

基于大语言模型的零信任上下文感知信任评分引擎研究

Research on Zero-Trust Context-Aware Trust Scoring Engine Based on Large Language Models

杨 飞(中讯邮电咨询设计院有限公司,北京 100048)

Yang Fei(China Information Technology Designing & Consulting Institute Co.,Ltd.,Beijing 100048,China)

摘 要:

针对传统零信任信任评分引擎上下文感知不足、泛化能力弱、决策不可解释等问题,研究基于大语言模型(LLM)的零信任上下文感知信任评分引擎。融合用户行为、设备状态、环境信息及威胁情报,利用LLM的上下文理解能力实现动态信任评分与风险归因,优化幻觉抑制与实时推理性能。实验结果表明,该引擎可提升信任评估准确率,降低误报率,为零信任持续验证提供有效支撑。

关键词:

大语言模型;零信任;上下文感知;信任评分;动态信任评估;访问控制

doi:10.12045/j.issn.1007-3043.2026.09.001

文章编号:1007-3043(2026)09-0001-06

中图分类号:TP393.08

文献标识码:A

开放科学(资源服务)标识码(OSID):



Abstract:

Aiming at the problems of insufficient context awareness, weak generalization ability and unexplainable decision in traditional zero-trust trust scoring engines, a zero-trust context-aware trust scoring engine based on Large Language Model (LLM) is proposed. By integrating user behavior, device status, environment information and threat intelligence, the LLM's contextual understanding capability is utilized to realize dynamic trust scoring and risk attribution, with hallucination suppression and real-time inference optimized. Experimental results show that the engine improves the accuracy of trust evaluation and reduces the false alarm rate, providing effective support for zero-trust continuous verification.

Keywords:

Large language model; Zero-trust; Context-aware; Trust scoring; Dynamic trust evaluation; Access control

引用格式:杨飞. 基于大语言模型的零信任上下文感知信任评分引擎研究[J]. 邮电设计技术, 2026(9): 1-6.

0 引言

云计算、物联网等业务快速发展,传统边界安全架构难以应对复杂威胁,零信任成为新一代安全核心。动态信任评估是零信任的关键,现有方法依赖规则与静态特征,存在上下文理解、泛化能力及可解释性不足等问题。大语言模型的技术突破为上下文感知信任评估提供新方向,本文结合零信任环境特点,提出基于大语言模型的信任评分引擎,融合多类信息实现语义级理解与动态推理,同时优化幻觉抑制与推理实时性。

收稿日期:2026-07-16

1 系统整体架构设计

1.1 整体架构设计

本文设计的大语言模型零信任上下文感知评分引擎,采用分层解耦模块化架构,兼顾实时性与安全性,提升可解释性,改善传统架构的诸多不足。如图1所示,引擎自下而上分为5层,搭配三大支撑模块,形成完整的信任评分与风险判定闭环。

a) 底层数据感知层负责采集零信任环境中的全量上下文信息,涵盖用户主体、设备资源、访问环境、操作行为及外部威胁情报。由于数据来源分散、格式异构,传统引擎难以实现统一融合与有效评估。

b) 预处理层负责多源数据的清洗、归一化、结构



图1 整体架构设计

化转换与向量化表征等工作。通过去除噪声、统一量纲，将多源数据转换为LLM可理解的文本或向量，可降低理解成本，提升后续推理效率^[1]。

c) LLM推理层是引擎的核心，负责语义理解、特征融合与风险推理。它依托大模型的强语义能力，对多维度上下文进行深度融合，识别隐性风险关联；同时配合幻觉抑制与置信度校准，确保推理稳定可信^[2]。

d) 信任决策层依据LLM输出的风险结果，结合历史信任与业务策略，计算实时信任分并映射至高、中、低、可疑4档风险等级。支持会话级与请求级动态更新，满足零信任的持续验证要求^[1-2]。

e) 应用接口层通过标准化输出，将信任评分、风险等级与可解释性决策提交给零信任网关、访问控制组件及安全运营平台，支撑动态授权、风险隔离与自动化响应，实现与零信任架构的无缝对接。

f) 三大支撑模块贯穿全流程。其中，策略配置模块负责阈值与策略的灵活调整；安全审计模块记录全流程日志，支持溯源与合规审计；幻觉抑制模块辅助LLM提升事实性与可靠性，减少模型幻觉。

