世界青年科学家创新创业成长计划

安恒信息“AI 星火计划”技术攻坚榜单

编号	项目需求名称
01	人工智能全生命周期治理关键技术攻关
02	基于大模型的安全领域多源知识图谱智能构建——核心算法研究
03	推理结果确定性高性能算子开发
04	AI 驱动自动化网络攻防关键技术研究与实践
05	具身智能全链路技术创新与安全治理攻关
06	机密计算支持大模型安全训练和推理核心技术的攻关
07	异构环境下零侵入全链路溯源与跨模态数字水印技术研究
08	可信数据合成与隐私安全关键技术攻关

01 项目需求名称：

人工智能全生命周期治理关键技术攻关

一、揭榜方向与总体目标

围绕人工智能系统在研发、训练、部署、使用及持续演进全过程中面临的突出治理难题，开展系统化、可落地的技术创新研究。通过方法论、算法、系统框架与工程化方案的原创突破，构建可验证、可迁移、可持续的 AI 治理能力，为安全、可信、可控的人工智能发展提供技术底座。

二、课题背景与必要性

随着生成式 AI 与智能体技术迅速发展，AI 模型规模持续扩大、应用场景不断拓展，AI 风险呈现体系化、跨阶段与累积性特征。幻觉输出、语料污染、身份伪造、智能体不可控行为以及缺乏统一、有效的安全测评体系，已成为制约 AI 高质量发展的关键瓶颈。亟需突破一批前沿治理技术，形成通用框架、可复用工具和验证机制，推动我国在 AI 治理领域实现源头创新与全球竞争力提升。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向）

1. 幻觉抑制与稳健生成技术

- 大模型幻觉识别、追因与实时抑制方法
- 基于知识增强、可验证推理的稳健生成机制
- 多模态幻觉检测与语义一致性校验技术

2. 语料库安全与可信数据构建

- 训练数据来源审计与风险识别技术
- 数据污染防御、版权识别与敏感信息剔除机制
- 面向模型训练的可信、高质量语料构建与评价体系

3. 模型与智能体的身份验证与行为约束

- 防伪造的模型身份指纹、模型水印与追踪技术
- 智能体自主行为监控、越权检测与边界约束机制
- 大规模智能体群体的安全协调与冲突控制方法

4. AI 系统安全测评与治理框架

- 面向大模型与智能体的多维安全测评指标体系
- 自动化、场景化、对抗式的安全评测工具链
- 可解释、可验证、可追溯的治理技术与系统架构设计

5. 全生命周期治理一体化方案

- 从数据、模型、推理到应用的系统级治理框架
- 治理机制与大模型训练/推理框架的深度融合
- 可工程化、可部署的治理工具套件

四、成果要求

1. 形成不少于 1 套具有原创性与工程价值的关键技术、方法或体系框架
2. 提交可验证的样机、原型系统或工具链
3. 发布技术文档、安全测评报告或标准建议稿
4. 支撑可在重点行业或典型应用场景中进行试点验证

（具体成果形式可根据研究方向适度调整）

02 项目需求名称：

基于大模型的安全领域多源知识图谱智能构建
——核心算法研究

一、揭榜方向与总体目标

研究基于大语言模型的安全领域知识图谱核心构建算法，突破多源异构知识抽取、跨源实体对齐、知识冲突消解等关键技术瓶颈，形成一套完整的算法体系和原型系统，为安全知识图谱的自动化、智能化构建提供理论基础和技术支撑。

二、课题背景与必要性

· 网络安全领域知识爆炸（CVE 漏洞年增超 30%），人工管理存在四大痛点：知识分散于 CVE 库、威胁情报等，无统一体系，更新滞后难追威胁演进，多源知识冲突难处理，漏洞-攻击-防御关联分析弱。

· 技术瓶颈突出：多源异构知识抽取泛化不足，跨源实体对齐准确率低，知识冲突缺乏智能消解机制，安全领域本体（如攻击链）建模不完善。

· 国内外现状：国外 MITRE ATT&CK 等依赖人工维护，学术研究局限于单一任务；国内以人工维护情报库为主，缺乏端到端图谱构建体系。

· 技术趋势：大语言模型（LLM）带来突破（相关任务 F1≥90%），但需解决领域提示策略适配、跨源实体对齐、知识融合可解释性三大问题。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. 大模型驱动的安全知识抽取：研究面向安全领域的提示工程方法、Few-shot

