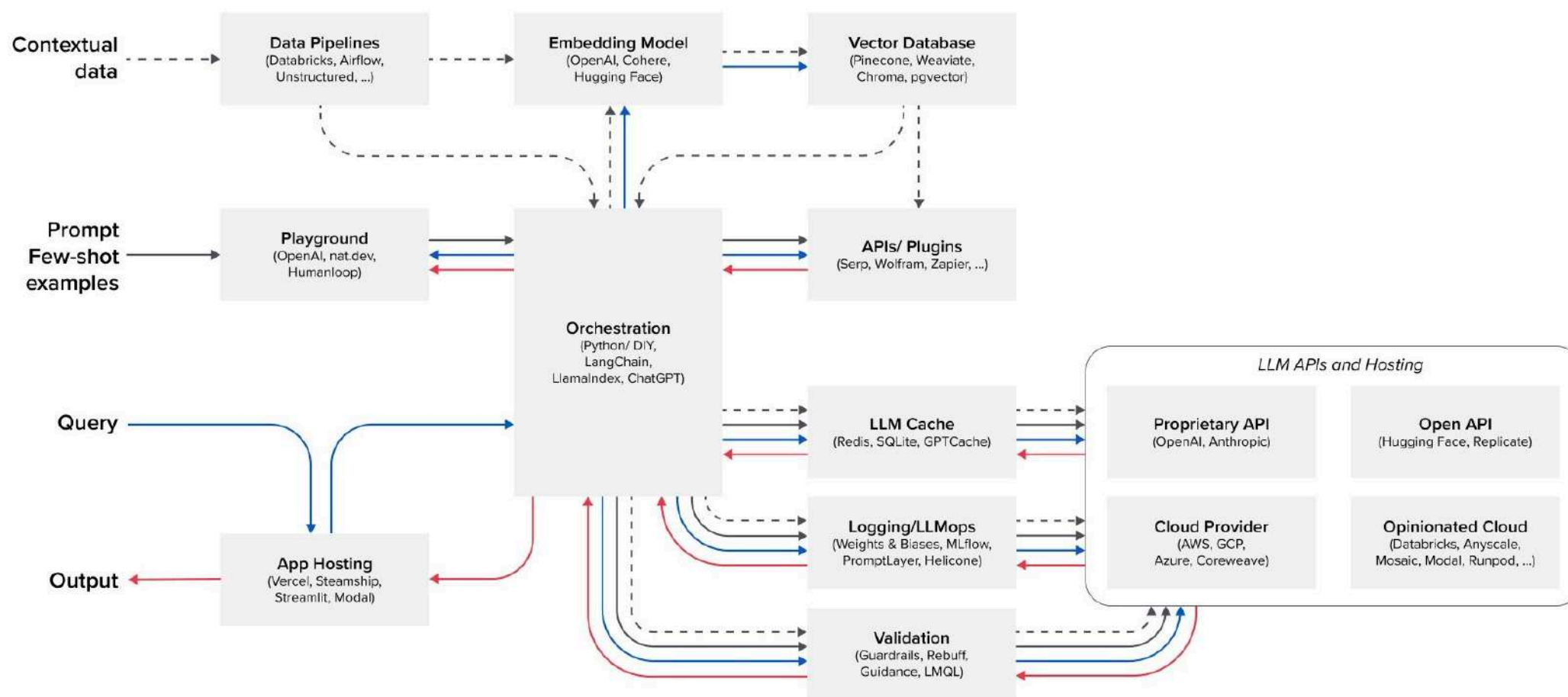


Emerging LLM APP Stack

新兴大模型应用程序堆栈



Emerging LLM App Stack



LEGEND

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

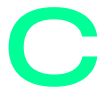
Arrows show the flow of data through the stack

- - - -> 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



Generative Artificial Intelligence

2023 生成式人工智能 能力解析



会看 Visual Understanding



图像识别 (Image Recognition)
人脸识别 (Face Recognition)
对象识别 (Object Detection)
场景理解 (Scene Understanding)
姿态估计 (Pose Estimation)
图像分割 (Image Segmentation)
视频分析 (Video Analysis)
视觉问答 (Visual Question Answering)
跨模态学习 (Cross-Modal Learning)
多模态 (Multimodal)

会读 Reading



语义理解 (Semantic Understanding)
文本分类 (Text Classification)
文本分析 (Text Analysis)
机器翻译 (Machine Translation)
信息提取 (Information Extraction)
文本摘要 (Text Summarization)
信息检索 (Information Retrieval)
问答系统 (Question Answering Systems)

会听 Listening



语音识别 (Speech Recognition)
语音转文本 (Speech-to-Text)
语音翻译 (Speech Translation)
情感分析 (Speech Emotion Analysis)
对话系统 (Dialogue Systems)

会说 Speaking



语音合成 (Speech Synthesis)
语音识别 (Speech Recognition)
对话系统 (Dialogue Systems)
语音翻译 (Speech Translation)
情感识别 (Emotion Recognition)

会画 Drawing



图像生成 (Image Generation)
图像编辑 (Image Editing)
图像增强 (Image Enhancement)
动画生成 (Animation Generation)
3D 建模 (3D Modeling)
数据可视化 (Data Visualization)
手写识别与生成 (Handwriting Recognition and Generation)
场景重建 (Scene Reconstruction)
视频生成 (Video Generation)

能理解 Understanding



自然语言理解 (Natural Language Understanding)
自然语言推理 (Natural Language Inference)
语义分析 (Semantic Analysis)
情感分析 (Sentiment Analysis)
语义搜索 (Semantic Search)
语境理解 (Context Understanding)
实体识别 (Entity Recognition)
关系抽取 (Relation Extraction)
多模态理解 (Multimodal Understanding)

懂逻辑 Logical Reasoning



推理引擎 (Inference Engine)
约束满足问题 (Constraint Satisfaction Problems)
归纳推理 (Inductive Reasoning)
演绎推理 (Deductive Reasoning)
数据分析 (Data Analysis)
模式识别 (Pattern Recognition)
知识表示 (Knowledge Representation)
知识图谱 (Knowledge Graphs)
形式验证 (Formal Verification)

会创意 Innovation and Creativity



音乐生成 (Music Generation)
音效生成 (Sound Effects Generation)
自动编曲 (Automated Composition)
创意生成 (Creative Generation)
产品创新 (Product Innovation)
内容推荐与个性化 (Content Recommendation and Personalization)
智能模拟与预测 (Intelligent Simulation and Prediction)

会写 Writing



自然语言生成 (Natural Language Generation)
文本修正 (Text Correction)
代码生成 (Code Generation)
语音转文本 (Speech-to-Text Writing)

会思考 Thinking



决策支持 (Decision Support)
问题解决 (Problem Solving)
自动规划 (Automated Planning)
机器学习 (Machine Learning)
强化学习 (Reinforcement Learning)
优化算法 (Optimization Algorithms)
模式识别 (Pattern Recognition)
推荐系统 (Recommendation Systems)
自适应学习 (Adaptive Learning)
仿真模拟 (Simulation Modeling)
异常检测 (Anomaly Detection)
知识图谱 (Knowledge Graphs)
预测分析 (Predictive Analytics)
智能体 (Agent)
具身智能 (Embodied Intelligence)

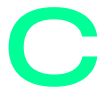


Prompt Engineering 提示工程



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





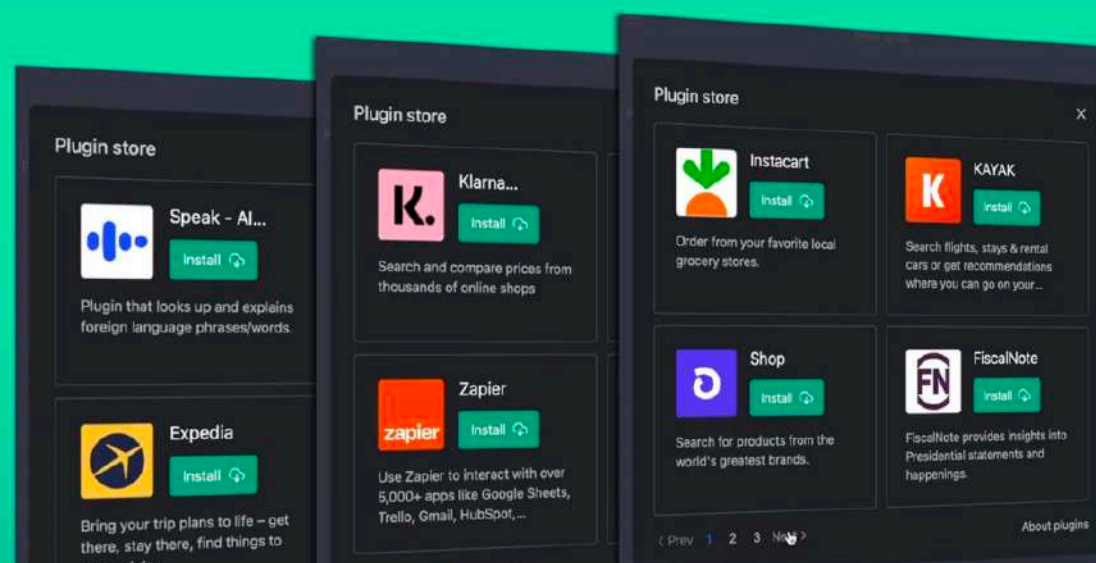
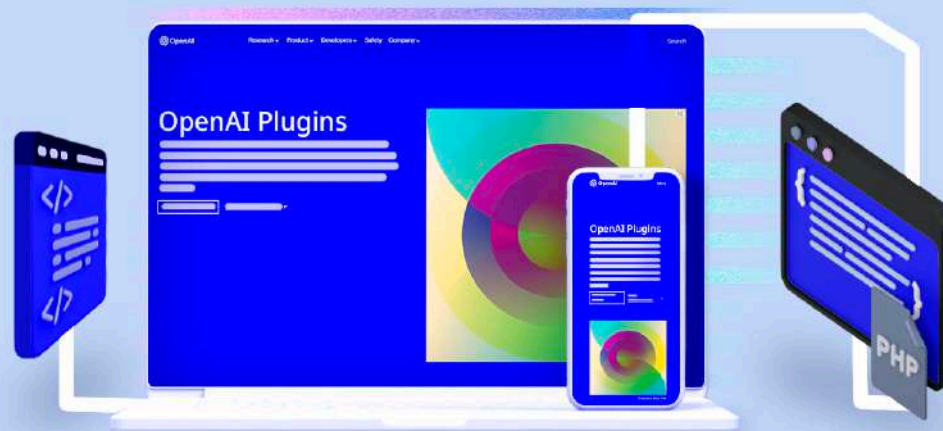
LLM Plugins 模型插件

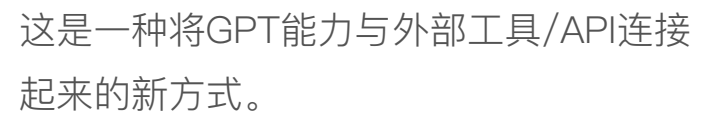
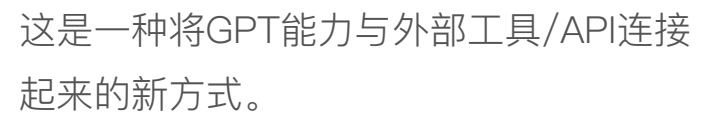
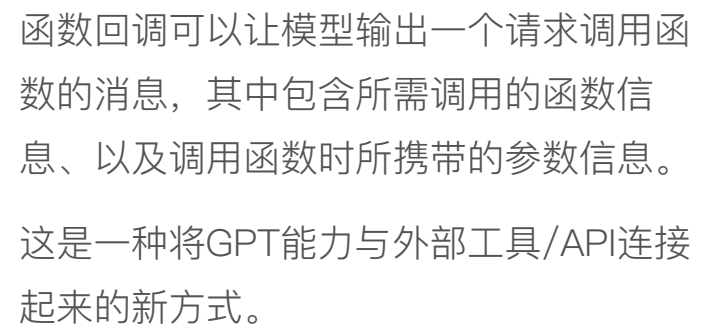
Plugins are tools designed specifically for language models with safety as a core principle, and help LLMs access up-to-date information, run computations, or use third-party services. 插件是专为以安全为核心原则的语言模型而设计的工具，可帮助大语言模型访问最新信息、运行计算或使用第三方服务。

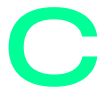
检索实时信息

检索知识库信息

代表用户执行操作

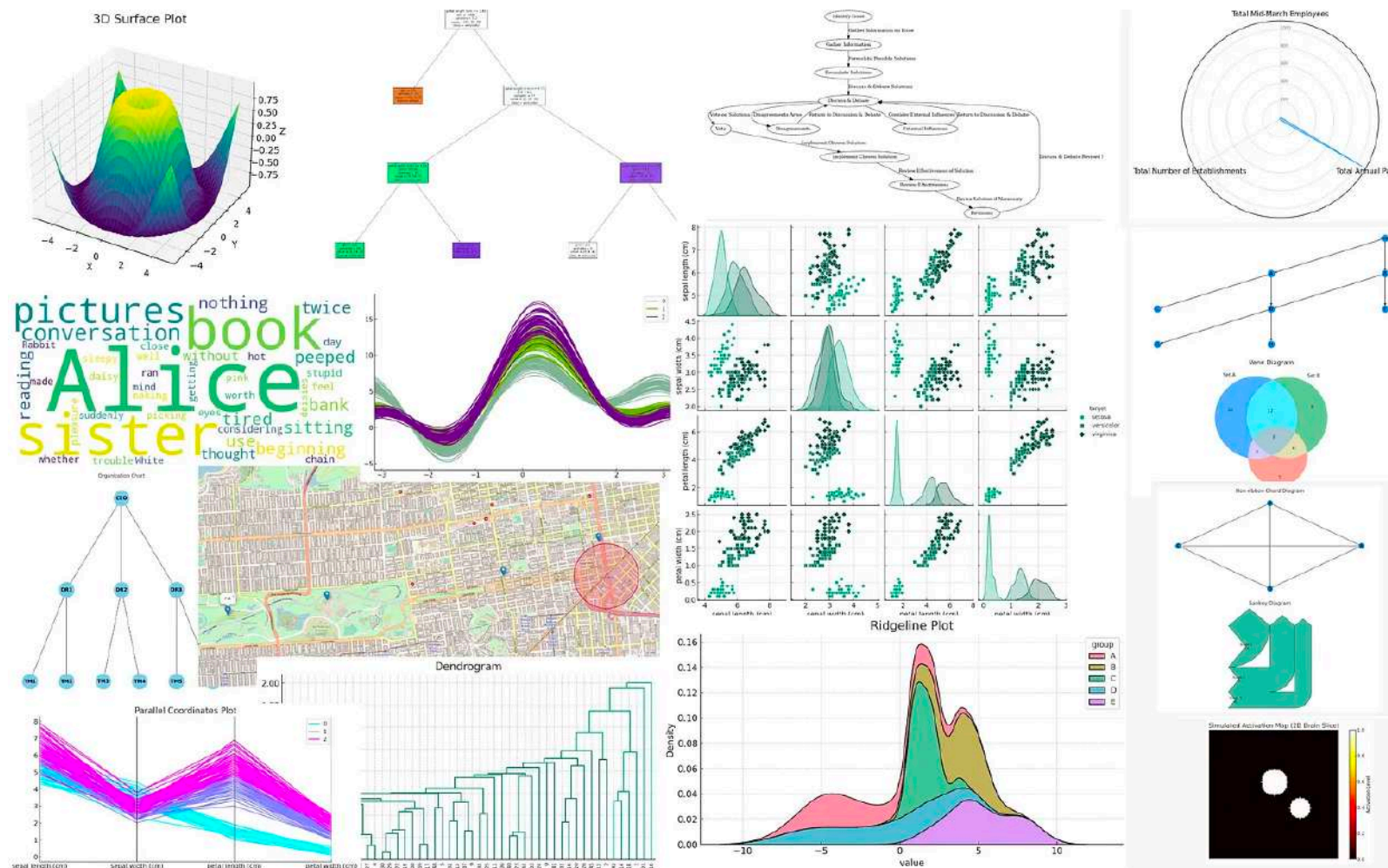






Advanced Data Analysis 高级数据分析

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

Model
Code Interpreter ALPHA

Plugins ALPHA

Default (GPT-3.5)

GPT-4

Legacy (GPT-3.5)

Default (GPT-3.5) with browsing ALPHA

Code Interpreter ALPHA

ChatGPT PLUS

```
To resolve this issue, we need to pass the "-pthread" option to the compiler and linker. However, it appears that the "execute_cpp_program" function does not allow us to specify custom compiler options.

As an alternative, I can write a new function that uses the "subprocess" module to compile and run the C++ program with the "-pthread" option. Let's proceed with this approach.

Finished working Hide code

python
def compile_and_run_cpp_program_with_pthread_option(cpp_code):
    """
    Compiles and runs a C++ program with the -pthread option.

    Args:
        cpp_code (str): The C++ source code to compile and run.

    Returns:
        tuple: A tuple containing the program's output and any error message.

    """
    # Create a temporary C++ source file
    with tempfile.NamedTemporaryFile(mode='w', suffix='.cpp', delete=False) as cpp_file:
        cpp_file.write(cpp_code)
        cpp_file_path = cpp_file.name

    # Create a temporary executable file
    with tempfile.NamedTemporaryFile(mode='w', suffix='.out', delete=False) as exe_file:
        exe_file_path = exe_file.name

    # Compile the C++ program with the -pthread option
    compile_command = ['g++', '-pthread', '-o', exe_file_path, cpp_file_path]
    compile_process = subprocess.run(compile_command, capture_output=True, text=True)
    if compile_process.returncode != 0:
        # Compilation failed
        return (compile_process.stdout, compile_process.stderr)

    # Run the compiled executable
    run_command = [exe_file_path]
    run_process = subprocess.run(run_command, capture_output=True, text=True)

    # Clean up temporary files
    os.remove(cpp_file_path)
    os.remove(exe_file_path)
```



知识向量引擎方案



P35 Parsing and Chunk

P36 Embedding

P37 Recall

P38 Knowledge Base

P39 RAG & Text2SQL

Parsing and Chunk 解析分块

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

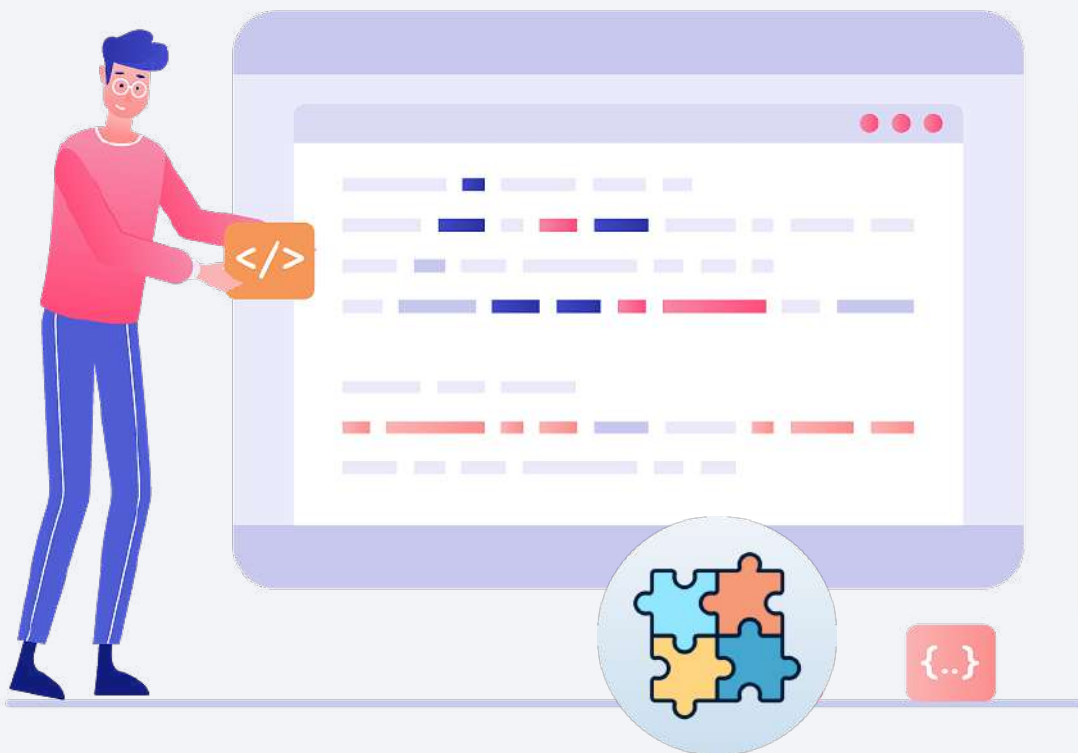
KMeans

聚类

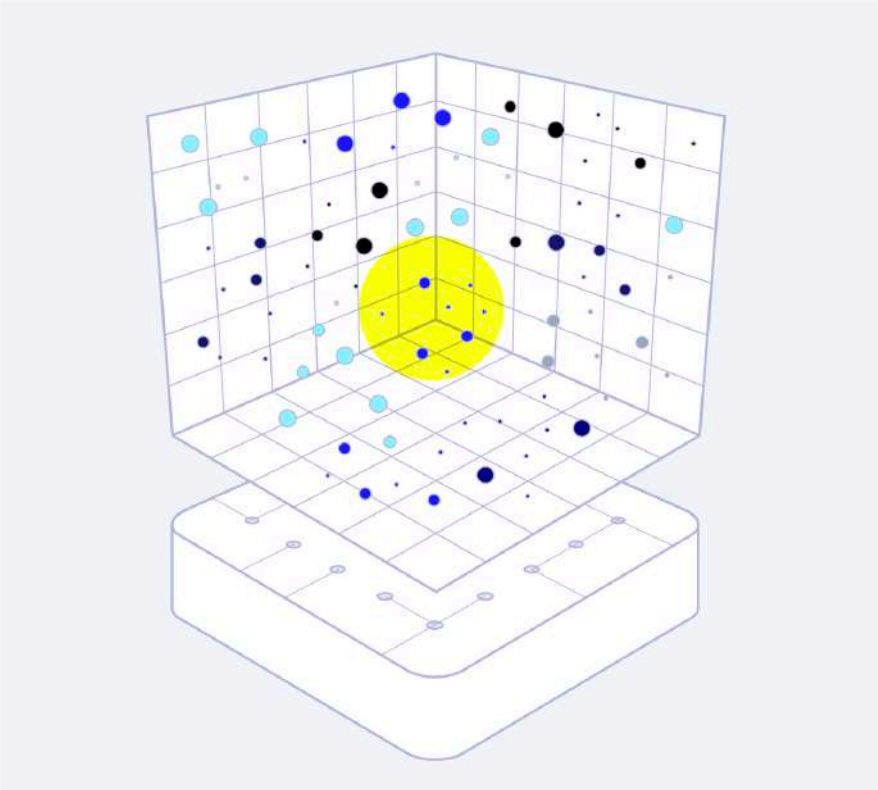
基于语义相似度对句子进行分组的技术。通过使用句子嵌入和 K-means
等聚类算法，可以实现句子聚类。

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

更多访问 www.xrunda.com

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

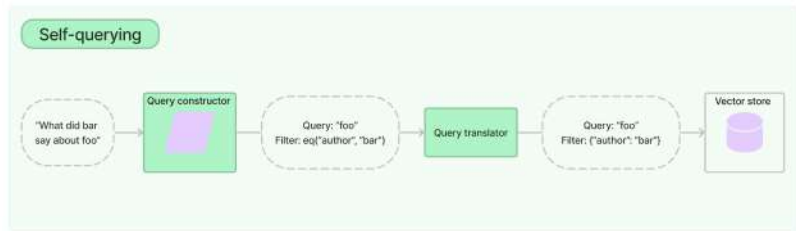
Hamming distance
汉明距离

生成结果方案

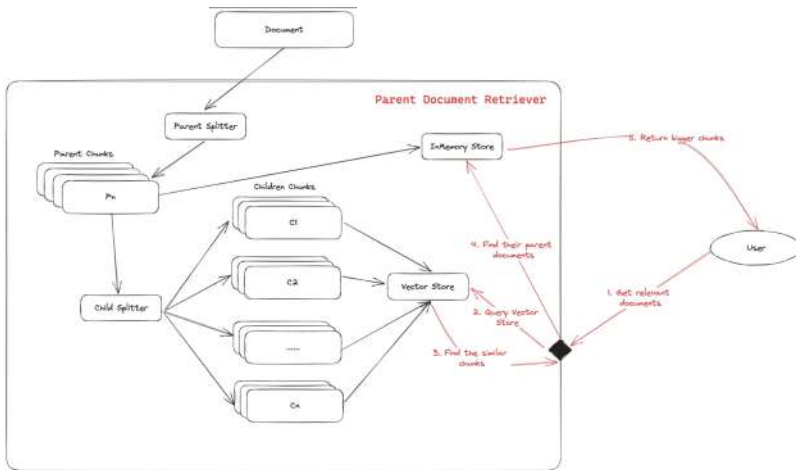
Stuff Refine Map Reduce Map Rerank



优化召回



- 关键词提取
- 父本目录
- Text2QA
- 聚类技术
- 项目表征任务
-



xRunda.com

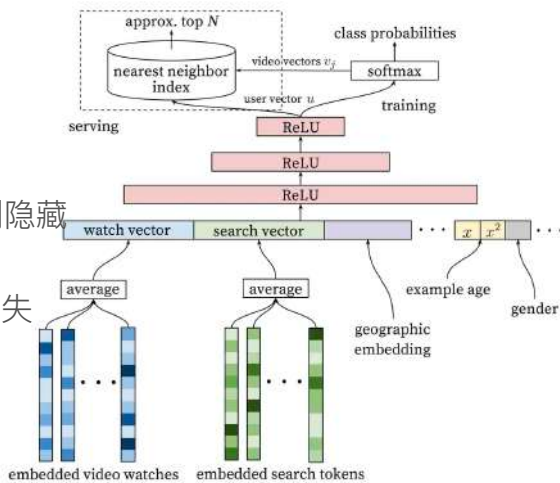
推荐系统等算法召回

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





Knowledge Base 知识库



数据预处理

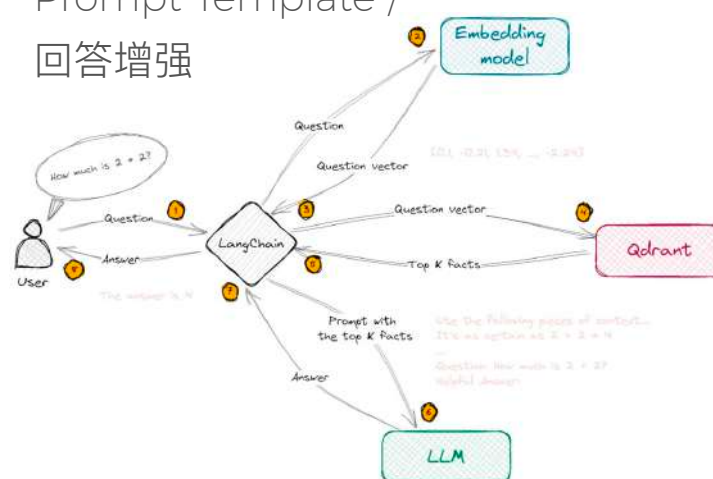
PDF / MARKDOWN /
WORD / PPT / HTML /
CSV / FROM

