



---

BEIJING

# 生成式AI在网络安全领域的应用

学习、构建、应用

## 个人简介



**王志刚**

**致力于技术加速AI成为新质生产力**

michael7736(微信、Twitter、Linkedin、Github、HuggingFace)

**网络安全专家、架构师、AI与数据科学家、生成式AI通用人工智能技术探索先趋**

**个人使命：推动AI助力网络安全**

### AI：大模型训练、AI推理应用培训、建设、和咨询服务

- 数据科学：数据平台、数据科学、数据安全与治理
- 应用和基础架构：云平台、云原生、DevOPS、微服务、分布式
- 网络安全：评估、防御体系建设、密码学

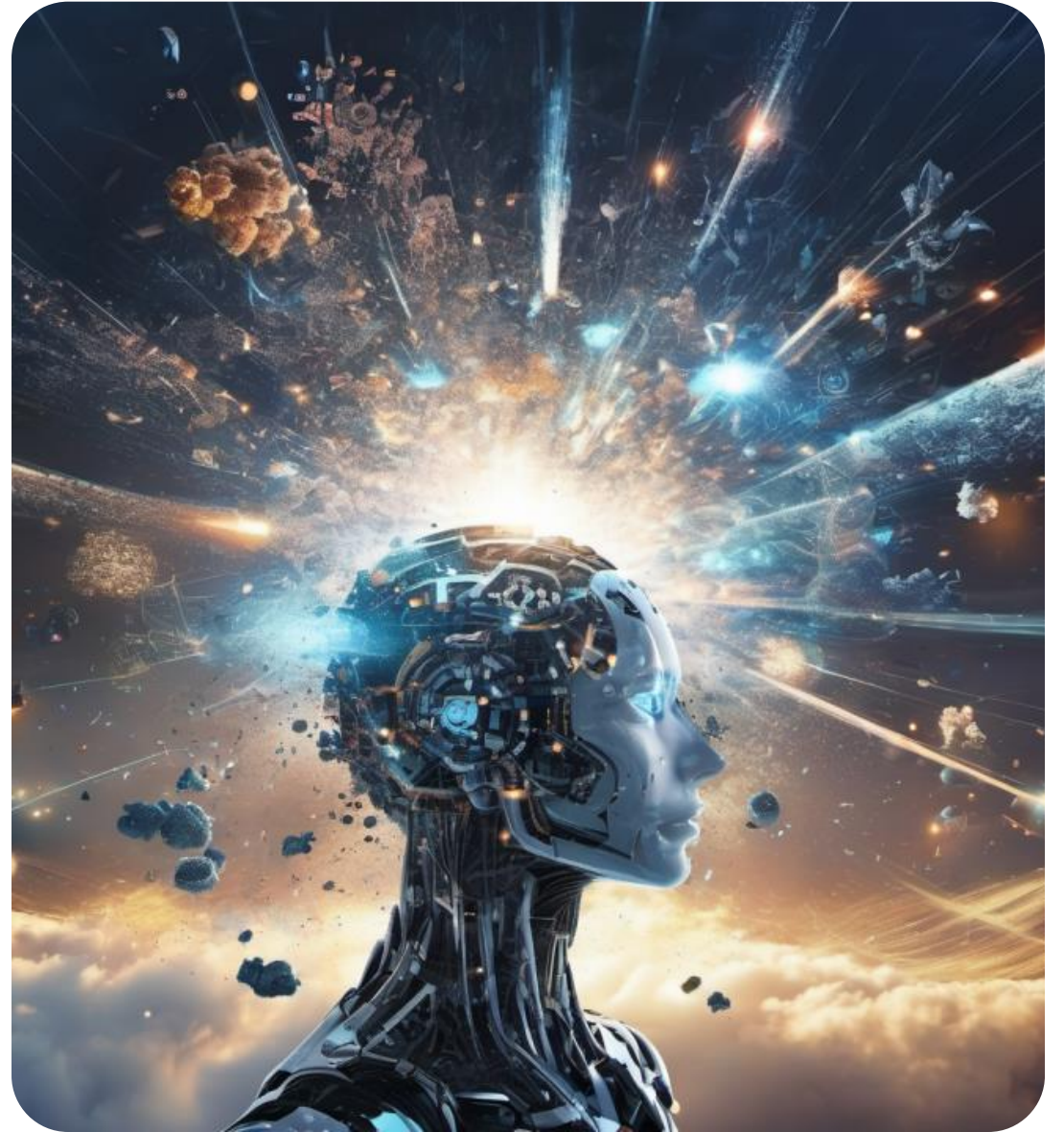
### 安全大模型、多模态嵌入、LLMOPS、大语言模型Zero2Hero

- 2008年奥运会熟悉安全专家
- 百度、亚马逊、华为、阿里云安全架构师，美团首席安全架构师，端到端筹建安全防御体系、加密体系和数据安全

# 目录

- AI大爆炸-大模型发展概览
- AI+安全：大模型重构安全业务
- How AI：安全企业如何构建自己的AI系统

# AI大爆炸 - 大模型发展概览



# 回答几个问题

- AI泡沫有多大?
- AI发展的现实情况?
- 如何获得AI的客观认知?
- 从商业视角看待AI



## AI 第二次技术革命

- 人工智能的发展与微处理器、个人电脑、互联网和手机的发明一样重要。它将改变人们工作、学习、旅行、就医和相互交流的方式。**整个行业都将围绕人工智能重新定位。企业将通过对人工智能的运用来脱颖而出。**
- 让尽可能多的人受益
- 最后，我们应该记住，**人工智能的成就才刚刚开始。**它目前所具有的任何局限性，在我们尚未意识到之前都将消失



