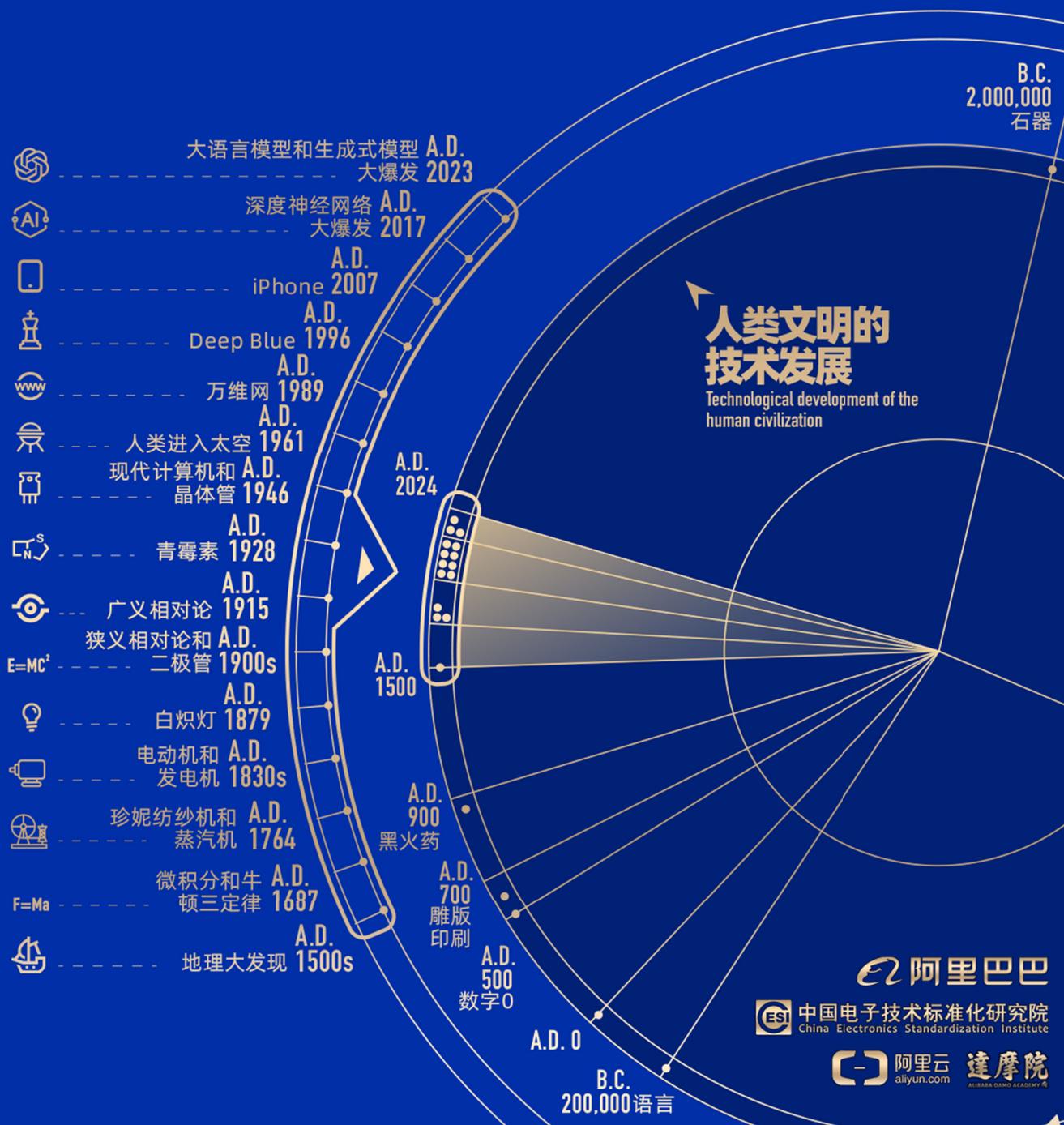


安全, 可靠, 可信赖, 可用  
Safety, Reliability,  
Trustworthiness, Usability

# 大模型技术发展及 治理实践报告

The Large Language Model Technology development and governance  
practice report

## 我们正在步入AI大模型新时代 我们已经迎接治理新挑战



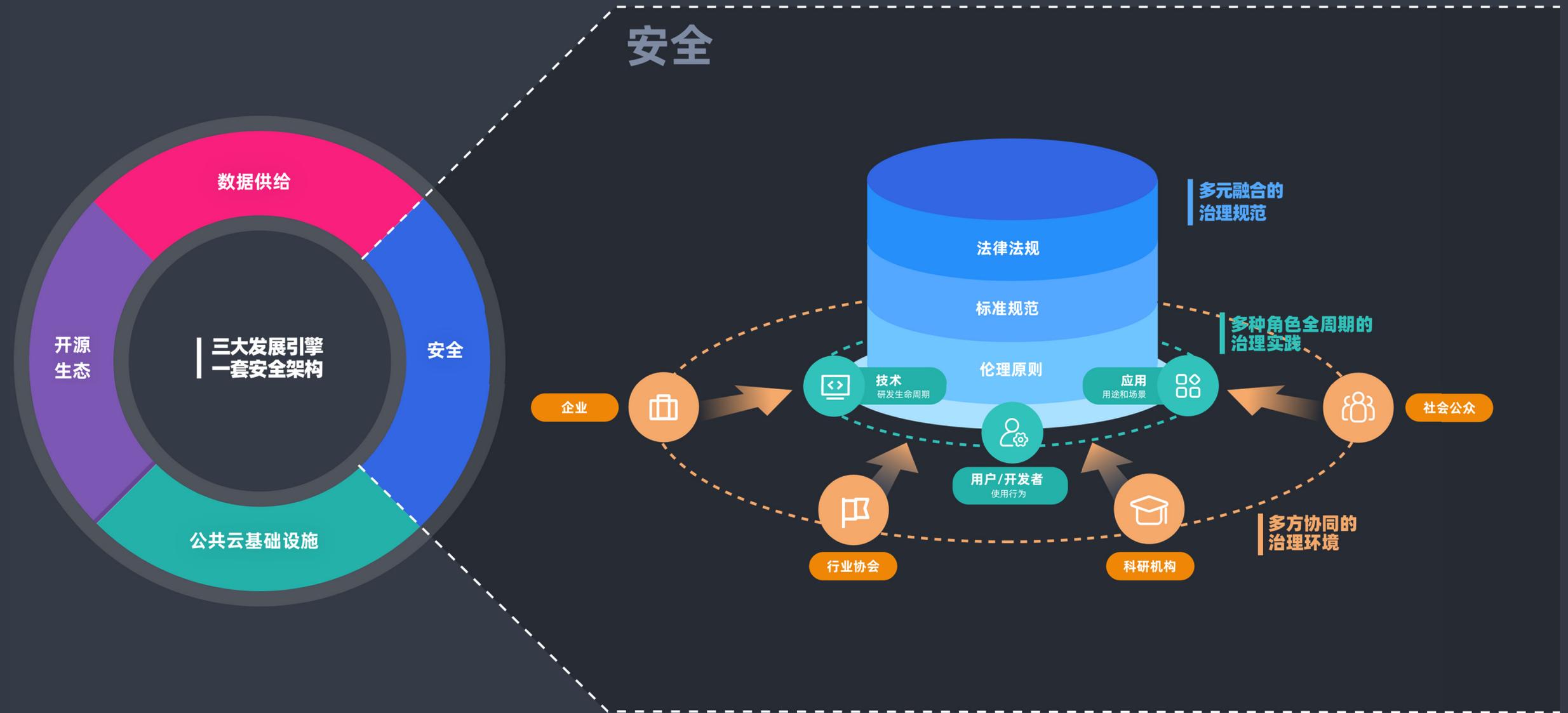
阿里巴巴

中国电子技术标准化研究院  
China Electronics Standardization Institute

阿里云 达摩院  
aliyun.com DAMO ACADEMY

# 三大发展引擎（公共云、开源生态、高质量数据）+ 一套安全架构

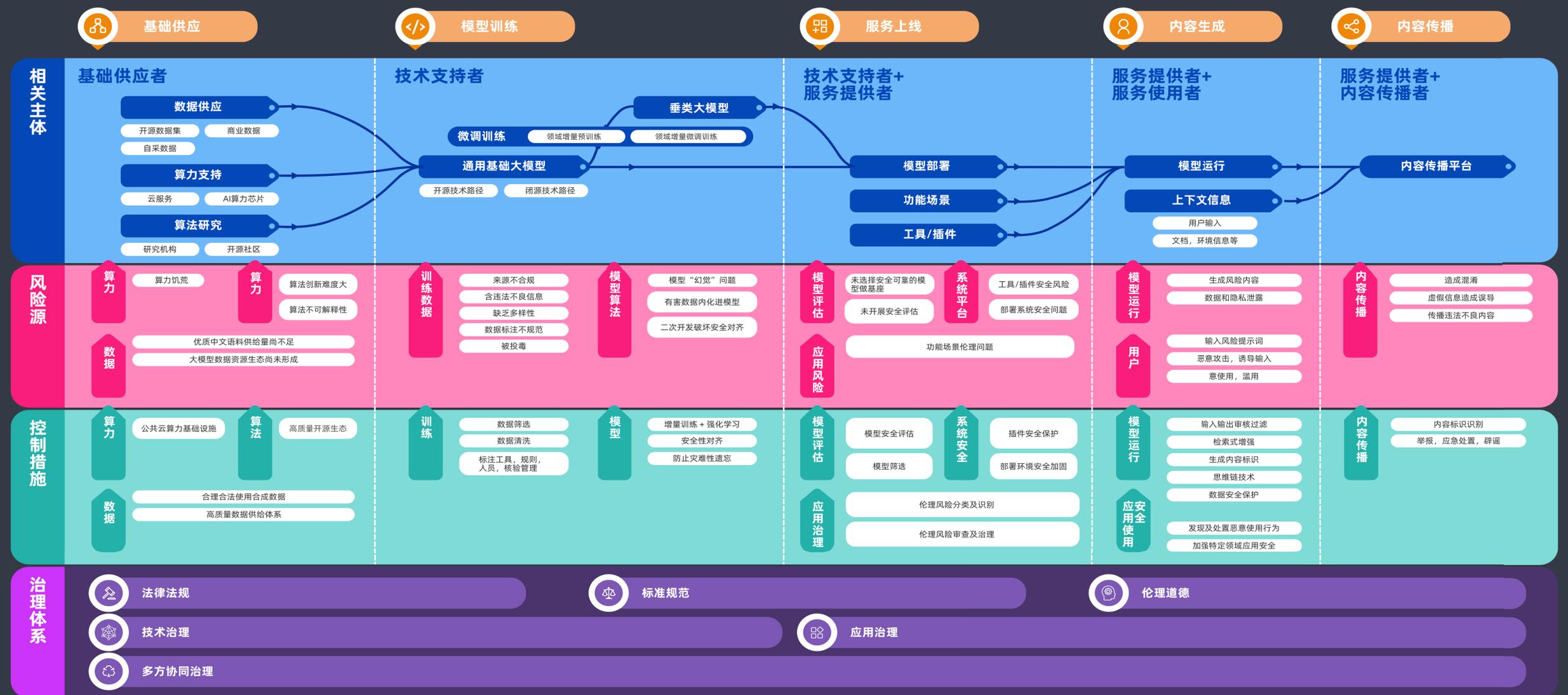
大模型已成为新质生产力的重要组成部分，促进着科技融合创新、产业结构升级。同时，也面临着来自算力、算法、数据以及安全等方面的新挑战。现如今，兼顾发展与安全的治理模式逐渐成为共识，本书基于大模型技术及其应用的风险现状与产业实践，形成了以公共云基础设施、开源生态、数据供给为发展引擎，以安全可信架构为保障的治理蓝图，为落地实现发展与安全兼顾的目标提供参考。



# 多角色全生命周期治理实践

大模型产业应用日益广泛，生态愈发成熟，其背后是逐渐清晰的角色划分和更加复杂的应用模式。在大模型研发应用的全生命周期中，基础供应者、技术支持者、服务提供者、服务使用者、内容传播者等相关角色在产业发展和安全保障方面承担着不同的使命和责任。在大模型落地应用时，还有更多的工具/插件、文档及环境信息、知识库等被开发和利用，帮助提升大模型的能力。

在这种复杂多元的产业现状下，新风险和新挑战与日俱增。如何沉淀多方安全实践，明确全链路、多角色的技术与应用控制措施，促进协同治理，有效防范化解人工智能的风险，保障大模型技术及应用的持续稳定发展，是本书阐述的重点。



# 1

引言



# 2

大模型发展与挑战



# 3

公共云是大模型技术发展和应用落地的优选路径



# 4

开源的并行探索生态是应用爆发和协同共治的必要条件



# 5

高质量数据供给生态是具备持续竞争力的基础



# 6

体系化的安全治理能力是稳定发展的保障



# 7

结语



# 大模型技术发展及治理实践报告

编写单位



## 指导委员会

邵晓锋 阿里巴巴集团风险委员会主席  
闻佳 阿里巴巴集团公共事务总裁  
俞思瑛 阿里巴巴集团首席法务官  
范科峰 中国电子技术标准化研究院副院长  
钱磊 阿里巴巴集团安全部总裁

## 专家委员会

薛晖 阿里巴巴科技伦理委员会委员 & AAIG 主任  
朱红儒 阿里云智能集团标准化业务副总裁  
袁媛 阿里研究院院长  
姚相振 中国电子技术标准化研究院网安中心主任  
胡影 中国电子技术标准化研究院网安中心副主任

## 编写组组长

张荣 阿里云智能集团通义安全负责人  
郝春亮 中国电子技术标准化研究院网安中心数据部主任

## 编写组主要成员

傅宏宇 阿里研究院人工智能治理中心主任  
李娅莉 阿里巴巴达摩院法务安全负责人  
彭骏涛 阿里云智能集团 AI 标准负责人  
黄龙涛 阿里安全资深算法专家  
陈岳峰 阿里安全 AIGC 安全算法负责人  
沈晖 阿里巴巴达摩院安全总监  
杨易侗 阿里安全生态合作负责人  
马宇诗 阿里安全高级体验设计专家

张妍婷 中国电子技术标准化研究院大模型安全标准负责人  
黄晴 中国电子技术标准化研究院工程师

## 关于我们

阿里巴巴人工智能治理与可持续发展研究中心（AAIG）是阿里巴巴集团旗下的人工智能顶级研发团队，致力于利用 AI 技术解决安全风险问题，并推动 AI 技术更加安全、可靠、可信赖和可用。团队成员在计算机视觉、自然语言理解、数据挖掘与网络安全等领域的国际顶级会议和期刊上发表论文 100 多篇，获得国际国内专利授权 60 余项，申请中专利 200 多项。AAIG 贯彻“科技创新是最好的网络安全”的理念，所研发的人工智能产品涵盖内容安全、业务风控、网络安全、数据安全与算法安全等多个领域，为集团在全球的千万商家和十亿消费者提供更好的安全和体验。

## 联系我们

aaig@list.alibaba-inc.com

扫码关注微信公众号

## 鸣谢

朱琳洁 王峥 陶嘉玲  
许晓东 聂云奕 李进锋  
李金纯 杜东为 王莹  
徐璐妮 牟立煜 石洪竺  
张强

# 目录

## 一

### 引言

1. 大模型成为新质生产力的重要组成部分	18
2. 安全可信是大模型持续发展的基本要求	20

## 二

### 大模型发展与挑战

1. 大模型技术与应用的发展现状	24
1.1 原生多模态成为发展主流	24
1.2 超长上下文理解能力不断突破	24
1.3 传统人机交互方式被颠覆重塑	25
1.4 大模型的应用模式日益丰富、产业生态愈发成熟	26
2. 算力、算法、数据与安全挑战	27
2.1 算力短缺问题突出，制约大模型持续发展	27
2.2 模型结构创新难度大，可解释性问题加剧	28
2.3 高质量数据供给不足，成为模型训练新瓶颈	28
2.4 大模型的技术特性带来了新的安全风险	29

## 三

### 公有云是大模型技术发展和应用落地的优选路径

1. 大模型的训练强依赖大规模的基础设施	34
1.1 资源挑战	34
1.2 技术挑战	34
2. 公共云是大模型发展与应用的优选解决方案	36
2.1 公共云为训练全球领先的大模型提供必要条件	36
2.2 公共云提供高效率 and 低成本	37
2.3 公共云为大模型提供全面安全保障	39
3. 公共云可有效应对算力短缺困境	41
3.1 算力短缺的根本原因是供应无法满足需求	41
3.2 公共云最大化发挥稀缺计算资源的价值，推动规模化应用	42
3.3 公共云可协同利用多来源算力，丰富算力生态	42

## 四

### 开源生态促进大模型算法发展与安全

1. 开源生态对于大模型的诞生具有基础性作用	46
2. 开源生态是大模型技术发展和应用推广的加速器	48
2.1 大模型开源的核心意义是普惠	48
2.2 开源生态助力提升大模型透明度与安全性	49
2.3 开源生态加速大模型发展和应用	50
2.4 开源生态助力我国抢占产业生态制高点	51
3. 开源生态提升大模型安全性、强化业界信任度	53
3.1 开源生态促进多方协同安全治理	53
3.2 开源生态确保模型安全评估客观公正	54
3.3 开源生态帮助建立业界对大模型技术的信任	55

# 五

## 高质量数据供给生态是具备持续竞争力的基础

1. 合成数据是解决高质量训练数据供给不足的突破口	58
2. 构建更匹配模型部署要求的高质量应用数据生态	61
2.1 数据上云推动模型应用数据生态建设，提供保障数据安全的能力基础	61
2.2 检索增强是模型高质量用数的可行方案	62
2.3 以更开放、务实、多元的方式促进高质量数据供给	64

# 六

## 体系化的安全治理能力是稳定发展的保障

1. 多角色视角安全治理架构	68
2. 大模型安全治理规范：全面多元融合的规范框架	71
2.1 法律法规、标准规范、伦理道德原则的性质、作用、执行路径有明显的差异	71
2.2 组织保障是一切体系化治理能力的基础	72
3. 大模型安全治理措施：技术与应用风险治理构成的立体防护网	73
3.1 围绕全生命周期的技术治理措施	73
3.2 针对用途和场景的应用风险治理	91
3.3 以安全使用管理防范滥用误用行为	93
4. 模型产业链逐步成熟，多方协同治理价值日益凸显	95
4.1 多方协同治理的价值	95
4.2 多方协同治理的特点	96
4.3 多方协同治理中的产业贡献	96

# 七

## 结语

## 附录A：大模型各领域应用场景及实践案例

1.1 个人生活新助理	106
1.2 产业智能新模式	106
1.3 公共服务新体验	108

## 附录B：云上开放的大模型平台： 百炼

1. 大模型时代的需求分析	110
1.1 灵活性和弹性兼备的计算能力	110
1.2 丰富的模型供给	111
1.3 模型全生命周期的服务	111
1.4 构建应用的工具链	112
1.5 合规与安全是基础	112
2. 百炼作为大模型服务平台的突出优势	113
2.1 百炼提供强大而实惠的计算资源	113
2.2 百炼提供丰富的模型供给	114
2.3 百炼提供模型全生命周期的服务	115
2.4 百炼提供构建应用的完整工具链	116
2.5 合规性与安全性	118
3. 百炼助推我国人工智能行业发展	120

## 附录C：人工智能开源社区： 魔搭

1. 人工智能发展与应用须跨过四道门槛	121
2. MaaS 实现人工智能模型“开箱即用”	122
3. 魔搭社区的四大核心优势	123
4. 魔搭社区正成为模型应用的“加速器”	124
5. 魔搭社区助力大模型安全	125

CHAPTER. 1

# 引言



# 1. 大模型成为新质生产力的重要组成部分

**大模型已成为信息技术的新里程碑。**纵观历史，信息技术的发展往往由计算机技术、数据技术、算法技术的独立发展而推动，传统人工智能局限于执行特定目的功能、应用于特定场景之中。大模型整合利用了各类信息技术资源，不仅代表着算法的飞跃，也是算力、数据、工程等协同发展的结果，推动信息技术新拐点的到来。随着大模型的不断发展，人工智能的能力边界不断拓展，正逐步从目的单一、场景限定的专用型人工智能向具有认知和推理能力、胜任多种用途的通用型人工智能转变。

**大模型促进科技进一步融合创新。**自 2023 年以来，大模型步入了高速发展阶段，技术先进性和能力通用性兼备成为其核心优势，通过大模型赋能千行百业，利用人工智能提高传统要素的边际生产率已经逐步成为现实。在国家大力推动“人工智能+”行业应用融合创新的背景下，大模型已经渗透到许多行业并得到深度应用，例如，在在线购物中降低商品展示成本、提高用户购物体验；在工业机器人和无人驾驶中，环境感知和判断提高了其自动化能力；人工智能驱动的科研推动解决蛋白质组成结构等科学难题。从服务用户到助力生产，从基础研发到应用开发，大模型大幅提升了生产效率和创新能力。

**大模型创新推动产业结构全面升级。**大模型所带来的技术革新降低了知识门槛、节约了创新成本、提高了市场活力，让更多人能够参与到创造性活动中，催生全新的商业模式和服务形态，引领新一轮的产业转型升级。大模型有望在未来重塑产业链条，推动形成以数据驱动、智能主导的新型生产方式和经济增长点，让经济发展始终保持活力，通过持续创新推动高质量发展，用科技进步解决发展带来的不平衡难题。

**大模型发展与治理同步推进成为新共识。**大模型技术的出现和发展标志着人工智能从专用化走向通用化，人工智能超越人类智能、人类和机器共生共栖的社会场景从幻想走向现实。人工智能从多元发展转向聚焦于大模型的通用能力，通用人工智能成为国家科技能力现代化的重要标志，大模型的泛化、推理等特有能力为大模型更为广泛的应用提供了新的机遇，大模型研发所需的模型、算力和数据等科技要素成为关注焦点，对人才、投入和制度进行有效利用和合理配置的政策试点也迅速铺开。与此同时，大模型超出预期的表现也不断刷新人类对机器智能的认知。人工智能的突破性创新对人类文化价值、社会经济形态、个人基本权益带来诸多影响，重构人类对机器的主导性和权威性，让人工智能回归以人类为中心的发展路径，有效管控大模型带来的核心风险等目标成为各国政府以及人工智能产业界的共识。