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



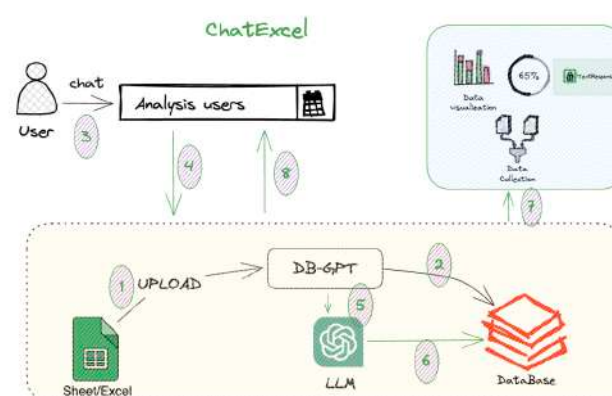
数据接入

- 系统改造 / 数据改造 -> 迭代优化

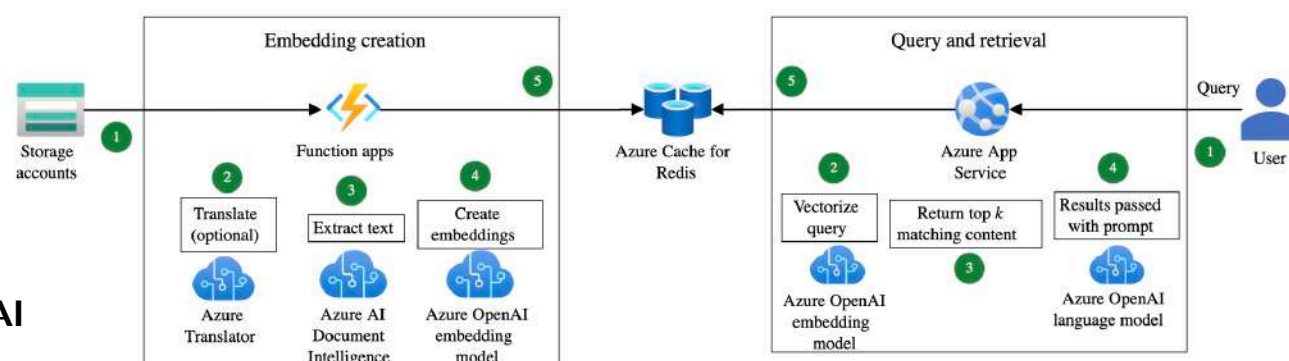
方案参考

DB-GPT

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



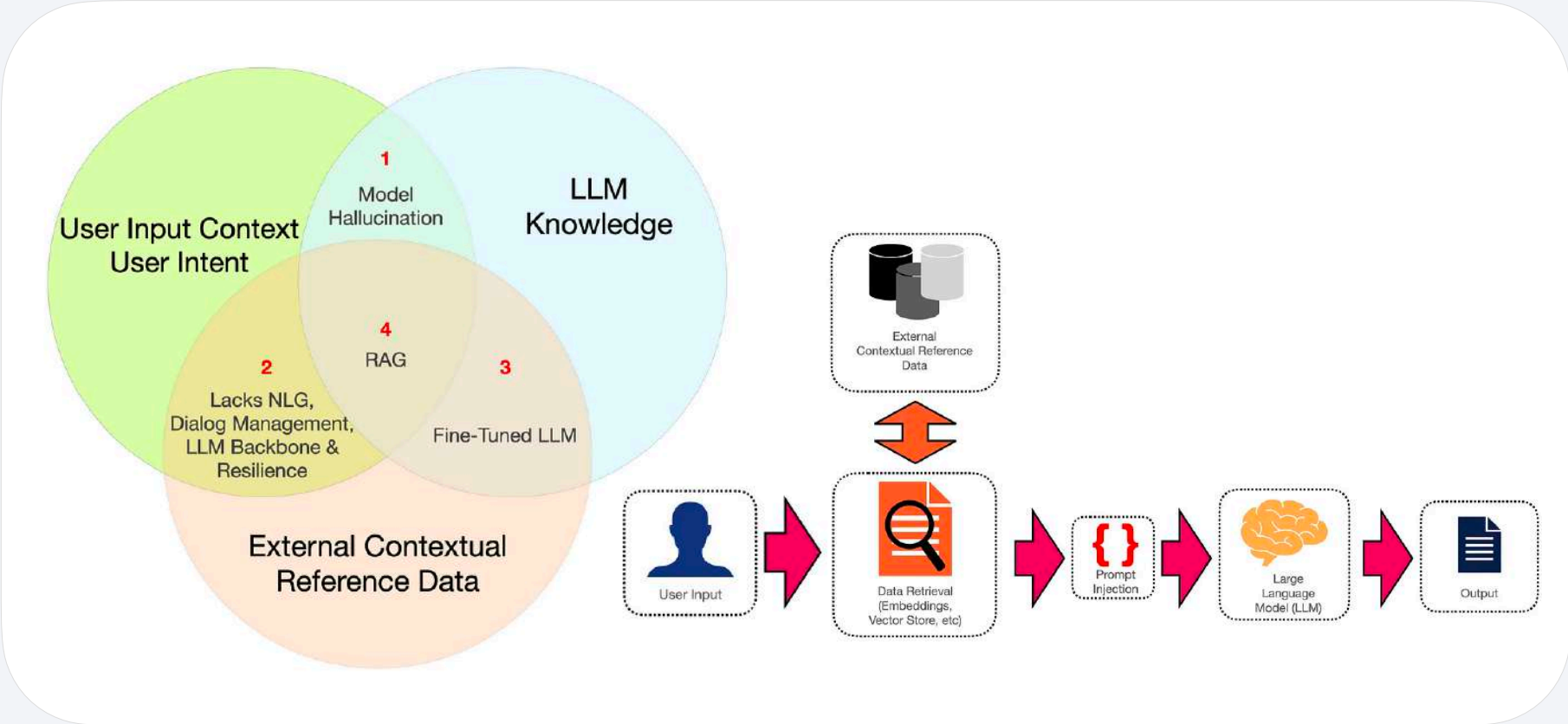
Azure OpenAI



RAG & Text2SQL 检索增强生成

RAG, Retrieval-Augmented Generation

检索增强生成 一种将检索和生成相结合的框架，用于处理知识密集型的生成任务，比如问答、摘要等。它可以利用外部知识库，根据输入的问题或文本，检索出相关的文档或段落，然后将它们作为上下文，用一个预训练的生成模型来生成答案或摘要



Text2SQL

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

TableQA

SPACE-T 表格问答引擎 @AliCloud

金融	政务	零售	能源	工业	办公	...
公共云			专有云			
阿里云智能客服 - 表格问答引擎						
Text-to-SQL			TableNLG			
预训练模型SPACE-T						
大规模表格知识						

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



P41 Pre-training

P42 Instruction Tuning

P43 SFT

P44 Freeze

P45 PEFT

P46 Prompt Tuning

更多访问 www.xRunda.com

P47 Prefix Tuning

P48 Adapter Fusion Tuning

P49 P-Tuning

P50 LoRA

P51 QLoRa

P52 Knowledge Distillation

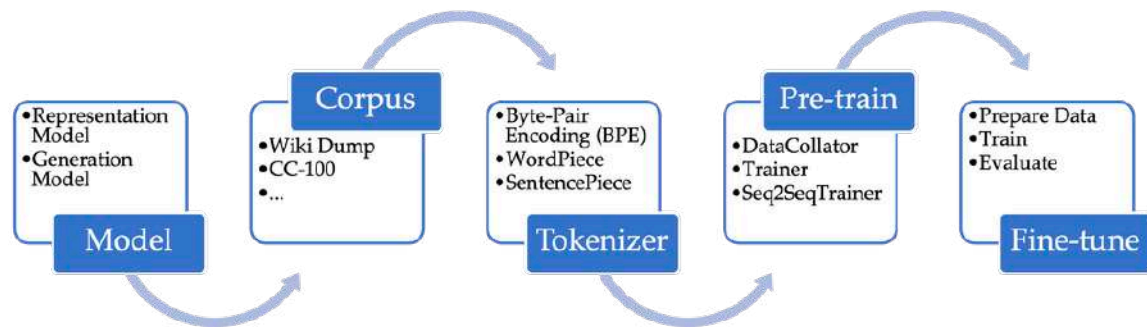
P53 Reinforcement learning



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

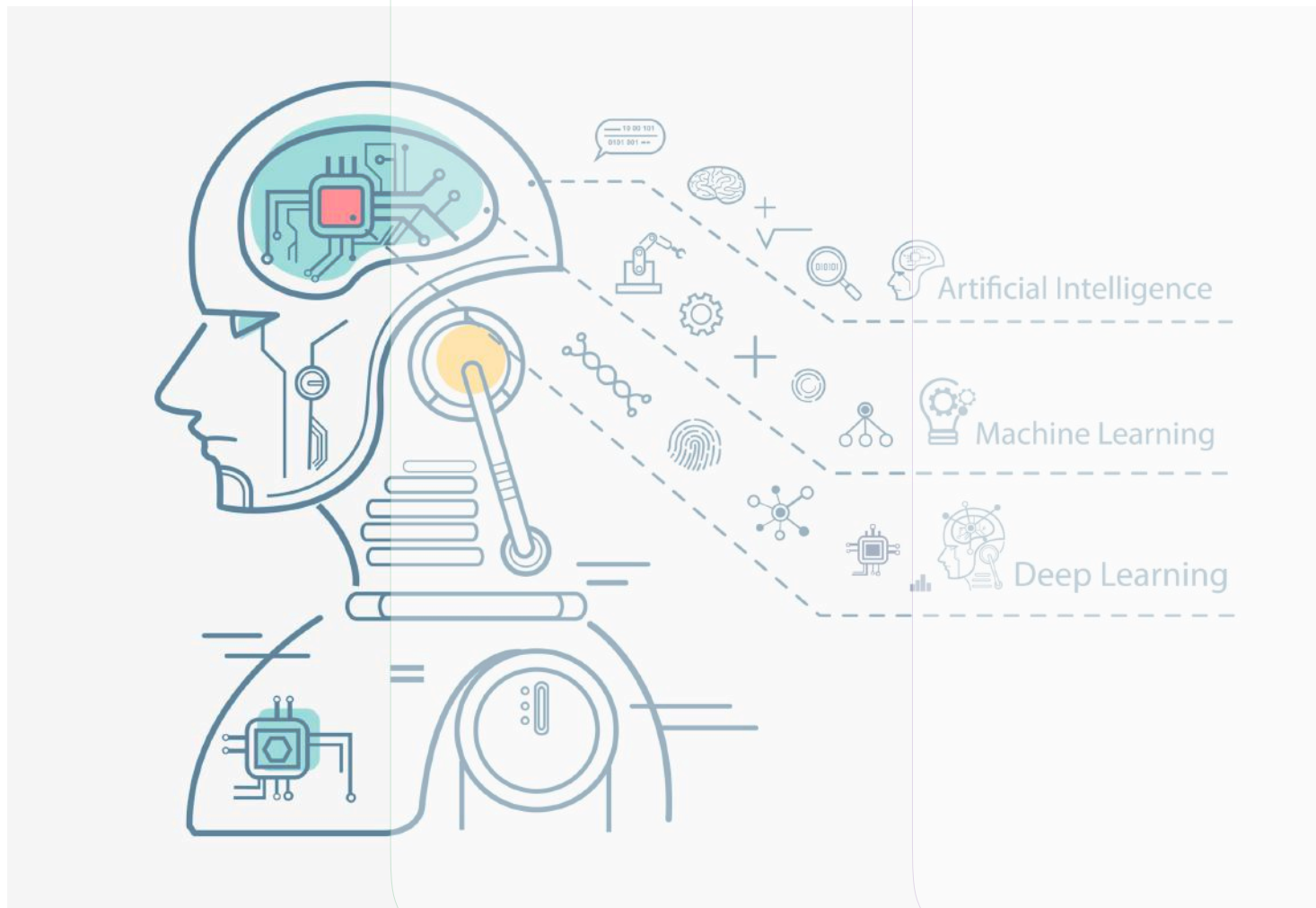


预训练（共建）行业大模型

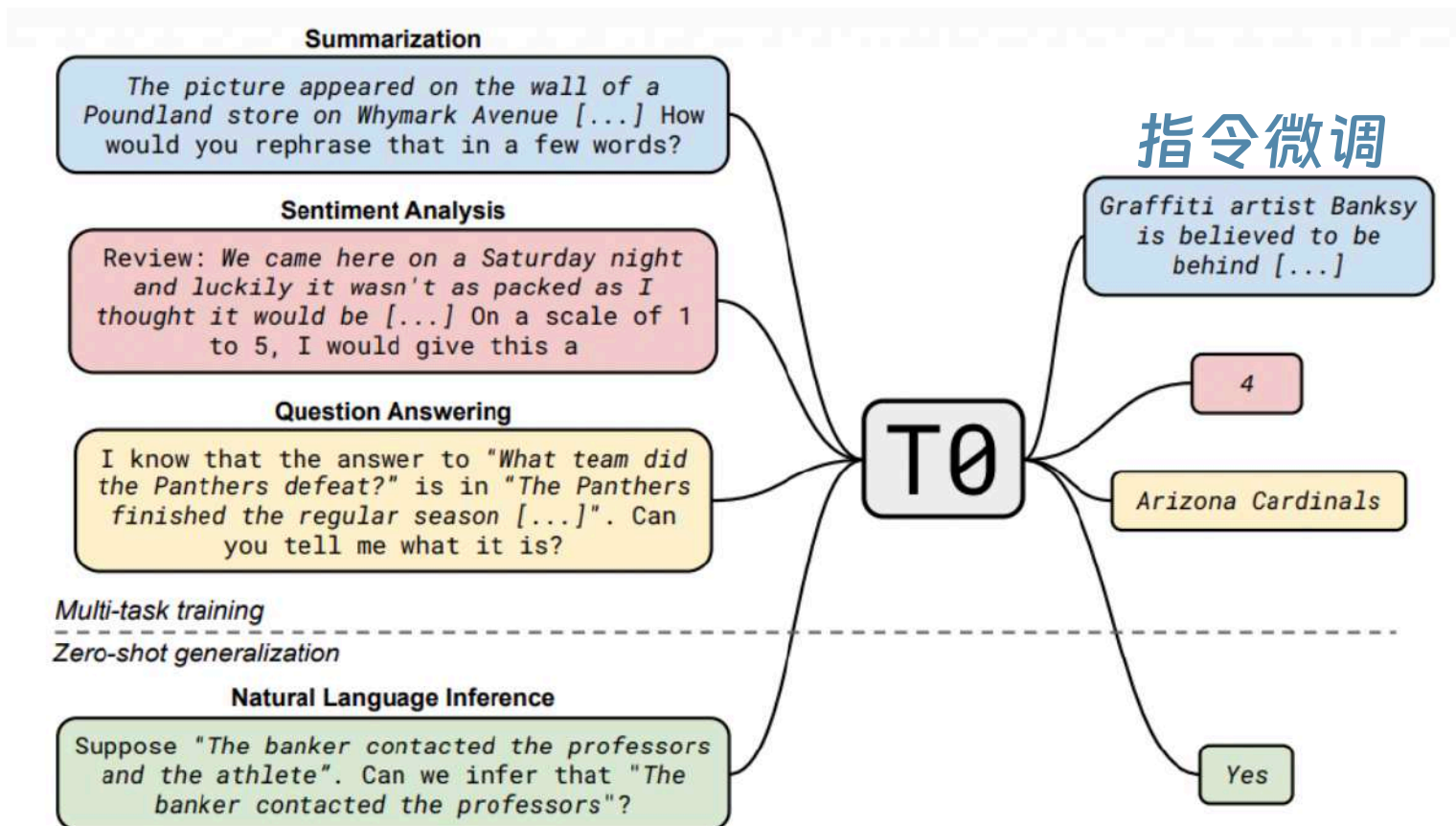
大规模行业知识

海量无结构数据

行业数据挖掘



Instruction Tuning



Instruction Tuning

结合了 Pretrain-finetune 和 Prompting 范式的优点，通过使用微调进行监督，改善语言模型对推理时文本交互的响应能力。

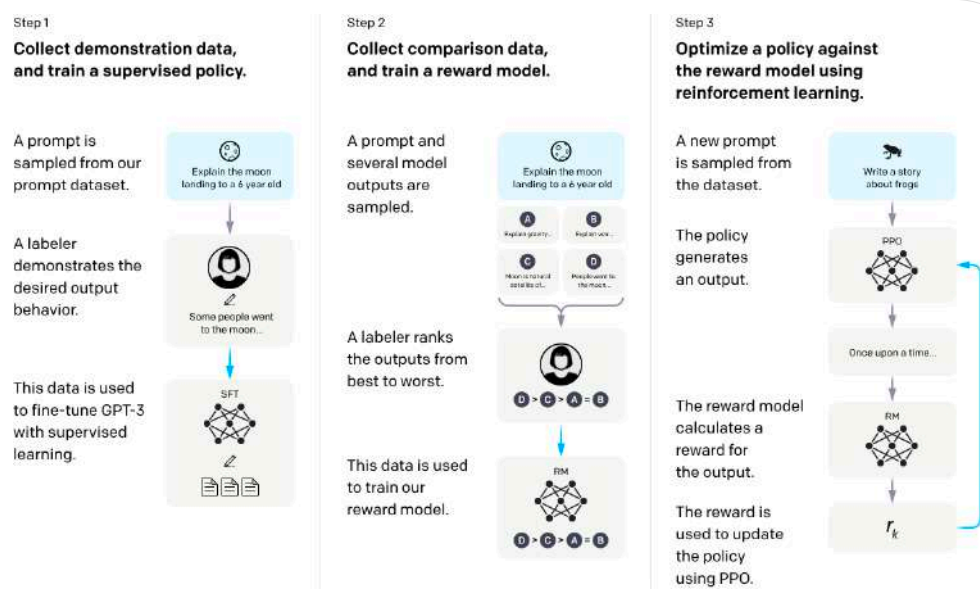
针对每个任务，单独生成 Instruction (hard token)，通过在若干个 Full-shot 任务上进行微调，然后在具体的任务上进行评估泛化能力 (zero shot)。其中，预训练模型参数是 Unfreeze 的

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)



Supervised FineTune

监督微调

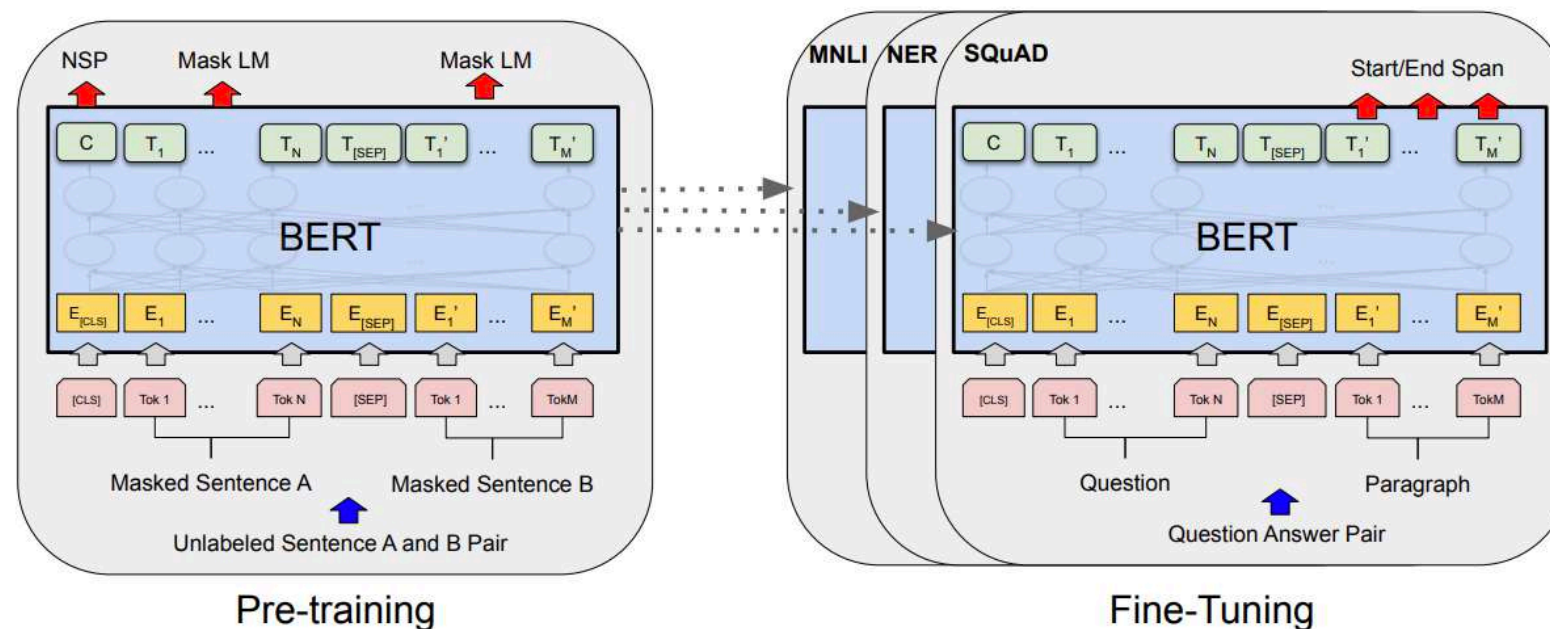
SFT

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

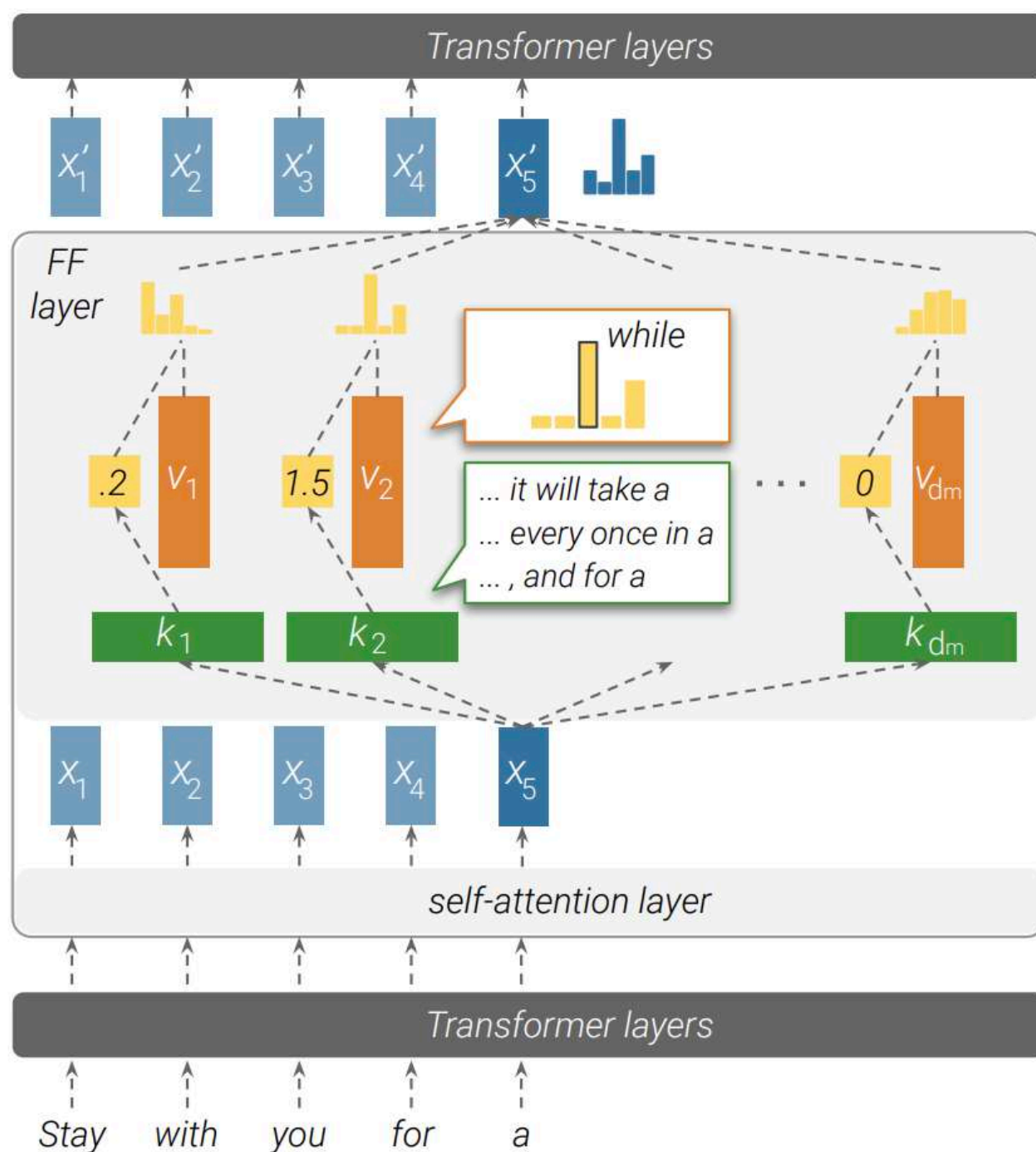
Language Models are Few-Shot Learners (NIPS2020)



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

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

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

Prompt-Tuning

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

LoRA

$W \in \mathbb{R}^{d \times d}$, $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times d}$, $h = Wx + ABx$

