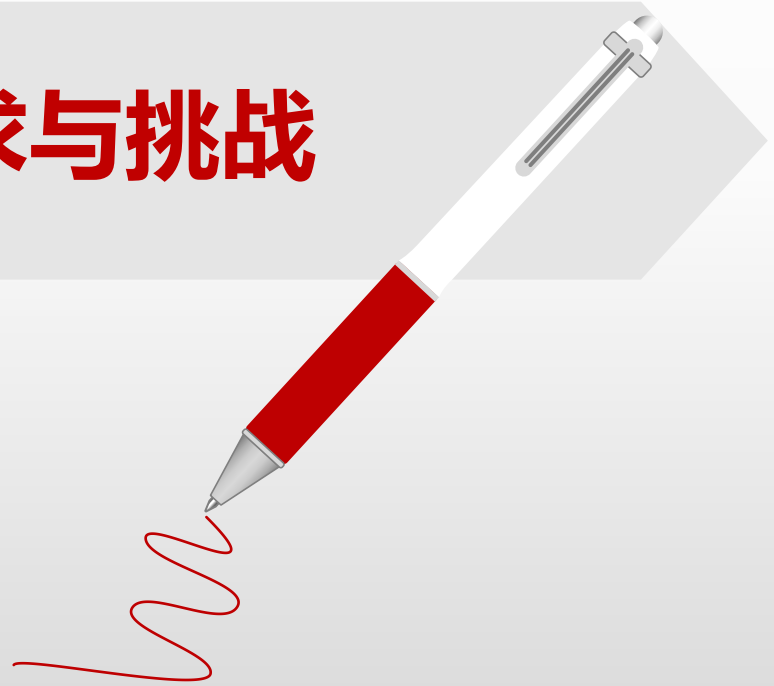
机密计算保障人工智能系统安全研究报告

2025年1月

CCiD赛迪
诚信·担当·唯实·创先

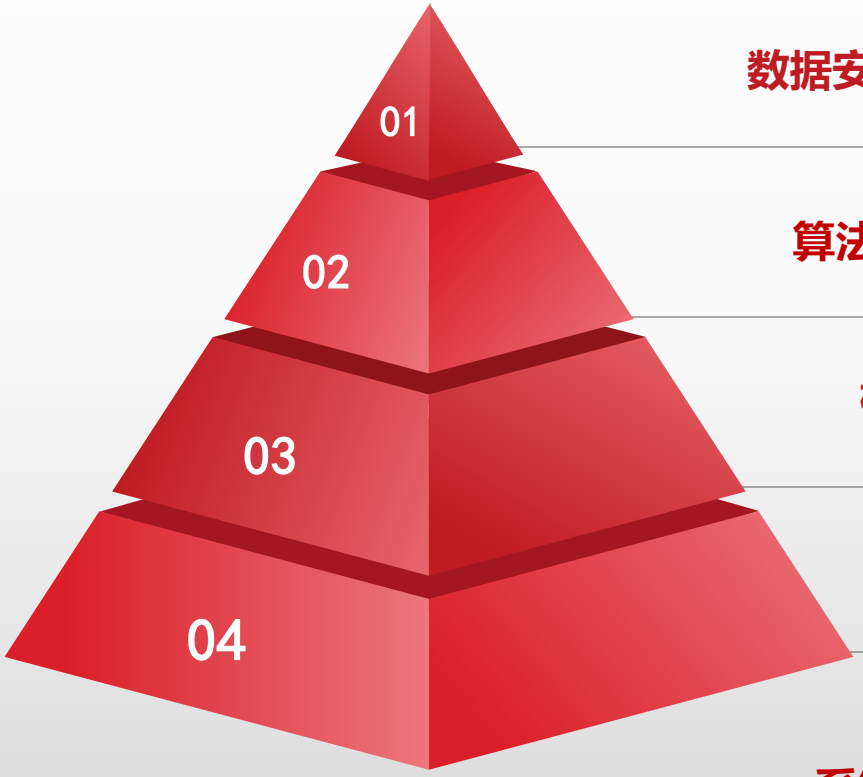
目录

1 AI安全需求与挑战



1、大模型时代的安全需求

- 新一代AI逐渐渗透到了各行各业，在显著提高生产力和效率的同时，也带来了前所未有的安全挑战，亟需建立一个涵盖各个层次的AI安全框架，有效控制AI安全风险



数据安全

确保AI系统中的数据在存储、传输和使用时，其机密性、完整性、可用性不会受到泄露、篡改或恶意使用影响

算法安全

确保AI算法具有公正性和鲁棒性，保障决策过程与结果的合法合规，避免受到恶意干扰

模型安全

确保AI模型在开发和部署过程中的机密性和完整性，防止对AI模型的盗窃、非法复制和滥用

系统安全

确保AI系统赖以运行的软硬件、服务、网络设备等基础设施的安全性，防止未经授权的访问和篡改

系统安全是最基础和最重要的，是确保业务层安全措施得以实施的基础；
《生成式人工智能服务安全基本要求》的相关要求也体现了“真正的安全要从系统层开始”这一理念

2、AI安全的主要风险和挑战

CCiD赛迪
诚信·担当·唯实·创先

01

系统安全

- 平台可信性风险
- 内存攻击
- 存储攻击
- 侧信道攻击
- 瞬态执行攻击
- 微架构数据采样
- 框架和组件的安全漏洞
- 网络攻击风险
- ……

02

模型安全

- 模型窃取攻击
- 模型后门攻击
- 模型越狱攻击
- ……

03

算法安全

- 缺乏可解释性
- 对抗性攻击
- 算法的歧视偏见
- ……

04

数据安全

- 数据投毒攻击
- 梯度反演攻击
- 成员推理攻击
- 用户输入数据泄露
- ……

3、AI安全现有解决思路的局限

◆ 现有AI安全解决方案主要集中在数据安全、算法安全和模型安全三个层级，常见的措施包括模型加密、对抗样本检测、差分隐私等，一定程度上提高了安全性，但应对更复杂的安全威胁时往往显得力不从心。

