

大模型应用开发的技术架构

智能架构李教头

阿里云最有价值专家

自我介绍下

阿里云最有价值专家

著有《Spring Cloud Alibaba微服务开发入门到实战》、《Java编程入门任务式学习指南》、《大数据技术入门到实践》等书。

做过开发，带过技术团队，做过课也讲过课，大模型方向创业中

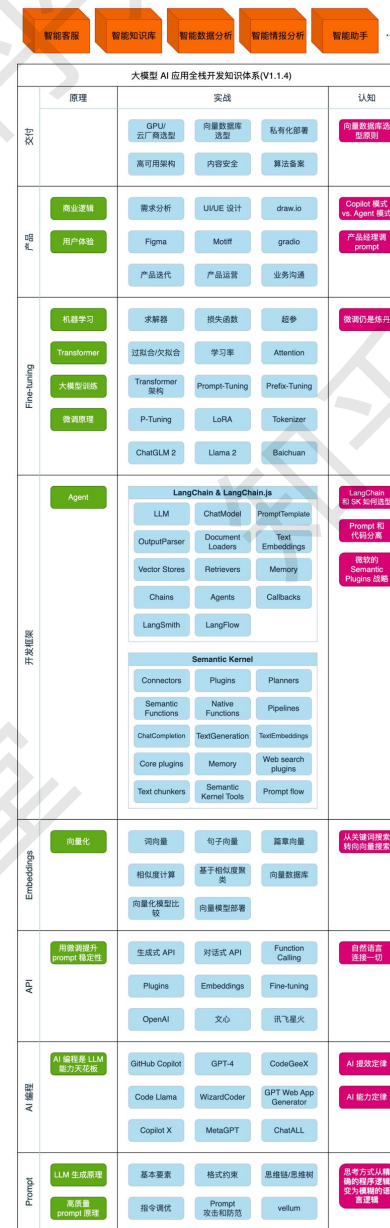


今日内容

大模型带来的行业变化

大模型应用的技术架构

大模型应用的知识体系



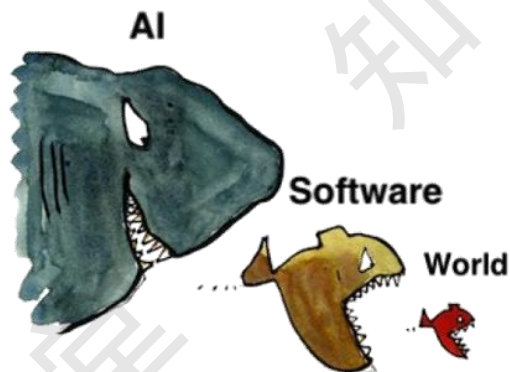
今日内容

大模型应用的技术架构

大模型应用的知识体系

大模型 AI 应用全栈开发知识体系(V1.1.4)			
交付	原理	实战	认知
		GPU/云厂商选型 向量数据库选型 高可用架构 内容安全 算法备案	私有化部署 向量数据库应用案例 算法备案
产品	商业逻辑	需求分析 UI/UE 设计 draw.io	Capitol 模式 vs Agent 模式
	用户体验	Figma Molif gradio 产品迭代 产品运营 业务沟通	产品特性与 prompt
Fine-tuning	机器学习	求解器 损失函数 超参	微调仍是伪命题
	Transformer	过拟合欠拟合 学习率 Attention	
	大模型训练	Transformer 架构 Prompt-Tuning Prefix-Tuning	
	微调原理	P-Tuning LoRA Tokenizer ChatGLM 2 Llama 2 Baichuan	
开发框架	Agent	LangChain & LangChain.js LLM ChatModel PromptTemplate OutputParser Document Loaders Text Embeddings Vector Stores Retrievers Memory Chains Agents Callbacks LangSmith LangFlow	
		Semantic Kernel Connectors Plugins Planners Semantic Functions Native Functions Pipeline ChatCompletion TextGeneration TextEmbeddings Core plugins Memory Web search plugins Text chunkers Semantic Kernel Tools Prompt flow	
Embeddings	向量化	词向量 句子向量 段落向量 相似度计算 基于相似度聚类 向量数据库 向量化模型比较 向量模型部署	从关键词检索转向向量检索
	用微调提升 prompt 稳定性	生成式 API 对话式 API Function Calling Plugins Embeddings Fine-tuning OpenAI 文心 讯飞星火	自然语言连接一切
AI 编程	AI 编程是 LLM 能力天花板	GitHub Copilot GPT-4 CodeGeex Code Llama WizardCoder GPT Web App Generator Copilot X MetaGPT ChatALL	AI 编程定律 AI 能力定律
	LLM 生成原理	基本要素 格式约束 思维链/思维树	思考方式从精确到模糊是思维链/思维树
Prompt	高质量 prompt 原理	指令微调 Prompt 攻击和防范 vellum	