# AI从产、学、研究领域都得到迅猛发展

## 产业界

- 全世界最顶尖的科技公司All-In AI，算力、风头AI成为当下最火热的投资
- 社区发展迅猛，生态繁荣
- 此外AI风潮正席卷全部行业

## 学术界

- AI成为创新最密集领域
- 全世界最顶尖的科学家、学术机构  
Stanford、MIT、CMU成为AI科研应用先趋



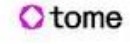


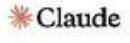


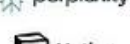
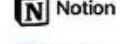
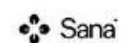
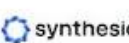

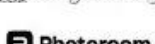
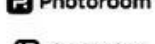

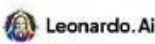



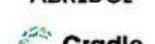
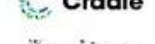
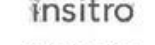

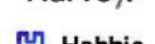

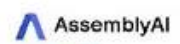



## 科学研究

- AI正在生物学Alpha Fold，材料学、分子物理学、基因学和医学领域产生创新性突破，并蔓延到艺术创作领域
























[2024 AI 技术报告](#)

[State of AI 2023人工智能现状报告](#)

## Apps

CONSUMER USES		ENTERPRISE STACK			INDUSTRY VERTICALS				
ENTERTAINMENT	<b>character.ai</b>  Pika	GENERAL PRODUCTIVITY	<b>ChatGPT</b> ●  <b>glean</b> 	<b>ADEPT</b>   <b>WRITER</b>	CREATIVE	DEFENSE	HEALTH & BIO	INDUSTRIAL	PROFESSIONAL SERVICES
PRODUCTIVITY	<b>ChatGPT</b> ●     	LEARNING & DEVELOPMENT	 		    Pika IIElevenLabs 	 	   	<b>FIGURE</b> <b>TRACTION</b> 	<b>Harvey.</b> 
		CUSTOMER EXPERIENCE	<b>CRESTA</b>						
		DEVELOPER & DATA TEAMS	 	 					

## Infrastructure

INFERENCE PROVIDERS		APP DEV FRAMEWORKS	MODEL HUBS	FOUNDATION MODEL PROVIDERS	
 anyscale	 databricks	 LangChain	 Hugging Face	<b>ANTHROPIC</b>	 cohere
together.ai	 baseten		 Replicate	 MISTRAL AI	 OpenAI
STORE & COMPUTE				HARDWARE	
LABEL / PROCESS DATA		CLOUD DATA PROVIDERS		CLOUD SERVICE PROVIDERS	
 Cleanlab	<b>scale</b>	 databricks	 Pinecone	 Weaviate	 Google Cloud
	 UNSTRUCTURED	 MongoDB	 snowflake	 aws	 Azure
					 cerebras
					 NVIDIA
					 AMD
					 intel
					(etc.)

