基于智能体工作流的高级钓鱼邮件检测方法

金建栋，黄正，胡占宇，邹远鑫，秦辉东，赖清楠，杨加，周昌令
（北京大学计算中心，北京 100871）

摘　要：为了应对日益复杂的高级持续性威胁（APT）及钓鱼邮件攻击，提出了一种基于智能体工作流的钓鱼邮件检测方法——PhishingAgent。该方法结合多源知识库和安全工具，充分发挥LLM的推理能力，提升对复杂钓鱼邮件攻击的识别精度与推理深度。智能体工作流基于双系统推理技术：首先通过快速检测系统实现高效的初步威胁识别，随后利用深度推理系统进行精细的语义分析和上下文推理，显著增强结果的可解释性。实验结果表明，PhishingAgent在保证检测精度的前提下提高了检测效率，并在APT钓鱼邮件检测中表现优于现有主流邮件安全防护机制。

关键词：钓鱼邮件；LLM；智能体工作流；双系统推理
中图分类号：TP181
文献标志码：A
DOI：10.11959/j.issn.1000−436x.2024243

PhishingAgent: an agentic workflow method for advanced phishing email detection

JIN Jiandong, HUANG Zheng, HU Zhanyu, ZOU Yuanxin, QIN Huidong,
LAI Qingnan, YANG Jia, ZHOU Changling
Computing Center, Peking University, Beijing 100871, China

Abstract: To address the increasing complexity of advanced persistent threat (APT) and phishing email attacks, an intelligent agentic workflow method for phishing email detection called PhishingAgent was proposed. PhishingAgent integrated multi-source knowledge bases and security tools to fully leverage the reasoning capabilities of large language model (LLM), enhancing the precision and depth of identifying complex phishing email attacks. The agentic workflow was built on a dual-system reasoning framework, a rapid detection system facilitates efficient preliminary threat identification, followed by a deep reasoning system that conducted detailed semantic analysis and contextual inference, significantly improving the interpretability of results. Experimental results demonstrate that the PhishingAgent increases detection efficiency without sacrificing accuracy and outperforms existing mainstream email security mechanisms in detecting APT-related phishing emails.

Keywords: phishing email, LLM, agentic workflow, dual-system reasoning

0　引言

钓鱼邮件攻击是一种社会工程学手段，攻击者通过伪造电子邮件，诱使受害者泄露敏感信息或执行恶意操作。近年来，钓鱼邮件攻击技术显著提升，攻击者通过伪造邮箱地址、混淆内容以及利用认证协议漏洞等手段，成功绕过邮件防护，增强了邮件的欺骗性。传统检测机制难以应对这些不断演变的策略，导致信息安全面临严重威胁，因此，对钓鱼邮件的检测方法受到广泛关注。

钓鱼邮件攻击广泛应用于高级可持续性威胁（APT, advanced persistent threat）技术中，用于获

收稿日期：2024−10−28
通信作者：周昌令，zclfly@pku.edu.cn

取目标系统的初步访问权限[1]。攻击者通过研究目标组织的沟通习惯，发起高度定制化的钓鱼攻击，显著提升攻击载荷的隐蔽性和触发率。此类定制化攻击手段又称为鱼叉攻击。如图1所示，这些邮件通常来自经过认证的域名和IP地址，恶意内容常隐藏在超链接或Word宏文档等附件中，意图规避检测机制。另外，为了提升伪装性，钓鱼邮件的格式和表达口吻等与目标的内部沟通邮件类似。从常见的邮件特征中难以分辨钓鱼邮件，只有上下文中的多个部分组合才显现出恶意意图。因此，此类攻击策略对传统的基于恶意签名或简单规则的检测系统构成了严峻挑战。

现有的大语言模型（LLM, large language model）在自然语言处理领域，如文本生成和情感分析中，已展现较大潜力。然而，尽管LLM在文本分析方面取得了进展，但其在钓鱼邮件检测中的应用研究仍显不足。在钓鱼邮件检测任务中，当前LLM主要依赖于单一的邮件内容分析，对传统攻击特征，如恶意网络统一资源定位地址（URL, universal resource locator）、可疑发件人来源、异常邮件头信息等，存在一定局限性。此外，这些模型在处理具有隐蔽意图的高级钓鱼攻击时，容易受制于缺乏上下文理解和深层次语义推理的能力，判断结果缺乏推理过程，可解释性较差。同时，现有的LLM方案在与其他安全工具，如网页扫描、代码分析、二进制逆向分析等的交互能力上存在不足。