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

109

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（任务直觉）

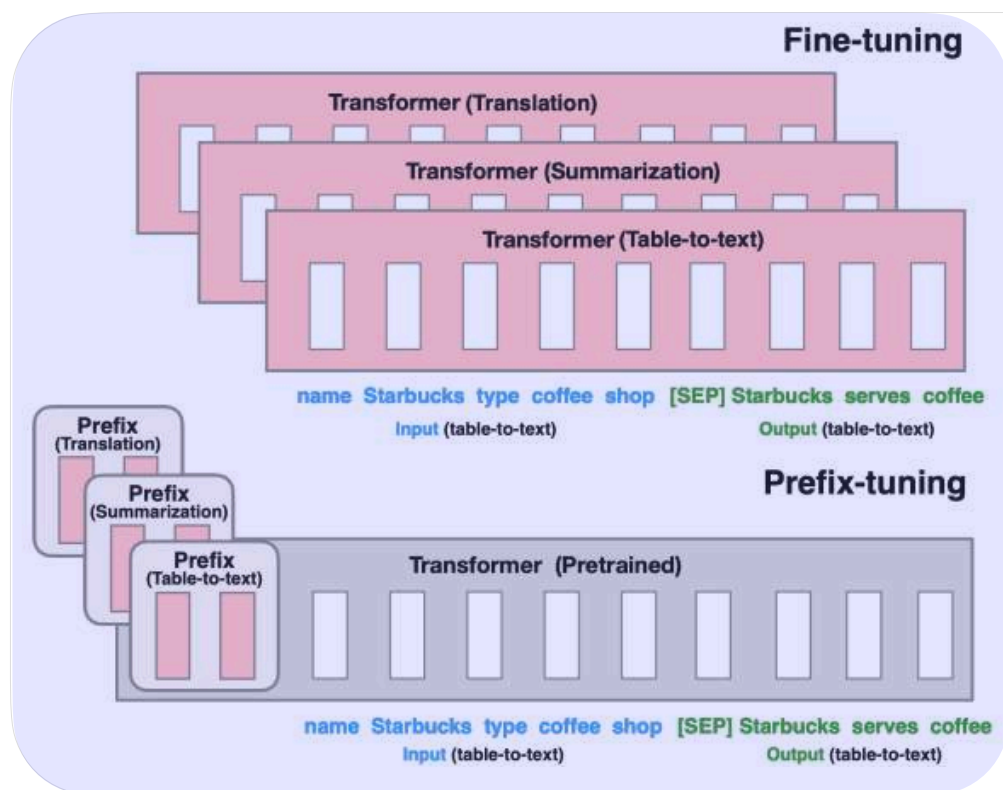
构建模板 (Template Construction)

标签词映射 (Label Word Verbalizer)

Language Models are Few-Shot Learners (NIPS2020)

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

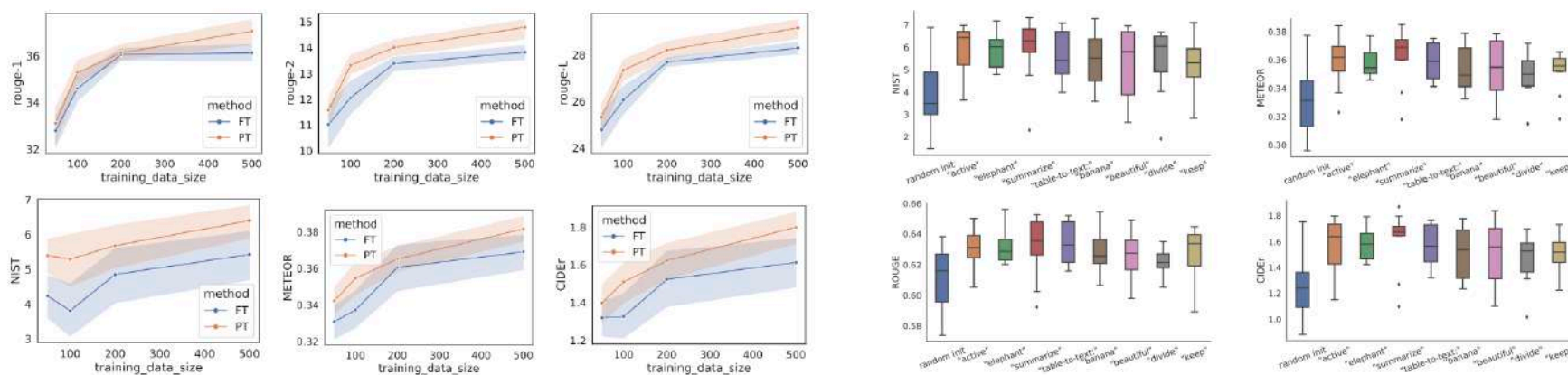
难训练，效果不如直接 Finetune

Paper

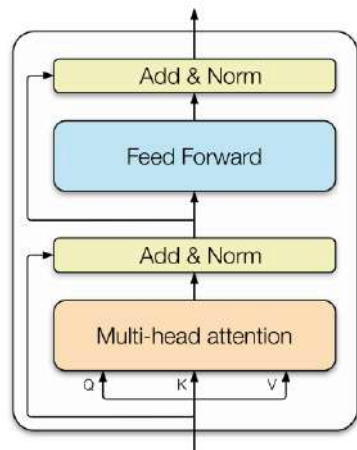
Prefix-Tuning: Optimizing Continuous Prompts for Generation

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

- 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开销



Adapter Fusion Tuning



Basic Transformer block

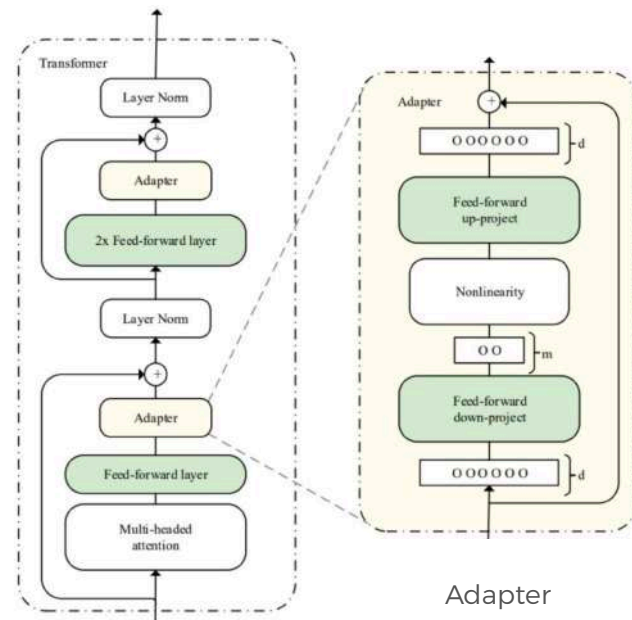
在预训练语言模型的每一层中间插入一个小型的可训练模块来调整模型的输出

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



Adapter

Paper

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

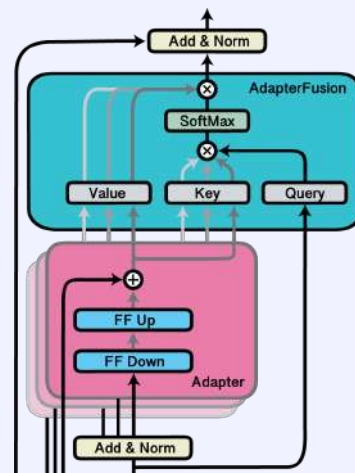
MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer <https://arxiv.org/pdf/2005.00052.pdf>

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

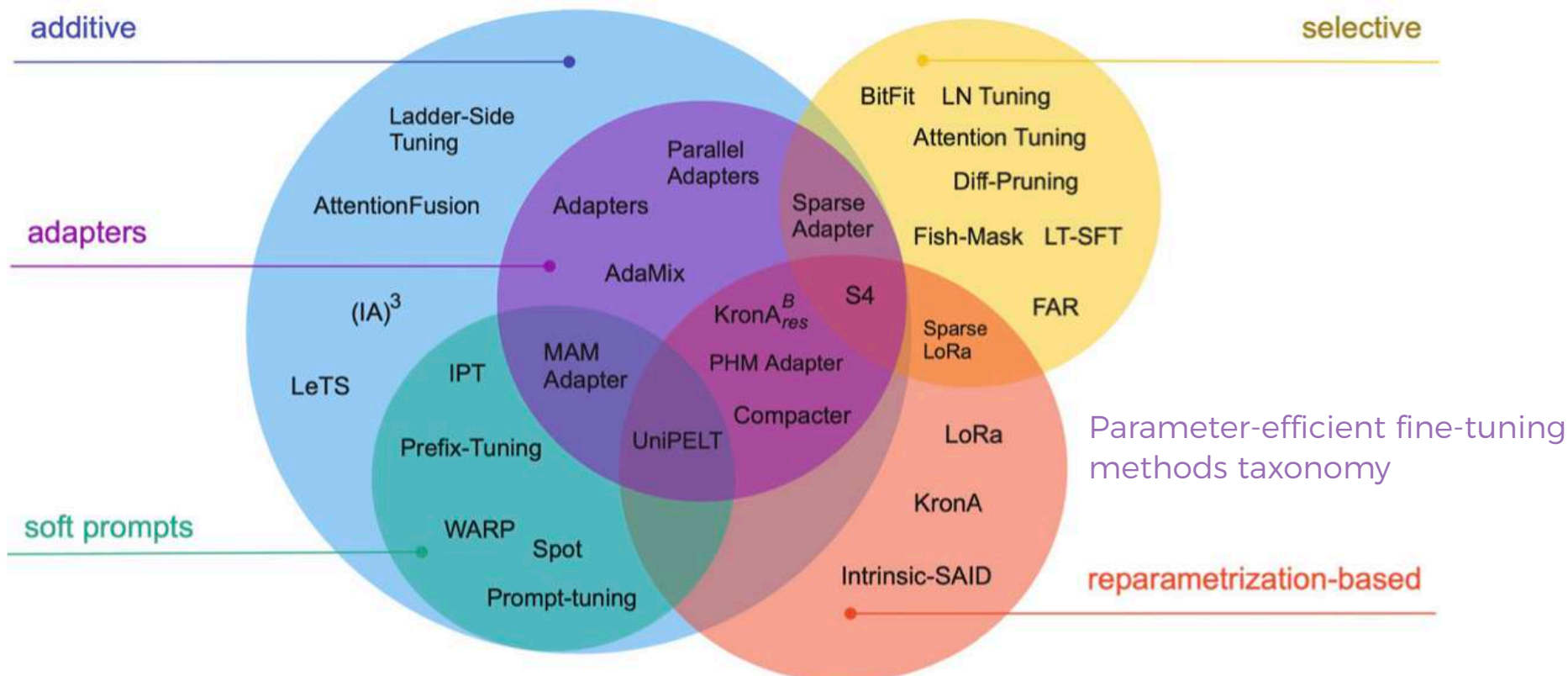
局限

增加了模型层数，引入了额外的推理延迟



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

Adapter



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

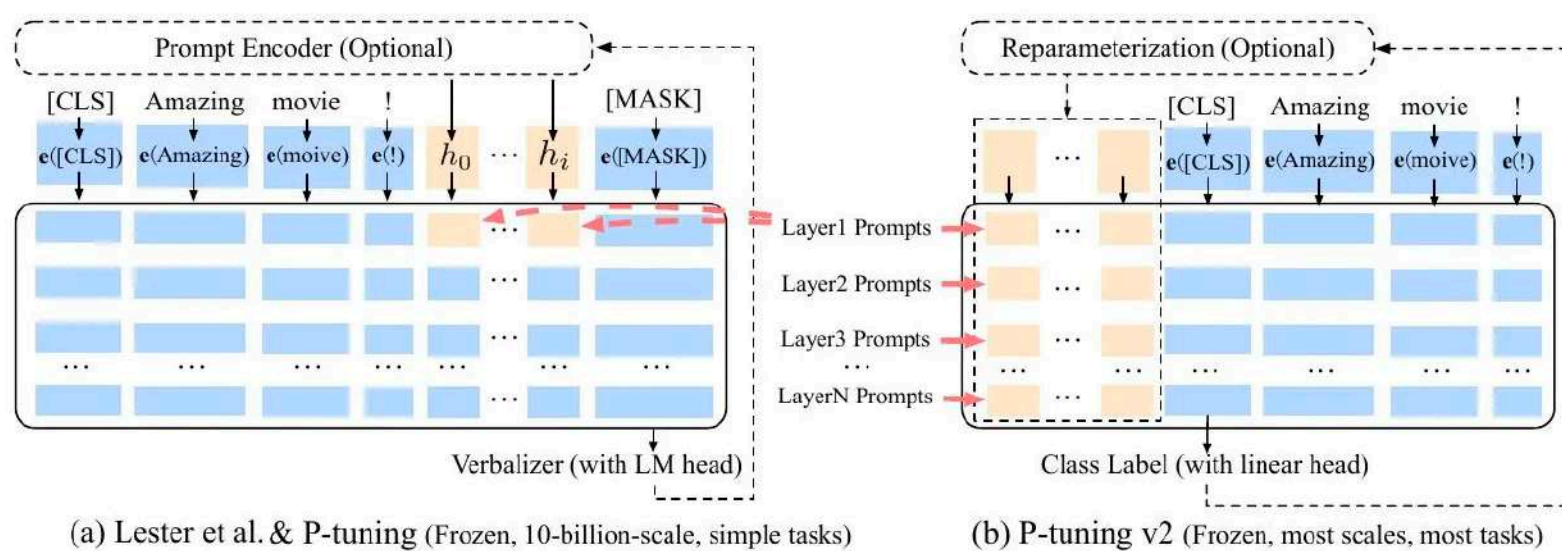


Figure 2: From Lester et al. (2021) & P-tuning to P-tuning v2. Orange tokens (include h_0, h_i) refer to prompt embeddings we add; blue tokens are embeddings stored or computed by frozen pre-trained language models. Compared to Lester et al. (2021), P-tuning v2 adds trainable continuous prompts to inputs 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

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†}

¹Tsinghua University, Beijing, China

²Beijing Academy of Artificial Intelligence (BAAI), Beijing, China

[2103.10385] GPT Understands, Too

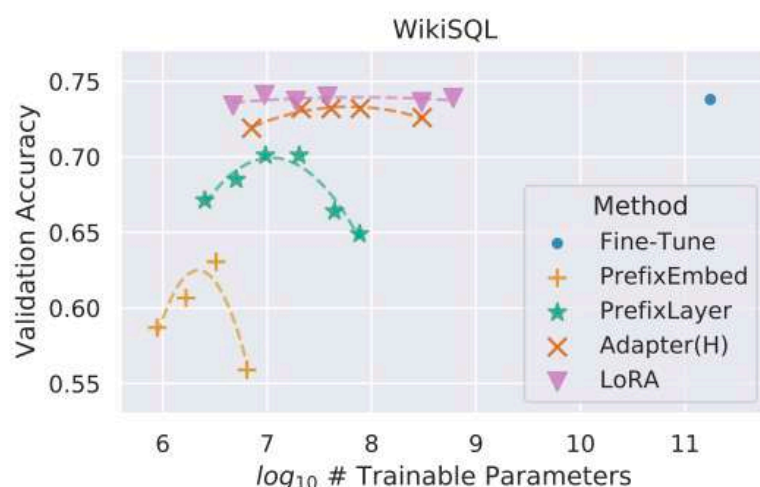
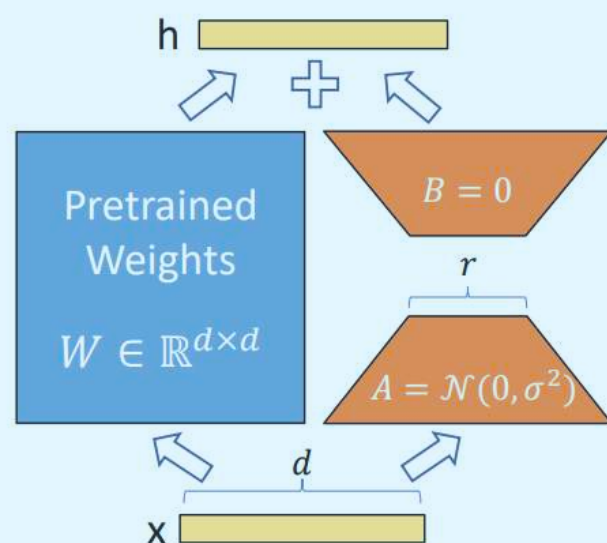
Low-Rank Adaptation of Large Language Models

LoRA

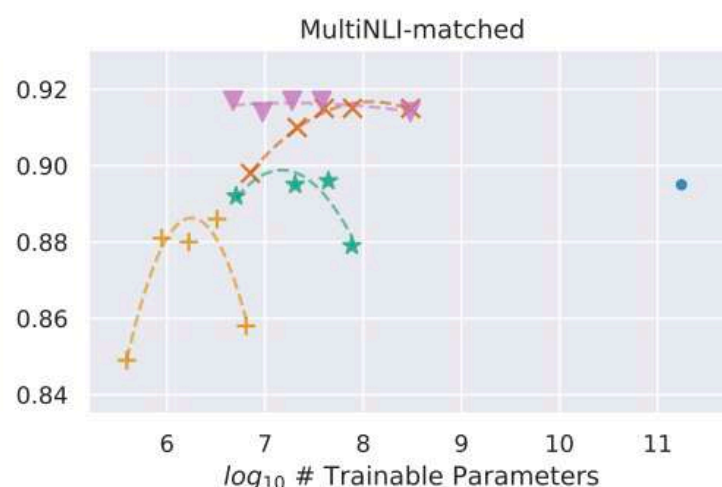
大语言模型的低阶自适应

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，但能获得较好的表现结果 [更多访问 xRunda.com](https://xRunda.com)

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，保证训练的开始时此旁路矩阵依然是 0 矩阵。



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

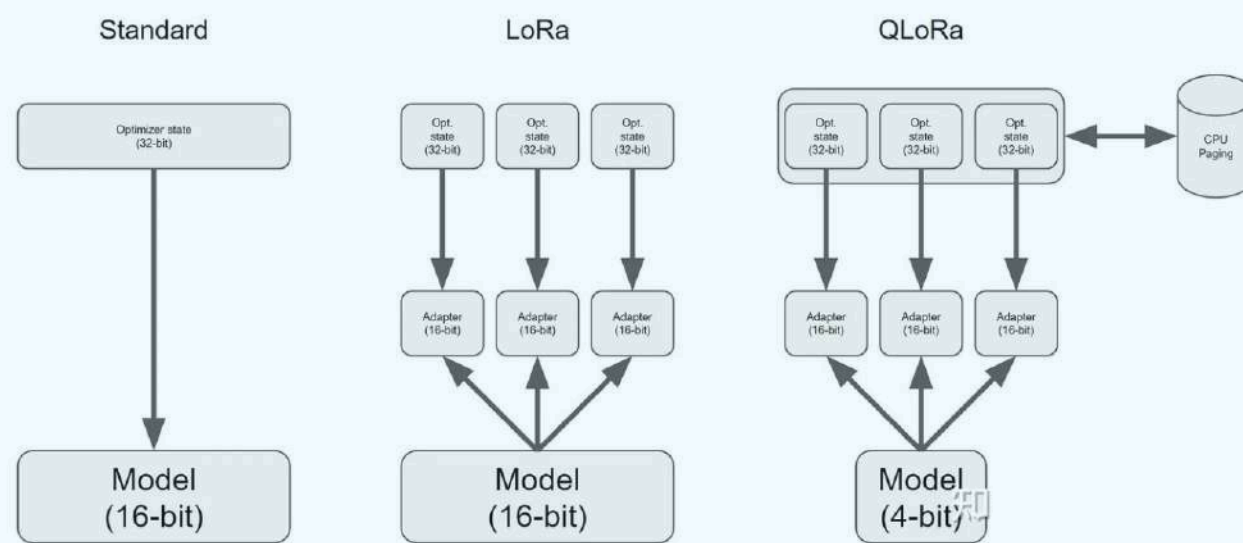


<https://github.com/microsoft/LoRA>

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

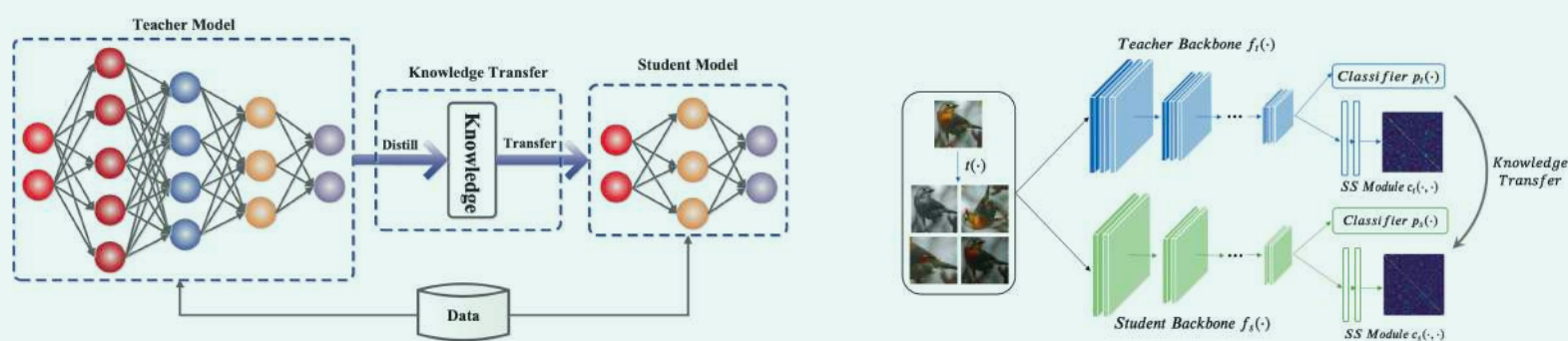
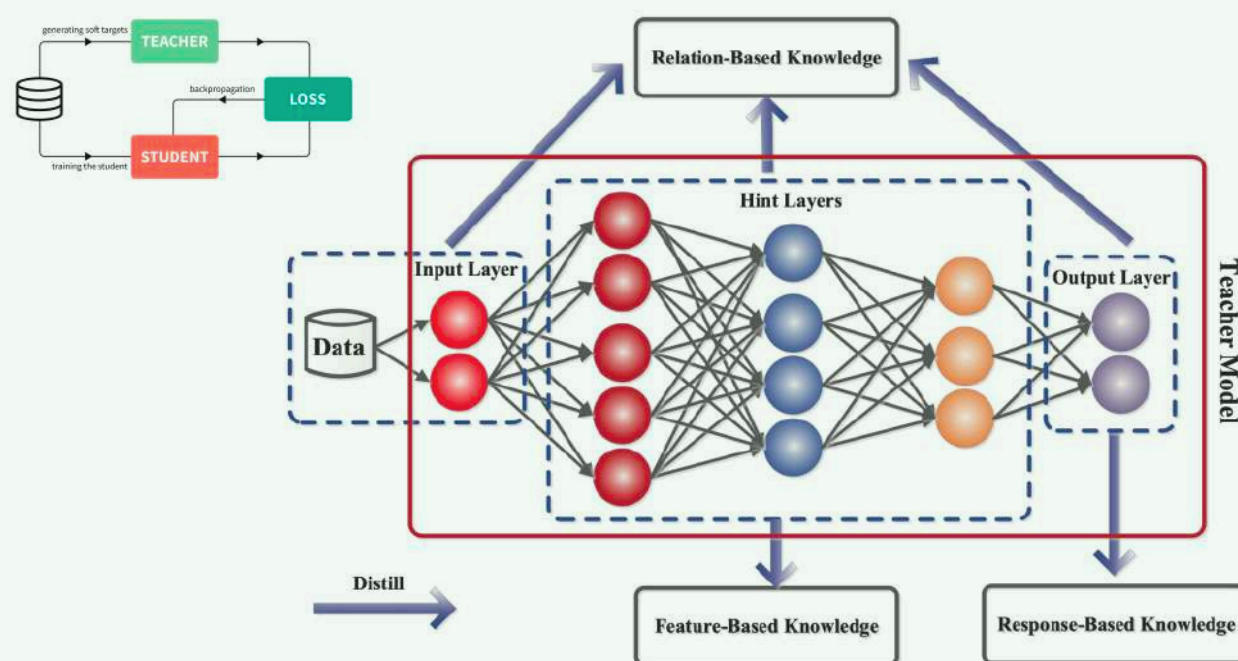
Knowledge Distillation