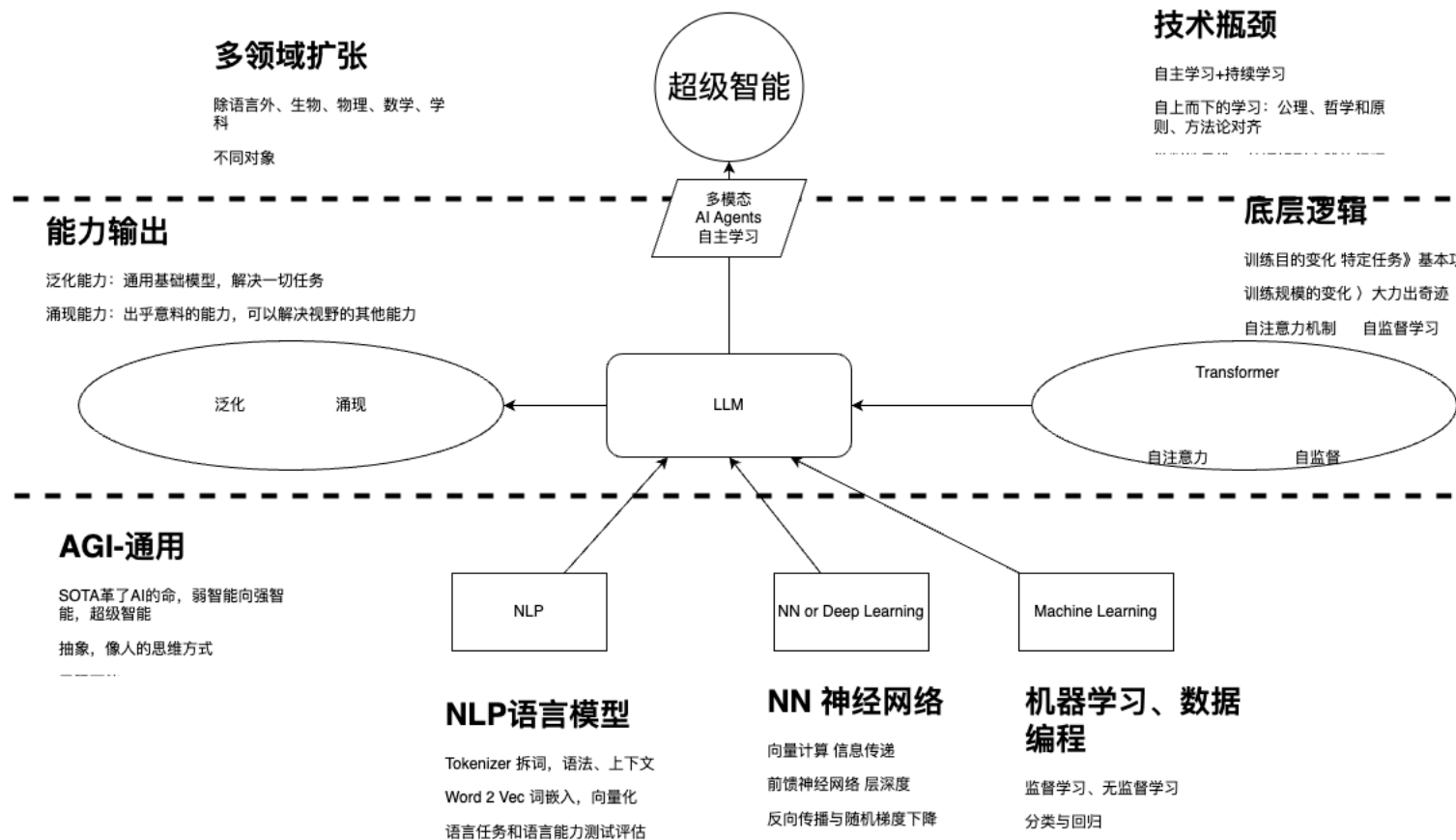


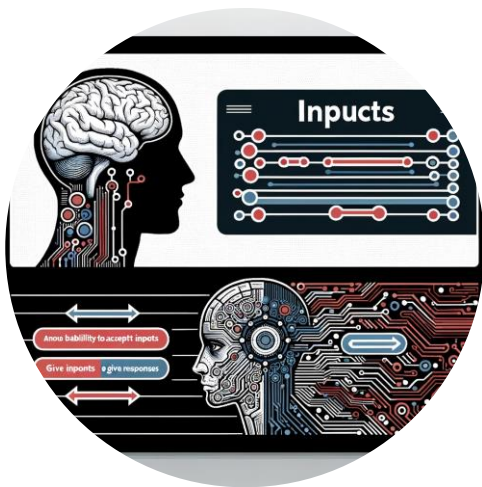
# AI将如何颠覆我们的产业？



# 什么是大模型

# 大模型的前世今生





## Scaling Law

### 超大规模参数、超大规模数据集

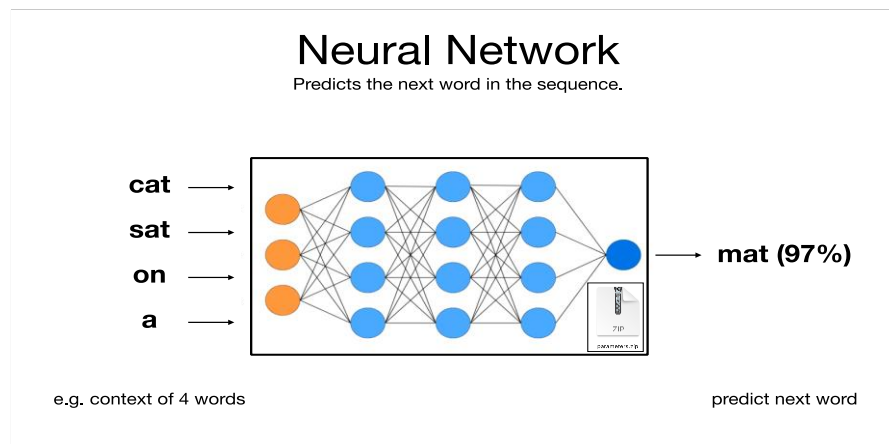
- 超大规模参数深层神经网络，可以理解事物一切特征、细节
- 超大规模的数据：构造出与其他事物的千丝万缕的联系，探究出足够抽象的理论和顶层逻辑

## 泛化与涌现

- 在自监督、无特定目标的简单猜词游戏基础上，洞悉文字和其代表的实体之间的复杂关系。从而实现举一反三，通用能力
- 当数量积累到足够的量，层数足够深，产生涌现-灵光乍现的能力

