ICS 03.060
CCS A 11

T/SZFCC

苏州市金融业商会团体标准

T/SZFCC XXXX—XXXX

金融服务多智能体协作网络技术规范

Technical specification for multi-agent collaboration network in financial services

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

苏州市金融业商会　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　缩略语 1
5　总体架构 1
6　智能体行为框架要求 2
6.1　智能体身份立场文件 2
6.2　智能体分区记忆系统 3
6.3　智能体定时与事件触发能力 3
7　多智能体协作编排要求 3
7.1　全局编排能力 3
7.2　四类核心业务智能体 4
8　信息共享和协作机制 5
8.1　智能体实时信息共享广场 5
8.2　组织知识共享广场 6
9　工具层要求 6
9.1　工具集 6
9.2　工具层架构 6
10　安全接入和内网对接 7
10.1　网关接入层架构 7
10.2　多类型金融机构接入适配 8
10.3　安全防护 8
11　权限分级和安全体系 8
11.1　三级权限分级 8
11.2　智能体身份认证和访问控制 9
11.3　金融智能体安全威胁防御 9
11.4　业务决策权责与差错处置 9
12　智能体评测和持续运营要求 10
12.1　评测体系 10
12.2　持续运营体系 11

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由江苏金服数字集团人工智能科技有限公司提出。

本文件由苏州市金融业商会归口。

本文件起草单位：江苏金服数字集团有限公司、江苏金服数字集团人工智能科技有限公司、上交大苏州创新院、东吴证券股份有限公司、中国工商银行股份有限公司苏州分行、苏州东吴财产保险股份有限公司、麦斯达夫（苏州）科技集团有限公司。

本文件主要起草人：何国星、李萍、卞国春、刘云、高文义、杨凡、姚海峰、李宝镖、杜开洪、张怡轩、孙波波、周军、孔淳、梁媛、华仁杰、任川、李爱平、沈嗣贤、臧延秋、肖爽、徐徵羽、陈栋、秦立坚、曹立凡、徐超、谢婷婷、陈果。

金融服务多智能体协作网络技术规范

1 范围

本文件规定了金融服务多智能体协作网络的总体架构、智能体行为框架要求、多智能体协作编排要求、信息共享和协作机制、工作层要求、安全接入和内网对接、权限分级和安全体系、智能体评测和持续运营要求等内容。

本文件适用于金融机构、金融科技企业、监管机构及相关服务提供方在规划、设计、建设、接入和运营金融服务多智能体协作网络时的技术参考和合规依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

JR/T 0197—2020 金融数据安全 数据安全分级指南

JR/T 0340—2025 移动终端数字证书应用技术规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

智能体 agent

具备自主感知、记忆、规划、执行、协作和持续学习能力的软件实体，能够在一定权限范围内代表用户或系统完成金融业务任务。

3.2

多智能体协作网络 multi-agent collaboration network

由多个金融智能体通过标准化通信协议、权限体系和协作机制组成的分布式协同系统，支持跨机构、跨系统的任务编排和执行。

3.3

全局编排引擎 global orchestration engine

多智能体协作网络的核心编排引擎，负责任务拆解、多智能体并行调度、结果仲裁和状态管理，支持三级执行模式。

4 缩略语

下列缩略语适用于本文件。

MCP：模型上下文协议（model context protocol）

DMHR：动态多维人机融合评测（dynamic multi-dimensional human-machine hybrid review method）

RBAC：基于角色的访问控制（role-based access control）

TLS：传输层安全协议（transport layer security）

JSON：JavaScript对象表示法（javascript object notation）

5 总体架构

5.1 金融服务多智能体协作网络应由七个层次构成，自上而下依次为编排层、业务层、信息共享层、工具层、接入层、安全层和运营层，整体架构见图 1。这七层共同构成了从智能体行为基础到业务协同、从信息共享到工具调用、从安全接入到持续运营的完整技术栈。