01

- 越来越多企业在云服务上进行AI训练和推理
- 现有的安全解决思路通常假设云服务本身是可信的，忽视了其可能存在的风险
- 使得AI系统在云服务环境中的安全性面临严峻挑战

02

- 现有方案在保护用户隐私数据传输方面存在不足，TLS协议缺少端到端的用户隐私数据保护
- 用户的隐私数据可能会被不可信的AI服务获取，从而导致敏感数据的泄露

03

- 在系统层面，特别是对操作系统、硬件以及云服务等关键软硬件和服务的保护上，安全性往往容易被忽略或未能得到应有的重视
- 系统层的薄弱环节可能导致整个系统的安全性失效

2 机密计算现状与趋势

1、机密计算技术原理与应用价值（1/2）

◆ 机密计算是利用具有通用计算能力的硬件可信执行环境（TEE）来保障“使用中”的数据安全，TEE应具备可编程性，可确保数据的可用且不可见，从而保护使用中的数据并维护数据的完整性和隐私。

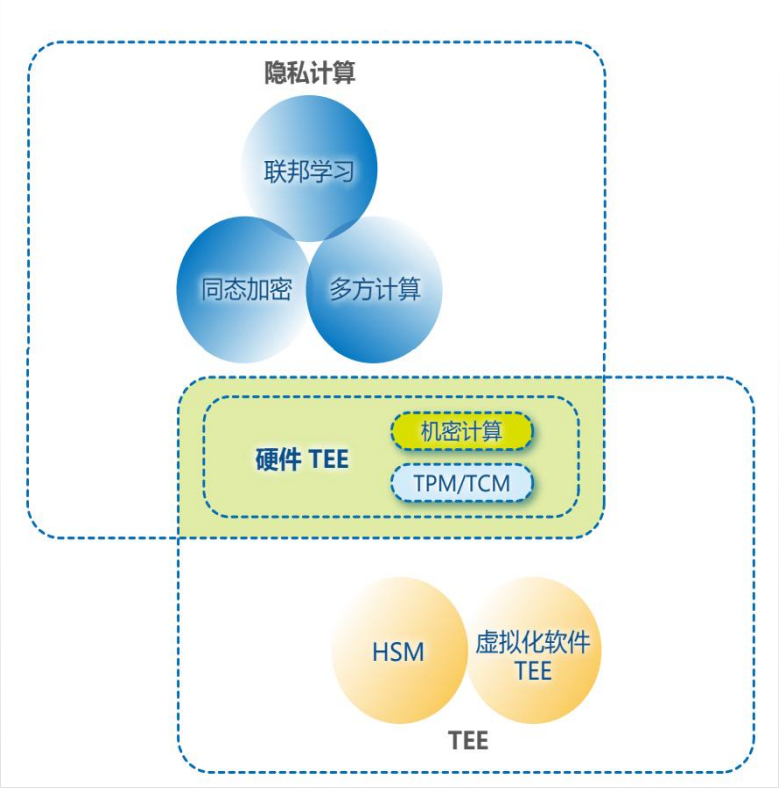


图 机密计算与隐私计算的关系

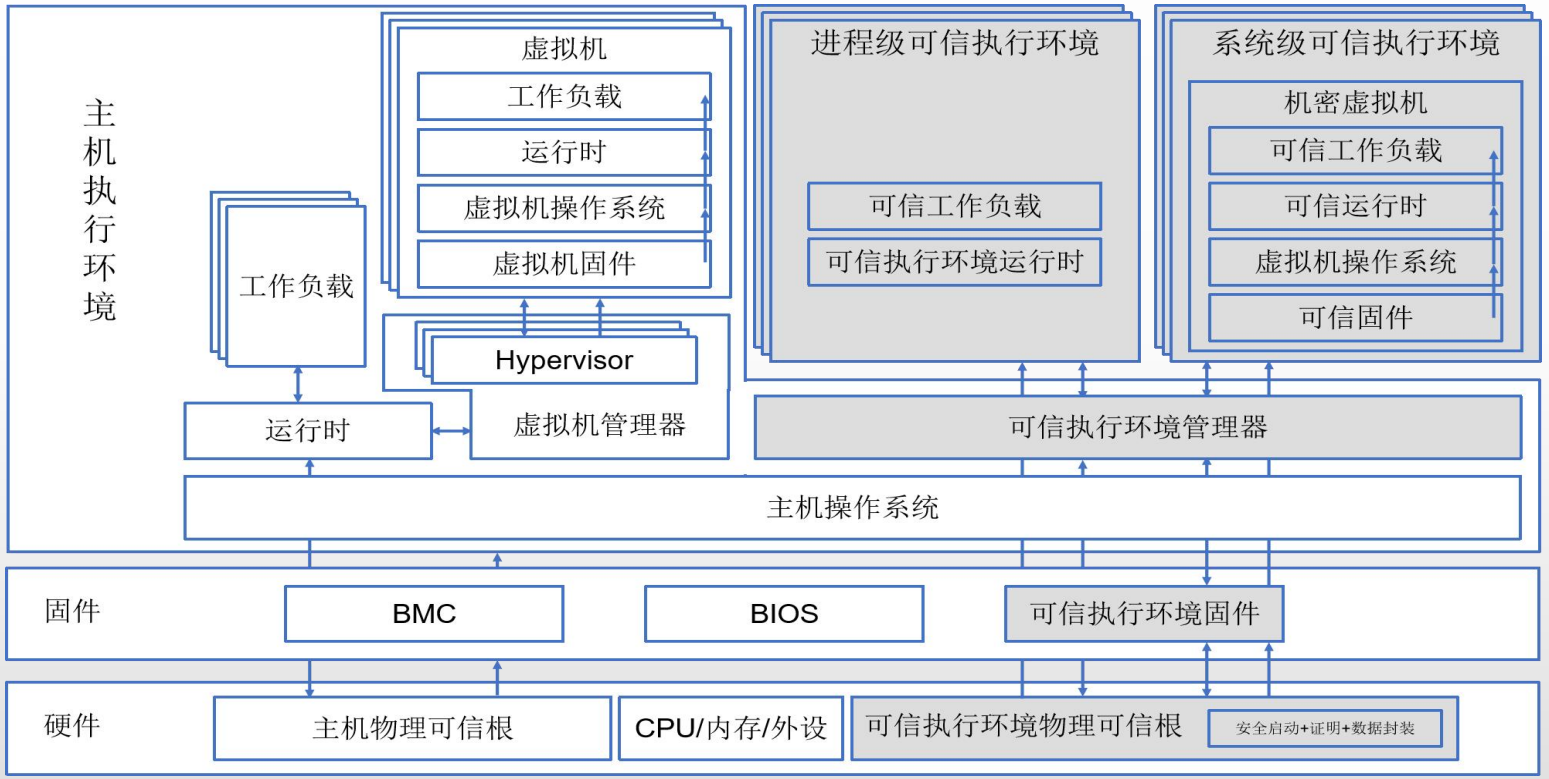


图 TEE实现方案

1、机密计算技术原理与应用价值（2/2）

机密计算为数据全生命周期提供保护

- 长期以来加密技术被用于为“传输中”数据和“静态存储”的数据提供防护，“使用中”的数据缺乏有效的保护手段，使用户数据暴露在未经授权的访问、篡改及窃取的巨大风险之下
- 机密计算全面覆盖数据生命周期的三大阶段——“传输中”、“静态存储”及“使用中”，确保数据在任何时刻都安全无虞

机密计算对云上用户数据提供强大的安全保障

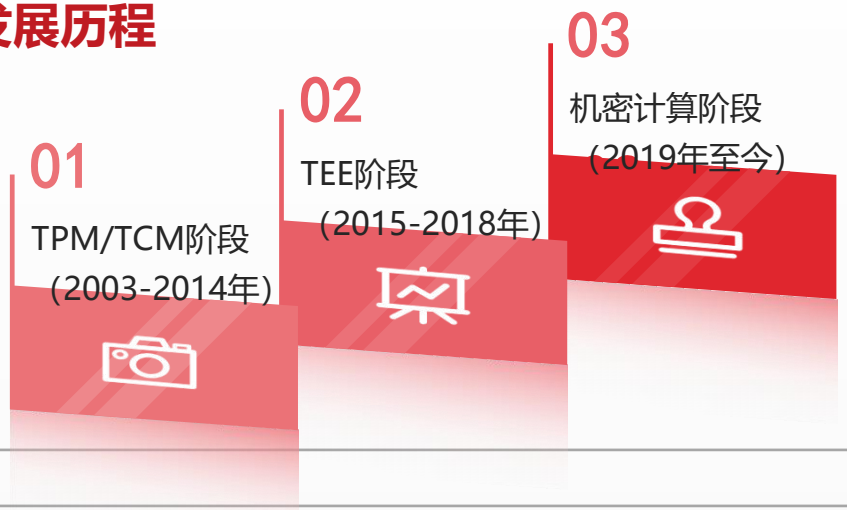