针对新型钓鱼邮件攻击的检测挑战及现有LLM在此任务中的局限性，本文提出一种基于智能体工作流的钓鱼邮件检测方法，并开发了一个原型系统PhishingAgent，如图1所示。该方法赋予LLM任务分解与规划的能力，提升其在复杂上下文中的深层推理性能。此外，引入智能体工作流可还原完整的推理流程，增强模型判断结果的可解释性，便于安全专家进一步分析与优化。通过整合传统攻击特征数据库、来源权威性数据库、威胁情报信息及自定义数据库，提升模型对传统钓鱼邮件攻击特征的识别能力。同时，集成网页浏览、代码分析、文件扫描等多种安全工具，扩展LLM的能力边界，应对新型钓鱼邮件的挑战。PhishingAgent采用双系统推理，对待测目标按照钓鱼邮件攻击指标进行初步判别后再进行详细推理，从而提高整体识别效率。本文主要有以下3点贡献。

1) 提出了一种创新的钓鱼邮件检测方法，该方法有效利用LLM的高阶推理能力，通过整合外部知识库和专业安全工具，实现了对邮件的多维度分析，验证了LLM在网络安全垂直领域的应用潜力。

2) 设计了一种基于双系统推理的智能体工作流框架，包括快速检测系统（System 1）和深度分析系统（System 2），允许根据具体场景需求灵活调整检测深度和效率的平衡，提高了方案在实际应用中的适应性。

3) 实验结果表明，与主流邮件厂商的安全防护机制相比，PhishingAgent在检测常见APT钓鱼邮件样例时展现出更高的准确率，并且具有更强的可解释性。

1 背景和相关工作

1.1 钓鱼邮件检测技术

现有的钓鱼邮件检测技术主要分为基于特征的方法和基于内容的方法。如图1所示，基于特征的方法通过从电子邮件中提取各类特征，如发件人信息、主题、正文及附件等，进而训练分类器以区分钓鱼邮件与合法邮件[2]。相对地，基于内容的方法则着重分析电子邮件的内容，包括文本、图像和链接，以识别可疑模式和异常情况[3]。现有方法主要关注检测结果本身，如精度和准确度等，而缺乏对输出结果的说明。

随着钓鱼邮件攻击的复杂性与频率不断增长，研究更先进的检测方法显得尤为重要。传统检测技术难以有效应对攻击者采用的新型手段，例如电子邮件地址伪造和内容混淆等[4]。这些伪装手段令钓鱼邮件在传统特征上表现得与正常邮件相似，从而有可能绕过检测系统，并使收件人放下警惕。

1.2 APT技术在钓鱼邮件中的应用

APT组织采用多种复杂的钓鱼技术以获得对目标网络的初始访问权限[1, 5-6]。这些技术包括使用混淆的文件扩展名来伪装恶意附件，例如APT32、Inception Framework和Patchwork APT等组织都曾采用此类策略[7]。APT28、Lazarus Group、TA505和APT-C-25等APT组织经常在附加于钓鱼邮件的Microsoft Word文档中嵌入恶意宏代码。当收件人启用这些宏时，恶意代码便会执行，从而危害系统安全[8]。此外，APT32、Patchwork APT、Lazarus