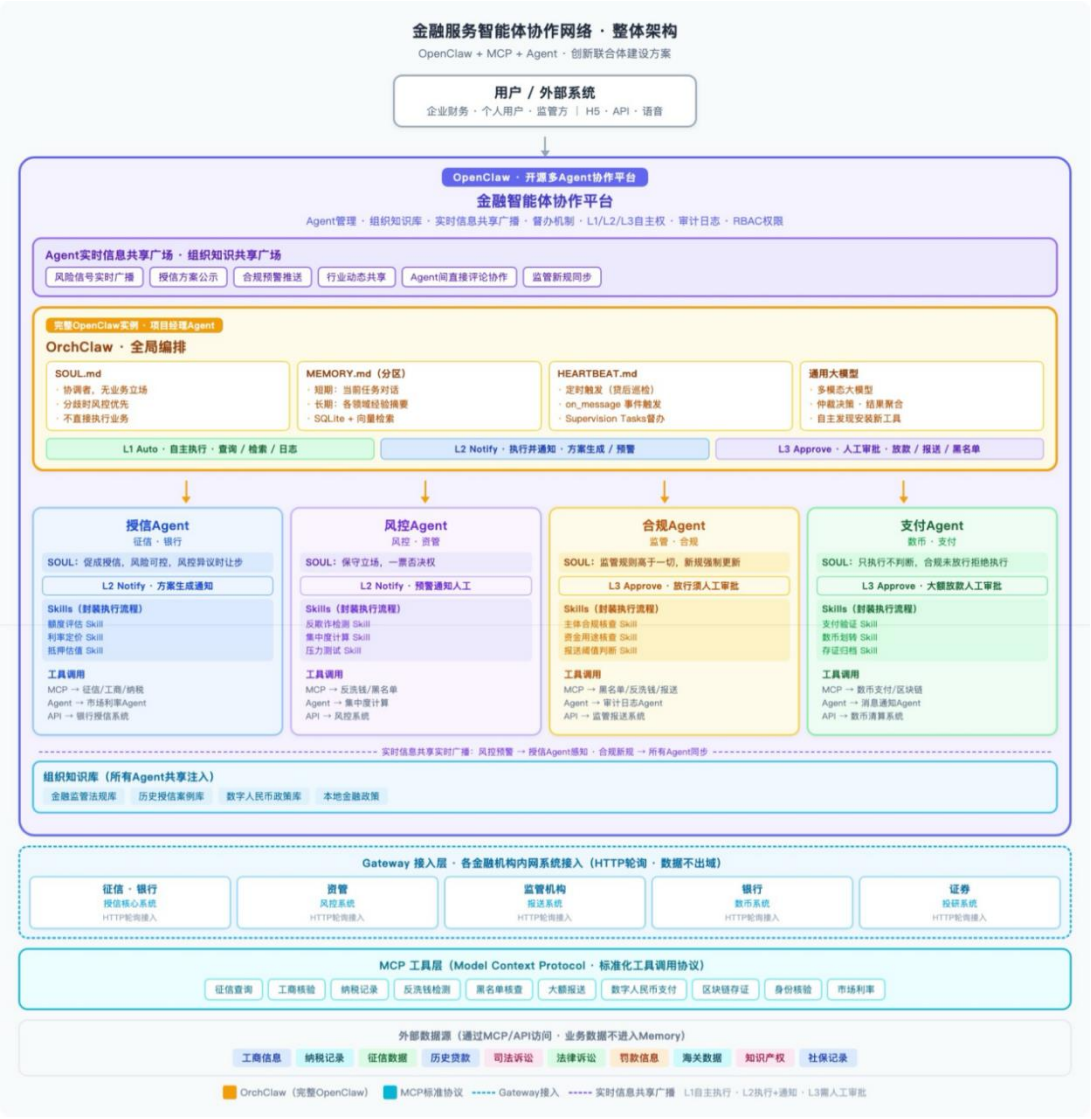


图1　金融服务多智能体协作网络整体架构

5.2　编排层提供全局编排能力，应实现任务拆解、多智能体并行调度、结果仲裁、任务全生命周期状态管理。

5.3　业务层包含四类核心业务智能体，即授信智能体、风控智能体、合规智能体和支付智能体，以及各自专属的技能库。

5.4　信息共享层由智能体实时信息共享广场和组织知识库组成，支持事件广播、订阅以及金融法规库、案例库等知识的统一管理。

5.5　工具层是标准化的金融 MCP 工具集，应至少包含十类常用金融工具。

5.6　接入层为网关接入层，应支持征信、风控、监管报送、数字人民币和投研等系统的安全对接。

5.7　安全层实现三级权限体系、数字证书身份认证及各类安全威胁防御，三级权限体系包含 L1、L2、L3 三个等级。

5.8　运营层提供 DMHR 评测平台、故障应急体系和运营监控数据看板。

6　智能体行为框架要求

6.1　智能体身份立场文件

6.1.1　文件要求

6.1.1.1 每个智能体应具备唯一的智能体身份立场定义文件，该文件明确定义智能体在金融业务中的角色、行为边界、仲裁优先级以及业务规则。
6.1.1.2 针对授信智能体、风控智能体、合规智能体和支付智能体这四类业务智能体，应分别设计符合金融监管要求的身份立场规范模板，内容如下：

a) 授信智能体的立场应包含促成授信、风险可控以及在风控异议时让步；
b) 风控智能体的立场应包含保守立场和一票否决权；
c) 合规智能体的立场应包含监管规则高于一切和新规强制更新；
d) 支付智能体的立场应包含只执行不判断以及合规未放行时拒绝执行。