- 随着云计算和边缘计算的蓬勃发展，众多用户倾向于在云服务器的设备上部署工作负载，这一转变使得数据保护面临全新挑战。
- 机密计算将防护重心明确转向确保用户数据免遭任何潜在的系统控制者攻击，尽管工作负载依旧依托云上软件进行管理，但这些软件却被严格限制，无法窥视或篡改用户数据

机密计算为“使用中”的数据提供全方位安全保护

- 任何未经授权的实体，无论是主机上的应用程序、操作系统、Hypervisor，还是系统管理员，甚至是对硬件拥有物理访问权限的其他人员，都无法窥探到在TEE和内存中的数据加密方式。这意味着，即便内存数据不幸被窃取，也绝不会发生信息泄露。
- TEE极大地增强了数据的安全性和可证明性，为用户提供了更高的信任度

2、机密计算技术路线与产业生态

1 发展历程



2 技术路线

- 针对不同的体系架构出现了多种机密计算技术路线，x86体系相对比较成熟，ARM和RISC-V体系也正积极跟进
- 系统级TEE方案备受产业界瞩目，是未来的重要发展方向
- 不同技术路线在数据机密性和完整性保护具体实现上存在较大技术差异，但在安全能力上差异较小

3 平台互联互通

- 不同的技术路线内存加密、隔离机制等实现方式有所不同，导致不同平台的机密计算环境无法直接互通，阻碍了跨平台应用的无缝协同运行和管理
- 行业联盟和标准化组织为达成互操作性，制定了统一的机密计算技术标准和规范，显著提升了产品的易用性、安全性和兼容性

4 产业生态

- 技术研发方面，机密计算掀起了创新浪潮，引领传统安全体系结构、模型、策略及措施的全面升级
- 标准化建设方面，当前机密计算缺乏标准，造成技术障碍，各方正加快构建机密计算标准体系
- 产业应用方面，协同发展、复合应用是机密计算产业发展的核心特征