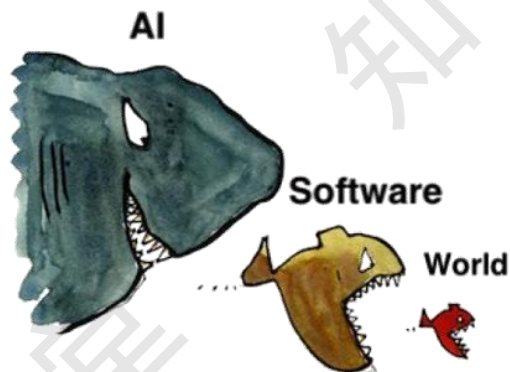
大模型应用的技术架构



“Software is eating the world”
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”
Jensen Huang, Nvidia CEO (2017)

AI 世界，有哪几种人？

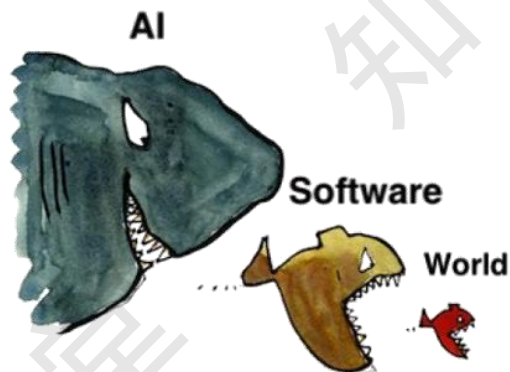


“Software is eating the world”
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”
Jensen Huang, Nvidia CEO (2017)

- AI 世界，人可以分为三类：
 - AI 使用者：使用别人开发的 AI 产品
 - AI 产品开发者：设计和开发 AI 产品
 - 基础大模型相关：训练基础大模型，或为大模型提供基础设施

为什么说 AI 产品开发者，是机会？

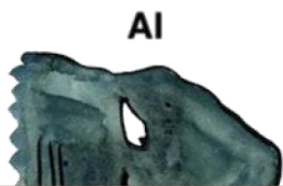


“Software is eating the world”
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”
Jensen Huang, Nvidia CEO (2017)

- AI 世界，人可以分为三类：
 - AI 使用者：使用别人开发的 AI 产品
 - AI 产品开发者：设计和开发 AI 产品
 - 基础大模型相关：训练基础大模型，或为大模型提供基础设施
- AI 产品开发者，是我们最大的机会
 - AI 使用者太普通，人人都是
 - 基础大模型相关门槛高、从业人数少，难获得机会

为什么说 AI 产品开发者，是机会？



- AI 世界，人可以分为三类：
 - AI 使用者：使用别人开发的 AI 产品

智能手机世界，人分为三类：智能手机使用者、APP 开发者、智能手机底层系统相关

基础设施

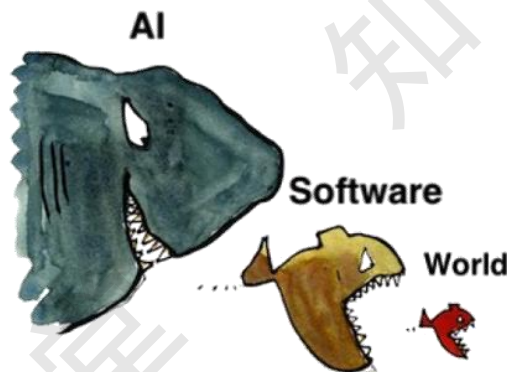
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”

Jensen Huang, Nvidia CEO (2017)

- AI 使用者太普通，人人都是
- 基础大模型相关门槛高、从业人数少，难获得机会

为什么说 AI 产品开发者，是机会？



“Software is eating the world”
Andreessen Horowitz, HP (2011)

“Software is eating the world, but AI is going to eat software”
Jensen Huang, Nvidia CEO (2017)

- AI 世界，人可以分为三类：
 - AI 使用者：使用别人开发的 AI 产品
 - AI 产品开发者：设计和开发 AI 产品
 - 基础大模型相关：训练基础大模型，或为大模型提供基础设施
- AI 产品开发者，是我们最大的机会
 - AI 使用者太普通，人人都是
 - 基础大模型相关门槛高、从业人数少，难获得机会

大模型应用的技术架构

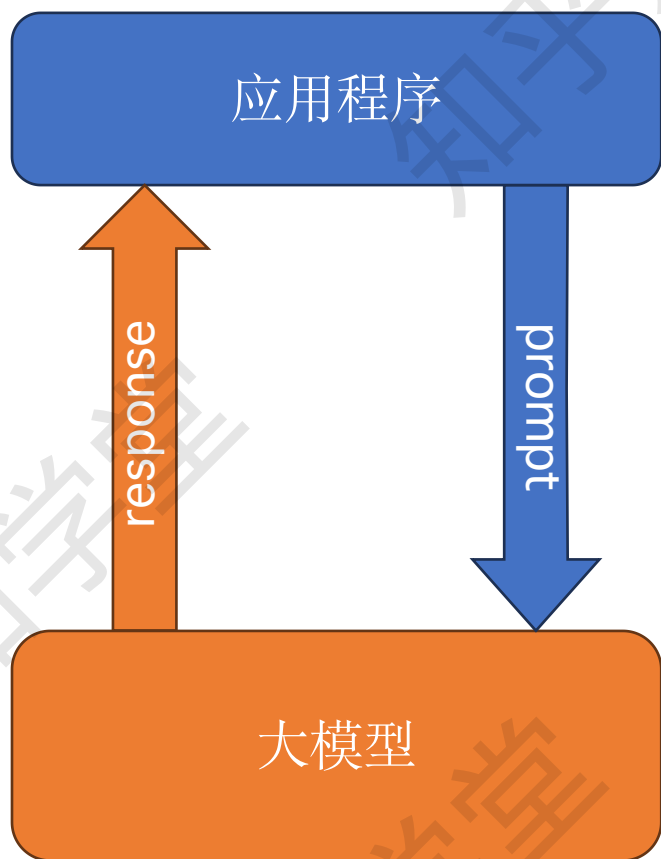
- 调用大模型，相当于调动一个人
 - TA 懂人话
 - TA 说人话
 - TA 直接给结果，但不一定对
- 核心心法：把 **AI** 当人看
 - 范式变化所在，所有方案的源泉