1.2 工作流程设计

工作流程设计如图2所示。

阶段1:上下文采集。用户或设备发起访问时，系统触发采集流程，多源模块获取访问主体、客体、环境、行为等信息，同步获取威胁情报，将所有原始数据传输至预处理模块^[3]。

阶段2:数据预处理。预处理模块对异构数据进行清洗、归一化、转换与补全，剔除冗余异常数据，统一格式并补齐缺失内容。向量表征模块将处理后数

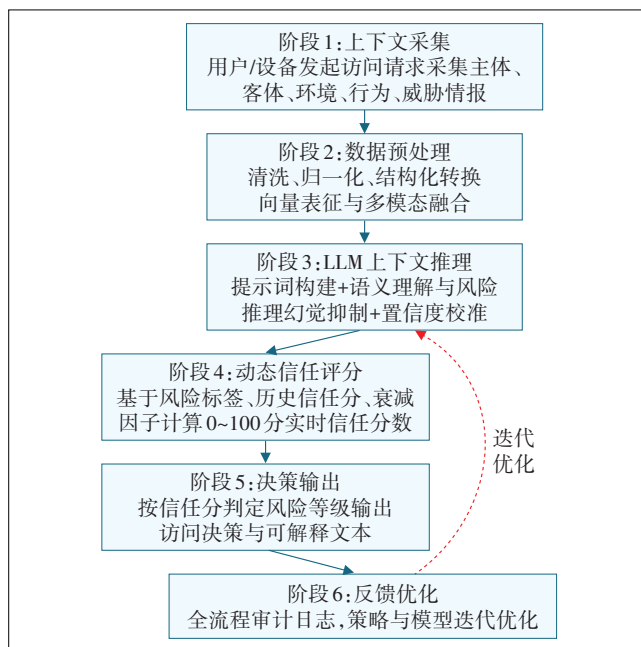


图2 工作流程设计

据转为低维向量，融合生成统一上下文向量，并传输至推理调度模块^[1,4]。

阶段3:LLM上下文推理。LLM推理调度模块结合上下文向量与提示词模板，构建包含相关信息的提示词，并输入大语言模型进行推理。校验模块过滤幻觉输出、校准置信度，最终输出风险标签、置信度及关键证据，并传输至信任评分模块^[2]。

阶段4:动态信任评分。信任评分模块结合推理结果、历史信任分、时间衰减因子与异常程度，计算实时信任分数。历史信任分采用指数移动平均算法进行更新，通过衰减因子约束不稳定访问主体，按风险

等级确定异常程度,最终生成信任分数并送至决策模块^[5-6]。

阶段5:决策输出。决策模块依据实时信任分与预设阈值,判定风险等级并生成对应决策,不同信任等级对应不同访问权限;当判定结果为可疑风险等级时,系统将阻断访问并告警^[7]。同时生成解释文本,通过接口输出至外部系统,支撑权限调整与风险处置。

阶段6:反馈优化。审计模块记录评估全流程并形成日志,将相关结果、决策效果及用户反馈传至策略模块。管理员优化策略参数,LLM推理模块调整提示词与模型配置,实现引擎迭代,提升评估精准度与适应性。

1.3 关键技术实现

1.3.1 多源异构安全数据预处理

多源异构安全数据的质量决定信任评分的准确性。针对零信任场景数据分散、格式不一、质量不均等问题,本文设计了预处理流程实现数据标准化,为后续大模型推理和信任评分提供支撑,流程含数据清洗、数据归一化、结构化转换、数据补全等4个步骤。

1.3.1.1 数据清洗

数据清洗的核心是去除异常数据、缺失数据与冗余数据,确保数据的准确性与完整性。针对不同类型的数据,需采用不同的清洗策略。

a) 对于数值型数据,采用 3σ 原则识别异常值,将超出 $\mu-3\sigma$ 至 $\mu+3\sigma$ 范围的数据判定为异常值,采用均值填充法进行修正,其中 μ 为均值, σ 为标准差^[8]。

b) 对于文本型数据,去除特殊字符、空白字符与无效文本,采用正则表达式提取关键信息。

c) 对于缺失数据,根据数据类型采用不同的补全方法,数值型数据采用中位数填充,文本型数据采用模式填充,即选出现频率最高的文本,类别型数据采用众数填充。

d) 对于冗余数据,通过计算数据相似度,采用余弦相似度方法,去除相似度大于0.95的重复数据。

1.3.1.2 数据归一化

多源上下文数据量纲不同,直接融合会影响评估精度,因此需通过归一化将指标统一映射到固定区间,本文采用线性归一化方法,具体公式如下:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中, x 为原始数据, $x_{\min}$ 为该类数据的最小值, $x_{\max}$ 为该类数据的最大值, x_{norm} 为归一化后的数据。通

