智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

分层多智能体框架下的黑灰产聊天记录分析方法

韦科顺, 石磊, 石拓, 田一帆, 程刚

引用本文:
韦科顺, 石磊, 石拓, 等. 分层多智能体框架下的黑灰产聊天记录分析方法[J]. 智能系统学报, 2026, 21(5): 1221-1236.
WEI Keshun, SHI Lei, SHI Tuo, et al. Analysis method of black-gray industry chat records under hierarchical multi-agent framework[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1221-1236.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202603037

DOI: 10.11992/tis.202603037
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260630.1507.002

分层多智能体框架下的黑灰产聊天记录分析方法

韦科顺[1], 石磊[2], 石拓[3], 田一帆[4], 程刚[5]
（1. 对外经济贸易大学 法学院, 北京 100029; 2. 中国传媒大学 媒体融合与传播国家重点实验室, 北京 100024; 3. 北京警察学院 公安管理系, 北京 102202; 4. 中国人民公安大学 信息网络安全学院, 北京 100038; 5. 应急管理大学 计算机与信息安全学院, 北京 101601）

摘　要: 网络黑灰产聊天记录作为地下产业链交互的直接载体, 蕴含了大量潜在犯罪意图与组织脉络线索。如何从海量非结构化文本中精准识别黑话术语, 深度挖掘犯罪类型与行为模式, 成为突破传统研判瓶颈、实现黑灰产情报智能化感知的关键课题。为此, 本文提出一种基于大模型的多智能体协同分析方法。方法采用工作流式编排与任务分工机制, 将黑话识别、类别判定、模式挖掘、报告生成四大核心环节按预定流程串联协同, 同时引入检索生成增强 (retrieval-augmented generation, RAG)、思维链 (chain-of-thought, CoT)、批处理、高层次模式提取等技术, 突破了传统单一模型在处理长链路复杂逻辑时的能力瓶颈, 实现从原始数据到智能分析报告的端到端转化, 提升了黑灰产聊天记录分析研判的智能化水平。实验结果显示, 该方法在网络黑灰产聊天记录数据集上具有良好表现, 检索生成增强、思维链等相关技术的引入, 有效提升了各智能体在子任务中的表现, 具备良好的实际应用价值。

关键词: 网络黑灰产; 大模型; 多智能体; 聊天记录分析; 检索增强生成; 思维链; 文本挖掘; 犯罪情报分析
中图分类号: TP391.1　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)05−1221−16

中文引用格式: 韦科顺，石磊，石拓，等．分层多智能体框架下的黑灰产聊天记录分析方法 [J]．智能系统学报，2026，21(5): 1221–1236.
英文引用格式: WEI Keshun, SHI Lei, SHI Tuo, et al. Analysis method of black-gray industry chat records under hierarchical multi-agent framework[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1221–1236.

Analysis method of black-gray industry chat records under hierarchical multi-agent framework

WEI Keshun[1], SHI Lei[2], SHI Tuo[3], TIAN Yifan[4], CHENG Gang[5]
(1. School of Law, University of International Business and Economics, Beijing 100029, China; 2. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; 3. Department of Public Security Management, Beijing Police College, Beijing 102202, China; 4. College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China; 5. School of Computer Science and Information Security, University of Emergency Management, Beijing 101601, China)

Abstract: As direct carriers of underground industry chain interactions, cyber black market chat records embed substantial implicit criminal intentions and organizational clues. The accurate identification of jargon from vast amounts of unstructured text along with the deep mining of criminal types and behavioral patterns has emerged as a critical challenge in overcoming traditional analytical bottlenecks and enabling intelligent insights into cyber black market intelligence. This study proposes a large-model-based multi-agent collaborative analysis method. The method employs a workflow-oriented orchestration and task division framework that sequentially coordinates four core processes in a predefined workflow: jargon identification, category determination, pattern mining, and report generation. It also integrates technologies such as retrieval-augmented generation (RAG), chain-of-thought (CoT), batch processing, and high-level pattern extraction. This approach overcomes the limitations of traditional single models in handling long-chain, complex logical tasks, achieving an end-to-end transformation from raw data to intelligent analysis reports and enhancing the analytical capabilities for cyber black market chat records. Experimental results demonstrate that the proposed method achieves strong performance on datasets containing cyber black market chat records. The integration of technologies such as RAG and CoT improves the performance of individual agents in their respective subtasks, indicating promising real-world applicability.

Keywords: cyber black market; large model; multi-agent; chat record analysis; retrieval-augmented generation; chain-of-thought; text mining; criminal intelligence analysis

收稿日期: 2026−03−26.　**网络出版日期:** 2026−06−30.
基金项目: 国家自然科学基金项目 (62406023, 42377200).
通信作者: 石磊. E-mail: leiky_shi@cuc.edu.cn.

网络黑灰产活动是指具有一定独立性、常业性的非法牟利型网络违法犯罪[1]。其主要分为黑

产与灰产 2 部分：前者主要包括利用互联网技术实施的攻击、窃取、诈骗等严重违法犯罪活动，如黑客勒索、木马植入、电信网络诈骗等；后者则指处于违法犯罪边缘、为黑产提供上游技术支持、下游资金结算等辅助服务的行为，如个人信息买卖、恶意注册账号、引流推广、虚假广告等，二者相互交织，共生互促，形成了一条分工明确且链路复杂的地下产业链。因此，网络黑灰产已成为威胁网络空间安全和社会公共秩序的突出问题[2-4]。

为预防和打击网络黑灰产活动，公安机关和有关监管部门持续保持高压态势，然而受限于该类违法活动隐蔽性强、犯罪链条纵横交织、技术管理分工明确等特点，传统人工研判的治理方式存在范围判别有限、时间滞后性强和关联交叉识别弱等不足。随着数智技术的发展，新技术在强化监管手段的同时，也被犯罪组织用于规避监管、遮盖犯罪行为。因此，如何在数字经济蓬勃发展的时代背景下，研究探索网络黑灰产的智能化治理路径，已成为提升治理效能、维护网络空间安全的核心议题[5-7]。

网络黑灰产聊天记录作为地下产业链交互的直接载体，成为研判与治理网络黑灰产活动的关键依据。该类信息记录通常具有非结构化、信噪比低等特征，表现出极强的冗余性与稀疏性。同时，为逃避监管，嫌疑人会在交流中使用大量隐晦的黑话暗语。这种复杂的文本特征与强对抗的语言伪装相互叠加，给关键信息的提取与研判带来了极大挑战。因此，构建集黑话识别、类别判定、模式挖掘与报告生成于一体的文本挖掘框架，对开展网络黑灰产活动精准研判具有重要意义。其中，黑话识别有助于消除对抗性文本的语义歧义，模式挖掘能够从海量数据中重构非法交易的行为链路与组织架构，类别判定可实现对犯罪类型的系统归类与多层次总结，报告生成则可完成碎片化数据到结构化情报信息的转化，上述过程有效提升了公安机关对网络黑灰产活动态势的感知能力。

为此，本文提出一种分层多智能体协同的黑灰产聊天记录分析方法。该方法通过工作流式编排与任务分工机制，将黑话识别、类别判定、模式挖掘、报告生成四大核心环节按预定流程进行串联协同，并引入检索增强生成、思维链、批处理、高层次模式提取等技术。与现有面向单任务的深度学习模型、传统同构多智能体架构以及通用 LLM(large language model) 提示工程方法相比，本方法针对黑灰产分析各环节的异质性需求构建了专职异构智能体协同框架，通过动态知识增强的 RAG(retrieval-augmented generation) 与 CoT(chain-of-thought) 融合推理机制，实现可解释的证据支撑与专家级逻辑推理的融合，同时设计了面向实战的增量迭代分析与标签层级归并方法，有效解决了长链路复杂逻辑推理中的能力瓶颈与工程化落地难题。

1　相关研究

网络黑灰产聊天记录分析是一项高度复杂的文本挖掘任务，其难点不仅在于文本的非结构化与低信噪比，更在于黑话术语的强对抗性与犯罪模式的动态演化。当前研究主要沿 3 条技术路径展开：1) 面向单一分析任务的文本语义理解模型；2) 面向复杂任务分解的多智能体协同框架；3) 引入外部知识增强的大语言模型推理技术。本章将围绕上述 3 条路径梳理研究现状，并在此基础上凝练现有方法在面对全流程、端到端黑灰产分析任务时存在的局限性。

1.1　面向单任务的黑灰产文本分析模型

针对黑灰产聊天记录的自动化分析，学界早期主要采用基于规则匹配与统计机器学习的文本分类方法。王媛媛等[8] 围绕聊天记录这一特殊文本类型，系统梳理了其不同于普通文本的特点，并对语义分析中涉及的关键技术进行了研究。基于关键词库与正则表达式的匹配方法虽实现简单，但在面对谐音、缩写、隐喻等对抗性黑话变体时漏报与误报率极高。以支持向量机、随机森林为代表的机器学习方法通过特征工程提升了分类性能，但仍受限于浅层特征的表征能力与领域适应性。在上述研究中，黑灰产从业者频繁使用近义词替代、同音字变换、符号嵌入等对抗手段，使得传统方法逐渐失效。已有研究指出，业界正逐步从“关键词匹配”转向“语义理解”，通过融合无监督新词发现、上下文建模与对抗训练来应对语言变体。近年来，以 TextCNN(text convolutional neural network)、BiLSTM(bidirectional long short-term memory)、BERT(bidirectional encoder representations from transformers) 为代表的深度学习模型在文本分类、序列标注等任务中取得突破，为黑灰产文本的语义解析提供了更强大的技术基座。Kim 等[9] 提出的 TextCNN 通过多尺度卷积核捕获文本局部特征，有效提升了短文本分类的准确性；Huang 等[10] 设计的 BiLSTM-CRF(bidirectional long short-term memory conditional random field) 架构在序列标注任务中展现出对长距离依赖关系的