大模型技术很仿人

要解决的问题	举例	人的思路	大模型的思路
布置任务	查数据	对话	Prompt Engineering
新知识/记不住	报税	学习资料	RAG
深度理解	学新语言	好好学习	Fine-tuning
对接外界	获知天气	各种工具	Function Calling
解决复杂问题	工程项目	能力综合	Agent

给大模型布置任务 – Prompt Engineering

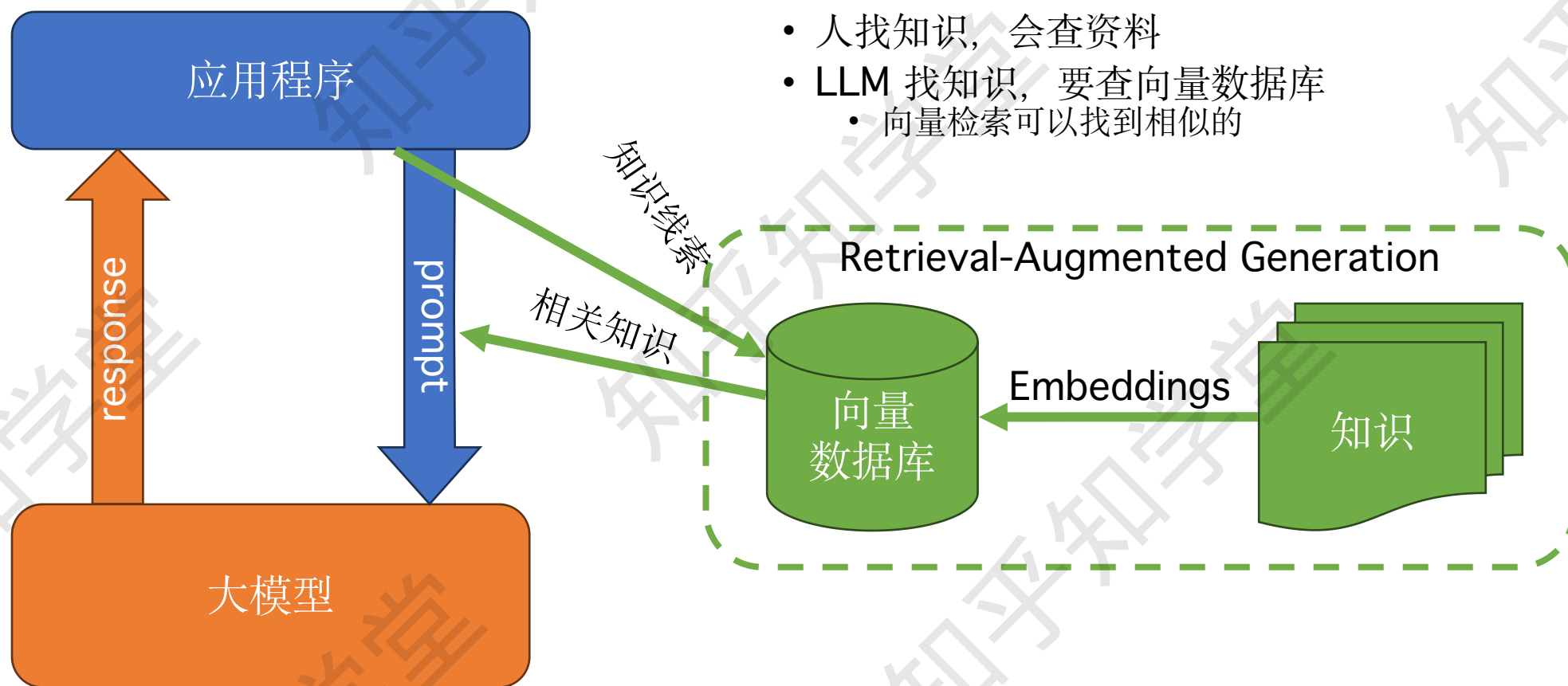
适用场景：知识问答，情报分析，写作，编程，文本加工处理.....