学习策略和领域自适应技术，实现实体识别准确率≥90%、召回率≥85%，关系抽取 F1 值≥85%

2. 跨源知识对齐算法：研究基于语义嵌入和大模型推理的实体对齐方法，突破命名不一致、描述差异等挑战，实现跨源实体对齐准确率≥88%

3. 知识融合与冲突消解机制：研究多源知识的冲突检测、可信度评估和智能消解策略，实现冲突检测准确率≥85%，支持增量知识的高效融合

4. 安全领域本体建模方法：设计涵盖漏洞、攻击技术、防御措施、威胁行为者等核心要素的安全领域本体模型，形成≥10 类节点类型和≥20 类关系类型的本体规范

5. 动态知识更新策略：研究增量数据的高效处理算法、知识时效性评估方法和图谱一致性维护机制，支持知识更新延迟≤24 小时

通过本项目研究的核心算法，可支撑以下安全应用场景：

1. 威胁情报分析：通过图谱关联分析，快速定位漏洞的利用方式、受影响资产和防御方案

2. 攻击溯源：基于攻击链知识图谱，辅助安全分析师进行攻击路径还原和归因分析

3. 安全知识问答：支持自然语言查询，如 CVE-2023-1234 的攻击利用难度如何？有哪些防御手段？

4. 漏洞优先级评估：结合图谱中的漏洞利用链、资产暴露面等信息，智能评估漏洞修复优先级

四、成果要求

1. 成果形式

· 算法原型系统：基于 Python 实现的知识图谱构建核心算法包，包括知识抽取、实体对齐、知识融合、动态更新等模块，代码结构清晰、文档完善、可复用性强

· 技术报告：详细的算法设计文档、实验评估报告、研究总结报告

· 实验数据集：构建面向安全领域的知识抽取、实体对齐、知识融合等任务的标注数据集，支持算法评估和后续研究

2. 交付物清单：算法代码与文档、研究成果文档、数据集与资源

3. 验收标准

· 算法性能指标：知识抽取准确性、实体对齐准确性、知识融合准确性、动态更新效率

· 功能完整性指标：支持≥5 类安全数据源（CVE、NVD、安全论文、技术博客、威胁情报）的知识抽取；支持≥10 类实体类型和≥20 类关系类型；提供完整的算法 API 和调用文档；代码可复用性强，模块化设计，支持灵活配置

· 研究成果指标：完成≥2 篇学术论文的投稿(至少 1 篇投稿 CCF B 类及以上)；构建≥1000 条标注数据的实验数据集；提交完整的技术文档（算法设计、实验报告、API 文档）

· 代码质量指标：代码结构清晰，符合 Python 编码规范（PEP 8）；关键模块有单元测试，测试覆盖率≥70%；提供完整的 README、安装文档和使用示例；代码开源或交付完整源码（含注释）

（具体成果形式可根据研究方向适度调整）

03 项目需求名称：

推理结果确定性高性能算子开发

一、揭榜方向与总体目标

开发一套支持主流推理框架的确定性高性能算子库与执行机制，实现：

1. 相同输入张量、相同执行配置下，不同并发度及不同请求顺序的推理结果完全一致（浮点逐位相等）；

2. 在保证确定性的前提下，性能不低于对应非确定性实现（pytorch）的 95%。

3. 选择 GPU、DCU、NPU 或其他国产化显卡架构中的一种作为目标平台进行开发与验证，确保算子在国产自主算力环境下具备完整可运行与可复现性。

二、课题背景与必要性

随着大模型在云端推理、在线服务及边缘部署中的广泛应用，推理结果的不确定性已成为可信 AI 的关键痛点之一。

当前主流推理框架 vLLM、Pytorch 在不同并发、不同调度顺序、不同批次合并策略下，即使输入张量与执行配置完全相同，输出结果的浮点数仍可能存在微小差异。这种计算非确定性主要源于以下原因：

- 并行归约中线程竞争及累加顺序差异；
- 并发请求调度与执行次序变化；
- 底层库（cuBLAS、cuDNN）算法路径的非确定性选择；
- 内核浮点舍入与融合优化差异；