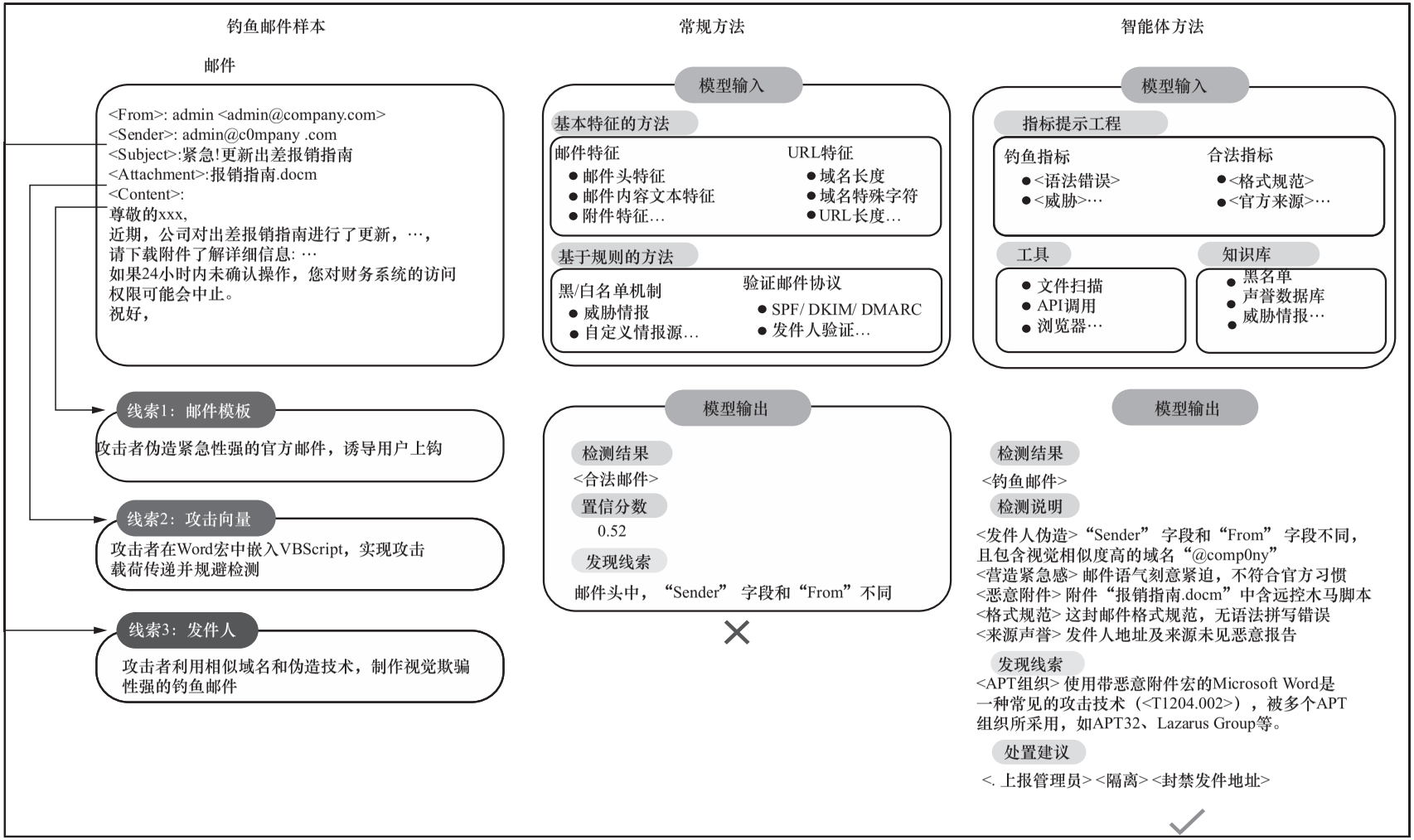

图 1　典型钓鱼邮件攻击与传统及智能体识别方法

Group和MuddyWater等APT组织常在钓鱼邮件或附件中包含恶意脚本，这些脚本可下载其他恶意软件、窃取敏感数据或在受感染系统上建立持久性控制[9]。APT-C-44等APT组织则经常在钓鱼邮件中使用恶意URL，诱导收件人访问钓鱼网站或恶意软件托管平台。这些URL可能被相似混淆，或被超链接到不一致的URL文本上[10]。社交工程策略，如冒充可信个人或组织、制造紧迫感或提供诱人的激励措施，常被APT组织用于操纵收件人泄露敏感信息或采取破坏安全的行动[11]。

1.3 LLM的工具集成

LLM在文本分类和异常检测等自然语言处理任务中表现出色[12]。近期研究表明，将外部工具与LLM整合，能够显著提升其在多种任务中的表现[13]。通过集成外部工具，LLM不仅能访问知识库、具备计算能力，还能够与现实世界进行交互，从而有效应对更复杂的问题。

现有研究探讨了将搜索引擎[14]、计算器[15]以及特定领域API[16]等外部知识和工具与LLM集成的可行性。这些工具已被证明可以提高LLM在问答、数学推理以及基于API的问题解决等任务中的表现。然而，专门针对电子邮件钓鱼检测的工具集成尚未得到充分研究。本文旨在开发一个基于LLM的智能体，利用信息安全和钓鱼邮件检测领域相关的外部知识和安全工具增强其推理能力，提升钓鱼邮件检测的准确性。

1.4 小结

本节分析揭示了钓鱼邮件检测领域的3个关键趋势：首先是攻击者，尤其是APT组织采用的钓鱼技术日益复杂，给现有检测方法带来了新的挑战；其次，对钓鱼邮件攻击，LLM能够依赖其推理能力来分析和识别复杂的恶意意图；最后，通过整合外部知识和工具，LLM在垂直领域任务中的表现得到了显著提升。然而，目前尚未有专门针对钓鱼邮件检测的LLM应用研究。

2 方案设计与实现

2.1 整体设计

在APT钓鱼邮件检测领域，主要面临2个关键挑战。首先，如何准确模拟安全研究人员的决策过程，实现对APT钓鱼邮件的精准识别和分析。这是一项需要丰富经验和深入洞察的精细化工作，对系统的检测精度提出了极高要求。其次，如何显著缩短APT钓鱼邮件的分析时间，提高决策效率。传统上，APT钓鱼邮件的分析往往需要安全专家投入时间进行深入研究，这在面对大规模邮件流量时变得尤为困难。因此，如何在保持专家级分析质量的同时，大幅提升决策速度，成为本文的另一个关键挑战。