- 人 and 人对话，程序和 LLM 对话，都要
 - 指令**具体**
 - 信息**丰富**
 - 尽量**少歧义**

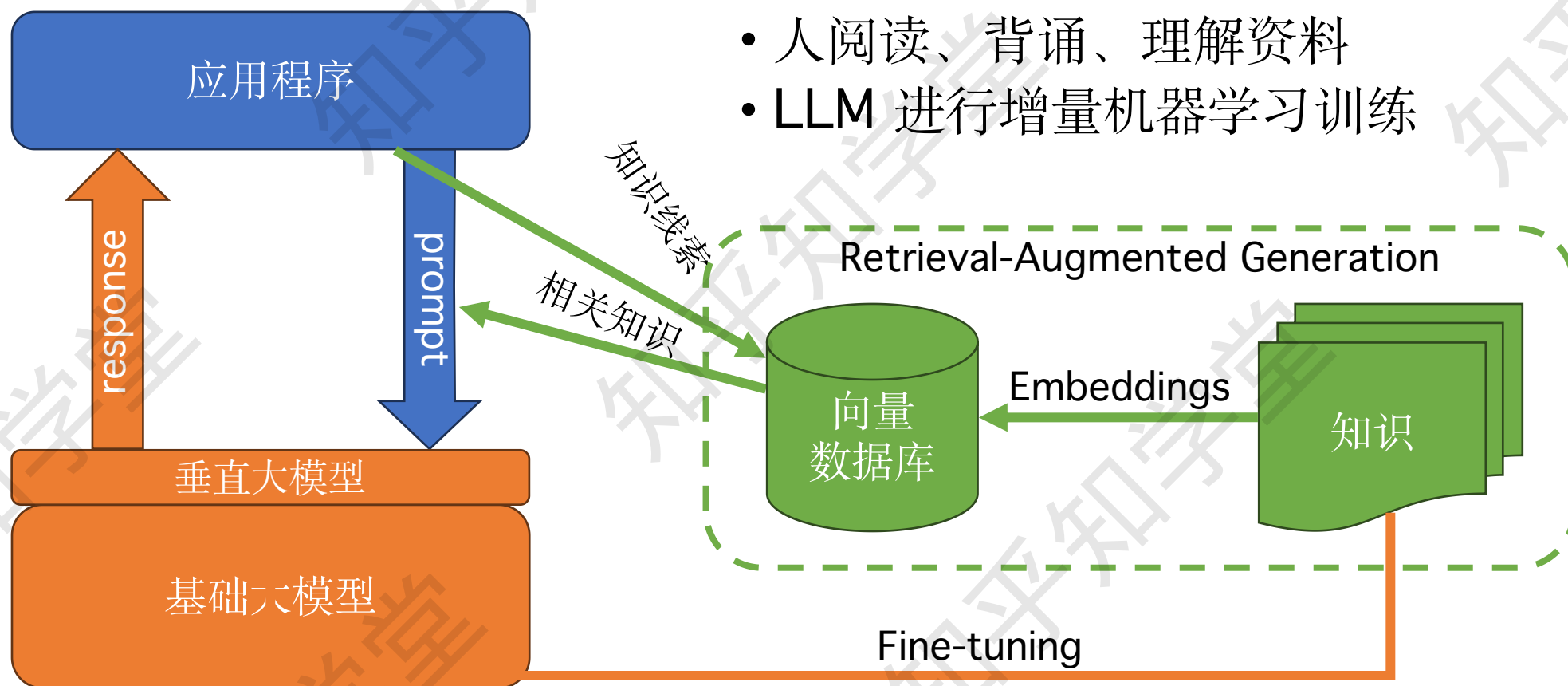
给大模型新知识 – RAG

适用场景：智能知识库，智能诊断，数字分身，带例子的 Prompt Eng.(Few-shot).....