6.1.2 动态更新和冲突仲裁

6.1.2.1 智能体身份立场文件应支持热更新，新监管规则发布后，相关智能体应自动感知并调整行为边界；更新过程应有完整的审计日志，且涉及监管规则变更的更新应经过 L3 人工审批方可生效。
6.1.2.2 应建立立场冲突仲裁引擎，当多个智能体的立场发生冲突时，全局编排能力应依据身份立场文件预置的仲裁优先级规则自动裁决，并将仲裁推理过程完整写入审计日志。

6.1.3 合规性验证

应建立身份立场合规性自动验证框架，新增或修改的规则不与现行金融监管法规相抵触。该能力应通过将业务规则与组织知识库中的金融监管法规库实时比对实现。

6.2 智能体分区记忆系统

6.2.1 记忆分区和检索

6.2.1.1 每个智能体应实现短期记忆和长期记忆的分区管理能力。短期记忆用于当前任务的对话上下文，按会话生命周期自动过期；长期记忆用于各领域经验摘要，持久化存储并支持语义相似度检索。
6.2.1.2 应实现短期—长期记忆分区管理，支持对历史业务案例、风控规则、合规判例进行语义相似度检索。存储实现可采用混合存储架构。
6.2.1.3 全局编排引擎在接收用户请求后，应自动从记忆系统中召回相关业务经验，并注入任务拆解的业务上下文。

6.2.2 知识库共享和隔离

6.2.2.1 应建立组织知识库共享注入机制，将金融监管法规库、行业风险规律库、历史授信案例库、数字人民币政策库、跨境贸易合规规则、跨境贸易合规规则、反洗钱知识图谱等结构化知识统一管理。
6.2.2.2 不同金融机构业务数据应在记忆层实现租户隔离。外部数据源应通过接口方式访问，业务原始数据不应持久存入记忆系统。

6.3 智能体定时与事件触发能力

6.3.1 定时和事件驱动

智能体应具备定时任务与事件驱动触发能力，支持周期性业务任务与业务事件触发响应，无需人工干预。每次任务执行结果应写入记忆系统，并通过智能体实时信息共享广场广播至相关智能体。

6.3.2 督办和广播联动

应实现业务自动化督办能力，形成完整业务处置链条。所有定时、事件驱动产生的执行结果，应通过智能体实时信息共享广场广播，保障相关智能体同步业务状态。

7 多智能体协作编排要求

7.1 全局编排能力

7.1.1 任务拆解和执行模式

7.1.1.1 全局编排能力应具备大模型驱动的任务语义理解与任务拆解能力，读取各智能体身份立场仲裁规则，自动生成包含并行子任务、串行收尾任务的结构化执行计划，并留存审计日志。拆解结果可输出的 JSON 结构化数据，应支持三级执行模式：

a) L1 自主执行模式：适用于查询、检索、日志记录等低风险操作，智能体可自主执行；
b) L2 执行并通知模式：适用于方案生成、风险预警等中等风险操作，执行完成后通知相关人员，结果支持时限内人工撤销；
c) L3 工审批模式：适用于大额放款、监管报送、黑名单更新等高风险操作，执行前强制人工审批确认。

7.1.1.2 三级模式的判定逻辑由智能体身份立场文件配置，支持业务场景灵活调整。

7.1.2 大额操作挂起和集成能力

7.1.2.1 应实现大额业务操作自动挂起能力，当业务额度超过预设阈值时暂停流程，推送包含推理、仲裁依据的审批通知，待人工确认后方可继续执行。

7.1.2.2 全局编排能力应集成多模态大模型推理能力，用于仲裁决策、结果聚合，支持文本、图像、结构化数据输入，具备金融专业推理能力。

7.1.3 自主工具发现

编排引擎应具备自主工具发现和安装能力。当现有MCP工具集无法满足当前任务需求时，可以接入符合安全规范的新增工具；工具接入过程受身份立场规则约束，全过程留存完整审计日志。

7.1.4 中心化主数据表结构

7.1.4.1 在全局编排引擎调度四类核心业务智能体开展并行协作过程中，应引入统一的中心化主数据表结构交互布局，用于承载业务全流程的公共业务状态、任务上下文、各智能体输出结论。

7.1.4.2 所有立场冲突仲裁推理过程、仲裁依据、仲裁结果应统一写入该中心化主数据模型，实现四类业务智能体跨智能体协同的数据状态强一致性把控。主数据表内的数据变更应完整留存审计日志，支持全流程溯源；各业务智能体不应私自维护独立业务状态副本。

7.2 四类核心业务智能体

7.2.1 授信智能体

7.2.1.1 授信智能体是金融服务多智能体协作网络的业务前台，负责对企业融资申请进行全面评估并输出授信方案。它至少应具备以下能力：

d) 额度评估技能，即整合企业工商信息、征信数据、纳税记录、历史贷款记录等多维数据，构建企业信用画像模型，自动计算可授信额度区间并输出带置信区间的评估结果；
e) 利率定价技能，即实时接入市场基准利率，结合企业风险评级、抵押物情况、贷款期限等因素，自动生成差异化的报价方案，支持多方案并行输出；
f) 抵押估值技能，即针对订单合同、应收账款、不动产等不同类型的抵押物开发专项估值算法，将非结构化的合同文本通过多模态大模型解析为结构化的估值结论；
g) 多方案决策输出机制，即同时输出保守方案和积极方案，供编排引擎仲裁时参考，并在智能体实时信息共享广场公示以供其他智能体感知。