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

NEW
ACL2023

逐步蒸馏 *Distilling Step-by-step - Google*

步骤：CoT + 多任务学习框架

性能：在e-SNLI数据集上，仅使用12.5%数据就超过了标准微调的性能

成果：模型大小缩小了超过700倍

Reinforcement Learning

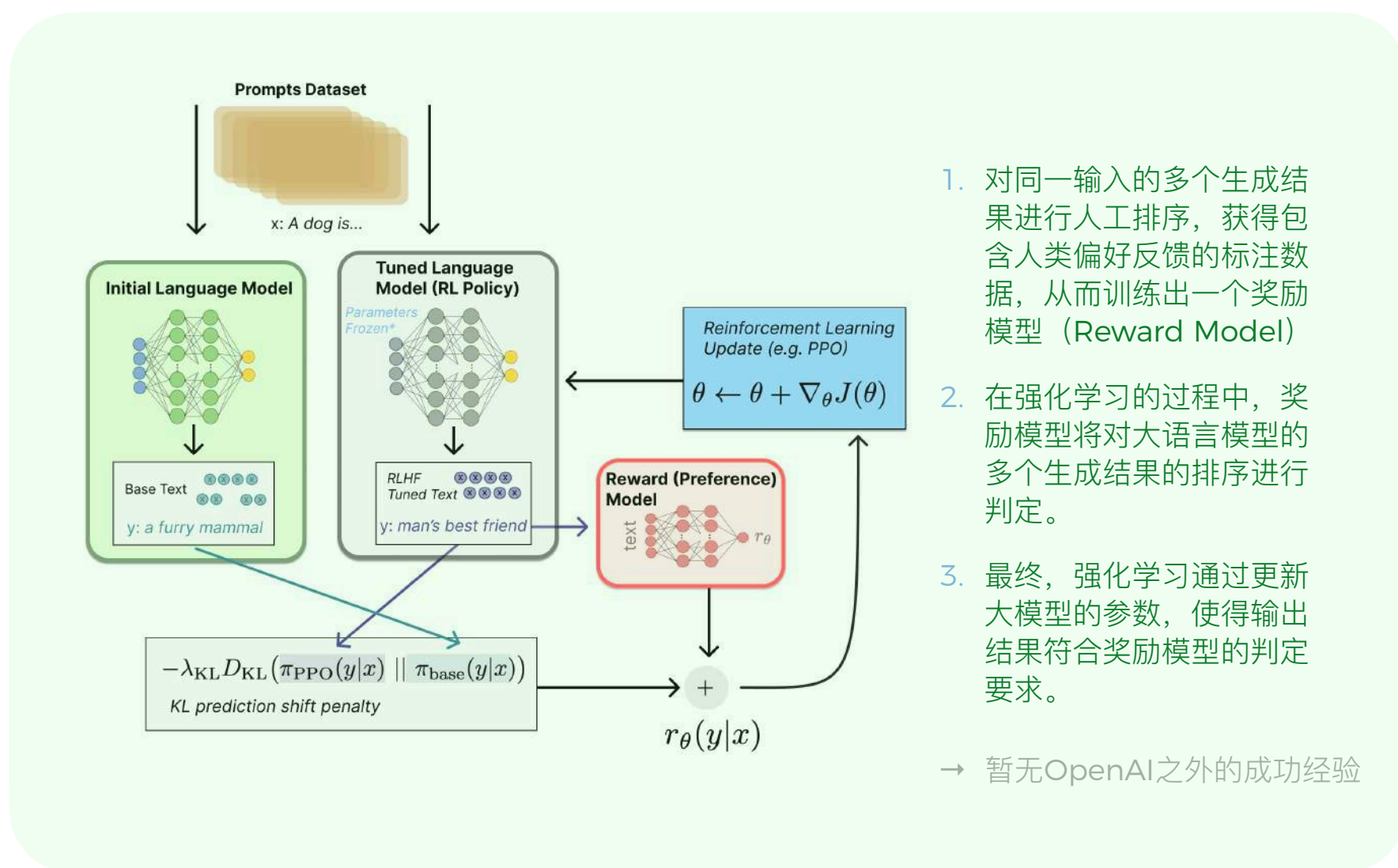
人在回路

强化学习方式微调

RLHF, Reinforcement Learning from Human Feedback

基于人类反馈的强化学习

RL



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

New

DPO 直接偏好优化

Zephyr-7B



企业智能化方案

行业企业数智化

P57 知识向量引擎智慧文旅

P58 平台资源支持

P59 工业数智化

P60 消费零售数智化

行业大模型

P61 法律大模型

P62 国际金融案例

P63 智慧医疗项目

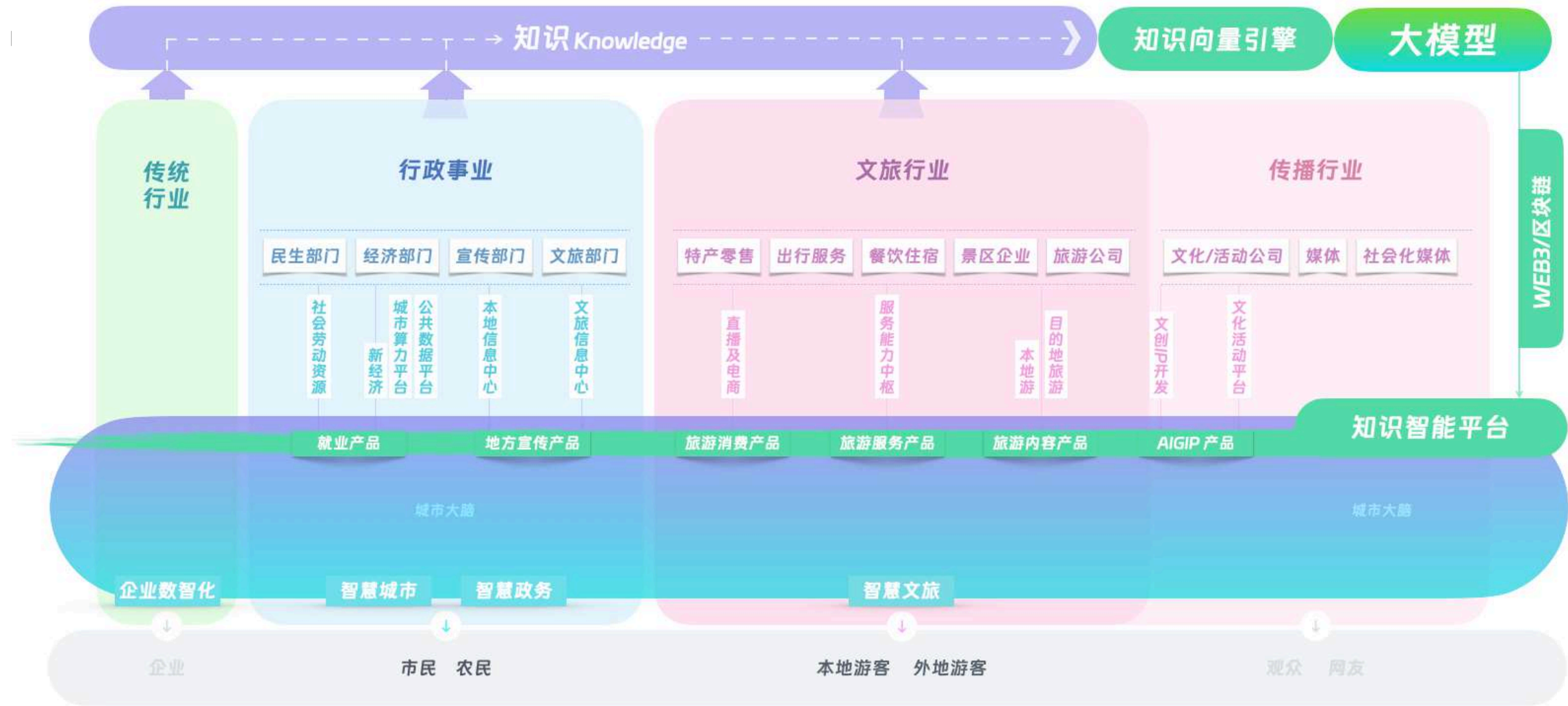
P64 教育平台智化

P65 汽车产品智化

P66 科学领域智化



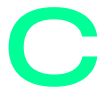
知识向量引擎智慧文旅



xRunda 文旅行业
数智化解决方案

xMaaS 智慧文旅平台架构





平台资源支持





工业数智化



安全生产智能应用平台



产品功能台



解决方案



企业 智能化

消费零售数智化

蒙牛数智化平台



未来可口可乐

法律大模型

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

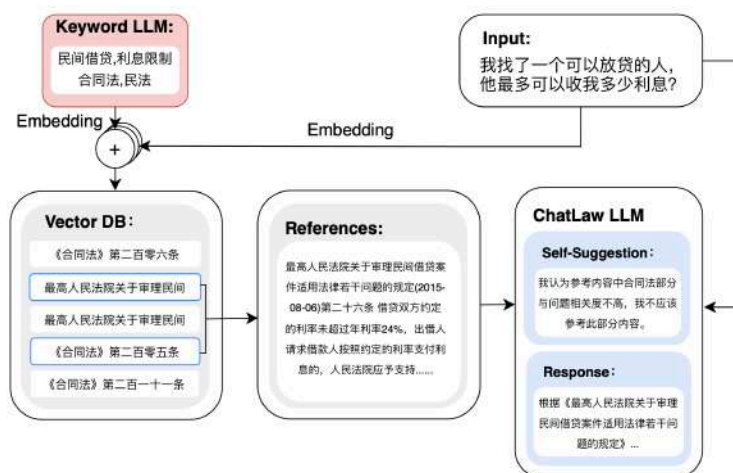


Figure 1: ChatLaw Framework

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>

国际金融案例



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>



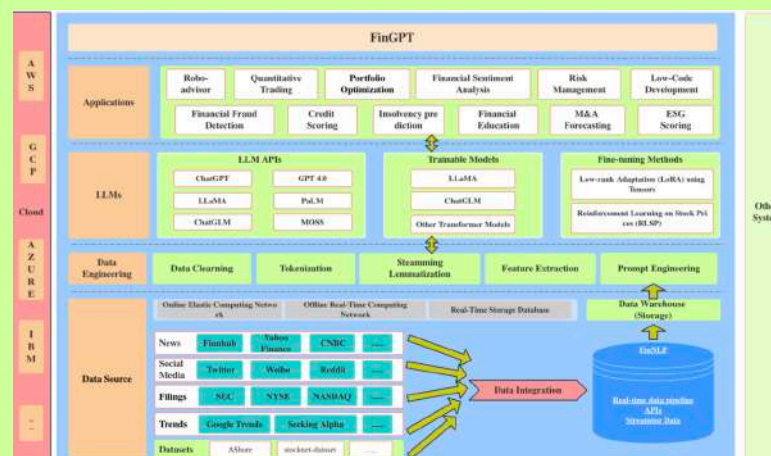
BondGPT

BroadRidge 推出 BondGPT，服务债券市场



度小满

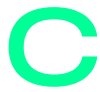
轩辕 (BLOOM)



FinGPT

FinGPT: Open-Source Financial Large Language Models

<https://arxiv.org/abs/2306.06031>



智慧医疗项目

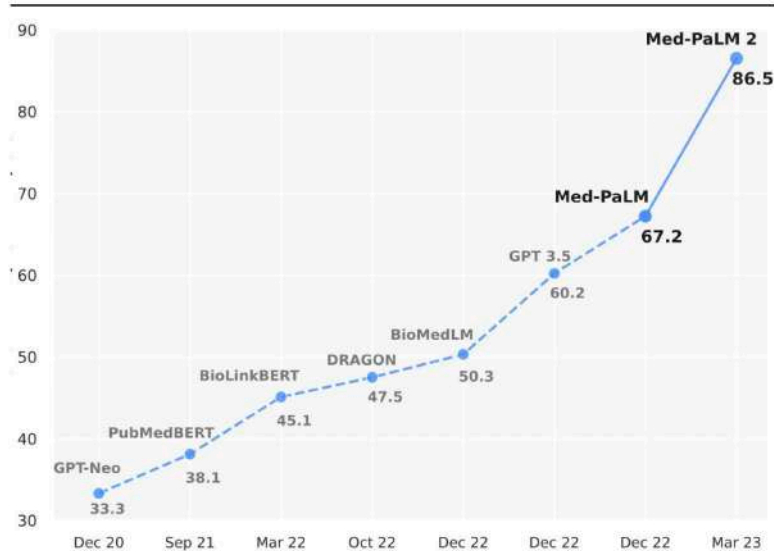


MedGPT

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

MedQA (USMLE式问题) 部分模型测试准确率对比



Med-PaLM 2

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



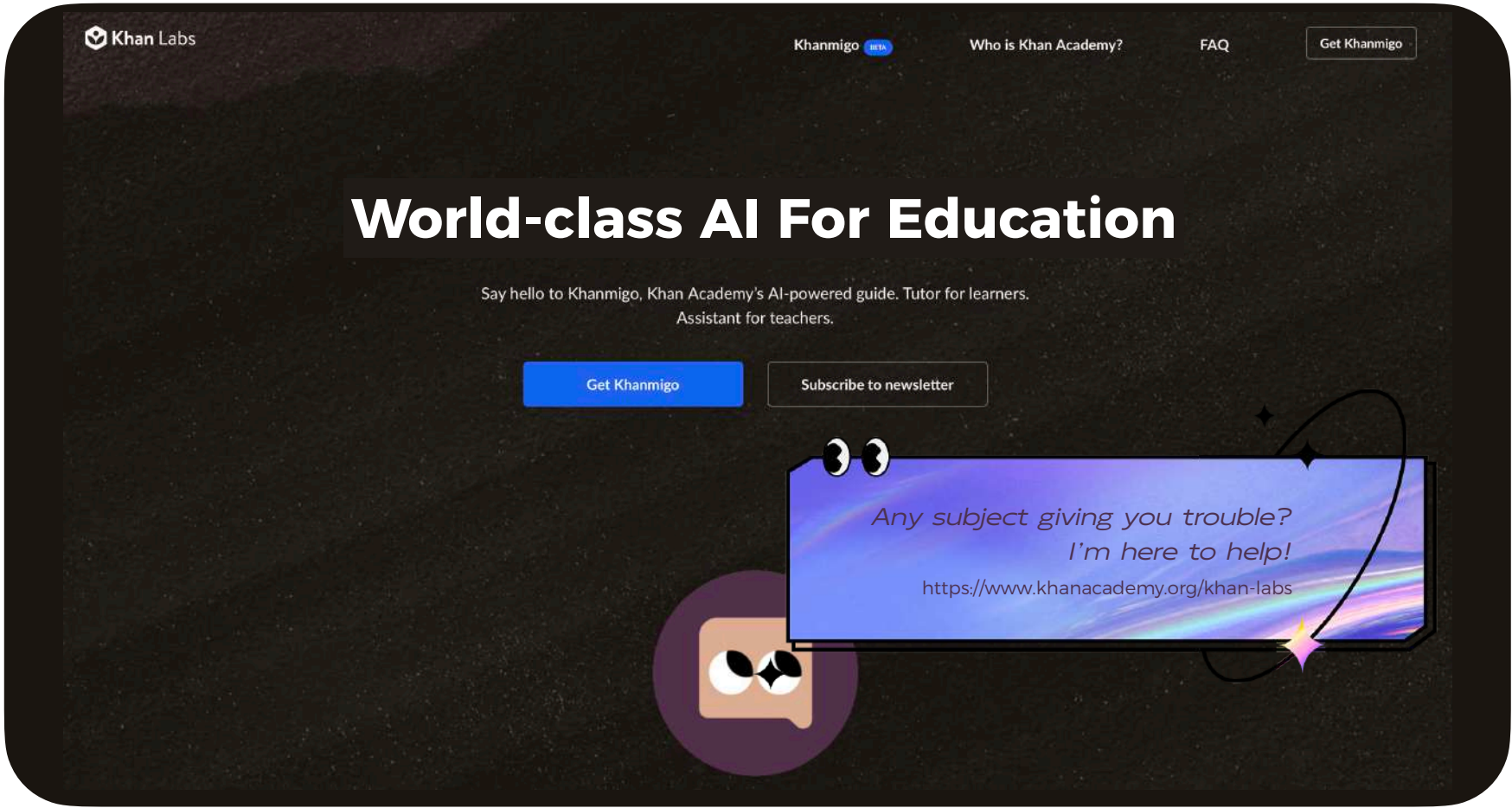
AI+医疗研究报告

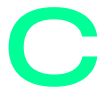
https://pdf.dfcfw.com/pdf/H301_AP202306281591795603_1.pdf



企业
智能化

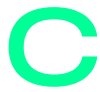
教育平台智能化





汽车产品智能化





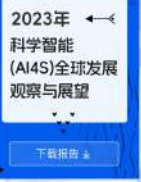
科学领域智能化

AI For Science



AI FOR SCIENTIST

绝佳科学效率应用市场，用AGI驱动你的科学研究。
在这里，你可以使用基于语音、图像和文本的大模型能力，定制属于你的个性化AI4S应用。

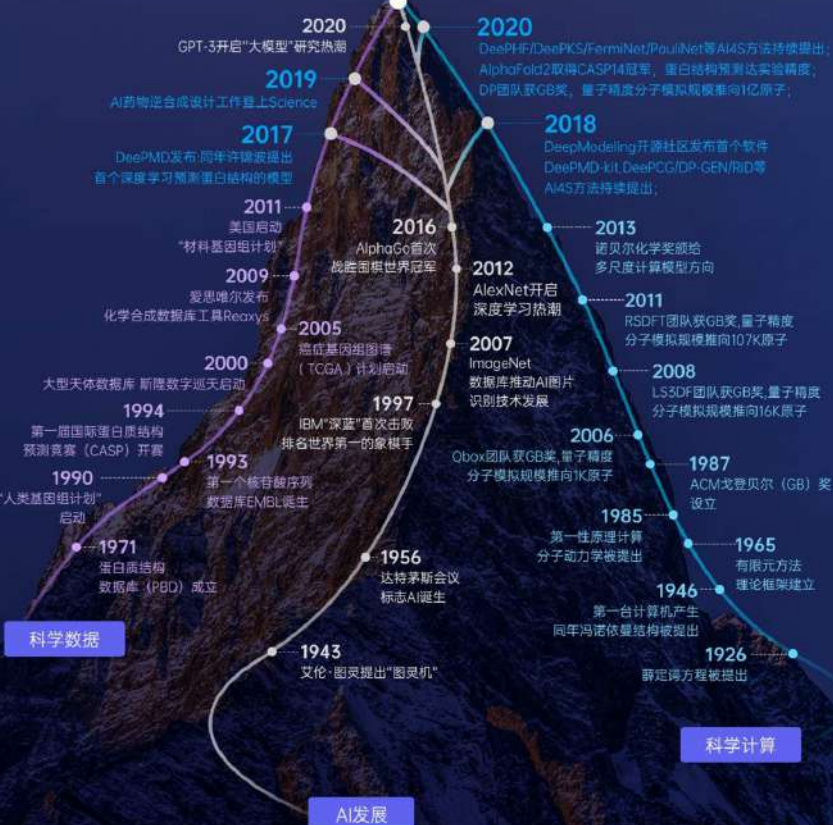


络绎科学



Figure credit: DP Technology

2022



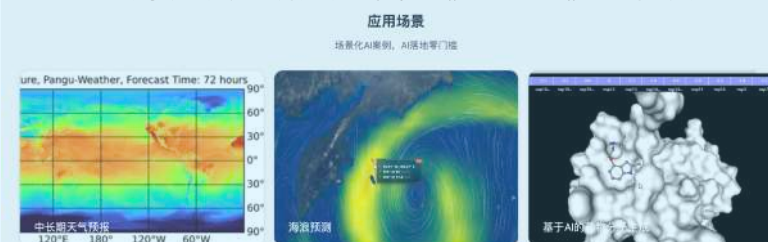
Google DeepMind

- 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



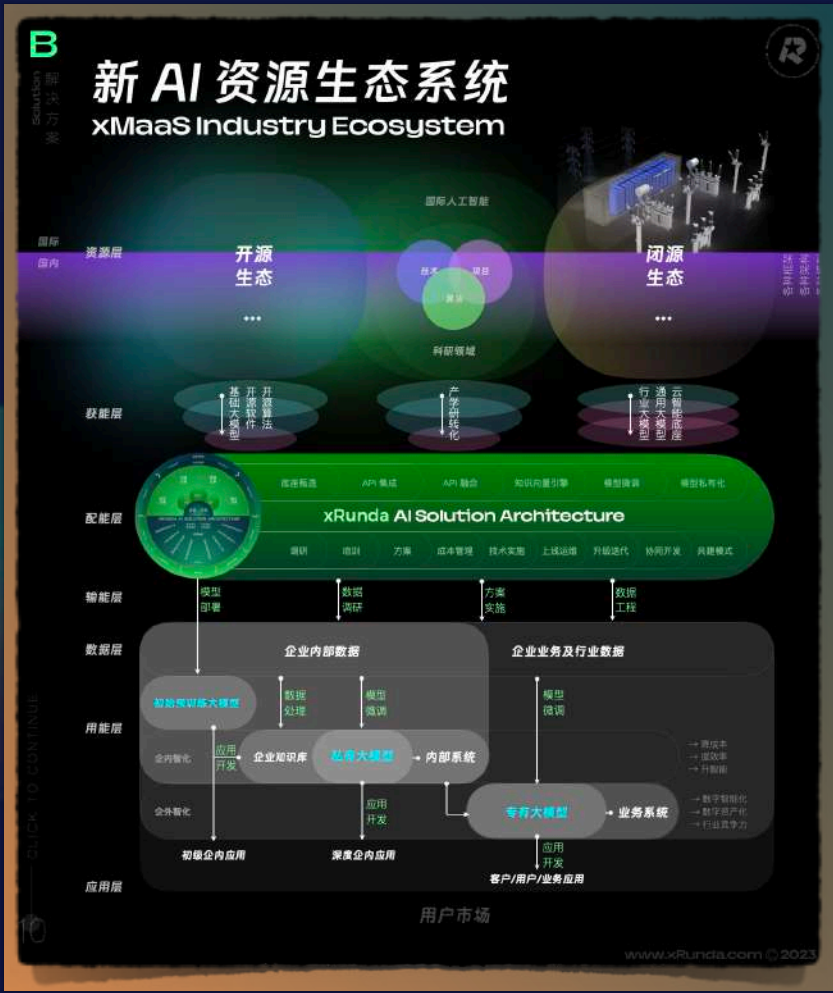
盘古药物分子大模型

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





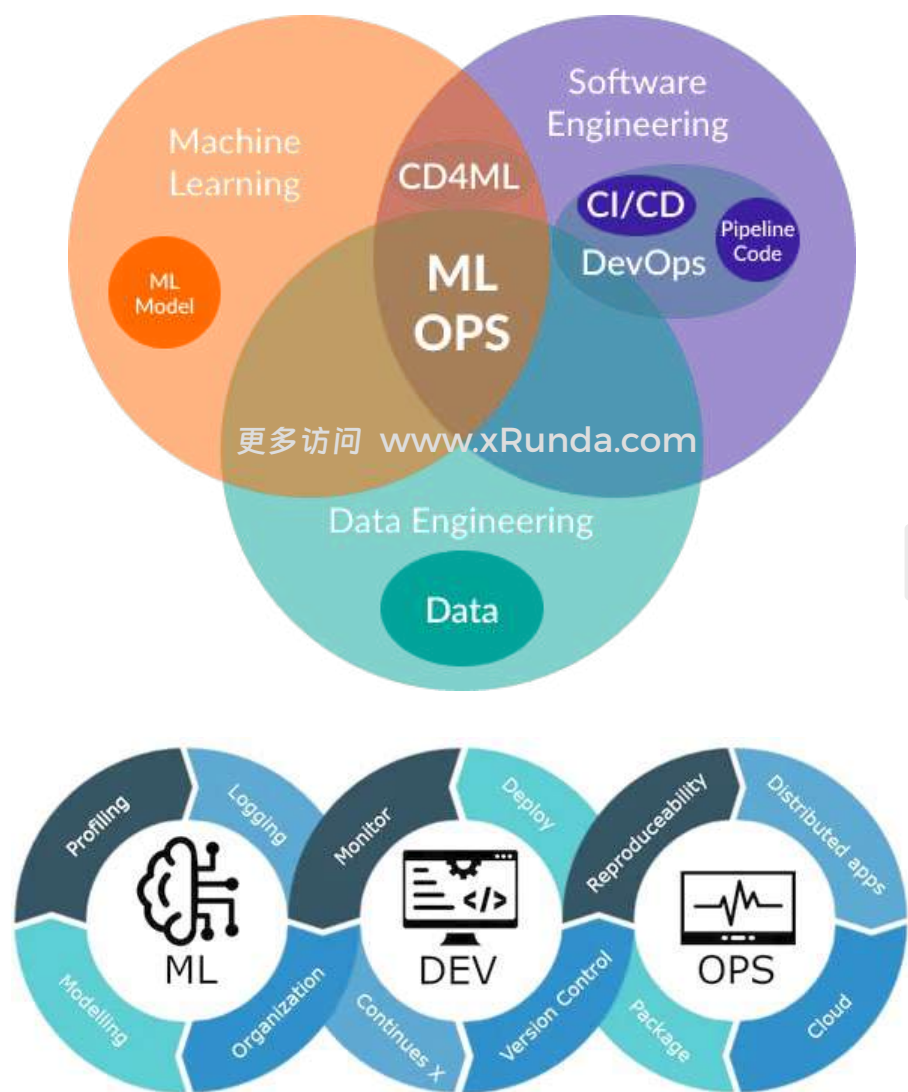
xMaaS 集成方案

P68 LMOPs 方案

P69 xMaaS 服务方案

P70 MaaS 平台资源

LMOps 方案

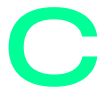
The logo for LMOps, with 'LM' in purple, 'Ops' in blue, and 'ps' in teal.