3、机密计算技术方向与发展趋势

硬件方面

机密计算将更加倚重CPU和GPU的专用安全功能，通过指令集扩展或独立安全核等方式，加速加解密操作，并为安全多方计算等复杂任务提供坚实的硬件支持

软件方面

技术演进将引领软件开发流程的深刻变革，与机密计算紧密结合的新兴编程语言和工具将极大简化安全程序的开发流程，并充分发挥以TEE为代表的硬件安全特性

平台化能力

云服务提供商正致力于将机密计算功能深度融入其云平台，使用户能够以即用即付的方式轻松享受到机密计算服务

通用范式

机密计算技术正推动安全多方计算、同态加密、零知识证明和联邦学习等一系列隐私保护技术的革新

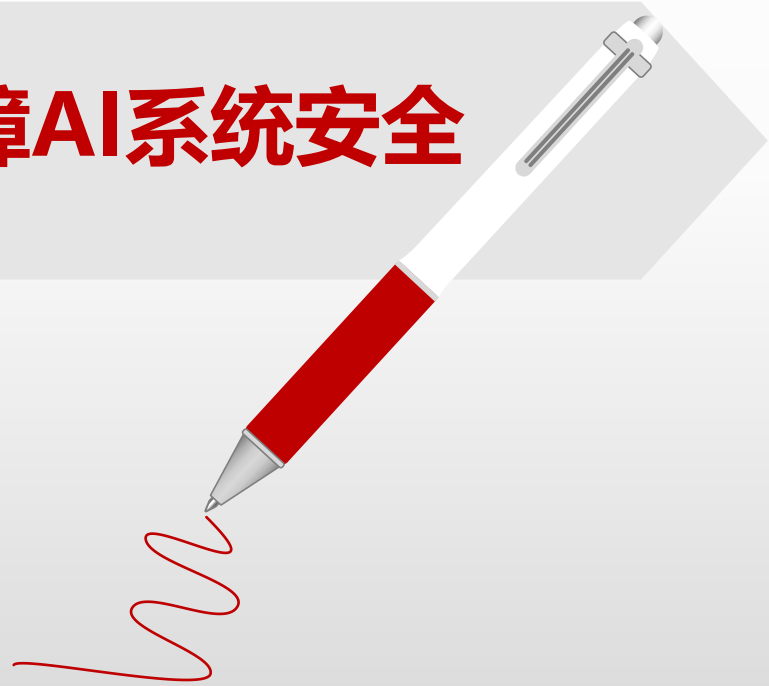
安全合规

机密计算技术需要具备高度的灵活性与适应性，以满足不同国家和地区的特定需求

技术融合

机密计算正逐步与人工智能、区块链、5G、物联网、云计算等新兴技术相互融合，催生出跨领域的综合性技术解决方案

3 机密计算保障AI系统安全



1、机密计算保障AI系统安全的思路

CCID赛迪
诚信·担当·唯实·创先

1

AI技术栈的四层结构

系统、数据、算法、模型四个核心层次结构，系统层如果存在安全漏洞，训练数据和推理数据就有可能从底层被泄露，从而对整个AI服务的安全构成严重威胁。因此，需要在系统层引入密态计算技术

2

系统层AI安全让用户减少信任实体验证

在传统的AI服务中，用户往往需要信任多个实体以确保数据的安全。这种多层次的信任结构不仅复杂，而且容易引发安全隐患。引入系统级AI安全技术，可为用户提供一种更为简洁且有效的信任模式

3

机密AI成为模型训练推理的“保险箱”