建模优势; Devlin 等[11] 提出的 BERT 模型通过双向 Transformer 预训练机制, 显著增强了文本的深层语义表征能力; Sun 等[12] 提出了引入知识图谱实体信息的 ERNIE 模型, 进一步强化了语义理解的知识关联能力; 针对黑灰产特定场景, 朱恩德等[13] 融合 BiLSTM 与 CNN 对推特平台的黑灰产推文进行分类, 验证了混合神经网络在对抗性文本识别中的有效性。然而, 上述模型均为面向单一任务的孤岛式设计, 其核心局限在于: 1) 模型架构与标注数据高度绑定, 难以在同一框架内兼容黑话识别、类别判定、模式挖掘等多环节的异质任务; 2) 模型推理过程缺乏可解释性, 输出结果难以直接转化为具备证据属性的研判信息; 3) 在处理长文本、多轮对话及语义动态演化场景时, 静态架构的模型难以实现跨会话的上下文推理与知识迁移。上述局限表明, 单纯依赖单一模型的性能提升, 难以支撑从原始语料到结构化情报的全流程转化需求。从数据安全法益的视角看, 张婷[14] 指出, 信息互联网向价值互联网转型催生了数据安全风险的变异与升级, 传统的“静态数据–动态信息”双核法益保护模式已难以有效涵盖数据的全部属性。这一论断进一步印证了本文所提出分层多智能体框架在应对动态演化、语义复杂的数据环境方面的必要性与前瞻性。

1.2　面向任务协同的多智能体分析方法

为应对复杂任务中的多阶段、多维度协同需求, 多智能体系统通过任务分解、角色分工与动态交互机制, 为长链路分析任务提供了新的解决范式。在网络安全领域, 传统多智能体技术的应用高度集中于欺诈检测的工具链层面: 如启明星辰的安星系统通过四大智能体协同实现攻击全生命周期的主动防御; AAMAS 2025 会议提出的分层多智能体强化学习框架通过主–子策略的层级协作, 显著提升了复杂网络攻击恢复决策的效率; Aldaihan 等[15] 构建了“首席审查员–调查员”分层架构的 Clouseau 系统, 通过智能体间的迭代调查与交叉验证, 实现了安全日志的自动化取证。在公共安全领域, Fard 等[16] 基于多智能体博弈模型, 实现了犯罪团伙的跨平台关联挖掘; 胡今鸣等[17] 则将强化学习与贝叶斯网络相结合, 构建智能体推演超高层建筑非法入侵路径, 为安防系统动态优化提供了决策支持。针对黑灰产共谋行为, MultiAgent4Collusion 框架通过模拟百万级智能体协同作恶, 揭示了信息附和、证据伪造等异常协同模式的内在机理。此外, 罗双春等[18] 提出基于目标识别与主题引导对话的黑灰产威胁情报挖掘方法, 通过构建对话智能体实现与黑灰产人员的主动交互, 体现了智能体在交互式情报采集中的潜力。上述研究充分证明, 多智能体系统在处理动态对抗、多目标优化的复杂任务中具有独特优势。近年来, 研究者进一步将多智能体协同拓展至更广泛的公共安全与复杂系统优化场景。芮兰兰等[19] 针对高动态车联网环境中的服务迁移问题, 提出基于多智能体深度强化学习的优化方法, 有效应对多用户资源竞争与节点可用性动态变化等挑战。然而, 当前研究多集中于同构智能体的平行协作或单一主–从层级架构, 而在黑灰产聊天记录分析这一特定场景中, 各分析环节对智能体的能力要求存在显著异质性——黑话识别依赖细粒度的语义消歧, 类别判定需要宏观的罪名体系映射, 模式挖掘则要求跨会话的时序逻辑重构。如何针对上述异质任务构建具备专项能力且可协同编排的异构分层智能体工作流, 仍是当前研究亟待填补的方法空白。值得注意的是, 多智能体协同方法在其他复杂决策领域已展现出良好的应用前景。黄艳等[20] 将决策思维链与多智能体协同技术引入防洪调度决策, 构建了衔接“预报–预警–预演–预案”业务逻辑的防洪大模型, 显著增强了复杂任务拆解与精准解答能力。这一思路为本文构建分层异构多智能体框架提供了有益的跨领域借鉴。此外, 马晓飞等[21] 提出的通用人工智能驱动的多智能体仿真方法 (artificial general intelligence+multi-agent system, AGI+MAS), 通过融合大语言模型、跨模态数据融合与多智能体深度强化学习, 推动建模机制从“规则依赖”向“自学习驱动”转变, 进一步印证了多智能体系统在复杂情境下的自适应学习与动态决策能力, 对本文方法具有重要启示意义。

1.3　大语言模型与检索增强生成技术

以 ChatGPT、DeepSeek 等为代表的大语言模型凭借其强大的语义理解、上下文建模与文本生成能力, 为安全情报分析开辟了新的技术路径。在引入大语言模型早期, 研究者大多以单一大模型为核心, 依赖提示词实现文本分类、意图识别等任务[22]。李衡峰等[23] 提出基于大模型的诈骗文本分类方法, 通过句向量聚类和自动提示词生成优化技术, 在少量样本下实现 81.6% 召回率和 84.9% 精确率。研究者通过微调领域模型、引入外部知识库等方式, 持续提升 LLM 对黑灰产对抗性文本的解析性能。多智能体技术与大语言模型的深度融合, 为复杂任务提供了新的解决路径。Zhu 等[24] 通过融合 LLM-MAS 模拟了人类社会的

协作分工，有效克服单智能体模型的性能瓶颈；窦路遥等[25]基于多智能体技术构建高价值专利的价值自动解释模型，通过微调 DeepSeek-V3 大模型，实现了从单一指标到多维融合的范式跨越；张益伟等[26]融合 LLM 语义生成与图注意力网络，实现了黑灰产文本与图结构特征的协同感知；Alhuzali 等[27]提出的 RAGIntel 系统采用混合检索算法，通过重排序与上下文压缩显著降低了 LLM 在威胁情报分析中的幻觉发生率；Blefari 等[28]进一步开发了基于中心 LLM 智能体的 CyberRAG 框架，通过迭代检索−推理循环实现对领域知识库的动态查询，在攻击分类任务中达到 94.92% 的准确率。检索增强生成技术通过引入黑话词典、犯罪模式库等外部结构化知识，为大模型推理提供显式证据支撑，有效提升了输出结果的可靠性与可解释性。石拓等[29]融合大语言模型思维链提示与语义聚类方法，精准解析电信网络诈骗引流话术的主题模式与语义策略，为诈骗被害预警提供了有力决策依据；尚思佳等[30]则从图挖掘角度出发，构建黑灰产网络资产图谱，通过子图挖掘与关键链路识别，揭示黑灰产团伙的运作模式。然而，现有 RAG 应用多聚焦于单一模态的静态知识检索，缺乏对动态演进知识的持续吸纳机制。在黑灰产语境下，黑话术语日新月异、犯罪模式迭代频繁，若无法将前期分析产出的新知识实时沉淀并反哺后续推理任务，检索增强的效果将随时间衰减。针对这一问题，李明锋等[31]提出的多智能体无监督事件骨架生成框架 MAU-GES, (multi-agent unsupervised model for generating event schema)，通过设计基于多智能体的图抽取算法和语义意义段拆分算法，实现了从原始文本到事件骨架图的端到端生成，在低资源场景下展现出优异的无监督学习能力。这一思路为本文构建动态知识增强机制提供了有益借鉴，即通过智能体协作实现对动态演进知识的持续吸纳与反哺。

1.4　现有方法局限性与本文创新定位

综上所述，当前黑灰产聊天记录分析方法可归纳为 3 条技术路径：面向单任务的深度学习模型、传统同构多智能体系统，以及基于 LLM 的通用提示工程方法。然而，面对黑灰产分析“黑话识别−类别判定−模式挖掘−报告生成”这一长链路、异构需求、知识动态演进的全流程任务，现有方法存在的共性局限详见表 1。

针对上述局限，本文提出一种基于分层多智能体的协同分析方法，通过黑话识别智能体输出结果动态构建 RAG 知识库，并利用思维链提示增强类别判定与模式挖掘智能体的推理深度，旨在实现黑灰产聊天记录从原始数据到研判情报的全链条自动化转化。本质创新与增量贡献体现为：

1) 架构创新：不同于传统同构多智能体，本文首次面向黑灰产全链条设计黑话识别、类别判定、模式挖掘、报告生成 4 大专职异构智能体，并通过工作流式编排形成分层协作机制 (黑话层−研判层−报告层)，实现了端到端转化。

2) 推理机制创新：突破现有方法中 RAG 与 CoT 独立使用的局限，创新性地将黑话识别智能体的输出动态构建为 RAG 知识库，并采用基于字符 N-gram 的精准检索策略，同时将 CoT 嵌入类别与模式智能体，形成“可解释证据支撑−专家级逻辑推理”的融合链路。消融实验 (3.3.3 节) 证明二者具有显著的互促提升效应。

3) 工程适配创新：针对数据动态接入实战需求，设计了“冷启动−增量迭代”两阶段框架与基于 SBERT 的标签层级归并技术，解决了大模型输出标签不一致与类别体系收敛的工程化难题，提升了法的实战可用性。

表 1　现有方法存在的共性局限

Table 1　Common limitations of existing methods

方法类别	代表性工作	核心局限
单任务深度学习	BiLSTM-CNN、BERT	孤岛式设计，无法兼容多环节异质任务；缺乏可解释性与跨会话推理能力。
传统多智能体	Clouseau、MultiAgent4Collusion	多为同构平行协作或主从架构，缺少面向黑灰产特定业务流的专职异构设计；知识传递依赖硬编码，缺乏动态增强机制。
通用LLM提示工程	单模型+CoT、单模型+RAG	单一模型难以同时胜任细粒度语义消歧与宏观模式抽象；CoT与RAG各自独立使用，未形成针对黑灰产任务的深度融合与证据闭环。
通用LLM多智能体	AutoGen、CAMEL	智能体角色设定通用，缺乏面向黑灰产黑话、犯罪模式的专业能力定制；工作流编排与领域知识库动态构建机制缺失。

1.5　研究方法

本文选取 DeepSeek-V3-671B 作为智能体构建的基座大模型，融合检索生成增强、思维链、聚类算法等技术，围绕黑话识别、类别判定、模式挖掘 3 方面内容进行分析，并将分析内容形成报告；同时针对上述环节，构建面向黑灰产聊天记录分析的智能体工作流，通过链式编排，实现任务分解、逐步推理与结果汇总，为公安机关针对网络黑灰产活动的线索发现、风险研判、打击治理、证据固定等工作提供了有力智能化支撑。