这种现象导致模型在多次运行中结果不一致，使得系统难以复现推理结果、无法进行精确调试与可信验证，严重影响了模型的可靠性与监管可追溯性。尤其

在高并发在线服务场景下，结果波动不仅降低可信度，还可能引发业务风险。国际上，开源社区已经开发出部分确定性算子，但仍然存在多线程并发推理场景推理吞吐降低30%多，和缺乏系统化解决方案的问题。因此，亟需突破“高性能并发推理的确定性保障”技术瓶颈，构建自主可控的确定性算子与执行规范，支撑可信大模型推理生态的建设。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. 确定性计算原理研究

- 浮点误差传播与舍入一致性分析；
- 归约顺序控制；
- 并发请求调度一致性策略研究。

2. 高性能确定性算子实现

- 针对矩阵乘、LayerNorm、Softmax、Attention 等核心算子开发 deterministic kernel；
- 采用 CUDA Graph、Warp-Level Reduction、Memory Barrier 等技术优化性能；
- 选择 GPU、DCU、NPU 或其他国产化显卡架构中的一种作为目标平台进行开发与验证，确保算子在国产自主算力环境下具备完整可运行与可复现性。

3. 并发确定性执行机制

- 实现固定化调度顺序的执行计划与动态批处理一致性策略；
- 消除线程随机性，实现“不同并发、相同结果”。

4. 确定性验证与性能测试平台

- 开发自动化验证框架，对不同并发度、不同请求顺序进行逐位比对；
- 生成确定性报告与性能对标结果。

5. 典型应用验证与集成

- 在 Qwen、GLM 等模型上验证确定性；
- 输出应用性能报告与标准化开发接口文档。

四、成果要求

1. 成果形式

- 自主可控的确定性高性能算子库（源码与编译包）在所选算力平台（GPU / DCU / NPU / 国产显卡）上可稳定运行；
- 并发确定性验证与性能测试平台；
- 技术标准与开发文档；
- 大模型应用验证报告。

2. 交付物与验收标准

- 确定性算子库：≥10 个高频算子；同输入同配置下不同并发与请求顺序结果逐位一致
- 验证平台：支持多并发测试与自动化比对；输出确定性检测报告
- 技术文档：完整 API、执行规范与环境要求说明
- 应用报告：≥2 个大模型推理示例，展示确定性与性能指标

（具体成果形式可根据研究方向适度调整）

04 项目需求名称：

AI 驱动自动化网络攻防关键技术研究与实践

一、揭榜方向与总体目标

核心目标是构建能够理解并执行类似 MITRE ATT&CK 战术技术过程的智能 Agent（智能体），实现从漏洞发现、利用到链路攻击的自动化闭环，解决当前安全工具“误报高、利用难、链路断”的痛点。

二、课题背景与必要性

随着网络攻击手段的复杂化和智能化，传统的基于规则和人工的渗透测试/漏洞挖掘已难以应对海量资产和未知威胁。本项目旨在征集基于人工智能的网络攻防自动化技术，构建能够理解并执行类似 MITRE ATT&CK 战术技术过程的智能 Agent（智能体），实现从漏洞发现、利用到链路攻击的自动化闭环，解决当前安全工具“误报高、利用难、链路断”的痛点。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. AI 驱动的智能模糊测试（Smart Fuzzing）与漏洞挖掘