本文提出的基于LLM的智能体工作流钓鱼邮件检测方法，旨在解决上述2个挑战。本文方法的核心是由LLM驱动的智能体工作流，采用双系统推理实现高效且深入的分析。首先通过工具集成和知识整合对邮件进行多维度分析，快速得到结构化的初步分析结果，再引入LLM的深度语义分析能力增强对复杂上下文的感知，并对邮件意图进行推理判断，提供可解释的输出。图2展示了该方案的整体架构，主要包括以下3个部分。

1) 知识与工具集成：整合外部知识库（如来源声望数据库、威胁情报库等）和专业安全工具（如协议校验、文件扫描、网页浏览等），拓展模型的知识空间和分析能力，为决策提供全面、精确的参考信息。

2) 语义化检测指标：将常见APT钓鱼邮件攻击策略总结为语义化的检测指标，通过提示工程构造推理思维链，指导模型进行深入分析。

3) 双系统推理：采用快速检测阶段（System 1）和深度推理阶段（System 2）相结合的工作流，其中快速检测阶段进行快速初筛和基础分析，深度推理阶段则针对可疑内容进行深度推理，生成准确、精细且可解释的检测结果。

2.2 智能体工作流

本文的智能体工作流设计借鉴了认知科学中的双过程理论，特别是Sloman[17]和Kahneman[18]等研究者提出的System 1和System 2概念。在认知科学中，System 1代表快速、自动、直觉的思维过程，而System 2则代表缓慢、深思熟虑、分析性的思维过程。在LLM的背景下，System 1一般被定义为直接根据输入生成响应的过程，不涉及中间token的生成，而System 2可以定义为任何生成中间token的方法，包括执行搜索或多次提示等技术，用于生成最终的响应。近年来，研究者提出了多种System 2技术，如Chain-of-Thought[19]、Tree-of-Thoughts[20]、Graph-of-Thoughts[21]等。这些方法通

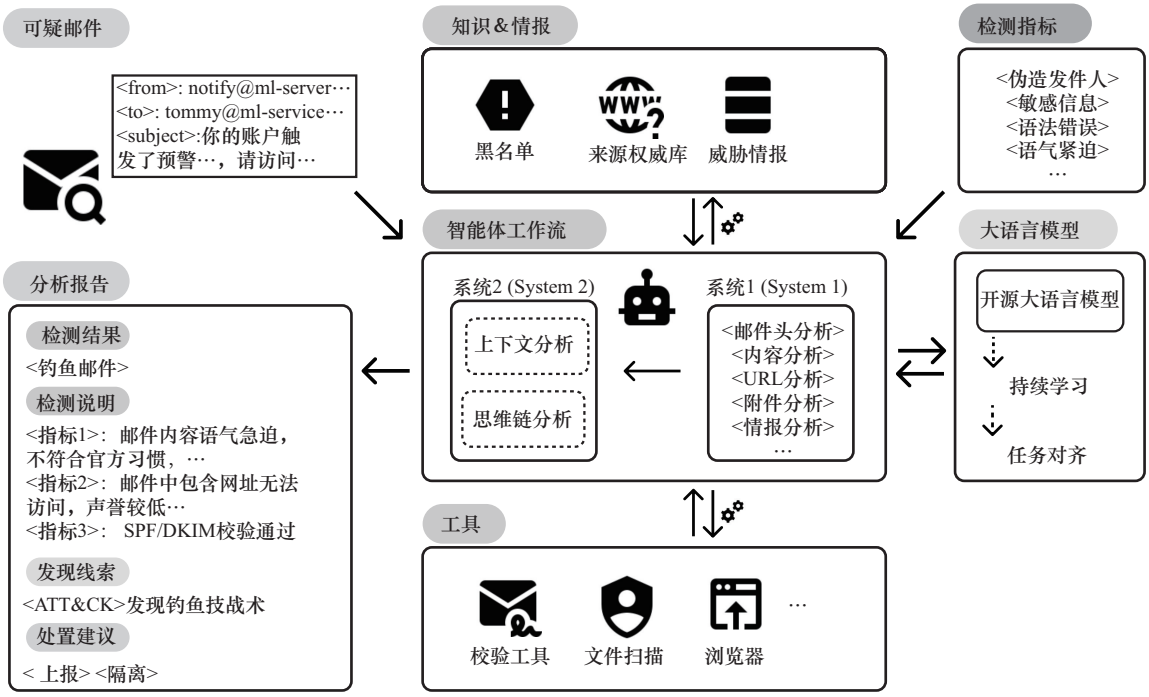

图2　智能体工作流架构示意

过显式推理提高了结果的准确性，但通常会增加推理成本和时延。

本文的核心问题在于：如何在 APT 钓鱼邮件检测中实现类似人类安全专家级别的分析精度，同时显著提高决策效率，使系统能够快速、准确地处理复杂的 APT 威胁。为此，本文将双过程理论应用于钓鱼邮件检测领域，设计了一个基于双系统推理的智能体工作流。这种设计旨在模拟人类专家处理潜在威胁的认知过程：首先快速扫描明显的危险信号（System 1），然后对可疑情况进行深入分析（System 2）。这种方法能够在增强系统应对复杂、隐蔽攻击的能力，同时平衡了推理成本和响应时间。该工作流主要包括以下几个关键阶段。