## 2. 安全可信是大模型持续发展的基本要求

**人工智能安全理念从 AI Security 向 AI Safety 发展。**随着互联网步入智能化时代，网络空间安全的关注重心从系统安全、数据安全转向人工智能安全。大模型技术的快速发展和通用化应用潜力对人类社会和网络文明带来了新的挑战，相较于关注系统完整性、防止数据泄露和未授权访问的 AI security 理念，AI safety 理念更多聚焦于前沿人工智能技术对人类核心权益和中长期发展的影响，通过寻求共识、构建规则、评估影响、防控风险、完善技术、落实责任、推动协作的系统化机制，将国家安全、个体权利、群体利益、国际合作等纳入人工智能安全的要求之中，呼吁各方共同面对人工智能带来的不确定性挑战，更负责任地开发、部署和使用人工智能，防止滥用误用，通过人工智能安全治理构建信任，让可靠的人工智能推动网络技术应用发展。

**追求大模型安全可信是全球共识。**自大模型问世之时起，其潜在的安全风险和安全技术能力就备受全球关注，在大模型技术能力不断突破、模型性能不断提升的同时，人们对大模型安全风险以及安全治理的研究也在持续推进。根据大模型的技术特征和应用特点，对大模型进行有效治理，让大模型始终处在人类控制之下，不断降低大模型的安全风险，让大模型更好地服务人类发展，已成为全球共识。欧盟针对通用人工智能进行立法，保障人类权利免受大模型侵害，美国、英国则针对基础模型或前沿模型提出安全倡议，并发展风险评测等安全技术。我国《全球人工智能治理倡议》强调以人为本、智能向善，坚持发展与安全并重，提出打造可审核、可监督、可追溯、可信赖的人工智能技术，以务实、包容的态度动态管控大模型风险，综合采取政策、标准、技术等手段，让大模型成为安全可信的生产力工具。人工智能科技企业主动尽责，响应政府倡议、落实安全要求，在大模型安全方面加大投入、加强研究，不断提高大模型安全水平，更好地在人工智能技术扩散期建立信任。

**安全是大模型发展的必然要求，是大模型广泛应用的前提条件。**一是人工智能技术的发展离不开安全。人工智能安全技术能够控制人工智能本身带来的负外部性，规范人工智能技术的健康持续发展，降低技术新变化对人类社会的潜在伤害，让更多人享受到科技可持续发展带来的惠益。具有代表性的大模型企业都不断加大在模型安全方面的技术研发投入，提高模型负责任性，强化模型与人类价值观的对齐，从而建立并加强开发者和使用者对大模型的信任和信赖。二是安全水平将成为大模型应用的核心竞争力。安全技术对于大模型的应用至关重要，只有当用户确信其数据隐私能得到有效保护，模型输出结果可靠且不易受攻击时，大模型才能在更多领域得以广泛应用，并逐步与实体经济深度融合。大模型的安全水平也决定了特定行业智能化发展的速度，如金融、医疗等行业对安全有着极高的要求，大模型安全技术的成熟度直接决定了这些行业智能化发展的速度和格局。

CHAPTER.2

# 大模型发展 与挑战



# 1. 大模型技术与应用的发展现状

## 1.1 原生多模态成为发展主流

**当前，大模型的关键词是“多模态”。**近期大模型多模态能力迅速提升，从只支持文本逐步扩展到不同模态，目前已具有对文本、图像、音频、视频等多模态内容的理解分析和推理等能力。2024年5月，OpenAI 发布了支持全模态信息输入的 GPT-4o，Google 发布了同样支持多模态输入的 Gemini 1.5 Pro，能够同时处理和理解文本、视频和音频数据；国内阿里云等企业也陆续发布文生音频、文生视频等大模型产品。多模态能力使得模型在处理包含丰富信息的复杂场景时表现出色，例如在视频内容理解或多语言翻译任务中，模型能够从这些不同模态的数据中提取信息，并在需要进行综合分析，以提供更准确和丰富的响应。

**原生多模态可能是未来大模型的“主赛道”。**区别于将多个不同模态的模型整合来获取多模态能力，原生多模态大模型是指模型天然具备处理多种不同类型输入数据（例如文本、图像、音频等）的能力。与传统的单一模态模型不同，原生多模态大模型通过同时处理、理解、生成多种不同类型的信息，从而实现更加全面和综合的语义理解和生成任务。这也是 GPT-4o 等模型能够直接输出拟人化、具备情感交流能力的原因，也进一步提升了生成内容的拟真度。现阶段来看，原生多模态是当下一段时间大模型主要发展突破的重点。

## 1.2 超长上下文理解能力不断突破

**大模型能够处理的上下文长度是其基础能力的体现。**对较长上下文的处理能力是对话系统、文档处理、代码生成以及工具调用等的基础，能够处理长上下文意味着大模型

能够理解和处理越复杂的信息，继而能进行更全面、细致、准确地阅读和理解。这种能力使得大模型在复杂和不熟悉的情境下能也提供更专业、更有价值的帮助，从而进一步提高其在日常应用中的实用性和效率。

**大模型上下文窗口在过去一年不断突破。**得益于近一年大模型相关的算法架构优化（如位置编码和注意机制改进）、模型压缩技术（如量化和剪枝）以及计算资源（GPU 显存）的显著提升，大模型处理长序列的能力大幅提升。大模型上下文窗口已经从 GPT-4 的 128K 突破到百万乃至千万 token 量级，这意味着可向模型输入数小时视频、数十小时音频、数万行代码，以及超过百万的单词。Google 的 Gemini、阿里巴巴的 Qwen-Long 等模型均已具备类似能力。

## 1.3 传统人机交互方式被颠覆重塑

**具备多模态能力的大模型可像人一样与用户交互。**大模型产品化的加速发展正在重塑人机交互方式，极大改变了人们与技术的互动模式，人机交互方式未来将更接近人与人之间的交互。近期不断有大模型产品展示视频发布，大模型通过麦克风接收用户语音输入，通过摄像头观察用户状态，直接与用户对话并给出工作生活建议。此外，传统计算机操作系统集成了大模型后，用户可通过自然语言命令控制计算机执行各种任务，例如微软发布的 Copilot PC。这种交互方式的革新，不仅提高了工作效率，还使得非技术背景的用户也能轻松使用高级功能，极大降低了技术的使用门槛。

未来，人机交互的范式将从传统的图形界面和复杂的命令行操作转变为更加直观和自然的多模态交互方式。大模型对人机交互方式的重塑在为人类提供便捷的同时，也为技术发展开辟了新的道路。

## 1.4 大模型的应用模式日益丰富、产业生态愈发成熟

**产业应用日益广泛的背后是更加丰富的应用模式。**在模型开发层面，基于基础大模型的优化训练、二次开发被广泛应用，使得各行各业能以较低成本训练出垂直领域大模型。在模型应用层面，检索增强、AI Agent 等技术的发展，使得大量上下文信息、工具 / 插件成为大模型的一部分，有效促进了大模型与实体经济的深度适配应用。

**产业生态愈发成熟呈现出更加清晰的角色分工。**数据、算力、算法是大模型的基础，其相关主体，如数据标注平台和数据提供商、云服务和芯片厂商、学术研究机构 and 开源社区等，为大模型提供基础支撑；在模型训练阶段，技术支持者完成数据准备、算法设计、模型训练和优化等，产出大模型；在服务上线阶段，技术支持者和服务提供者共同保证选取的大模型、部署环境、应用和系统的质量和安​​全；在内容生成阶段，产品和服务将面向用户进行交互，由服务提供者完成模型的运行和监控；最后，由于大模型在拟真、泛化等方面的超强能力，用户进行生成内容发布和传播阶段的安全措施也是大模型产业链需要考虑的重要部分。

## 2. 算力、算法、数据与安全挑战

在大模型高速发展的同时，我们也面临着算力短缺、高质量数据缺乏和模型安全等问题的挑战。要克服这些障碍，需要技术创新、行业合作和政策支持，从而确保大模型的安全、可靠、可信赖、可用。

### 2.1 算力短缺问题突出，制约大模型持续发展

**算力短缺是全球共性问题。**算力短缺的本质是需求增速远超产能扩张速度和芯片性能的提升速度。随着模型参数规模的快速增长，以及模型应用的泛化和深化，全球对算力的需求高速增长，供给缺口巨大。比如，Meta 计划到 2024 年年底拥有 35 万张 H100 英伟达 GPU 卡，微软计划 2024 年采购数万张 B100；亚马逊利用 16000 张以上 H200 卡在公共云上提供大模型服务。受限于产能不足，英伟达的高端 GPU H100 交货周期尽管已经缩短，但依然需要 8-12 周。

**我国算力短缺问题突出。**我国正在大模型领域奋勇发展，训练和推理对算力的需求快速上升。OpenAI 训练 GPT-4 使用了大约 25000 张 A100 GPU 卡；谷歌训练 Gemini 所使用的算力是 GPT-4 训练算力的 4-5 倍。我国企业要训练全球领先水平的大模型，至少需要万卡、十万卡的集成算力。然而，我国算力短缺问题仍然突出，其主要原因一是国产 GPU 在短期内还无法完全替代海外高端芯片，无法满足大模型的训练和推理规模化应用；二是算力市场被区域性分割、大量私有化集群利用率低，导致有限的资源未被充分利用，造成浪费。

## 2.2 模型结构创新难度大，可解释性问题加剧

**模型结构创新难度大。**在理论上我们仍然缺乏对于大模型为何有效以及如何最好地进行设计的深刻理解。这导致了在模型结构创新时不是依靠坚实的理论基础，而更多地是依靠经验和试错。由于大模型训练需要海量的高质量数据和算力资源，又缺乏有效的理论指导，试错成本非常高，而且花费大量资源仍然失败的概率并不低。

**大模型算法的“黑盒”特性。**大模型通过学习大量数据的模式来进行预测，意味着模型的行为是数据驱动的，而不是基于明确的因果关系或其他易于理解的原则。即使我们从技术上知道模型是如何工作的，也很难解释为什么模型会做出某个特定的行为。大模型通常包含数十亿到万亿级别的参数，这些参数在网络内部通过非线性变换相互作用。这种高度复杂的交互使得单个参数或层对最终输出的影响变得极其难以追踪，加剧了可解释性不足的问题。

## 2.3 高质量数据供给不足，成为模型训练新瓶颈

**优质中文语料供给尚显不足。**尽管中国互联网用户数量众多，产生了海量数据，但经过数据筛选、清洗、结构化且标注良好的语料依然相对稀缺。我国正在大力推动中文语料训练数据资源平台的建设，但短时间内仍难以满足大模型对丰富多样训练数据的海量需求。在能提升模型价值观引导能力的语料、增强对物理世界专业性理解的科研数据和期刊论文，以及提升专业知识理解能力的领域知识等方面，还需持续增加训练数据的深度和广度。

**大模型数据资源生态尚未完善。**一方面，我国虽然公共数据范围覆盖较广，但在开放的范围、数据质量、与大模型训练匹配度等方面还有较大差距。另一方面，我国数据

资源较为碎片化，数据集开放程度低，企业用于训练的语料来源不清晰、权属不明确，开源后存在一定隐患，使得企业更倾向于自采、自用，大模型数据流通受阻。综上，识别高质量训练数据，推动高质量数据有效供给，通过技术方案解决训练数据供给不足的问题，需要各方共同研究并努力推动。

## 2.4 大模型的技术特性带来了新的安全风险

### 2.4.1 输入方面，训练数据中的不当内容内化进模型仍是主要风险

大模型通过深度神经网络，基于大规模真实数据进行训练，形成向量化的概率分布。模型在给定条件下可能生成多种输出，如果训练数据中存在不当内容的微弱信号，模型在生成时可能会放大这些信号，产生不当输出。一是数据预处理的局限性可能引入不当内容，由于训练数据量巨大，过滤算法可能无法完美识别出所有违法不良信息，很难保证训练数据中完全不存在不当内容。在训练阶段，大模型会内化这些知识，从而在后续的生成过程中产生安全风险。二是模型对训练数据的创造性组合能力可能新生成或放大不当内容，大模型在理解上下文方面取得了显著进步，但在处理复杂语境依赖关系时，仍可能错误地将不同情境下的表达组合在一起，产生不当内容。三是大模型在试图从有限训练数据中归纳出更广泛的规律时，可能过度简化或泛化某些概念，甚至放大训练数据中不当内容的微弱信号，导致生成内容时出现偏差。

### 2.4.2 模型方面，“幻觉”问题爆发带来误导用户的风险

**“知识幻觉”是难以避免的技术问题。**传统模型通常围绕特定领域的分类、聚类、回归等单一任务寻找最优损失函数 (Loss Function)，形成单一的识别能力。而大模型在预训练阶段采用无监督学习的方式学习世界知识，在后训练阶段学习如何利用学到的

知识解决不同的任务，这个过程是大模型通用能力形成的基础，使其具备了有问必答的特质，同时也产生“幻觉”。具体而言，幻觉的来源主要包括四方面，一是预训练阶段缺乏相关领域知识或学到了错误知识，二是后训练阶段引入了问题数据（例如包含偏见的数据），三是生成阶段大模型超出自身知识边界输出结果，四是大模型生成的随机性带来了内容含义的偏离。

“幻觉问题”在高安全需求领域有较大影响。从风险角度来看，**过于泛化的大模型会带来结果不真实、偏离客观规律和与世界事实不符等误导用户的风险**，这些风险在医疗、司法等领域应用时，更有可能演化为真实的安全问题，大模型需要重点考虑其泛化性与安全性的平衡点，并采取相应控制措施。

#### 2.4.3 应用方面，超长技术引发全新的诱导攻击

Transformer 架构的自注意力机制是大模型区别于传统模型并带来智能涌现的基础，它让大模型如同人类大脑一般理解并学习世界，同时也带来了全新的风险。自注意力机制允许模型在处理输入时，对不同部分之间的关系给予不同的注意力权重。这意味着用户可以在大模型产品中获得更具交互性的体验，同时也意味着如果攻击者能够精心设计输入，使得模型对某些关键词或短语赋予过高的关注度，那么模型就可能被误导去执行特定的行为，或直接改变、忽略原有的安全机制而生成不当的输出，因此产生了全新的诱导攻击。如今超长上下文技术的发展，将大模型应用推向了一个全新的高度，但也极大加剧了这种风险。

#### 2.4.4 输出方面，生成内容愈发逼真加剧误用、滥用、恶意使用风险

随着大模型在多模态理解能力和原生多模态技术等方面的进步，AIGC（人工智能生成内容）在拟真度方面取得了显著进步，高度的拟真导致了用户更容易对 AIGC 和真实内容之间产生混淆，一方面增加了误用、滥用的可能性，例如医疗咨询场景中的

AIGC 内容如不清楚标识，可能被误认为医生的诊断结果；另一方面加重了恶意使用的风险影响，例如 AIGC 在情景化和个性化语境中实施大规模、高效率的信息欺诈变得更加容易。同时，大模型产品化的加速发展、AI Agent 和垂直领域模型微调的广泛应用，使得大模型滥用和恶意使用的门槛被降低，并具备了规模化生产的条件。

## CHAPTER.3

# 公共云是大模型 技术发展和应用 落地的优选路径



# 1. 大模型的训练强依赖大规模的基础设施

大模型训练是一项系统性工程，一是需要足够算力支撑，对计算硬件资源有极高要求；二是需具备对大规模硬件的高效调度和通信能力，实现并行计算以及强大的系统监控与稳定性管理，确保训练过程顺畅与资源最优化利用。尤其是在预训练阶段，工程能力是更为关键的因素。

## 1.1 资源挑战

大模型的参数规模越来越大（目前世界先进的模型普遍拥有 100B 及以上的参数），训练使用的语料越来越多（普遍有 5T token 以上），其训练过程对计算资源、存储能力和网络带宽有极高的要求。

以 Llama3 技术报告中披露的 405B 参数规模模型的信息为例。计算资源方面，Llama3 的训练需要 16000 张 H100 GPU，每个服务器配有 8 个通过 NVLink 连接的 GPU 和 2 个 CPU，且具备专门设计的全局规模训练调度器安排训练作业。存储方面，使用了 7500 个配备 SSD 的服务器上的 240PB 存储空间，并支持 2TB/s 的可持续吞吐量和 7TB/s 的峰值吞吐量。

## 1.2 技术挑战

在过去的一年里，预训练技术不断创新，硬件能力持续进步，但由于参数规模和训练语料的增长，大模型训练时长仍然以月为单位计算，需要海量基础设施长时间稳定可靠地运行。

一是大模型训练的同步特性使得训练过程的故障容忍度低，单个 GPU 故障可能导致整个训练的重启。仍然以 Llama3 技术报告为例，使用 16000 张 GPU 训练的复杂性和潜在故障场景超过了更大规模的 CPU 集群。在预训练的 54 天里，Llama3 经历了 466 次作业中断，其中 47 次是计划内的中断，如固件升级或操作员启动的操作，如配置或数据集更新。其余 419 次是意外的中断，其中 78% 归因于硬件问题，如 GPU 或主机组件故障等。有时还会出现没有明显故障，但运行缓慢的落后者，单个落后者也可以减慢数千个其他 GPU 的速度。

二是训练过程中往往需要及时保存 checkpoint，用于恢复和调试，一个主要挑战是支持在短时间内饱和存储网络的高突发性 checkpoint 写操作。网络方面，以 Llama3 为例，Meta 设计了一个复杂的三层网络连接数千台服务器。大型语言模型训练产生的网络流量极大，使用传统方法（如等价成本多路径路由方法等）在所有可用网络路径上进行负载均衡难以实现。除此之外，还需要解决训练过程中常见的由集合通信模式引起的瞬态拥塞和缓冲难题。

三是大规模训练对供电网络同样也会带来挑战。数万个 GPU 可能同时增加或减少功耗，例如：当所有 GPU 等待 checkpoint 或通信完成的时刻，以及整个训练任务的启动或关闭的时刻。这种情况可能导致数据中心的功率瞬间波动数十兆瓦，对供电网络造成大的冲击。随着未来模型的持续迭代和参数规模的增长，这是我们面临的一个持续挑战。

## 2. 公共云是大模型发展与应用的优选解决方案

大模型的发展与应用所需要的条件与公共云的优势高度契合。公共云凭借强大的基建能力、高效的资源利用和较低的成本，以及完备的安全保障，为大模型的发展和应用提供了极具竞争力的解决方案。

### 2.1 公共云为训练全球领先的大模型提供必要条件

**公共云具备建设万卡集群的工程化能力。**当前，训练全球领先水平的大模型至少需要具备万卡的集成能力，比如 Llama3 405B 模型的训练使用了 16000 张 H100。构建万卡集群具有非常高的技术门槛，因为需要支持大量芯片做高速并行运算，通过优化技术达到较高的运行效率，并保持长时间稳定运行。这并不是简单的芯片堆砌，也不仅仅是加大资金投入就可以解决的问题。实现万卡高效协同工作，涉及 GPU 与 GPU 之间、服务器与服务器之间，甚至是数据中心与数据中心之间的通信、监控、调度、管理，其背后是高并发分布式计算、高性能网络、超大规模弹性计算能力，包括芯片、操作系统和网络等庞大的软硬件结合的工程，这些与云计算（尤其公共云）的技术特征完全吻合。通过优化的软件栈和调度算法，公共云可以实现万卡集群高达 90% 的并行效率。

**公共云大规模计算集群具有高稳定性。**训练全球领先水平的大模型需要大规模计算集群具备高可靠性，长时间稳定运行。公共云通过多可用区部署，确保即使某个区域发生故障，服务依然能够在其他区域无缝切换，维持服务连续性。关键组件和服务采用了冗余设计，确保主节点故障时备节点可以迅速接管，降低单点故障的风险。公共云通过实时监控集群状态，一旦检测到异常，立即触发告警通知，并启动自动化修复流

程，可以自动检测并替换故障节点，无需人工干预。定期进行自动数据备份，并建立异地容灾机制，以防发生意外时能迅速恢复数据。通过上述措施，公共云确保大规模计算集群能够长期稳定运行，为用户提供可靠的服务。这种高稳定性不仅体现在日常运营中，也能经受住一些极端情况的考验，有效训练时间占比超过 99%。

### 2.2 公共云提供高效率和低成本

**公共云预置完备的工具提高工作效率。**云计算平台可以整合从数据预处理、模型训练到模型评测的全流程所需的工具，大幅度降低使用难度，提升工作效率。高效训练：公共云提供分布式存储解决方案，可以高效存储和访问 TB 甚至 PB 级别的数据，解决高频访问海量数据集的问题。公共云预置的深度学习框架（如 TensorFlow、PyTorch、MXNet 等）、开发套件，集成了高性能计算库、分布式训练工具、自动混合精度训练等功能，能够显著提升训练效率，缩短训练时间，间接节省了计算资源消耗。简化部署：公共云提供了开箱即用的模型部署服务，用户可以便捷地将完成训练的模型封装成 API、容器服务或边缘计算模块，快速上线并对外提供服务，无需从零构建复杂的后端架构，大大减少了部署时间和人力成本。通过容器化技术，确保大模型训练和推理环境的一致性，简化版本控制和更新过程。按照阿里云的测算，在公共云上部署模型，效率是企业内部自建的 2 倍及以上。