1.1 对应 ATT&CK 阶段：Reconnaissance （侦察）， Initial Access （初期访问）

1.2 核心痛点：传统 Fuzzing 盲目性大、代码覆盖率低、针对逻辑漏洞无力。

1.3 技术要求：

· 利用 AI 模型（如 LLM 或代码大模型）对目标源代码或二进制文件进行理解，生成高质量的种子变异策略。

· 实现针对特定协议或应用接口的语义感知 Fuzzing，而非单纯的随机字节翻转。

· 能够自动识别崩溃（Crash）并进行初步的可利用性评估（Triage）。

2. 自动化脆弱性扫描与漏洞利用生成（Auto-Exploit Generation）

2.1 对应 ATT&CK 阶段：Execution（执行），Persistence（持久化），Privilege Escalation （提权）

2.2 核心痛点：扫描器误报率高，有了 POC（验证代码）往往难以转化为 EXP（利用代码），难以适应不同环境。

2.3 技术要求：

· 构建基于 AI 的漏洞验证引擎，自动编写或修改 Payload 以适应目标环境（如绕过 WAF、适配不同 OS 版本）。

· 利用强化学习（RL）或规划算法，在模拟环境中自动化尝试提权路径。

· 实战化要求：必须证明工具能完成“发现漏洞 -> 生成 EXP -> 获得 Shell”的全自动流程。

3. 基于攻防知识图谱的漏洞链式应用（Vulnerability Chaining）

3.1 对应 ATT&CK 阶段：Lateral Movement（横向移动），Collection（收集），Impact （危害）

3.2 核心痛点：单点漏洞危害有限，缺乏将多个低危漏洞组合成高危攻击路径的能力。

3.3 技术要求：

· 构建攻防知识图谱，让 AI Agent 具备“战术规划”能力。

· 实现多跳攻击决策：例如，利用 Web 漏洞进入内网 -> 扫描内网弱口令 -> 横向移动到数据库。

· 场景要求：在给定的靶场环境中，自动规划并执行一条包含至少 3 个攻击步骤的完整攻击链。

四、成果要求

1. 核心考核指标

· ATT&CK 覆盖度： 工具或算法需至少覆盖 ATT&CK 矩阵中 3 个以上战术阶段和 10 个以上技术点。

· 自动化程度：在攻击/检测过程中，人工干预次数需 ＜2 次（理想情况为 0 干预）。

· 准确性：漏洞验证的误报率需 ＜5%；漏洞验证的漏报率需 ＜10%（基于标准靶场测试）。

· 效率指标：针对标准 C 级网段资产的自动化决策与探测时间，较传统人工渗透测试需缩短 50%以上。

· 可解释性：系统需输出完整的攻击/防御路径图，清晰标注每一步利用的 ATT&CK 技术编号（如 T1190）。

2. 交付物要求

· 可运行的原型系统/工具：提供 Docker 镜像或可执行程序，便于在隔离环境中进行复测。

· 技术白皮书：详细阐述 AI 模型架构（如使用了何种 LLM、强化学习算法）、训练数据集来源，以及针对相应攻防阶段的映射逻辑。

· 测试报告：基于不少于 3 个典型靶场环境（如 DVWA、Metasploitable3 或自建实网仿真环境）的实测报告。

（具体成果形式可根据研究方向适度调整）

05 项目需求名称：

具身智能全链路技术创新与安全治理攻关

一、揭榜方向与总体目标

聚焦具身智能系统在全链路的感知、决策、动作执行、系统升级与多智能体协作等关键环节，开展系统化技术创新研究。通过方法论、算法、工具链与框架设计，实现具身智能在复杂真实环境中的高效感知、稳健决策、精确执行及可控协作。

在全链路治理中，安全与合规性是重要环节，包括控制指令验证、行为合规性监控、OTA 升级安全、边界风险防控等。课题旨在形成可验证、可工程化的方案，实现具身智能系统的安全、可信、可部署与高效运行，为智能体在工业、服务、科研等应用场景的落地提供技术支撑。

二、课题背景与必要性

具身智能系统在现实环境中自主感知、推理和执行任务，带来效率与功能的提升，但也引入了多维技术挑战：

- 感知与理解：多模态信息融合、动态场景理解、对象与关系预测
- 动作规划与执行：高精度控制、不确定环境下稳健动作、多任务协作
- 系统升级与边界控制：OTA 升级漏洞、软件与固件安全、智能体行为边界约束
- 安全与行为监控：控制指令异常检测、行为合规性验证、可追踪与可解释

目前具身智能系统仍缺乏覆盖全链路的技术体系和安全治理方案，亟需突破感知、决策、执行及安全管理的关键技术，形成可落地、可验证、可部署的解决

方案，推动具身智能在复杂场景中的安全可靠应用与产业化落地。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. 场景感知与理解