2　多智能体工作流构建

智能体作为近年来人工智能领域涌现出来的关键技术范式, 凭借其强大的任务协同与上下文编排能力, 为高复杂度、长链路的工作流构建提供了新的实践路径。黑灰产聊天记录的文本结构复杂、表述隐晦, 且黑话术语以碎片化的形式夹杂其中, 传统的分析方法往往无法兼顾文本底层语义解析的深度与高层次逻辑推理的全局性。本文紧密围绕公安机关打击黑灰产活动中所面临的痛点难点问题, 如黑话术语对抗性强、线索信息碎片化程度高、犯罪模式迭代升级快等, 构建了一个分层多智能体协同的分析研判框架, 精准对齐情报研判的核心链路, 分别为黑话识别、类别判定、模式挖掘、报告生成 4 个关键环节设立专职智能体, 并利用黑话识别智能体的分析结果动态构建检索生成增强知识库, 保证语义还原与分析研判工作的认知协同。同时, 为提高专职智能体的逻辑稳定性与研判精度, 本文在类别判定、模式挖掘智能体中分别构建嵌套式智能体模块, 进一步强化了智能体的自动流转与自主规划的能力。具体流程如图 1 所示。

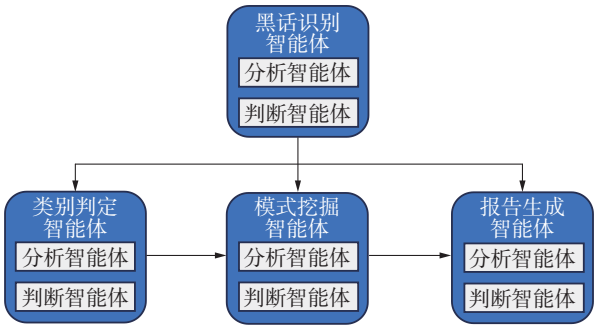

图 1　分层多智能体流程

Fig. 1　Flowchart of hierarchical multi-agent system

分层多智能体结构相比单模型端到端处理, 能够降低复杂任务的推理负载, 减少黑话语义误解在后续分类和模式分析中的累积影响, 同时保留各环节中间结果, 增强分析过程的可解释性和可追溯性。

2.1　智能体主要类别

2.1.1　黑话识别智能体

黑话是黑灰产活动中, 违法犯罪嫌疑人为逃避监管而使用的具有较强隐蔽性与对抗性的词语。面向公安实践, 黑话的强更新性、旧词新义等特点, 致使传统基于关键词库检索的方式在一定程度上存在滞后, 在缺乏上下文关联的前提下易出现漏报、误报问题; 针对新型黑话的识别往往依赖于专家经验, 存在较强的主观性且缺乏广泛适用性, 难以在全局研判中形成标准化的情报分析体系。黑话识别智能体则通过识别聊天记录中看似日常但逻辑异化的表述, 推理出隐语背后的真实指向并构建黑话知识库, 为后续分析任务提供数据支撑。该智能体的输出包含“词条、推断含义、原文片段、置信度”4 个字段。

2.1.2　类别判定智能体

面向黑灰产聊天记录, 对风险信息进行类别判定, 有助于公安机关对目标嫌疑人或犯罪团伙的精准定位。黑灰产活动犯罪手段复合, 涉及罪名交织, 传统的深度学习模型多受限于对标注数据的依赖以及黑盒化的特征提取机制, 难以在对抗性语境下捕捉深层次犯罪意图, 且在处理新型黑灰产犯罪时的泛化能力不足。类别判定智能体通过整体把握语料的异常特征, 要求模型在对既有犯罪类型精准归类的同时, 主动挖掘新型犯罪类型, 实现“已知类型精准识别、未知类型及时识别”的双重能力闭环。该智能体的输出包含“类型名称、出现频次”两个字段。

2.1.3　模式挖掘智能体

黑灰产犯罪具有线索高度碎片化且行为模式迭代迅速的特征, 传统犯罪模式挖掘方法多受限于硬编码的规则匹配或频繁项集挖掘, 对隐藏在时序对话中的潜在逻辑和协作范式的挖掘能力不足。模式挖掘智能体专注于从海量非结构化对话中解构从事黑灰产活动违法犯罪嫌疑人及犯罪团伙的标准作业流程和组织演化规律, 探索犯罪模式的共性特征, 为公安机关针对特定犯罪模式的阻断与打击提供关键情报支撑。该智能体的输出包含“低层次犯罪模式、高层次犯罪模式”两部分内容。

2.1.4　报告生成智能体

报告生成智能体负责汇集上述 3 个智能体的分析产出, 基于生成式语言模型技术自动梳理信息脉络, 将结构化信息转化为具备逻辑性、可读性和证据属性的研判报告, 辅助指挥决策层快速掌握黑产动向并制定针对性打击方案, 解决了人工撰写报告耗时费力、撰写标准不统一、证据链条梳理困难等问题。报告生成智能体的输出包括“报告标题、黑话识别、类别判定、模式挖掘、研判详情、处理建议”几方面内容。

2.2　智能体协作机制

在黑灰产聊天记录分析任务中, 4 类智能体并非简单的并列关系, 而是围绕公安业务流程, 形成具有明确分工和递进逻辑的层级协同关系。

从协作链路上看, 黑话识别智能体位于语义

解析层，负责对聊天记录中的隐晦表达、替代表述等黑话术语进行语义还原，为后续类别判定与模式挖掘提供基础语义支撑；类别判定与模式挖掘智能体位于分析研判层，在黑话语义还原基础上完成犯罪类型映射，前者负责对碎片化线索的分类整合，后者负责对黑灰产活动的行为链路与协作方式开展深度挖掘，其分析结果构成研判报告的核心内容；报告生成智能体位于结果输出层，负责将上述 3 个智能体的分析产出汇总形成具备较强可读性的研判报告，为公安机关的指挥决策和定向打击提供依据。

从任务分工的合理性上看，黑灰产聊天记录分析涉及多种异质任务，单一智能体难以同时兼顾细粒度识别、宏观分类、复杂推理和规范表达。黑话识别任务强调对短词化、符号化、隐喻化表达的上下文消歧，需要较强的局部语义推断能力；类别判定任务强调将分散线索映射至相对稳定的犯罪类型体系，需要较强的归纳概括与类别约束能力；模式挖掘任务强调跨文本、跨线索的流程重构和关系抽象，需要较强的链式推理与结构化分析能力；报告生成任务则强调信息整合、逻辑组织和面向实战的表达规范，需要较强的文本生成与证据组织能力。上述任务在输入形式、输出结构、推理深度和评价目标上均存在明显差异，因此采用专职智能体分别承担不同环节，能够降低单一模型的任务负载，减少不同目标之间的相互干扰，提高各环节输出结果的稳定性和可解释性。

2.3　智能体工作模式

本文在类别判定、模式挖掘环节的智能体均构建基于同构智能体工作流，即以统一的智能体类别与交互范式为基础，通过差异任务提示与规则约束，完成不同维度的分析任务。黑话识别智能体则略有不同，其输出结果将被构建成为外部知识库，为后续分析任务提供可复用的知识支撑。此处需要说明的是，为节约篇幅，本章节仅选择“类别判定”环节作为样例呈现，具体流程如图 2 所示。

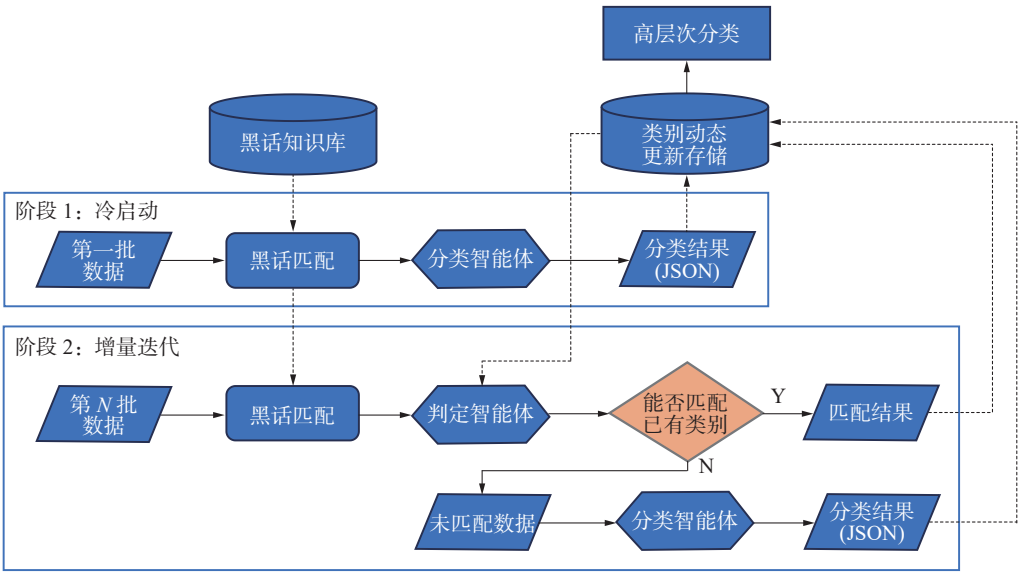

图 2　类别判定流程

Fig. 2　Flowchart of category determination

该流程面向黑灰产聊天数据的批量接入与轻量化更新需求，构建了冷启动−增量迭代两阶段分析框架。冷启动阶段以第一批数据为输入，通过外部黑话知识库获得证据辅助，再由分类智能体完成初始类别判定，并将 JSON 格式的输出结果写入类别动态更新存储；增量迭代阶段首先通过同样方式接收后续批次数据并进行黑话匹配，之后通过判定智能体匹配已有类别并将未匹配数据传递至分类智能体，将新类别提取结果与匹配结果同时写入类别动态更新存储。增量迭代阶段反复执行直至全部数据处理完成。最后，使用基于字符 N-gram 的标签聚类算法对细粒度类别标签进行相似性聚合与归并，形成“高层次类别−细化类别”的层级分类体系，为公安机关开展基于黑灰产聊天记录的针对性治理与精准打击提供有力依据。

