2024 年隐私计算专项科研基金课题目录

一、可信机密计算

1. 考虑模型加速的硬件异构大模型安全推理技术
2. 机密计算安全监控模块的漏洞挖掘与安全分析
3. 云端机密计算环境的隐私保护研究
4. 跨平台信任根虚拟化技术研究
5. 应用 NbSP 安全范式系统研究 TEE 设计与实现
6. 可信应用代码透明化研究

二、 密码学

1. 可证可信混合计算加速研究
2. 密态关系型算子加速研究
3. 可信执行环境新型证明算法研究

三、隐私+大模型

1. 核身场景大模型身份信息隐私保护
2. 大模型数据合成中的隐私计算技术
3. 针对语言大模型的高效水印技术研究
4. 高效实时可验证计算与隐私计算融合系统研究
5. 基于隐私计算的 Deepfake 检测模型开发与应用
6. 针对在黑盒情况下垂域大模型价值窃取攻击的防御技术
7. 满足最小可用原则的数据脱敏/安全蒸馏技术

一、可信机密计算

1. 考虑模型加速的硬件异构大模型安全推理技术

背景：基于密码学的大模型安全推理技术存在效率不高的问题，该问题解决思路转向基于可信执行环境（TEE）的大模型安全推理方案。目前 TEE 设备的现状是现存大量仅支持 CPU 计算的 TEE，单纯用仅支持 CPU 运算的 TEE 和明文 GPU 推理还存在数十倍的效率差距，对于长文本的情况差距会更大。如何利用 TEE（CPU）的安全能力和明文 GPU 设备的快速计算能力是研究热点，典型的方法是把非线性计算放入 TEE 内，线性运算经过 one-time-padding 的轻加密方法放到明文 GPU 设备上运算。但这样会造成大模型逐层切割，TEE（CPU）和明文 GPU 设备的 IO 通信大大增加。虽然使用模型量化或稀疏化的操作可以减少异构设备间的 IO 通信，但大模型推理过程中的加速框架（vLLM 等）就不可用，最终使得硬件异构的 TEE（CPU）和明文 GPU 方案的实际时效性比明文 GPU 推理效率还差很多。本课题希望在异构硬件的环境下，利用轻加密算法、大模型优化算法，同时使能或部分使能大模型加速框架（vLLM 等）完成高效实用的大模型安全推理技术。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

2. 机密计算安全监控模块的漏洞挖掘与安全分析

背景：数据已成为驱动科技进步、政策制定和经济发展的新型生产要素和战略性资源。而机密计算，作为解决数据要素流通安全问题的关键技术，相较于其它隐私计算技术，具有性能好、普适性广、易用性强的特点。然而相较于其它隐私计算技术，机密计算需要用户信赖机密计算的信任根。因此机密计算的信任根的正确性与安全性显得格外重要。

目前主流的机密计算技术，例如 TDX，SEV-SNP、ARM-CCA、HyperEnclave 等，都依赖于一个运行在最高特权级的软件作为安全监控器，为具体的 TEE 实例提供机密性、完整性、可验证的保证。安全监控器是由软件编程实现，一旦安全监控器存在漏洞，那么整个机密计算提供的保证将不复存在。

针对于机密计算安全监控模块的软件（HyperEnclave hypervisor， TDX module， SEV-SNP firmware 等），本项目需要通过系统的理论分析（例如形式化验证）以及系统的漏洞挖掘等方式（例如 黑盒测试、fuzzing 测试、程序静态分析等）提供安全的背书。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1-2 篇。

返回首页

3. 云端机密计算环境的隐私保护研究

背景：机密计算技术在云端的部署可以让它面向更广泛的应用场景以及参与方，同时也对机密计算环境的隐私保护能力提出了更高的要求。例如苹果提出了 Private Cloud Compute 云端隐私计算架构，用于服务广泛的终端设备用户。在隐私计算、移动 App 等应用场景下，同样需要一套完善的隐私保护框架以及相应的算法协议设计。本项目希望基于 HyperEnclave、Occlum 等蚂蚁现有的开源机密计算方案，探索云端部署时所面临的加密保护、任务不可追踪、任务可验证、身份匿名化等隐私保护需求，通过合理使用密码算法、硬件安全设计等资源，设计相应的安全架构方案。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1-2 篇。

返回首页

4. 跨平台信任根虚拟化技术研究

背景：随着云计算和隐私计算技术的飞速发展，密算中心成为实现数据全链路安全保障的首选方式，如何在多样化的密算中心平台及其虚拟化环境中构建统一且可信的安全架构，特别是在多租户环境中实现隔离的信任根，已经成为保障密算中心数据隐私和计算安全的关键问题。然而，现有的信任根虚拟化技术(vTPM)依赖于特定的硬件实现（如 SGX、