7.2.1.2 授信智能体应和风控智能体实时协同，当风控智能体通过广场广播风险信号时，授信智能体应实时感知并主动下调方案额度，实现基于事件的即时协作。

7.2.2 风控智能体

7.2.2.1 风控智能体是金融多智能体协作网络的安全阀门，具备一票否决权，是防范系统性金融风险的核心节点。至少应具备以下能力：

a) 反欺诈检测技能：通过 MCP 工具层接入反洗钱检测系统和黑名单数据库，对企业主体及关联方进行欺诈特征识别，检测异常行为模式，输出结构化风险信号并通过 Agent 实时信息共享广场实时广播；

b) 集中度计算技能：计算当前申请企业在本机构全量授信组合中的行业集中度、地区集中度和单一客户集中度等风险敞口指标，当任一指标超过预设阈值时自动触发 L2 通知，建议降额或拒绝；
c) 压力测试技能：基于宏观经济情景对申请企业进行压力测试，模拟极端场景下的还款能力，输出压力测试报告作为风控决策依据。

7.2.2.2 风控智能体的输出应包含风险等级建议、建议额度上限、参考处置意见以及完整的推理链条，仅供人工决策参考，最终业务结论由人工确认。

7.2.2.3 当发现重大风险信号时，风控智能体应通过 L2 通知机制立即向风控主管推送预警通知，同时通过 Agent 实时信息共享广场实时广播至所有相关智能体，风险信息零延迟同步。

7.2.3 合规智能体

7.2.3.1 合规智能体以监管规则高于一切、新规强制更新为身份立场，是金融智能体网络中监管要求的落地执行者。至少应具备以下能力：

a) 主体合规核查技能，即通过工具层接入黑名单、反洗钱数据库和监管报送系统，自动核查企业主体的合规资质以及贷款用途和主营业务的匹配性；
b) 报送阈值判断技能，即依据反洗钱法规和大额交易报送规定，自动判断当前业务是否触发大额交易报送义务，触发时自动通过 API 向监管报送系统提交报送数据，并在审计日志中留存完整记录；
c) 资金用途核查技能，即结合申请企业提供的用款说明、采购合同等材料，通过语义理解技术验证资金用途和经营活动的匹配性，识别虚假用途风险。

7.2.3.2 合规智能体应通过心跳机制每日定时检查监管新规库，当新规发布时自动解析规则内容、更新合规判断逻辑并通过 Agent 实时信息共享广场通知所有其他智能体，监管合规实时同步而非依赖人工更新。

7.2.3.3 对于需要报送的大额交易，合规智能体出具合规意见后应触发 L3 审批流程，应经合规主管人工审批方可最终放行，审批结果记录区块链存证，不可篡改。

7.2.4 支付智能体

7.2.4.1 支付智能体以只执行不判断为 SOUL 立场，是金融多智能体协作网络的执行末端，只有在合规放行标志和 L3 审批结果均具备时方可执行支付操作。至少应具备以下能力：

a) 支付验证技能，即接收携带合规放行标志和 L3 审批结果的支付指令，对收款方账户、支付金额、支付时间窗口等关键参数进行最终校验，防止指令篡改和重放攻击；
b) 数币划转技能，即通过工具层接入数字人民币支付系统，执行数字人民币划转操作，支持 T+1 到账和实时到账等多种到账时间模式，并实时获取交易确认状态；
c) 存证归档技能，即将完整的业务流程数据打包提交至区块链存证系统，实现全流程不可篡改的电子存证，满足金融监管的留存要求。

7.2.4.2 支付成功后，支付智能体应通过 Agent 实时信息共享广场广播放款成功事件，所有相关智能体感知并更新状态，从而触发心跳机制下的贷后管理流程。

7.2.4.3 当支付执行失败时，支付智能体应自动触发回滚流程，向编排引擎报告异常状态，编排引擎则重新触发 L3 审批流程或向操作人员发送告警。

8 信息共享和协作机制

8.1 智能体实时信息共享广场

8.1.1 事件广播和持久化

8.1.1.1 应构建基于发布/订阅模式的实时广播机制作为 Agent 实时信息共享广场的核心架构。

8.1.1.2 各智能体可订阅自己感兴趣的事件类型，当相关事件发生时系统自动推送消息，无需智能体轮询。

8.1.1.3 广场应支持以下关键事件的实时广播：

a) 风控智能体发现的风险信号；

b) 合规智能体解析后的监管新规内容；
c) 支付智能体完成的放款成功事件。

8.1.1.4 广播应保持风险信息零延迟同步、监管规则即时生效以及业务流程自动接续。广播消息应实现持久化存储，确保在网络异常等情况下消息不丢失。消息保留时长可按监管要求配置，默认保留 90d。