2.3.1　基于字符 N-gram 的检索生成增强

检索生成增强是指在大模型生成过程中引入外部知识库，通过检索获得与当前输入相关的证据片段，将其与原始提示共同组合成为提示输入，提高大模型的推理能力并为其输出结果提供

可解释的证据支撑。本文将黑话识别环节的输出结果构建形成黑话知识库, 通过基于字符 N-gram 的词法相似度匹配方法, 检索当前批次数据中与知识库词条相匹配的关键词, 并将其对应的完整数据条目整合形成证据片段用于提示增强, 使模型能够在显式证据约束下完成语义消歧与含义推断, 有效提升了对黑话隐语的理解一致性与可靠性。

本文通过黑话识别智能体动态构建黑话知识库, 每一条黑话记录由“词条、推断含义、原文片段、置信度”4 个字段构成, 同样通过冷启动阶段直接识别、增量迭代阶段先判定匹配后识别的方式, 实现黑话数据持续接入与动态更新。

本文使用的基于字符 N-gram 的检索生成增强与传统方法略有不同, 传统方法通常依赖向量表示与向量数据库的语义近邻检索, 具有较强的语义泛化能力, 但需要额外的 Embedding 模型、向量索引构建与存储维护成本, 且针对领域黑话、缩写变体、错别字等“形式变化强但语义稳定”的短词条场景, 可能出现向量表示不稳定或召回漂移的问题。相比之下, 基于字符 N-gram 的检索方式以词形重叠为主要依据, 更适用于字符串短、存在大量变体与拼写扰动的场景, 具有实现简单、计算开销低等优势, 为方法向实战应用落地的转化提供了较好的实践路径。

考虑到黑灰产黑话普遍具有短词化、符号化和扰动化特征, 本文设置 N-gram 中的 N 值为{2, 3, 4}, 同时提取二元、三元和四元字符片段, 以兼顾短词黑话的召回能力和较长隐晦表达的区分能力; 设置最低相似度阈值 θ 为 0.05, 仅保留相似度不低于阈值的候选词条; 同时, 为控制冗余证据干扰, 设置 Top-k 值为 5, 即仅保留得分最高的 5 条知识记录。

字符 N-gram 检索的具体流程为:

1) 对黑话词条和待分析文本进行规范化处理, 减少大小写、符号、空格等噪声差异。

2) 采用连续字符窗口对文本进行 N-gram 切分, 得到不同粒度的字符片段集合。

3) 基于 Jaccard 相似度计算待分析文本与知识库词条之间的字符片段重合程度, 并按照相似度得分进行排序。

4) 结合相似度阈值与 Top-k 策略筛选相关黑话记录, 将其组织为外部证据片段, 输入后续智能体。

字符 N-gram 检索的公式为

$$\mathrm{sim}(x,y)=\frac{|G_{\mathrm{N}}(x)\cap G_{\mathrm{N}}(y)|}{|G_{\mathrm{N}}(x)\cup G_{\mathrm{N}}(y)|} \tag{1}$$

式中: $G_{\mathrm{N}}(x)$表示待分析文本x的字符 N-gram 集合, $G_{\mathrm{N}}(y)$表示知识库词条y的字符 N-gram 集合。相似度得分越高, 表示二者在字符片段层面的重合程度越高。

2.3.2 基于动态窗口的数据批量处理方法

由于大模型本身存在输入长度限制, 无法一次性处理全量数据, 而传统的单样本逐条处理方法会显著增加 API 调用次数, 带来不必要的计算开销与延迟。基于此, 本文提出了一种基于动态窗口的数据批量处理方法, 该方法根据大模型可支持最大输入长度, 将原始数据划分成为若干个批次, 并在每个批次内并行处理多条聊天记录, 以提高整体推理效率。具体而言, 单批次可处理的数据条数并非固定值, 而是由模型最大上下文长度、提示词模板长度、检索增强证据长度、输出预留长度以及当前样本平均长度共同决定。当样本文本较短时, 单批次可容纳的数据条数相对较多; 当样本文本较长或检索证据较多时, 系统则自动缩小批次规模, 以避免超过模型上下文长度限制。伪代码如算法 1 所示。

算法 1　基于动态窗口的数据批量处理算法

输入　D, $L_{\max}$, P, R, L_{out}

输出　Batch set Batches

1) Initialize Batches = ∅ and B = ∅。

2) Compute the available window length:

$L_{\mathrm{sample}} = L_{\max} - \mathrm{length}(P) - \mathrm{length}(R) - L_{\mathrm{out}}$。

3) For each sample x_i in D:

4) If total_length(B) + length(x_i) ⩽ L_{sample}, then:

5) Add x_i to the current batch B。

6) Otherwise:

7) Add the current batch B to Batches。

8) Reset the current batch as $B = \{x_i\}$。

9) End if。

10) End for。

11) If $B \neq \varnothing$, add B to Batches。

12) Return Batches。

同时, 考虑到生成式语言模型输出一致性受限的特点, 在处理不同批次的同一类别数据时, 模型输出的分类标签可能会存在近义但略有偏差的表述, 这种偏差会造成类别标签冗余, 影响分类结果的精准度。针对上述问题, 本文引入了专门用于语义映射判定的智能体。该智能体通过接收当前已完成批次数据的分类标签, 判定当前批次数据能否完成与已有标签的语义映射, 将能够完成映射的数据合并至目标标签, 同时将不能完成映射的部分传递至分类智能体, 进行新类别标

签的重新提取，提升了分类结果的一致性和可用性。

2.3.3　思维链提示策略

思维链是一种通过引导大模型生成中间推理步骤来提升复杂任务表现的提示方法[32]。其核心思想在于，在上下文学习中，通过提示词构建从问题到结论输出的逻辑推理链路，激发模型的逻辑推理潜能。在黑灰产聊天记录分析领域，由于原始文本中包含大量隐晦代词、行业黑话以及对抗性表达，直接进行类别提取往往面临精准度低、可解释性差等问题。通过思维链提示显式构建从原始文本到语义还原、逻辑拆解、再到标准判定的逻辑链路，有助于模型在处理非结构化、高模糊性的黑灰产文本时，具备类似于专家级的研判思维过程，取得更佳的分类效果。此处以分类智能体中使用的思维链提示作为展示，如图 3 所示。

分类智能体

角色设置：你是一名网络黑灰产情报分析专家。

任务描述：你的任务是严格按照以下思考流程，识别当前批次文本所涉及的黑灰产犯罪类型。

思维链：
第一步：黑话去混淆
1.识别关键词：扫描文本，锁定非标准用语、隐晦代词或行业黑话。
2.语义映射：结合黑话补充知识，将识别到的黑话翻译为通用业务术语。
3.语境还原：在替换黑话后，重新通读句子，还原对话的真实含义。
第二步：意图与要素结构
1.提取核心对象：分析交易或互动的标的物是什么?如个人隐私数据、银行卡等。
2.判定核心动作：分析针对该标的物执行了什么操作?如买卖、技术支持等。
第三步：确定角色立场：
1.判断发言者是需求方还是供给方，这决定了类别的方向，如求购数据、出售数据等。
第四步：标准化标签生成
1.合成初始标签：采用“核心动作+核心对象”的格式组合标签。
2.术语规范化：将口语化的组合转化为公安或行业通用的治理术语。
3.颗粒度校验：检查标签是否足够具体，若不够具体，如仅为“诈骗”或“灰产”等，则需回溯第二步提取更细的特征。

输入文本:{texts_block)

输出格式:请返回 JSON 格式如下：
{
 "classification":[
{
  "index":文本编号,
  "crime_types":犯罪类型}]
}

图 3　基于思维链的类别判定提示词示例

Fig. 3　Example of chain-of-thought prompt for category determination

2.3.4　基于 SBERT[33] 表征与层次聚类的标签层级归并

细粒度类别标签种类繁多，虽然有助于刻画更为细致的线索信息，但直接应用于公安实战会增加研判和管理成本，也不利于对风险的宏观识别和整体把握。对细分标签的高层次归纳有助于相关部门掌握黑灰产活动的特征与趋向，识别潜在的风险，为前置预警和治理工作提供数据支撑。针对上述问题，本文在完成全量数据的类别提取后，引入基于 SBERT 表征与层次聚类的标签层级归并方法。SBERT 是一种基于 BERT 的改进网络架构，能够将文本转换为固定大小的向量表示，为语义相似度计算、文本聚类、信息检索等任务提供了高质量的嵌入向量起点。同时，相较于传统的交叉编码器，SBERT 采用的解耦编码机制显著降低了模型推理复杂度，提升了计算效率。层次聚类算法是一种无监督学习方法，通过计算余弦相似度衡量嵌入向量之间的语义距离，并用平均联动准则合并语义距离最小的两个簇，并在不同切分层级下计算平均轮廓系数，选择较优聚类粒度。赵颖等[34] 提出的黑灰产网络资产图谱构建方法同样采用图结构与聚类技术，实现了核心资产与关键链路的有效识别。

本文将大模型直接生成的黑灰产类别标签集合$V_{\text{key}}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$进行向量嵌入，通过 M3E 模型将标签映射至高维度语义空间，即$\boldsymbol{v}_t=[e_1\ \ e_2\ \ \cdots\ \ e_d]$，其中$d$为嵌入向量特征维度。

在向量嵌入的基础上，使用层次聚类算法进行文本标签的归并。首先，任意选取两个嵌入向量标签v_i与v_j计算余弦相似度，计算公式为

$$\text{sim}\left(\boldsymbol{v}_i,\boldsymbol{v}_j\right)=\frac{\boldsymbol{v}_i\cdot\boldsymbol{v}_j}{|\boldsymbol{v}_i|\,|\boldsymbol{v}_j|} \tag{2}$$

式中：$\boldsymbol{v}_i\cdot\boldsymbol{v}_j$表示向量的点积，$|\boldsymbol{v}_i|$和$|\boldsymbol{v}_j|$分别表示向量的模长。

之后，采用自底向上的策略，通过平均联动准则合并语义距离最小的簇，计算公式为

$$D(\boldsymbol{A},\boldsymbol{B})=\frac{1}{|\boldsymbol{A}|\cdot|\boldsymbol{B}|}\sum_{\boldsymbol{v}_i\in A}\sum_{\boldsymbol{v}_j\in B}\left(1-\text{sim}\left(\boldsymbol{v}_i,\boldsymbol{v}_j\right)\right) \tag{3}$$