## 范式转换

- GPT前：小模型或专有模型：特定任务，低复用，低扩展，重开发
- GPT后：大模型：通用模型，高复用，高扩展，轻开发，



# 生成式AI具体能干什么？

## LLM的能力

- 具有语言的能力
- 具有世界知识
- 具有专家经验
- 具有推理能力
- 具有创造能力
- 强大的信息处理能力

# 安全风险三要素

## 业务

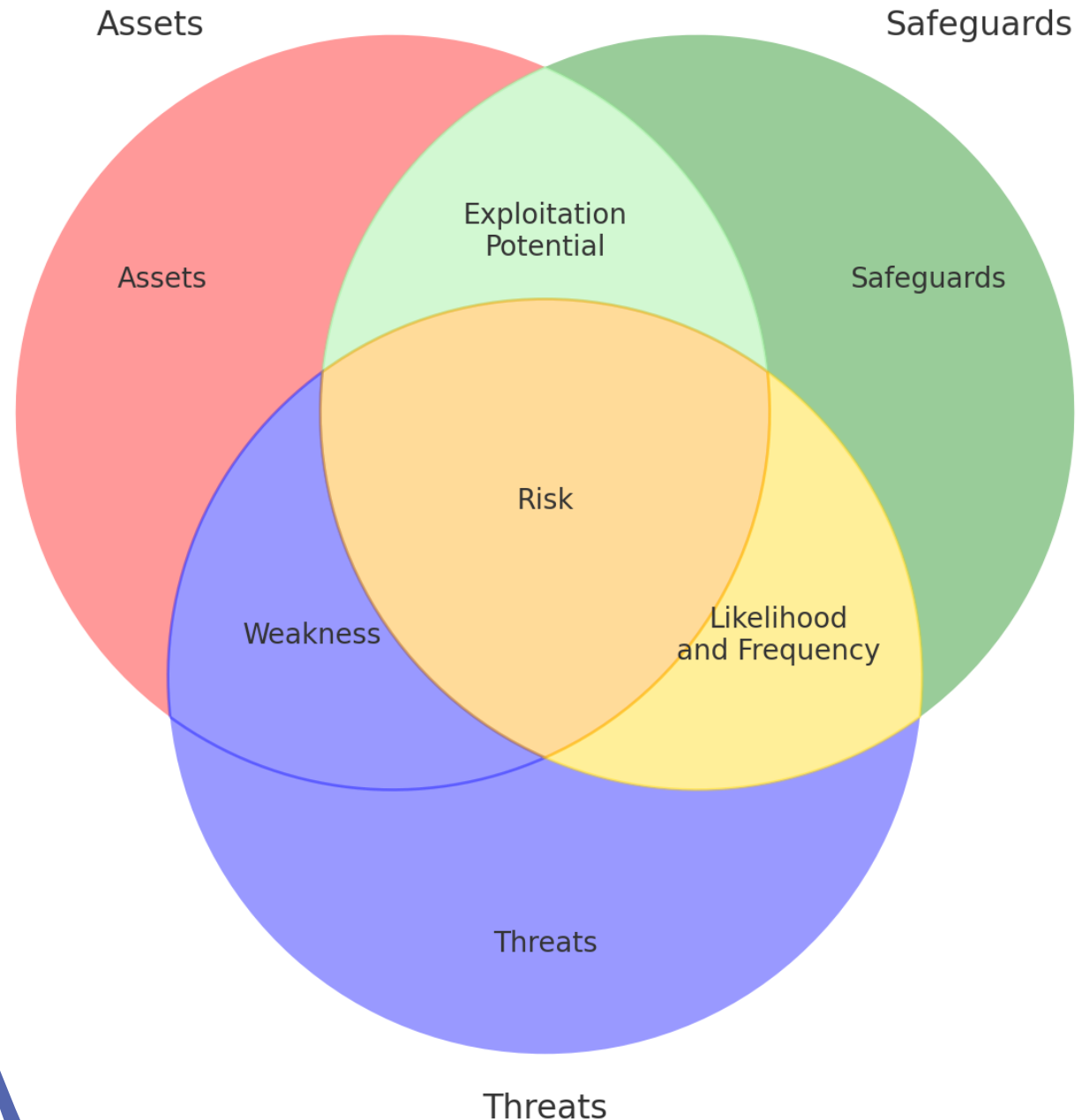
- 顶层设计：业务核心痛点与诉求。隐私合规？竞争对手？资产损失？信用？（业务review）
- 中层：业务数据流转与应用使用逻辑（架构review, case study）
- 底层：资产治理、识别盘点，细粒度资产盘点、IT基础设施和关系数据建设（tools + 人）

## 威胁

- 威胁建模与攻击面分析（安全评估）
- Attacks 框架与威胁情报
- 威胁实体特征与情报库，入侵检测，画像分析（SOC + AI）

## 防护

- 纵深防御体系
- 防御原则
- 防御技术





# 安全业务

安全运维（依托Soc/sims和资产治理）

- 资产管理、配置管理、安全指标管理
- 漏洞、补丁和供应链管理
- 安全风险分析、趋势分析
- 安全事件应急响应、追查

对抗组

- 安全研究与漏洞挖掘
- 蓝军
- 情报

安全架构、解决方案


- 主机服务器安全
- 办公与终端安全
- 移动安全
- 应用和代码安全
- 服务安全
- 基础架构安全
- 数据安全
- 网络安全
- 账户安全
- 业务风控

## • 安全工具

- 漏洞管理
- 身份认证（人、设备、服务）
- 访问控制（ABAC）
- 密码箱Secret Manager
- KMS（密钥管理、加解密SDK/API）
- PKI数字证书（身份、签名）
- 去标识化
- 传输加密tls,https
- WAF、HIDS、NIDS
- 防火墙、Jumper、Proxy
- 端安全：DLP、杀软、EDR
- SIMS/SoC
- 数据安全治理

# 痛点和需求

- 覆盖率
- 准确性
- 实时性
- 成本控制
- 预测能力



亲者痛，仇者快  
能力提升：  
智能化、自动化、透明化

# 安全智能项目

## 编程大模型Coding:

- 代码审核、漏洞挖掘、供应链分析、访问关系识别、后门识别
- 数据血缘、API治理、安全合规检查

## 安全知识库与知识图谱（数据）

- 资产盘点、配置基线
- 威胁情报、攻击样本
- 攻击特征与画像

## 安全运维大模型（模型）

- 攻击检测：通过包括安全日志、网络流量日志、应用日志
- 风险识别：通过3要素的识别，行为、画像分析

## 攻击智能体（综合渗透测试智能体）

规划与反思：基于项目范围、目标和限制条件制定渗透策略，并构建整个项目编排

记忆：存储攻击

## 安全合规审计（分类模型）

敏感数据分类、分级和标注

代码审核、第三方软件审核

## 安全运维AI智能体（综合）：

攻击检测：从情报、安全知识库和知识图谱，识别出攻击、异常信号，并规划、启动攻击检测任务。确定特征和范围，调用工具启动信息确认和调查。必要时可调用风控或者远程命令实现具身验证、测试。保留记录以便事后追查。