- 多模态感知融合与环境建模技术
- 动态环境中语义理解与任务推理
- 复杂交互场景下对象识别、关系理解与预测方法

2. 动作规划与执行优化

- 高精度动作控制与路径规划算法
- 不确定环境下的稳健执行与自适应策略
- 多任务协作与冲突调度方法

3. 行为安全检测与监控

- 控制指令异常检测与行为合规性验证
- 基于规则与数据驱动的安全监控框架
- 智能体行为可追踪、可解释与可验证机制

4. OTA 升级与系统边界安全

- 升级过程安全性检测与漏洞防护
- 智能体软件与固件安全管理策略
- 多智能体系统协作中的安全边界约束

5. 具身智能全链条安全治理

- 从感知、决策到执行的端到端安全治理框架

- 安全策略与控制机制与系统架构的深度融合
- 工程化验证与模拟环境中安全评测工具链
- 提供模拟环境与真实场景可部署的验证方案

四、成果要求

1. 全链路技术创新：提出具身智能系统在感知、决策、动作执行及多智能体协作等环节的原创方法、算法或框架。

2. 安全治理能力：实现控制指令验证、行为合规性监控、OTA 升级安全及边界风险防控的可验证方案。

3. 工程化与部署：构建可在模拟环境或真实场景验证的原型系统或工具链，支持分阶段部署与迭代优化。

4. 文档与标准化：提交技术文档、评测报告或可复用的工具模板和方法指南。

（具体成果形式可根据研究方向适度调整）

06 项目需求名称：

机密计算支持大模型安全训练和推理核心技术的攻关

一、揭榜方向与总体目标

旨在通过机密计算技术突破，实现可以实际落地的、成本可控的，可以为大模型预训练、微调、推理阶段的数据以及模型提供安全保护的技术能力。

二、课题背景与必要性

大模型（LLM）的性能高度依赖于训练数据的规模和多样性。然而，最有价值的数据——如医疗影像、金融交易记录、个人身份信息，由于合规要求、商业机密或隐私顾虑，被严格限制在各自的“数据孤岛”中 ，无法实现跨域流通和融合分析。

大模型的训练和推理的成本动辄上千万美元，其训练的数据、推理的数据、模型的权重都是开发者最核心的知识产权。这个过程中涉及数据提供方、模型提供方、算法提供方等多个参与方，各方均具有不同的安全考量。在云上进行训练或提供推理服务时，需要确保各方提供的模型和数据不会被包括云服务商、恶意特权用户等在内的潜在恶意攻击者窃取。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. 机密计算与大模型训练和推理的安全架构

- 机密计算环境下的安全评价模型与安全体系架构设计
- 机密计算环境下的多租户服务的数据安全隔离与数据合规管控机制

· 机密计算环境下的大模型训练与推理全生命周期威胁建模与安全策略

2. 机密计算技术与大模型框架的融合技术

· 机密计算下的主流模型底层调度框架深度融合方案

· 大模型训练/推理流水线在机密计算环境下的编排与抽象

· 面向开发者的机密计算+大模型一体化开发与调试工具链

3. 机密计算技术的性能瓶颈优化

· 机密计算环境下大模型计算与内存开销分析与优化

· 机密计算环境下大模型 I/O 与网络瓶颈优化

· 硬件加速与系统级协同优化

4. 机密计算在大模型场景的互联互通技术

· 异构机密计算平台之间的可信、安全互联与协同调度

四、成果要求

1. 核心考核指标

为确保技术的可应用性，所有方向均需交付可运行原型系统、技术报告与可复现测试材料；指标覆盖安全体系与评价、性能对比优化、互联互通与可移植性，以及工程化质量与示范落地。

· 安全架构与功能完备性：构建面向机密计算场景的安全体系与评价模型，形成覆盖“数据—代码—运行时—访问控制—审计追溯”的可落地方案，并提供原型实现与验证结果。

· 性能与开销控制：针对主流机密计算技术栈的关键性能瓶颈（如内存加密开销、I/O、网络、证明流程、密态容器/虚拟化等），提出并实现优化方案，给出可复现的对比验证，证明方案具备工程可用的性能水平。

· 互联互通与可移植性：形成跨平台、跨环境、跨实现的对接能力，支持与现有业务系统/数据基础设施的标准化集成，降低迁移成本，确保方案可在不同机密计算硬件与软件栈上部署运行。