式中：$D(\boldsymbol{A},\boldsymbol{B})$表示簇$\boldsymbol{A}$与簇$\boldsymbol{B}$之间的关联距离，$|\boldsymbol{A}|$和$|\boldsymbol{B}|$为簇内样本数。

最终生成的类别数目k由所有样本的平均轮廓系数$S(k)$确定，以确保分类体系的收敛性，计算公式为

$$S(k)=\frac{1}{n}\sum_{i=1}^{n}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{4}$$

式中：$a(i)$为样本i到同簇其他点的平均距离，$b(i)$

为样本 i 到最近其他簇的平均距离。$S(k)$ 越接近 1 表示聚类效果越理想。

最后, 通过计算关键词向量与簇质心的相似度, 筛选各类簇内的核心主题词, 计算公式为

$$\mathrm{sim}(\boldsymbol{v}_t, \boldsymbol{C}_d) = \frac{\boldsymbol{v}_t \cdot \boldsymbol{C}_d}{|\boldsymbol{v}_t||\boldsymbol{C}_d|} \tag{5}$$

式中: $\boldsymbol{C}_d = \frac{1}{N_d}\sum_{\boldsymbol{v}_t \in d} \boldsymbol{v}_t$ 表示簇 d 内所有标签向量的质心。至此完成面向黑灰产聊天记录的高层次类别归并, 形成“高层次类别−细化类别”的层级分类结果。基于 SBERT 表征与层次聚类的标签层级归并算法伪代码如算法 2 所示。

算法 2　基于 SBERT 表征与层次聚类的标签层级归并算法

输入　T, E, Sim, K

输出　Hierarchical label merging result H

1) Initialize the label vector set $V = \varnothing$。

2) For each label t_i in T:

3) Encode t_i into vector v_i using embedding model E。

4) Add v_i to V。

5) Compute the pairwise cosine similarity matrix S based on V。

6) Perform hierarchical clustering with average linkage。

7) For each candidate cluster number k in K:

8) Compute the average silhouette coefficient。

9) Select the optimal cluster number k* with the best silhouette coefficient。

10) Obtain high-level label clusters C according to k*。

11) Compute cluster centroids and select representative topic words。

12) Construct the hierarchical label structure H。

13) Return H。

本文选择层次聚类作为标签层级归并方法的原因在于, 该方法能够更好适应黑灰产细粒度标签数量多、表达多样、目标类别数量难以确定等任务特征。与 K-means 和密度聚类方法对比来看, 一方面, 黑灰产高层次类别会随着黑话表达、犯罪形态、作案模式等变化而发生动态演化, K-means 等方法需要预设聚类数量, 易因类别数设定不合理造成标签过度合并或拆分, 且 K-means 方法更适合近似球形、规模相对均衡的簇结构, 难以适应黑灰产标签中长尾分布明显、类别规模差异较大的情况; 另一方面, 密度聚类方法对邻域半径、最小样本数等参数较敏感, 当部分类别标注数据有限或语义距离分布不均衡时, 参数轻微变化可能导致聚类结果波动较大, 不利于形成稳定、可解释的标签层级体系。而层次聚类方法能够无需预设类别数量, 通过自底向上的方式逐步合并语义相近的标签, 并形成较为直观的树状结构。

本文选择平均联动准则作为簇间合并依据的原因在于, 相较于单联动准则, 平均联动能够避免因局部相似关系导致的链式合并, 相较于全联动准则, 平均联动能够降低边界样本对聚类结果的过度影响, 从而在链式合并与过度拆分之间取得平衡, 更适合语义相近但表达多样的黑灰产标签归并场景。

从方法收敛性上看, 自底向上的层次聚类方法通过在每轮迭代中合并语义距离最近的两个簇, 使簇数量单调减少, 并最终根据平均轮廓系数选择较优聚类粒度, 能够在有限步内完成标签归并; 从方法稳定性上看, SBERT 语义表征降低了标签字面差异对相似性判断的影响, 余弦相似度削弱了标签长度和词频规模差异的干扰, 平均联动准则通过综合考虑簇间全部样本的平均距离, 减少单个边界样本或局部相似样本造成的聚类偏差。因此, 该方法能够在标签表达存在噪声、类别规模不均衡和长尾分布明显的情况下, 获得相对稳定且具有解释性的标签层级归并结果。

3　实证研究

3.1　数据预处理

本文所使用的实验数据来源为 Telegram 平台的公开黑灰产交流频道, 通过关键词过滤的方式, 从 315 个活跃社群中回溯抓取了近 36 个月的历史消息, 并对所获取数据进行了重复项剔除、噪声过滤以及语料清洗标准化预处理, 同时考虑到隐私安全, 所有实验数据均已在预处理阶段完成匿名化操作, 最终获得可用样本 15 623 条。

3.2　黑话识别分析

3.2.1　黑话分布统计分析

基于对 15 623 条网络黑灰产聊天记录原始数据的挖掘, 本文共识别并提取具有明确指向的黑灰产黑话实体 297 例 (去重后), 且在 14 676 条数据中均出现了黑话表达, 占原始数据条数的 93.93%, 说明黑灰产聊天记录中黑话的使用并非偶发现象, 呈现普遍、稳定的特点。本文对识别出的黑话基于其语义以及所服务的黑灰产业链条进行高层次分类, 共划分为“诈骗与引流”“资金链/洗钱

支付”“博彩赌博”“账号与身份资源”“担保结算与交易”“风控规避与执法风险”“色情相关”“组织角色与渠道”“数据隐私与黑客服务”“其他”10 类，黑话种类分布统计如图 4 所示。

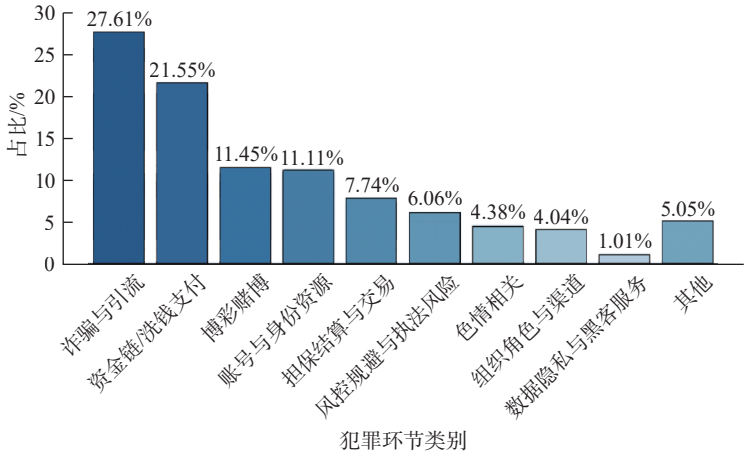

图 4　网络黑灰产黑话种类分布

Fig. 4　Distribution of slang types in cyber black and gray industries

从比例分布上看，本文语料中黑话主要集中于“诈骗与引流”和“资金链/洗钱支付”两大环节，二者合计占比高达 49.16%，说明在黑灰产语言体系中，涉及诈骗及其手段以及资金流转等语言表述往往容易触发平台审核或形成证据链的关键节点，最需要被隐匿表达，例如：“精聊”指代诈骗转化环节的深度话术沟通；“引流”指代将目标导入赌博/诈骗平台的获客动作；“U”指代 USDT 结算单位，表明加密资产结算已成为黑灰产常见支付方式的一部分。

除诈骗、资金结算外，博彩赌博活动同样是黑灰产的重要场景之一，而账号、身份信息等资源则构成黑灰产活动的运作基础，特别是“账号与身份资源”类黑话往往与实名制体系、风控机制高度相关，其隐匿表达不仅有助于规避平台对账号交易、身份冒用等行为的敏感检测，也能降低沟通中直接暴露违法要素的风险。

在低频类别中，“色情相关”“组织角色与渠道”和“数据隐私与黑客服务”类别占比相对较低，表明黑话语义场景更偏向诈骗与资金链条，更多集中于“交易型、运营型”黑灰产业态。

黑灰产黑话呈现上述分布的主要原因在于，诈骗与引流环节处于黑灰产链条前端，直接面向潜在受害者或流量资源，相关沟通具有高频次、强扩散和跨平台流动等特点，因此更需要能够规避平台审核和降低识别风险的替代表达；资金链与洗钱支付环节则处于黑灰产链条后端，承担非法收益转移、结算、变现等功能，是执法监管和平台风控重点关注的对象，因此相关主体通常会采用缩写、符号、隐喻等方式隐藏支付工具、结算方式和资金通道。二者分别对应黑灰产活动中的入口和出口，也是整个违法链条中最依赖沟通、最容易留下痕迹的环节，因此在黑话分布中占据主要比例。其他低频类别则更多受业务场景、社群类型和样本来源影响，呈现出较强的长尾特征，反映出黑灰产语义体系中既有高频共性环节，也存在大量分散化、场景化的细分表达。

3.2.2　黑话形式特征分析

本文通过对所识别黑话实体的进一步分析发现，其词条呈现出明显的短词化、符号化和编码化的特征。短词因其具备输入成本低、传播速度快、可组合性强等特点，在黑话列表中占据了极高的比例；而符号、编码、字母−符号的混合表达方式则展现出了更强的对抗性与隐蔽性，能够在一定程度上干扰传统基于关键词检测方法的检测效果。基于此，本文在同一语料库分别采用基于关键词匹配方法与本文基于思维链提示的方法开展了对比实验，结果详见表 2。