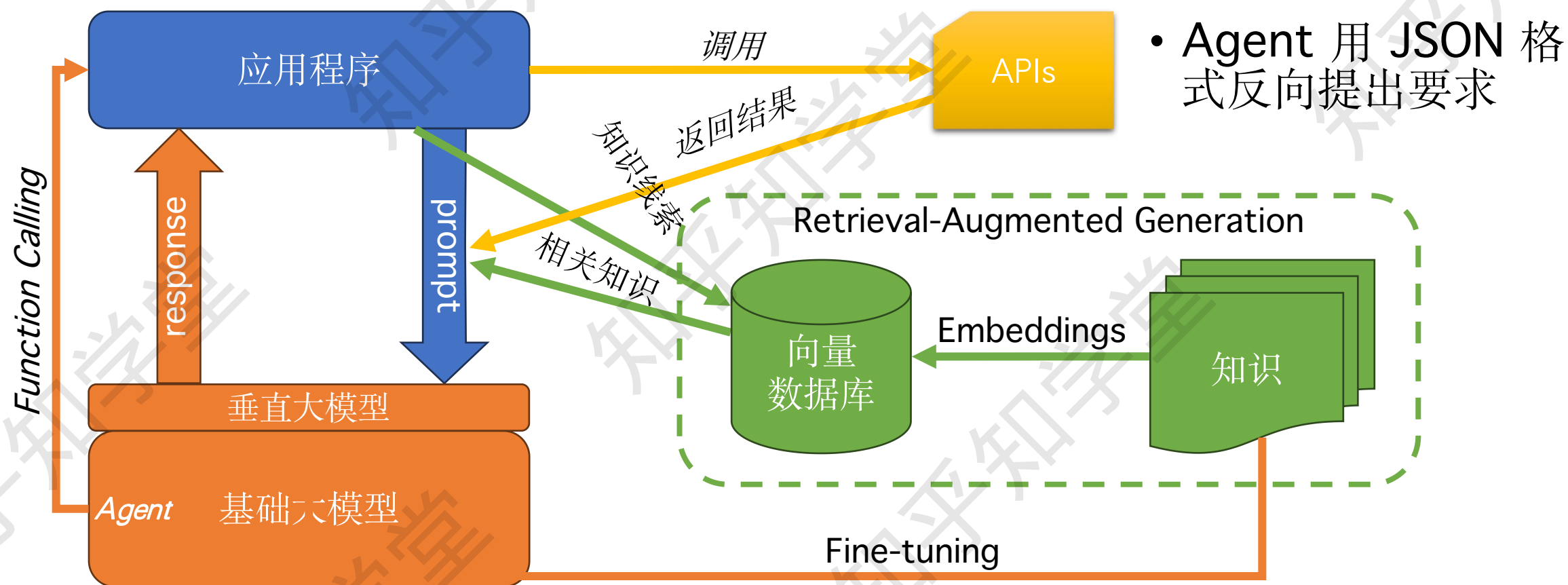
让大模型深度理解知识 – Fine-tuning

适用场景：智能知识库，智能诊断，数字分身……



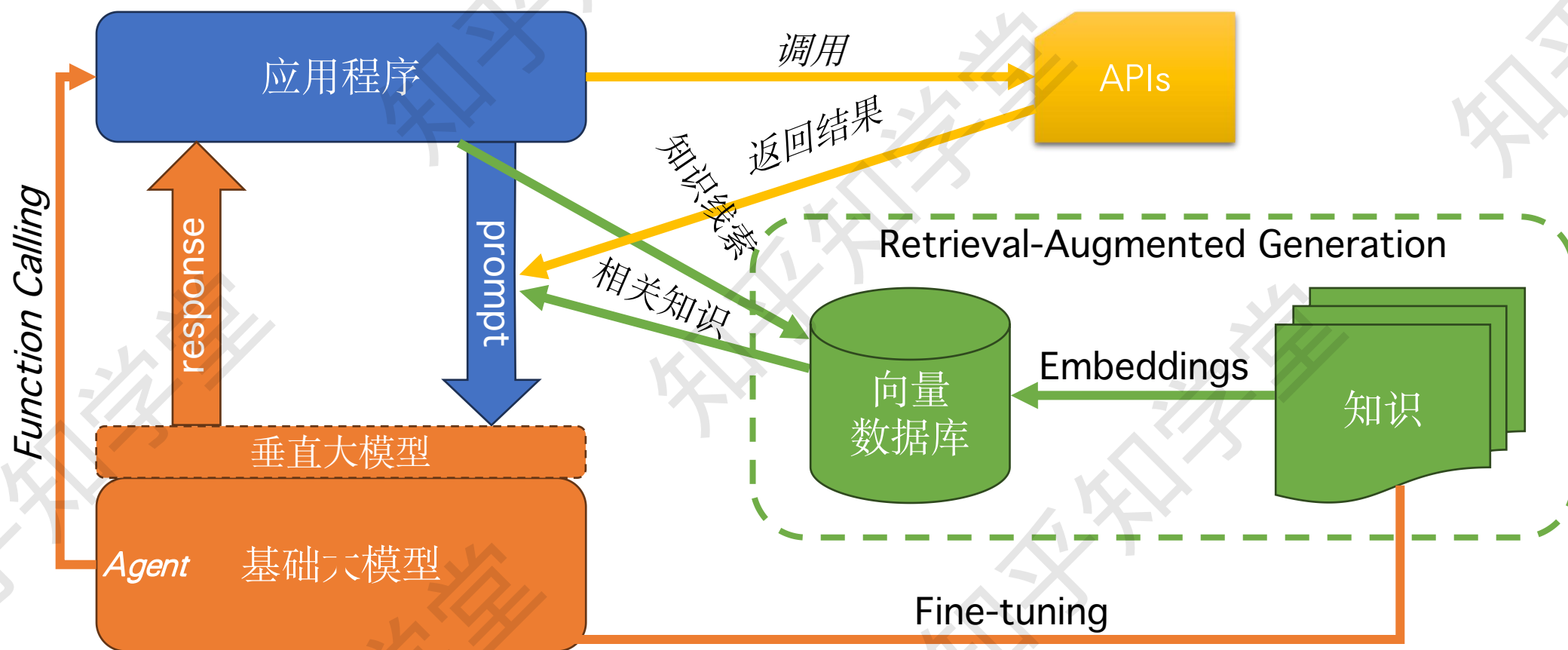
让大模型对接外界 – Function Calling

适用场景：智能助手，下一代搜索引擎，机器人，Agent.....