机密AI作为一种前沿的技术解决方案，巧妙地融合了机密计算的可信执行环境与模型数据安全的理念，构成了一个综合性的软硬一体技术框架

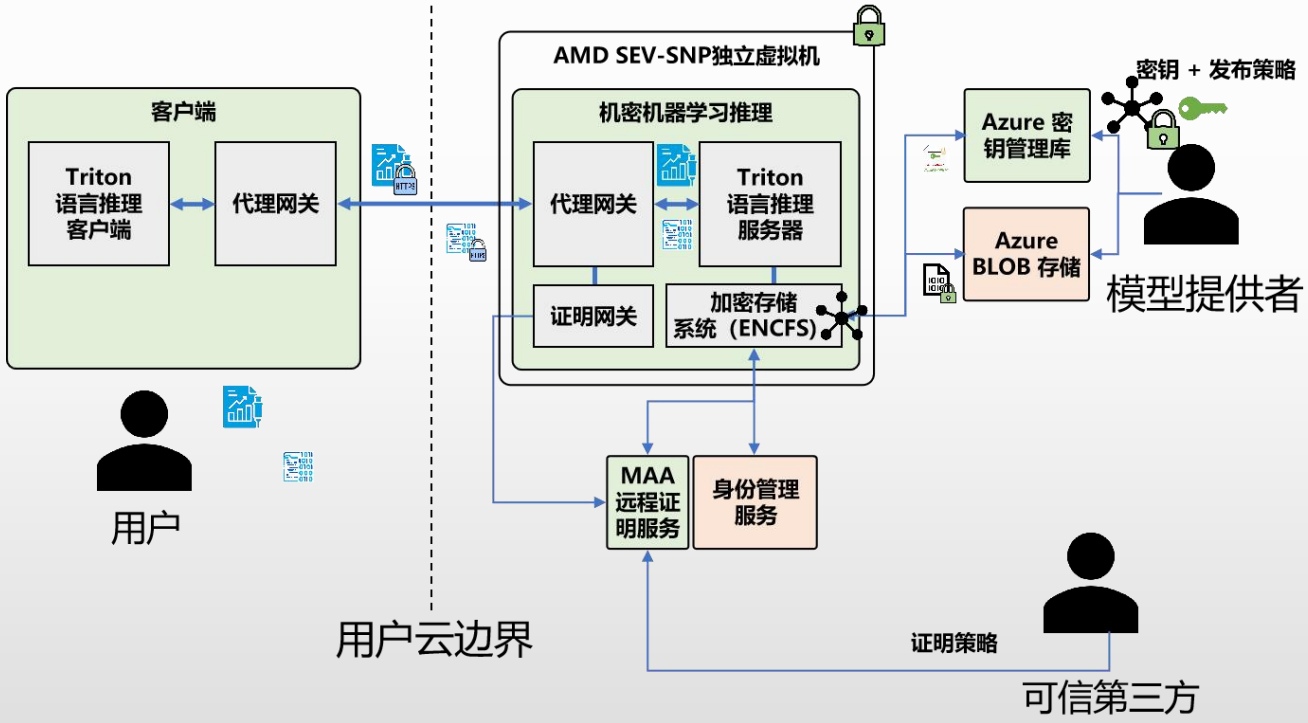
4

机密AI面临的问题与挑战

一是AI全过程保护
二是AI模型的透明使用
三是AI过程完整性保护
四是数据隐私和合规性
五是计算资源和效率

2、微软Azure机密AI技术

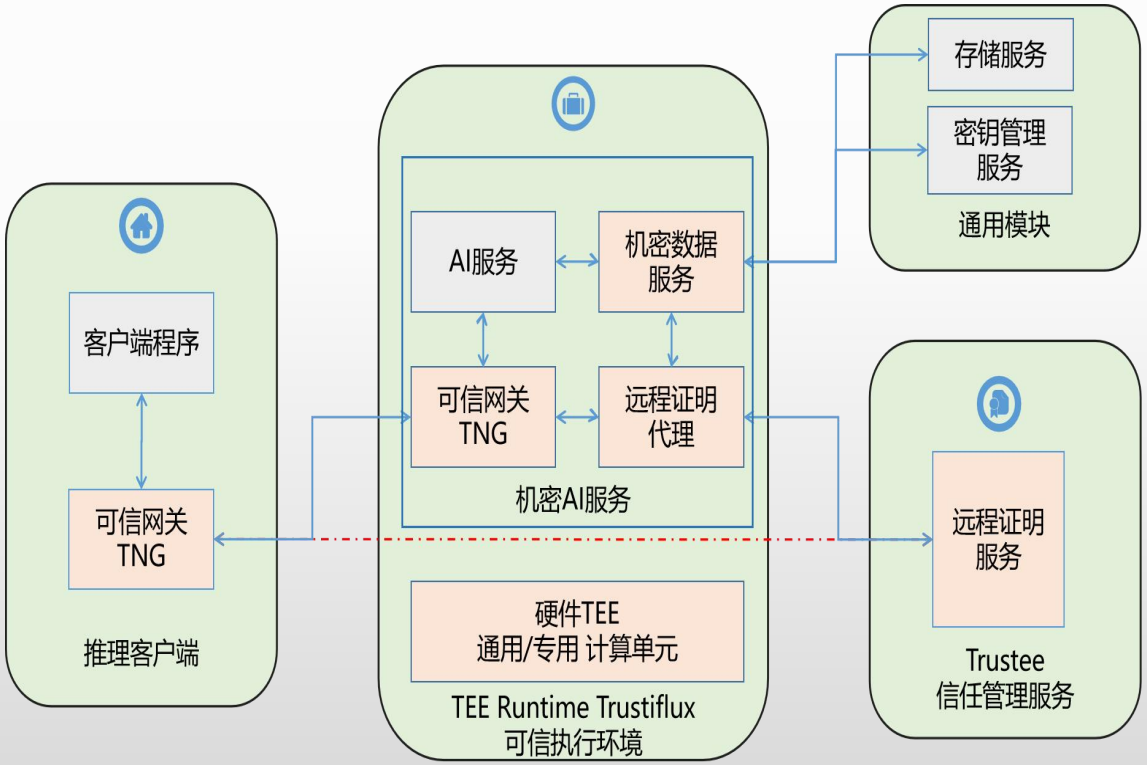
◆ 微软Azure机密AI技术基于AMD的 SEV-SNP，由运行可信执行环境中的机密机器学习推理模块、客户端以及通用模块三部分组成



- 微软Azure机密AI架构引入了可信第三方角色，将部分信任问题从Azure的云服务转嫁给可信第三方负责，从而避免了模型提供者和用户相互不信任的问题。
- 微软通过远程证明服务将Azure中已经存在的ACI（Azure容器实例）、AKV（Azure密钥管理库）和AAD（Azure目录控制）云产品进行了整合，使之在单独发挥功能的同时共同组成一套安全解决方案。
- 使用代理网关的模式，在不对推理客户端程序和服务端程序进行修改的前提下，实现了二者间的可信安全通信。
- 微软Azure应用了英伟达Hopper架构的GPU机密计算加速功能，实现了将GPU加速纳入机密AI体系架构中。

3、阿里云机密AI技术架构与实现

◆ 阿里云机密AI运用机密计算可信执行环境分离开了模型数据的所有权和使用权，并结合TEE Runtime Trustiflux和信任管理服务Trustee等软件框架，在允许模型提供者能够安全可信地将模型数据授予其他多个实体使用的同时，还能继续保持对模型数据的独占性



- **机密AI的核心TEE Runtime Trustiflux，**主要作用于以下两个执行过程：一是模型数据在可信执行环境中的解密过程，二是用户隐私数据隧道的建立过程。
- **机密AI保护模型的基石-信任管理服务Trustee，**通过远程证明功能，能够代表模型使用者等对目标可信执行环境的真实性存疑的角色实现基于远程证明过程的认证与授权机制；此外，还提供可信审计日志、可信本地存储、可信运维通道、自主身份管理等功能
- **机密AI保护模型使用者隐私的窗口-可信客户端**

4、阿里云机密AI技术的核心优势

01

系统硬件层高强度安全：
机密AI利用机密计算硬件平台提供的可信执行环境，确保AI计算过程所在的CPU与GPU可信执行环境中的内存是加密且访问隔离的

02

可审计证明的安全过程：
机密AI充分利用机密计算远程证明技术，强制保证实施模型数据在可信执行环境中的解密过程和用户隐私数据隧道的建立，必须经由模型提供者和模型使用者等角色事先配置的远程证明策略的显式同意，才能得到合法的认证与授权

03

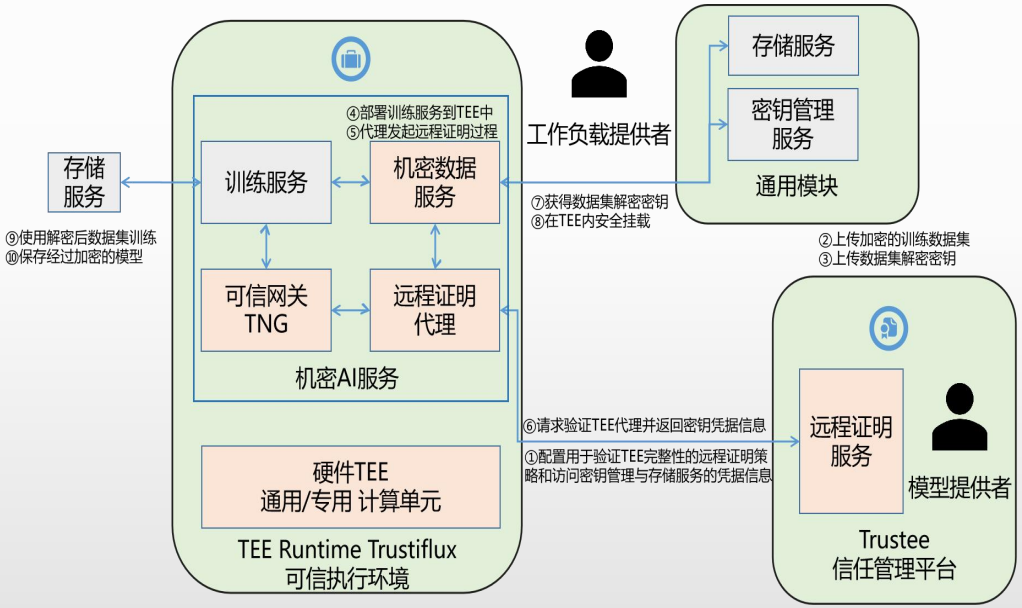
用户数据“使用中”加密：
机密AI能够帮助模型提供者提供端到端的模型数据安全防护能力，为模型使用者提供端到端的用户隐私数据安全防护能力

