全频谱v2.0技术实施

—— 从协议到生产线 ——

信贷AI治理框架工程化落地手册

汇报人/部门：架构组 | AI治理委员会

目标与受众

核心目标

对齐边界

明确业务系统与治理层的职责划分，消除模糊地带，确保各司其职。

统一接口

定义5大核心组件的交互契约与标准协议，建立高效协作的技术基础。

落地路径

确立分阶段的系统集成方案与详细验收标准，保障项目稳步推进与交付。

涉及团队

架构组

负责整体架构设计与规范制定

后端组

负责核心组件的开发与集成

ML组

负责模型的解释性与效果优化

QA组

负责全流程的质量控制与验证

DevOps

负责持续集成、部署流水线搭建与系统运维支持

架构全景（系统拓扑）

业务应用层 (Business Layer)

- 信贷审批系统 (Core Business)
- 智能风控引擎 (Risk Engine)
- 全渠道营销系统 (Marketing)

治理组件层 (Governance Layer) - 独立部署

- **Collector**
  数据采集
- **API Gateway**
  流量入口
- **BSRM**
  服务治理
- **MQ Broker**
  异步解耦
- **TDP Plugin**
  协议转换
- **Envoy Proxy**
  高性能代理

核心治理原则：全频谱非侵入式介入

治理层独立于业务层部署，不侵入核心业务代码。通过Sidecar和Filter模式透明介入，在保障业务连续性的同时，实现流量监控、熔断降级与服务治理的全生命周期覆盖。

组件1 - FSHI健康指数引擎

输入源 (Input)

业务指标：RT / 错误率 (Prometheus)
合规指标：隐私 / 全链路溯源 (Audit Log)
自定义指标：S_meaning (业务系统回调)

计算核心 (Core)

$$FSHI = 0.4S + 0.3R + 0.3M - P$$

S=Survival(存活) R=Relation(关联)
M=Meaning(意义) P=Penalty(惩罚)

输出 (Output)

智能告警：触发阈值推送 (AlertManager)
状态快照：全量指标存储 (Redis)
实时流：增量数据订阅 (Kafka)

FSHI 健康指数引擎是AI模型的“实时体检中心”。它聚合了来自监控系统、审计日志及业务侧的多维数据，通过预设的加权算法模型计算出量化的健康分数。该分数不仅用于即时触发告警以响应异常，还会将结果快照持久化存储，并通过消息流实时推送，为后续的趋势分析和模型优化提供核心数据支撑。

组件2 - 分层治理控制器

细胞 (Cell)

判定阈值：FSHI < 0.2
当前状态：系统完全隔离

器官 (Organ)

判定阈值：0.2 ⩽ FSHI < 0.7
当前状态：动态观察监控

身体 (Body)

判定阈值：FSHI ⩾ 0.7
当前状态：系统正常运行

严格只读权限

仅允许对外部请求进行读取、解析和数据记录，严禁执行任何可能改变系统状态的操作。

受控执行权限

仅允许执行经过预定义的低风险操作，且操作过程受实时监控，一旦异常立即回滚。

完全自治权限

拥有系统的完整操作权限，支持执行复杂的规划与执行任务，系统处于高度自治的安全模式。

组件3 - TDP可信解释网关

AI 推理的“安检员”与“解说员”

在 AI 服务的入口处，为每一次请求建立完整的可信链路。通过动态注入身份标识、实时生成决策解释、精确记录资源消耗，解决 AI 黑盒决策不可追溯、不可解释的痛点。

核心功能特性

- **身份注入** — 自动为请求分配唯一DID，确立数据权属。
- **实时解释** — 调用模型服务，同步生成推理逻辑与决策依据。
- **成本计量** — 量化计算资源消耗，生成精准的成本账单。

标准交互链路

用户请求 (API Gateway) → 网关处理 (注入DID / 生成解释) → 后端服务 (附加成本标签)

最终返回：带有身份凭证、推理依据与成本明细的可信响应结果

组件4 - SMP考试服务

01 / 场景题库

基于真实信贷业务场景构建
涵盖高风险拒贷解释等
典型考核案例库

02 / 沙箱执行

基于Docker构建隔离运行环境
支持断网运行与超时熔断
确保模型执行的安全性

03 / 多维评分

客观题结果自动匹配判定
主观题采用人工双评+仲裁
建立严谨公正的评分机制

04 / 链上发证

考核通过颁发VC格式数字证书
基于区块链技术实现链上存证
全流程可追溯、可查验

核心价值：构建 AI 模型的标准化准入与能力认证体系，从源头确保上线模型的可靠性、安全性与合规性

组件5 - BSRM 黑天鹅熔断器

主动防御 · 危机阻断核心机制

基于多源实时监测与动态规则引擎，在系统风险阈值触发时，毫秒级执行保护动作

监测输入 SOURCES

- FSHI 市场异常流动指数监测
- 全网突发外部风险新闻舆情
- 监管机构实时风险预警信号

规则判定 DSL LOGIC

```
IF regional_fshi < 0.1 AND news_contains('重大风险') THEN throttle_down()
```

执行动作 ACTIONS

- **节流 (Throttle)**：动态降低业务流量峰值
- **降级 (Scale)**：切换至备用简化规则集
- **熔断 (Break)**：毫秒级暂停高危服务调用

数据流与审计

事件捕获
全链路操作实时监听

日志持久化
结构化数据不可篡改

合规接口服务
权限管控与数据隔离

事件总线架构

核心 Topic：gov.events
统一采用 CloudEvents 标准格式，确保事件结构的规范性与跨服务的兼容性。

审计日志 Schema

核心关键字段：
timestamp, agent_did, action, fshi_before/after, penalty, tdp_shown

监管合规接口

提供只读 GraphQL API
专门面向监管机构开放，支持复杂条件的审计日志查询与历史数据追溯。

集成策略（三步走）

Phase 1
影子模式 (Shadow)

核心部署
采集组件 + TDP网关接入

关键动作
仅监控与日志记录，不干预业务运行

预计周期
4 周 (摸清现状与基线)

Phase 2
权限拦截 (Block)

核心部署
分层权限管控 + SMP考试认证体系

关键动作
关键岗位必须持证上岗，违规操作拦截

预计周期
8 周 (建立标准与规范)

Phase 3
自动熔断 (Auto-Fuse)

核心部署
BSRM 熔断规则库全面配置生效

关键动作
系统自动降级（初期辅以人工二次确认）

预计周期
12 周 (实现自动化治理闭环)

里程碑与验收计划

M1 交付周期：2周

FSHI采集Pipeline上线

核心指标数据全量入库
实现自动化采集与清洗

M2 交付周期：4周

TDP身份注入生效

用户侧UI身份信息实时可见
鉴权逻辑与数据打通

M3 交付周期：8周

首轮SMP考试完成

完成首批模型守庙人认证
确保5+核心人员持证上岗

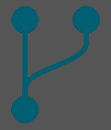
M4 交付周期：12周

熔断演练顺利通过

高并发场景下自动降级生效
验证系统的稳定性与容错能力

资源与下一步

资源链接 RESOURCES



核心代码库

github.com/blackswan-ai-immunity/full-spectrum-ethics

项目接口文档

Confluence / 接口规范 v2.0 (已更新)

所有资源已同步至团队内部知识库

下一步行动 NEXT STEPS

01

组件负责人认领

各组需在本周三前完成核心组件Owner的认领与登记。

02

接口契约细化

基于 AsyncAPI / Protobuf 标准，完成所有微服务接口定义。

03

架构方案评审

下周一下午召开架构评审会，确认最终技术落地实施方案。