1) 快速检测阶段（System 1）。在快速检测阶段，系统首先利用邮件解析服务解析邮件内容，提取邮件头、邮件体、附件和统一资源定位符（URL）等关键信息。随后，系统调用邮件头分析器、URL 分析器和附件分析器等模块，对邮件的各个组成部分进行快速分析。这些分析器基于预定义的规则和阈值，结合威胁情报库和安全工具进行快速匹配和检查。其中，邮件头分析器会验证 DKIM 和 SPF 记录，URL 分析器会检查 URL 的信誉度，而附件分析器则会扫描附件中的潜在威胁。在快速检测阶段，目标是快速识别明显的威胁指标，为下一阶段的深度分析提供初步筛选结果。

2) 深度推理阶段（System 2）。核心是引导 LLM 获取合适的上下文并结合场景进行推理。PhishingAgent 的模型底座选用了主流的开源模型（Llama 3.1 70B）。系统首先将 System 1 的分析结果与原始邮件进行整合，形成完整的分析上下文。为了提高模型在复杂场景下的分析能力，本文引入了基于规则的语义化检测指标，定义了一系列与钓鱼邮件检测相关的指标，如发件人可信度、紧急性程度、敏感信息请求等。这些指标被转化为语义化规则，用于指导语言模型的推理过程。随后，系统引导语言模型进行多轮推理：第一轮对整体邮件内容进行初步评估，识别潜在的钓鱼指标；后续轮次则针对特定的可疑点进行深入分析，重点关注风险指标识别技术标识的关键特征。在深度推理阶段，目标是准确地捕捉到细微的攻击特征，同时提供更具可解释性的分析结果。

3) 结果生成与解释。系统基于快速检测阶段和深度推理阶段的结果给出最终的钓鱼邮件检测结论。LLM 生成详细的推理过程说明，包括关键判断依据、使用的工具和知识，以及潜在的攻击指标。这些信息被整合成一个结构化的 JSON 输出，包含评估结果、命中指标和补充说明等字段，便于后续的自动化处理和人工审核。此外，系统还提供

了一个可视化模块，将复杂的推理过程转化为易于理解的图表，极大地提升了结果的可解释性。附录1展示了PhishingAgent对一封钓鱼邮件的分析结果。

在整个工作流中，LLM不仅作为核心的推理引擎，还能够协调各组件的工作，确保检测过程的连贯性和有效性。通过这种设计，系统能够灵活应对各种复杂的钓鱼攻击场景，同时保持对高威胁攻击的精细、高效检测。

2.3 知识整合与工具集成

本文中的智能体工作流通过整合多元化的知识来源和专业安全工具，增强LLM在钓鱼邮件检测中的性能。

在知识整合方面，本文构建了一个多层次的知识体系，为语言模型提供了多源分析视角。本文接入了PhishTank等权威的传统攻击特征数据库，利用其API服务检索丰富的已知钓鱼URL特征。这使得系统能够快速识别常见的钓鱼模式，如可疑的URL结构或已知的恶意域名；其次，本文开发了一个基于WHOIS服务的来源声望验证系统，能够实时查询和分析域名的注册信息，有效识别可疑的新注册域名或匿名注册行为；此外，本文还整合了包括VirusTotal等威胁情报源，使语言模型能够获取最新的攻击向量，已知的APT组织战术和攻击指标等，从而有效识别新兴的攻击模式。

在工具集成方面，本文开发并集成了3种工具，每种工具都专注于邮件安全的特定方面。

1) 邮件头分析器：用于分析邮件认证信息，验证DKIM和SPF记录。利用DKIM和SPF库进行身份验证检查，输出是发件人身份验证结果和域名信誉评估。

2) URL分析器：用于分析邮件中的URL。ur对指定URL进行2类检测，一种是调用浏览器工具对其进行可达性访问，即在沙箱内启动无头浏览器访问，爬取其页面内容；另一种调用知识库对其进行情报的检索，基于VirusTotal威胁情报库查询相关报告。可达性访问和情报检索的结果会作为阶段性结果，用作后续分析。

3) 附件分析器：用于分析邮件附件。同样进行2类检测，一种是调用文件扫描工具对附件内容进行分析，且支持自定义工具接入，本文实现了对office宏代码的快速扫描器；另一种检测方式同样调用威胁情报库，获取文件哈希对应的报告。附件内容分析结果和相关威胁情报作为阶段性结果，用作进一步分析。

本文将知识库和工具的操作接口封装为结构化的API，与PhishingAgent系统进行交互，使其能够快速检索和利用相关信息，接收具体的操作请求并返回结构化的分析结果，确保分析过程的一致性和可解释性。