表 2　不同方法黑话识别效果对比

Table 2　Comparison of jargon recognition performance by different methods

方法种类	出现黑话表达的样本数/条	黑话命中频次
基于关键词匹配方法	8 842	34 967
基于思维链提示方法	14 676	130 981

实验结果表明，相较于关键词匹配方法，本文方法在覆盖率与命中频次两个维度均表现出明显优势，表明该方法具有更强的识别能力，具备潜在的实战应用价值。

此外, 本文对黑话替换规律进行分析后发现, 替换规律可大致分为“谐音编码、缩写符号、隐喻、风控承诺”4 种模式, 且这些模式往往叠加出现。在谐音替换模式中, 最典型的即为“菠菜=博彩”, 即用常用词替换违法词, 类似的还有“掴内/掴外=国内/国外”等。谐音替换的优势即为学习成本低、跨群体应用稳定性强, 是黑话体系中较为普遍的应用模式。在缩写符号模式中, 常见的黑话表示有“U=USDT”“BC=博彩”等, 这种缩写的目的不仅限于规避平台审查, 更能较好地提高沟通效率, 且易达成行业共识, 与外界形成良好隔离。在隐喻模式中, 违法分子通常将参与或退出犯罪活动表达为“上车/下车”, 将用于资金转移的行动小组表达为“车队”, 将对受害者的不法侵害表达为“杀鱼”等。这种隐喻将违法表述生活化, 难以被关键词检测方法覆盖。风控承诺模式则体现为, “上押/双押”指保证金担保制度, “叔叔不找”指低风险、不易被查, “不爆红”用颜色隐喻表达平台不出事、不被查封, 这类黑话的出现, 恰恰印证了黑灰产内部最担心的不是“能不能做”, 而是“做了以后会不会出事、会不会被坑”的心理状态。

3.3　类别判定分析

3.3.1　实验结果

类别判定智能体共提取细粒度黑灰产犯罪类别 95 类, 并在此基础上进行高层次类别提取, 最终形成“资金链与洗钱结算”“诈骗与资金盘业务”“网络赌博与博彩”“综合型跨境与黑产服务”“数据隐私与网络入侵”“伪造货币与证件票据”“组织卖淫与色情活动”“非法烟草酒类活动”8 个类别, 犯罪类型数量分布如图 5 所示。

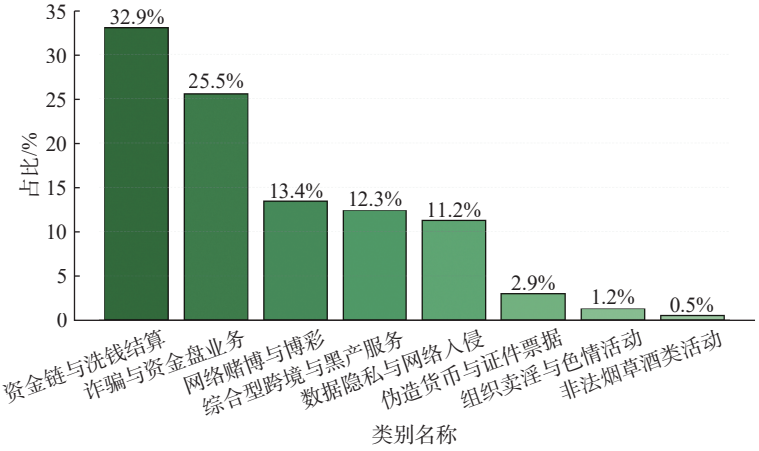

图 5　网络黑灰产犯罪类型数量分布

Fig. 5　Distribution of the number of crime types in cyber black and gray industries

从图 5 中可以看出, 黑灰产犯罪活动分布呈现头部集中、长尾分布的特征, 高频类别高度聚焦于资金流转与诈骗变现等活动。具体而言, “资金链与洗钱结算”占比位居首位, 达到 32.9%, 能够看出资金转移、支付通道、结算方式等内容, 因直接关系到非法收益的落地与风险控制, 已成为黑灰产的核心话题。这一发现与许浩[35] 对金融黑灰产新样态的分析结论相一致, 其研究指出不法贷款中介、非法代理维权、职业背债人骗局等金融黑灰产活动均围绕资金链条展开, 呈现出产业化运作的显著特征。从法律规制视角看, 张新平等[36] 进一步提出, 后互联网时代网络黑灰产治理需要法治体系健全与技术优势赋能的二元协同路径。“诈骗与资金盘业务”次之, 占比 25.5%, 说明诈骗类活动仍是黑灰产生态中主要的业务形态之一, 该类别与“资金链与洗钱结算”共同构成高频组合, 二者合计占比达 58.4%, 展现出较强的“诈骗实施–资金转移–完成变现”的闭环链路特征。

第三梯队主要包括“网络赌博与博彩”(13.4%)、“综合型跨境与黑产服务”(12.3%) 以及“数据隐私与网络入侵”(11.2%)。其中, 博彩类占比相对突出, 说明博彩业务活跃度强, 并与诈骗链条可能存在交叉渗透; 而数据隐私与网络入侵类别保持一定比例, 表明个人信息、黑客服务等上游资源在黑灰产体系中发挥支撑作用, 为诈骗引流、账号运营等提供关键原料。

黑灰产犯罪类型呈现上述分布的原因在于, 各类黑灰产活动最终均需要通过资金转移、支付结算和洗钱变现完成收益闭环, 因此“资金链与洗钱结算”占比最高; 诈骗与资金盘业务具有复制成本低、传播速度快、变现路径短等特点, 容易与引流、账号资源和支付结算等环节结合, 因此占比较高。相比之下, 伪造货币与证件票据、组织卖淫与色情活动、非法烟草酒类活动等类别具有更强的场景专属性, 受样本来源和社群类型影响较大, 因此呈现相对低频的长尾分布。由此可见,

该犯罪类型分布反映了黑灰产生态中核心变现环节集中、外围支撑环节分散的结构特征。

3.3.2　知识图谱

本文对 8 种类型的黑灰产犯罪分别构建知识图谱，使用 DeepSeek-V3-671B 大模型对原始语料进行信息抽取形成三元组，为知识图谱构建提供结构化数据支持。本节以“资金链与洗钱结算”犯罪类型为例进行展示，如图 6 所示。

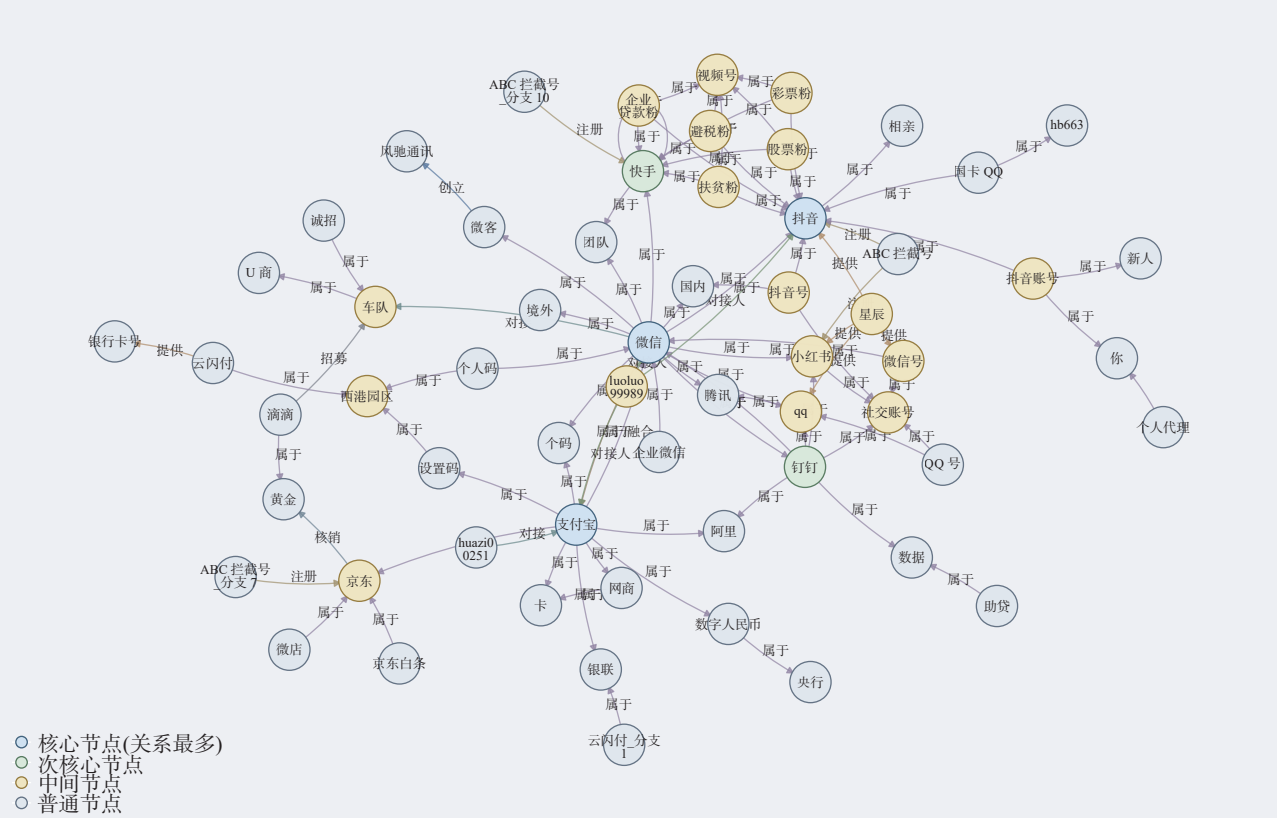

图 6　资金链与洗钱结算知识图谱

Fig. 6　Knowledge graph of capital chain and money laundering settlement

在该类犯罪的知识图谱中，本文以节点度数作为衡量实体在图结构中重要性的指标。节点度数是指某一节点与其他节点发生关联的次数，即连接边的数量。度数越高，表明该实体在语料中被提及并与其他实体产生关联的次数越多，通常对应业务链条中的“枢纽节点”或“高频基础设施”。本文对排名前十位的节点及其度数进行统计，结果详见表 3。

结合知识图谱分布与排名前十的节点度数统计结果可以看出，微信与支付宝位居前两位，二者作为直接的资金流转平台，门槛低、覆盖面广，说明该类犯罪实际运转过程中，存在面向新人、非专业人士或是更广泛人群的交易场景，选择通用性更强的沟通或交付方式能有效降低运转成本。排名 3~10 位的节点以短视频、社交和电商平台为主，说明该类犯罪存在跨平台引流获客的导向，也在一定程度上反映了该类犯罪门槛低、可复制、易扩散的特点。