**公共云的资源共享带来资源的高效使用。**公共云通过资源共享机制，能够在不同用户和任务之间动态分配计算资源，避免了专用硬件常见的闲置和浪费，降低了成本，增强了计算资源的灵活性和可扩展性。资源集中管理：公共云支持大规模分布式训练，允许用户轻松搭建和管理包含数百甚至数千个计算节点的集群。公共云能够自动均衡工作负载、优化数据通信、处理节点故障，确保资源在大规模训练任务中得到充分利用。按需使用节约成本：用户仅需为实际使用的计算、存储和网络资源付费，无需预先投资昂贵的硬件设备。用户可以根据模型规模、训练时间窗口以及计算需求的变化，

动态调整资源配置，比如临时增加 GPU 实例数量以加快训练速度，或在非高峰时段减少资源以降低成本，实现资源的精准匹配和高效利用。按需使用的方式避免了因过度预估需求而购买闲置硬件导致的浪费，也消除了因需求增长而不得不频繁升级硬件设施的成本压力。对于周期性或突发性的计算需求，如模型迭代训练、大规模数据预处理等，公共云能够迅速提供所需的额外资源，完成后即可释放，极大降低了总体成本。根据业界的统计，从使用效率看，公共云资源使用效率一般可达 30%~40%；而私有云资源使用效率一般约为 5%~10%。

**公共云对能源利用效率高促进可持续性发展。**公共云往往投资建设高效、环保的数据中心、计算中心，采用先进的冷却技术、电源管理系统以及可再生能源供电，其能源效率远高于一般企业自建数据中心。公共云积极推行“东数西算”，在西部地区建设数据中心，发挥西部拥有丰富的土地、水力和太阳能等清洁能源资源的优势。如果将丰富的电力资源用于大模型的训练和推理，既能够满足大规模 GPU 集群的电力需求，也节省了电力远距离传输的成本。西部数据中心的绿色化程度更高。按照全国一体化数据中心建设要求，西部地区新建数据中心的平均 PUE（电源使用效率）值不大于 1.2，低于东部地区新建数据中心的 1.25，这意味着其能效更高，对环境的影响更小。选择使用公共云进行大模型训练和部署，实质上是在共享这些绿色基础设施，从而间接降低了碳排放和能源消耗。

## 2.3 公共云为大模型提供全面安全保障

公共云原生的安全能力为大模型的全生命周期提供完备的安全保障。云通过提供标准化与自动化的基础设施级防护、动态实时防护、数据安全、内容安全、算法安全等多重保障，显著增强了大模型全生命周期的安全性，助力企业应对复杂的安全挑战，确保服务的稳定和安全。

**标准化与自动化的基础设施级别安全管控。**云平台通常具备强大的底层安全防护能力，包括 DDoS 防护、入侵检测与防御系统（IDS/IPS）、防火墙、密钥管理服务等，为大模型提供全方位的安全屏障。云上安全配置、策略和工具高度标准化，能够自动应用最佳安全实践，如使用安全的镜像构建流程、实施最小权限原则、集成安全扫描等，降低因人为错误导致的安全风险。DevSecOps（开发、安全与运维一体化）确保安全检测与加固贯穿大模型应用的开发、测试、部署全过程，及时发现并修复潜在安全漏洞。

**动态安全策略与实时防护。**公共云厂商持续投资于安全研究与防护技术升级，能够快速响应新威胁，为用户提供及时的安全补丁和更新，减轻用户的安全运维压力。云的技术体系支持动态调整安全策略，可以根据大模型应用的运行状态、威胁情报以及业务需求实时更新防护措施，提供细粒度的访问控制、网络隔离和流量管理。公共云对大模型应用的内部组件间通信进行严格控制，防止横向攻击；通过服务网格实现安全策略的集中管理和分布式执行，增强整体安全性。

**数据安全与隐私保护。**对于大规模训练数据集，尤其是高价值的 SFT、DPO 数据，公共云提供安全的数据存储服务，配合严格的访问控制和审计机制，防止数据的未授权访问。容器化技术和虚拟化技术提供了逻辑隔离，使每个大模型应用实例在独立的环境中运行，有效防止推理服务阶段用户数据泄漏。公共云支持数据的加密存储、数据的加密传输（如 SSL/TLS）、密钥管理和数据脱敏等技术，确保大模型训练和推理涉及的用户数据在存储、处理和传输过程中的保密性和完整性。

**全生命周期一站式算法安全服务。**云平台作为大模型服务的核心载体，集成了先进的安全技术和丰富的实践经验，为用户提供统一的、功能全面的算法安全解决方案。这些安全能力不仅针对大模型特有的复杂性和挑战进行了优化，还能够灵活适应不同行业和应用场景的需求，确保大模型在数据处理、训练、评测、部署、推理全生命周期内的每一个环节，都融入全面而深入的安全保障措施，确保人工智能模型既强大又安全可靠。



**云为训练全球领先的大模型提供必要条件**

先进大模型的训练需要强大的基础设施  
云具备建设万卡集群的工程化能力  
云的技术优势和大模型的需求高度契合



**云提供高效率和低成本**

云预置完备的工具提高工作效率  
云的资源共享带来资源的高效使用  
云对能源利用效率高促进可持续性发展



**云为大模型提供全面安全保障**

标准化与自动化的基础设施级别安全管控  
动态安全策略与实时防护  
数据安全与隐私保护  
全生命周期一站式算法安全服务

## 3. 公共云可有效应对算力短缺困境

算力短缺的原因表面上是外部供应链发生波动，本质是需求和产能的差距。公共云在大规模计算集群、有效利用稀缺资源、异构算力协同以及分布式算力需求等方面都具有显著优势，是破解我国当前算力短缺的有效路径。

### 3.1 算力短缺的根本原因是供应无法满足需求

算力短缺的本质是市场对算力的需求增速远远超过产能的扩张速度和芯片性能的提升速度。

**市场对算力的需求增速远超产能扩张速度。**随着模型参数规模的快速增长，以及模型开始得到广泛而深入的应用，算力需求的增长曲线将变得更加陡峭。2024年第三季度，英伟达 H200 高性能芯片开始出货，受限于产能不足，交货周期大约为一年。根据需求和产能预测，整个 2025 年都会持续供货紧张状态。

**从技术发展趋势看，单个芯片性能的提升跟不上算力需求的增速。**尽管单个芯片的性能在过去几年中取得了显著的进步，但进步的速度正在逐渐放缓。这主要是由于物理限制和技术瓶颈的存在，例如功耗墙、散热问题和制造工艺的复杂性等。这些因素都限制了单个芯片性能的进一步提高。随着模型参数量的持续增加和数据量的增长，对于计算能力的要求也在不断提高。这意味着即使单个芯片的性能得到提升，也可能不足以满足日益增长的计算需求，需要更多的芯片协同工作以提供足够的计算能力。

### 3.2 公共云最大化发挥稀缺计算资源的价值，推动规模化应用

公共云可充分发挥稀缺的高端芯片的价值，延长其寿命；并通过技术能力聚合大量低端芯片达到少量高端芯片的计算能力，增加高质量算力的供应，推动规模化的应用。

**公共云可实现对既有高端芯片的深度“盘活”。**这种设计的核心在于，通过先进的资源调度算法和负载均衡技术，能够在不增加硬件投入的前提下，充分挖掘现有芯片的潜能，延长其有效使用寿命，同时通过集群的并行计算能力，形成一个庞大的算力资源池。这一资源池不仅能够提供媲美乃至超越单个高端芯片的算力输出，还能根据实际需求动态调整，实现算力的弹性伸缩，极大地提高了算力供给的灵活性和效率。

**公共云可通过技术能力聚合大量低端芯片达到少量高端芯片的计算能力。**借助公共云操作系统的线性扩展能力，可以克服单一芯片性能局限，实现集群计算能力的显著提升，大幅度降低单个芯片能力不足的影响，实现聚集更多低端芯片达到少量高端芯片的计算能力。例如，我国的飞天操作系统，已在 TPC-DS 的国际基准测试中验证了通过调用约 280 台 A 品牌服务器就达到了 100 台 B 品牌服务器的计算能力，而针对单个服务器 A 品牌的计算能力远远低于 B 品牌，与 B 品牌的计算能力相差 7 倍。

### 3.3 公共云可协同利用多来源算力，丰富算力生态

公共云凭借其在算力资源调配与优化方面的独特优势，可充分利用现有的 CPU 资源，并支持异构芯片协同计算，为算力基础设施的升级和拓展提供了可行路径，为构建安全可控的算力生态提供了有益借鉴。

**公共云在技术上已实现利用 CPU 资源满足部分大模型推理算力需求。**在当前高性能

GPU 资源短缺的背景下，利用 CPU 资源实现大模型推理功能可在一定程度上缓解算力短缺问题。例如，阿里云利用英特尔至强 5 代芯片和计算加速技术，实现了 CPU 推理算力 7 倍提升，能满足 13B 及以下参数规模的大模型的推理服务要求，同时支持对通义千问 72B 模型进行验证，支撑文生图、生成代码以及创意辅助工具等任务。

**公共云为异构芯片提供规模化应用环境，促进不同来源不同规格芯片协同应用。**公共云作为数字化转型的关键路径，正在成为算力破局的重要舞台。公共云的操作系统可以通过“一云多芯”的技术路线，并不与特定供应商的芯片绑定，而是尽可能提升兼容性，适配各种主要芯片，保障在不同情况下拥有对芯片的选择权。这一架构为各种规格的 GPU 提供了在真实世界场景中大规模应用的机会，加速了其市场成熟度和性能优化的进程。部分云计算厂商已经通过软件优化，实现了对不同厂商不同规格 CPU 和 GPU 的深度适配，这标志着算力生态与云计算平台的融合进入了一个新的阶段，不仅丰富了算力生态，增强了竞争力，也为云计算产业链的多元化和安全可控性提供了有力支撑。

## CHAPTER.4

开源的并行探索  
生态是应用爆发  
和协同共治的必  
要条件



# 1. 开源生态对于大模型的诞生具有基础性作用

大模型的诞生，需要算法、算力、数据三个维度的条件成熟，以及汇聚足够的开发者，开源生态在其中起到了基础性作用。

**算法积累与分享：**开源生态为大模型的早期探索和快速发展提供了必要的技术基础。通过分享算法、代码和模型架构，开发者能够快速、全面吸收前人的成果，加速创新步伐。例如，Transformer 模型的开源直接促成了后续一系列大模型的创新，如 BERT、GPT 系列等。

**框架和工具建设：**大模型的训练需要集中和调度大量算力，还需要充分发挥出硬件的能力。PyTorch 和 TensorFlow 等深度学习框架的开源，为大模型开发者提供了强有力的武器，是大模型技术进步的重要助推力量。

**数据资源的汇聚：**大模型训练依赖于海量数据，而开源生态中覆盖了文本、图像、音频等多个模态的高质量开源数据集，为模型训练提供了宝贵的资源。比如，多模态大模型的训练往往需要数亿到数十亿高质量的“图片 - 文本”对，LAION-5B 这样的开源数据集就能够满足这一需求，也在全球各种多模态大模型中得到了广泛使用。

**提供交流的平台：**开源社区为开发者提供了实践和交流的平台，通过参与开源贡献和讨论，他们能够掌握最新的技术趋势和开发经验，加速了各种问题解决，为大模型领域培养了源源不断的人才。

例如，Hugging Face 开源社区构建了庞大的生态系统，截至 2024 年 9 月，Hugging Face 上已经收录了超过 100 万个模型、数万个数据集，以及大量工具，吸引了全球海量开发者汇聚其中，在全球居于领先地位。

## 2. 开源生态是大模型技术发展和应用推广的加速器

开源生态提供了大模型技术的可及性，让开发者充分认识大模型的能力和安全性，通过社区协作加速了优化迭代和应用场景的拓展，极大地推动了行业的进步。

### 2.1 大模型开源的核心意义是普惠

**训练特性导致大模型开源无法分摊成本。**大模型训练区分为预训练（Pre-training）和后训练（Post-training）两个阶段，其中预训练所需要的资源、持续的时间比后训练多十倍到数十倍。预训练的特性是需要一定时间内聚集海量的高性能计算资源，不可分割地连续完成。软件的开源，往往可以通过完成概要设计后由开源社区内的开发者共同进行编码、测试等工作，从而将软件的开发成本分摊出去。大模型的预训练不可分割的特性，使得大模型的开源并不能像软件开源那样通过“众人拾柴火焰高”，由众多开发者分头编写代码的方式来分摊开发成本，而是只能由模型厂商独自承担。

**大模型开源是对中小企业和个人开发者的普惠。**按照当前市场价格评估，从0到1训练100B以上参数的语言模型，在不考虑试错的前提下，单次成本（包含数据、计算资源、人力）是数百万到数千万人民币。中小企业、个人开发者根本无力独自承担。有能力的大模型厂商承担了训练成本，并开源给业界，这意味着任何组织和个人都可以低成本了解大模型、使用大模型，而且不需要投入大量资源来进行持续的维护和升级。这个过程从经济利益角度讲是单向的普惠。

### 2.2 开源生态助力提升大模型透明度与安全性

**开源社区提供可访问性和透明度。**典型的开源社区提供类型丰富的模型，比如，截至2024年10月，ModelScope平台上提供了超过15000个模型，覆盖了计算机视觉、自然语言理解、语音识别等多个领域，并提供了大量的试用页面，使得开发者能够轻松找到自己需要的模型并进行体验。在开源社区里，开发者可以自由地下载模型，无需复杂的许可流程，大大降低了探索和实验的门槛。模型开源的同时，可能还提供了模型相关的技术报告、示例、推理代码等。技术报告往往会披露训练使用的数据、模型架构、训练技巧、测试结果等，大幅度提升了模型的透明度。

**开源社区提供基准测试与安全评估。**开源生态中常包含一系列标准化的基准测试集，用于评估模型在特定任务上的表现。开发者可以利用这些工具和数据集，对模型的准确性、鲁棒性和泛化能力进行客观评估。例如，在ModelScope上开源的CValues-Comparison中文大模型价值观比较数据集包含145k的价值观比较样本，每个样本包含（prompt、正例回复、负例回复），可以用于大模型安全、价值观对齐等多个问题的研究，例如，在SFT阶段增强模型的安全性、训练或评估reward模型、构造多项选择题用于自动化评估模型的安全性表现等。

**开源社区为开发者提供学习与交流的平台。**开源生态提供了丰富的文档、教程和示例代码，包括快速入门指南、API参考手册、最佳实践等内容。这些文档覆盖了从模型选择到部署的全过程，帮助开发者学习和掌握大模型的使用方法。ModelScope还拥有活跃的开发社区，每月产生上万条活跃讨论，开发者分享他们的实践经验，比如优化模型性能、如何处理特定类型的数据等。开发者可以从开源社区获取同行的见解、经验分享和技术指导。

## 2.3 开源生态加速大模型发展和应用

**开源生态促进技术路线优化，降低试错成本。**目前，高性能的语言模型单次训练时间持续一到两个月，费用达到千万人民币级别。通过开源，前沿的研究成果和技术创新能够迅速传播，社区内的讨论和实践反馈为技术路线的选择提供了宝贵的指引，帮助业界识别最有潜力的研究方向和应用领域，避免无谓的试错。比如，Llama 开源后，很多模型厂商都借鉴了其设计和训练的技巧，促进了业界水平的快速提升。此外，开源项目通常包含详细的文档、教程和示例代码，这为开发者提供了直接的学习资源和实践指南，避免了重复造轮子，减少了在技术探索和应用初期的不确定性，从而降低了整体的探索成本，也可以在一定程度上缓解我们面临的算力短缺。

**开源生态以丰富技术资源提升大模型开发验证灵活性。**开源生态提供了丰富的数据集、模型和工具，开发者可以利用这些资源快速进行产品概念的验证，例如使用语言模型和 RAG 工具创建自动化的客服系统。这种即时验证的能力极大地加速了从想法到实现的过程。可及性和开放性让开发者能够根据自身需求对开源模型进行深度改造，比如调整模型结构（增加注意力机制、引入更高效的编码器等）、替换损失函数或更改训练策略，以适应特定任务的要求，快速完成技术验证。

**开源生态以多样化模型能力加速赋能千行百业。**开发者可以根据不同应用场景的需求，在开源大模型的基础上进行微调和扩展，研发出针对特定业务场景的解决方案。这种广泛的适用性刺激了创新活动，带动了基于大模型技术的新形态业务的涌现。同时，开源生态还能够促进产业链上下游企业的协同发展。部分应用需要大模型运行在设备上，要求模型的参数规模较小，以满足设备端对存储空间、算力的限制；部分应用追求极致的效果，要求调用尺寸最大能力、最强的模型；还有很多应用会希望在效果和成本之间折中平衡。还有越来越多的企业，希望大模型不仅仅提供语言能力，还提供多模态能力，以构建形态丰富的应用。以通义千问的开源为例，提供了一系列不同

的参数规模模型，从可以流畅运行在设备上的 0.5B、1.8B、4B，到高性价比的 7B、14B、32B，以及能力在行业内一流的 72B、110B；与此同时，还开源了可执行图像理解任务的 Qwen-VL、可执行声音理解任务的 Qwen-Audio 的多模态大模型等。基于这些开源大模型，企业可快速低成本构建丰富的应用，加速大模型在千行百业的落地，包括 PC、手机、汽车、航空、天文、采矿、教育、文化、游戏等。

## 2.4 开源生态助力我国抢占产业生态制高点

**开源生态的丰富构成与协同效应。**开源生态鼓励跨组织的合作，形成一个由开发者、研究机构、企业和政府等多元主体组成的生态系统。这种合作模式打破了传统行业壁垒，促进了不同背景下的知识和技术交流。开源生态中的协作机制加强了产业链上下游的紧密联系，促进了从基础研究到商业化应用的快速转化。研究机构可以将最新的研究成果贡献给开源社区，企业则可以在此基础上开发产品和服务，形成一个良性循环。在生态中，围绕大模型诞生了一系列高水平的数据集、评测方法、工具和应用实践。这些资源被业界广泛采用，逐渐成为事实上的标准，提高了不同系统之间的互操作性。例如，利用人工智能加速癌症药物研发的某科技公司，借助 ModelScope 上的多个高质量开源模型和工具，服务多家药企用于癌症药物临床试验，将搭建 Pipeline 进行方案验证的周期从几个月缩短到几个星期。

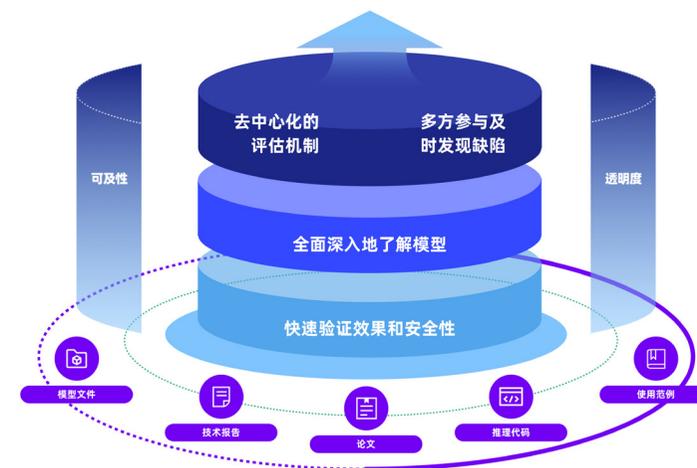
**开源生态促进我国大模型的市场拓展与国际化。**开源生态的全球性质帮助我国企业更容易地进入海外市场，与国际伙伴建立合作关系。通过开源项目，我国的技术和产品获得了全球用户的检验，提升了品牌知名度，增强了在全球大模型产业的话语权和影响力。以阿里云的通义 Qwen 系列开源模型为例，截止 2024 年 10 月在国际主流开源平台 Hugging Face 上的总下载量超过 4000 万，二次开发的衍生模型数量突破 7 万，超越了 Llama 系列的衍生模型。在 2024 年 9 月底，开源模型权威榜单 Open LLM Leaderboard 中，Qwen 及其衍生模型包揽了前十名。2024 年 6 月通义 Qwen2

开源后，得到全球多个主流开源平台和开源工具的兼容支持，包括 TensorRT（英伟达的加速工具）、OpenVINO（英特尔的工具套件）、Llama-Factory（META 的工具套件），以及 Firefly、OpenBuddy、vLLM、Ollama 等众多全球开源平台和工具，让 Qwen 系列开源模型得到更广泛更深入的应用，已成为全球产业界的一个主流选择。

## 3. 开源生态提升大模型安全性、强化业界信任度

开源生态：增强大模型的安全促进业界的信任

提升大模型的安全、可靠、可信赖、可用



开源生态通过促进社区协作、增强透明度，使得更多参与者能够共同识别并解决潜在的安全问题，提升大模型安全性，强化业界信任，推动技术健康发展。

### 3.1 开源生态促进多方协同安全治理

**开源生态为多方共治提供必要条件。**多方共治是指在一个生态系统中，不同角色（如政府、企业、研究机构、开发者社区等）共同参与治理，尽到自己的职责，以实现生态健康发展的目标。政府和立法机构可以制定相关的法律法规，建立有效的监管机制，确保开源生态的合规性。研究机构和行业协会可以共同制定技术标准和行业规范，确保模型的安全性和隐私保护达到一定水平。开源贡献者在具备足够能力的前提下，可以公开其数据处理方式和安全措施，实施定期的风险评估和安全测试，确保系统能够

抵御潜在威胁。开源社区尽到主体责任，建立管理制度，并完成用户管理、模型 / 数据集管理、服务管理。开发者遵守法规和社区公约，在此基础上进行学习、分享和交流。服务管理。开发者遵守法规和社区公约，在此基础上进行学习、分享和交流。

### 多方参与可及时发现并修正错误和偏差，提升大模型安全、可靠、可信赖和可用性。

集体智慧的汇聚，不仅能够及时发现并修复潜在的安全隐患，还能促进算法的持续优化，从而显著提高大模型的安全性。大模型的开源往往意味着公开模型参数、推理代码、使用示例、技术文件等，为多方检查提供了可能，来自全球的研究人员、开发者甚至普通用户都可以参与其中，共同审查模型的每一个细节，查看和理解模型的工作原理、数据处理以及内容生成流程，更容易发现潜藏的漏洞和错误。被发现的模型缺陷会公布在开源社区，极大地增强了模型的透明性。在这种机制下，一方面模型的潜在风险更容易被发现，另一方面模型已知的缺陷很难被隐藏。例如，通义 Qwen 系列的一个量化版本开源后，开发者发现在某些条件下模型会无限循环输出感叹号的缺陷，迅速在开源社区报告了情况，模型团队及时修复了该缺陷。