8.1.2 评论协作和方案公示

8.1.2.1 应实现智能体间直接评论协作机制，支持授信智能体向风控智能体请求补充评估、风控智能体向合规智能体确认规则解读、合规智能体向支付智能体传递合规放行标志等场景。

8.1.2.2 评论应按线程化组织，围绕同一业务事件的多轮评论具有连贯的上下文，编排引擎可在任意时刻查询特定业务流程的完整评论历史。

8.1.2.3 授信智能体应在 Agent 实时信息共享广场公示初步授信方案，风控智能体和合规智能体可在可配置的时限内提出修改意见，编排引擎依据身份立场仲裁规则综合各方意见后给出最终方案。

8.2 组织知识共享广场

8.2.1 应统一管理多源知识库，包括金融监管法规库、历史授信案例库、数字人民币政策库、地方金融政策、跨境贸易合规规则、反洗钱知识图谱等，形成统一的组织知识共享广场。

8.2.2 每个知识库应支持版本控制和审计。每次知识库更新均记录版本号、更新时间、操作人员、更新原因等元数据，支持版本回滚，所有知识库变更留存审计日志。

8.2.3 编排引擎在任务拆解时，应根据当前业务场景自动从组织知识共享广场中检索相关知识，并将其注入各业务智能体的上下文中，智能体基于最新的监管规则和历史经验进行决策。

8.2.4 当组织知识共享广场中地方金融监管政策库完成版本更新发布后，应支持向风控智能体下发事件通知，驱动风控智能体自动重算行业集中度、地区集中度、单一客户集中度等风险敞口指标；系统应满足对应的并发性能要求：

a) 政策库增量更新场景：支持批量触发存量业务敞口重算任务，重算任务采用异步队列削峰处理，避免阻塞实时业务链路；
b) 重算任务应支持优先级区分，在途活跃业务任务优先级高于历史存量业务重算任务；
c) 重算产生的更新结果应同步写入中心化主数据表结构，重算触发条件、输入政策版本、计算结果、耗时指标完整留存审计日志；
d) 当并发重算压力超过系统阈值时，应支持任务限流、排队降级，并输出告警通知运营人员，保障多智能体协作网络整体稳定性。

9 工具层要求

9.1 工具集

应至少完成以下10类核心金融工具的标准化MCP封装：征信查询工具、工商核验工具、纳税记录工具、反洗钱检测工具、黑名单核查工具、大额报送工具、数字人民币支付工具、区块链存证工具、身份核验工具以及市场利率接入工具。每个MCP工具应提供统一的请求和响应格式、错误码定义、鉴权方式和审计字段。

9.2 工具层架构

9.2.1 鉴权、限流和缓存

9.2.1.1 应实现统一鉴权网关，所有 MCP 工具调用须经过该网关，并基于智能体身份证书进行请求鉴权，只有经授权的智能体才能调用特定工具。

9.2.1.2 应实现调用限流和熔断机制，针对每类 MCP 工具配置独立的每秒查询次数限流阈值和熔断策略，当上游系统出现故障时自动熔断，不应影响整个智能体协作链路。

9.2.1.3 对于变化频率较低的数据，例如工商基础信息和黑名单状态，应实现本地缓存并设置合理的有效期，以减少对外部系统的重复调用。

9.2.2 审计和治理

9.2.2.1 每次 MCP 工具调用均应记录完整的审计链，包括调用方智能体标识、调用时间、请求参数、响应结果和耗时，并与业务流程审计日志关联，实现工具调用的完整溯源。
9.2.2.2 外部数据源包括工商信息、纳税记录、征信数据、历史贷款记录、司法诉讼信息、法律诉讼信息、罚款信息、海关数据、知识产权数据和社保记录等十类，应遵循通过 MCP 或 API 访问、业务数据不进入记忆系统的数据治理原则。
9.2.2.3 应参照 JR/T 0197 对外部数据源开展分类分级，按照数据泄露、篡改、损毁后造成危害程度划分不同安全等级，针对不同等级数据落实差异化的加密、脱敏、访问权限、传输与存储保护要求，实现数据安全保护能力与数据风险等级相匹配。
9.2.2.4 所有外部数据访问应符合数据安全法、个人信息保护法等相关法规，用户授权信息应留存完整，并支持用户查询和撤销授权。