2.4 语义化检测指标

为了提高LLM在识别和分析复杂钓鱼邮件时的准确性和可解释性，本文设计了一套语义化的检测指标。这些指标不仅涵盖了常见的APT钓鱼邮件攻击策略，还考虑了邮件的语言学特征和上下文信息。本文将这些指标转化为结构化的提示，指导模型进行深入分析和推理。表1展示了部分语义化检测指标。

表1 部分语义化检测指标

指标	描述
可疑发件人	发件人地址异常，与声称不符，可能存在伪造
敏感信息请求	要求提供登录凭证或个人隐私信息
可疑链接	链接显示与实际地址不符，或指向可疑站点
可疑附件	附件格式异常或无关附件，可能导致系统损害
语言/格式错误	存在明显不符合组织通讯规范的语气或格式

这些指标被编码为结构化的提示，例如：“分析邮件中的紧急性表达是否存在设定不合理截止日期、暗示严重后果要求立即行动的语句？如果有，请指出具体内容并解释其可能的意图。”等。通过这些语义化的检测指标，本文引导LLM进行多维度、深入的分析，使模型不仅能够识别威胁指标，还能提供详细的推理过程和判断依据。

3 实验评估

本文设置了一系列试验，以评估PhishingAgent在应对钓鱼邮件，尤其是新型钓鱼邮件攻击时的检测效果。本文在收集的自定义数据集上进行了试验评估。数据集中邮件来自网络和真实APT钓鱼攻击案例，其中的钓鱼邮件包含多个类别，如混淆扩展名、Microsoft Word宏、恶意脚本、恶意URL等，每一个类别中包含5至10封电子邮件。

3.1 有效性评估

本实验在自定义数据集上将PhishingAgent与

主流邮件服务运营商的检测系统进行了对比分析。通过比较各邮件系统对不同类型钓鱼邮件的检测结果，评估 PhishingAgent 在各类钓鱼邮件检测中的有效性。本实验选取了知名公共邮件服务平台（如 Gmail、Outlook、iCloud Mail）和企业级邮件服务平台（如 QQ、163、Coremail）作为对照组。

实验首先从自定义数据集中进一步筛选出钓鱼邮件样本，并按照 APT 战术类型进行分类，包括拓展名混淆、微软 Office 恶意宏、恶意脚本和恶意 URL 等类别。每个类别选取最具代表性的一封邮件作为测试对象。为避免其他条件干扰，实验从多个不同 IP 地址、邮箱账号及客户端向多个目标地址发送这些邮件，并记录明显多数的检测结果。实验主要统计 2 个检测指标：邮件拦截和 APT 攻击预警。邮件拦截指邮件被检测系统拦截，未进入收件人邮箱；APT 攻击预警指系统向收件人发出提示，警示该邮件可能存在钓鱼或 APT 攻击迹象。

表 2 记录了各检测系统对不同类型钓鱼邮件的检测结果。对于 4 种不同类型的钓鱼邮件，PhishingAgent 均能识别其恶意意图并生成报告。相比之下，主流邮件服务平台的检测系统在面对包含微软 Office 恶意宏和恶意 URL 的钓鱼邮件时，普遍无法有效识别和拦截此类攻击。整体而言，PhishingAgent 在检测钓鱼邮件攻击，尤其是新型钓鱼邮件攻击方面有效性更高。

3.2　案例分析

为了提升钓鱼邮件检测结果的可解释性，PhishingAgent 在判断邮件是否为恶意邮件的基础上，附带该结论的详细推理过程。本实验选取了数据集中具有代表性的典型钓鱼邮件和合法邮件，详细分析其推理过程，以评估检测结果的合理性与可解释性。其中，钓鱼邮件完整的样本及分析结果详见附录 1。

针对钓鱼邮件样本，PhishingAgent 从邮件头、URL、附件和情报 4 个维度进行了深入分析和推理，得出了以下多个线索。

1) 可疑发件人：发件人与实际发件地址不符，存在发件人伪造行为。

2) 可疑链接：包含未知域名，且解析失败，可能指向恶意网站。

3) 邮件正文：限定截止日期，营造紧迫感，这是一种常见的钓鱼邮件社会工程手段。

4) 邮件格式：组织结构较差，不符合通常的官方邮件格式。

综合上述线索推理，LLM 判断该样本为恶意邮件。通过对邮件的人工分析，证明 PhishingAgent 的线索判断正确且具有条理性。同样，对于合法邮件样本，通过来源权威性验证、邮件格式验证等维度，确认其合法性，整体逻辑正确。