大模型应用的技术架构

ChatGPT、Copilot、GPTs、LangChain、MetaGPT 等都是对此架构的封装

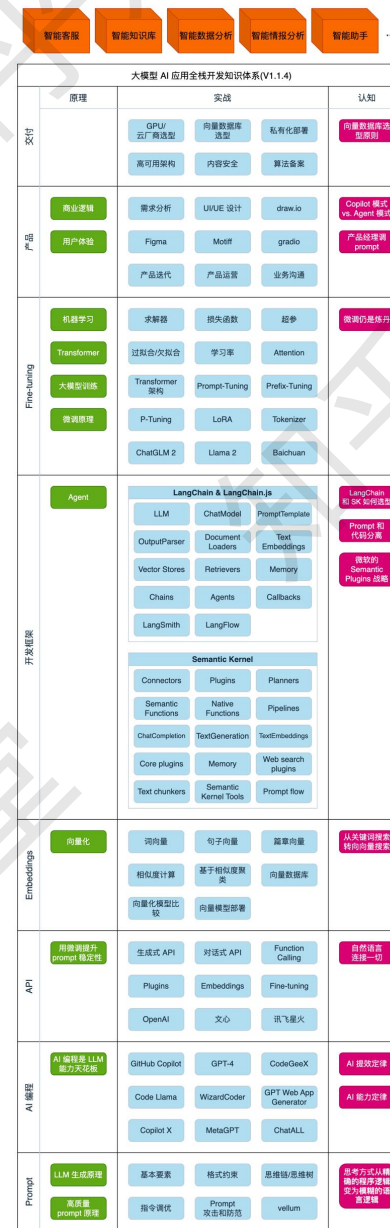


今日内容

大模型带来的行业变化

大模型应用的技术架构

大模型应用的知识体系



今日内容


大模型应用的技术架构

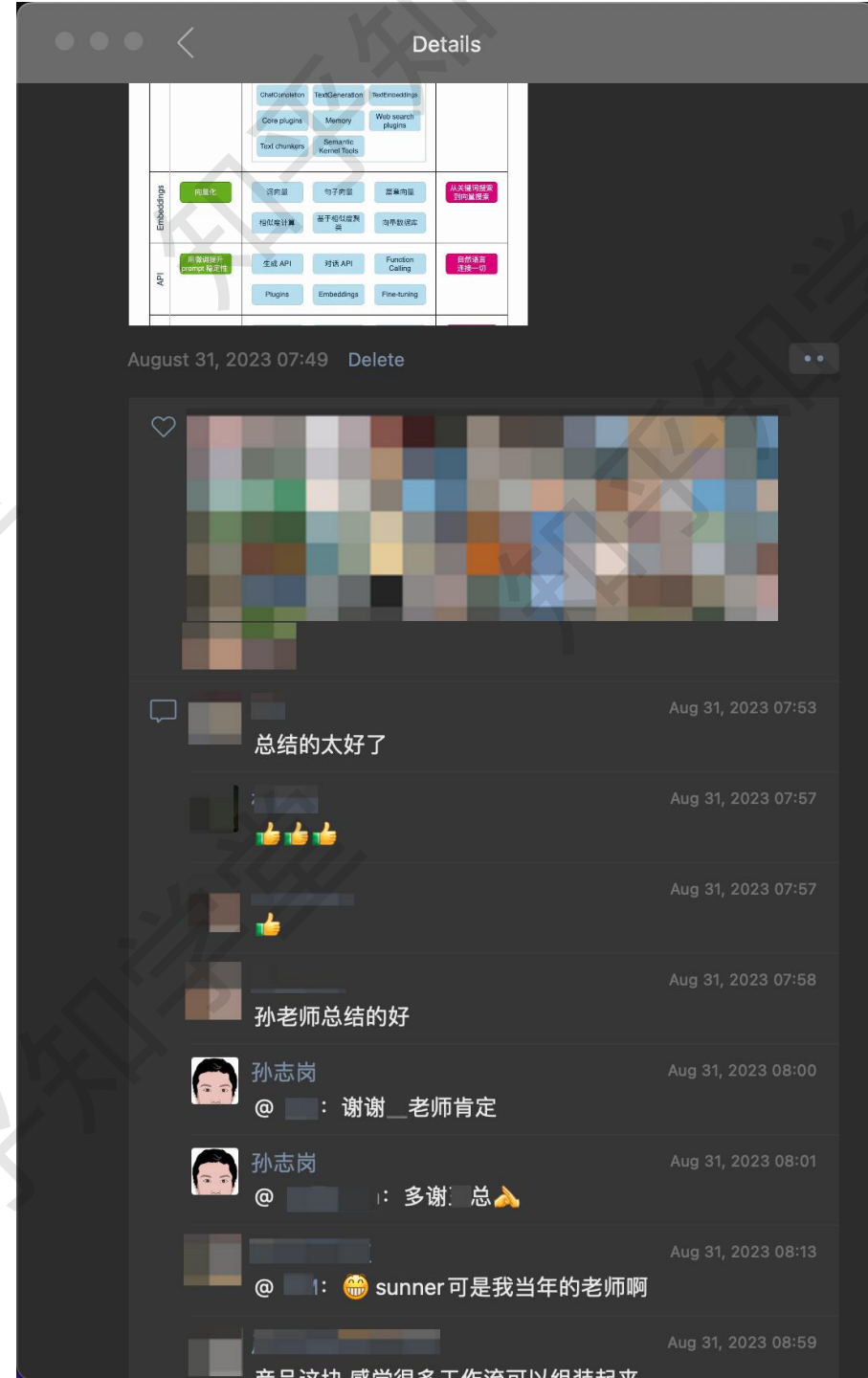
大模型应用的知识体系

大模型 AI 应用全栈开发知识体系(V1.1.4)			
交付	原理	实战	认知
		GPU/云厂商选型 向量数据库 私有化部署 高可用架构 内容安全 算法备案	向量数据库选型 Copilot 模式 vs Agent 模式
产品	商业逻辑	需求分析 UI/UE 设计 draw.io	Copilot 模式 vs Agent 模式
	用户体验	Figma Miro gradio 产品迭代 产品运营 业务沟通	产品特性与 prompt
Fine-tuning	机器学习	求解器 损失函数 超参	微调仍是伪命题
	Transformer	过拟合欠拟合 学习率 Attention Transformer 架构 Prompt-Tuning Prefix-Tuning P-Tuning LoRA Tokenizer ChatGLM 2 Llama 2 Baichuan	
开发框架	Agent	LangChain & LangChain.js LLM ChatModel PromptTemplate OutputParser Document Loaders Text Embeddings Vector Stores Retrievers Memory Chains Agents Callbacks LangSmith LangFlow	LangChain 和 GPT 选型 Prompt 和 代码生成 独特的 Semantic Plugins 功能
	Embeddings	Semantic Kernel Connectors Plugins Planners Semantic Functions Native Functions Pipelines ChatCompletion TextGeneration TextEmbeddings Core plugins Memory Web search plugins Text chunkers Semantic Kernel Tools Prompt flow	
API	向量化	词向量 句子向量 段落向量 相似度计算 基于相似度聚类 向量数据库 向量化模型比较 向量模型部署	从关键词检索转向向量检索
	用微调提升 prompt 稳定性	生成式 API 对话式 API Function Calling Plugins Embeddings Fine-tuning OpenAI 文心 讯飞星火	自然语言连接一切
AI 编程	AI 编程 LLM 能力天花板	GitHub Copilot GPT-4 CodeGeex Code Llama WizardCoder GPT Web App Generator Copilot X MetaGPT ChatALL	AI 编程定律 AI 能力定律
	LLM 生成原理	基本要素 格式约束 思维链/思维树 指令微调 Prompt 攻击和防范 vLLM	思考方式从精确到模糊 受大模型的思维逻辑

大模型应用的知识体系



智能客服					智能知识库					智能数据分析					智能情报分析					智能助手					...				
大模型 AI 应用全栈开发知识体系(V1.4.3)																													
交付	原理					实战										认知													
	GPU/云厂商选型					向量数据库选型					私有化部署					向量数据库选型原则													
	推理性能优化					高可用架构					内容安全																		
	算法备案																												