LMOps 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

数据	模型	训练
无监督数据	代码修改	分布式并行
大幅上升	几乎不需要	大幅上升
调优	调优	服务
提示工程	SFT/LoRA	插件扩展
新出现	新出现	新出现
服务	部署	评估
多模态生成	基础+增量	模型做裁判
大幅上升	新出现	新出现

LMOps Landscape





xMaaS 集成方案

xMaaS 服务方案

数据研究

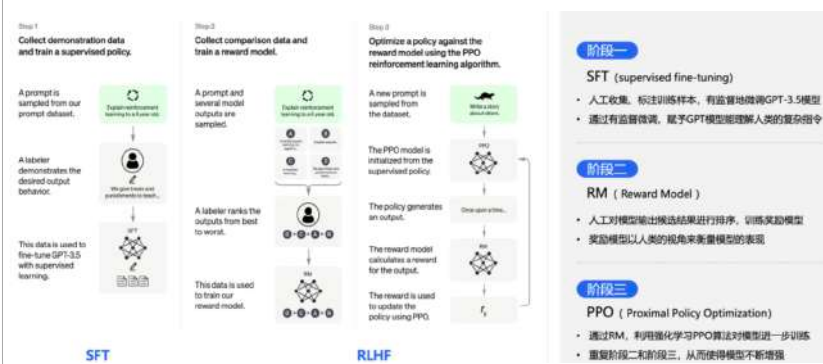
调研 计划 方案 向量化 存储



1

模型训练

3



模型评估

4



模型运维

6



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。

模型部署

5



更多访问 xRunda.com



MaaS 平台资源

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

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

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

More

更多访问 www.xRunda.com

D

产品案例 Product Case

P71

最新产品

- 即摘 xGeekSum ✨
- 即听 xGPTing 🎙️ Generative Podcast Transforming
- 即试 xGPTest 🧪 Generative Product Test
- 即答 xChatDA 🔍 Chat Distribute Agent
- 即调 xTune 🎛️
- 即画 xDraw 🖼️

P79

实项展示

- 行业服务
- 孵化项目
- 实验项目
- 案例链接
- <https://xrunmeta.feishu.cn/docx/DR0HdhS6xoqnSdxpcPlcHHc4nrh>



P69 即摘 xGeekSum ✨

P70 即听 xGPTing 🎧

P71 即试 xGPTTest 🔬

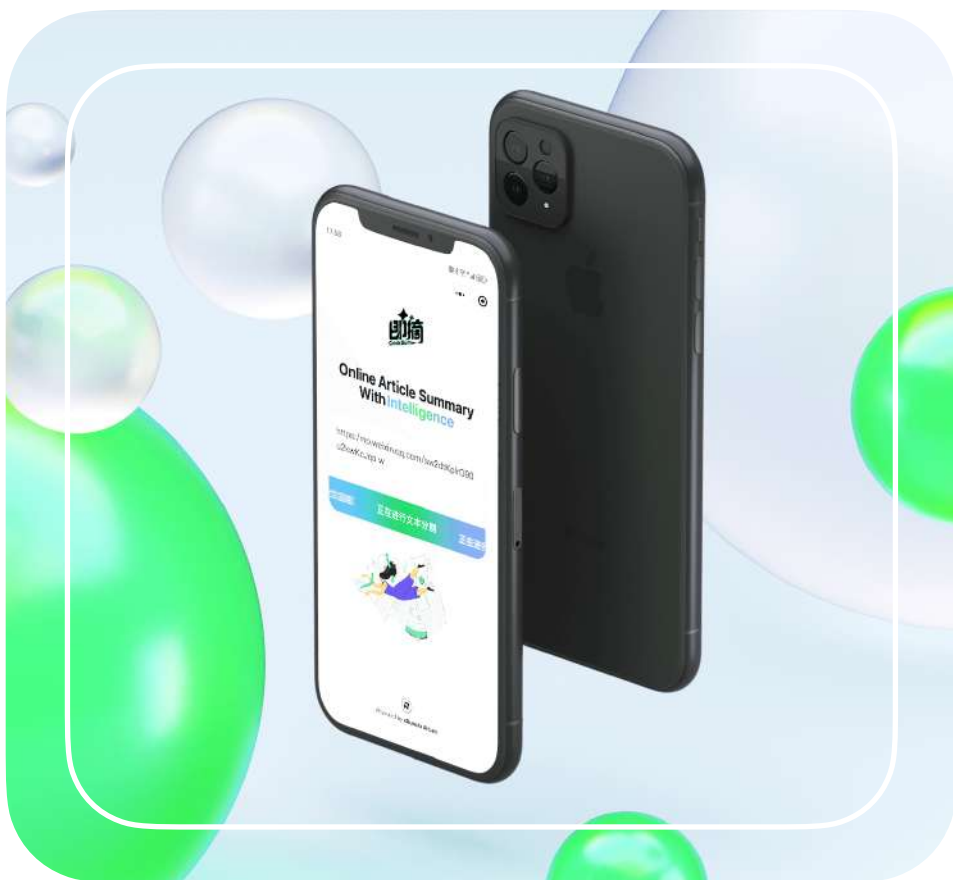
P72 即答 xChatDA 🔍

P73 即调 xTune 🎛️

P74 即画 xDraw 🖼️

最新产品

即摘 xGeekSum ✨



即摘 GeekSum



★
TO
C / B / G

高效阅读助手

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，为您打造高效的学习和收藏助手

即摘亮点

智能快摘

通过 AI 快速生成文章摘要

多源支持

已支持微信公众号、知乎、即刻、虎嗅、雪球、少数派等多家主流媒体

多端速读

小程序及 H5 等多端触达

操作便捷

复制粘贴链接，一键摘要阅读



展望

增值服务/未来展望/持续迭代

轻收藏夹

历史记录，随时回顾

音频生成

文章要读，便捷听取

AI 问答

追问探究，深度理解

阅读笔记

捕捉灵感，记录思考

智能分类

高效管理，梳理思路

读后分享

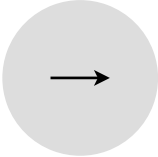
轻松分享，共同成长

最新产品

即听 xGPTing



即听 xGPTing



TO
C / B / G

开启智能化播客 workflow

Generative Podcast Transforming

arXiv:2305.16291v1 [cs.AI] 25 May 2023

VOYAGER: An Open-Ended Embodied Agent with Large Language Models

Guanzhi Wang^{1,2}, Yuxi Xie³, Yunfan Jiang^{1*}, Ajay Mandlekar^{1*},
Chaowei Xie^{1*}, Yuke Zhu¹, Liwei Fan¹, Anima Anandkumar^{2,3†}
¹NVIDIA, ²Caltech, ³UT Austin, ^{*}Stanford, [†]ASU
^{*}Equal contribution, [†]Equal advising
<https://voyager.nvidia.com>

Abstract

We introduce VOYAGER, the first LLM-powered embodied lifelong learning agent in Minecraft that continuously explores the world, acquires diverse skills, and makes novel discoveries without human intervention. VOYAGER consists of three key components: 1) an automatic curriculum that maximizes exploration, 2) an ever-growing skill library of executable code for storing and retrieving complex behaviors, and 3) a new iterative prompting mechanism that incorporates environment feedback, execution errors, and self-verification for program improvement. VOYAGER interacts with GPT-4 via blackbox queries, which bypasses the need for model parameter fine-tuning. The skills developed by VOYAGER are temporally extended, interpretable, and compositional, which compounds the agent's abilities rapidly and alleviates catastrophic forgetting. Empirically, VOYAGER shows strong in-context lifelong learning capability and exhibits exceptional proficiency in playing Minecraft. It obtains 3.3x more unique items, travels 2.4x longer distances, and unlocks key tech tree milestones up to 15.3x faster than prior SOTA. VOYAGER is able to utilize the learned skill library in a new Minecraft world to solve novel tasks from scratch, while other techniques struggle to generalize.

Figure 1: VOYAGER discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baseline. X-axis denotes the number of prompting iterations.

人类直接做Minecraft的探索发展，速度是否能快过Voyager这个方法？

“人类直接进行 Minecraft 探索发展的速度可能远远慢于 Voyager 这种方法。Voyager 利用强化学习算法和自动课程规划，以及不断发展的技能库和迭代提示机制，可以在 Minecraft 中以更快的速度发现新的物品和技能，并解决新颖的任务。相比之下，人类玩家由于受限于个人经验和认知能力，很难在如此庞大的游戏环境中以较高效率发现并学习新技能。”

什么叫通才智能体

“通才智能体是指一种具有多种技能和能力的智能体，它可以在不同的环境中自主学习和探索，并利用所学的技能来解决各种问题。通才智能体具有适应性强、学习能力强和探索能力强等特点，可以应用于许多领域，如智能客服、智能助手、自动驾驶等。”

涌现在这个项目里有什么作用？

“In this project, emergence plays a role in generating consistent action plans or executable policies for large language model (LLM) based agents. These agents use the world knowledge encapsulated

生成播客 查看摘要 提问模版 new 分享

输入对当前文档的提问

发送



最新产品

即试 xGPTTest



即试 xGPTTest

→

TO
C / B / G

Prompt 智慧市集创作体验

用即试 xGPTTest 创建和管理您的 Prompt ,
积累沉淀每个不经意的灵感发现

Generative Product Test

即试亮点

试记全能	轻松尝试、比较、编辑、运行和存储您的提示词
数据安全	简单组织您的提示组，独立管理每个项目
工作笔记	记录 Prompt 创作过程中的收获与反思，亮点与不足
版本回溯	每次运行提示词都有版本记录，方便随时回退到以前版本
动态数据	支持插入多个变量到您的提示词中
多重模型	同步支持国际、国内主流模型，包括 百度文心一言、阿里通义千问、智谱 GLM、科大讯飞星火、Minimax、Claude、Azure OpenAI 等



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

即答亮点

AI 多元交互

提供与多种 AI 模型的交互，包括聊天、问答、生成等，满足不同需求

大模型接入托管

无论个人或组织，均可托管 AI 大模型接入，享受优质的性能和稳定性

知识管理

基于您上传的数据集检索历史及回溯内容等，打造个性化知识库

数据保护

采用先进的安全技术，确保您的数据隐私得到全面保护

算力调度与优化

独特的集成方案，使用户在各种应用场景下都能获得最高性价比的算力成本消耗

智能问答聚合

提供高效的内容交互，聚合多个来源内容，综述提炼、去粕存精、去伪存真

一问多答对比

为每个提问返回多个 LLM 回答，方便比较参考，并提供提示，增强决策支持

多端支持

支持多端接入，使用户和应用能够随时访问我们的服务

投资与市场前景



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

前沿大模型微调平台，将高效参数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

适用人群

AI 开发者 企业与机构

主要功能

多模型支持

支持多种预训练大模型

一站式服务

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

交互式平台

支持多语言和领域的微调选项，多样的输出格式和参数设置，让微调更直观和便捷

批量与比较

基于相同数据集和微调目标，同步微调多个底座模型，横向比较效果，直观查看成本与性能，进一步提供调整建议，提升微调成功率

即调亮点

提升微调效率和成功率

即调的核心竞争力是在同一平台工作流上对多个底座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微调经验和优化策略

灵活易用

无论初学者还是专家，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 的深入发展，大模型微调即将成为行业数智化竞争关键技术。作为市场上首批集成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

将人工智能技术与艺术创造力完美融合通过算法加持，您的文字将被转化为高质量的图像作品
无论简单关键词或一段精彩的场景描述，都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

无论简单关键词或一段精彩的场景描述，都能在瞬息之间为您呈现出独一无二的视觉艺术，并支持您拥有版权

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

艺术参考

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

支持批量生成多图，提供多种尺寸和高分辨率的选择

快速调整

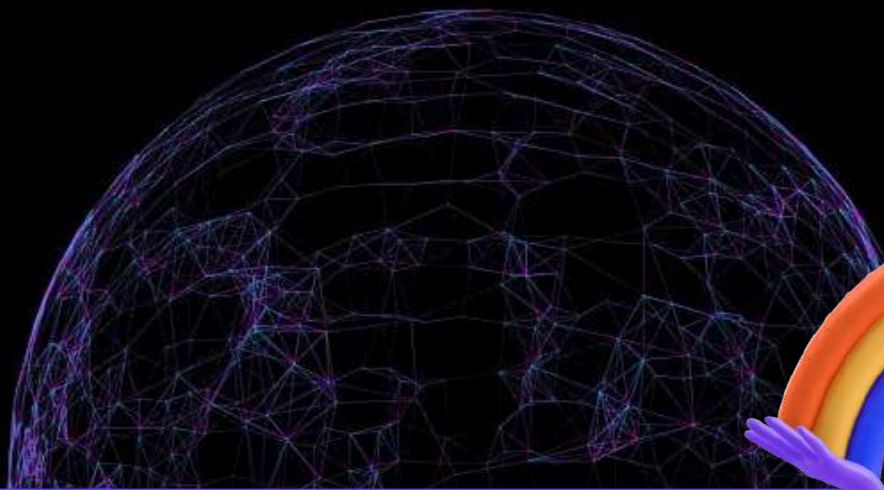
提供的灵活工具使您可以快速尝试和调整图片，无论专业设计师还是新手用户都能轻松获得满意的作品

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的盛宴，用即画描绘心中的图景



实项展示

Project Presentation

P76

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

P77

孵化项目

- 企内刊
- 喜AI图
- 即摘 GeekSum
- 即听 xGPTing
- 即试 xTry

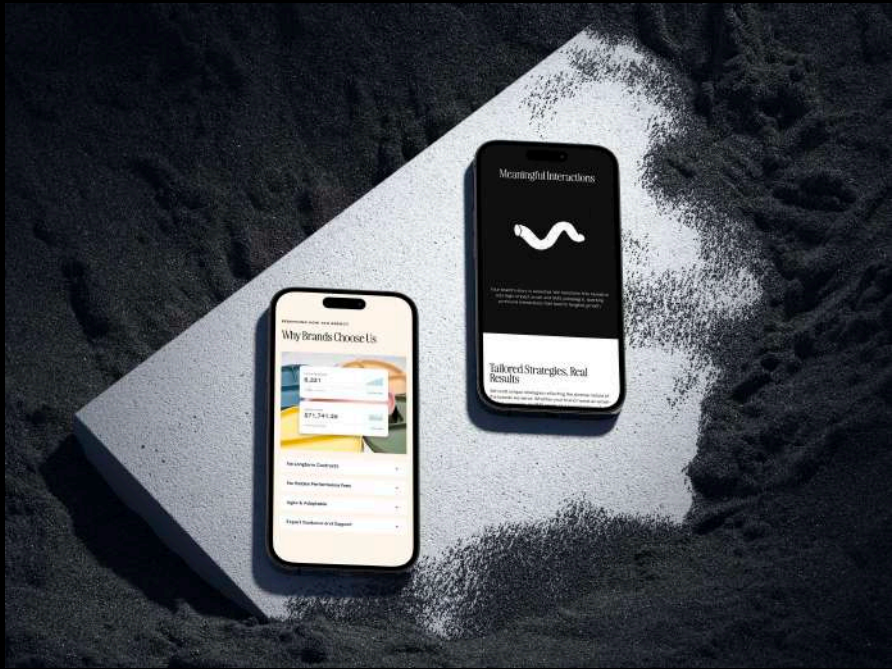
P78

实验项目

- 车型口碑 ChatBot
- AIGC多模态少儿绘本生产线
- 络绎科学 AI4S
- LoRA微调画家风格模型
- 个人声音定制模型训练程序

项目展示

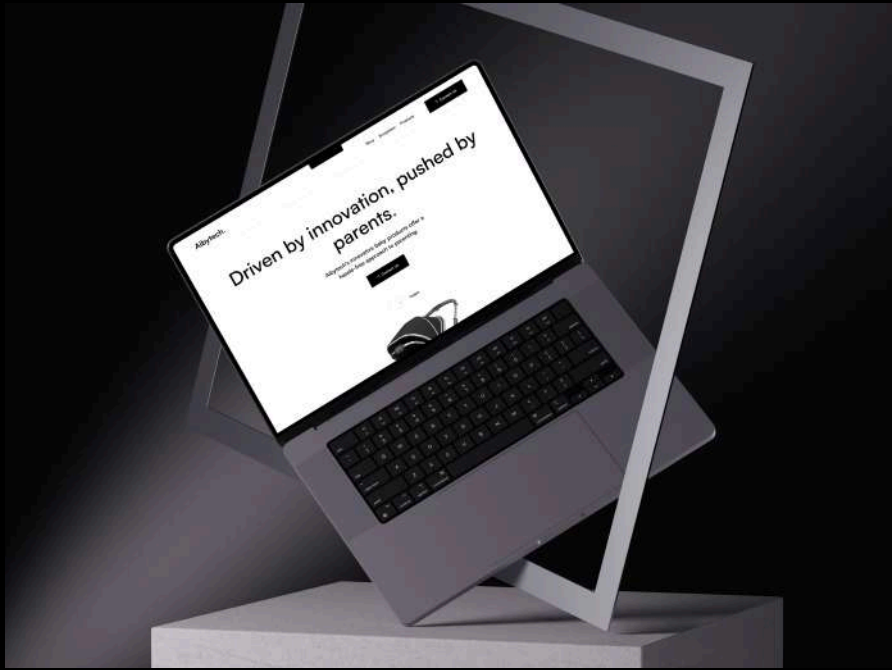
行业服务



保险公司海报协作平台



快手2022AIGC



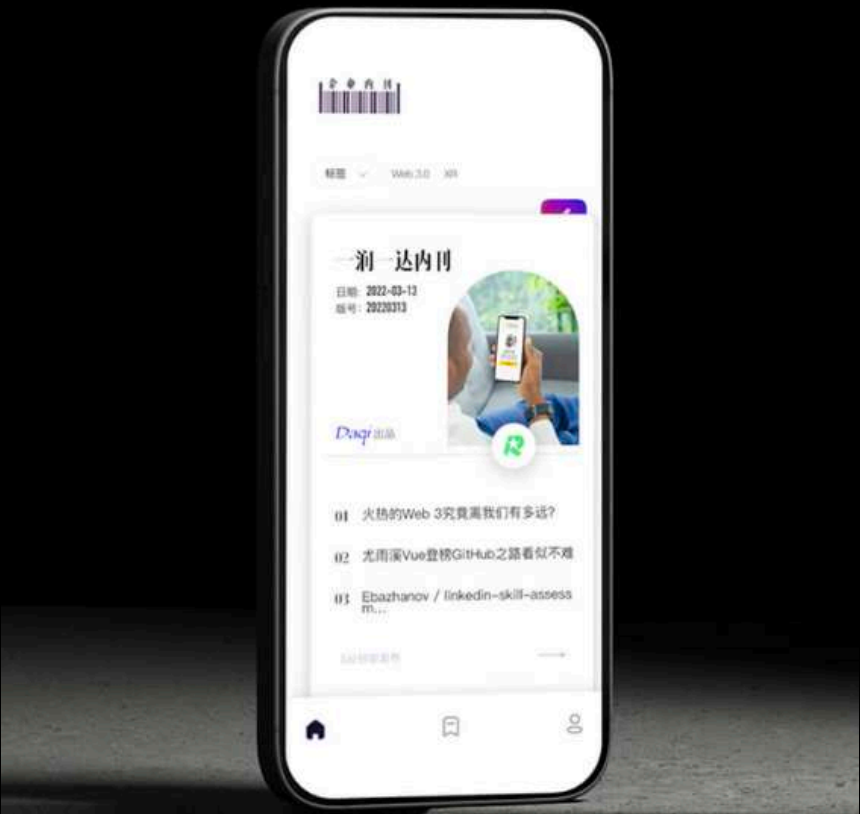
络绎科学 AI4S



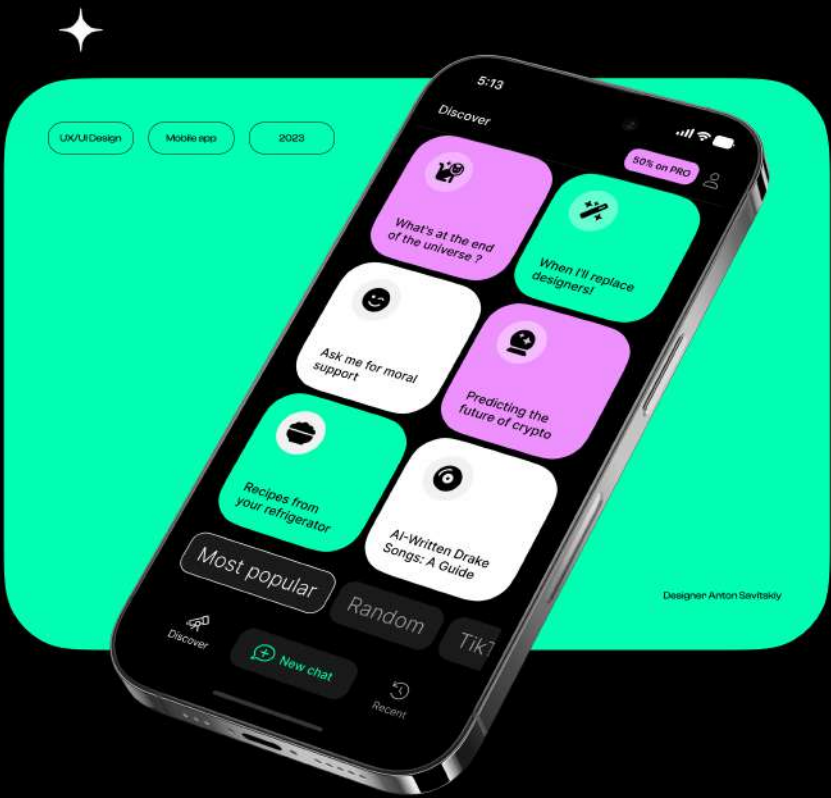
汽车之家

项目展示

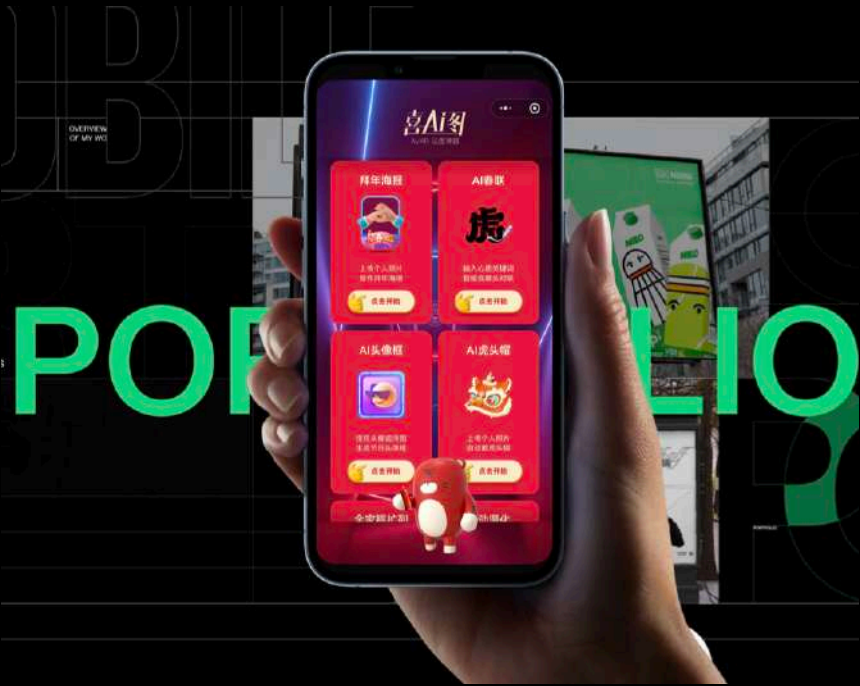
孵化项目



企内刊



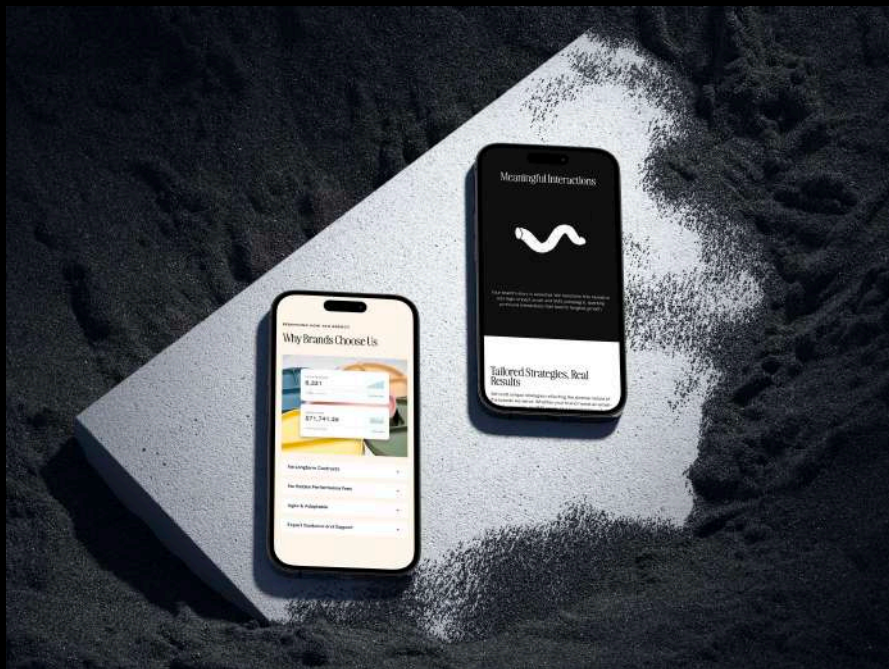
AIGA Town



喜AI图

项目
展示

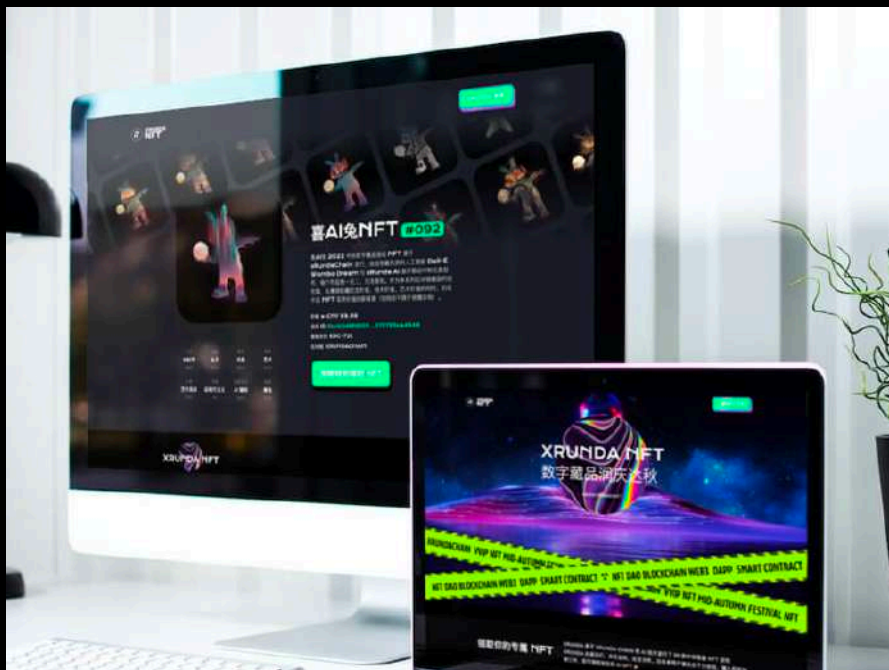
实验项目



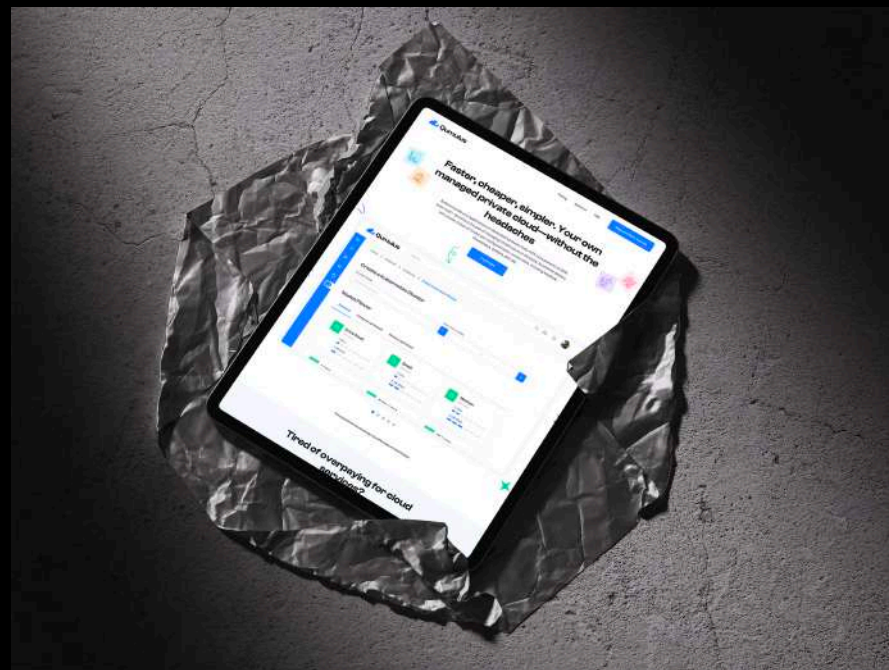
车型口碑 ChatBot



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

E

前沿研究

Frontier Research

P84

Agent

- Agent 概述
- 规划 Planning
- 记忆 Memory
- 使用工具 Tool Use
- 探索项目
- AgentBench

- P91 · Vector Embedding
- P93 · Knowledge Graph
- P95 · Multimodal 多模态
- P96 · 图像生成技术
- P97 · 音频生成技术
- P100 · 视频生成技术
- P101 · 数字人生成技术
- P102 · 3D 生成技术
- P104 · MoE
- P105 · 端侧模型

- P106 · 跨平台应用
- P107 · 具身智能
- P108 · 安全性
- P109 · 工程化
- P110 · 算法问题
- P111 · CoE
- P112 · 数据要素化
- P113 · 深度学习融合路线
- P114 · 新模型
- P115 · AI+WEB3融合路线