事件溯源：依据相关的日志、调查结果数据，结合外部情报、人员信息进行事件调查，取证和溯源。

## 画像模型（分类与特征提取模型）

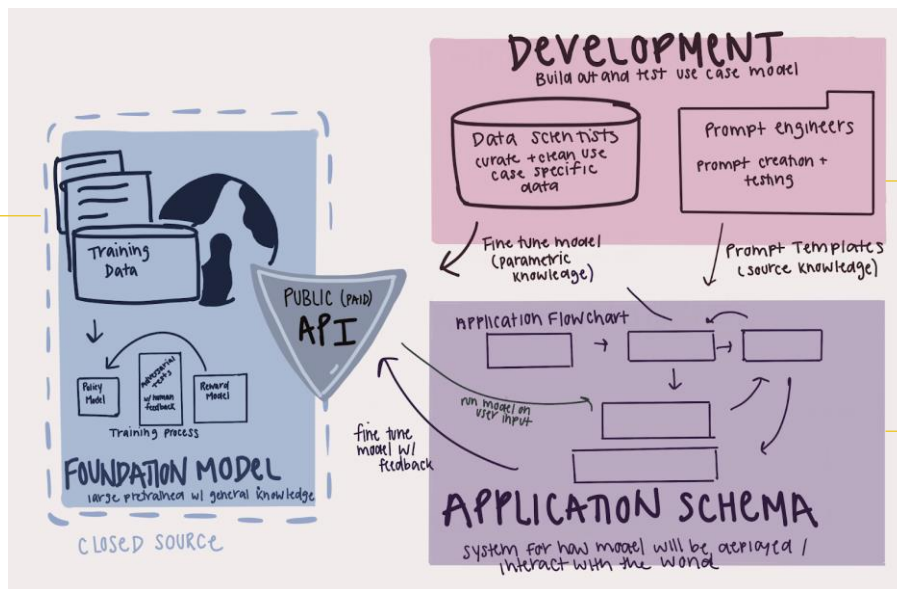
依据个体特征、群体特征、关联关系、行为特征构建出不同的攻击异常画像  
APT、病毒木马、挖矿、信息窃取、后门、黑、灰产特征