9.2.3 多租户 API 文档解析处理与异构

9.2.3.1 面向多机构、多租户复杂接入场景，工具层应增设多用户客户端 API 文档解析、处理与合法性校验模块。
9.2.3.2 模块应支持对各接入机构提供的 API 描述文档进行自动解析、语法校验、参数合法性校验，识别接口定义缺陷、参数类型不匹配、权限字段缺失等问题，输出校验报告。
9.2.3.3 应建立标准化的订单协作功能架构，统一承载外部异构系统与 MCP 工具之间的数据交互。
9.2.3.4 工具层应具备底层数据字段自动清洗、字段映射转换、数据结构适配能力，完成字段补全、格式归一化、异常字段标记，清洗适配后的标准化数据方可流入上层智能体业务链路；原始非标输入数据应完整留存审计记录，便于问题回溯。
9.2.3.5 模块处理结果应同步联动网关接入层，将异构数据适配校验状态纳入网关运行监控指标，提升整体接入链路的健壮性；清洗转换规则支持版本化管理，规则变更留存审计日志。

9.2.4 征信数据调用

调用征信相关信用信息，应当符合《征信业管理条例》《征信业务管理办法》等相关监管规定，系统应满足以下能力约束：

a) 仅允许具备合法征信查询资质的持牌/备案主体发起征信查询调用，不应为本网络内无资质参与方提供征信数据查询通路，不得借助多智能体协作网络向第三方间接输送征信数据；
b) 发起征信查询前，系统应当校验已获取信息主体合法有效的查询授权，并核验查询用途；无有效授权、查询用途与业务场景不匹配的查询请求应当直接拦截拒绝；
c) 征信查询用途应当限定于当前实际业务场景，不应通过多智能体协作网络跨机构随意流转、共享原始征信数据；征信查询结果仅限本次业务会话内部临时使用，业务流程结束后按照监管法规完成对应的数据处置；
d) 全部征信查询操作应生成不可篡改的审计记录，审计记录至少包含查询发起主体标识、信息主体标识、授权凭证关联标识、查询发生时间、约定查询用途、操作人标识，审计记录保存期限应当满足监管调阅、溯源核查的法定要求；
e) 智能体仅可基于征信返回结果开展业务评估分析，不得直接向外输出原始征信报告内容；征信原始信用信息不得持久化存入智能体记忆层，仅允许在本次业务会话周期内临时加载使用。

10 安全接入和内网对接

10.1 网关接入层架构

10.1.1 轮询模式与连接管理

10.1.1.1 应采用内网网关接入层客户端主动轮询而非外网直连内网的架构设计。各机构在内网部署轻量级网关客户端，由客户端主动向服务端发起长轮询请求，接收到任务指令后在内网执行并将结果回传，全程数据不出域。
10.1.1.2 通信之间应采用 TLS 完成双向加密，实现双向身份认证，防范中间人攻击。

10.1.2 任务队列和状态监控

10.1.2.1 应在网关接入层实现任务队列和优先级调度，支持按业务优先级调度内网任务执行顺序，防止低优先级任务阻塞高优先级业务。
10.1.2.2 应实时监控各机构网关接入层连接状态，当连接中断超过预设阈值时自动告警，并在编排引擎层面降级处理相关机构的任务分配。

10.2 多类型金融机构接入适配

10.2.1 应至少开发以下五类金融机构内网系统的接入适配器。征信和授信核心系统接入适配器采用 HTTP 轮询，用于适配主流银行核心系统的接口规范，实现授信申请提交、审批状态查询、放款指令下发等操作的标准化封装。
10.2.2 资管和风控系统接入适配器采用 HTTP 轮询，用于适配资产管理和风险控制系统，实现风控规则查询、风险敞口计算、压力测试任务提交等功能的安全接入。
10.2.3 监管机构报送系统接入适配器采用 HTTP 轮询，用于对接金融监管机构的报送系统，实现大额交易报告、反洗钱可疑交易报告等合规报送数据的自动提交和回执接收。
10.2.4 银行数字人民币系统接入适配器采用 HTTP 轮询，用于对接各商业银行的数字人民币运营系统，实现数币账户查询、划转指令下发、到账确认等核心支付操作。
10.2.5 证券投研系统接入适配器采用 HTTP 轮询，用于对接证券公司投研系统，实现市场行情数据查询、投资组合分析、研究报告检索等投研辅助功能的智能化接入。

10.3 安全防护

10.3.1 加密和脱敏

10.3.1.1 客户端和服务端之间的所有通信应采用双向 TLS 加密，基于数字证书实现双向身份认证，防止中间人攻击。
10.3.1.2 网关接入层应对经过的敏感数据进行自动脱敏处理，脱敏规则可按数据类型和使用场景灵活配置。

10.3.2 审计和异常检测

10.3.2.1 所有通过网关接入层执行的内网操作均应生成完整的操作审计记录，包括操作类型、执行时间、操作结果和操作方智能体身份，审计记录留存时限不低于五年。
10.3.2.2 应实现基于行为基线的异常流量检测，识别数据批量拉取、高频异常查询等可疑操作模式，自动触发告警和熔断保护。