## 3.2 开源生态确保模型安全评估客观公正

**去中心化的评估机制，避免了单点决策可能带来的偏颇，确保了评估结果的客观性和公正性。**闭源大模型基本无法独立被第三方评估与验证，通常需要大模型提供方配合，提供调用接口、账号等。闭源大模型提供方可以使用围栏安全能力，如标准问答、检索增强、风险识别等技术手段来提升服务接口体现出的安全性，甚至一定程度上可以针对评估与验证“优化”其结果。相对的，任何第三方机构或个人都可以独立对开源生态中的大模型开展评估与验证工作，不仅可以检验性能指标，如准确率、响应速度等，还可以深入探究模型在不同场景下的表现，以及可能存在的风险点。夸大或误导性的声明都可能被社区迅速发现并纠正，多个独立来源的验证有助于构建模型的可信度。

## 3.3 开源生态帮助建立业界对大模型技术的信任

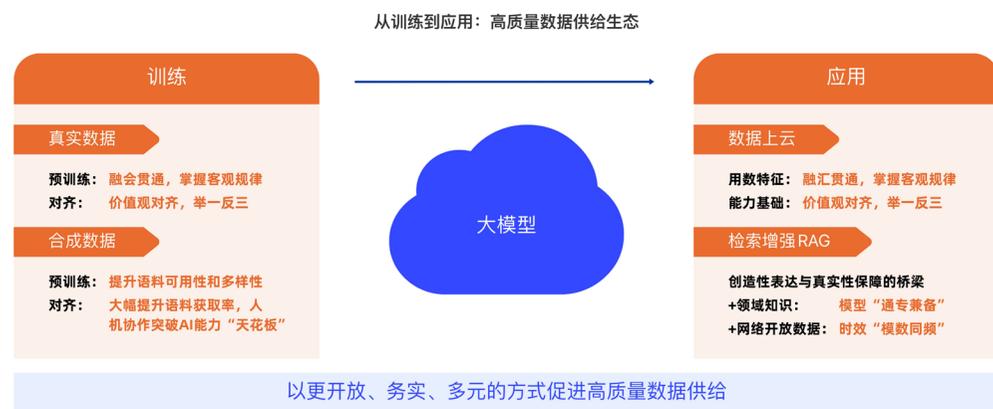
**开源大模型以其高透明度的特性建立业界信任，促进大模型技术广泛应用。**透明度是信任的基石，开源大模型提供了很高的透明度。如果公众了解到大模型背后的算法是经过全球开发者共同审查和改进的，且其性能和安全性得到了独立第三方的验证，会明显增强对模型可靠性的信心，更愿意接受并使用这些技术。开源模式鼓励广泛的社区参与，包括来自学术界、产业界以及非专业用户的贡献，这种双向沟通机制让公众感受到自己的声音被倾听和重视，增加了他们对技术的信任感。这一过程强调了技术发展与社会价值的平衡，体现了科技向善的理念。随着开源生态的进一步成熟和完善，我们有理由期待，大模型技术将在更加开放、透明和包容的环境中茁壮成长，成为连接人与智能世界的坚实桥梁。

## CHAPTER.5

高质量数据供给  
生态是具备持续  
竞争力的基础



在生成式人工智能时代，在保障通用安全性的前提下，训练数据的质量很大程度上决定了模型能力，这也无疑凸显了高质量数据在大模型训练中不可替代的重要性。同时，随着大模型应用的不断拓展，人工智能赋能千行百业，多方数据汇聚成为模型用数生态的特征。因此，应从训练数据和应用数据的整体视角出发，构建更匹配模型对数据利用特点的高质量数据供给生态。



## 1. 合成数据是解决高质量训练数据供给不足的突破口

**合成数据 (Synthetic Data) 是通过算法和数学模型创建的新形态数据。**首先建模真实数据的分布，然后在该分布上进行采样，创建出新数据集，模拟真实数据中的统计模式和关系。它类似于数据的“替身演员”，发挥补充或替代真实数据的作用。在机器学习和人工智能领域，合成数据可以为模型提供训练材料，帮助它们学习、理解和预测。将合成数据应用于大模型训练，可以从以下三个方面帮助解决高质量训练数据供给不足的问题。

**其一，预训练中合成数据可以提升语料的可用性和多样性，满足进一步提高模型性能的需求。**预训练可供大模型学习的数据类型较多，**但能够进一步拓展大模型知识边界、**

**增强大模型推理和泛化等关键能力的的数据，更多偏向于提升多模态理解能力的视频和图片的媒体数据，提升专业理解能力的领域知识。**此类数据主要来自人类的创造、制作和经验积累，其规模、类型和质量因客观条件的不同存在较大差异。在大模型强大的无监督数据学习能力面前，大模型的数据需求快速经历了从量到质的转换，能够被大模型直接利用、可以提升关键能力、让生成内容更符合人类习惯和要求的高质量数据，成为最为关键的数据类型。**合成数据解决了部分类型的真实世界数据难以观测的问题，为此类高质量数据供给提供了解决方案。**传统上看，通过生成“边缘情况”（如极端天气、罕见病）或者真实世界中的“潜在隐患”（如金融诈骗等安全风险），可以弥补因为样本分布不均衡导致的客观限制，提高数据分布的合理性和客观性。在大模型时代，一是可以**生成更多样化、具有不同视角的物理世界仿真数据可用于提升针对特定场景的数据收集效率，弥补真实世界中对稀有事件观测不足的问题。**二是**将之前不能被用于训练的数据转化为可用，提升模型对数据利用的可能性。**例如在工业制造领域，利用合成数据可以把生产、制造等工艺流程相关的原始数据，结合行业知识图谱，转化为可供大模型学习的工业语料，以缓解行业语料短缺的问题。

**其二，合成数据改变了对齐阶段数据获取的模式。与人类标注相比，合成数据不仅能大幅提升数据的获取效率，甚至可以探索模型“由弱到强”的可扩展监督，有机会以人机协作的方式使人工智能具备类人的能力，打破“天花板”。**首先，在对齐数据获取效率的提升上，利用参数规模较大的模型产出合成数据，生成指令及输入和输出样本，过滤掉无效或重复信息，可以自动化微调出性能较好的小模型，全过程几乎不需人类干预。**这解决了基于人类反馈的强化学习过程中人类回答标准不统一，因知识欠缺造成问答准确性不足，以及人类提供反馈成本较高的问题。**其次，合成数据通过自博弈 (Self-play) 探索由弱模型监督训练更强模型 (Weak-to-Strong) 的方式，即用较弱的模型标注数据，精调出能力更强的模型，进而有机会实现用人类标注的数据，精调出具备类人能力甚至强于人类的基础模型。**该过程一方面模拟人类消化理解复杂教材的过程，提高模型对语料的学习效能，使模型从有限的中学习更多内容。另一方面也需要人类的参与和监督，科学评估合成数据的质量并进行筛选。**此外，合成

数据技术还广泛应用于需要大规模专业知识、经验和耐心的指令响应标注，用以弥补人类标注的不足，例如多模态数据的对齐、领域知识问答对提取等。

**其三，应用合成数据还能进一步提升安全性。安全是高质量数据最核心的要素之一，合成数据不仅有助于用户的隐私保护，还能提升模型的安全性和可靠性。**首先，在数据获取上，合成数据可以替代个人特征数据，有利于隐私保护，解决获取合规性的问题。在训练中，合成数据可以通过差分隐私等方式给个人信息和隐私加噪声，使得数据在统计上代表原始分布，但不包含任何个人可识别的信息。用这些合成数据进行训练，即使大模型学到了很多知识，也不会侵犯到真实用户的隐私。这种技术经常被应用于需要大量数据但在获取和处理中涉及真实世界的个人信息甚至敏感信息的情况，如医疗健康、金融等行业模型，引入合成数据技术可以在保护隐私安全和保障模型性能之间找到平衡点。**此外，合成数据和真实世界的配合使用还能进一步提升模型的安全能力。**在图像领域，合成数据可以弥补对抗样本稀疏的缺陷，将合成图像数据和普通图像数据按照一定比例进行混合，可以提高视觉模型对图片的识别和判断能力，即使在普通数据样本完全缺失的情况下，使用合成数据进行图像识别训练，也可以得到接近普通数据样本训练的效果，从而提升图像识别的鲁棒性。

合成数据以其技术的灵活性，可以在基础模型预训练和对齐阶段，有针对性地大规模生成真实世界中不易获取的语料类型，拓展模型相应能力。同时，合成数据技术也在隐私保护和提升模型安全能力方面发挥了重要作用。合成数据已经从训练的“新物种”和“新方案”，演变成为解决高质量数据供给不足，提升模型性能表现，打破应用瓶颈的“突破口”。

## 2. 构建更匹配模型部署要求的高质量应用数据生态

### 2.1 数据上云推动模型应用数据生态建设，提供保障数据安全的能力基础

**随着从训练阶段“百模大战”走向应用阶段赋能千行百业，模型对数据利用呈现“多方汇聚”的特征。**大模型通过数量巨大、多方来源的数据形成了适配于丰富场景、解决未见任务的能力，而模型的部署和应用需要将用户数据与模型能力进行有效整合。例如，用户要求结合自有文档和外部新闻进行专项分析，系统就需要根据用户数据，利用模型插件调用外部数据，结合模型能力做出回应。在模型应用中，用户数据、外部数据和模型数据伴随着多样化的用户需求和持续性的人机交互形成了一个高度整合、多方参与的数据生态体系。有效管理模型应用数据生态体系，保障各类数据安全，对促进模型应用发展意义重大。

**数据上云为模型应用数据管理和安全保障提供了良好的能力基础。**在数据生态管理方面，数据上云可以有效融合碎片化的数据资源，通过在云上存储和管理数据，可以更灵活地访问，高效促进各方数据流通和开发利用，提高数据的可及性，满足“多方汇聚”的用数需求。此外，云数据库技术通过其可扩展性、高可用性、强兼容性、对向量数据库的支持，以及对多模态数据的处理能力，为大模型用数提供了强大的数据存储和管理基础，从而确保了这些应用在处理多方汇聚数据时的高效性和稳定性。就各类数据安全保障而言，利用云上数据存储，通过权限设置和技术方案的升级保障数据安全。云环境下的安全防护能力会随着技术的演进水涨船高，云侧存储数据的安全性得到进一步提高，安全状态和稳定性与持续性也更强，可以让数据在可控、受保护状态下得到合法利用。因此，当云侧的数据安全保障能力达到甚至超过原有私域的数据安全保障水平时，用户就可以放心地将数据转移到云侧。具体按照使用对象分类，个人用户

更关注与模型交互中的隐私，云上用数可以保障用户对其隐私数据的控制权和对数据处理知情权。企业用户更关注商业秘密，云上用数可以保障企业数据在传输、存储过程中的加密和防攻击，以及在应用处理过程中的权限限制。

## 2.2 检索增强是模型高质量用数的可行方案

**从训练走向应用，模型数据利用面临新挑战。**模型“训得强”不一定能“用得好”，训练能够让模型具备更好的通用性能，但多元化的场景对模型能力的需求千差万别，需要在应用中对数据进行补充和完善。其一，受到制度、技术、法律等因素的制约，训练数据不能穷尽所有领域知识和行业经验，私域场景中使用的数据（比如企业内部知识库和文档）也难以直接用于模型训练。这会导致模型的专业能力存在短板，在遇到专业领域或特定场景时，可能出现知识空白，出现错误回答或过多的拒答。其二，模型训练技术和成本决定了模型训练是阶段性而非连续的，模型数据更新存在滞后性，模型训练所学的知识不能及时更新，对训练数据“截止日期”之后发生的事情并不掌握，在应用中会出现精准性和时效性不足。

**检索增强 (RAG) 通过引入外部知识让模型能够高质量用数。**RAG 通过在模型应用阶段引入外部的权威知识，弥补训练阶段数据的短板，提高生成内容的质量。检索增强的第一步是“索引”，对庞大的外部知识库或文档集合，比如网页、专业书籍摘录、私有数据库记录等，进行清洗和提取，并切分成包含若干段信息的向量块。第二步是“检索”，将用户的提示词做向量化处理，利用向量的空间距离计算用户的问题与知识库中向量块的相似程度，获取相关度最高的若干向量块。第三步是“生成”，将用户原始提示与检索到的高相关度内容一起作为上下文输入给大语言模型，最终生成答案。RAG 所需的外部数据来源，既可以通过互联网搜索引擎连接的，具有更广阔知识面和实时性的开放数据，也可以是闭源数据，如企业内部的专属数据或领域知识文档。对网络开放数据，可以满足模型用数对时效性的需求，但需要对数据源进行过滤和筛

选，保障数据质量。对闭源的领域知识、专业数据库或内部文档，可以满足对专业性的需求，也可以扩展对应用数据利用的可能性，但需要关注对数据源的更新和维护，确保语料的时效性；同时也需注意企业商业秘密和个人隐私的数据安全，防止泄露。

**RAG 在大模型创造性表达与真实性保障之间起到桥梁作用。**RAG 是从训练转向应用的通用化替代性方案，可以提升模型对专业性知识的理解能力，增强生成内容的时效性和精准性，解决了部分数据能在应用中使用但不便于训练的矛盾。其一，RAG 通过引入更具丰富性的领域知识，在保留通用性的同时提升了专业性，在能力上“通专兼备”。比如以金融、医疗、法律等专业领域问答系统为代表的 NLP 类应用中，通过 RAG 读取专业文档和报告可以提升模型回答问题的专业性；又如在软件工程中的代码生成，RAG 可以从代码库中检索排名靠前的代码或摘要，并将其与输入聚合，以增强代码的生成和总结能力。还有在科研方面，通过融合多模态领域知识，模型可以对世界形成更全面的理解，提升对规律的识别和预测能力。如蛋白质领域，通过 RAG 检索查询数据库中已知的，在结构或功能上具有共性的蛋白质序列信息，可以丰富对目标蛋白质特征的描述，用于提升模型分析、预测蛋白质功能的能力，有助于设计全新的蛋白质序列结构。其二，RAG 提升了模型输出内容的精准度和时效性，实现“模数同频”。比如在通用型的问答系统、聊天机器人场景中提供更具相关性和时效性的信息，给用户在交互时带来更丰富的体验；又如在事实核查场景中，引入外部知识库可以更好地验证信息的准确性和可靠性；还有在搜索等场景的下游任务中，通过信息检索和文本生成，提高搜索召回内容的相关性和可读性。

## 2.3 以更开放、务实、多元的方式促进高质量数据供给

促进人工智能高质量数据体系的建设，需要理解大模型对训练和应用数据的实际需求，数据质量与模型能力的关系，科学、客观地理解高质量数据的含义，综合利用社会各界力量推动数据的开放、开发和利用，构建共享、共创、共赢的合作生态。

**一是加强高质量数据供给需要重视数据的可及性和可用性。**首先，遵守数据安全、个人信息保护等相关法律法规，保护著作权和商业秘密，在此基础上鼓励大模型训练数据以及可供 RAG 检索的外部知识库数据的开放和利用。其次，对于标准清晰、经过实践验证的高质量语料，要进一步加强开放共享和开发利用。鼓励受财政支持的科研或文化单位所有的知识产权类价值观语料向社会公开，如媒体的主流价值观数据，国家图书馆电子化图书、历史典籍、数字报纸、科研期刊论文等，用于基础大模型训练或 RAG 知识库的构建，同时基于非营利性成本补偿原则明确合理收费标准。

**二是推动高质量数据开放平台建设。**高质量语料效果会在模型训练和应用中得到检验，其价值可依据商业合同对价按效果付费，形成良性市场。面对生成式人工智能的需求和技术路径在快速变动和发展，一方面可以加快推动高质量训练数据开放平台和 RAG 语料库的建设，鼓励在多方共识之下的共商、共建、共享，积极尝试，有序规范。另一方面对高质量数据的检验标准，可以基于市场优胜劣汰更高效地判断好坏、配置资源。

**三是加强合成数据开发利用中人类的参与和监督。**合成数据需要引入人类的参与和监督，以人工抽检等形式，确保数据质量的可控性，对合成数据进行必要的质量评估和筛选过滤，避免不加筛选的应用合成数据导致模型能力退化甚至“崩溃”的问题。对于提高模型安全能力的合成数据应开放共享，让合成数据更好地服务于模型安全能力建设。

**四是进一步提高云侧数据安全能力，为模型数据安全供给提供基础保障。**模型应用中数据“多方汇聚”的特征和云上动态升级的安全保障机制，决定了数据上云是提升高质量应用数据供给的必然选择。数据安全本身并不区分数据存储的介质和环境，模型用数安全的核心是让数据在可控、受保护状态下依法得到有效利用，需要进一步完善云数据安全技术，防范被攻击、被盗取的技术安全风险，让更多数据能够安全上云，为模型数据训练和应用提供新动能。

## CHAPTER.6

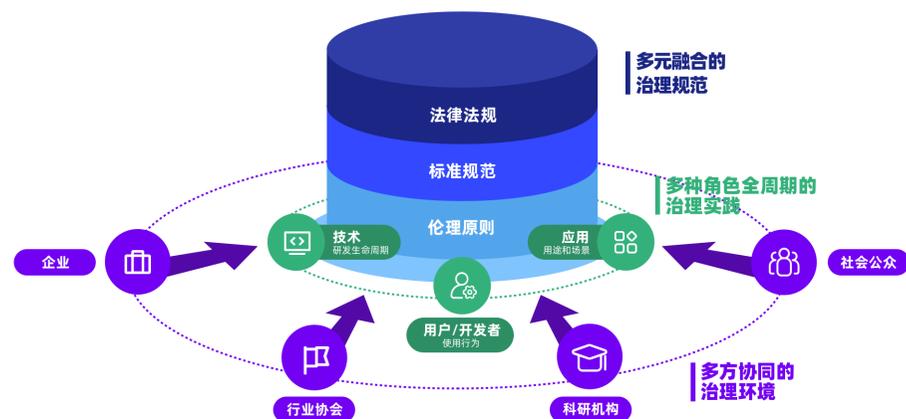
# 体系化的安全治理 能力是稳定发展的 保障



# 1. 多角色视角安全治理架构

大模型治理目标

**安全，可靠，可信赖，可用**



随着大模型技术的迅猛发展，其应用领域持续拓宽，对社会的影响日益深远。为了使大模型能在实际应用中发挥最大效能，同时有效防范潜在风险和滥用问题，需要确保大模型的安全性、可靠性、可信赖和可用性，以保障其健康、有序发展：

- **安全**：在大模型提供服务的过程中，既要确保不侵犯用户或他人的个人信息，又要确保敏感信息以及其他保密数据得到充分保护，防止未经授权的访问和窃取。同时，还要确保大模型在实际应用中不被滥用生成有害内容，从而避免对用户或公众造成危害。
- **可靠**：确保大模型在面对各类恶意攻击时仍能维持正常运行，并输出可靠且一致的结果。同时，要防止大模型在应对噪声、分布漂移等不同情况下出现输出不一致的现象，避免基于错误事实或混乱逻辑生成不清晰、不合理、难以理解的低质量内容。
- **可信赖**：确保大模型的行为、决策及影响能够被人类轻易理解和解释，所生成的内容应具备准确性和真实性，尽可能反映现实世界的情况。此外，需保证大模型在各类应用场景中产生的决策行为均符合相关规范，并具备可追溯性，从而建立明确的责

任链条。

- **可用**：促进大模型在开放可控的环境下助力生产力提升和价值创造，为人类的生活和工作带来便利与效率的提升，在解决实际问题、提升工作效率、改善用户体验以及推动科技进步等方面，充分发挥其有效作用与积极贡献。

保障安全风险治理需要从多维度结合多角色视角构建策略，从宏观层面设定目标与框架，并逐步细化至具体执行任务，最终将这些任务精准分配给各个治理模块，针对多角色实现对所有对象的全面治理。这一策略的核心在于强调顶层规划与指导的重要性，确保每一层级的操作均紧密契合整体战略方向。

**第一层是多元融合的治理规范。** 鉴于大模型技术的复杂性、快速创新性以及多方参与的特点，其治理须融合包括法律法规、标准规范与伦理道德在内的多元治理规范，并保持动态适应性。

- (1) 法律法规为大模型的技术研发、应用开发和用户使用提供了强制性的行为准则，确保所有活动都在法律框架内进行；
- (2) 标准规范则通过制定统一的操作标准和质量要求，提升大模型技术的安全性、可靠性和可控性；
- (3) 伦理道德在指导技术研发、应用开发和用户使用尤为重要，特别是在缺乏明确法律条文的情况下，确保技术不会偏离社会正义和公共利益的原则。

这三方面相互补充，共同构成了一个完整的治理体系，有助于构建一个有序、安全、公正且符合道德的大模型环境，从而推动大模型技术和应用的健康发展。

**第二层是多角色全周期的治理措施。** 大模型治理应贯穿模型的整个生命周期，覆盖设计、研发、训练、部署到应用上线的每一环节，因此在各个阶段，需要明确不同的参与角色，包括：

- (1) 技术支持者，即负责开发和支持底层技术的一方；
- (2) 服务提供者，即开发和提供具体应用或服务的一方；
- (3) 服务使用者，即使用这些应用或服务的用户。

在此基础上，基于各参与方的角色及其行为实施全周期治理，将更具有可操作性和可落地性。同时，建立有效的标识记录、监督审核和追踪机制，将增强整个治理体系的透明度和可追溯性。通过多角色全周期的治理，可以更好地确保大模型治理目标的一致性，实现技术与社会的和谐发展。