# 如何构建？

# 生成式AI生态

## 基础模型预训练

- 模型架构设计
- 大批量无标注数据
- 大集群、并行训练
- 监控、跟踪和评价



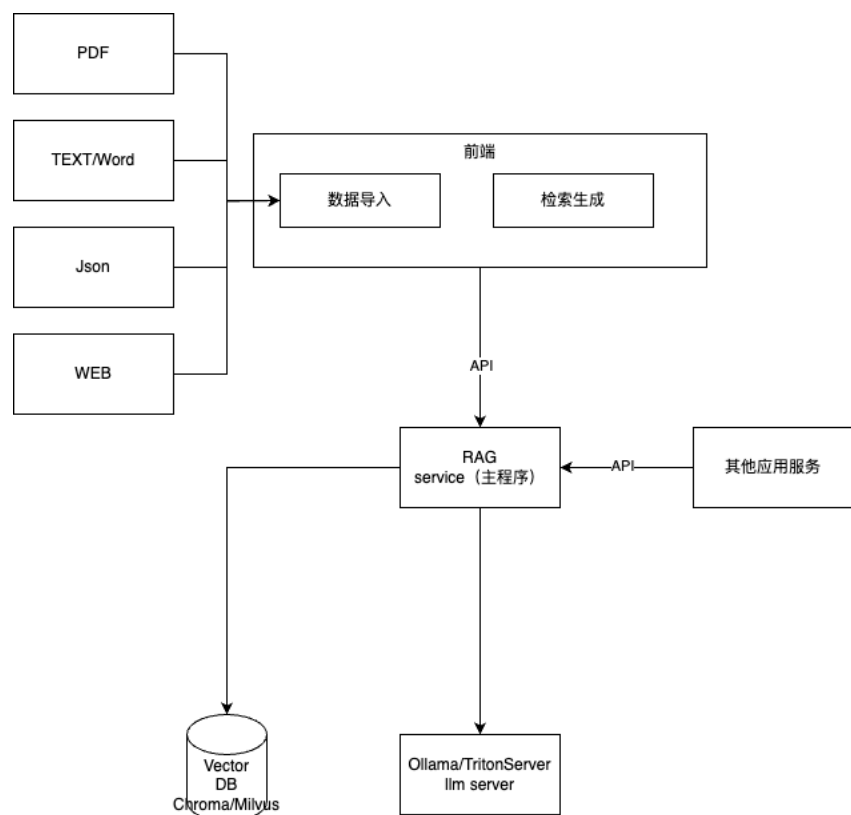
## 模型定制化开发

- 模型评估
- 小批量标注数据 ()
- 模型压缩、优化
- 模型对齐

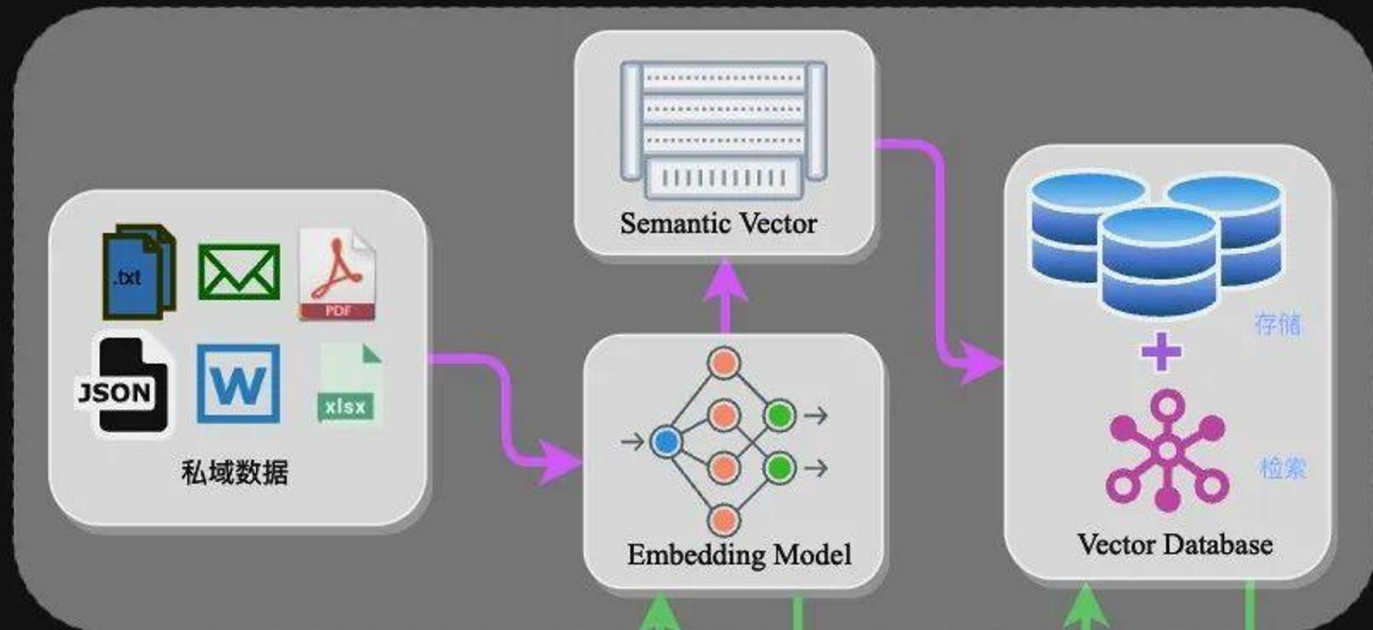
## 应用集成

- 针对企业私有数据、特定场景应用集成