SEV 等），在跨平台和跨服务中的应用面临诸多局限性，尤其是在资源受限的轻量级虚拟化平台上，如何构建通用跨平台信任根成为亟待解决的问题。

为了应对这些挑战，本项目旨在突破现有信任根绑定特定硬件的局限，实现更灵活的跨平台安全能力，形成更为通用性的信任根虚拟化(vTPM)生态。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 1 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

5. 应用 NbSP 安全范式系统研究 TEE 设计与实现

背景：传统的安全系统构建范式是基于攻击的，即根据已知的攻击手段构建防御系统。零越范式（Non-bypassed Security Paradigm）尝试从攻击路径和控制点角度重新审视了安全系统，要求安全系统在所有攻击路径上增加不可绕过的审查点。

当下围绕 TEE 设计漏洞的新攻击不断涌现，如：侧信道攻击，亡羊补牢的现象尤为突出，究其本质是设计之初缺乏系统化分析。本研究希望应用 NbSP 安全范对 TEE 设计展开系统化梳理，进而让现有或未来 TEE 设计，有据可依，有迹可循，更加从容的应对未知攻击。

目标：

1） 基于 NbSP 安全范式对业界主流硬件 TEE TDX、SEV 等，以及蚂蚁自研 TEE HyperEnclave 进行的系统分析，具体包含：1) 已知与潜在攻击路径 2) 访问控制点 3) 完备性；

2）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

6. 可信应用代码透明化研究

背景：构建以数据为关键要素的数字经济已经成为国家的重要战略， 而数据的安全流通涉及到端到端各个层面安全可信机制的保障，其中可信应用的代码透明化具有重要的意义，代码越透明，数据的流通越值得信任。 虽然可信执行环境 TEE 和远程证明机制能保证应用和使用中的数据相对 TCB 以外软硬件组件和人员都是黑盒隔离和安全的，但

是仅仅这些是不够的，我们也希望运行在 TEE 内部的可信应用程序的逻辑是透明和无害的。
可信应用透明化研究是构建完整可信透明化信任体系的一个重要环节，目前常规的做法是开放源码给专家用户审核，但这对非专业用户几乎是不可能的，我们希望可以借助有效的工具解决和改善这个问题； 另外，我们还需要保证可信应用度量值和实际运行的代码对应关系，严格意义上只能通过可复制构建过程验证，但是这个方式对用户有专业能力和资源依赖，我们同样希望可以通过工具抽象可复制构建过程，延迟验证过程到审计追责阶段。 具体的讲：可信应用代码透明化需要解决以下问题：1). 生命周期管理流程和安全透明化 2). 供应链安全和可信申明 3). 威胁检查， 源码或者二进制级别 4). 配置，环境变量等敏感外部输入透明化和验证技术 5). 相关的存证、溯源技术

目标：

1）POC 源代码：设计与相关的原型代码；

2）申请发明专利至少 1 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

二、 密码学

1. 可证可信混合计算加速研究

背景：数据已成为新的生产要素和战略性资源，推动着科技进步、政策制定和经济发展。隐私计算作为解决数据要素流通中的关键技术，越来越受到重视。然而，单一的加密计算（如多方安全计算和同态加密）由于性能的局限性难以处理大规模数据；而单一的机密计算由于依赖硬件信任根，端侧部署成本高，短时间内难以大面积推广。因此，本课题拟研究如何结合加密计算和机密计算技术，特别是在部分机构拥有可信硬件的情况下，如何通过软硬结合的手段加速整体隐私计算性能。旨在设计一种高效、安全、可扩展的隐私计算框架，从而解决现有技术在性能和安全性上的瓶颈，推动数据要素的安全流通与应用。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

2. 密态关系型算子加速研究

背景：数据已成为新的生产要素和战略性资源，推动着科技进步、政策制定和经济发展。隐私计算作为解决数据要素流通中的关键技术，越来越受到重视。然而，目前的加密计算技术（如多方安全计算和同态加密）主要聚焦于线性代数和机器学习相关算子的加速，对常见的关系代数-数据分析算子（例如排序和洗牌）的支持较差。本项目旨在设计创新的协议，以解决数据分析中关键算子的性能问题，从而提升隐私计算在数据分析领域的实用性和效率。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