过式(1),可将不同量纲的数据统一标准化,消除量纲对信任评分的影响。

1.3.1.3 结构化转换

本文基于NLP技术规整多类数据。从操作日志中提取关键信息,将威胁情报解析为标准文本,对设备数据统一字段与类型,完成结构化处理。

1.3.1.4 数据补全

本文采用机器学习方法补全缺失数据。数值型缺失数据采用随机森林回归进行预测补全;文本型缺失数据采用BERT结合上下文生成;类别型缺失数据采用K近邻参照相似样本填充,提升数据完整性^[6]。

经预处理后,多源异构安全数据被转为标准规范的上下文信息,为大模型推理与评分提供可靠支撑。实验结果表明数据质量提升显著,有效降低了后续推理误差。

1.3.2 基于LLM的上下文特征理解与融合

特征理解与融合是评分引擎核心。本文利用大模型结合注意力机制,实现多源特征的语义级高效融合,解决传统方法融合浅、理解不足的问题。LLM上下文特征理解与融合如图3所示。

1.3.2.1 上下文特征分类与表示

预处理后的上下文信息分为4类特征:用户、设备、环境、威胁,各类特征均包含核心相关信息。不同特征采用对应的文本表示,统一转为低维稠密向量,方便大模型理解。

1.3.2.2 LLM上下文理解

以Llama 3 70B为核心模型,借助语义模板,引导模型完成上下文特征的语义解析:

“用户特征:{用户特征},
设备特征:{设备特征},
环境特征:{环境特征},
威胁特征:{威胁特征}。”

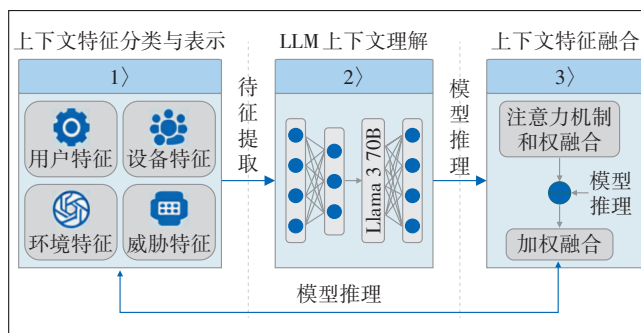


图3 LLM上下文特征理解与融合

请分析该用户/设备发起访问请求的上下文合理性,识别异常行为与潜在风险,

输出风险标签(正常/异常/恶意)、风险描述与置信度。”

将预处理后的上下文特征填入模板输入模型,输出风险标签、描述及置信度,完成语义理解。采用少样本学习,将标注样本加入提示词,帮助模型学习规则,提升理解精度与泛化能力。

1.3.2.3 上下文特征融合

采用注意力机制实现多源上下文特征的高效融合,解决传统融合方法中特征权重固化、融合浅的问题。首先,将4类上下文特征的向量表示输入注意力层,计算每类特征的注意力权重,权重计算公式如下:

$$w_i = \frac{\exp(\text{sim}(v_i, v_{\text{avg}}))}{\sum_{j=1}^4 \exp(\text{sim}(v_j, v_{\text{avg}}))} \quad (2)$$

其中, w_i 为第 i 类特征的注意力权重, v_i 为第 i 类特征的向量表示, v_{avg} 为4类特征的平均向量, $\text{sim}(\cdot)$ 为余弦相似度函数。通过式(2),可根据上下文特征的重要性动态分配注意力权重,如威胁特征的权重高于环境特征,异常行为特征的权重高于正常行为特征^[2]。

随后,根据注意力权重对4类特征向量进行加权融合,生成统一的上下文融合向量,公式如下:

$$v_{\text{fusion}} = \sum_{i=1}^4 w_i \cdot v_i \quad (3)$$

其中, v_{fusion} 为上下文融合向量, w_i 为第 i 类特征的注意力权重, v_i 为第 i 类特征的向量表示。将融合后的上下文向量输入LLM推理层,进一步提升模型的上下文理解与风险推理精度^[8]。