· 其他：为鼓励不同技术路线与创新方案参与揭榜，除上述方向外，允许并鼓励揭榜方在不降低安全与可用性前提下提出差异化技术路径与扩展能力，并以可交付成果。

2. 交付物要求

2.1 原型系统（可运行）

· 提供可部署的原型（含安装包/镜像/部署脚本），可在指定硬件/云环境完成部署与演示。

· 输出核心模块清单、接口说明、运行手册、故障排查手册。

2.2 技术报告（可审阅）

· 体系架构与威胁模型、安全设计、关键机制说明、实现细节、性能评测方法与结果、边界与限制、适用场景与风险提示。

2.3 测试与验收材料（可复现）

· 性能基准与安全测试用例、测试数据集/生成脚本、测试环境配置说明。

· 形成验收演示脚本：一键部署→功能演示→安全验证→性能对比→接口互通。

（具体成果形式可根据研究方向适度调整）

07 项目需求名称：

异构环境下零侵入全链路溯源与跨模态数字水印技术研究

一、揭榜方向与总体目标

针对企业在无法获取应用源码（无代码审计权限）的“黑盒”环境下，解决数据在“应用—数据库—文件”之间流转的链路断层问题。核心目标是构建一套非侵入式的溯源体系，利用大语言模型（LLM）的逻辑推理与因果推断能力，实现基于流量和日志的动态血缘重构；同时结合跨模态水印技术，打通结构化数据（DB）到非结构化数据（文档/图片）的追踪路径，实现全生命周期的可视、可管、可溯。

二、课题背景与必要性

1. 实际痛点（零侵入需求）：现代企业安全建设中，安全团队往往难以介入开发流程或获取源码权限，导致传统的代码扫描（SAST）无法落地。而在微服务架构下，数据库审计日志与 API 网关日志割裂，无法还原业务侧的真实数据流向。

2. 跨模态监管空白：数据从数据库查询导出为 Excel、PDF 或被截图后，原有安全标记丢失，导致泄露后无法追查源头。

3. 技术结合点：传统基于规则的日志关联误报率高，利用 LLM 强大的语义理解能力，可以精准推断“API 请求”与“SQL 执行”之间的因果关系，并深入理解数据流转背后的业务意图。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑

战方向。）

1. 基于全流量解析与多源日志因果关联的零侵入血缘重构（核心重点）采集数据库通信协议流量与应用层 HTTP 流量，利用 LLM 对 API 日志与 SQL 审计日志进行高维特征提取。基于时间窗口、参数相似度及访问模式，构建因果推断模型，自动化推导“API 接口–数据库表/字段”的映射关系，在不触碰代码的前提下补全应用级血缘。

2. 微服务链路下的数据流转拓扑自动发现与异常检测针对 Trace ID 缺失的旧系统或复杂微服务链，研究基于流量指纹的无代理（Agentless）追踪技术。结合图神经网络（GNN）还原微服务间的数据流转拓扑，识别非业务逻辑的数据聚合或绕行访问。

3. 结构化数据到非结构化文档的跨模态隐式指纹嵌入聚焦数据导出场景（ETL/BI 报表），研究在结构化数据转换为 Excel/CSV/PDF 瞬间生成抗攻击隐式指纹的算法。确保指纹能抵抗格式转换与截断攻击，并能反向解析出导出者身份与时间。

4. 基于图数据库的超大规模动态血缘存储与秒级查询针对全流量分析产生的海量瞬时血缘关系，研究高效时序图谱存储架构。解决实时写入与历史回溯难题，支持对敏感字段“过去 30 天流转路径”的秒级检索。

5. 数据流转意图识别与风险语义分析利用 LLM 对多源日志进行联合语义分析，不仅仅识别“发生了什么”，更侧重推理“为什么”。区分“正常批量导出”与“恶意爬取”“合规运维”与“特权滥用”，为溯源图谱附加动态风险语义标签。

四、成果要求

1. 学术/理论成果：

· 发表 CCF-A/B 类会议/期刊或 SCI 期刊论文不少于 2 篇（重点涵盖日志因果推断、意图识别、跨模态水印方向）。

· 申请或授权国家发明专利不少于 3 项。

· 形成一份《企业零侵入式数据血缘构建与风险溯源技术白皮书》。

2. 工业/落地成果：

· 原型系统：提供一套全流量数据血缘分析平台，展示从 API 点击、DB 访问到文件导出的完整链路。

· 核心组件/SDK：日志关联引擎，无 Trace ID 环境下，API 与 SQL 关联准确率>85%；意图识别模块，对典型高危流转行为的意图识别召回率>90%；跨模态水印工具，支持 Excel/PDF/图片的水印嵌入与提取。