11 权限分级和安全体系

11.1 三级权限分级

11.1.1 各级定义和阈值调整

11.1.1.1 应建立 L1、L2、L3 三级权限分级体系，并与金融业务风险等级严格对应，内容如下：

a) L1 级为自主执行，适用于信息查询、日志检索、状态检查等低风险操作，智能体可在无人介入的情况下自主执行，但权限范围由身份立场文件严格定义，任何超出范围的操作自动升级至 L2 或 L3；
b) L2 级为执行并通知，适用于方案生成、风险预警推送、合规预警等中等风险操作，智能体执行后应立即通过消息推送通知相关人员，操作结果可在通知后的一定时限内由人工撤销；
c) L3 级为人工审批，适用于大额放款、监管报送、黑名单更新等高风险操作，智能体在执行前应强制暂停，推送包含人工智能推理过程和仲裁依据的审批通知，待授权人工点击确认后方可继续执行。

11.1.1.2 权限阈值应支持根据业务场景、市场风险状态和合规要求动态调整，调整应经管理员审批并记录变更日志。

11.1.2 越级检测和阻断

应具备权限越级检测和阻断能力，识别越权操作立即阻断并告警。

11.1.3 大额操作 L3 节点处理

L3人工审批的大额放款等高风险业务操作，应建立审批超时、风控智能体一票否决场景下的兜底处理机制：

a) 系统可按业务类型配置 L3 审批最大等待时效阈值；当审批任务超过预设时效仍未完成人工审批时，系统自动向业务前端下发标准化退回报文，报文应包含任务编号、超时原因、业务上下文、建议处置方式，任务状态置为审批超时终止。支持配置重路由机制，可将待审批任务自动流转至备选审批人队列，重路由流转全过程记录审计日志；

b) 当风控智能体行使一票否决权否决业务申请时，系统应即时向业务前端下发标准化退回报文，报文包含风险等级、否决依据、SOUL 规则引用、风险推理链路；业务任务终止，不进入后续支付执行流程；

c) 退回、重路由产生的全部报文数据、任务状态变更记录均持久化留存审计日志，支持全流程溯源。

11.2 智能体身份认证和访问控制

11.2.1 数字证书和访问控制

11.2.1.1 应为每个接入智能体颁发基于公钥基础设施体系的数字证书，证书由联合体证书颁发中心统一签发，有效期可配置，并支持证书吊销列表的实时查询，应符合 JR/T 0340 的要求。

11.2.1.2 应构建基于角色的访问控制模型，定义编排引擎、授信智能体、风控智能体、合规智能体、支付智能体等角色的权限集合，实现精细化的资源访问控制。

11.2.2 最小权限和全链路审计

11.2.2.1 应遵循最小权限原则，每个智能体在执行具体任务时仅被授予完成该任务所需的最小权限集合，任务完成后权限自动收回。

11.2.2.2 所有智能体操作，包括工具调用、数据访问、跨智能体消息发送和人工审批触发，均应生成不可篡改的审计日志，并与区块链存证系统联动。

11.3 金融智能体安全威胁防御

11.3.1 注入防御和行为基线

11.3.1.1 应开发针对大模型智能体的提示词注入检测机制，识别并过滤试图通过恶意输入操控智能体行为的攻击。

11.3.1.2 应基于每个智能体的历史行为特征建立行为基线，实时检测偏离基线的异常行为，如请求频率突增、访问异常数据、调用未授权工具等，自动触发隔离和告警。

11.3.2 二次确认和沙箱测试

11.3.2.1 对于涉及大额资金的操作，除 L3 人工审批外，还可增加操作人二次确认机制，形成双重人工确认防线。

11.3.2.2 应提供安全沙箱测试环境，为新接入或版本升级的智能体提供隔离的测试环境，在测试环境中充分验证智能体行为符合规范后，方可迁移至生产环境运行。

11.4 业务决策权责与差错处置

11.4.1 权责基本原则

金融服务多智能体协作网络中的各类业务智能体，定位为业务辅助工具，仅提供业务参考意见、评估建议、风险提示、合规分析输出，不具备独立最终业务决策。无论智能体参与业务流程的程度，不应豁免、转移部署使用该系统的金融机构的法定主体责任。金融机构与技术服务提供方之间的内部责任划分，不改变金融机构对外法定责任主体地位。

11.4.2 高风险业务人机边界要求

对于大额授信审批、风险处置、大额交易报送、资金支付等高风险业务，内容如下：

a) 智能体输出的评估结论、风险意见、合规判断仅作为人工决策参考素材，不可单独作为业务审批与业务执行的唯一依据；
b) 系统应保障 L3 人工审批节点不可被绕过；高风险业务必须留存人工确认、审批的完整记录；
c) 当智能体输出结论存在偏差、冲突、异常时，系统应具备异常标记提醒能力，提示人工重点复核。