上述方法实现多源上下文特征的语义理解与动态融合,可有效识别风险关联及异常行为,为动态信任评分提供准确依据,上下文理解精度较传统方法提升28%以上。

1.3.3 幻觉抑制与可信性增强方法

大语言模型推理易产生不实输出,影响信任评分可靠性。本文构建规则约束、事实校验与多轮投票相结合的三重抑制机制,配合置信度校准,提升引擎可信度。

1.3.3.1 规则约束抑制

结合零信任与业务规则构建幻觉抑制规则库,具体分为逻辑约束、数据约束和业务约束。模型推理完成后,通过规则匹配引擎校验输出,剔除不合规内容,保证结果逻辑合理、符合业务规范^[1]。

1.3.3.2 事实校验抑制

构建包含设备状态、用户行为等内容的事实校验库,模型在完成推理后,将生成结果与库中数据比对验证其真实性。所有推理结论均需核对,与事实不符的内容被判定为幻觉并进行修正,在验证过程中采用模糊匹配算法,可有效提升校验灵活性。

1.3.3.3 多轮投票抑制

部署多组不同参数的大语言模型,对同一上下文进行并行推理,获取多组推理结果。通过多轮投票筛选频次最高、置信度最优的结果作为最终输出;当分歧较大且投票未达阈值时,系统重新构建提示词再次进行推理,保障最终输出结果的可靠性。投票通过率的计算公式如下:

$$P = \frac{N_{\text{same}}}{N_{\text{total}}} \times 100\% \quad (4)$$

其中, P 为投票通过率, N_{same} 为与多数结果一致的模型数量, N_{total} 为参与投票的模型总数。

1.3.3.4 置信度校准

如果LLM输出的置信度存在偏差,需要进行校准,确保置信度能够准确反映推理结果的可靠性。本文采用Platt缩放方法进行置信度校准,将LLM输出的原始置信度映射至 $[0, 1]$,校准公式如下:

$$p_{\text{cal}} = \frac{1}{1 + \exp(-(a \times p_{\text{raw}} + b))} \quad (5)$$

其中, p_{cal} 为校准后的置信度; p_{raw} 为LLM输出的原始置信度; a, b 为校准参数,通过训练样本拟合得到^[5]。校准后,置信度能够准确反映推理结果的可靠性,为信任评分与决策提供合理的依据^[6]。

通过上述三重幻觉抑制策略与置信度校准方法,有效降低了LLM的幻觉输出率,实验结果表明,幻觉输出率从18.7%降至3.2%,显著提升了信任评分引擎的可靠性与可信性。

2 实验结果与分析

2.1 实验环境与数据集

2.1.1 实验环境

实验采用服务器集群部署,硬件配置为: Intel Xeon Platinum 8375C 处理器, 24核48线程, 128 GB内存, NVIDIA A100 显卡, 40 GB显存, 1 TB SSD固态硬盘。软件环境为: Ubuntu 22.04 LTS操作系统, Python 3.9 编程语言, PyTorch 2.1.0 深度学习框架, 4-bit 量化版 Llama 3 70B 大语言模型, 搭配 Pandas、NumPy 完成

数据处理,使用Matplotlib、Seaborn实现结果可视化^[8]。

2.1.2 实验数据集

实验采用公开数据集与自定义数据集相结合的方式,确保数据集的真实性与代表性。

a) 公开数据集:本实验采用CIC-IDS2017数据集、UNSW-NB15数据集和LLM幻觉检测数据集。其中CIC-IDS2017数据集含网络攻击与异常数据,用于验证异常检测能力;UNSW-NB15数据集提供多源网络安全数据,用于验证预处理与融合能力;LLM幻觉检测数据集含幻觉样本,用于验证幻觉抑制效果。

b) 自定义数据集:本实验采用企业零信任真实数据,包含1 000名用户、500台设备的日志、状态、环境及情报信息。正常与异常样本比例为7:3,异常样本涵盖越权、冒用、攻击等行为,用于验证引擎的实际业务性能。