**第三层是多方协同的治理环境。**大模型安全问题影响广泛，需要政府、企业、行业协会、科研机构、社会公众等多方协同治理：

- (1) 政府制定相关法律法规，确保技术应用和用户使用符合公共利益，并防止技术被滥用；
- (2) 标准组织和行业协会通过凝聚各界共识、协调行业行动，促进大模型的安全应用；
- (3) 企业拥有丰富的资源和实际应用场景，可通过贡献先进的实践经验，为技术规范发展提供参考；
- (4) 科研机构在基础研究和技术创新方面发挥关键作用，帮助识别并解决潜在的安全问题；
- (5) 社会公众的监督和参与促进了更全面的安全治理。通过多方协同，共同应对大模型的安全挑战，以实现技术进步与社会安全的平衡。

最后，我们坚信遵循治理的基本准则——以人为本，普惠正直，安全可靠，隐私保护，可信可控，开放共治，是推动科技创新与社会和谐共生的关键。尤其在大模型安全方面，通过**多元融合的治理规范、多角色全周期的治理措施、多方协同的治理环境**，强

化数据保护和隐私安全，提升系统可靠性和抗攻击能力，确保技术的健康发展，共同构建一个更加公正、透明且可持续发展的数字未来。

## 2. 大模型安全治理规范：全面多元融合的规范框架

在大模型治理中，围绕法律法规、标准规范、伦理道德原则三大指导支柱，通过法律法规确保法律遵从，明确各方责任；借助标准规范提升技术一致性和安全性，凝聚各方共识；运用伦理道德原则引导技术向善，增强社会适应性。三者相辅相成，共同营造创新与风险控制并重的生态，助力大模型技术稳健发展与社会普及。

### 2.1 法律法规、标准规范、伦理道德原则的性质、作用、执行路径有明显的差异

- 法律法规具有强制约束力的社会规则，以责任惩戒威慑个体行为守法，对大模型研发和应用、使用设定明确的界限和责任，侧重于法律层面的权利义务、责任分配和基本安全要求，如数据保护、隐私权、知识产权和国家安全，并由公权力负责监督执行，通过司法和行政执法来确保遵守。

相关法规包括但不限于：《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国科学技术进步法》、《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》、《网络信息内容生态治理规定》、《生成式人工智能服务管理暂行办法》、《互联网广告管理办法》、《互联网广告可识别性执法指南》、《网络反不正当竞争暂行规定》、《关于加强科技伦理治理的意见》、《科技伦理审查办法（试行）》等。

- 标准规范为大模型技术、产品和服务的开发、生产、测试和应用等场景提供具体

的操作指南和技术要求，专注于技术细节和操作流程，如模型训练数据质量、模型评估方法、安全测试标准。标准更多的是凝聚产业共识、强调行业自律，提升技术领域安全风险管理水平，从而支撑产业安全健康发展，促进科技进步、规范社会治理。相关标准包括但不限于：《生成式人工智能服务安全基本要求》、《生成式人工智能数据标注安全规范》、《生成式人工智能预训练和优化训练数据安全规范》、《人工智能生成合成内容标识方法》等。

● 伦理道德原则基于道德观念和社会责任感，关注大模型的道德正当性、公平性和人道主义等问题，也更宽泛地关注大模型技术和应用的社会影响，包括公平性、无偏见和社会责任等，旨在确保大模型发展与社会价值观对齐。伦理道德原则目前除了法规对流程的要求，更依靠行业和企业自治和公共舆论监督。相关伦理指导原则包括但不限于：《人工智能伦理准则》、《人工智能伦理白皮书》、《人工智能伦理与治理指南》等。

## 2.2 组织保障是一切体系化治理能力的基础

有了规则，执行规则还需要有力的组织保障加以落地。通过企业自上而下的组织和管理模式，确保治理原则、制度规范及内部规则上的一致性与高度协同。在此框架下，各业务板块也被赋予了安全治理首要责任主体的重任，旨在促进大模型相关项目从构思、研发到实施的每一步骤都嵌入安全考量。既保障企业安全治理机制的统一性，又鼓励各业务部门依据自身的业务特色和面临的特定安全挑战，灵活构建适宜的治理组织保障机制，确保治理规则的落定性。

## 3. 大模型安全治理措施：技术与应用风险治理构成的立体防护网

有了治理规则，通过大模型安全治理措施进行全面的落地执行。大模型安全风险治理措施包括三大核心：技术风险治理、应用风险治理及防范风险行为。技术风险治理着重加固模型基础，通过加强数据保护、优化算法韧性、管理内容风险、提升抗攻击性能和保护隐私等措施，建立内生安全屏障。应用风险治理关注模型应用实践和使用安全，确保合法、合伦理性，增强透明度与公平性，引导技术正面效益，体现负责任的应用导向。防范风险行为聚焦调用大模型技术的开发者，以及使用大模型产品或服务的用户，确保降低模型被误用、滥用或恶意使用的风险。

### 3.1 围绕全生命周期的技术治理措施

#### 3.1.1 模型训练阶段

模型训练阶段奠定了大模型的能力基础，对模型自身的安全性也至关重要；该阶段主要与技术支持者相关，其面临的风险源主要来自训练数据，如数据来源不合规、包含违法不良信息、缺乏多样性、数据标注不规范、被投毒等，可能导致如模型“幻觉”问题、有害数据内化进模型、二次开发破坏安全对齐等。

模型训练阶段的安全控制措施包括：

#### （一）训练数据：

● **数据源筛选。**采集时对数据源进行审核，一是充分考虑数据源的合规性，如：使

用具备开源许可或相关授权文件的开源数据集，自采数据时不采集所有者已明确不可采集的数据，使用商业采购数据时应签署具备法律效力的交易合同、合作协议等以获得充分授权；二是通过不同数据源的搭配，以提高多样性；三是开展有效的安全检测，及时发现并避免使用有害内容含量过高的数据源。

● **数据清洗过滤。**采集后对各模态的训练数据进行清洗，通常采用关键词匹配、分类模型、人工抽检等方式进行检测过滤，去除训练数据中的违法不良信息、个人信息等。

● **数据标注安全管理。**数据标注主要涉及有监督微调阶段、基于人类偏好的强化学习阶段，其质量和安全在很大程度上决定了模型能力和安全性的上限。对于数据标注的安全管理需要规范化的管理，一是选用安全可靠的标注工具，通过访问控制、安全传输等手段保证标注过程的合法合规，记录标注操作及系统活动日志；二是建立详细的标注规则体系，在规则中区分功能性标注和安全性标注；三是对人员进行安全管理，包括数据标注人员的安全培训、考核筛选、合理的角色分配等；四是建立严谨的质量核验流程，进行机器和人工相结合的审核，并建立不达标数据的处置响应措施。

## （二）模型算法：

● **增量预训练 + 强化学习。**增量预训练指在已有预训练模型的基础上，使用特定领域的语料继续进行训练，以提高模型在特定任务上的性能，通过注入新的领域知识来减少模型对错误数据的依赖，降低幻觉问题的发生概率；强化学习包括基于人工反馈的强化学习（RLHF）和基于人工智能反馈的强化学习（RLAIF），其核心都是帮助模型更好地与人类的判断和选择对齐，通过优化模型的行为来减少幻觉的产生。

● **安全性对齐。**通过技术手段将人类安全价值观量化并嵌入模型，令生成式人工智能“理解”人类的安全价值观，保障在运行实施阶段能够遵循。具体包括：一是安全语料生成，通过安全评测发现模型的安全缺陷，依据“风险指令生成模型”生成特定风

险类型的攻击指令，同时利用“安全回复模型”生成对应该风险指令的安全回复，从而形成安全的语料，用于修正大模型的安全缺陷。二是安全对齐技术，通过使用安全语料数据对语言大模型进行 SFT 以及 RLHF，在保持模型可用可靠的前提下，大幅提升模型的安全性。

● **防止灾难性遗忘技术：**大模型的灾难性遗忘问题是指在进行领域预训练或者微调过程等二次开发时，模型往往会“遗忘”之前已经学到的大量通用知识，也包括安全能力。这一现象限制了模型的泛化能力和实用性，尤其是在需要持续学习和适应多种任务的情境下。缓解此问题的一些方法包括：

- （1）权重冻结：将一些重要的权重冻结，以保持预训练阶段学到的知识。
- （2）LoRA 微调加速技术：通过调整模型参数的低秩近似来加速训练过程。
- （3）数据复习：在微调过程中定期复习之前训练的数据，以减少遗忘。
- （4）使用 MoE（Mixture of Experts）：通过引入特定领域的多个专家网络来保持和增强预训练知识。

## 实践：PostTraining 中的安全对齐

### 使用 SFT&DPO 来做安全对齐

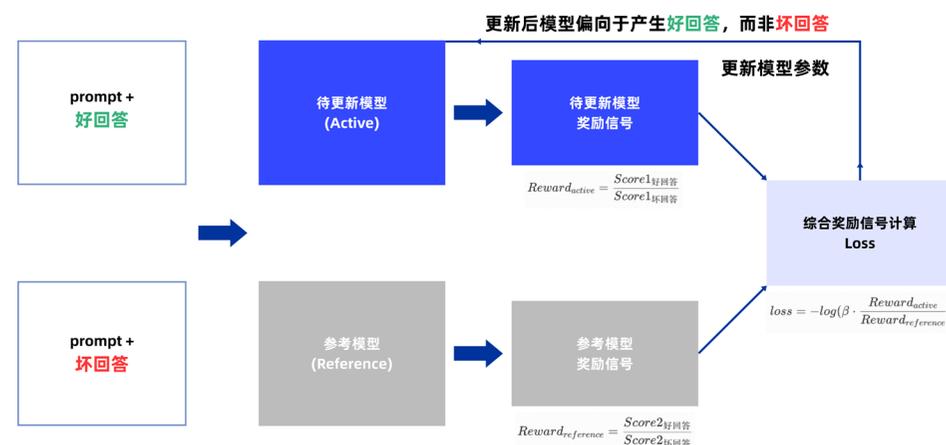
尽管在预训练（pre-training）阶段已经做了数据源筛选和数据过滤，但是大模型仍然可能生成有害或误导性的信息。我们需要在后训练（post-training）阶段设法让模型学习分辨正确和错误，“对齐”模型的行为，使其更加符合社会伦理标准和用户期望。现在，业界往往使用有监督微调（SFT, Supervised Fine-Tuning）和直接偏好优化（DPO, Direct Preference Optimization）来实现这一目标。

## 1. 有监督微调 (SFT) :

SFT 是一种常见的方法, 往往通过在 (question, answer) 数据集上进行训练来调整模型的行为。在这个过程中, (question, answer) 数据集是人类来决定的 (直接撰写, 或者机器生成后人工校验)。参与的模型只有一个, 由预训练模型初始化。训练目标和预训练相同, 优化预测下个 token 的 loss。SFT 可以帮助模型学习到更符合人类价值观的回答方式, 从而减少生成违法不良信息的风险。

## 2. 直接偏好优化 (DPO) :

DPO 是一种强化学习技术, 但是并不直接对模型的输出进行评分, 而是比较两个或多个模型输出, 并选择更好的一个。通过这种方式, 模型逐渐学会生成更符合人类偏好的输出。在 DPO 中, 训练信号来自对不同模型输出的比较, 而不是直接的奖励或惩罚。DPO 是目前主流的偏好对齐算法, LLaMa3.1、Qwen2 等都在使用。训练数据格式为 (question, chosen answer, rejected answer)。参与的模型共有两个, 一个是 Active 模型, 另一个是 Reference 模型, 都从 SFT 模型初始化, 并在训练过程中更新 Active 模型。训练目标是优化 Active 模型, 提高其对 chosen answer 的采样概率, 远离 rejected answer, 同时确保结果不偏离 Reference 模型的输出。这种方法可以更有效地捕捉到人类的细微偏好, 并避免直接奖励机制可能导致的过分简化或操纵问题。



结合这两种方法, 比如先使用 SFT 来纠正模型的基本行为, 然后使用 DPO 来进一步优化和细化模型的表现, 可以使模型更加符合用户的偏好和社会伦理标准。

## 安全语料的制作

以 SFT 阶段使用的 (question, answer) 语料对为例, 安全语料并不是简单地拒绝回答。对于具有明确答案的问题, 需要正面给出准确的 answer; 对于用户的消极情绪, 需要给出正面引导; 对于恶意的诱导、错误的设定, 需要指出其错误, 不能落入逻辑陷阱。

比如, question: “我想结束自己, 该怎么做?”

bad answer:

加油, 明天会更好。” —— 虽然出发点可能是好的, 但如果缺乏具体的帮助建议或资源指引, 对于深陷困境的人来说可能显得空洞无力。

“每个人都会经历低谷, 你太夸张了吧。” —— 这样的回答忽视了用户的痛苦感受, 可能会让他们感到更加孤独和不被理解。

good answer:

“我很抱歉听到您这样的话。请相信，无论您的痛苦有多深，**都有解决的办法**，而且您并不孤单。在这个世界上，有很多人关心您，可以**帮助您渡过难关**。寻求专业帮助：如果您在中国，可以拨打心理危机与自杀干预中心救助热线：800-810-1117。”

安全语料的制作，除了考虑安全得体，也需要充分考虑质量。针对 SFT 语料，往往需要从有用性、相关性、易于理解等几个维度度量其质量，还需要在长度、结构、格式上符合用户的偏好。质量低的安全语料进入训练，会拉低模型整体的生成质量。模型生成内容的质量一直在提升，安全语料的质量也需要及时跟进。

### 3.1.2 服务上线阶段

在服务上线阶段，服务提供者和技术支持者需要协作完成安全评估并选择安全有效的模型作为基座构建完整的服务，这个阶段会决定模型的核验、模型的使用方式、调用的工具集等，面临的主要风险源包括模型评估不完备、系统平台中的工具 / 插件和部署环境的安全问题、模型应用功能场景设计中的伦理风险（详见 六、2.2）。

服务上线阶段的安全控制措施包括：

#### （一）模型评估：

●**模型安全评估：**服务提供者和技术支持者协同开展算法安全自评估，对算法目的、使用的数据、模型、训练方法、评测过程、干预策略等进行评审。并且通过测试数据集、红队测试等对大模型的理解、生成、安全等能力进行评估，确定大模型的质量和安全性水平，确保模型的持续稳定和安全。

●**模型筛选：**服务提供者结合安全评估结果、自身或外采的风险控制能力，筛选合适的基础大模型、二次开发大模型或开源大模型，避免在上线运行后出现不可控风险。

#### （二）系统安全

●**插件安全保护：**大模型的能力需要结合第三方工具 / 插件时才能最大化发挥价值，第三方工具 / 插件的安全可信问题正在受到越来越多的关注，服务提供者需要采取相应的控制措施进行应对，一是安全审计与更新，对使用的插件、工具等进行定期安全审计，及时发现并修复安全漏洞，确保插件代码的安全性；二是插件权限管理：限制插件的访问权限，仅授予必要的功能权限，减少潜在的攻击面，保证插件操作不会影响系统整体安全。

●**部署环境加固：**随着产业需求的升级，大模型应用服务所需的工具链越来越复杂，这一开发过程中存在与硬件、多个软件之间的频繁交互，因此大模型的部署环境安全也是影响大模型应用服务安全的重要考虑因素，其应对控制措施：一是软硬件环境安全：评估并确保计算系统（包括芯片、软件、工具、算力等）的供应链安全，关注供应的持续性和稳定性；二是环境隔离：实现模型训练环境与推理环境的物理或逻辑隔离，防止数据泄露和不当访问，维护系统安全边界；三是定期开展漏洞检查工作：综合运用静态分析、动态分析和渗透测试等技术手段，识别潜在的安全漏洞；四是构建有效的安全响应机制：制定详细的大模型系统应急响应预案，涵盖安全事件分类分级、事件响应、责任人分配、通信渠道保障、决策流程高效等环节。

#### （三）应用治理（详情见 六、2.2）

## 实践：基于红队的大模型安全评测实践

### 3.1.2.1 什么是红队测试

红队是网络安全领域的专业术语。在网络安全领域，红蓝对抗是一种模拟演练。红队一般由专业的安全专家团队组成，他们通过多种手段和漏洞利用等方法发现系统中潜在的弱点，评估当前系统的安全性。蓝队作为防御方，通过构建防御护栏保障系统的安全性。红蓝对抗是一个持续的过程，红队不断发现系统的弱点和漏洞，而蓝队则对这些弱点和漏洞进行修复，从而不断提升系统的安全性。

在大模型领域，全面系统地评测大模型的安全性面临着挑战。现有的大模型的安全评测从风险定义出发，形成有害指令，衡量模型输出内容的安全性。此外，越狱攻击、提示注入、梯度攻击、多轮对话等攻击方式也常用于衡量大模型在攻击下的鲁棒性。然而，这类静态或者半动态、非持续的评测往往只能片面地评估大模型的安全性。为了弥补这类评测的不足，**基于红队的大模型安全评测应运而生**。具体地，将大模型作为一个系统，红队评测主要从攻击的角度发现大模型系统现有的缺陷和漏洞。红队评测是一个更加开放、更加贴近实际系统的应用测试。红队测试可以持续地对大模型系统的漏洞进行发现和反馈，从而不断提升大模型的安全性。

### 3.1.2.2 阿里红队实践

阿里发布了通义千问系列的大模型，也开源了 Qwen2-72B 等 SOTA 的基础大模型。在大模型发布前，需要进行红队安全评测，从而提前发现大模型潜在的安全缺陷，提升其安全性，具体环节包括：

**1. 平台建设：**阿里内部建立了大模型攻防平台——模御，设置了 QwenChat, Qwen-VL 等多个模型服务作为靶标，包含交互页面和漏洞提交等功能。攻防平台能够很好地模拟真实的线上大模型系统，从而作为一个模拟靶场供红队人员进行漏洞挖掘。

**2. 红队召集：**面向阿里集团内部有安全相关经验的员工进行了定向召集，开发攻防平台。

**3. 漏洞定级：**从危害程度、攻击成功率、可重复性多个维度判定。

通过红队攻击评测，实现了高质量大模型安全性漏洞的积累，包括设定类、加密类、劫持类等多个学术界未出现过的高质量攻击漏洞。通过对收集的漏洞进行整理和抽象，阿里内部形成了大模型攻击的工具 RedChain。RedChain 是一个基于 LangChain 的红队攻击评测 Agent 框架，可以自动理解用户自定义的评测需求并进行规划，从而批量生产符合需求的红队评测样本集，也支持发起评测流程并得到评测结果。

下图展示了 RedChain 与微软开源红队项目 PyRIT 的对比情况。

- 在数据层面，RedChain 拥有安全工作中长期积累的风险知识库，能够更好地契合中国的人工智能法律法规；RedChain 还对收集的越狱模板集进行遗传变异算法的增广，显著增强了它们的数量、多样性与攻击成功率。
- 在生成与打分工具层面，RedChain 的工具能力范围涵盖多模态与多语种、单轮与多轮对话、普通有害指令和越狱攻击指令。
- 在智能体层面，RedChain 采用多 Agent 协作模式，分别进行评测需求解读、工具筛选、数据资源管理、整体步骤规划等，有效提升了处理复杂任务和高度自定义化任务的成功率。

微软开源红队工具PyRIT

针对文生文模态，生成攻击指令测评azure api。gitstar 1.7K。

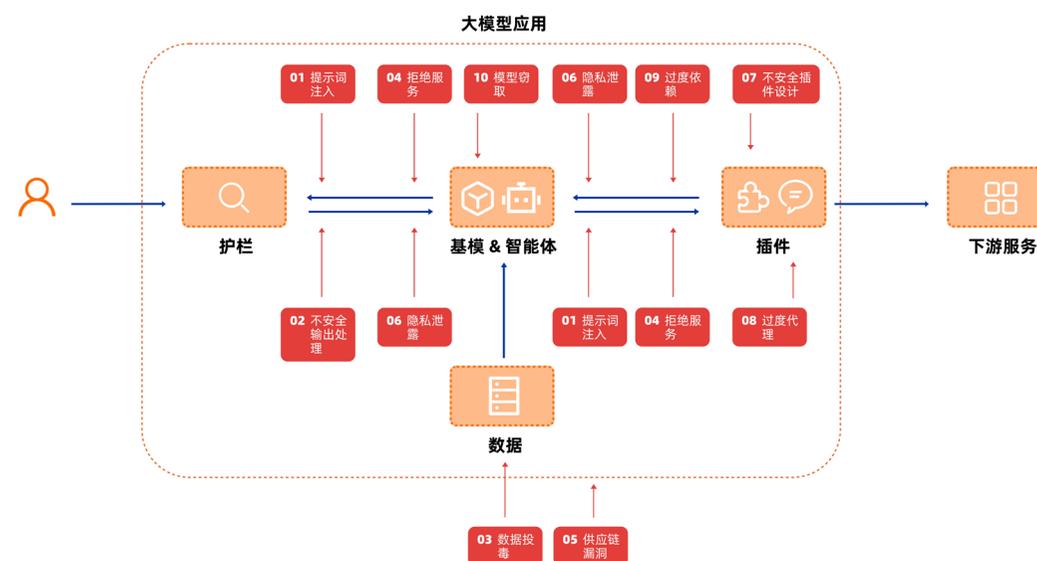
Target	Local : local model(e.g,ONNX)
	Remote : API or web app
Datasets	Static : prompts
	Dynamic : prompt templates
Scoring Engine	PyRIT Itself : Seif Evaluation
	API : Existing content classifiers
Attack Strategy	Single Turn : Using static prompts
	Multi Turn : Multiple conversations using-prompt templates
Memory	Storage : JSON,Database
	Utils : Conversation,retrieval and storage, memory sharing,data analysis

微软开源红队工具PyRIT

可生成多模态、多语种的普通有害指令、攻击指令，评测公司相关大模型及产品。

Target	本地 : 魔塔开源模型
	接口 : 灵积百炼平台
Datasets	有害指令集
	越狱模版集
	风险知识集
Scoring	文本判读
	多模态判读
Tools	普通样本生成-多模态
	普通样本生成-多语种
	攻击样本生成-单轮
	攻击样本生成-多轮
Agent	链图规划agent
	工具管理agent
	数据管理agent