· 验证报告：在至少 1 个实际企业环境（日均亿级 SQL 日志）完成部署验证，证明方案有效性。

（具体成果形式可根据研究方向适度调整）

08 项目需求名称：

可信数据合成与隐私安全关键技术攻关

一、揭榜方向与总体目标

旨在通过合成数据的数据生成、质量评估与隐私保护等关键方向的技术突破，形成可实际落地、成本可控、可监管可追溯的合成数据生产与治理能力。

二、课题背景与必要性

大模型（LLM、VLM 等）对高质量数据的依赖日益增强，但医疗、金融、政务等重点领域的真实数据受制于隐私保护、合规政策与商业机密，难以在机构之间直接共享，形成“数据可见不可用”的突出矛盾。合成数据被视为缓解数据稀缺、提高数据质量、合规约束的重要手段，但现有技术在以下方面仍存在明显短板：一是合成数据在统计分布、结构特征与业务语义上的保真度不足，易导致模型偏差扩大或性能退化；二是缺乏系统化的隐私风险量化与可验证机制，难以证明合成数据不会泄露原始个体信息；三是在真实业务场景下尚未形成标准化的质量评估指标体系与工程化工具链，难以支撑跨行业、跨场景复用。如何在隐私安全前提下大幅提升合成数据的真实性、有效性与可用性，并与大模型训练与评测流程深度融合，是当前亟待攻克的关键问题。

三、重点攻关任务

（申请团队可围绕下列方向选择一个或多个切入点，也可提出全新、高价值的挑战方向。）

1. 高保真合成数据的生成机制

· 研究能够有效捕捉真实数据高阶特征关系、跨字段结构依赖与时间模式的生

成方法，实现对不同数据类型（表格、时序、文本、图像）的高保真合成。

· 研究可控生成机制，通过特征约束、稀疏事件建模、分布对齐等策略，实现对关键统计规律和业务结构的精准表达。

· 针对长尾样本、罕见事件和异常模式等易缺失区域，研究增强式生成策略，提升合成数据的覆盖性和多样性，缓解模式坍缩与分布退化问题。

2. 多维合成数据质量评估体系

· 构建评价合成数据质量的指标体系，覆盖统计一致性、结构依赖保持度、样本多样性、稀疏特征保留能力、异常模式还原度等多个维度。

· 建立适用于不同数据类型的结构一致性检测方法，有效识别字段关系破坏、时间错序、关键特征丢失等生成缺陷。

· 建立面向典型行业场景（如医疗、金融、网络安全、政务服务等）的合成数据评测基准与规范化流程，形成可复用评测样例库与脚本工具。

3. 合成数据的隐私风险建模、攻防验证与安全使用边界

· 建立面向合成数据的隐私泄露风险模型，系统分析不同生成机制对成员推断、属性推断、分布反演等攻击的敏感性。

· 研究隐私保护与信息保留之间的量化关系，明确在不同隐私预算或生成策略下敏感特征的泄露概率与可恢复程度。

· 融合差分隐私、联邦学习、机密计算等技术，探索合成数据的隐私风险缓解机制，包括特征敏感度抑制、合成空间限制、扰动注入、训练结构改造等方法，提高合成数据应用的安全可靠性。

四、成果要求

1. 核心考核指标

· 合成数据质量与可用性：在典型下游任务中，基于合成数据训练/微调的大模型性能与基于真实数据训练模型的关键指标差距不超过约定阈值(如精度、召回、AUC 等），合成数据对模型坍塌的影响不超过约定阈值。

· 隐私与合规性：在成员推断、属性推断等典型隐私攻击场景下，攻击成功率显著低于真实数据基线，满足差分隐私或相关合规要求中约定的隐私预算/风险上限。

· 完整与系统性：合成数据的质量评估体系选取的评估对象需要具备真实场景意义，质量评估体系的构建需要具备系统性，能充分评估约定场景下所有数据情况；能给出质量评估体系的理论性依据与推导。

· 可迁移性与可扩展性：合成数据生成与评估能力可迁移至至少 2 个行业或区域场景，在不同数据规模下具备良好的扩展性与稳定性。

2. 交付物要求

· 可运行的合成数据生成与评估原型系统/工具：提供便于复测的原型平台或工具包，支持至少 1 类结构化数据与 1 类非结构化数据（如文本或图像）的合成与评估，鼓励提供开源源码或可复现脚本。

· 典型行业示范数据集与实验报告：形成覆盖若干典型业务场景的合成数据样例及其与真实数据对比的实验报告，包含质量评估指标、隐私风险评估结果与典型应用案例。

· 技术白皮书/报告：详细阐述所解决的问题、技术路线与系统架构设计，给出合成数据生成方法、评估体系、隐私保护机制、工程化实现路线及实践经验总结。

· 测试报告：基于不同业务场景的实测结果，系统展示合成数据的综合效果。

（具体成果形式可根据研究方向适度调整）