04

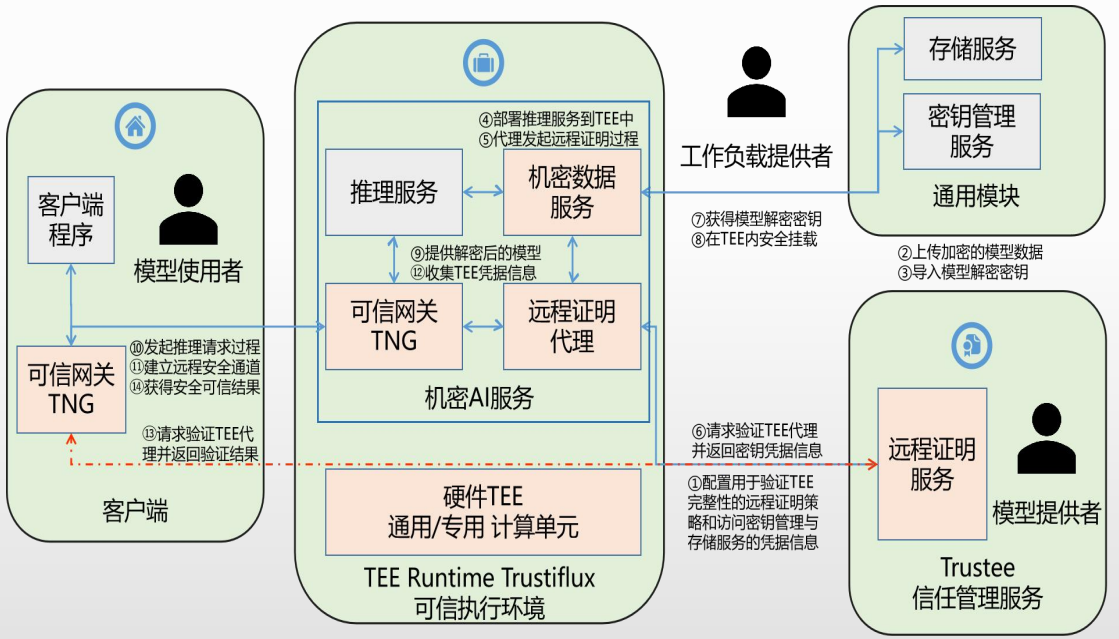
对上层应用的透明性：
机密AI在设计上要求其复杂性对AI应用来说必须是无感知的，体现在应用适配和应用部署两个方面