表 3　本类型知识图谱中度数排名前十节点

Table 3　Top 10 nodes by degree in this type of knowledge graph

序号	名称	度数
1	微信	223
2	支付宝	187
3	抖音	102
4	微收云	91
5	三网手机卡	72
6	钉钉	66
7	快手	60
8	京东	52
9	QQ	38
10	小红书	37

3.3.3　消融实验

为证实本文方法中基于字符 N-gram 的检索生成增强和思维链提示策略的正向作用，本文基于类别判定智能体，构建 2×2 因子消融实验。从原始数据中挑选 5000 条数据构建专用数据集，邀

请黑灰产研判领域专家进行类别标注, 且为保证标注结果的可靠性与一致性, 标注过程中专家需结合上下文语境对每条样本的风险类型进行判定, 并统一映射至本文构建的 8 类高层次类别体系, 对于存在分歧的样本, 采用复核讨论机制进行一致性修订, 最终形成用于性能评估的标注数据集, 为模型性能的量化对比提供客观依据。本文采用宏平均 (Macro-average) 的 Precision、Recall 与 F_1-Score 作为评价指标。首先分别计算每个类别的各项指标, 再对所有类别取算术平均, 使得每个类别在总体指标中具有相同权重, 能够更好地反映模型在类别分布不均衡场景下对小样本类别的识别能力。

设共有C个类别, 对第i类有真阳性T_{P_i}、假阳性F_{P_i}、假阴性F_{N_i}。则该类的精确率、召回率与 F_1-Score 定义为

$$\begin{aligned} P_i &= \frac{T_{P_i}}{T_{P_i}+F_{P_i}}, \\ R_i &= \frac{T_{P_i}}{T_{P_i}+F_{N_i}}, \\ F_{1_i} &= \frac{2P_iR_i}{P_i+R_i} \end{aligned} \tag{6}$$

宏平均指标定义为

$$\begin{aligned} P_{\text{macro}} &= \frac{1}{C}\sum_{i=1}^{C}P_i, \\ R_{\text{macro}} &= \frac{1}{C}\sum_{i=1}^{C}R_i, \\ F_{1_{\text{macro}}} &= \frac{1}{C}\sum_{i=1}^{C}F_{1_i} \end{aligned} \tag{7}$$

实验结果详见表 4。

表 4　消融实验结果

Table 4　Ablation experiment results

模型	Precision	Recall	F_1-Score
RAG+CoT	0.9311	0.9290	0.9294
仅RAG	0.9278	0.9260	0.9263
仅CoT	0.9166	0.9150	0.9153
均不使用	0.8877	0.8840	0.8846

实验结果可以看出, 当仅保留 RAG 或 CoT 时, 3 项指标相比均不使用取得较明显提升, 说明 RAG 与 CoT 均具有正向作用, 当同时使用 RAG 与 CoT 时, 模型性能继续取得提升, 说明了 CoT 与 RAG 在模型推理中形成协同互补关系, RAG 为 CoT 提供了可靠的推理前提, 避免链式推理建立在错误的语义理解之上; CoT 则对 RAG 的检索结果进行解释、筛选和整合, 避免外部证据停留于简单的拼接层面。

3.3.4　对比实验

为进一步验证本文方法的有效性, 本文选取基于少样本上下文学习的大语言模型方法 (LLM-Fewshot) 和 BERT 预训练语言模型作为基线方法, 开展对比实验。实验使用与消融实验相同的数据集, 同样采用宏平均 (Macro-average) 的 Precision、Recall 与 F_1-Score 作为评价指标, 实验结果如表 5 所示。

表 5　对比实验结果

Table 5　Comparison experimental results

模型	Precision	Recall	F_1-Score
本文方法	0.9311	0.9290	0.9294
LLM-Fewshot	0.8920	0.8880	0.8889
BERT	0.9240	0.9263	0.9240

由表可见, 本文方法在 3 项指标中均取得了最优结果, 说明本文方法具有较好的整体性能; BERT 方法的结果仅次于本文方法, 说明 BERT 预训练模型已经具备较强的上下文语义表征能力, 能够较好地捕捉黑灰产聊天记录中的语义特征; LLM-Fewshot 方法结果相比其他方法有所差距, 说明仅依靠少量示例进行上下文学习, 虽然能够在一定程度上提高大语言模型的语义理解能力, 但在面对黑话、缩写以及类别边界模糊的黑灰产文本时, 仍易出现示例覆盖不足、提示词适配性不足等问题, 导致分类结果不佳。

3.4　模式挖掘分析

本文基于思维链提示, 要求大模型围绕成员招募模式、洗钱模式、作案流程、组织架构 4 个维度, 面向黑灰产聊天记录开展犯罪模式挖掘, 共分析提炼出细分的犯罪模式 277 种, 并在此基础上总结形成了 43 个高层次模式, 具体如图 7 所示。

图 7 中, 不同犯罪维度中犯罪模式的丰富程度呈现差异。其中, 洗钱模式的高层次模式 (14 个) 和细分模式 (108 个) 数量均处于首位, 说明资金流转环节不仅流转链路长, 且手段多样, 在规避监管和藏匿资金流向方面存在较强的对抗性, 是黑灰产流转链条中较为复杂的关键节点。成员招募则表现为细分模式 (73 个) 较多而高层次模式 (6 个) 较少, 说明该环节的核心范式相对稳定, 如固定的话术框架或渠道等, 但在具体实践中仍具有高度的可变化性, 如引流方式、身份包装的频繁更替等。作案流程作为黑灰产的实施环节, 其高层次模式 (12 个) 和细分模式 (52 个) 的数量均处于中间区域, 说明其既存在相对清晰的流程框架, 又会因业务类型、违法环境、工具条

件等因素产生流程分支。组织架构的细分模式(44 个) 数量最低，进而说明黑灰产业层级关系相对明确，角色配置较为稳定，结构关系与职责分工具有较强的可复用性。

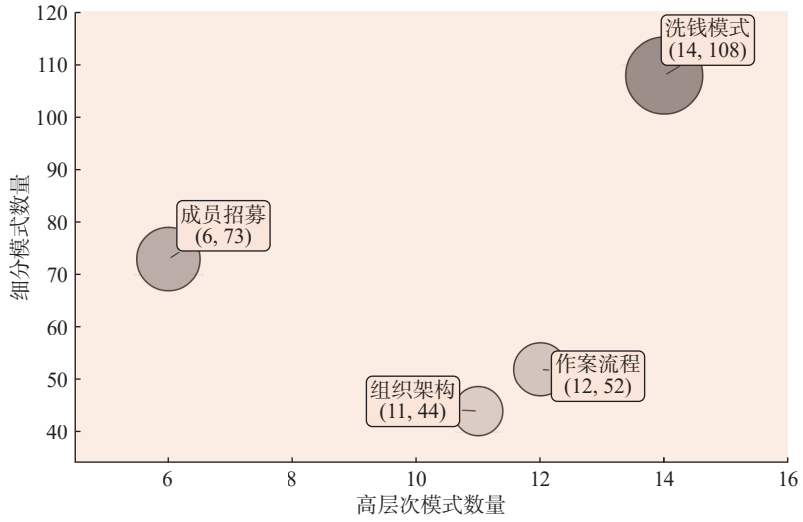

图 7　犯罪模式细分与高层次数量统计

Fig. 7　Statistics of crime mode subdivision and high-level quantities

黑灰产犯罪模式呈现上述分布的原因在于，洗钱模式直接关系到非法收益能否安全转移和最终变现，且需要不断应对平台风控、账户冻结、资金追踪等风险，因此最容易衍生出多样化的操作路径和规避手段；成员招募虽然底层目标较为固定，但会根据平台环境、目标群体和业务类型不断调整话术、渠道和身份包装方式，因此细分模式较多而高层次模式较少；作案流程受具体犯罪类型影响较大，不同业务会形成不同实施路径，因此呈现中等规模的模式数量；组织架构则更多体现为分工关系和层级配置，整体相对稳定，可复用性较强，因此模式数量相对较少。由此可见，犯罪模式数量差异反映了不同环节在黑灰产链条中的动态程度和对抗复杂度差异。

4　结束语

本文提出了一种分层多智能体协同的黑灰产聊天记录分析方法，通过检索增强生成与思维链提示协同，有效提升了黑话识别、类别判定、模式挖掘、报告生成 4 个环节的工作质效。

1) 提出了面向黑灰产全链分析的分层异构多智能体协同框架，设计了黑话识别、类别判定、模式挖掘、报告生成 4 大专职智能体，通过工作流式编排与分层协作机制实现了从原始数据到结构化情报的端到端转化；

2) 构建了基于动态知识增强的 RAG 与 CoT 融合推理机制，将黑话识别智能体的输出动态构建为外部知识库，引入基于字符 N-gram 的检索增强生成方法，同时在类别判定与模式挖掘智能体中嵌入思维链提示策略，显著提升了模型在对抗性语境下的语义消歧与深度推理能力；

3) 设计了面向实战的增量迭代分析与标签层级归并方法，构建了“冷启动−增量迭代”两阶段分析框架，并引入基于 SBERT 表征与层次聚类的标签层级归并技术，在保证线索精细度的同时降低了宏观研判的管理成本。

上述成果为公安机关和有关监管部门打击防范网络黑灰产违法犯罪活动提供了有力的技术支持。未来，可结合跨领域数据源，如社交平台、支付平台等，形成更广泛的数据联动和情报共享机制，进一步提升网络黑灰产的识别、打击与预警能力。

参考文献：

[1] 皮勇. 网络黑灰产刑法规制实证研究[J]. 国家检察官学院学报, 2021, 29(1): 18−40.
PI Yong. An empirical study on the criminal regulation of cyber black and grey industries[J]. Journal of National Prosecutors College, 2021, 29(1): 18−40.

[2] BASHEER R, ALKHATIB B. Threats from the dark: a review over dark web investigation research for cyber threat intelligence[J]. Journal of computer networks and communications, 2021(1): 1302999.

[3] CHEN Guangxuan, LIU Qiang, CHEN Guangxiao, et al. Exploring illicit personal information trading behind telecom fraud in China[J]. Humanities and social sciences communications, 2025, 12(1): 1−11.

[4] 斯彬洲, 孙海春, 吴越. 基于大语言模型和事件融合的电信诈骗事件风险分析[J]. 数据分析与知识发现, 2025, 9(7): 38−51.