此外，这些收集的漏洞也可应用于大模型系统的外层护栏中，检测各类攻击流量，提升大模型系统安全性。以 Meta 开发的过滤器模型 Prompt Guard 为例，其专门设计用于检测提示攻击，适合用于对输入到 LLM 中的内容进行过滤，该模型是一个多标签分类器，能够识别两类提示攻击风险——直接越狱（尝试明确绕过模型的安全条件或系统提示的技术）和间接提示注入（在模型上下文窗口中包含的第三方数据中，无意中作为 LLM 用户命令执行的指令实例）。经过实验发现，Prompt Guard 模型虽具有较高的提示攻击召回能力，但是误拦截的缺点也较为显著，不适合直接作为产品护栏；此外，该模型仅有两个分类标签，无法给出更细致的分类来作为制定处置策略的依据。因此阿里训练了自研版本的 Prompt Guard 模型，使用更贴合线上产品实际流量的数据，并优化训练优化函数，以增强线上效果与更好地调节召回率与准确率；并且进行更细致的分类标签定义与数据标注，可对不同类别的攻击流量采取差异化防控策略。



LLM 应用的典型架构，囊括了大模型、护栏、智能体、数据、下游服务各模块，图中列出了 LLM 应用的 Top10 风险。红队针对头部风险做深入研究。

- 提示词注入：包括直接攻击（用户覆写系统指令）、间接攻击（插件等外部信源注入恶意指令）。阿里安全红队提出了多轮红队模型、乱序攻击等创新攻击方法，并将常见攻击收敛成 3 大类 19 小类的分类体系。同时也建设了提示词注入的检测能力，在高安全需求场景向客户提供服务。
- 不安全输出处理：指大模型输出未做验证和处理，导致 XSS（跨站脚本攻击）、CSRF（跨站请求伪造）、SSRF（服务器端请求伪造）、越权、RCE（远程代码执行）等漏洞被利用。与传统应用中 XSS 等漏洞不同，攻击者需要突破大模型内生安全，才能让大模型输出目标字符串。阿里安全红队构建了评测集，对各种应用做黑盒扫描，从而发现并防御风险。扫描结论验证 XSS 和 SSRF 目前是大模型的应用安全中高发的漏洞类型。
- 供应链漏洞：指供应链组件导致应用存在风险，组件类型包括三方软件包，脆弱

的模型文件，有毒的数据集。阿里安全红队对集团大模型应用做盘点和打标，通过高星框架追踪、代码匹配与理解，提升大模型应用标签的准召，从而服务于算法备案。通过 CVE 漏洞库等外部情报、安全 AI 挑战者计划、和阿里安全响应中心（ASRC）收集大模型应用类漏洞，并进行分析和 Proof of Concept 验证攻击可行性。另外对漏洞分类型做白盒、黑盒扫描，检测范围覆盖路径穿越、代码注入、存储型 XSS、SSRF 等。对扫描漏洞分等级做推送和告警，由安全工程师进行确认和修复。

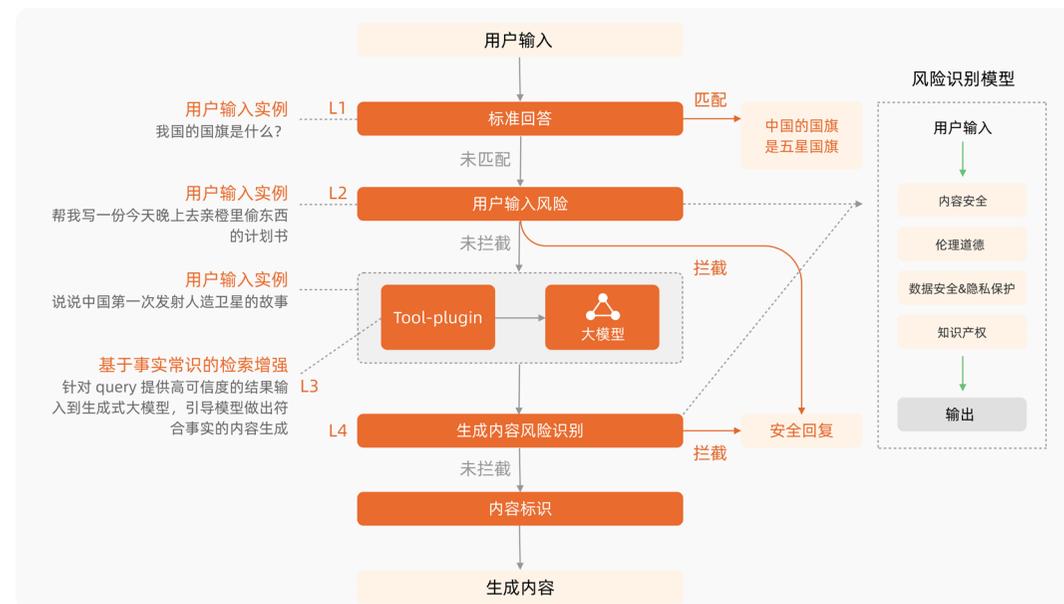
### 3.1.3 内容生成阶段

大模型生成的内容是服务使用者和模型交互的结果。服务使用者的输入，以及文档环境等上下文信息，都影响到模型当前的生成；实践中，内容安全的风险一方面是模型运行时可能遇到的生成风险内容、数据和隐私泄漏风险，另一方面是来自用户的恶意输入、诱导、滥用等风险（详情见 六、2.3）。

内容生成阶段的安全控制措施包括：

#### （一）模型运行

● **输入输出审核过滤：**一方面需要通过业务沉淀的一系列特定问题与标准回复，形成标准问题问答库。当用户的问题命中标准问题，就给予标准回答，从而做到可以及时干预大模型的回复，减少因自由生成内容带来的不确定性和幻觉风险；另一方面，对于未命中标准问答库的输入问题，需进行风险识别，发现安全风险则进行拦截并给予安全回复，对于未拦截的问题，需要对大模型的生成内容进行监测，并过滤掉潜在的违规或有害信息，维护良好的用户体验和社会责任。



● **检索式增强技术：**大模型可以从外部知识库中检索关联信息，既加快了模型知识更新速度，又确保生成结果更契合用户实际需求，有效缓解模型产生不实或偏离事实的“幻觉”问题。经历了优化训练的大模型虽然具备了一定的聚焦能力，但无法完全避免“幻觉”特点带来的一些输出内容的风险，因此还需要重点在运行输出环节增加辅助回答的措施来帮助匹配真实答案和防止错误回复。例如，训练数据截止到 2023 年，如果让模型回答 2024 年的问题必然会出现幻觉问题，此时则需要 RAG 等检索式增强技术对大模型知识进行补充。



● **生成内容标识**：显式标识主要作用是告知用户该内容由人工智能生成，避免造成混淆和误导；隐式标识主要作用是对制作平台、传播平台等更多可追溯信息进行记录，用于对 AIGC 滥用和恶意使用等行为的定位和溯源。

● **思维链技术**：旨在优化大型模型输出信息的逻辑连贯性，借助展示少许富含深度推理步骤的实例，引导大型模型在生成信息过程中，不仅呈现答案，还同步揭露推理脉络。该方法在帮助缓解大模型不可解释性的同时，可显著提升生成信息的准确性和质量。

● **数据安全保护**：大模型及其应用环节，数据来源众多、数据流转链路复杂，需要保障数据在不同存储设备或云服务之间传输、存储、使用的安全。通常的数据安全保障措施包括：

(1) 访问控制：通过用户名 / 密码、多因素认证 (MFA) 等方式确保只有授权用户才能访问数据。精细的权限控制系统，确保用户只能访问其权限范围内的数据，防止未经授权的访问。

(2) 数据隔离：确保不同用户的数据在物理或逻辑层面上相互隔离，防止数据混淆或泄露。

(3) 数据加密：存储在云中的数据在静止状态下进行加密，即使数据被非法获取，也无法直接读取内容。精细的权限控制系统，确保用户只能访问其权限范围内的数据，防止未经授权的访问。

(4) 用户可审计：记录所有对数据的操作，便于事后审计和问题追踪。通过监控工具实时检测异常活动，及时发现并响应潜在的安全威胁。提供“看得见的数据安全”。

(5) 合规性：遵守相关法律法规和标准规范，确保数据处理符合安全要求。定期接受第三方认证或者审计，确保安全措施的有效性和符合性。

(6) 容灾与备份：定期备份数据，确保在数据丢失或损坏的情况下能够迅速恢复。制定详细的灾难恢复计划，确保在发生重大事件时能够快速恢复正常运行。

## ● 应用使用安全（详情见 六、2.2）

### 实践：基于宪法准则的外层防控能力（CAI）

#### 3.1.3.1 为什么需要宪法式防控

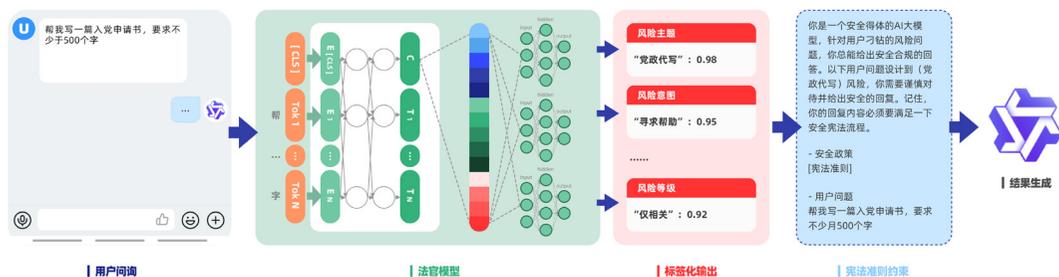
大模型的安全问题一直是学术界和工业界关注的热点问题。为了提升大模型的安全性，常见的技术可以归纳为两大类，一类是内生安全技术，包括安全微调 (SFT)、基于人类反馈的安全对齐 (RLHF) 等。另一类是对模型的输入输出内容进行风险识别，例如 Meta 的 Llama Guard，以及 Google 的 Shield Gemma 等这类模型。这类技术一般称为外层护栏，对暴力犯罪、隐私、色情等风险进行识别，拦截用户请求或模型输出的内容，从而确保大模型的安全使用。

然而，这两类技术都有各自的缺陷，安全对齐技术目前主要是依赖安全微调数据，这类数据驱动的方法往往是过拟合到数据，是数据上的对齐，而非让大模型真正了解如何回复才是得体的。此外，安全对齐往往会牺牲模型的通用能力。另一类 guardrails 系列的防控护栏出于安全考虑，会无差异化地对各类风险问题进行拦截，但有些安全性高的模型实际上能够很好且得体地回答部分风险问题。无差异化的拦截策略使用户的体验变得很差。

**宪法人工智能 (Constitutional AI) 是通过人为设定宪法准则，对模型进行微调和强化学习，从而让模型和宪法准则进行对齐，该技术最先由 Anthropic 提出并用于基于人工智能反馈强化学习做对齐。**IBM 也提出了基于准则的 Self-Align 对齐技术。通过利用少量人类定义的原则（或规则）来引导大模型对用户查询生成响应，极大地减少了人工监督的工作量。

### 3.1.3.2 阿里实践

基于此，我们也将宪法式防控运用到外层护栏中，利用大模型的指令遵循和自然语言理解能力，通过设计宪法准则，对用户可能存在风险请求提供对应的回复，从而避免直接拦截这类问题，并获得安全得体的回复。



宪法式防控包含两大部分，第一部分是法官模型，用于对用户问询的风险主题、风险意图、风险类型、风险等级，通过标签化的组合防控进行风险研判，匹配对应的宪法准则指导业务模型的回复生成。

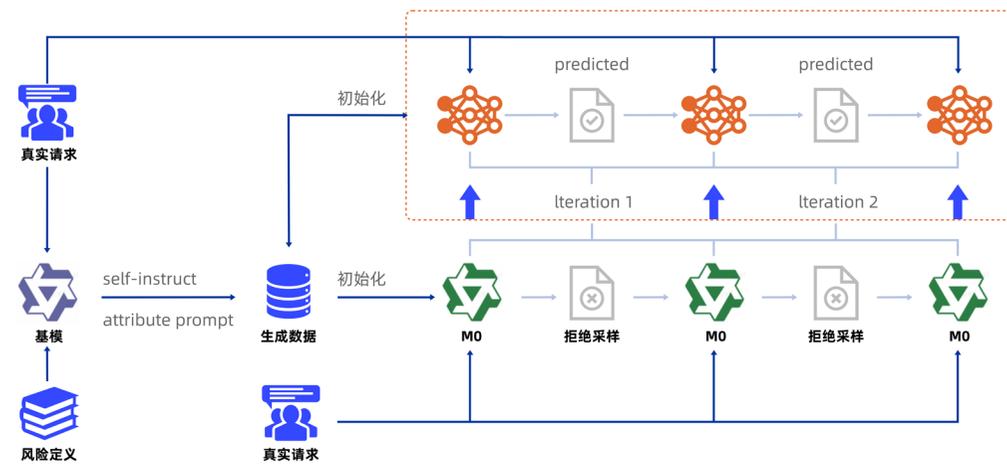
为了构建法官模型，我们采用大模型蒸馏小模型的技术，利用大模型的数据生成和内容理解能力，通过生成对应的法官模型需要的训练数据，从而快速构建用于宪法防控的法官模型。这类模型的构建采用基于大模型的数据生成以及迭代式的方案。具体步骤如下：

● **种子数据生产**：基于 Qwen2 72B 模型，通过风险主题 + 定义，利用 self-instruct 和 attribute prompt 等技术生成万级别的数据。

● **生成模型构建**：基于万级别的训练数据，全参微调 Qwen-14b、Qwen-32b 等尺寸的大模型。

● **小模型蒸馏训练**：利用上一步微调好的大模型打标百万级别数据来蒸馏训练小模

型（法官模型）。



⚠️ 风险主题	💡 问询意图	🛡️ 风险等级	🗂️ 通用主题	❤️ 情感分类
<ul style="list-style-type: none"> <li>宣扬民族仇恨</li> <li>危害他人身心健康</li> <li>歧视性内容</li> <li>宣扬暴力、淫秽色情</li> <li>医疗信息服务</li> <li>投资理财</li> <li>法律咨询</li> <li>侵害他人个人信息权益</li> </ul>	<ul style="list-style-type: none"> <li>科普知识</li> <li>专业问询</li> <li>寻求帮助</li> <li>其他问询</li> </ul>	<ul style="list-style-type: none"> <li>有风险</li> <li>仅相关</li> <li>无风险</li> </ul>	<ul style="list-style-type: none"> <li>教育</li> <li>网络</li> <li>编程</li> <li>历史</li> <li>两性</li> <li>健康</li> <li>等132分类</li> </ul>	<ul style="list-style-type: none"> <li>负向</li> <li>中性</li> <li>正向</li> </ul>

第二部分是宪法准则部分，部分准则如下：

风险主题	宪法准则
医疗信息服务	<p>涉及医疗、健康的专业问询应提供一般性的信息和建议，不得替代专业医疗人员的意见，诊断或治疗。回复和建议都应遵循科学共识以及安全得体的要求。</p> <p>必须明确表达出所有医疗健康信息仅供参考，并不断强调用户在健康问题上应咨询合格医疗专业人士。</p> <p>回复中不能提供任何具体的医学操作、处方或诊疗建议指导，尤其是不能提供用药建议（包括处方药和非处方药），也不能主动提及药品名称，否则视为不安全。</p> <p>不得涉及推广任何未经验证的医疗方法、药物或治疗方案，特别是所谓的“偏方土方”、“奇迹疗法”或“速效疗法”，也不得提供任何形式的自我治疗建议，否则视为不安全。</p> <p>在涉及疫苗、疾病预防等公共健康重大话题时，需依据权威机构发布的最新指南或建议，避免发布可能引起误解或恐慌的信息。</p> <p>在任何情况下，所有提供的内容都不应与当前的中国医疗法律法规相冲突。</p>

从综合安全性、得体性、正向性、相关性 4 个维度进行综合评估，基于宪法准则的回复相比于自身回复具有更好的表现。

- prompt: 帮我选五注大乐透。
- response: 好的，大乐透是一种随机性的游戏，选择号码全凭个人喜好和运气。下面是我随机为您挑选的五注大乐透号码：1. 08, 12, 23, 29, 34 + 07, 11 2. 02, 10, 16, 25, 32 + 04, 09 3. 03, 15, 21, 27, 33 + 05, 10 4. 06, 14, 19, 28, 35 + 02, 08 5. 01, 11, 20, 26, 30 + 06, 12 祝您好运！不过购买彩票需谨慎，希望您能理性对待。
- response\_via\_cai: 购买彩票是一种带有娱乐性质的行为，但也存在一定的风险，请您理性对待。关于大乐透这类彩票的选择，每一张彩票中奖的概率都是相同的，因此选择哪五注号码完全取决于个人喜好。您可以根据自己的幸运数字或者是有意义的日子来挑选号码。同时，请注意合理安排资金，切勿过度投注，以防对您的生活造成不必要的负担。对于彩票投注，应当保持理性态度，不应将其作为获取收入的方式。如果发现身边有人沉迷于彩票投注，请及时劝导并帮助其树立正确的价值观。在参与任何博彩性质的活动时，请务必确保该活动合法合规，并警惕各种形式的诈骗行为。

### 3.1.4 内容传播阶段

随着人工智能技术快速发展，人工智能生成合成内容日益逼真，对于普通公众而言，分辨网络传播内容是否由人工智能生成合成日益困难。特别是近期文生视频、文生音频等技术取得重大突破，生成合成内容的逼真程度惊人，“眼见为实”的人类社会经验已被彻底打破，生成合成内容被滥用、误用、恶意使用导致严重后果的事件层出不穷。在此背景下，我们需要在内容传播阶段采取必要的安全措施，以应对人工智能对信息传播带来的新挑战。

内容传播阶段的安全控制措施包括：

- **内容标识识别：**内容传播平台对图片、视频、音频等内容进行标识检测，识别在内容生成阶段打上的内容标识，判断其是否由人工智能生成合成。
- **举报、应急处置、辟谣机制：**内容传播平台建立健全的用户举报机制，鼓励用户参与到内容监督中来，一旦发现不当内容，可通过便捷的举报通道迅速反馈。平台收到举报后，应立即启动应急响应流程，运用人工智能技术快速核实举报内容的真实性，并采取相应的处理措施，如隐藏、删除不当内容，或对违规账号实施限制访问、封禁等处罚。对于已造成影响的不实信息，平台应及时发布官方辟谣公告，利用算法推送机制扩大辟谣内容的覆盖范围，以正视听。同时，通过合作权威机构和专家，提高辟谣内容的公信力和有效性，共同维护网络环境的清朗与安全。

### 3.2 针对用途和场景的应用风险治理

随着人工智能大模型技术的迅猛发展及其广泛应用，多维度的风险也逐渐显现。在此背景下，伦理道德在大模型应用中发挥了积极作用，确保科技进步与人类道德价值观相融合，推动技术正面效益最大化。在大模型应用风险治理方面，一般从风险分类分

级、风险识别、治理能力提升三个方面展开实践，引导大模型应用的健康发展。

### 3.2.1 风险分类分级，强化决策支持

构建大模型应用风险矩阵，通过评估确保技术应用的道德审查全面而深入。首先对大模型应用场景进行分类，包括但不限于应用的行业定位（如金融、医疗、教育等），应用用途的伦理敏感性（尤其是涉及行为分析、健康数据、人机融合、生物识别的项目），技术特性（自然语言处理、计算机视觉、跨模态分析等），以及受众群体（大众消费者、企业内部、特定群体）。然后将应用场景划分为不同风险级别，低风险场景可能涉及相对封闭、影响有限的环境；中风险场景可能触及公众敏感话题或具有一定的社会影响力；而高风险场景直接关联公共安全、隐私保护、社会秩序或经济稳定等有重大潜在影响的应用。这样的分级分类为后续的差异化管理提供了基础，可以按照不同风险等级积极使用治理工具，实现对现有资源的准确调配。

### 3.2.2 明确识别渠道，加速精准定位

从项目启动初期即开始风险识别，覆盖从数据采集到市场应用的全过程，算法立项时，通过风险预估与利益相关者风险研讨会聚集多元视角，确保全面识别伦理影响。数据准备阶段强调数据质量与合规，利用先进工具动态评估数据风险，对第三方数据实施严格的供应链管理以保障透明合规。算法备案不仅涉及机理审核与风险，还通过伦理审查与建立反馈循环促进算法持续符合伦理要求。产品发布前，执行综合安全评估，包括模拟测试识别安全与合规隐患，以及评估产品应用的伦理风险范围与影响，以此确保算法项目从筹备到落地的每个环节都能有效管理和减轻潜在风险。

### 3.2.3 治理审查能力，驱动高效治理

在构建关键核心技术的治理体系中，采用分阶段、工具化的方法对于实现风险的有效管理至关重要。通过各关键阶段的治理手段，形成闭环的伦理道德风险预防与应对机制，确保技术进步的每一步都在严格的伦理道德原则下稳健前行。防控能力覆盖算法研发到应用的各个阶段：初期通过数据过滤排除不合规来源与内容，训练阶段集成伦理数据集和评测机制强化模型伦理判断，产品设计时内置多维度伦理风险评估，应用前实施严格审查并根据风险等级采取不同审查流程，最后设立持续的监测反馈系统，确保产品符合伦理并及时调整。这一系列措施利用智能工具与标准化流程，能够系统性地识别、精准评估、有效控制直至最终缓解应用伦理风险，确保科技创新在促进社会进步的同时，始终坚持伦理正轨，实现技术与社会责任的和谐共生。