- 应用开发
- 工程能力

# RAG



## 数据准备



## 数据检索

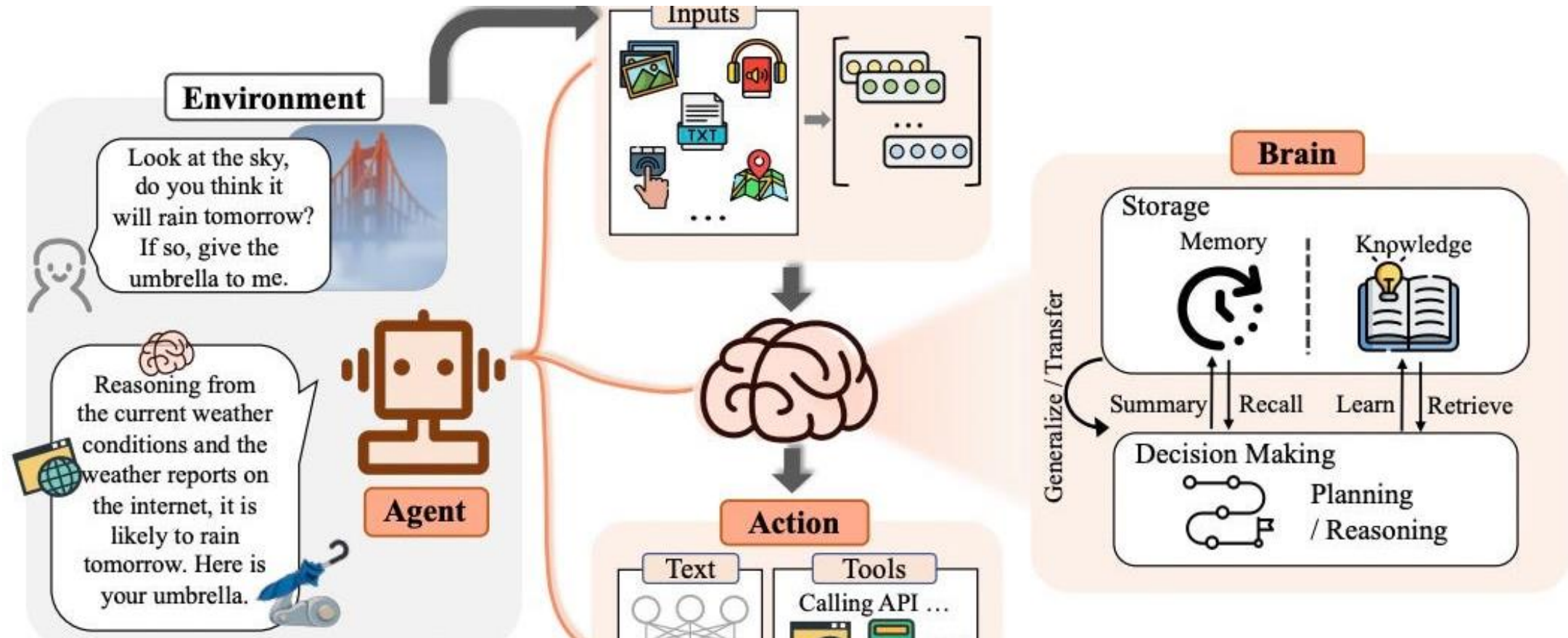


## LLM 生成



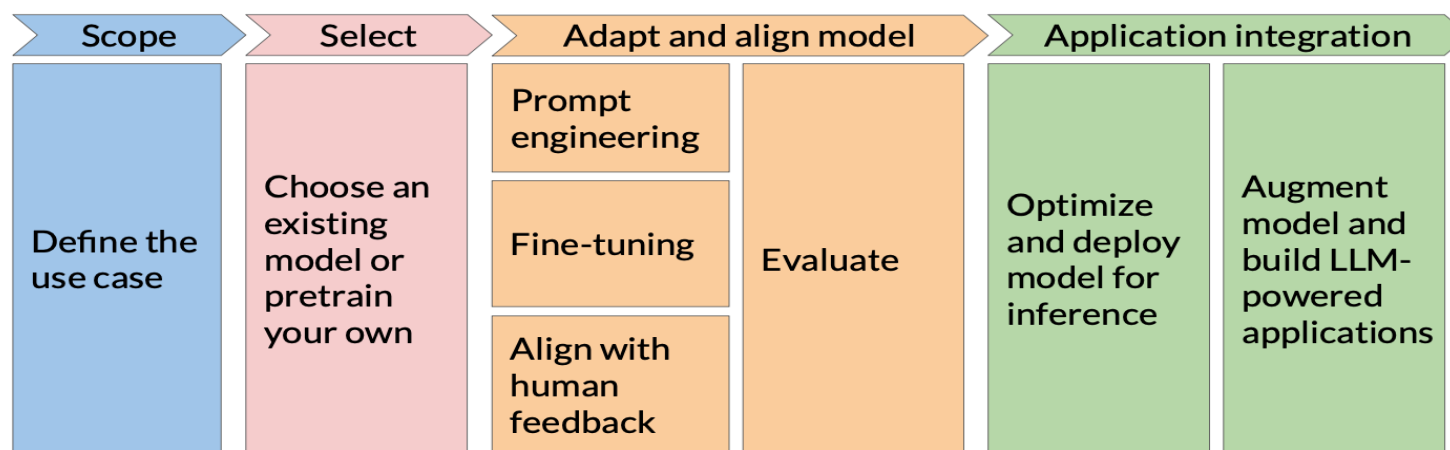


# AI Agents

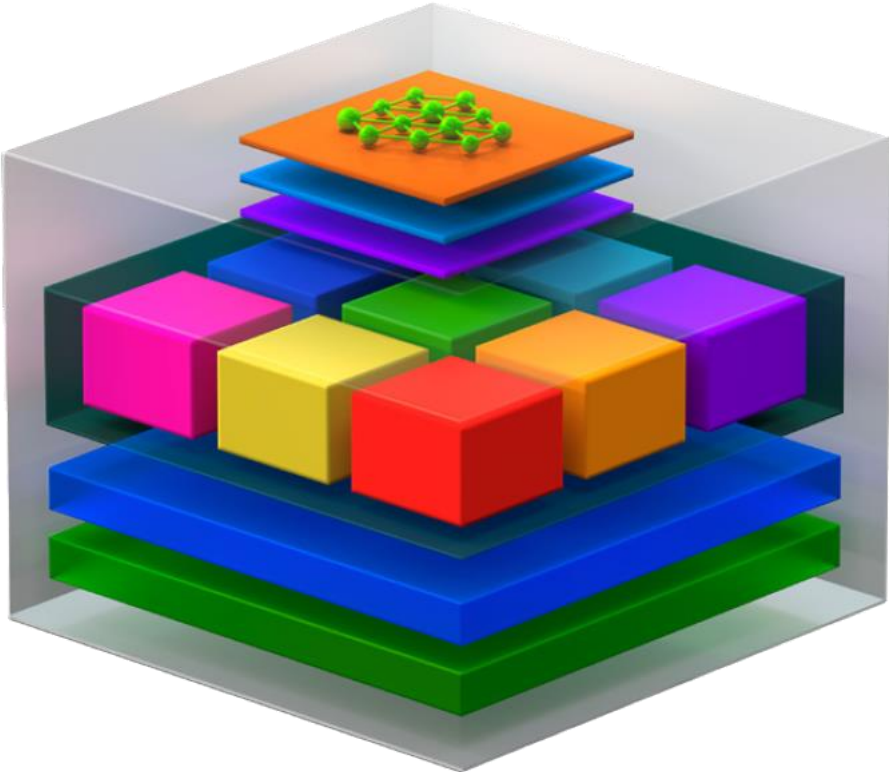
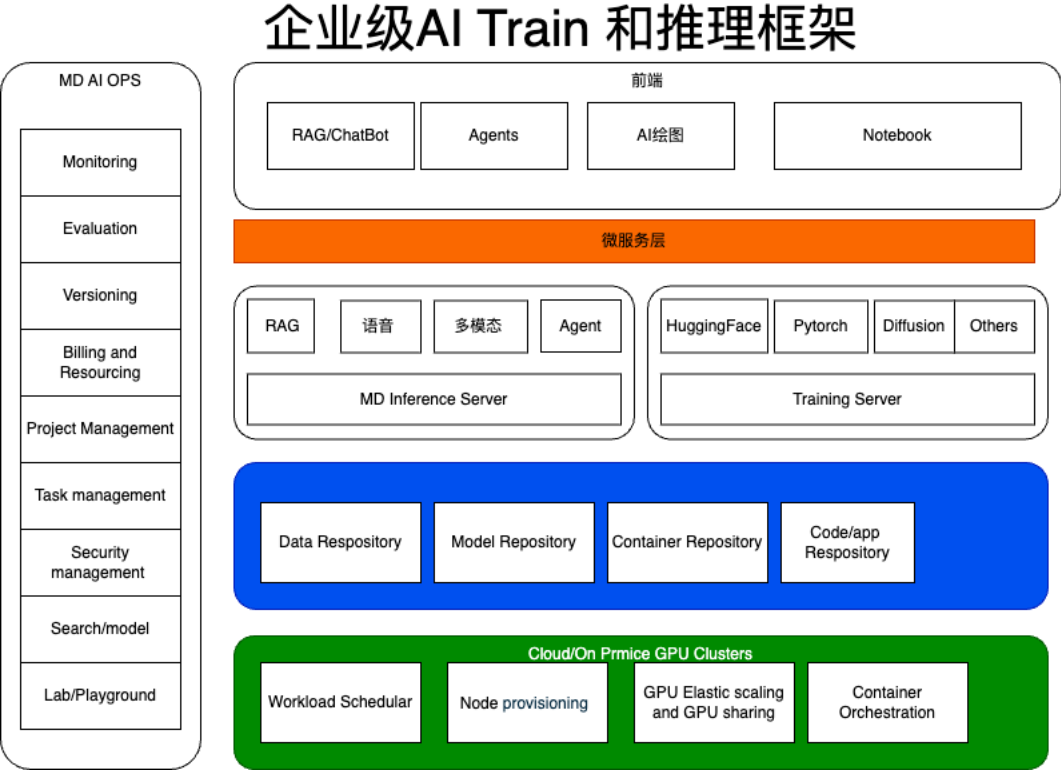