SI Binzhou, SUN Haichun, WU Yue. Risk analysis of telecom fraud events based on large language models and event fusion[J]. Data analysis and knowledge discovery, 2025, 9(7): 38−51.

[5] XU Zequan, LI Lianyun, LI Hui, et al. Self-supervised graph representation learning for black market account detection[C]//The Sixteenth ACM International Conference on Web Search and Data Mining. Singapore: ACM, 2023: 330−338.

[6] WANG Qichao, MA Huan, WEI Wentao, et al. Attention paper: how generative AI reshapes digital shadow industry?[C]//Proceedings of the ACM Turing Award Celebration Conference-China 2023. Wuhan: ACM, 2023: 143−144.

[7] GUMPHUSIRI P, TRIYASON T. Synthetic data for scam detection: leveraging LLMs to train deep learning models[C]//2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology. Bangkok: IEEE, 2024: 868−873.

[8] 王媛媛, 范潮钦, 苏玉海. 面向聊天记录的语义分析研究[J]. 信息网络安全, 2017(9): 89−92.
WANG Yuanyuan, FAN Chaoqin, SU Yuhai. Research on semantic analysis for chat records[J]. Netinfo security, 2017(9): 89−92.

[9] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACM, 2014: 1746−1751.

[10] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015−08−09)[2025−01−01]. https://arxiv.org/abs/1508.01991.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: ACL, 2019: 4171−4186.

[12] SUN Yu, WANG Shuohuan, LI Yukun, et al. ERNIE: enhanced representation through knowledge integration [EB/OL]. (2019−04−19)[2025−01−01]. https://arxiv.org/abs/1904.09223.

[13] 朱恩德, 王威, 高见. 融合 BiLSTM 与 CNN 的推特黑灰产分类模型[J]. 计算机工程与应用, 2025, 61(1): 186−195.
ZHU Ende, WANG Wei, GAO Jian. Twitter black and grey industry classification model integrating BiLSTM and CNN[J]. Computer engineering and applications, 2025, 61(1): 186−195.

[14] 张婷. 数字经济时代数据犯罪的风险挑战与理念更新——以数据威胁型网络黑灰产为观察对象[J]. 法学论坛, 2022, 37(5): 121−128.
ZHANG Ting. Risk challenges and concept renewal of data crime in the digital economy era: taking data-threatening cyber black and grey industries as the observation object[J]. Legal forum, 2022, 37(5): 121−128.

[15] ALDAIHAN A, ALOTAIBI F , MAFFEIS S. Clouseau: a hierarchical multi-agent approach for cyber attack investigation[C]//2025 IEEE Annual Computer Security Applications Conference. Honolulu: ACM, 2025: 516−532.

[16] FARD A M, ESTER M. Collaborative mining in multiple social networks data for criminal group discovery[C]// Proceedings of the 2009 International Conference on Computational Science and Engineering . Vancouver: IEEE, 2009: 582−587.

[17] 胡今鸣, 胡啸峰, 石磊, 等. 基于强化学习的超高层建筑非法入侵情景推演方法[J]. 智能系统学报, 2025, 20(4): 958−968.
HU Jinming, HU Xiaofeng, SHI Lei, et al. Scenario deduction method for illegal intrusion into super high-rise buildings based on reinforcement learning[J]. CAAI transactions on intelligent systems, 2025, 20(4): 958−968.

[18] 罗双春, 黄诚, 孙恩博. 基于目标识别与主题引导对话的黑灰产威胁情报挖掘[J]. 信息安全学报, 2025, 10(3): 35−47.
LUO Shuangchun, HUANG Cheng, SUN Enbo. Threat intelligence mining for black and grey industries based on target recognition and topic-guided dialogue[J]. Journal of cyber security, 2025, 10(3): 35−47.

[19] 芮兰兰, 邓淑予, 陈子轩, 等. 基于多智能体深度强化学习的智能网联汽车服务迁移优化方法[J]. 通信学报, 2026, 47(1): 141−155.
RUI Lanlan, DENG Shuyu, CHEN Zixuan, et al. Service migration optimization method for intelligent connected vehicles based on multi-agent deep reinforcement learning[J]. Journal on communications, 2026, 47(1): 141−155.

[20] 黄艳, 李小龙, 王权森, 等. 基于决策思维链驱动和多智能体协同的防洪大模型构建与实践[J]. 水利学报, 2026, 57(6): 918−931.
HUANG Yan, LI Xiaolong, WANG Quansen, et al. Construction and practice of a flood control large model driven by decision-making chain-of-thought and multi-agent collaboration[J]. Journal of hydraulic engineering, 2026, 57(6): 918−931.

[21] 马晓飞, 王中钰, 王佳. “人工智能+”公共政策研究: 基于通用人工智能驱动的多智能体仿真方法 (AGI+MAS)[J/OL]. 电子政务, 1−12[2026−04−01].
MA Xiaofei, WANG Zhongyu, WANG Jia. “Artificial Intelligence+” Public Policy Research: A Multi-Agent Simulation Method Driven by Artificial General Intelligence (AGI+MAS)[J/OL]. E-Government, 1−12[2026−04−01].

[22] HE Junda, TREUDE C, LO D. LLM-based multi-agent systems for software engineering: literature review, vision and the road ahead[J]. ACM transactions on software engineering and methodology, 2025, 34(5): 1−30.

[23] 李衡峰, 郑榕, 邓菁, 等. 基于大模型的电信网络诈骗预

警技术研究[J]. 警察技术, 2025(4): 13−18.
LI Hengfeng, ZHENG Rong, DENG Jing, et al. Research on early warning technology for telecom network fraud based on large models[J]. Police technology, 2025(4): 13−18.

[24] ZHU Mingjie , LI Mingcheng , CUI Mingzhang, et al. A survey on large language model-based multi-agent systems: paradigms, applications, and challenges[EB/OL]. (2026−02−20)[2026−04−01]. https://www.techrxiv.org/doi/full/10.36227/techrxiv.177160638.89229642/v1.

[25] 窦路遥, 魏凤, 邓阿妹, 等. 高价值专利的价值自动解释研究——基于多智能体技术[J/OL]. 图书情报工作, 1−20[2026−04−01].
DOU Luyao, WEI Feng, DENG Amei, et al. Research on automatic value interpretation of high-value patents based on multi-agent technology[J/OL]. Library and information service, 1−20[2026−04−01].

[26] 张益伟, 梁丽芝, 周璧. 基于 LLM 语义生成和 GAT 图结构表征的黑灰产情报检测[J]. 情报杂志, 2026, 45(4): 91−101.
Zhang Yiwei, Liang Lizhi, Zhou Bi. Black and Grey Industry Intelligence Detection Based on LLM Semantic Generation and GAT Graph Structure Representation[J]. Journal of Intelligence, 2026, 45(4): 91−101.

[27] ALHUZALI A. LLM-powered threat intelligence: a retrieval-augmented generation approach for cyber attack investigation[J]. Peer J computer science, 2025, 11: e3371.

[28] BLEFARI F, COSENTINO C, PIRONTI F A, et al. CyberRAG: an agentic RAG cyber attack classification and reporting tool[EB/OL]. (2025−07−03)[2025−07−10]. https://arxiv.org/abs/2507.02424.

[29] 石拓, 曾昭龙, 韩娜. 融合大语言模型和语义聚类的电信网络诈骗引流话术文本分析[J]. 情报杂志, 2025, 44(10): 105−112.
SHI Tuo, ZENG Zhaolong, HAN Na. Text analysis of telecom network fraud inducement scripts integrating large language models and semantic clustering[J]. Journal of intelligence, 2025, 44(10): 105−112.

[30] 尚思佳, 陈晓淇, 林靖淞, 等. 基于图挖掘的黑灰产运作模式可视分析[J]. 信息安全研究, 2024, 10(1): 48−54.
SHANG Sijia, CHEN Xiaoqi, LIN Jingsong, et al. Visual analysis of black and grey industry operation patterns based on graph mining[J]. Journal of information security research, 2024, 10(1): 48−54.

[31] 李明锋, 李欣, 王军杰, 等. 面向情报文本的多智能体无监督事件骨架生成方法[J]. 情报杂志, 2026, 45(4): 110−120.
LI Mingfeng, LI Xin, WANG Junjie, et al. A multi-agent unsupervised event skeleton generation method for intelligence text[J]. Journal of intelligence, 2026, 45(4): 110−120.

[32] WEI J, WANG Xuezhi, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824−24837.

[33] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese bert-networks[EB/OL]. (2019−08−27)[2025−01−01]. https://arxiv.org/abs/1908.10084.

[34] 赵颖, 付铄雯, 赵鑫, 等. 黑灰产网络资产图谱构建与可视化[J]. 计算机辅助设计与图形学学报, 2024, 36(8): 1181−1193.
ZHAO Ying, FU Shuowen, ZHAO Xin, et al. Construction and visualization of cyber asset graph for black and grey industries[J]. Journal of computer-aided design & computer graphics, 2024, 36(8): 1181−1193.

[35] 许浩. 金融黑灰产的新样态与刑法规制[J]. 法律适用, 2025(5): 115−131.
XU Hao. New forms of financial black and grey industries and criminal law regulation[J]. Journal of law application, 2025(5): 115−131.

[36] 张新平, 肖正隆. 后互联网时代网络黑灰产的法律应对与技术治理[J/OL]. 中南民族大学学报 (人文社会科学版), 1−9[2026−04−01].
ZHANG Xinping, XIAO Zhenglong. Legal response and technological governance of cyber black and grey industries in the post-internet era[J/OL]. Journal of South-Central Minzu University (Humanities and Social Sciences), 1−9[2026−04−01].

作者简介：

韦科顺，博士后，主要研究方向为网络安全、应急管理及数字法治。E-mail: keshun1218@163.com。

石磊，副研究员，博士，中国人工智能学会智能服务专委会委员，主要研究方向为智能信息处理、大数据分析与挖掘、社交网络搜索及人工智能。E-mail: leiky_shi@cuc.edu.cn。

石拓，教授，博士，主要研究方向为数据警务、人工智能。发表学术论文 40 余篇。E-mail: shituo@bjpc.edu.cn。

[责任编辑：孟玮]