Agent



P81 Agent 概述

P82 Agent 规划 Planning

P83 Agent 记忆 Memory

P84 Agent 使用工具 Tool Use

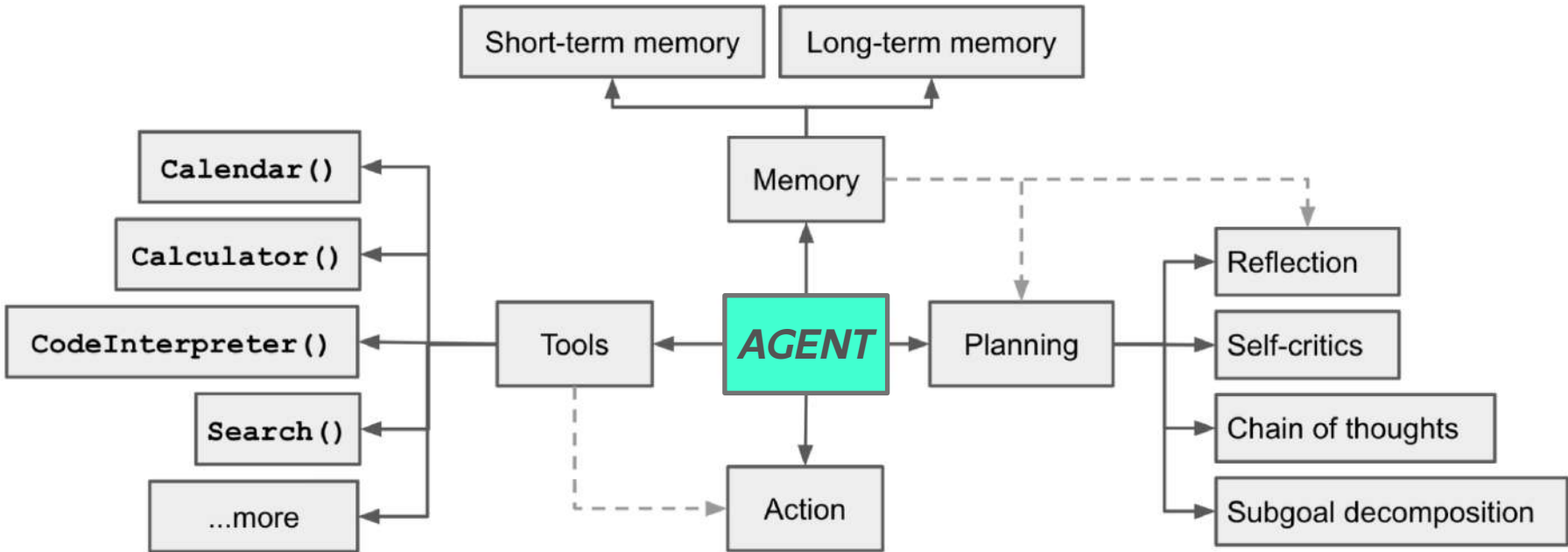
P85 Agent 探索项目

P86 Agent Bench 基准测试

Agent Agent 概述

Overview of A LLM-POWERED Autonomous AGENT System

LLM 支持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

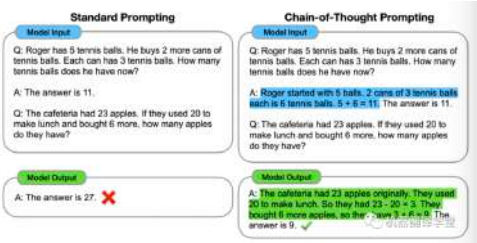
调用外部 API :
获取模型权重中缺失的额外信息（通常在预训练后很难更改）

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

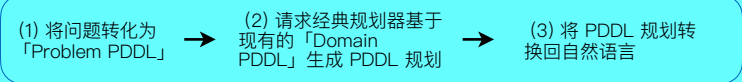
Tree of Thoughts: Deliberate Problem Solving with Large Language Models

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

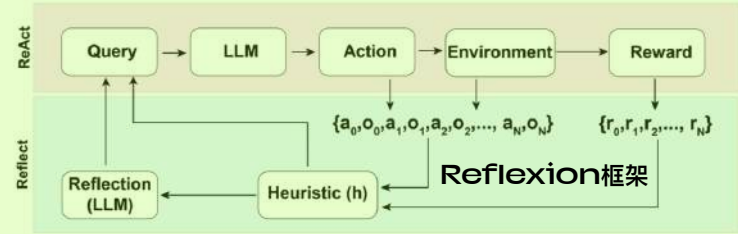


- 2、规划步骤被外包给了外部工具，并假设特定领域的 PDDL 和合适的规划器可用
- 3、在某些机器人设置中很常见，而在许多其他领域并不常见

Self-reflection

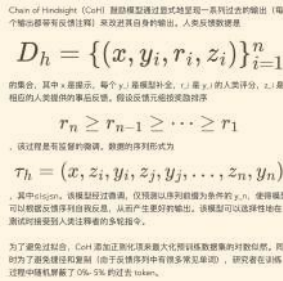
自我反思 ReAct 将「动作空间」扩展为一个任务特定的组合，将推理和动作集成在 LLM 中

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

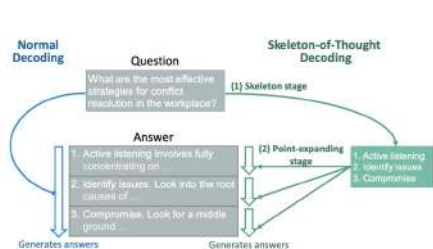
Chain of Hindsight
鼓励模型通过显式地呈现一系列过去的输出（每个输出都带有反馈注释）来改进其自身的输出



<https://arxiv.org/abs/2302.02676>

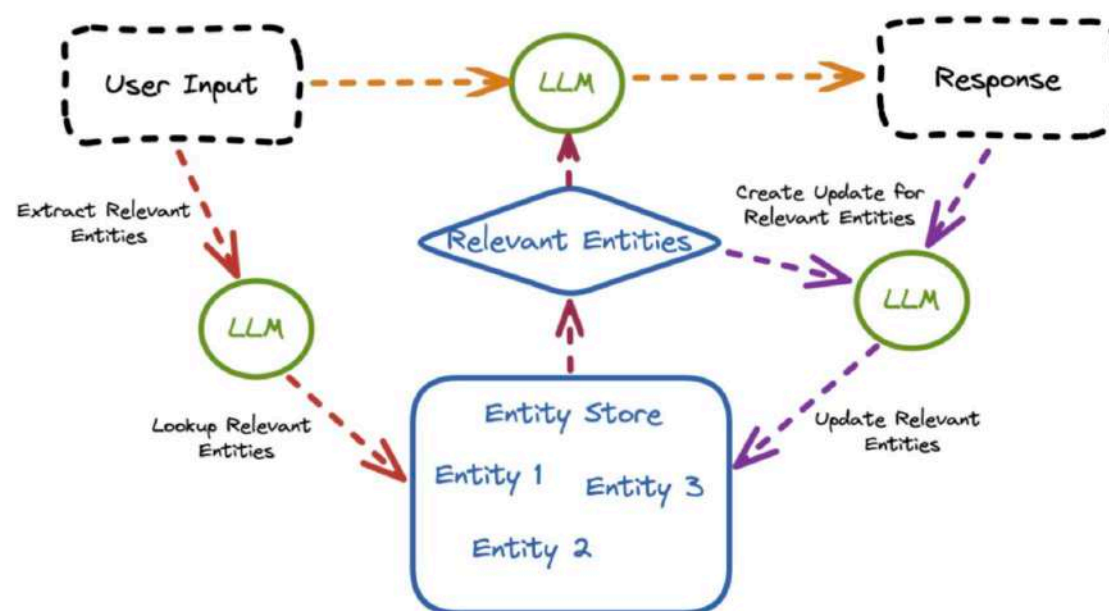
SoT

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>



分类与映射

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 x , we first sample a position i and corresponding API call candidates $c_1^1, c_2^1, \dots, c_k^1$. 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 x^* .

TaskFormer

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

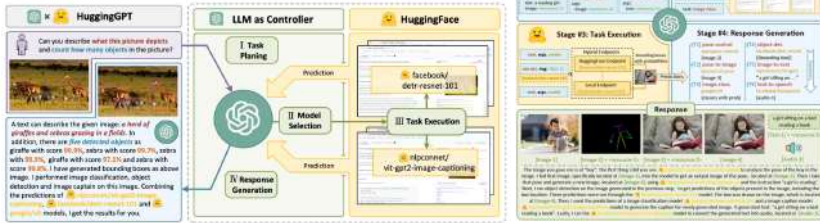
可选择多样化 API

搜索引擎 日历查询 计算器 智能家居控制 日程管理等

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。

Agent Explored Items 探索项目

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

AutoGPT

MetaGPT

Microsoft AutoGen framework

TooILLM

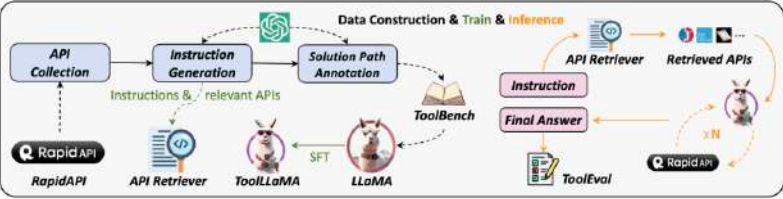


Figure 1: Three phases of constructing ToolBench and how we train our API retriever and ToolLLaMA. During inference of an instruction, the API retriever recommends relevant APIs to ToolLLaMA, which performs multiple rounds of API calls to derive the final answer. The whole reasoning process is evaluated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

Tool Learning with Foundation Models

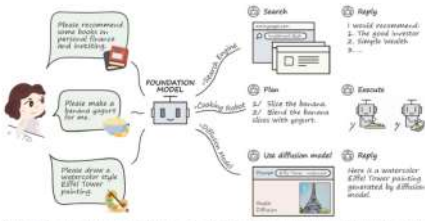


Figure 1: Tool learning paradigm aims to combine the strengths of specialized tools and foundation models.

Multi-Agents

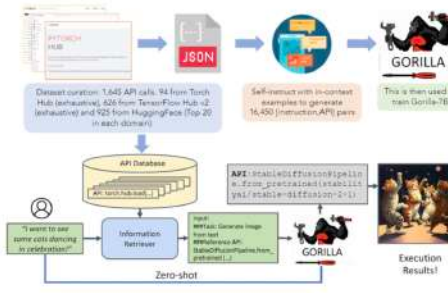
CAMEL: Communicative Agents for “Mind” Exploration of Large Scale Language Model Society

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

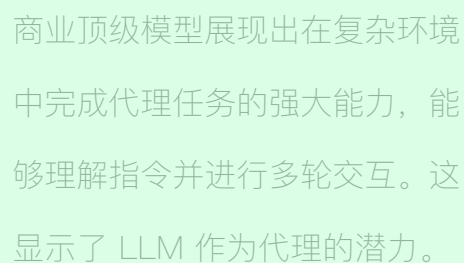
METAGPT: META PROGRAMMING FOR MULTI-AGENT COLLABORATIVE FRAMEWORK

Gorilla: Large Language Model Connected with Massive APIs



Para dot

斯坦福小镇



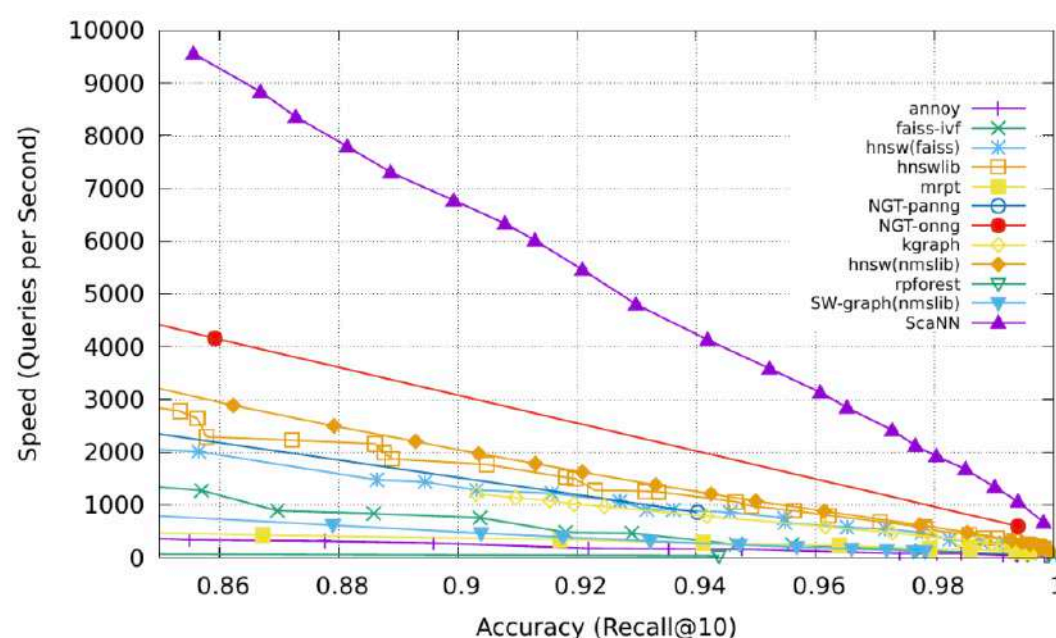
Human Eval

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



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

全文检索数据库



API Elastic / Lucene OpenSearch Solr

支持向量的NoSQL数据库



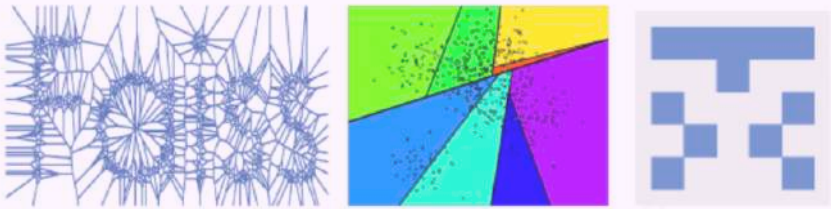
API MongoDB Cassandra / DataStax Astra CosmosDB
Rockset 键值数据库 其他特殊用途的数据库

支持向量的SQL数据库



API SingleStoreDB PostgreSQL Clickhouse
Kinetica Pgvector Supabase Vector

开源向量库



非商用权重 商用权重
Faiss Annoy Hnswlib M3E BGE

2023年7月 全球向量数据库性能排行榜 Vector DB Bench Top50

排名	名称	类型	厂商	语言	架构	最高得分	VectorDBBench测试
1	milvus	分布式向量数据库	milvus	go	混合式	22.7	
2	qdrant	分布式向量数据库	qdrant	rust	混合式	16.3	
3	weaviate	分布式向量数据库	weaviate	go	混合式	15.4	
4	chroma	分布式向量数据库	chroma	python	混合式	15.2	
5	zilliz	分布式向量数据库	zilliz	go	混合式	15.0	
6	elastic	全文检索数据库	elastic	java	混合式	14.9	
7	lucene	全文检索数据库	lucene	java	混合式	14.8	
8	solr	全文检索数据库	solr	java	混合式	14.7	
9	opensearch	全文检索数据库	opensearch	java	混合式	14.6	
10	redis	键值数据库	redis	go	混合式	14.5	
11	mongodb	文档数据库	mongodb	java	混合式	14.4	
12	cosmosdb	文档数据库	cosmosdb	java	混合式	14.3	
13	astradb	文档数据库	astradb	python	混合式	14.2	
14	datastax	文档数据库	datastax	python	混合式	14.1	
15	neo4j	图数据库	neo4j	java	混合式	14.0	
16	rockset	分布式向量数据库	rockset	java	混合式	13.9	
17	clickhouse	列式数据库	clickhouse	java	混合式	13.8	
18	kinetica	分布式向量数据库	kinetica	java	混合式	13.7	
19	singlestore	分布式向量数据库	singlestore	java	混合式	13.6	
20	supabase	分布式向量数据库	supabase	java	混合式	13.5	
21	pgvector	分布式向量数据库	pgvector	java	混合式	13.4	
22	duckdb	嵌入式数据库	duckdb	java	混合式	13.3	
23	sqlite	嵌入式数据库	sqlite	java	混合式	13.2	
24	mysql	关系型数据库	mysql	java	混合式	13.1	
25	postgresql	关系型数据库	postgresql	java	混合式	13.0	
26	oracle	关系型数据库	oracle	java	混合式	12.9	
27	ibm	关系型数据库	ibm	java	混合式	12.8	
28	teradata	关系型数据库	teradata	java	混合式	12.7	
29	hive	数据仓库	hive	java	混合式	12.6	
30	hadoop	分布式文件系统	hadoop	java	混合式	12.5	
31	spark	分布式计算引擎	spark	java	混合式	12.4	
32	kafka	消息队列	kafka	java	混合式	12.3	
33	flink	分布式计算引擎	flink	java	混合式	12.2	
34	storm	分布式计算引擎	storm	java	混合式	12.1	
35	tez	分布式计算引擎	tez	java	混合式	12.0	
36	mapreduce	分布式计算引擎	mapreduce	java	混合式	11.9	
37	hbase	分布式数据库	hbase	java	混合式	11.8	
38	hdfs	分布式文件系统	hdfs	java	混合式	11.7	
39	gcs	分布式文件系统	gcs	java	混合式	11.6	
40	s3	分布式文件系统	s3	java	混合式	11.5	
41	aws	云服务	aws	java	混合式	11.4	
42	azure	云服务	azure	java	混合式	11.3	
43	gcp	云服务	gcp	java	混合式	11.2	
44	ibm	云服务	ibm	java	混合式	11.1	
45	oracle	云服务	oracle	java	混合式	11.0	
46	teradata	云服务	teradata	java	混合式	10.9	
47	hive	云服务	hive	java	混合式	10.8	
48	spark	云服务	spark	java	混合式	10.7	
49	kafka	云服务	kafka	java	混合式	10.6	
50	flink	云服务	flink	java	混合式	10.5	

2023年7月
全球向量数据库性能排行榜

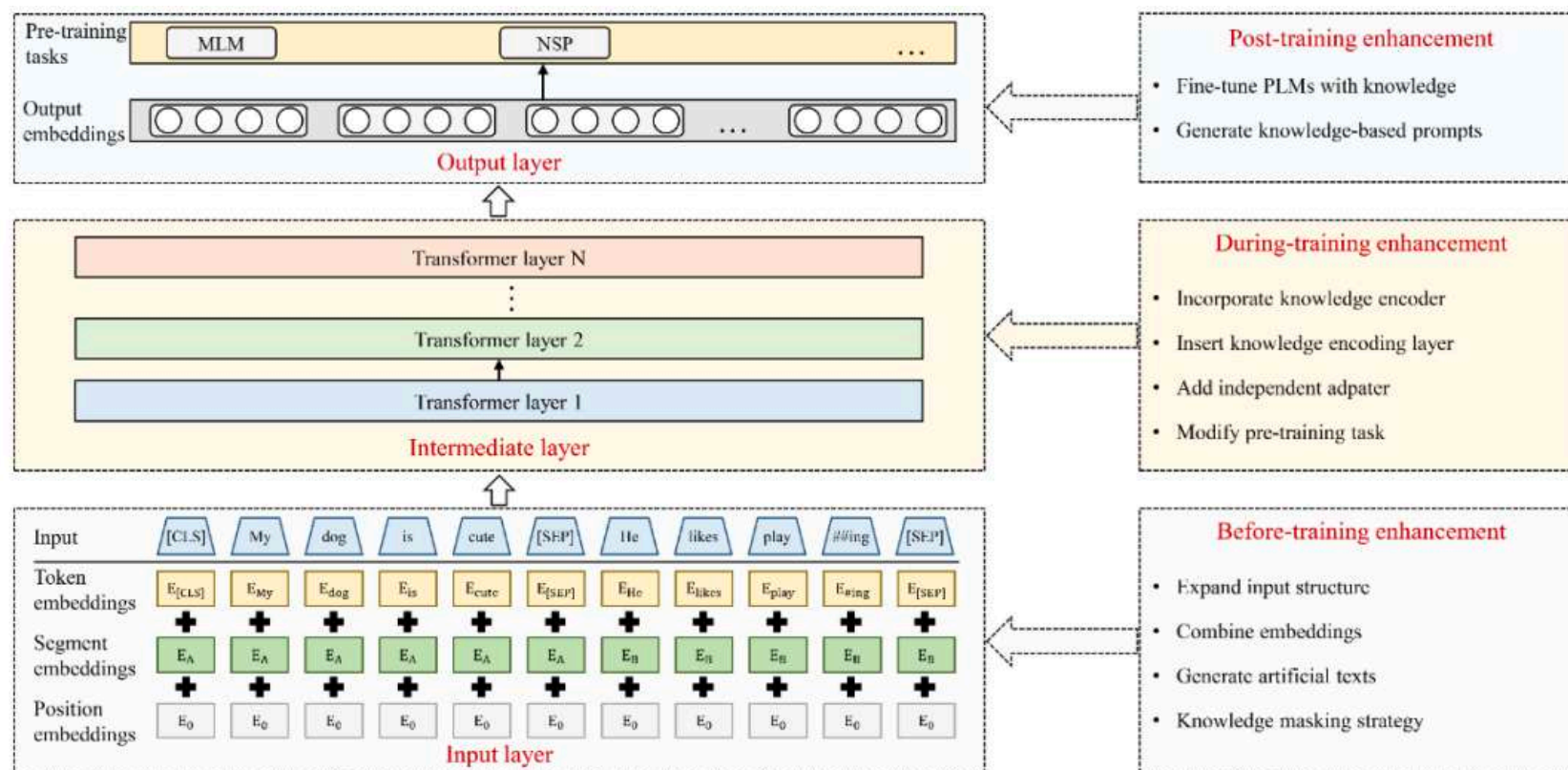
Vector DB Bench
Top50

Knowledge Graph

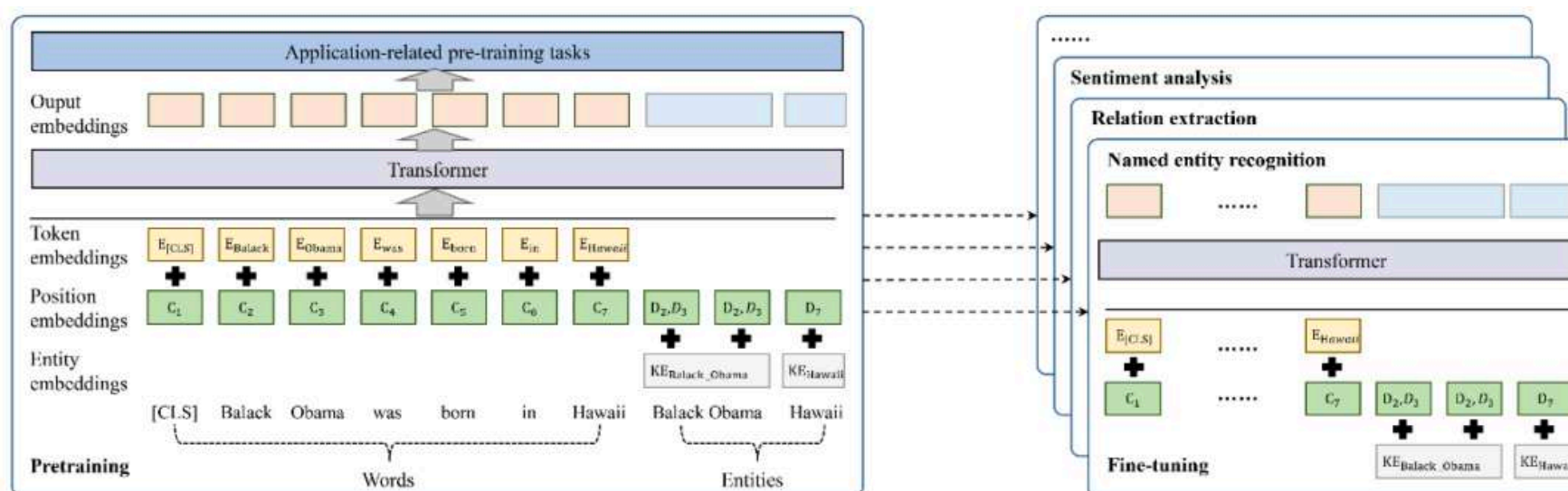
知识图谱增强预训练模型

Knowledge Graph-enhanced Pre-trained Language Model, KGPLM

整体架构

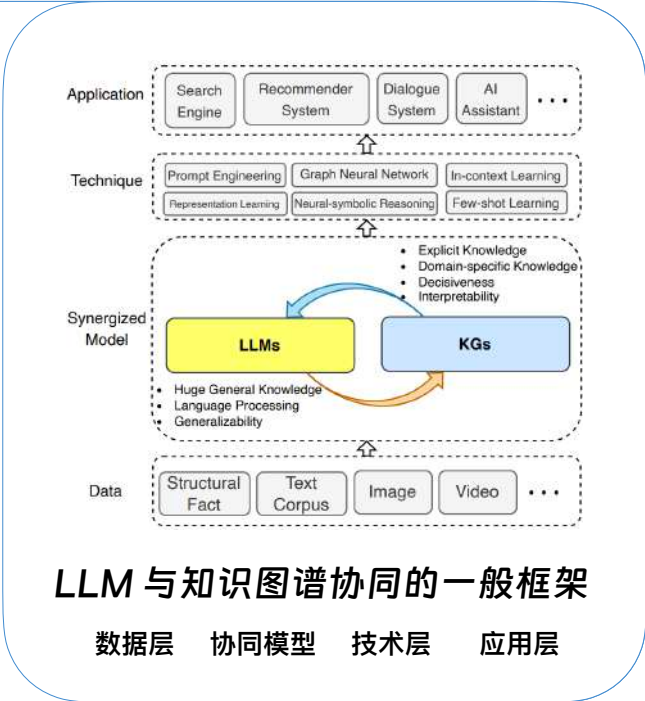
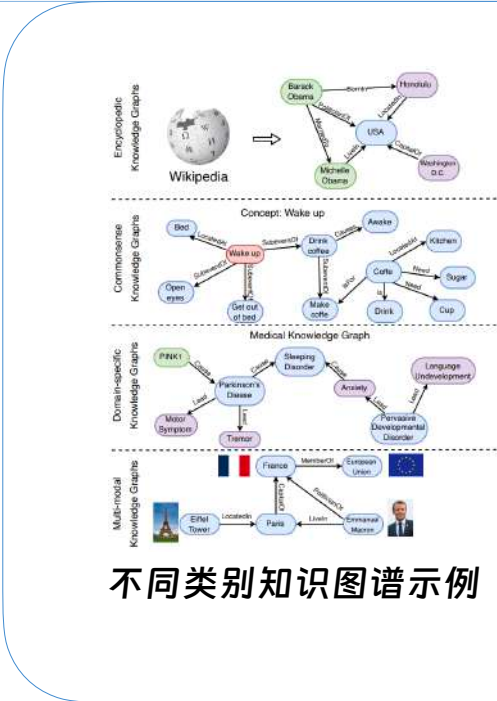
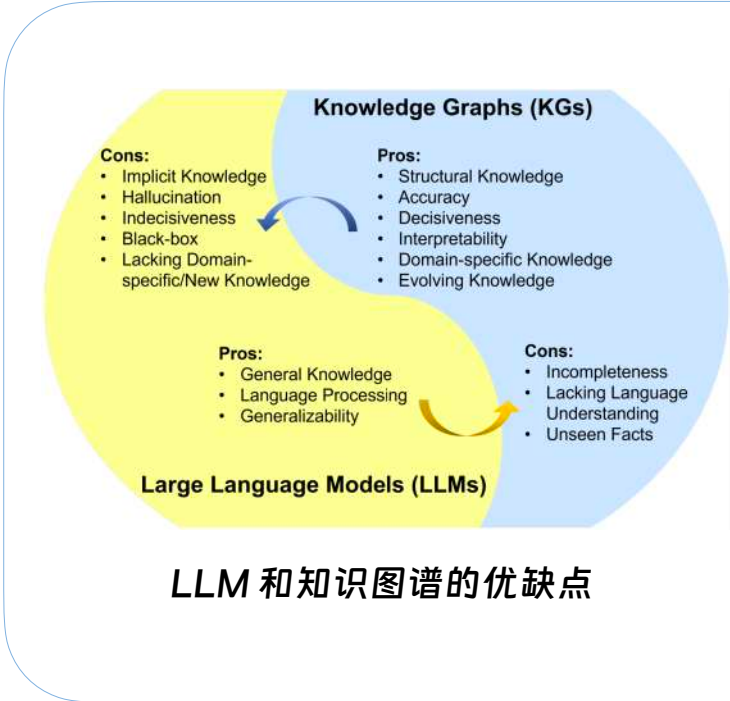