5、阿里云机密AI技术的典型场景

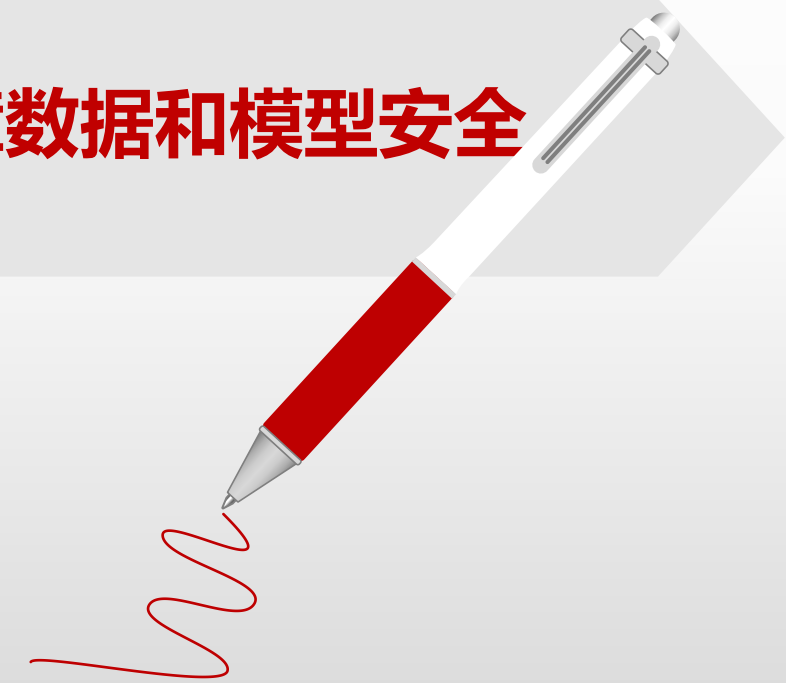
1、使用机密AI训练模型场景



2、使用机密AI推理场景



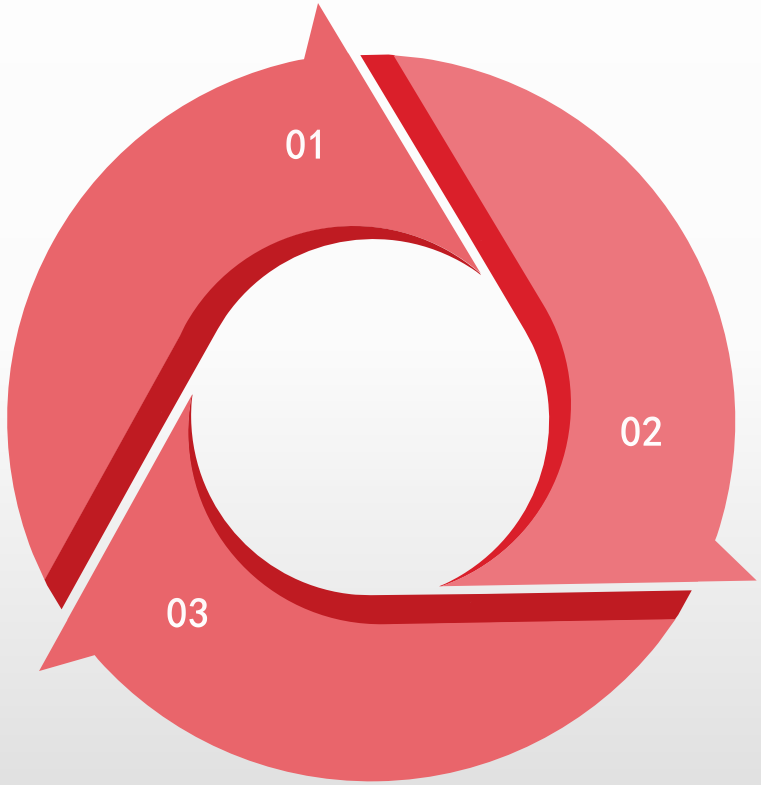
4 机密计算保障数据和模型安全



1、机密计算保障AI数据和算法安全

训练数据安全与用户隐私保护

在训练场景下，面对海量数据中可能潜藏的敏感数据泄露风险，机密计算通过TEE实现数据隔离与保护，在TEE内部进行模型训练，确保外部无法访问敏感数据。



实现数据最小化使用

基于硬件的可信执行环境，确保仅授权代码可运行，从而实现数据最小化使用

AI算法安全防护

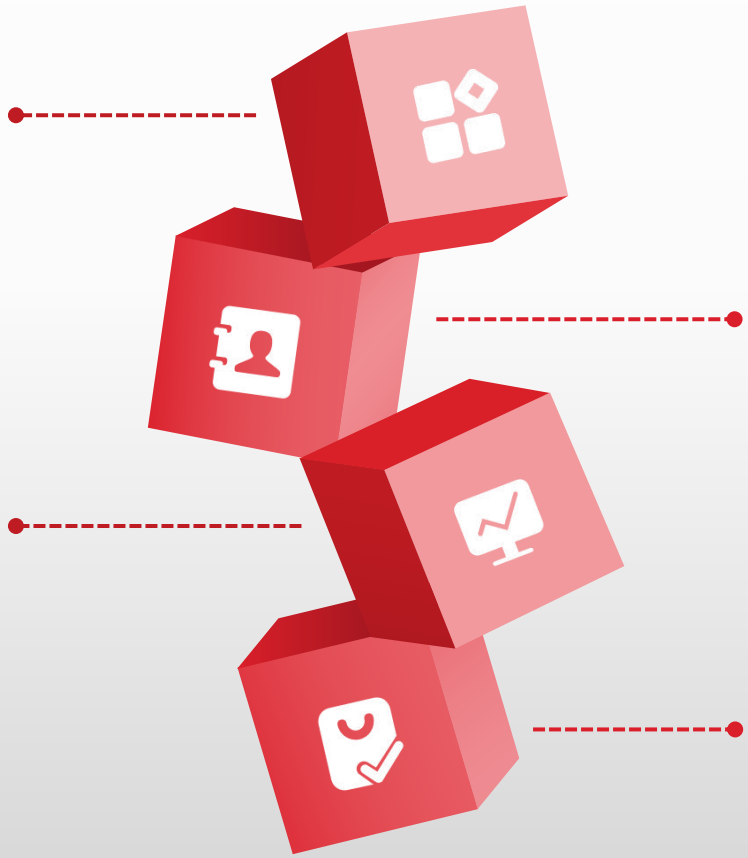
AI大模型算法面临对抗攻击、算法偏见歧视等多种安全风险，而机密计算作为系统层安全方案，为缓解这些风险提供了有力支持，特别是在白盒攻击方面

2、机密计算全面保障AI模型安全

◆ AI模型作为重要的数字资产，其安全保护聚焦于模型知识产权，以应对盗窃、非法复制和滥用等风险当前，AI模型安全保护技术主要包括以下四种，其中基于机密虚拟机的保护方案展现出显著优势。

基于传统加密算法的模型保护。虽能保护模型存储与传输安全，但无法保障运行时的安全。

基于密态计算的保护。虽在密文形式下进行推理，但AI模型在密文形式下进行推理应用，计算开销或者通信开销很大，导致模型推理时延，影响推理效率。



基于TEE的AI模型保护。虽在特定算子/子图的推理等场景下有效，但通用性不强，且可能增大TEE攻击面。

基于机密虚拟机的AI模型保护。通过机密硬件虚拟机技术，为模型提供全生命周期的保护，有效抵御基础设施提供者、恶意系统管理员等多方威胁，确保在静态存储、传输及使用过程中的安全。