产品	商业逻辑					需求分析		KANO 模型			Galileo AI					AI 产品设计原则													
	用户体验					v0		真高设计 MasterGo			draw.io					产品经理 prompt													
						Figma		Motif			GPTs					产品经理 实现 demo													
						扣子 Coze		Gradio			产品迭代																		
多模态	特征对齐					多模态大语言模型										传统 CV 模型 仍存价值													
						Vision Transformer		CLIP			Q-Former																		
						GPT-4V		Gemini			LLaVA																		
						X-LLM		NEXT-GPT			LLaVA-Plus																		
Fine-tuning	机器学习					求解器		损失函数			超参					微调仍是趋势													
	大模型训练					过拟合/欠拟合		学习率			Attention																		
	微调原理					Transformer		RWKV			Mamba																		
						Prompt-Tuning		Prefix-Tuning			P-Tuning																		
开发框架和工具链	Agent					LangChain & LangChain.js										多框架组合开发													
						LCEL		LLM			ChatModel					Prompt 和 代码分离													
						PromptTemplate		OutputParser			Document Loaders																		
						Vector Stores		Retrievers			Text Embeddings																		
RAG	Embeddings					Chains		Memory			Agents					AGIClass.ai 制作													
						Tools		LangGraph			LangSmith					扫码获取最新版													
						LangFlow		LangServe																					
						LlamaIndex		MetaGPT			XAgent																		
	Embeddings					词/句子/篇章 向量		文档拆分			表格处理					从关键词检索 转向 向量检索													
						相似度计算		Embeddings 模型			OpenAI Embeddings																		
						BAAI BGE Embedding		向量数据库			向量检索																		
						混合检索		Elasticsearch			Chroma																		
						FAISS		Weaviate			Reranker																		
						表格处理																							
	用词保通					生成式 API		对话式 API			Assistants API					自然语言 理解一切													



大模型应用的知识体系



大模型 AI 应用全栈开发知识体系(V1.4.3)			
交付	原理	实战	认知
		<div>GPU/ 云厂商选型</div> <div>推理性能优化</div> <div>算法库</div> <div>向量数据库 选型</div> <div>高可用架构</div> <div>私有化部署</div> <div>内容安全</div>	<div>向量数据库 选型原则</div>