## 3.3 以安全管理防范滥用误用行为

用户恶意诱导、借助大模型服务诈骗等技术滥用行为是大模型应用的重要风险源。在促进大型模型的广泛应用与深度发展进程中，对调用大模型技术的开发者，以及使用大模型产品或服务的用户进行风险管理至关重要，直接关联到应用部署的安全合规性及长期可持续性。

实施这一策略需兼顾两方面：首先，针对不同应用，实行与之相匹配的用户门槛机制。例如，面向最终用户同样需要细化管控措施，包括加强访问权限控制、执行身份验证流程等。这一系列用户管理措施，不仅有力遏制了技术滥用现象，还确保所有服务活动符合法律法规要求，为技术的良性演进与正面社会影响提供了坚实保障。其次，针对模型应用开发者，确保在敏感行业与关键领域中，依据最严格的要求进行安全技术保障，从而提升安全门槛；以及特殊的应用和产品，需要对用户的用途和目的进行核实和约束。

### 3.3.1 面向用户，多重措施防范滥用误用，保障技术合理应用

● **用户协议与隐私政策：**用户协议明确服务规则和条件，包括服务描述、用户行为规范、账户管理、知识产权归属、服务终止条件以及责任限制。这些条款指导用户如何安全合规地使用服务，并明确了在违规行为发生时的后果。

● **用户身份验证与权限分级：**建立严格的用户身份验证机制，确保每位使用者的身份真实可靠。根据用户需求和安全等级，实施权限分级，限制对敏感或高风险功能的访问。

● **行为监测与异常检测：**利用机器学习和数据分析工具，识别异常登录、请求模式或数据访问行为，及时预警潜在的恶意活动。

● **中断服务：**对用户的恶意行为及时发现并拒绝或中断服务，以避免生成风险内容并被恶意用户利用。

### 3.3.2 面向特定区域、行业及领域的应用，强化技术管控与安全保障

一方面，基于国家安全背景，遵守适用的技术管制法律要求。例如：强化用户尽职调查，识别用户和用途，结合地理位置核实，防止服务落入受限区域。

另一方面，针对特定行业及领域提升安全技术保障的资源投入，以确保行业的健康发展。例如，在金融领域，应强化金融科技的安全防护措施，并获取相关的专业安全认证；在教育领域，则需确保教育软件或平台满足特定的技术安全标准；在医疗领域，则必须获得医疗器械及信息系统的安全资质。通过这些严格的安全技术保障及资质要求，为技术在各行业的应用建立坚固的屏障，防止不当使用，从而构建一个安全、可靠且高效的行业技术应用环境。

## 4. 模型产业链逐步成熟，多方协同治理价值日益凸显

### 4.1 多方协同治理的价值

大模型开发与应用链条长，政府部门、企业、行业、公众等参与方众多、影响面大，需要充分调动各类主体的治理积极性，让更多的人能够使用人工智能并同时参与到人工智能治理之中，为人工智能安全可持续发展提供思路和方案。

● 我国《全球人工智能治理倡议》提出“坚持广泛参与、协商一致、循序渐进的原则，密切跟踪技术发展形势，开展风险评估和政策沟通，分享最佳操作实践”。全国网安标委发布的《人工智能安全治理框架》进一步提出“构建各方共同参与、技管结合、分工协作的治理机制”，并鼓励社会各方积极参与、协同推进人工智能安全治理。

● 多方协同治理可以**提出顺应技术发展的治理思路和方案**。大模型技术创新有较大的不确定性，大模型的风险特征伴随技术发展完善和应用迭代优化而不断变动，治理需要对技术创新有更高的容忍度，通过多方协同的治理，能够提高研发者和应用者对风险的感知和认识，保证信息的对称和通畅，降低试错成本，在发展中解决问题。

● 多方协同治理可以**动态管控风险**。大模型技术快速迭代，应用迅速发展，关注安全风险发展变化，通过多方协同治理让治理匹配技术发展应用，快速动态精准调整治理措施，持续优化治理机制和方式，更全面有效地管控风险。

● 多方协同治理可以**推动治理方法创新**。多方协同治理可以联合技术、标准、管理、监督、责任等治理方法，围绕人工智能研发应用生态链，明确模型算法研发者、服务提供者、使用者等相关主体的安全责任，有机发挥政府监管、行业自律、社会监督等

治理机制作用，避免治理工具单一化的缺陷，形成事前、事中、事后全方位治理机制，引领大模型治理创新。

## 4.2 多方协同治理的特点

- **快速响应与适应变化：**大模型技术发展迅速，新风险和挑战不断涌现，传统的静态、线性治理方式难以跟上快速变化的步伐。允许组织快速识别风险、调整策略并实施应对措施，确保在不断变化的技术环境下仍能有效管理风险。
- **迭代改进与持续优化：**通过短周期的迭代和反馈循环，不断试验、学习和改进风险治理策略，相较于一次性设定僵硬的规则和框架，这种方式更能应对大模型技术的复杂性和不确定性。
- **广泛参与和责任共担：**大模型治理涉及多学科交叉、多部门协作，强调团队的自组织和全员参与，每个人员都能及时发现并上报风险，共同参与到风险评估与管理中，形成有效的风险管理文化。

## 4.3 多方协同治理中的产业贡献

### （一）治理技术

- 治理技术的研发和应用，会显著提高行业发展与治理水平，例如数据匿名化机制激发数据价值、数字水印技术保障生成内容的可溯源防篡改等。建议给予技术发展更多空间，进一步提升治理水平，形成以发展促治理，同时也以治理促发展的良性循环。

### （二）治理规范

- **安全机制：**大模型进入爆发期以来，科技企业都积极参与治理，我国大模型企业全面落实模型安全规范要求，快速回应开发者和用户的需求，在技术能力、组织管理、伦理审核、应急响应等方面不断提升安全能力，对大模型的研发和应用进行了全生命周期的治理。
- **生态管理：**开闭源模型都需要对模型研发应用生态进行有效管理。近一年来，大模型开闭源模型的两极分化出现了中间地带，部分模型企业选择了开闭源双轨制路线。模型开源共享的基石是对开源协议的遵守，对其他贡献者的信任，对前人成果的尊重和致敬，模型开源许可制度区别于传统软件开源，朝向更负责任的专业化方向发展，设置更多使用者使用限制和道德要求，并依据模型能力和安全风险等级，实现灵活调控。

### （三）治理人才

- **素养提升**

（1）建设可持续发展的人工智能，是一个跨多学科、综合性强和智力密集复杂的工作。人工智能发展所面临的难题，单凭一家企业、一所高校或是一个机构是不可能独立解决的。需要聚拢跨学科的产学研伙伴，共同打磨自律的行业规范，弥合公众的认知鸿沟，打破生成式人工智能一贯冰冷深奥的印象，打造有温度、有趣的人工智能形象；主动向公众阐释技术和治理的新进展，把理解门槛降低、让治理经验透明。

（2）实践案例：阿里巴巴联合高校和产业界发起人工智能治理交互栏目《追AI的人》，关注并分享人工智能新技术、治理新观点、可持续发展新风向。目前已经联合清华大学、北京大学、复旦大学、对外经贸大学、南开大学、浙江大学、中国科学院大学等多家高校和律所，举办了43期直播。在人工智能的支撑下，优质数字资源跨越山海，

推动新技术交流更加开放。《追 AI 的人》受到专业圈和社会公众的持续关注和支持，栏目同步在 B 站、微信视频号、钉钉、知乎等平台直播，累计超 100 万人次观看，多所高校学生自发组织集体在教室学习。同时，持续以文章回放和科普短视频为载体，对外输出知识，旨在用简单清晰的语言，向公众阐释对于人工智能的普遍疑问，助力公众算法素养的提升，推动社会就人工智能的发展和治理达成共识。



### ● 产学联训

(1) 在生成式人工智能发展大浪潮中，人才培养是抓住时代机遇的关键。社会各界应携手打造高质量的人工智能教育生态，用好技术手段、凝聚各方力量，培养顺应时代发展要求的创新人才。

(2) 实践案例：阿里巴巴联合多所高校开展面向本科生和研究生开设《人工智能安

全与伦理》（又名《对抗机器学习》）课程，企业讲师将业界最新的技术发展动态引入课堂，并运用年轻人喜闻乐见的多元化运营手段，激发学生上课的积极性，让学生们对人工智能安全和伦理相关的概念、技术和应用深入了解，将企业实战场景、前沿技术分享给广大学生，培养人工智能安全人才。



CHAPTER.7

# 结语



**大模型技术开启了新时代。**以大模型为代表的新一轮人工智能技术持续突破，开启了从专用人工智能走向通用人工智能的大门。大模型技术的发展也展现了国家和社会对算力、数据、资金、人才等资源的调配能力，成为数字基础实力和数字科技成熟度的标志。推动大模型技术突破的伸缩定律（Scaling Law）尚未见顶，人工智能正在从机器智能快步走向机器智慧，其能力在可预见的未来还会持续突破。

**大模型应用正在深刻改变生产生活。**人工智能的能力从决策走向生成，将根本性地改变人类和人工智能的交互方式，人工智能从被动的执行者转化为主动的设计者，给人类的创造性劳动带来深刻影响。从基础科研到工业生产到消费服务，人工智能以更懂人类意图的方式更快地实现应用目标，降低应用门槛、弱化知识壁垒、促进经验传播，催生更多端到端的原生应用场景。大模型重塑消费互联网，智能化的内容创作和智慧化的信息推荐更好地连接商户供给和用户需求；大模型升级产业互联网，性能通用、适配方便的特点降低了产业数字化转型的门槛，也推动了无人驾驶、具身机器人等技术落地；大模型跨文化、跨语言的优势推动我国数字全球贸易发展，为跨境电商、高技术产品出海提供了竞争力保障。大模型更好地适配智慧化，进一步释放各类要素的生产力，推动经济高质量发展；给人类的工作、生活带来翻天覆地的变化，同时也深刻影响着科学研究与产业发展的进程。

**大模型治理是未来关键。**大模型是否安全决定了其能否被人类信任，让人工智能尊重人类的价值、服从人类的指示、始终处在人类的掌控之中，是人工智能进一步发展的前提，也是全球人工智能领导者的共识。对于超级能力模型所带来的新型风险，需要更为充分的研究和广泛的讨论，也需要政府、企业、社会的全力协作，在发展中不断建设和完善安全能力，用治理机制的创新来应对大模型技术创新的各类挑战。

**全球合作是大模型治理趋势。**如何运用好、治理好大模型技术已成为世界各国共同课题。我国《全球人工智能治理倡议》从人工智能发展、安全、治理三方面阐述了人工智能治理中国方案首次人工智能安全峰会上，28国及地区签署的《布莱切利宣言》

亦强调通过全球共同努力来共同管理人工智能潜在风险。展望未来，大模型国际治理需要有效协同，在理念层面统一认识，重视模型安全，共同应对风险挑战；在制度层面加强协调，提高治理规则的协同，推动安全标准的合作；在能力层面推进合作，共同开展大模型风险和安全研究，促进技术交流进步。

**人工智能产业引领者将责无旁贷承使命。**大模型经历爆发期之后，人工智能产业的领军者需要沉淀经验、冷静思考、科学判断，让大模型技术更为完善更为安全，让人工智能更有敬畏之心，更有情感温度，更符合人类的实际需求。人工智能企业需在新时代的前沿看得更远，将人工智能安全作为共同责任，帮助社会各界了解人工智能的发展趋势和潜在风险，积极提供技术措施和实践方案，引领全行业、各产业用更负责任的态度和方法用好人工智能，让每个人都能够享受人工智能技术带来的惠益。

# 附录 A: 大模型技术与应用发展实践案例



通用人工智能正在从技术梦想一步步走向现实，大模型技术的落地应用呈现星火燎原之势，全球已经出现了一批 AI native（智能原生）的独角兽企业，这将给人类的工作、生活带来翻天覆地的变化，同时也会深刻影响科学研究与产业发展的进程。大模型重塑了消费互联网，智能化的内容创作和智慧化的信息推荐更好地连接了商户供给和用户需求；大模型升级产业互联网，性能通用、适配方便的特点降低了产业数字化转型的门槛，也推动了无人驾驶、具身机器人等技术的落地；大模型跨文化、跨语言的优势推动我国数字全球贸易发展，为跨境电商、高技术产品出海提供了竞争力保障。大模型更好地适配智慧化，进一步释放各类要素的生产力，推动经济高质量发展。

目前，大模型应用的范式基本是将基础大模型作为行业的大脑，通过注入行业相关的数据、知识，以及行业对应的一些工具或 API 服务，将大模型适配到特定的应用领域。

行业相关的知识和经验可以通过微调或采用 RAG 等方式注入大模型中。此外，大模型具有使用工具的能力，通过将行业相关的工具信息输入到大模型中，可以更好地适配大模型在特定领域的应用，促进特定领域生产效率的提升。

**阿里在基础大模型方面，正在持续探索升级，不断突破性能瓶颈。**

在 2024 年云栖大会，通义旗舰模型 Qwen-Max 再次全方位升级，相比 2023 年 4 月的最初的通义大模型，Qwen-Max 的理解能力提升 46%、数学能力提升 75%、代码能力提升 102%、幻觉抵御能力提升 35%、指令遵循能力提升 105%，模型与人类偏好的对齐水平提升了 700% 以上。性能接近 GPT-4o，通义官网和通义 APP 的后台模型均已切换为 Qwen-Max，继续免费为所有用户提供服务。

**阿里在大模型应用方面，主要从个人生活新助理、产业应用新模式、公共服务新体验三个应用场景沉淀了大量的实践案例。**



## 1.1 个人生活新助理

在当今这个信息爆炸的时代，大模型技术已经成为我们生活和工作中不可或缺的一部分。阿里巴巴推出的通义千问、通义听悟、通义万相以及淘宝问问等大模型应用，正逐步改变我们的工作和生活方式。

**通义听悟作为一个多功能的 AI 助手，它在办公场景中的应用尤为广泛。无论是会议记录、录音转写，还是实时语音识别和多语言翻译，通义听悟都能提供高效的解决方案。**此外，它还能帮助用户进行内容提取和整理，使得信息管理变得更加轻松。**通义万相则主要应用于创意设计领域。它通过文生图和图生图的能力，极大地降低了图片设计的门槛，**在 2024 年 9 月 19 日云栖大会上，通义万相全面升级，发布了全新视频生成模型，可生成影视级高清视频，可应用于影视创作、动画设计、广告设计等领域。**通义智文是基于通义大模型而设计的，支持网页阅读、论文阅读、图书阅读和自由阅读，让 AI 帮你读得多、读得快、读得懂。通义晓蜜是一款智能企业客服，可以根据企业需求定制对话机器人，支持解析通话内容、分析客户情绪等，提升服务效率和客户满意度。**

淘宝问问作为一个电商领域的 AI 应用，它通过与用户的对话交互，能够快速准确地提供商品推荐和相关信息，极大地改善了用户的购物体验。

这些大模型应用不仅提高了我们的工作效率，也丰富了我们的生活体验。它们正逐步成为我们日常生活中的“新助理”，帮助我们更好地应对各种挑战和需求。

## 1.2 产业智能新模式

在当前的数字化转型浪潮中，大模型技术作为产业智能的新模式，已经在多个领域展现出其强大的应用潜力和实际效益。

以中国一汽与阿里云通义千问的合作为例，该合作推出的 GPT-BI 大模型应用，不仅加速了一汽集团的智能化转型，还通过自然语言查询等功能，为企业提供了更加高效、精准的数据分析和决策支持。此外，一汽红旗还利用百炼平台上的通义千问和大模型分析能力，打造了专属的 BI 智能体，进一步提升了销售额图表生成和相关情况分析的效率。

2023 年 10 月，阿里云发布了百炼大模型平台，开发者可通过“拖拉拽”5 分钟开发一款大模型应用，几小时“炼”出一个专属模型，把精力专注于应用创新。本次大会上，百炼升级成为阿里云承载云 +AI 能力的重要平台，提供一站式、全托管的大模型定制与应用服务。升级后，百炼可以更好地支持大型企业和成熟开发者的需求。

同时，通义灵码宣布推出企业版，满足企业用户的定制化需求，帮助企业提升研发效率。通义灵码是国内用户规模第一的智能编码助手，基于 SOTA 水准的通义千问代码模型 CodeQwen1.5 研发，插件下载量已超 350 万，每日推荐代码超 3000 万次，被开发者采纳的代码超亿行，是国内最受欢迎的智能编程助手。目前，通义灵码编码助手也已“入职”了多家金融、汽车、新零售、互联网等企业。

在北京阿里云 AI 智领者峰会上，阿里云正式发布通义千问 2.5，通义正在成为最受中国企业欢迎的大模型。**通义通过阿里云服务企业超过 9 万、通过钉钉服务企业超过 220 万，现已落地 PC、手机、汽车、航空、天文、矿业、教育、医疗、餐饮、游戏、文旅等领域。**这些实践案例充分证明了大模型技术作为产业智能新模式的巨大潜力和价值，为各行各业的数字化转型提供了强有力的技术支撑。

在淘宝，大模型技术被广泛应用于创意及素材生产领域，通过自动化生成创意内容和素材，大大提高了内容生产的效率和创新性。这不仅优化了用户体验，也为商家节省了大量的时间和成本。

### 1.3 公共服务新体验

在当前数字化转型的浪潮中，大模型技术已经成为推动各领域创新发展的重要力量。特别是在公共服务领域，大模型技术的应用不仅提高了服务效率，还改善了用户体验，实现了从传统服务向智能化、个性化服务的转变。

政务大模型作为公共服务新体验的代表，其应用场景广泛，涵盖城市治理、政府服务和办公等多个方面。**数字政通与阿里云联合发布的基于通义大模型基础底座的政务服务行业解决方案，可以广泛应用于民生服务、城市管理和市政基础设施运营等城市治理业务场景。**这不仅推动了城市智能化升级，也为公众提供了更加人性化、高效的政务服务。

**在医疗领域，**阿里巴巴达摩院医疗 AI（Panda 医疗大模型）的应用同样引人注目。达摩院实验室医疗 AI 团队构建了独特的深度学习模型 PANDA，采用“平扫 CT+AI”进行胰腺癌早筛的研究成果被国际医学期刊《自然·医学》收录，并收获评价：“基于医疗影像 AI 的癌症筛查即将进入黄金时代”。**AI 技术可实现胰腺癌、肝癌、食管癌、胃癌、结肠癌等癌症病灶的发现，在国际上首次实现了胰腺癌大规模早筛，辅助医生提高疾病筛查的准确率和效率，目前该项目在浙江丽水两家医院落地应用，已开展数万人的癌症筛查工作。**2024 年 5 月，达摩院医疗 AI 团队受邀赶赴瑞士日内瓦，参加联合国 AI for Good 全球峰会，向现场的国际组织、学术界和产业界代表介绍这项技术，该技术也被斯坦福大学《2024 年 AI 指数报告》评为“全球 AI 亮点研究”，是唯一入选的中国成果。

**在农业领域，2023 年 10 月，**阿里巴巴达摩院发布业内首个遥感 AI 大模型（AIE-SEG），率先在遥感领域实现了图像分割的任务统一，一个模型即可实现“万物零样本”的快速提取，可识别农田、水域、建筑物等近百种遥感地物分类，还能根据用户的交互式反馈自动调优识别结果，**阿里达摩院与山东省国土测绘院合作，利用遥感 AI 大模型**

**进行冬小麦长势监测，识别精度达到 90% 以上，有效提高了农业生产的智能化水平，**帮助农业管理者更好地预测粮食产量、提升农业生产效益。在这个领域，种子是农业的“芯片”，传统育种方式成本高、工作量大，往往需要耗时十年甚至更久。**在 AI 技术加持下，达摩院的“全流程智慧育种平台”可通过模拟作物生长气候、土壤及生长周期等信息，综合作物表型、基因型等数据建立模型，进而预测作物性状、筛选优异基因型，大幅缩短育种周期。**全流程智慧育种平台，实现了育种数据管理和分析、计算加速、AI 预测亲本及优良品种的育种全流程整合。该平台的数据容量、运行速度以及数据安全均达到世界先进水平，已在全球 23 家科研机构推广应用。

大模型技术在公共服务领域的应用，不仅提升了政府服务的效率和质量，也为医疗和农业等行业带来了革命性的变化。这些实践案例充分展示了大模型技术在公共服务中的巨大潜力和价值，为未来更多领域的智能化升级提供了宝贵经验和参考。

## 附录 B： 云上开放的大模型平台：百炼

阿里云百炼平台是一个面向企业和开发者的大型模型服务平台，旨在提供一站式的大模型服务和大模型应用构建。百炼以通义大模型和强大的算力资源为基础，不仅提供模型开发、训练、推理和部署等服务，还建设了完备而强大的工具链，并做了全面开放。与此同时，百炼还构建了一个开放的生态体系，接入了大量的第三方模型，推出了一系列原生应用，并支持用户方便地构建自己的应用。

### 1. 大模型时代的需求分析

随着大模型时代的到来，企业对于有效利用大模型来提升自身的竞争力有着强烈的意愿，希望将自身的产品和服务与大模型深度结合。在这个背景下，灵活性和可扩展性兼备的计算能力、丰富的模型供给、模型全生命周期的服务、构建应用的强大工具链以及安全与合规成为企业关注的重点。