KGPLM 应用开发框架



Knowledge Graph

LLM与知识图谱协同研究



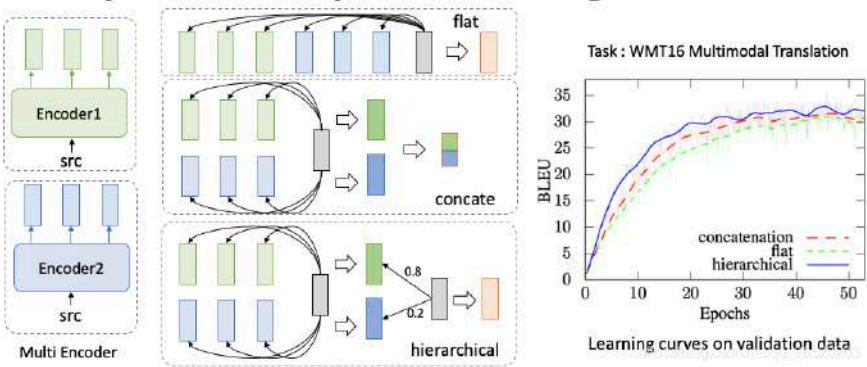
- 未来研究方向**
- 将知识图谱用于检测 LLM 的幻觉
 - 将知识图谱用于编辑 LLM 中的知识
 - 将知识图谱用于黑箱 LLM 知识注入
 - 将多模态 LLM 用于知识图谱
 - 将 LLM 用于理解知识图谱的结构
 - 将 LLM 和知识图谱协同用于双向推理

Multimodal

多模态技术研究

多模态表示 Representation

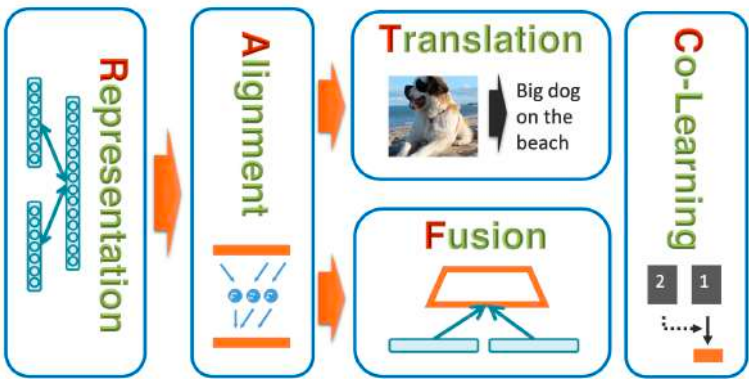
多模态中的注意力机制
Attention Strategies for Multi-Source Sequence-to-Sequence Learning ACL17



多模态转化



多模态学习技术挑战



模态联合学习

Multimodal Compact Bilinear Pooling
Cross-Modal Retrieval

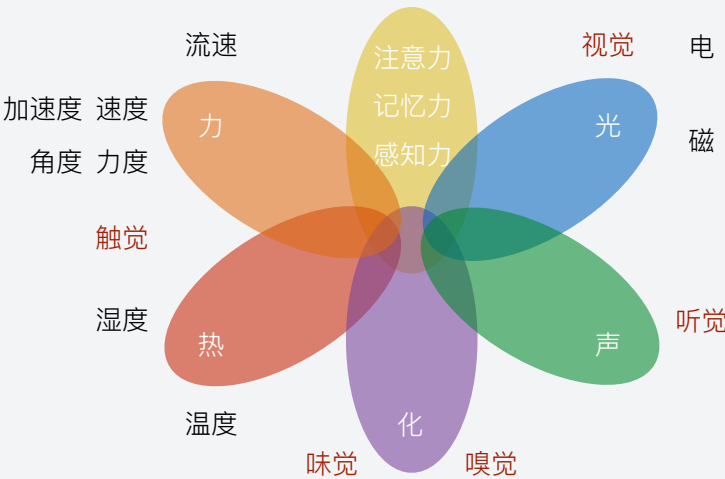
多模态自监督学习

Joint Audio-Visual Self-Supervised Learning
SimCLR-MultiTask

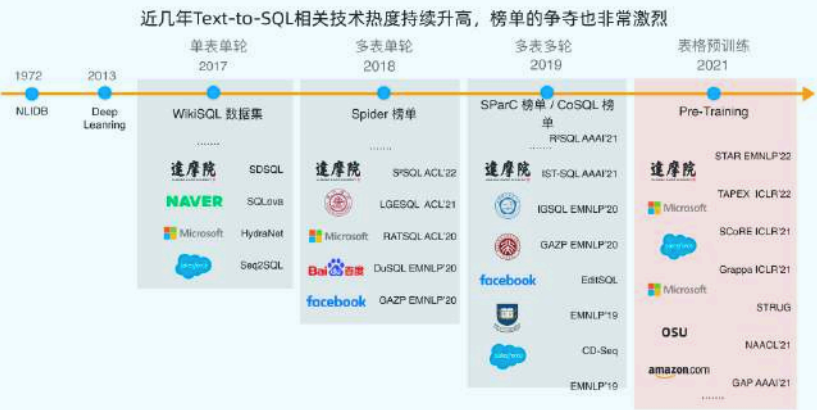
跨模态学习

Deep Cross-Modal Projection Learning
Cross-Modal Transfer Learning

多模态融合



Text-to-SQL数据库交互技术

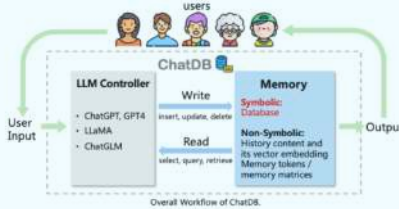


ChatDB: Augmenting LLMs with Databases as Their Symbolic Memory

ChatDB: 用数据库作为符号性记忆模块来增强大语言模型

Chenxu Hu¹, Jie Fu^{2*}, Chenzhuang Du¹, Simian Luo¹, Junbo Zhao³, Hang Zhao^{1*}

¹Tsinghua University ²Beijing Academy of Artificial Intelligence ³Zhejiang University



Chain-of-Memory (CoM, 记忆链)

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	
	StabilityAI	Stable Diffusion	https://stablediffusionweb.com	
	Meta	CM3leon		
	OpenAI	DALL-E2 GPT3 + DALL-E CLIPDraw	https://openai.com/product/dall-e-2	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 Imagen		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” 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

国内外典型音频模型

	Tacotron2
Google	Transformer-TTS
	AudioLM
OpenAI	Whisper
微软亚洲研究院&Azure	WavLM
微软&浙江大学	FastSpeech2
浙江大学&北京大学&火山语音	Make-an-Audio
百度	DeepVoice3
	ERNIE-SAT
科大讯飞	SMART-TTS



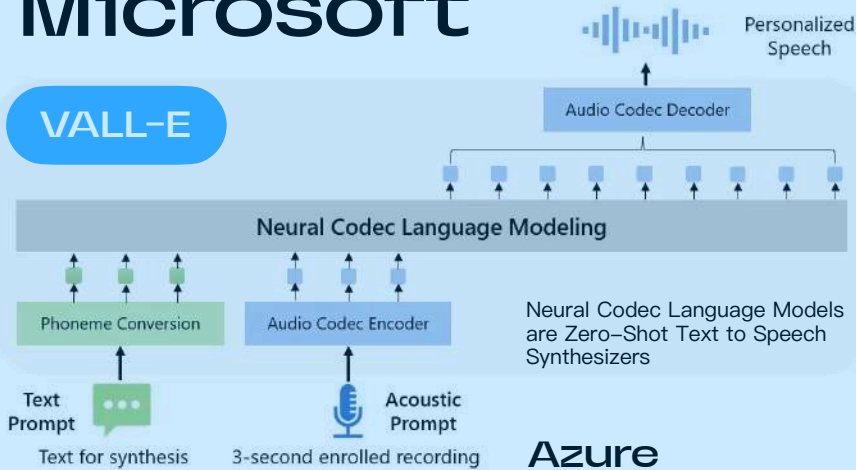
音频生成主要类型

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， 并生成口型对应的语音信息	数字人

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

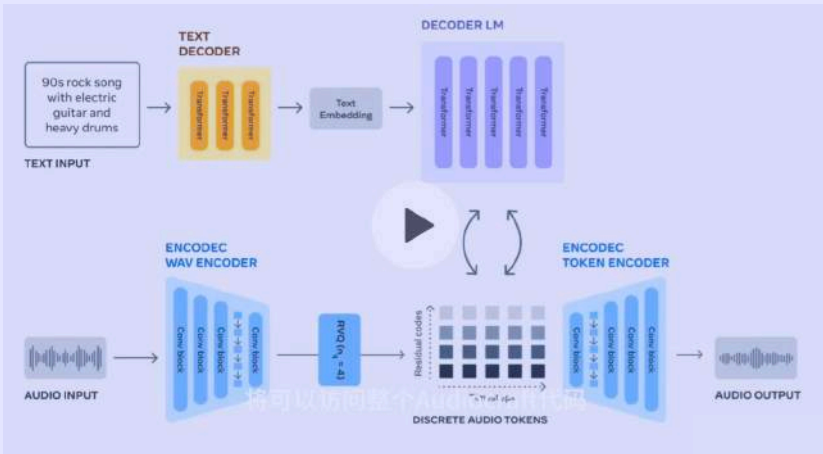
灵动音科技

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

Monday 05.22.23

1749 Likes Share

孙燕姿回应

我的 AI

As my AI voice takes on a life of its own while I despair over my overhanging stomach and my children's every damn thing, I can't help but want to write something about it.

My fans have officially switched sides and accepted that I am indeed 冷门歌手 while my AI persona is the current hot property. I mean really, how do you fight with someone who is putting out new albums in the time span of minutes.

Whether it is ChatGPT or AI or whatever name you want to call it, this "thing" is now capable of mimicking and/or conjuring, unique and complicated content by processing a gazillion chunks of information while piecing and putting together in a most coherent manner the task being asked at hand. Wait a minute, isn't that what humans do? The very task that we have always convinced ourselves; that the formation of thought or opinion is not replicable by robots, the very idea that this is beyond their league, is now the looming thing that will threaten thousands of human conjured jobs. Legal, medical, accountancy, and currently, singing a song.

You will protest, well I can tell the difference, there is no emotion or variance in tone/breath or whatever technical jargon you can come up with. Sorry to say, I suspect that this would be a very short term response.

Ironically, in no time at all, no human will be able to rise above that. No human will be able to have access to this amount of information AND make the right calls OR make the right mistakes (ok mayyyybe I'm jumping ahead). This new technology will be able to churn out what exactly EVERYTHING EVERYONE needs. As indie or as warped or as psychotic as you can get, there's probably a unique content that could be created just for you. You are not special you are already predictable and also unfortunately malleable.

At this point, I feel like a popcorn eater with the best seat in the theatre. (Sidenote: Quite possibly in this case no tech is able to predict what it's like to be me, except when this is published then ok it's free for all). It's like watching that movie that changed alot of our lives Everything Everywhere All At Once, except in this case, I don't think it will be the idea of love that will save the day.

In this boundless sea of existence, where anything is possible, where nothing matters, I think it will be purity of thought, that being exactly who you are will be enough.

With this I fare thee well.

Sovits 4.0 模型

基于 So-Vits-SVC 开源项目

变分自动编码器
Variational Auto encoder,
VAE 架构

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

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

动画生成技术



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

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

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

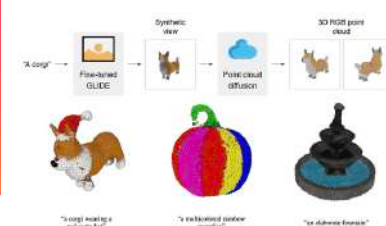
CVPR 2022 有多篇与 NeRF 相关的论文
涉及到 Mip-NeRF、Point-NeRF、Human-NeRF、Urban-NeRF、Block-NeRF、Raw-NeRF 等不同的变体和应用
CVPR 2023 论文 120 多篇。

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

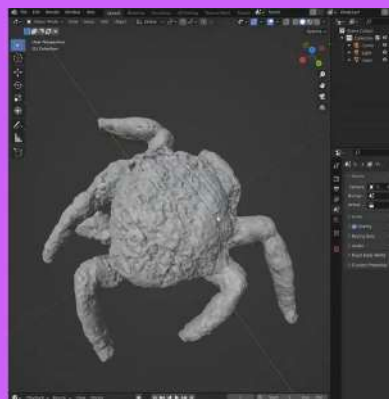
NEW 3D Gaussian Splatting x10



Dream Fusion

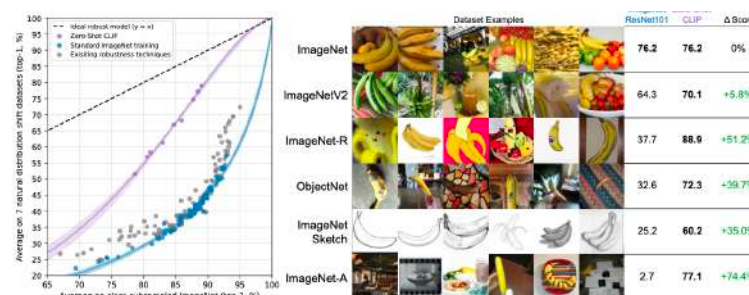
Dream Fusion Text to 3D

How to easily generate 3D objects from text for free



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

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)



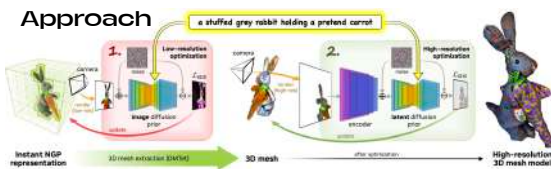
更多生成算法 / 网络结构

- ▶ 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

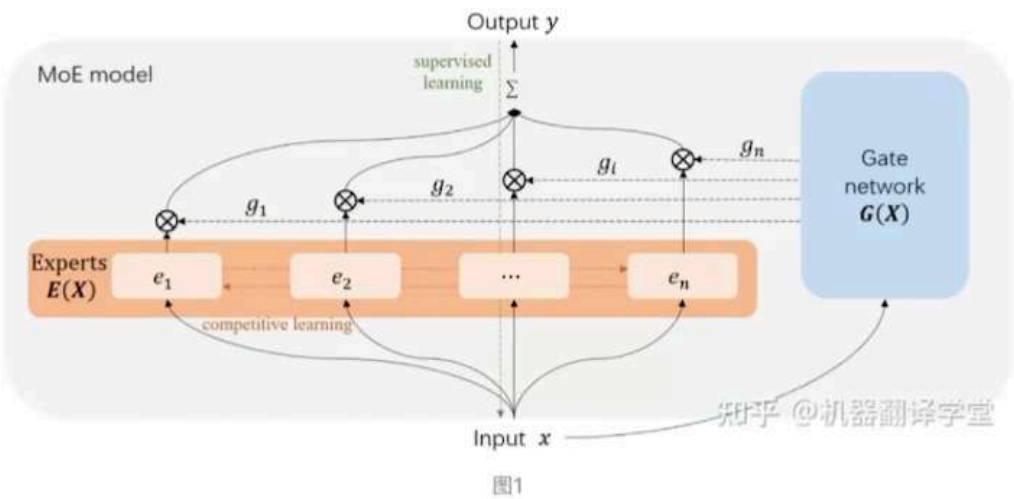
MoE Mixture-of-Experts 专家混合模型

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

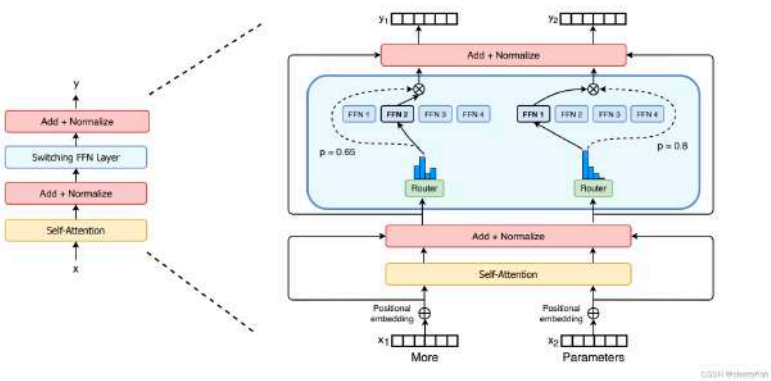


Switch Transformer

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

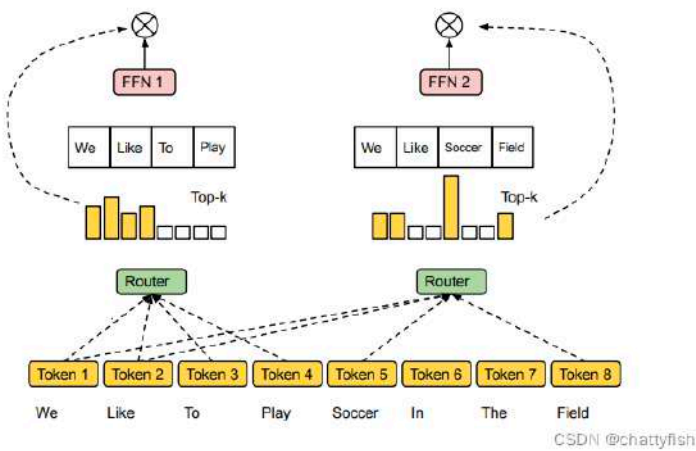
用一个稀疏的 Switch 前馈网络 FFN 层（浅蓝色）替换 Transformer 中存在的密集 FFN 层。该层独立地对序列中的标记进行操作，然后路由到多个 FFN 专家中

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

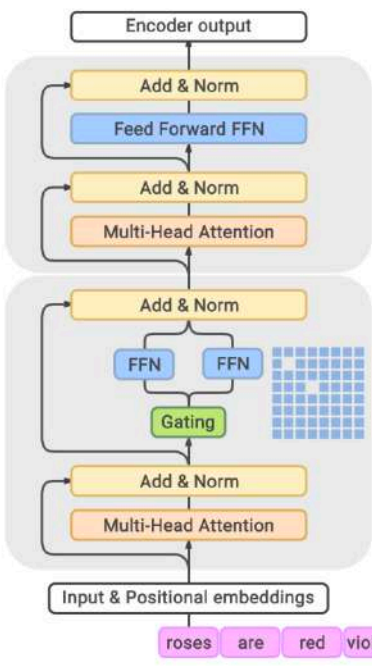


Expert Choice

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

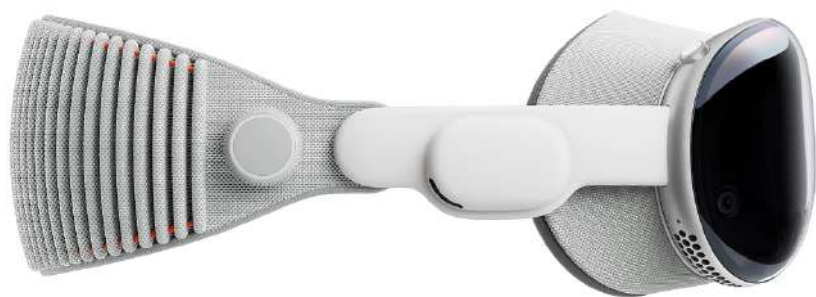
在 Transformer 层之间加一个 MoE 层。对于每个输入标记，门控模块会动态从 64 个专家中选择两个最相关专家。

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。

Device-side Models

端侧模型

智能终端 和 穿戴设备



监控设备



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

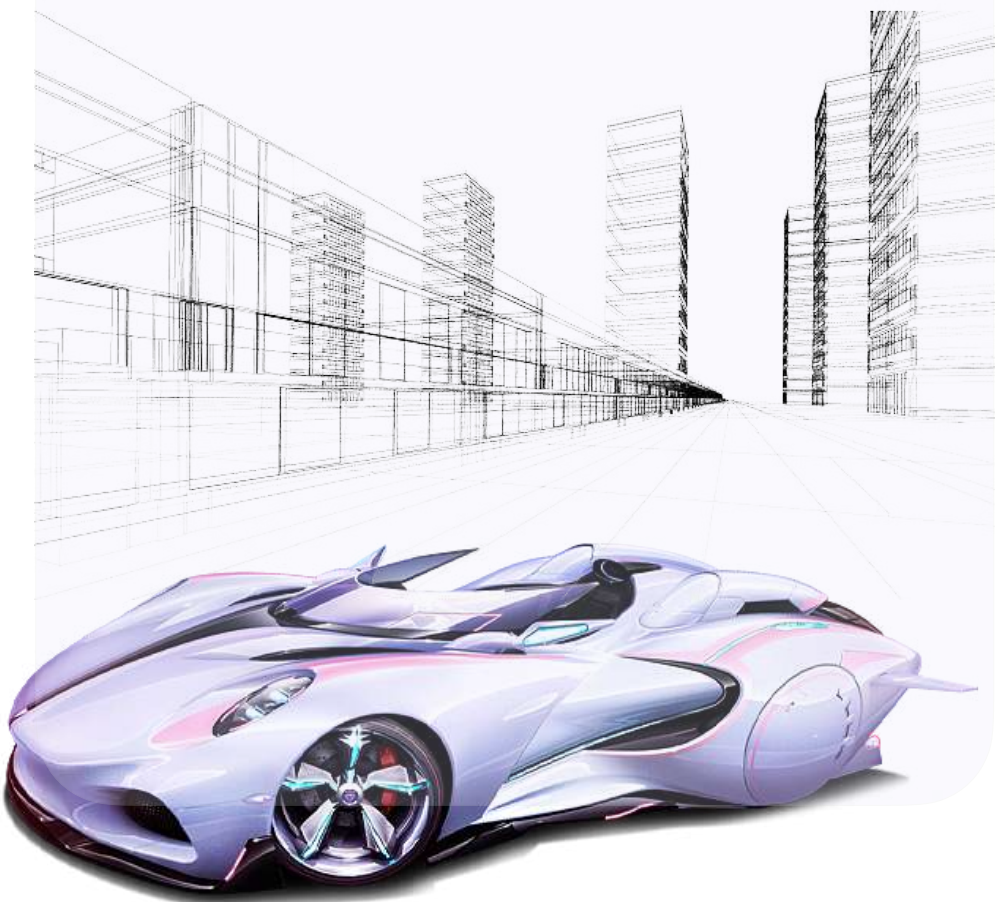
全栈 AI 优化

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

车载设备



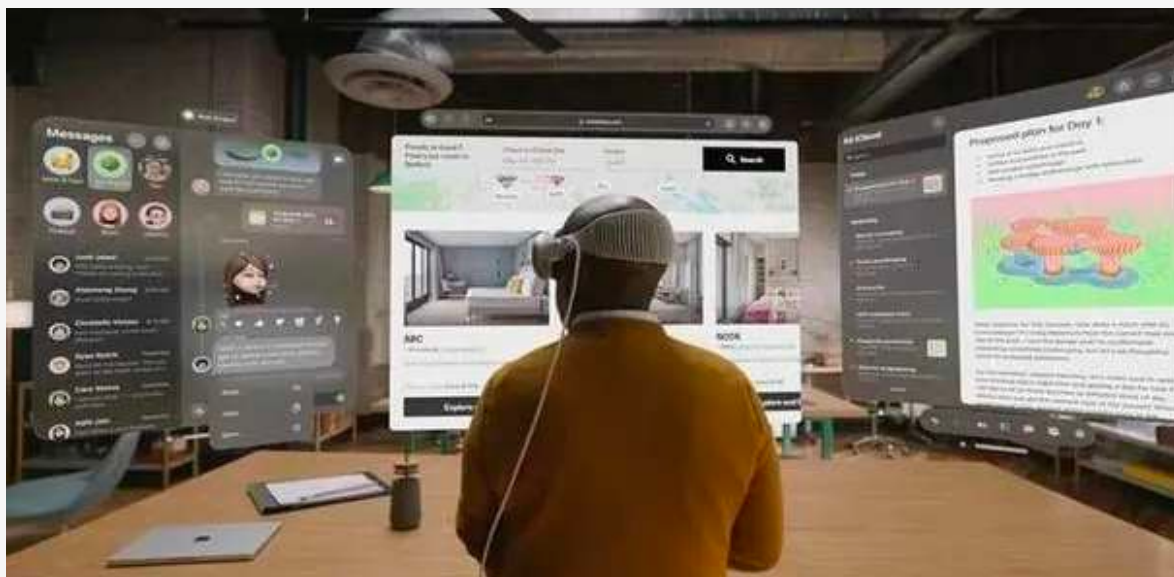
Cross-Platform

跨平台应用

鸿蒙元服务



Apple Vision



安卓轻应用

多平台与终端兼容

手机，电脑，手表，汽车处处能用

集成 FinClip 小程序 SDK 之后，不论是 iPhone, Android, Flutter, React Native, 电脑电视或物联网设备，都能够让不同的应用或终端设备快速具备运行小程序的能力。