AI 应用		<div data-bbox="733 25 1082 168">WizardCoder</div> <div data-bbox="1149 25 1498 168">MetaGPT</div> <div data-bbox="1567 25 1916 168">GPT Engineer</div> <div data-bbox="733 229 1082 372">MAGE: GPT Web App Generator</div>	
Prompt	<div data-bbox="242 515 591 658">LLM 生成原理</div> <div data-bbox="242 725 591 868">高质量 prompt 原理</div>	<div data-bbox="733 515 1082 658">基本要素</div> <div data-bbox="1149 515 1498 658">格式约束</div> <div data-bbox="1567 515 1916 658">风格控制</div> <div data-bbox="733 725 1082 868">思维链</div> <div data-bbox="1149 725 1498 868">自洽性</div> <div data-bbox="1567 725 1916 868">思维树</div> <div data-bbox="733 935 1082 1078">指令调优</div> <div data-bbox="1149 935 1498 1078">Prompt 攻击和防范</div> <div data-bbox="1567 935 1916 1078">vellum</div> <div data-bbox="733 1145 1082 1288">GPTs</div> <div data-bbox="1149 1145 1498 1288">Coze</div>	<div data-bbox="2058 515 2407 658">把 AI 当人看</div>

		<div>百度文心</div> <div>讯飞星火</div> <div>MiniMax abab</div>	
AI 编程	<div>AI 编程是 LLM 能力天花板</div>	<div>GitHub Copilot</div> <div>ChatGPT Plus</div> <div>CodeGeeX</div> <div>通义灵码</div> <div>Tabby</div> <div>Code Llama</div> <div>WizardCoder</div> <div>MetaGPT</div> <div>GPT Engineer</div> <div>MAGE: GPT Web App Generator</div>	<div>AI 提效定律</div> <div>AI 能力定律</div>
	<div>LLM 生成原理</div>	<div>基本要素</div> <div>格式约束</div> <div>风格控制</div>	<div>把 AI 当人看</div>

		<div>FAISS</div> <div>Weaviate</div> <div>Reranker</div> <div>表格处理</div>	
API	<div>用微调提升 prompt 稳定性</div>	<div>生成式 API</div> <div>对话式 API</div> <div>Assistants API</div> <div>Function Calling</div> <div>Actions</div> <div>Embeddings</div> <div>Fine-tuning</div> <div>Moderation API</div> <div>OpenAI</div> <div>百度文心</div> <div>讯飞星火</div> <div>MiniMax abab</div>	<div>自然语言 连接一切</div>
	<div>AI 编程是 LLM 的专属领域</div>	<div>GitHub Copilot</div> <div>ChatGPT Plus</div> <div>CodeGeeX</div>	<div>AI 提效定律</div>

Embeddings

词/句子/篇章
向量

文档拆分

表格处理

从关键词搜索
转向向量搜索

相似度计算

Embeddings
模型OpenAI
EmbeddingsBAAI BGE
Embedding

向量数据库

向量检索

混合检索

Elasticsearch

Chroma

FAISS

Weaviate

Reranker

表格处理

Agent

LangChain & LangChain.js

LCEL

LLM

ChatModel

PromptTemplate

OutputParser

Document
Loaders

Vector Stores

Retrievers

Text
Embeddings

Chains

Memory

Agents

Tools

LangGraph

LangSmith

LangFlow

LangServe

LlamaIndex

MetaGPT

XAgent

多框架组合开发

Prompt 和
代码分离

AGIClass.ai
制作

扫码获取最新版



机器学习

大模型训练

微调原理

求解器

损失函数

超参

过拟合/欠拟合

学习率

Attention

Transformer

RWKV

Mamba

Prompt-Tuning

Prefix-Tuning

P-Tuning

LoRA

QLoRA 量化

Tokenizer

ChatGLM 3

Llama 3

微调仍是炼丹

特征对齐

多模态大语言模型

Vision
Transformer

CLIP

Q-Former

GPT-4V

Gemini

LLaVA

X-LLM

NExT-GPT

LLaVA-Plus

MM-ReAct

图像生成模型

Diffusion
ModelStable
Diffusion

Midjourney

DALL·E

LoRA

ControlNet

传统 CV 模型
仍有价值

		算法备案			
产品	商业逻辑	需求分析	KANO 模型	Galileo AI	AI 产品设计原则
	用户体验	v0	莫高设计 MasterGo	draw.io	产品经理调 prompt
		Figma	Motiff	GPTs	产品经理 实现 demo
		扣子 Coze	Gradio	产品迭代	
		产品运营	与人沟通	Shape of AI	
	特征对齐	多模态大语言模型			传统 CV 模型 仍有价值



大模型 AI 应用全栈开发知识体系(V1.4.3)			
交付	原理	实战	认知
		<div>GPU/ 云厂商选型</div> <div>推理性能优化</div> <div>算法备案</div> <div>向量数据库 选型</div> <div>高可用架构</div> <div>私有化部署</div> <div>内容安全</div>	<div>向量数据库 选型原则</div>