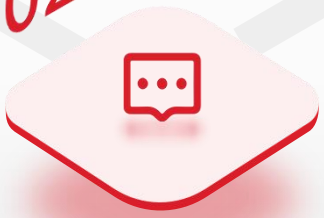
5 机密AI未来趋势展望

1、发展规模持续扩大

01

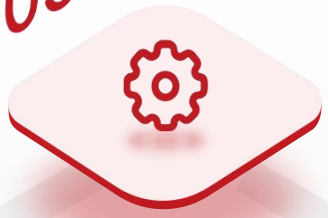
机密AI能够在数据处理过程中保障隐私和安全，从而更好地满足数据隐私法规要求

02



机密AI能够在云端环境中安全地处理敏感数据，有效降低数据泄露的风险

03



硬件技术的进步使得在隔离的硬件环境中执行AI算法成为可能，为机密AI的广泛应用创造了更多有利条件

04

关键行业对多方安全计算的需求不断增长，将为机密AI带来更为广阔的市场空间，助力其在这些行业中发挥更大价值

2、标准建设逐步完善

01 **算法与协议规范**

包括保护数据隐私的加密算法和协议等，确保数据在处理过程中的安全性和隐私性

02 **平台框架协议**

定义服务框架的架构与关键组件，并规定服务的生命周期管理，以确保服务连续性和安全性

03 **应用接口规范**

包括数据输入输出、模型训练等接口规范，确保应用程序与机密AI服务间的无缝交互，促进系统兼容性

04 **安全管理框架标准**

制定安全管理的政策和流程，以及规定安全事件响应和灾难恢复计划，以有效应对潜在的安全威胁

05 **软硬件加速技术规范**

包括硬件接口、性能指标和能效比等，规定硬件加速器与机密AI算法的集成标准，推动性能优化

06 **基于机密计算的应用标准**

针对特定应用场景的机密计算应用开发标准和最佳实践，以满足特定行业的安全需求

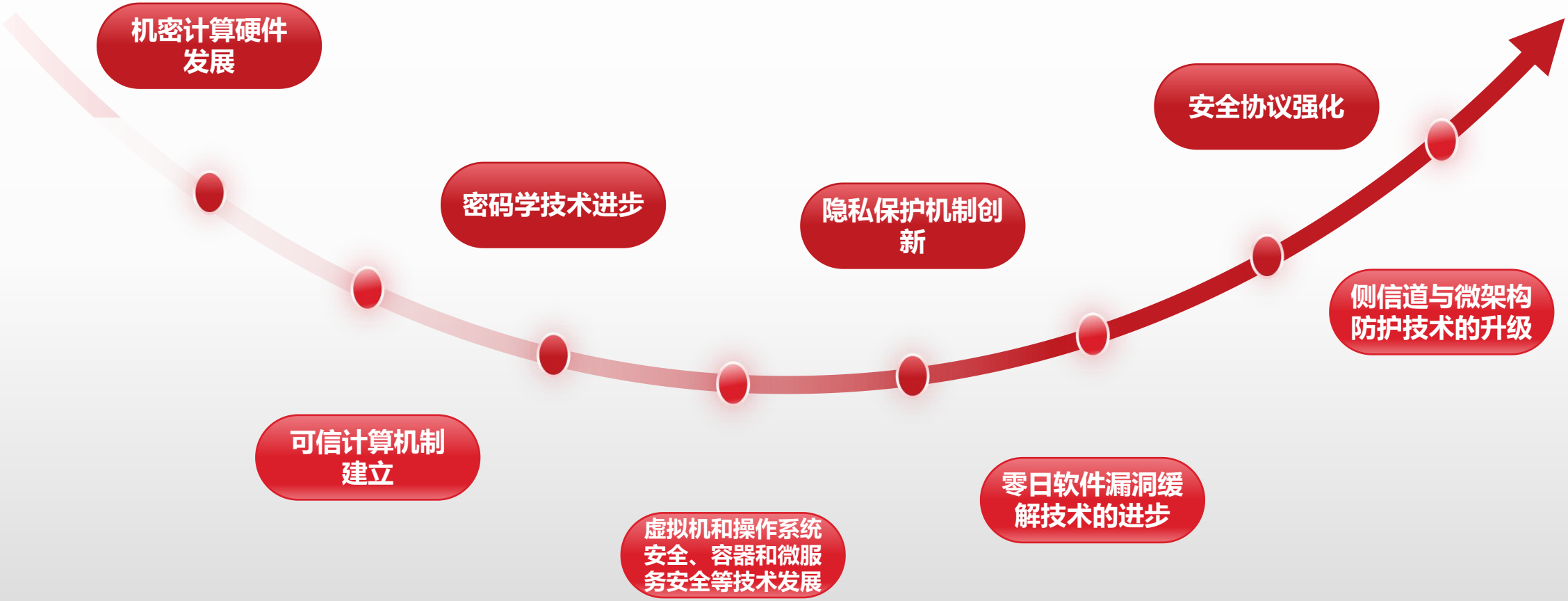
07 **平台功能规范**

明确平台必须提供的功能，及可扩展性、可维护性和高可用性要求

08 **检测评估标准**

规定评估和测试机密AI系统安全性和性能的方法，为机密AI系统的部署和维护提供科学依据

3、技术协同态势显著


机密计算硬件发展
可信计算机制建立
密码学技术进步
虚拟机和操作系统安全、容器和微服务安全等技术发展
隐私保护机制创新
零日软件漏洞缓解技术的进步
安全协议强化
侧信道与微架构防护技术的升级

4、行业应用不断深化

电子政务领域，利用机密AI实现不同政府部门间的数据安全共享，确保数据在共享过程中严格保护隐私，防止未授权访问和数据泄露，为政府数据的合规使用与高效流通提供坚实保障

金融领域，金融机构可以利用机密AI进行跨机构的风险评估和欺诈检测，安全地共享风险信息，从而显著提升金融系统的安全性和稳定性，为金融市场的健康发展保驾护航。

医疗领域，利用机密AI开展医学、药物和基因研究，在保护患者隐私的前提下分析医疗数据，助力疾病诊断、药物发现和个性化治疗，还可以支持跨机构的医学研究合作，使得大规模的临床试验和基因组研究成为可能。

工业领域，借助机密AI进行生产流程优化和预测性维护，保护知识产权和工艺数据安全，通过数据分析提高生产效率和产品质量，推动工业制造的智能化升级

商业领域，企业可以利用机密AI进行消费者行为分析和市场趋势预测，在保护商业机密的基础上，实现合作伙伴之间的数据安全共享，为更广泛的业务合作与市场洞察提供有力支持

诚信、担当、唯实、创先

赛迪研究院

思想型智库、国家级平台、全科型团队、创新型机制和国际化品牌建设

本报告由赛迪研究院网络安全研究所与阿里云、中科院软件所、南湖实验室联合编写，如有疑问和业务咨询，请联系：

温晓君 网络安全研究所所长

联系方式：13522369288，
wenxiaojun@ccidthinktank.com