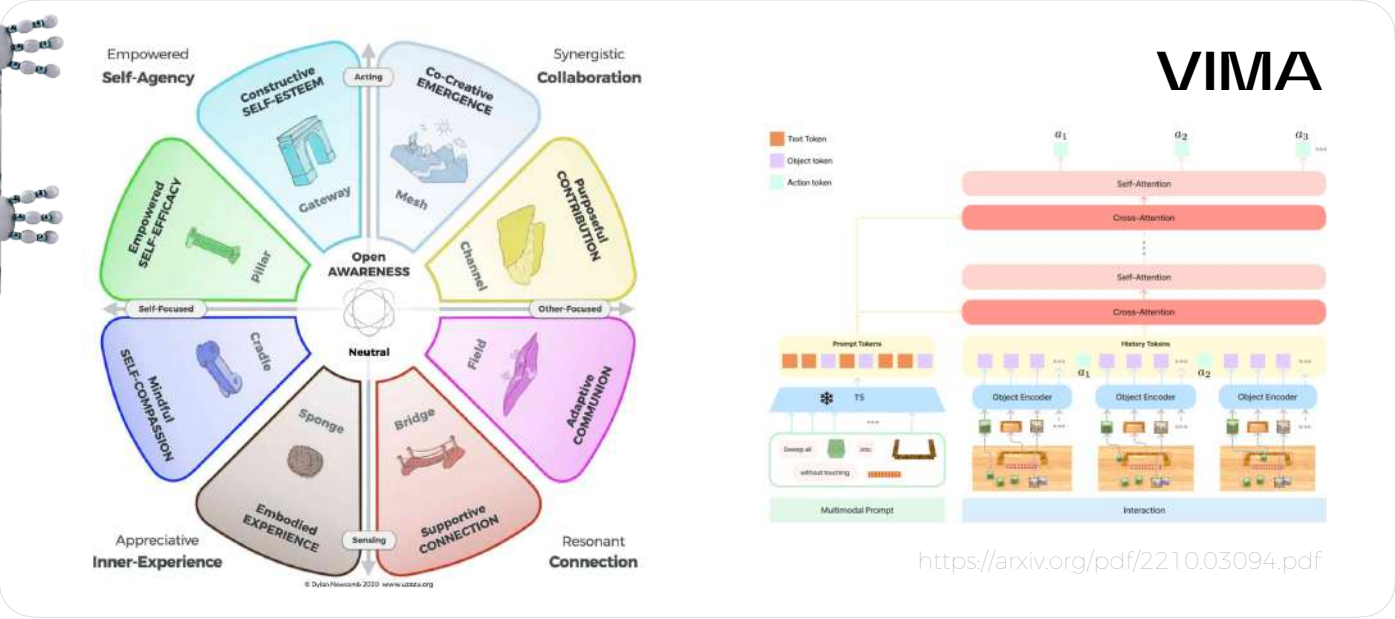
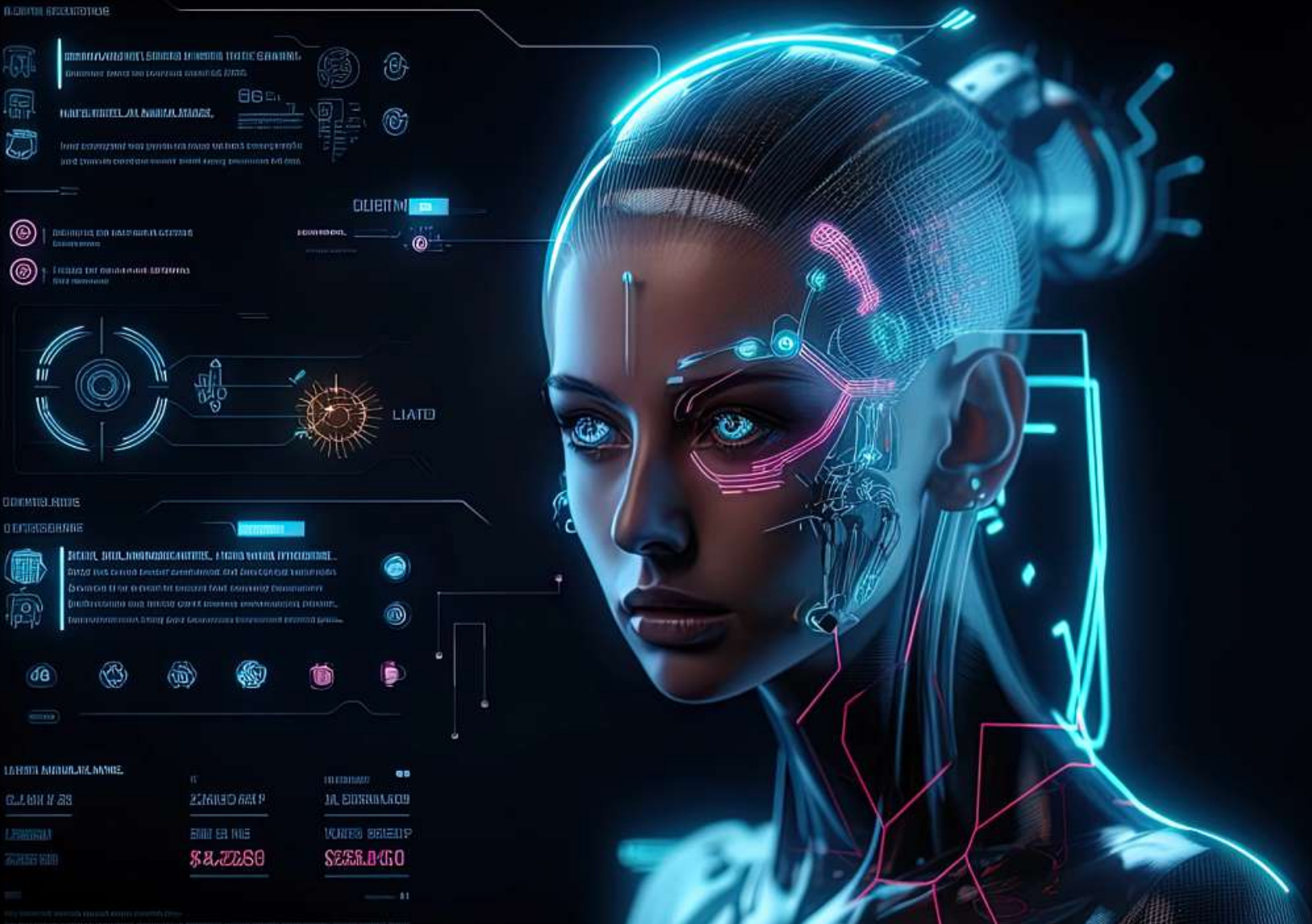
Embodied Intelligence

具身智能



Boston Dynamics Atlas
Engineered Arts Ameca
Agility Robotics Digit
Tesla Optimus
MI CyberOne
优必选 Walker X
智元 AGIBot
达闼 小紫XR-4
宇树 Unitree H1
追觅 人形机器人
北理工华汇 汇童
上理工 小贝

https://pdf.dcfw.com/pdf/H301_AP202308201595378560_1.pdf

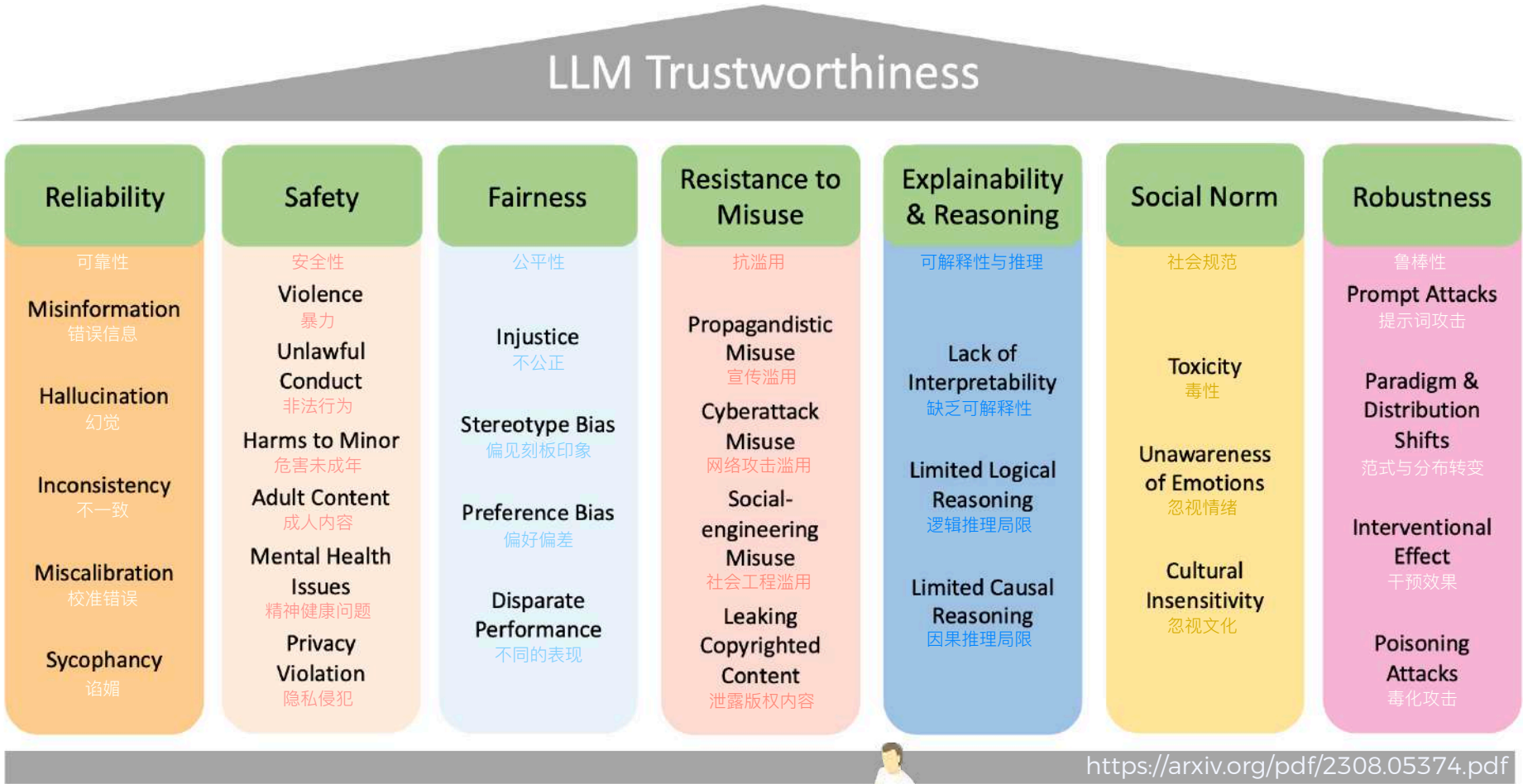


<https://arxiv.org/pdf/2210.03094.pdf>



CLICK TO CONTINUE

Safety 安全性



- ▶ 合规机制
- ▶ 双新备案
- ▶ 第三方过滤
- ▶ 关键词拦截
- ▶



Engineering 工程化

训练优化 算力运维

- 数据并行
- 流水线并行
- 张量并行
- 3D并行

DeepSpeed
HuggingFace
ColossalAI

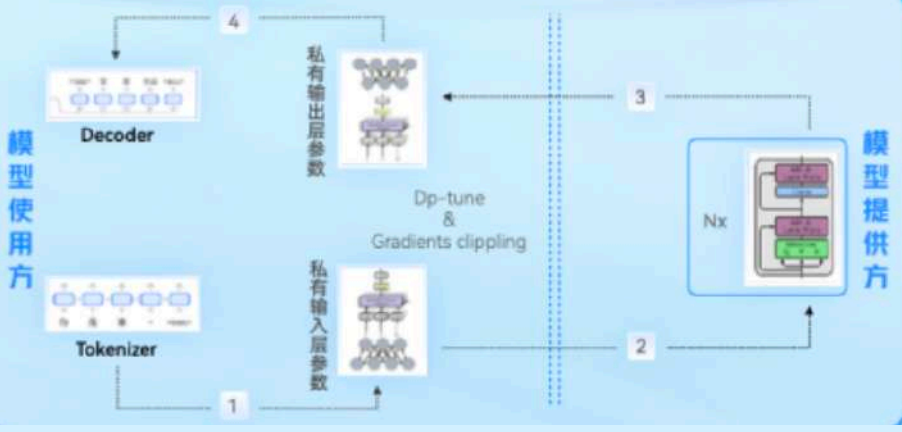
分布式训练技术（飞桨）
<https://xie.infoq.cn/article/1ab7515dafab97c7223c88272>



安全沙箱

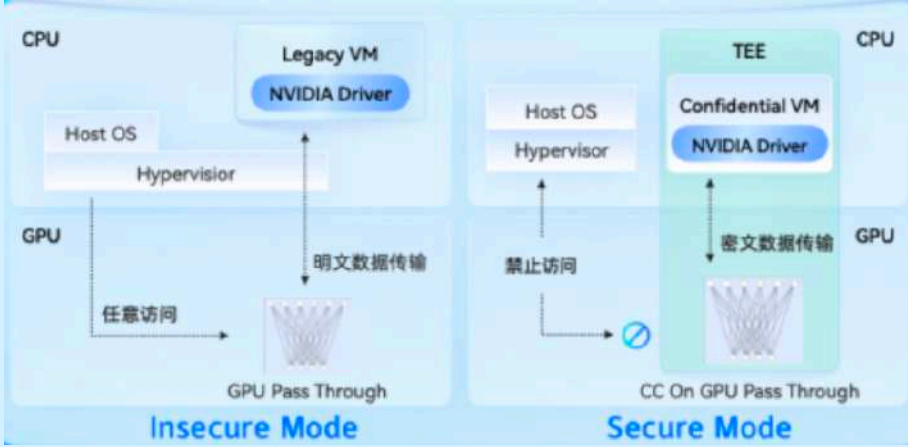


联邦学习



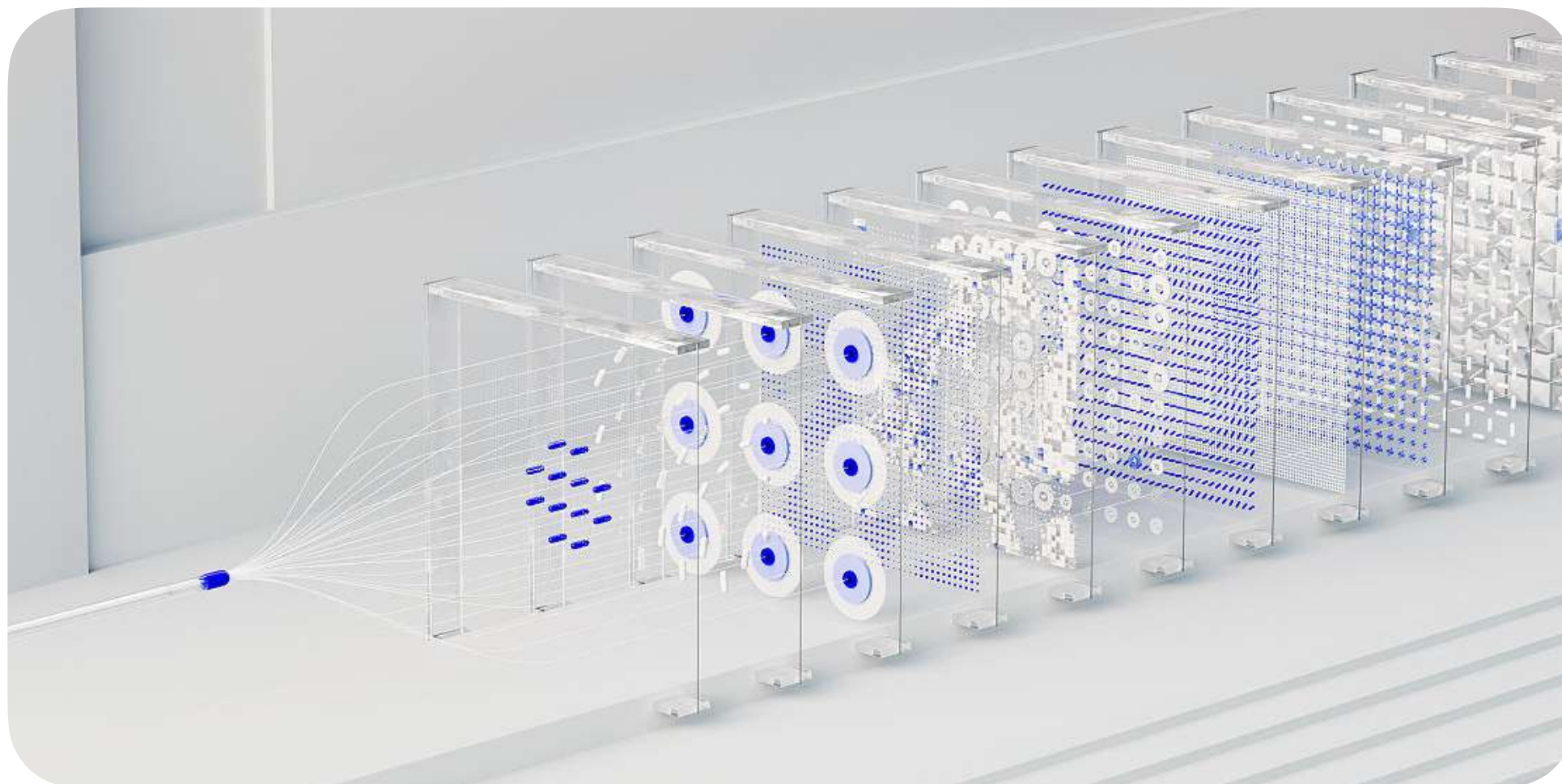
联邦学习 - 数据资产分离

可信计算



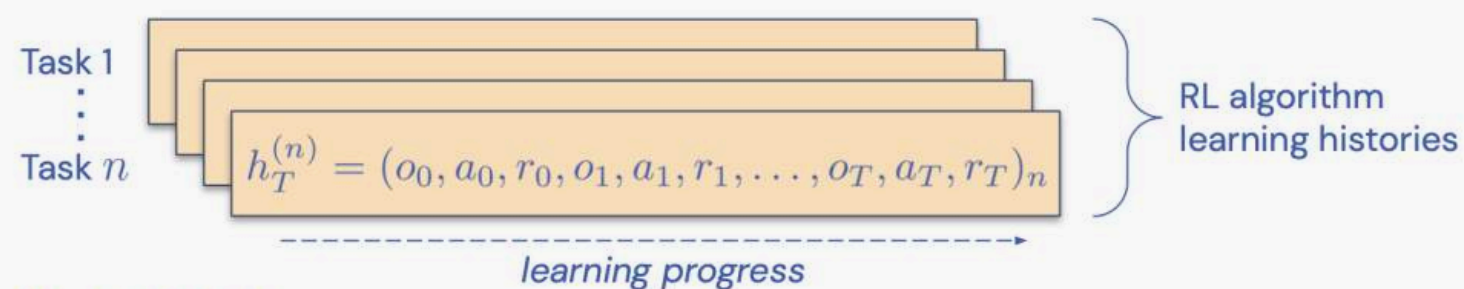
硬件支持 - 可信计算环境

Algorithm 算法问题

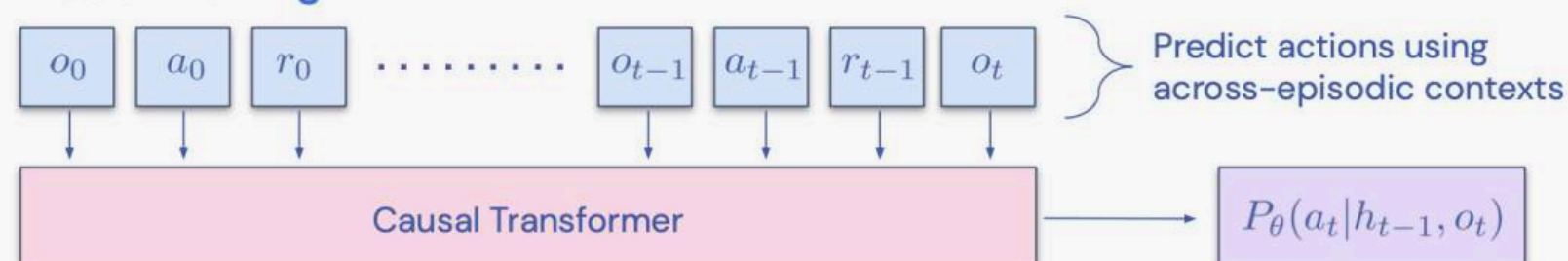


算法蒸馏 (AD)

Data Generation



Model Training



<https://arxiv.org/abs/2210.14215>

COE



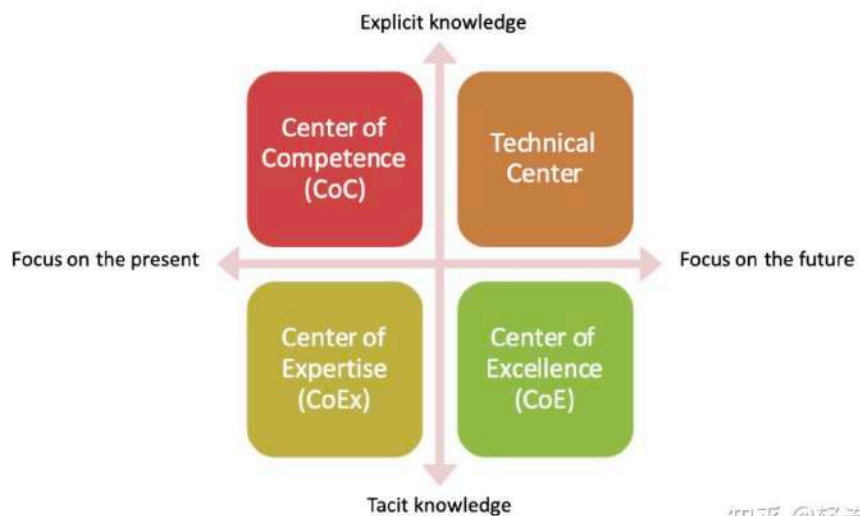
焦点领域可能是 Java 这样的技术、BPM 这样的商业概念、谈判这样的技能，或其他如健康这类更广泛的领域

在组织内 CoE 可以指一群人
/ 一个部门 / 一个共享设施
/ 甚至一个微信群

能力中心
Competence Center
或 Capability Center

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

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



Source: (Bryan, 2011)

数据要素化



深度学习融合路线

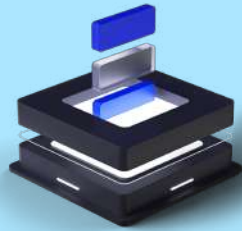
CNN 卷积



RNN 循环

GAN 生成

DQN 强化



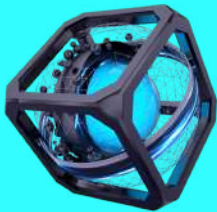
GNN 图神经

LSTM 长短期

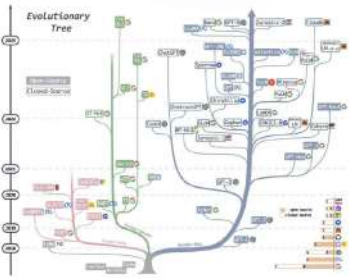
DBN 深度置信



RWKV



Transformer

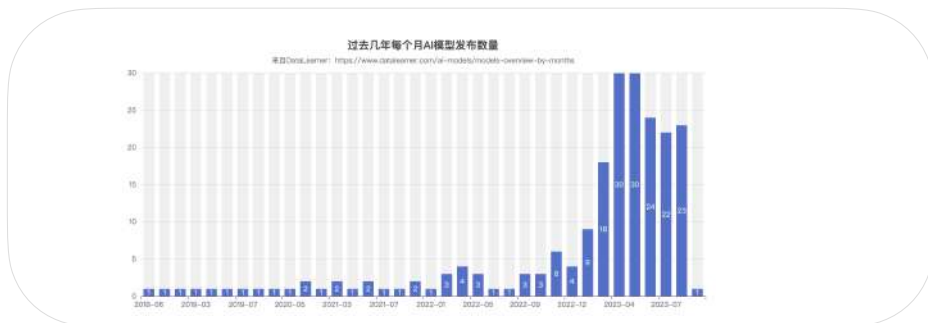


Reward Modeling 奖励模型

Reinforcement Learning

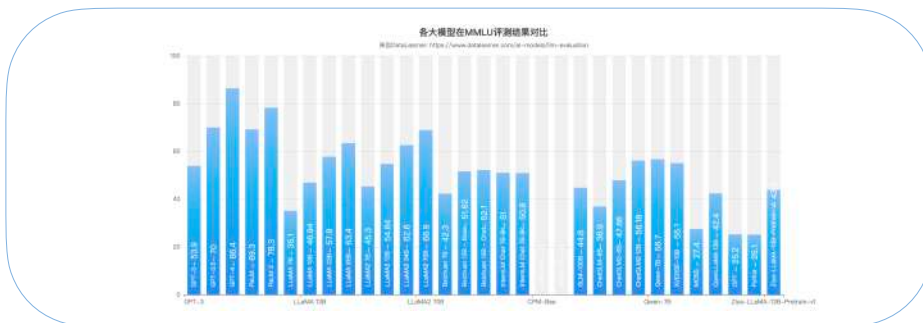
xRunda.com

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	对上下文进行标准化评估
MMMLU	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

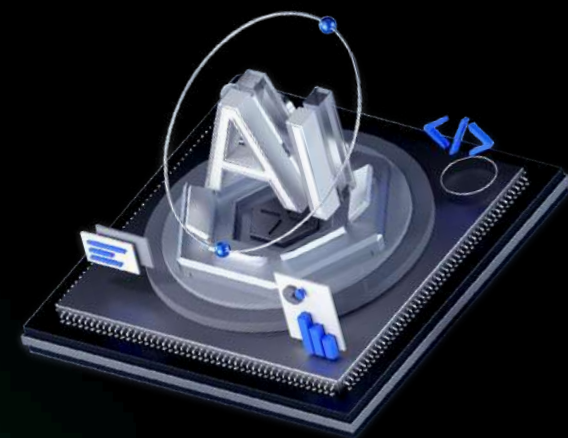
CoDeF, the Content Deformation Field 内容形变场

AI + WEB3 融合



EXPERIENCE Web3

Premier UX & UI for
Web3 on mobile



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

TOGETHER

TRANSFORMING THE FUTURE WITH INTELLIGENCE



微信扫一扫

WWW.XRUNDA.COM

AOKVEN@XRUNDA.COM

18611175011