表 2　　**各检测系统对不同类型钓鱼邮件的检测结果**

邮件服务商	APT 攻击类型	拓展名混淆	Office 恶意宏	恶意脚本	恶意 URL
Gmail	邮件拦截	是	否	是	否
	APT 预警	是	否	是	否
Outlook	邮件拦截	是	否	是	否
	APT 预警	否	否	否	否
iCloud Mail	邮件拦截	是	否	是	否
	APT 预警	否	否	否	否
QQ	邮件拦截	否	否	否	否
	APT 预警	否	否	是	否
163 邮箱	邮件拦截	否	否	是	否
	APT 预警	是	否	否	否
Coremail	邮件拦截	是	否	否	否
	APT 预警	是	否	否	否
PhishingAgent	APT 预警	是	是	是	是

综上所述，PhishingAgent对被测邮件样本的分析推理逻辑正确且有可解释性，既能有效帮助智能体检测分析复杂场景下的钓鱼邮件，又可以增强邮件的可解释性，帮助安全人员进一步分析。

3.3　效率评估

为了评估PhishingAgent在APT钓鱼邮件检测中的性能，本文设计了一个对比实验，选择了和表2实验中同一批APT钓鱼邮件样本进行评估，比较以下方法的检测效率。①基线方法，将原始邮件直接输入LLM获取分析结果，不经过快速检测的处理，为了引导模型产生正确的输出，基线模型结合了语义化检测指标；②PhishingAgent，如上文讨论，原始邮件经过快速检测阶段和深度推理阶段的输出结果；③人类专家评估，本文邀请5名安全分析人员对APT邮件样本进行评估，包括使用情报库、安全工具等梳理攻击样本等，并报告检测结果，同时记录其分析时间用作定性评估。

本文利用基线方法和PhishingAgent，对样本集进行了10次推理，并统计了平均推理时间等指标。表3展示了2种方法在保持100%准确率的同时，推理速度的对比结果。

表3　PhishingAgent与基线方法推理速度的对比结果

方法	推理时间/s			
	平均	最长	最短	标准差
基线	40.27	68.85	11.70	28.57
PhishingAgent	13.62	24.11	5.17	7.86

实验结果表明，PhishingAgent在处理APT钓鱼邮件时展现出显著的性能优势。与基线方法相比，PhishingAgent能够将平均推理时间从40.27 s缩短到13.62 s，实现了66.15%的性能提升，同时较低的标准差表明其具有更高的效率和稳定性。

此外，PhishingAgent能够将分析时间从人类专家所需的1 200~1 800 s缩短到平均13.62 s，凸显出在实际应用中的潜力。在实际情境中，其可以作为现有邮件安全检测机制的补充，针对可疑或上报邮件进行批量、自动化、深入的分析，极大降低人类安全专家的工作负担。

4　结束语

本文针对新型钓鱼邮件攻击的识别挑战，提出了一种基于LLM的智能体工作流方法PhishingAgent，并开发了相应的原型系统。智能体工作流在决策中整合多种知识来源，集成多种外部安全工具，以提高对复杂场景和新型钓鱼邮件攻击手段的检测能力。对该原型系统的评估结果表明，相较于主流邮件服务运营商使用的钓鱼邮件识别方案，PhishingAgent在识别新型钓鱼邮件攻击上表现更为出色。此外，通过增加推理过程，提升了模型结果的可解释性。通过构建一个能适应攻击者策略变化的检测系统，PhishingAgent为应对日益严峻的钓鱼邮件攻击问题提供了一种更有效且可靠的解决策略。

在后续工作中，笔者计划进一步完善模型的持续优化机制，包括实现交互式反馈收集系统，用于持续改进模型和规则，以快速学习和应对新型攻击模式；持续探索基于反馈数据对LLM进行任务对齐，增强其在安全垂直领域的能力，确保系统能够持续进化，适应不断演变的钓鱼攻击技术。

附录1　Phsihing Agent分析结果

表4　邮件头分析

项目	详情
发件人	<□@□.cn>
收件人	□@□□.cn
主题	税务稽查通知书 (Tax Inspection Notice)
日期	□□ 2024 23:12:11 +0800 (GMT+08:00)
发件人域名	□.cn
实际发送域名	未知
发件地址匹配	否

表5　URL分析

项目	详情
URL	https://□.cn/file//1
域名声誉	未知
内容扫描	错误：站点无法解析

表6　附件分析

属性	值
文件名	iom□Dfn.exe.zip
类型	应用/压缩包
恶意软件（VirusTotal）	无
包含VBScript	否
可疑/原因	是-可疑的文件大小
文件大小/B	11 157 533