3. 可信执行环境新型证明算法研究

背景：在隐私计算相关的技术路线中，可信执行环境（TEE）有相对较高的可落地性。然而 TEE 对硬件安全的依赖又一定程度影响用户对它安全性的信赖程度。在隐私计算的广泛应前景（例如 toC 隐私计算、密态大模型推理服务等）中，用户对安全性可能会提出更高的要求，但又缺乏对 TEE 硬件内部安全机制的审计验证能力。TEE 底层的密码协议相应需要提供更高的安全性，通过零知识证明等技术的结合，降低用户的安全度量机制对硬件安全的依赖，提升用户对自身数据安全的管控能力。本项目希望结合上述技术，为 TEE 设计更加完备的远程证明与应用度量算法，降低用户端安全验证的门槛，提升用户对 TEE 云服务安全性的信任程度。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1-2 篇。

返回首页

三、隐私+大模型

1. 核身场景大模型身份信息隐私保护

背景：蚂蚁核身平台是数字身份的基础设施，为业务提供可信便捷的身份验证服务 。应用生物识别(指纹、人脸、声纹等)、大数据、AI 技术在数字化时代和为消费者解决“我是我”、为机构解决 “你是谁”这个挑战。目前服务核心系统链路承载着亿级流量，服务国内外十亿用户。为数十亿支付宝用户的资金安全保驾护航。蚂蚁域内核身覆盖国际、网商、消金等多个场景，由于数据隐私合规的诉求需要在多端多地进行本地化部署。

身份相关模型尤其是目前的大模型的独立部署存在大量的隐私数据泄漏问题。攻击者可以通过逆向等恢复原有训练数据中的人脸，身份证件等信息。目前核身平台通过基于多模态大模型的实时视频流主动交互推动核身，随着 scaling-up 下的模型容量的增大，模型的信息量进一步增加，模型逆向的空间和风险更大，更为突出。

目标：

1）源代码：相关隐私保护代码；

2）POC 报告和相关的三方背书；

3）申请发明专利至少 3 项；

4）产出蚂蚁认可的 CCF A 类论文 2 篇。

返回首页

2. 大模型数据合成中的隐私计算技术

背景：在一体化大模型数据合成供给的业务实践中，面临着一个复杂而迫切的挑战：如何高效地利用客户私有的丰富数据资源及内部知识库，以合成高质量的训练数据，进而增强大模型的泛化能力和对特定任务的掌握度。这些数据不仅是企业宝贵的智力资产，同时也高度敏感，涉及严格的隐私保护与合规要求。因此，传统的数据集中处理方式已不再适用，必须寻找一种既能保障数据隐私，又能实现数据价值最大化的创新路径。“大模型数据合成中的隐私计算技术”聚焦于利用隐私计算与数据合成技术，在不暴露个人隐私或商业机密的前提下，对分散于不同客户间的私有数据进行高效、安全的整合与加工。本课题探索将联邦学习框架扩展到数据合成领域，通过在本地合成数据并仅分享合成模型或合成结果的差分更新，避免直接传输或集中处理原始数据，降低了隐私泄露的风险。

然而，与标注数据的联邦精调相比，大模型数据合成中的隐私计算技术需克服一系列独特的技术难关。首先，如何在保留数据真实性和多样性的同时，确保合成数据能够反映原始数据的统计特性及复杂结构，是提升合成数据质量的核心。其次，合成过程中的计算效率与通信开销，尤其是在跨域、跨机构合作时的异构数据兼容性问题，对算法设计提出了更高要求。此外，还需建立有效的评估机制，确保合成数据不仅在数量上满足大模型训练需求，更能在质量和隐私保护层面达到高标准。鉴于此，开展“大模型数据合成中的隐私计算技术”的课题研究，意味着要在隐私保护、数据合成算法的创新与优化、以及跨机构协同机制设计等多个维度上深入探索。通过与高校及研究机构的紧密合作，引入最新的理论成果与技术工具，旨在构建一个既安全又高效的合成数据供给体系，为大模型的持续进化提供强大动力，进而推动 AI 技术在尊重隐私、保障安全的基础上，更好地服务于各行各业的智能化转型与升级。

目标：

1）源代码：大模型跨域数据合成及迁移框架及模型；

2）申请发明专利至少 3 项；

3）产出蚂蚁认可的 CCF A 类论文 1-2 篇。

返回首页

3. 针对语言大模型的高效水印技术研究

背景：大语言模型（如 ChatGPT，GPT-4，LLaMA 等）在内容理解、文本生成、对话系统等领域的应用日益广泛，然而 LLMs 在生成高质量文本方面的卓越能力也带来了一些问题，例如生成内容的知识产权保护，合成虚假新闻、诈骗信息的泛滥等。在此背景下，水印技术作为一种有效的数字版权保护手段，日益受到研究者的关注。水印技术通过在生成内容中嵌入特定的标识信息，可以对内容来源进行认证，防止内容被篡改或未经授权的传播。这一技术在图片和视频领域已有较为成熟的应用，但在大语言模型生成的文本内容中，仍处于探索阶段。大语言模型文本水印的研究，不仅涉及到如何在不影响文本质量的前提下嵌入隐蔽信息，还需考虑如何在复杂多变的语言环境中实现稳健的水印提取。该研究不仅具有理论价值，对业务模型的防伪认证，知识产权保护等也有重要意义。