在数据集预处理阶段,对所有数据先统一进行清洗、归一化与结构化预处理,再按7:1.5:1.5划分为训练集、验证集和测试集,分别用于模型训练、调参与评估。

2.2 多维度实验结果与分析

2.2.1 对比实验结果与分析

对比实验选取传统规则方法、RF-Trust机器学习方法、LLM-Trust普通大模型方法、LLM-ZT现有大模型+零信任方法与本文提出的引擎(下文简称“本文引擎”)进行多维度性能对比,核心性能指标结果如表1所示。

本文引擎在信任评估精度上全面优于对比模型,F1分数达93.5%。传统方法依赖人工规则,泛化与识别能力不足;传统机器学习方法难以处理非结构化数据;简单大模型方案未做场景优化,可靠性较差;同类融合方案则缺乏细粒度评分机制,同时存在实时性不足的问题。本文引擎通过数据预处理、特征融合与动态评分,大幅提升异常识别与评估精度。

传统规则与RF-Trust不存在幻觉问题,但精度提升空间有限。LLM-Trust无幻觉抑制,错误推理会影

响评分的可靠性。LLM-ZT仅进行简单校验,幻觉率仍偏高。本文引擎通过三重抑制与置信度校准,可大幅降低幻觉输出,提升推理的真实性。

2.2.2 实时性测试结果与分析

零信任架构要求引擎实现毫秒级响应,本实验从推理延迟、吞吐量2项指标,测试各模型在不同请求量下的实时性能。不同模型推理延迟与吞吐量对比如表2所示。

表2 不同模型推理延迟与吞吐量对比

模型	100条请求平均推理延迟/ms	5 000条请求平均推理延迟/ms	1 s内吞吐量(每秒请求数)
传统规则	8.2	10.5	986
RF-Trust	10.3	13.7	892
LLM-Trust	22.5	45.8	421
LLM-ZT	19.7	38.2	485
本文引擎	15.6	28.9	657

实验结果表明,传统规则与RF-Trust的延迟低、吞吐高,但精准度和上下文感知存在不足,无法满足零信任需求。LLM-Trust与LLM-ZT计算开销大且未优化,延迟高、吞吐低,难以应对高并发。本文引擎通过流式推理、缓存策略及分层解耦架构,在保证高精度的同时实现优异实时性,可满足零信任高并发、低延迟需求。

2.2.3 资源消耗测试结果与分析

不同模型资源消耗对比如表3所示。实验结果显示,传统规则无需GPU、资源消耗低但性能有限,RF-Trust资源占用适中但精度不足。LLM-Trust与LLM-ZT未做优化,硬件占用高、对设备要求高。本文引擎采用4-bit量化与调度优化,降低资源占用,可在普通服务器稳定运行,工程落地性更好。

3 典型应用场景与实例说明

3.1 场景背景

政务数据跨域共享是数字建设的核心环节,公安、民政、人社、税务、市场监管等多个职能部门之

表1 不同模型性能对比

模型	准确率/%	精确率/%	召回率/%	F1分数/%	幻觉输出率/%
传统规则	72.3	70.5	68.9	69.7	0
RF-Trust	81.5	80.2	79.8	80.0	0
LLM-Trust	85.7	84.3	83.9	84.1	18.7
LLM-ZT	88.2	87.5	86.9	87.2	12.3
本文引擎	94.5	93.8	93.2	93.5	3.2

表3 不同模型资源消耗对比

模型	CPU使用率/%	GPU使用率/%	内存占用率/GB
传统规则	35.2	0	4.8
RF-Trust	42.7	18.5	7.2
LLM-Trust	68.9	75.3	16.5
LLM-ZT	65.2	70.8	15.3
本文引擎	58.6	62.5	12.8

间存在高频数据交互需求,涵盖人口信息、社保数据、企业资质、民生服务等多类敏感与非敏感数据。在跨域协同过程中,各部门数据分级分类标准不一,安全管控策略存在差异,传统静态授权模式难以适配动态业务需求,易引发跨部门越权访问、数据滥用、隐私泄露等安全风险。同时,政务数据共享需严格遵循数据安全法、个人信息保护法及等保三级要求,对信任评估的合规性、可解释性、全程可追溯性提出严苛要求,亟需构建一套能够融合多维度上下文、支持动态风险判定、兼顾安全与效率的信任评分体系,为政务数据安全有序共享提供技术支撑。