表 7　　PhishingAgent 输出结果

方面	详情
评估	潜在的网络钓鱼
指标 1	可疑的发件人或发件人欺骗 发件人的电子邮件地址（□@pk-u.cn）看起来可疑，与预期的发件人（可能是合法税务机关电子邮件地址）不匹配。 类型：网络钓鱼
指标 2	虚假的紧迫感或时间压力 该电子邮件制造了一种虚假的紧迫感，声明收件人必须在 2024 年□月 15 日之前完成税务稽查。 类型：网络钓鱼
指标 3	可疑或误导性的链接 该电子邮件包含一个可疑的链接（https://□.cn/file//1），其声誉未知，可能导致恶意软件或网络钓鱼网站。 类型：网络钓鱼
指标 4	语法、拼写或格式错误 该电子邮件包含一些语法错误和格式问题，这对于税务机关的官方通信来说是不寻常的。 类型：网络钓鱼
指标 5	威胁、恐吓或后果 该电子邮件暗示收件人必须遵守税务稽查，这可能被视为一种威胁或恐吓策略。 类型：网络钓鱼
附加说明	该电子邮件似乎针对说中文的收件人，内容以正式的语气书写，这可能会使一些收件人更容易相信。然而，可疑的发件人、虚假的紧迫感和可疑的链接表明存在潜在的网络钓鱼威胁。

参考文献：

[1] DEVORE M R, LEE S. APT (advanced persistent threat) s and influence: cyber weapons and the changing calculus of conflict[J]. The Journal of East Asian Affairs, 2017, 31(1): 39-64.

[2] SRIVASTAVA S, GUPTA S. Phishing detection techniques: a comparative study[C]//Proceedings of the 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). 2021:1-6.

[3] DEBARR D, RAMANATHAN V, WECHSLER H. Phishing detection using traffic behavior, spectral clustering, and random forests[C]//Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics. Piscataway: IEEE Press, 2013: 67-72.

[4] Email spoofing backlashes[R]. 2019.

[5] LEMAY A, CALVET J, MENET F, et al. Survey of publicly available reports on advanced persistent threat actors[J]. Computers & Security, 2018, 72: 26-59.

[6] MITRE Corporation. Phishing, technique T1566 - enterprise | MITRE ATT&CK®[R]. 2023.

[7] BANDLA K, CASTRO S. APT notes[R]. 2022.

[8] CIMPANU C. The scariest hacks and vulnerabilities of 2019[R]. 2019.

[9] Threat Intelligence Team. Patchwork APT caught in its own Web[R]. 2022.

[10] HUSS D, LARSON S. Triple threat: North Korea-Aligned TA406 steals, scams, and spies[R]. 2021.

[11] ADINEH R. Advanced persistent threats (APTs) and the MITRE ATT&CK framework[R]. 2023.

[12] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv Preprint, arXiv: 2005.14165, 2020.

[13] WANG Z, CHENG Z, ZHU H, et al. What are tools anyway? a survey from the language model perspective[J]. arXiv Preprint, arXiv: 2403.15452, 2024.

[14] LAZARIDOU A, GRIBOVSKAYA E, STOKOWIEC W, et al. Internet-augmented language models through few-shot prompting for open-domain question answering[J]. arXiv Preprint, arXiv: 2203.05115, 2022.

[15] PARISI A, ZHAO Y, FIEDEL N. TALM: tool augmented language models[J]. arXiv Preprint, arXiv: 2205.12255, 2022.

[16] LI M, ZHAO Y, YU B, et al. API-Bank: a comprehensive benchmark for tool-augmented LLMs[J]. arXiv Preprint, arXiv: 2304.08244, 2023.

[17] SLOMAN S A. The empirical case for two systems of reasoning[J]. Psychological Bulletin, 1996, 119(1): 3-22.

[18] KAHNEMAN D. Thinking fast and slow[M]. New York: Farrar, Straus and Giroux, 2011.

[19] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2024: 24824-24837.

[20] YAO S, YU D, ZHAO J, et al. Tree of thoughts: deliberate problem solving with large language models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2024: 11809-11822.

[21] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: solving elaborate problems with large language models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17682-17690.

[作者简介]

金建栋（1994– ），男，蒙古族，内蒙古通辽人，北京大学工程师，主要研究方向为网络空间安全、开源情报、智能推理决策等。

黄正（2002– ），男，浙江舟山人，北京大学硕士生，主要研究方向为网络空间安全、软件漏洞挖掘等。

胡占宇（1988– ），男，河北邢台人，北京大学硕士生，主要研究方向为网络空间安全、多智能体决策等。

邹远鑫（2002– ），男，安徽濉溪人，北京大学硕士生，主要研究方向为网络空间安全、自动化渗透测试等。

秦辉东（1994– ），男，河南鹿邑人，博士，北京大学工程师，主要研究方向为数据分析与可视化、工程应用。

赖清楠（1990– ），男，江西兴国人，北京大学工程师，主要研究方向为攻防对抗、人工智能、网络安全管理等。

杨加（1975– ），男，重庆人，博士，北京大学高级工程师、硕士生导师，主要研究方向为人工智能与网络安全。

周昌令（1977– ），男，重庆人，博士，北京大学高级工程师，主要研究方向为网络安全、人工智能、安全大数据分析等。