目标：

1）源代码：大模型水印技术的集成代码库以及新方案技术源码；

2）申请发明专利至少 2 项；

3）产出 CCF-A 类论文至少 1 篇；

4）性能指标：

效果方面，水印检测的 TPR 和 F1 在 0.99 以上；

鲁棒性方面，可以抵抗现有攻击手段的干扰；

通用能力影响方面，水印算法对模型生成文本质量不能造成过大影响。

返回首页

4. 高效实时可验证计算与隐私计算融合系统研究

背景：可验证计算的能力与隐私计算相结合，为数据价值流转带来新的协作模式。实现可验证计算的主要技术是零知识证明。在数据要素流通的大背景下，可验证计算有非常重的应用场景和价值，特别是在区块链与 RWA（实物资产证券化）的应用方向上，可验证性尤为重要。既包括复杂的如 AI 推理运算、也包括在计算资源受限的计算终端上数据处理的实时性证明。这对当前可验证计算从算法到架构乃至异构加速都提出了全面的挑战。当前，面向通用、专用领域计算提出的可验证计算算法已经成为该领域的研究热点，大量创新算法不断提出，但距离实时性证明的要求，还是有差距。本项目期望在特定的领域切入，构建实时性满足应用要求的证明系统。

目标：

1) 提出和落地新型证明协议、面向特定 AI 场景或者资源受限的端侧计算场景，设计新型专用可验证计算处理器，实现更优底层 IOP 协议、更快底层承诺方案等；

2) 基于创新的协议或虚拟机完成端到端原型验证（如区块链+IOT、可验证 AI 等场景）；

3) 产出通用零知识证明业界调研报告 及 CCF A 类论文 1 篇和相关专利。

返回首页

5. 基于隐私计算的 Deepfake 检测模型开发与应用

背景：随着公司国际业务的加速发展，对用户身份验证（KYC, Know Your Customer）的需求日益增长，特别是在线金融服务领域，确保交易安全与用户真实性成为至关重要的环节。传统的 KYC 流程包括用户证件上传及后续的活体人脸识别验证，然而，这一过程频

繁遭遇 Deepfake 技术的挑战。Deepfake 技术通过高精度的人脸合成，使不法分子得以伪造身份，绕过安全检查，对企业的反欺诈体系构成了严重威胁。为应对此类风险，构建高效准确的 Deepfake 检测模型成为当务之急。然而，模型训练需要广泛而多样化的真人人脸数据集，这在国际业务场景下意味着数据需跨国界流通，直面数据出境与隐私保护的法律约束。因此，迫切需要一种创新的解决方案，利用隐私计算技术，在保护个人隐私的同时，实现跨国数据的安全共享与模型训练。

目标：

1）Deepfake 检测模型：一个基于联邦学习框架的高精度 Deepfake 检测模型，能够在保护用户隐私的前提下，跨地域进行模型训练与优化，有效识别合成人脸；

3）申请发明专利至少 2 项；

4）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

6. 针对在黑盒情况下垂域大模型价值窃取攻击的防御技术

背景：垂域大模型比通用大模型蕴含更多的专业知识，训练垂域大模型的数据是有较高的商业价值。为了保护垂域大模型，可以把其部署在安全的云端环境内。虽然垂域模型部署在安全的环境中，但也不可避免地遭受恶意 prompt 的输入，让垂域大模型输出远超普通 query 对应的垂域知识输出。如果频繁地进行此类恶意 prompt 的输入，垂域大模型的价值会被恶意泄露。需研究此类价值窃取攻击的特点，根据其特点设计相应的检测算法，并通过相关技术手段增强大模型本身的针对价值窃取攻击的鲁棒性。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页

7. 满足最小可用原则的数据脱敏/安全蒸馏技术

背景：数据要流通，数据的安全性问题就需要被解决。结合密码学计算技术的方法在大模型应用中存在效率低的问题。数据脱敏/安全蒸馏技术是使得处理后的数据可以直接被大模型训练使用的一种安全数据处理技术。另外，大量的数据蕴含丰富的价值，在实现某一垂域大模型能力的需求下并不一定需要全部数据。如何在满足最小可用原则的条件下（可以是以选取的数据数量或每个数据信息大小等角度来考虑），对数据进行脱敏/安全蒸馏，是数据高效安全流通亟待解决的问题。

目标：

1）源代码：相关的原型代码；

2）申请发明专利至少 2 项；

3）产出蚂蚁认可的 CCF A 类论文 1 篇。

返回首页