# AI项目框架

## Generative AI project lifecycle



# AI 基础框架-工具支撑



# 大模型技术栈-LLM101

原则：大模型知识海洋，学什么、按什么顺序学！**选择比努力更重要**

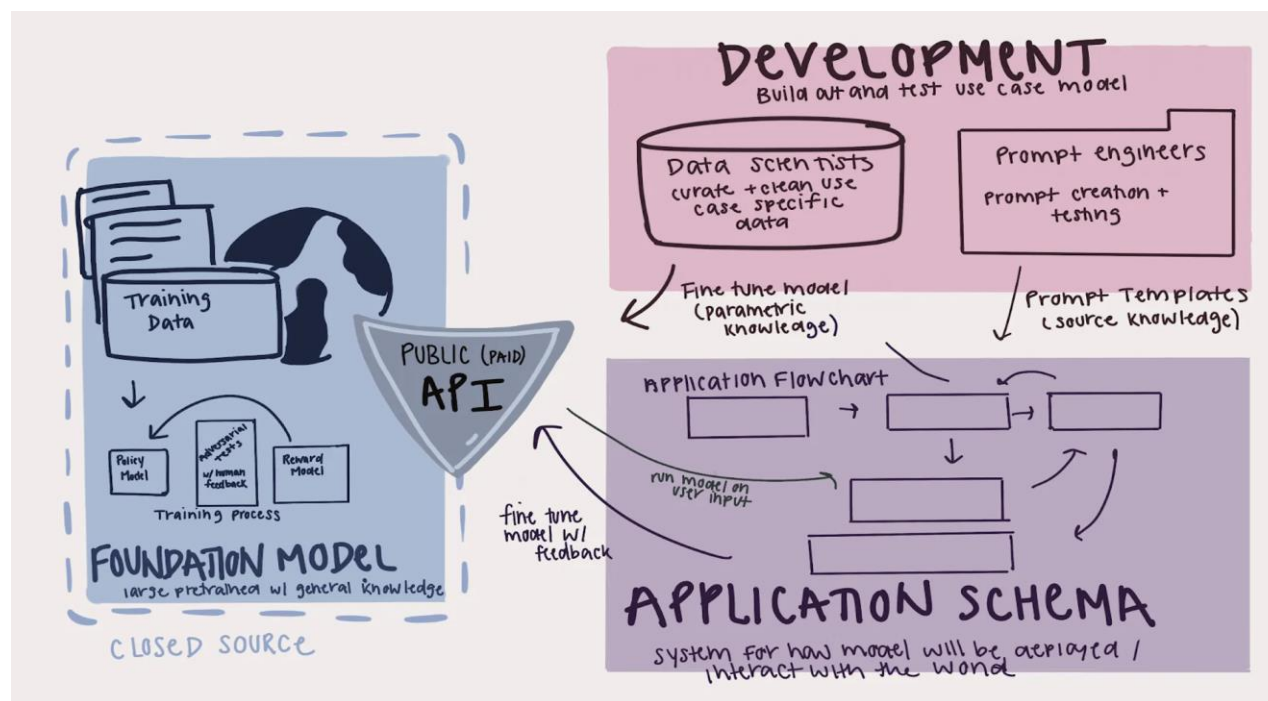
## 大模型学习全景：两条主线，一条辅助线

主线一：模型全生命周期技能

主线二：AI应用开发技术

辅助线：用来支撑大模型技术的技术

- ❑ 数据：数据科学、数据治理、数据安全
- ❑ 理论：NLP, ML, NN
- ❑ 工具：Python, Docker, Devops, Linux/Mac
- ❑ 工程：微服务和云原生, 分布式, llmops



<https://www.notion.so/LLM101-3d53f33e9f0d40eea8d678c6e8b72959>

## 参考链接

<https://zhuanlan.zhihu.com/p/4423339228>

<https://zhuanlan.zhihu.com/p/3555951416>

<https://zhuanlan.zhihu.com/p/712514706>

<https://zhuanlan.zhihu.com/p/702989158>

其他信息请关注后续课程

# Take Away

- AI新范式，让AI应用门槛降低，人人生产AI
- AI强大数据处理能力深度结合安全，大规模技术升级
- AI将对产业进行分化，头部玩家将最大获益
- AI学习是个技术活，既要聪明、也要努力



# Q& A



返町

北京 朝阳



扫一扫上面的二维码图案，加我为朋友。



BEIJING

Thank you!

michael7736