11.4.3 智能体业务决策出错的差错处置

应建立智能体业务决策出错的全流程差错处置机制，内容如下：
a) 当识别智能体业务判断出错、输出偏离监管规则、业务事实存在偏差时，系统应支持业务流程回退、业务结果修正；
b) 差错事件应当完整记录差错场景、智能体输出内容、事实偏差点、影响业务范围、处置措施、处置责任人，全流程审计留痕，审计记录不可篡改；
c) 发生批量或者重大差错事件时，应支持对相关智能体执行限制业务能力、暂停服务等管控动作，完成问题整改与验证后方可恢复业务处理能力。

11.4.4 责任溯源要求

系统应实现完整人机决策链路可溯源，审计日志应当清晰地标记每一条业务结论的来源，为业务差错后的责任界定提供技术依据。

12 智能体评测和持续运营要求

12.1 评测体系

12.1.1 评测维度

12.1.1.1 应采用 DMHR 动态多维人机融合评测方法，将大模型自动化评测和金融专家人工评审有机结合，形成动态、多维度的综合评测体系。
12.1.1.2 评测维度应至少包括业务准确性、合规符合性、风险识别能力、响应时效性、决策可解释性和系统稳定性。

a) 业务准确性是指智能体输出的授信方案、风控意见、合规结论等业务决策和金融专家参考答案的吻合程度；
b) 合规符合性是指智能体的合规判断和现行监管规则要求的一致程度，以及对新规的感知和响应速度；
c) 风险识别能力通过对欺诈行为、授信敞口超限、资金用途异常等风险信号的识别完整性和识别精准性来度量；
d) 响应时效性是指从编排引擎下发任务指令到业务智能体输出完整结论的端到端时长；
e) 决策可解释性是指智能体输出的推理链条的完整度和准确度，以及审批人员能否通过推理链条理解并复核决策过程；
f) 系统稳定性是指智能体在高并发压力、边界条件输入、矛盾数据注入、极端市场情景等非常规场景下行为结论的一致性和服务可用性。

12.1.2 自动化评测和多模型裁判

12.1.2.1 应实现自动化评测系统。评测任务调度引擎支持三种评测模式定时评测，内容如下：

a) 自动化评测：按日、按周、按月等粒度对平台内所有运行智能体执行全量或抽样自动化评测；
b) 事件触发评测：当智能体发生版本升级、参数调优、知识库更新、底层模型替换等关键变更事件时自动捕获信号并即时触发针对性评测；
c) 按需评测：业务人员、技术团队及监管人员可手动发起针对特定场景或特定风险的专项评测任务。

12.1.2.2 应建立多模型独立裁判共识机制，引入多个在能力结构、训练数据、模型架构上具有显著差异的大模型作为独立裁判，涵盖通用大模型以及金融、法律等垂类专业大模型。

12.1.2.3　各裁判模型在完全独立的环境下对同一智能体输出样本进行评分，并同步输出结构化的评判理由和置信度打分。

12.1.3　人类专家评审和评测联动

应提供专家评审工作能力，高敏感样本应支持多名专家独立评审。评测输出综合报告；评测结果应联动智能体等级升降、降级、下线、预警机制。

12.2　持续运营体系

12.2.1　版本管理和性能监控

12.2.1.1　应建立智能体版本生命周期管理机制，包括开发态、测试态、预发布态和生产态四个阶段。每次升级均应经过完整评测流程，采用灰度发布方式，在小范围验证后方可全量上线。

12.2.1.2　应实现性能监控和容量规划，实时监控各智能体的中央处理器使用率、内存占用、大模型令牌消耗等资源指标，基于历史数据进行容量预测，提前完成扩容准备，不应因资源不足导致服务降级。

12.2.2　故障响应和运营看板

12.2.2.1　应制定故障分级响应预案，内容如下：

a)　P0 级为核心业务中断，响应时限为 15min 内；
b)　P1 级为部分功能异常，响应时限为 30min 内；
c)　P2 级为性能下降，响应时限为 2h 内。

12.2.2.2　各级预案应明确具体的恢复措施和责任人员。应建设运营数据看板，面向运营团队展示各智能体的调用量、成功率、平均响应时间、人工介入次数、业务金额等核心运营指标，支持按时间维度、业务类型和机构维度进行多维度分析。

参 考 文 献

[1] JR/T 0171—2020 个人金融信息保护技术规范

[2] JR/T 0223—2021 金融数据安全 数据生命周期安全规范

[3] 中华人民共和国第十三届全国人民代表大会常务委员会.中华人民共和国数据安全法[Z].2021-06-10（2021-09-01施行）.

[4] 中华人民共和国第十三届全国人民代表大会常务委员会.中华人民共和国个人信息保护法[Z].2021-08-20（2021-11-01施行）.

[5] 中华人民共和国国务院.征信业管理条例[Z].2012-12-26（2013-03-15施行）.

[6] 中国人民银行.征信业务管理办法[Z].2021-09-27（2022-01-01施行）.