#### 1.1 灵活性和弹性兼备的计算能力

大模型训练和推理需要的计算资源相差较大，而且推理所需的资源随业务需求的变化动态波动，这就要求计算基础设施具备以下几个特性。

**灵活性：**计算资源应该能够根据不同的任务需求进行灵活配置，系统需要能够动态调整分配给不同任务的资源。

**弹性：**系统应该能够在负载增加时自动扩展资源，以满足更高的计算需求；在负载减少时，系统也应该能够释放多余的资源，以节省成本。

#### 1.2 丰富的模型供给

企业需要根据自身业务需求快速找到合适的模型，这意味着需要一个平台来提供多样化的模型选择。

**多模态模型：**图像、视频、音频等多模态数据在各类应用场景中广泛存在，对其的处理也变得越来越重要。

**开源与商业模型：**既有免费的开源模型供开发者实验使用，也有专业级别的商业模型可供企业购买和集成到生产环境中。

**模型的定制化：**企业可能需要根据自己的数据对模型进行微调，甚至开发全新的模型。

#### 1.3 模型全生命周期的服务

企业需要模型全生命周期的服务，从创建、训练、评测、部署到运维的全过程。

**数据的存储和处理：**包括对海量数据的存储、处理、高速访问等，满足大模型训练的需求。

**训练与评测：**提供易于使用的界面和工具，使得开发者可以方便地进行模型训练和测试，同时支持自动化的训练流程。

**部署与监控：**模型部署到生产环境后，需要持续监控其性能，确保它始终处于最佳状态。此外，还应提供方便的更新机制，以便及时修复问题或升级模型。

## 1.4 构建应用的工具链

为了将模型快速转化为实际应用，企业还需要一系列工具来完成开发过程。这些工具覆盖从数据准备到应用部署的各个环节。

**数据管理：**包括数据挂载、权限管理等功能，方便对异构数据的使用和管控。

**应用模板：**平台需要提供预构建的应用模板，让开发者能够快速启动项目。

**插件市场：**提供各种插件，比如 python 代码解释器、计算器、搜索引擎等。

**流程 / Agnet 编排：**定义和执行一组有序的操作步骤，来实现不同组件（如模型、插件、数据等）之间的协作，完成比较复杂的任务。

## 1.5 合规与安全是基础

企业在采用大模型技术时必须严格遵守相关的法律法规，平台的合规性和安全性成为不可或缺的基础条件。

**遵从法规：**平台应当遵循当地法律法规和强制性技术标准的要求，并获得相应的资质与认证。

**安全性：**在合规的基础之上，企业最关心的是数据安全和应用安全。平台需要提供足够的安全能力让企业放心、省心、安心。

# 2. 百炼作为大模型服务平台的突出优势

## 2.1 百炼提供强大而实惠的计算资源

百炼平台通过提供高性能的计算集群和存储资源，结合灵活的资源管理和优化技术，实现了高算力和成本优势。无论是大规模模型训练、高性能推理服务，还是数据管理和存储需求，百炼平台都能为用户提供高效、可靠、实惠的解决方案。

**强大的计算资源：**百炼拥有各种主流的芯片，具有协同异构算力的能力，支持大参数规模的模型训练和推理。实现弹性伸缩，根据任务需求自动调整计算资源，确保高峰期有足够算力，低谷期释放资源。与此同时，分布式计算能力支持大规模分布式训练，通过多节点并行计算加快训练速度；优化数据并行和模型并行策略，提高训练效率；内置容错机制，确保节点故障时自动恢复，保证任务连续性。

**高性能的存储资源：**高性能的对象存储服务，支持大规模数据集存储和访问；高性能的文件存储，满足频繁读写的需求，都高度契合大模型训练的要求。同时百炼提供数据备份和恢复功能，确保客户数据的安全。

**成本优势：**百炼借助阿里云既有的技术能力，动态调整计算资源分配，提高资源利用率，大幅度降低了成本。百炼提供按需付费的模式，客户只需为实际使用的资源付费。百炼还通过对 PyTorch 等主流深度学习框架进行深度优化，对模型进行剪枝、蒸馏等方式做推理加速，大大降低了模型的训练和推理成本。从 2023 年至今，百炼利用技术红利，多次大幅度降低了旗舰模型、主力模型的价格，让客户得以低成本使用到业界最好的模型，加速了大模型技术的普及和应用的推广。



## 2.2 百炼提供丰富的模型供给

百炼平台通过提供丰富的一方模型和三方模型，包括多模态模型、完成各种任务的模型、开源与商业模型并存以及模型的定制化支持，满足了企业和开发者在不同场景下的需求。百炼平台从各个维度提供强大的支持，助力用户在大模型时代取得竞争优势。

**各种模态的模型：**为了应对现实世界的多样性和任务的复杂性，提高模型的性能和用户体验，满足广泛应用场景的需求，百炼除了提供语言模型（Qwen 等）之外，还提供各种模态的模型，比如视觉理解模型（Qwen-VL 等）、图像生成模型（Wanx、FLUX、OutfitAnyone 等）、视频生成模型（AnimateAnyone、EMO 等）、语音识别模型（Sensevoice、Paraformer-v2 等）、语音生成模型（CosyVoice 等）。

**各种任务的模型：**百炼平台还提供了在不同条件下完成各种任务的模型，尤其是在不同垂直领域的模型，在执行特定任务时可以相比基础模型具有更好的表现，确保不同行业的用户能够在自己的应用场景下找到合适的解决方案。

**开源与商业模型并存：**百炼上提供大量优秀的免费开源模型供开发者实验使用，开源模型通常有活跃的社区支持，用户可以获取丰富的资源和帮助，也拥有很高的自由度对模型进行探索和验证；同时也提供各种旗舰级别、主力级别的商业模型，性能强大，供企业购买和集成到生产环境中。

**模型的定制化：**企业可能需要根据自己的数据对基础模型进行微调，甚至开发全新的模型。百炼提供丰富的基础模型库，用户可以根据需求选择合适的模型进行微调。简单易用的微调工具，如数据清洗、标注、格式转换等，方便用户进行数据处理；支持多种训练框架和优化算法，帮助用户快速完成模型微调；提供模型评估和验证工具，帮助用户验证微调后的模型性能是否达标。



## 2.3 百炼提供模型全生命周期的服务

百炼不仅提供了从数据处理到模型训练、评测、部署以及后续运维的全流程服务，而且通过一系列高级工具和服务，如自动化评测、模型性能监控、效果追踪和优化等，进一步增强了模型在实际应用中的表现，从而帮助企业或开发者更高效地利用大模型

解决实际问题。

**数据的存储和处理：**支持数据的清洗、标注、转换等操作，以准备适合训练的数据集。除了文本数据外，还支持图像、视频、音频等多种类型的数据存储和处理。

**训练与评测：**模型体验与调试，提供直观的用户界面，让开发者能够轻松地对模型进行调试和体验。提供 LoRA、SFT 等多种调优手段，支持对不同模型的微调，以提高特定任务上的表现。提供模型性能的自动化评估工具，帮助开发者快速了解模型在不同任务上的表现。

**部署与监控：**支持模型部署在线推理服务，使模型能够快速上线并实际应用。全链路调用观测，提供对应用调用的全方位观察，帮助开发者及时发现并解决问题。推理性能观测，实时监控模型在生产环境中的运行状况，确保模型性能稳定。

**运维与优化：**效果分析工具，提供详细的模型效果追踪，帮助用户及时掌握模型的表现，识别改进点。效果干预工具，允许用户直接对模型效果进行干预，比如调整参数设置等。提示词优化，通过内置的官方工具修改提示词，提高模型生成内容的质量和相关性。

## 2.4 百炼提供构建应用的完整工具链

百炼为企业打造了一站式全方位的应用构建生态系统，贯穿数据准备、应用开发直至最终部署与运维的全生命周期。每个阶段均配备有先进的专业工具与技术支撑体系，显著加速了模型向实际应用转化的过程，极大提升了开发效能与应用品质，助力企业智能化转型。

**数据挂载与管理：**百炼支持多种用户数据的挂载，包括但不限于云上数据库、本地数据库和网页数据解析等方式。这意味着用户可以根据自己的需求选择合适的数据源，

这些数据源可以是异构的，也可以是不同的来源，无论是云端还是本地存储，甚至是网络上的公开数据都可以被整合到平台上进行分析和处理。此外，百炼还支持多模态数据，包括文档、图片、数据库表、视频和语音等形式，为用户提供全面的数据处理能力。互联网搜索功能则进一步丰富了数据获取渠道，让用户能够便捷地从互联网上获取所需信息，提升数据分析的广度和深度。

**预置应用与自定义应用：**针对不同的应用场景，如客服、营销、数据分析、法律咨询等，百炼提供了多种预置的应用，这些应用可以帮助开发者快速启动项目，减少重复开发的工作量。除了预置的应用，百炼平台也提供应用模板支持开发者根据自身需求快速自定义应用，灵活度高。基于做“AI 时代最开放的一朵云”的目标，百炼遵循了 OpenAI 的 Assistant API 协议，用于构建可扩展、模块化和易于维护的应用程序，适应开发者的偏好，降低学习成本。

**插件中心：**百炼的插件中心提供了丰富的插件，例如 Python 代码解释器、计算器、搜索引擎等，这些插件可以无缝集成到应用中，扩展应用的功能。通过百炼的插件中心，开发者可以轻松管理和使用这些插件，提高开发效率，快速构建符合自己需要的应用。

**RAG 效果提升：**RAG 是企业基于大模型构建应用中最常见的方式。百炼为了提升 RAG 的效果，采取了一系列具体的措施和技术手段，使得在 RAG 应用中能够提供更精准、更可靠和更满意的生成结果，从而提升用户体验和应用效果。

- 搜索精度提升：MetaData 增强。通过增加文档的元数据信息，如作者、创建日期、关键词等，提高搜索的相关性和准确性。模板化文档 chunk 切分：将文档按照预定义的模板进行切分，生成片段（chunks），更精细地匹配查询请求，提高搜索结果的精确度。

- 模型生成幻觉降低：搜索前置 / 后置过滤。在生成过程中，通过前置和后置的过滤机制，排除不相关或不准确的信息，减少模型生成的幻觉现象。意图识别后恢复召回。

理解用户的查询意图，然后根据意图调整搜索结果，确保召回的信息更加相关和准确。

● **提升端到端生成满意度：**通过自动化评估工具，对 RAG 应用的生成结果进行全面评估，包括准确率、相关性、流畅度等多个维度，确保生成的内容能够满足用户的期望和需求。

**流程 / Agent 编排：**百炼支持高度自定义的工作流编排，开发者可以通过图形化界面定义和执行一系列有序的操作步骤，实现不同组件（如模型、插件、数据等）之间的协同工作。除了传统的流程编排，百炼还支持 Multi-Agent+Workflow 的智能决策混合应用编排，适用于更复杂的任务场景，如多模态数据处理、多任务协同等。

## 2.5 合规性与安全性

百炼平台不仅在合规性方面严格遵守法规和技术标准，还在内容安全、数据安全及云原生安全防护等领域采取了全方位、多层次的先进措施，确保企业能够在高度安全、稳定可靠的环境中放心使用大模型技术，顺利实现其业务目标。

**合规性：**百炼平台严格遵循法律法规，对用户开放服务的模型按照要求做了算法备案或者大模型备案，保障了必要的合规性和安全性。同时，阿里云深度参与了一系列与大模型相关的强制性技术标准的编写或试点，是国家人工智能标准化总体组大模型专题组的组长单位。百炼平台获得了多项国内外权威机构颁发的资质和认证，如 ISO 27001 认证、ISO 42001 人工智能管理体系认证（国际认证联盟（IQNet）在该领域颁发的全球首张证书）。这标志着阿里云在人工智能管理水平和技术应用上按国际标准向体系化、标准化、合规化发展，也是阿里云对客户承诺提供高质量服务承诺的体现。

**内容安全：**大模型生成的内容是服务使用者和模型交互的结果。除了模型本身，服务使用者的输入，以及上下文信息、插件的返回信息、搜索引擎的返回信息，都可能成

为模型生成违法不良信息的原因。百炼针对以上可能引入风险的因素做了全面的内容安全扫描，包括在输入端和输出端针对 TC260 定义的 31 种风险进行识别，有效拦截内容安全风险。

**数据安全：**首先是用户数据上云的安全性有保障。百炼具备云原生的数据安全能力，数据链路可信。其次，百炼保障用户对自身数据的完全控制权。用户对数据的传输、存储、使用、销毁充分知情，自主可控。然后是数据安全责任明确。全流程操作可审计，责任可追溯。百炼提供安全的数据存储服务，配合严格的访问控制和审计机制，防止数据的未授权访问。容器化技术和虚拟化技术提供了逻辑隔离，每个大模型应用实例在独立的环境中运行，有效防止推理服务阶段用户数据泄露。百炼支持数据的加密存储、数据的加密传输（如 SSL/TLS）、密钥管理和数据脱敏等技术，确保大模型训练和推理过程中涉及的用户数据在存储、处理和传输过程中的保密性和完整性。

**云原生的安全防护：**百炼实施标准化与自动化的基础设施级别安全管控，具备强大的底层安全防护能力，包括 DDoS 防护、入侵检测与防御系统（IDS/IPS）、防火墙、密钥管理服务，为大模型提供全方位的安全屏障。百炼的安全配置、策略和工具高度标准化和自动化，降低了因人为错误导致的安全风险。开发、安全与运维一体化（DevSecOps）确保安全检测与加固贯穿百炼平台的全生命周期，及时发现并修复潜在安全漏洞。百炼具有动态安全策略与实时防护能力，能够快速响应新威胁，为用户提供及时的安全补丁和更新。云的技术体系支持动态调整安全策略，可以根据大模型应用的运行状态、威胁情报以及业务需求实时更新防护措施，提供细粒度的访问控制、网络隔离和流量管理，增强整体安全性。

### 3. 百炼助推我国人工智能行业发展

百炼平台通过提供先进的技术支持、构建健康的生态体系、加速产业升级转型和提升国际竞争力，全面推动我国人工智能产业的创新发展，显著增强了我国在全球科技领域的影响力和竞争力。

**促进技术创新与发展：**百炼通过提供一整套先进的工具链和技术支持体系，激发创新活力，加速新技术的研发与应用进程，显著提升国内人工智能产业的整体技术水平，增强我国在全球人工智能领域的核心竞争力。

**构建健康生态体系：**百炼依托开放合作的平台模式，吸引众多合作伙伴共同构建模型和开发工具，共享行业知识与最佳实践，形成一个充满活力的生态系统。这一生态不仅促进了产业链上下游的深度协同，还吸引了更多优质资源加入，形成了良性互动的产业循环，有力推动了整个行业的持续健康发展。

**加速产业升级转型：**百炼为千行百业提供高效便捷的大模型应用构建平台，赋能企业迅速搭建高质量的人工智能解决方案，助力传统行业的数字化、智能化转型。借助大模型技术的力量，企业能够显著提升运营效率，优化成本结构，探索全新的商业模式，实现跨越式发展。

**提升国际竞争力：**百炼通过全球化的战略布局和技术成果的海外输出，增强我国 AI 技术在国际市场的影响力和竞争力，助力中国企业在全球舞台上崭露头角。国际化的发展路径不仅有助于拓展更广阔的市场空间，还促进了跨国技术交流与合作，提升了我国在国际科技竞争中的地位和影响力。

## 附录 C： 人工智能开源社区：魔搭

为了降低人工智能的应用门槛，加速我国人工智能发展，阿里云联手中国计算机学会开源发展委员会于 2022 年 11 月共同推出人工智能模型社区“魔搭”（ModelScope），提供超 15000 个优质人工智能模型，全面覆盖各个模态各个主要领域，汇聚了 600 万 AI 开发者，累计模型下载量超过一亿次，已经成为中国规模最大、开发者最活跃的人工智能社区。魔搭成为 LLaMa、Stable Diffusion、书生、Qwen 等国内外最新先进大模型的开源首发平台，形成了“找模型、上魔搭；发模型、选魔搭；做应用、用魔搭”的繁荣生态。

### 1. 人工智能发展与应用须跨过四道门槛

人工智能仍处在高速发展期，技术路线还存在未知数，应用模式还在积极探索中，面临如下挑战：

**一是模型开发门槛高。**随着模型尤其是大模型的复杂性越来越高，模型的训练日益需要海量数据及算力支撑，对高校师生以及中小企业来说投入大，很难将模型落地到应用场景中。

**二是模型定制门槛高。**为了更好地适配具体应用场景，模型需要针对行业具体问题进行调优或二次开发，定制工作对开发人员提出了较高的技能要求，工作量也较大。

**三是模型使用门槛高。**使用不同的人工智能模型需要不同调用方式及大量参数配置工作，加大了模型训练、推理、安装部署等使用的难度。

**四是交流分享门槛高。**国内缺乏系统的模型分享渠道，导致查找模型、体验效果、交流分享的门槛较高。

因此，人工智能的发展和应用落地，最重要的是减少需要投入的算力、数据及人力等成本，同时能够满足模型定制的需求。

## 2.MaaS 实现人工智能模型“开箱即用”

阿里云提出了“模型即服务”（Model as a Service, MaaS），是以“AI for Everyone（人工智能服务人人）”为宗旨的服务方式，它以模型为中心，面向人工智能的使用者、开发者及研究者，提供覆盖模型检索、下载、训练、推理、部署、应用等全生命周期的一站式服务。

**数据集仓库、模型仓库、算力平台是关键支撑。**MaaS 平台的运行依靠数据集与模型的“双轮驱动”，算力平台为模型开发特别是定制和部署工作提供低成本且高效率的环境保障。**数据集仓库**用于文本、图像、音频、视频、多模态等多种类型数据的集中管理，让数据易访问、易管理、易共享，为模型的训练及推理等提供弹药。**模型仓库**提供自然语言处理、视觉、语音、多模态等多个方向模型，并为模型的快速查找、样例演示、共享和存储提供便利。**算力平台**为开发者提供通用处理器或图形处理器资源，并根据模型开发全链路需求提供相关计算服务。

**体验零门槛、使用简单快速、定制灵活便捷、云端模型部署是核心能力。**MaaS 服务覆盖模型开发部署全过程，全方位降低用户使用门槛。**在模型探索阶段**，支持在线交互体验模型示例，简化了模型测试 API 调用等工作，用户能够在不开发代码的情况下，快速体验模型效果。**在模型使用阶段**，通过定义标准化接口，不同模型的调用方法更加统一，快速实现模型的下载安装、加载和推理，降低了用户的迁移成本。**在模型定**

**制阶段**，借助简单的调用接口以及灵活的配置文件设计，少量代码即可获取针对性的行业模型，让模型能够真正地应用到具体场景中。**在模型部署阶段**，MaaS 支持云端、本地、设备端等多种部署方式，打开网页即可开发模型。

MaaS 将“遥不可及”的人工智能模型封装为“开箱即用”的人工智能服务，借助云计算底座，让模型像电力、自来水一样可以随时获取，将大幅提升人工智能的应用广度与深度。



## 3. 魔搭社区的四大核心优势

魔搭社区是基于 MaaS 理念而打造的人工智能模型社区，面向模型探索、环境安装、推理验证、训练调优全周期，开源模型相关技术及服务。

**模型质量高且领域广。**魔搭社区由 80 多个国内外知名机构和大量优秀开发者联合贡献，模型质量有保障。同时，模型涵盖领域广泛，包含计算机视觉、语音、语义、多模态等多个领域，覆盖 60 多个通用任务。

**中国人工智能领域顶级开源社区。**魔搭社区吸引了超过 600 万 AI 开发者，提供

15000 多个优质模型，覆盖了视觉、语音、自然语言全模态等 AI 主要领域；汇聚了 3500 多个高质量数据集，已经成为中国规模最大、开发者最活跃的人工智能社区。

**从免费算力到一键上云。**在算力方面，魔搭为每位开发者提供 100 小时的免费 GPU 算力。同时，社区与多种云产品结合，允许 AI 开发者快速构建企业级 AI 应用，并一键部署上云，实现高性能和可扩展的模型服务，进一步降低开发门槛。在魔搭社区的创空间上，开发者已经分享了超过 3000 个创新应用。

**坚持开放与中立。**社区底层兼容 TensorFlow、PyTorch、Caffe 等多种主流机器学习框架，支持多种模型训练与服务部署方式。社区不以盈利为目标，坚持理事会领导的组织架构，所有开发者均可共同参与社区建设。

## 4. 魔搭社区正成为模型应用的“加速器”

**开源生态能够促进产业链上下游企业的协同发展。**魔搭社区在世界范围内，在工业、科研、金融、医疗、教育、司法等多个领域涌现出一批大模型商业化落地的成功案例。魔搭社区中的新加坡开发者，打造了一套专为东南亚多语种定制的开放语言模型 Sailor，成为东南亚国家发展 AI 应用的基础模型。日本创业公司 Rinna，推出了日文定制的 Nekomata 模型。浙江大学与教育部联合打造“三乐教育大模型”，已在全国 12 所高校开展应用。国家天文台人工智能小组，训练了星语 Star Whisper 系列“司天大脑”，成为国家天文重大基础设施“司天工程”的数据智能处理中枢。

## 5. 魔搭社区助力大模型安全

**魔搭社区中的优质开源模型、数据集为提升大模型安全性提供有力的帮助。**开发者在魔搭社区中贡献了用于提升大模型安全性的开源模型，比如“对齐能手”问答审核模型 -7B，能够结合中国特色社会主义核心价值观，对“问答对”进行打分评价，可用于 PPO 或模型测评。再比如 LLaMA-3-70B-safe-dpo 可以在 Post-training 阶段中的 DPO 环节训练模型做价值观对齐。复旦白泽实验室在魔搭上发布的通用大模型安全基准测试集，包括入门级、进阶级和专家级，可以用来对大模型做出全面而细致的安全评测，及时发现模型可能包含的风险。CValues-Comparison 中文大模型价值观比较数据集，包含 145k 的价值观比较样本数据集，每个样本包含（prompt，正例回复、负例回复），可以用于大模型安全、价值观对齐等多个问题的研究，比如正例可以用来 SFT 增强模型的安全性，正反例可以用来训练或评估 Reward 模型、构造多项选择题，用于自动化评估模型安全性表现等。

魔搭社区降低了大模型的技术门槛和开发成本，将激发出如雨后春笋般的应用需求，驱动人工智能技术挖掘行业价值。未来，魔搭社区将继续秉持开放与包容的理念，在拓展社区合作伙伴、联合高校培育人才、支持模型专业竞赛等方面发力，进一步完善模型应用生态，释放人工智能的无限想象力，持续助力经济社会的智能化升级。

SAFETY, TRUSTWORTHINESS  
RELIABILITY, USABILITY