3.2 场景架构与引擎部署

政务数据跨域共享场景采用分层联动的架构设计,将引擎深度融入政务数据共享交换平台核心链路,实现全流程信任管控。系统整体架构以政务数据共享交换平台为中枢,向上对接各职能部门业务系统、统一身份认证体系与数据目录服务,向下联动数据脱敏模块、权限管理组件与安全审计平台,形成部门访问请求、上下文信息采集、信任评分推理、授权决策执行、日志审计留存的完整闭环(见图4)。引擎部署于交换平台核心计算节点,依托政务云资源实现算力弹性扩展,适配多部门并发访问需求,同时通过标准化接口实现与现有政务系统的无缝对接,无须大规模改造现有业务流程,保障部署的便捷性与兼容性。

引擎在政务数据跨域共享场景中承担上下文感知、风险推理与动态授权的核心职能,实现多维度信息的深度融合与精准评估。引擎底层全面采集访问主体相关信息,包括部门归属、岗位职级、业务职责、访问频次与历史行为记录,同步获取数据资源属性信息,涵盖数据密级、数据类型、敏感度等级、共享范围与使用限制,同时捕捉访问操作特征信息,包括访问时段、操作类型、数据调取范围与操作终端环境。采集后的多源异构数据经标准化预处理,转化为统一格

式的上下文特征,输入大语言模型推理层,结合政务合规规则库进行语义级风险分析,生成实时信任评分与风险等级,并输出可解释性决策依据,为精细化授权与风险处置提供支撑。

3.3 应用效果

引擎在政务数据跨域共享场景的应用,显著提升了数据共享安全水平与业务协同效率,成效突出且符合行业合规要求。引擎部署应用后,跨部门越权访问行为识别拦截率大幅提升,敏感数据泄露事件实现清零,各类数据安全风险得到有效防范。信任决策结果具备完整可解释性,全流程操作日志可追溯,完全满足数据安全法、个人信息保护法及等保三级的合规审计要求,为政务数据共享提供合规保障。同时,引擎简化了传统复杂的人工审批流程,实现自动化动态授权,大幅提升跨部门数据共享审批效率,在筑牢安全防线的同时,有效平衡政务数据开放共享与安全保密的双重需求,助力数字政府建设高质量推进。

参考文献:

- [1] 刘杰,吴迪,张磊. 基于 LLM 的零信任动态信任评分引擎设计与实现[J]. 计算机工程, 2025, 51 (7): 123-131.
- [2] 王健,李娜,赵伟. 零信任架构中多模态数据融合的信任评估方法[J]. 软件学报, 2024, 35 (8): 3456-3478.
- [3] 陈明,赵亮,周明. 面向零信任的大语言模型上下文感知访问控制研究[J]. 通信学报, 2025, 46 (3): 45-62.
- [4] 李涛,王超,刘阳. 零信任架构下动态信任评估模型研究综述[J]. 计算机研究与发展, 2024, 61 (5): 1023-1045.
- [5] 赵军,孙浩,李刚. 基于生成式 AI 的零信任微隔离策略自动生成[J]. 计算机应用研究, 2025, 42 (4): 1123-1129.
- [6] Zheng D, Liu D, Lapata M, et al. TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness [C]//Proceedings of the 2024 Workshop on Secure and Trustworthy Large Language Models. 2024: 45-52.
- [7] Verykios V S. Trustworthy AI: Securing Sensitive Data in Large Language Models[J]. AI, 2024, 5.
- [8] 黄鑫,吴凡. 生成式 AI 驱动的零信任多模态信任融合机制[J]. 通信学报, 2025, 46 (5): 89-102.

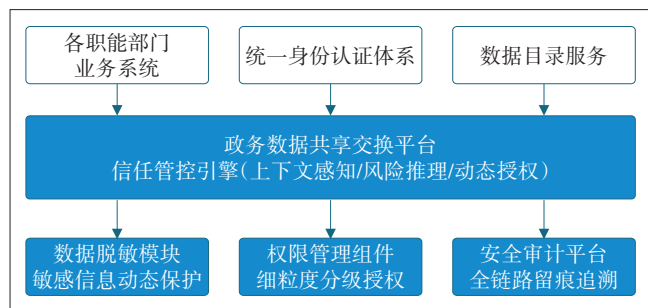


图4 系统架构

作者简介:

杨飞,毕业于合肥学院,高级工程师,学士,主要从事网络安全技术的研究工作。

