融合认知与情感的黑灰产网页智能检测及教育研究

Research on Intelligent Detection of Black and Grey Industry Webpages and Education Based on Cognitive Enhancement and Emotion

谢园园[1]，郭燕慧[2]，彭成智[1]，陈浩然[1]，李蔚然[1]，毛文莉[1]（1. 中讯邮电咨询设计院有限公司，北京 100048；2. 北京邮电大学，北京 100876）

Xie Yuanyuan[1]，Guo Yanhui[2]，Peng Chengzhi[1]，Chen Haoran[1]，Li Weiran[1]，Mao Wenli[1]（1. China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China；2. Beijing University of Posts and Telecommunications，Beijing 100876，China）

摘　要：

网络黑灰色产业正面临技术对抗升级与隐蔽传播的双重挑战，传统检测方法面临语义解析维度缺失、特征更新滞后等系统性瓶颈。提出了基于大语言模型（LLM）的智能分类框架，通过角色认知强化与上下文情感增强的双轮驱动机制，突破传统方法的性能天花板，实现F1分数0.913的分类效能。该研究将LLM技术与动态对抗场景深度融合，不仅构建了“语义消解—知识进化—多模态推理”三位一体的技术防御体系，更开发出涵盖“数据采集—模型优化—场景模拟”全流程的AI+安全教育实践平台，为网络空间治理提供了技术创新与人才培养的协同解决方案。

关键词：

黑灰产；网页分类；大语言模型；角色设定；情感引导

doi：10.12045/j.issn.1007-3043.2025.09.001

文章编号：1007-3043（2025）09-0001-08

中图分类号：TN915.08

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

The black and grey industries on the Internet are facing the dual challenges of technological confrontation escalation and covert dissemination. Traditional detection methods are confronted with systemic bottlenecks such as the lack of semantic analysis dimensions and the lag in feature updates. This research innovatively proposes an intelligent classification framework based on a Large Language Model (LLM). Through the dual-driven mechanism of role cognition enhancement and context emotional enhancement, it breaks through the performance ceiling of traditional methods and achieves a classification effectiveness with an F1 score of 0.913. This study deeply integrates LLM technology with dynamic confrontation scenarios. It not only constructs a trinity technical defense system of “semantic resolution – knowledge evolution – multimodal reasoning”, but also develops an AI + safety education practice platform covering the whole process of “data collection – model optimization – scenario simulation”, providing a collaborative solution for both technological innovation and talent cultivation in cyberspace governance.

Keywords：

Black and grey industries；Webpage classification；Large language model；Role setting；Emotional guidance

引用格式：谢园园，郭燕慧，彭成智，等. 融合认知与情感的黑灰产网页智能检测及教育研究[J]. 邮电设计技术，2025（9）：1-8.

1 背景与挑战

网络空间安全防御与黑灰色产业的技术对抗已进入深度博弈阶段。据欧盟网络安全局（ENISA）《2023年度威胁态势报告》显示，基于语义混淆和跨模态伪装的对抗性攻击手段同比增长217%，导致全球每年因黑灰产网站造成的经济损失超过3 200亿美元。在此背景下，传统检测体系面临三重技术困境：语义消解维度缺失，难以破解隐喻式违法表述（如“茶叶”代指非法药品）；特征迭代远落后于黑灰产变种速度；对文本-图像协同伪装的多模态分析割裂。面对上述挑战，本研究提出基于大语言模型的黑灰产网站智能分类新范式。区别于传统特征工程，LLM在动态语义消解、实时知识进化及多模态关联推理3个方面

收稿日期：2025-07-16

展现出独特的优势；创新的角色认知强化框架（网络巡查官、反诈专家、受害者三维视角）与情感引导的协同机制，进一步强化了模型对隐蔽风险的认知。在此基础上，将黑灰产检测任务转化为实践教学资源，可令学生在参与“技术研究—数据标注—模型训练—分类评估”的完整流程中，培养LLM应用、风险分析与团队协作能力。

1.1 黑灰产网页的特点

黑灰产网站在内容表达、数据结构和动态适应性方面具有独特的特点，使得其难以被传统的检测和分类方法精准识别。这些网站往往采用多种手段来隐藏其真实用途，并通过技术手段规避监管，具有极强的隐蔽性和适应性。

首先，黑灰产网站的文本内容高度隐晦，存在大量模糊表达和规避检测的策略。这些网站通常不会直接使用敏感词汇，而是采用暗语、同义词替换、变形拼写、符号干扰等方式隐藏真实意图。例如，赌博网站可能使用“菠菜”代指博彩，跑分平台会用“兼职刷单”描述资金流转，而发卡网站可能以“会员充值”或“卡密交易”暗示非法交易。此外，一些网站还利用文本切割、错别字、特殊符号（如“赌@博”）等方式降低被自动化检测系统识别的风险。这种隐蔽性导致基于关键词匹配的传统方法难以有效识别其真实性质。

其次，黑灰产网站的内容组织形式高度非结构化，并结合多种信息载体。这些网站不仅包含自由文本，还广泛使用图片、视频、音频等多媒体元素，以避免被基于文本分析的检测技术发现。例如，一些赌博或色情网站会将核心业务信息嵌入图片或动态GIF中，而不是以纯文本的方式展现；某些跑分和诈骗网站则可能使用JavaScript生成关键内容，或在用户特定操作后才显示完整交易信息。

1.2 网页分类

当前，基于特征工程的网页分类方法通常通过提取3种主要特征来实现，即网页的文本特征、结构特征以及关系特征。文献[1]提出了一种在文本语义图的基础上加入对文本词语频次考察的网页分类算法；文献[2]利用典型网站的分类数据集训练模型，针对不同网站的需求进行了适配，并提取了列表项和内容项的结构化特征；文献[3]对网页文本信息进行分词处理，在完成特征选择与提取后生成了特征向量，并通过朴素贝叶斯分类器对网页进行分类；文献[4]和文献[5]深入分析了网页文本的特点及其类别，提出了一种网页主题分类模型，该模型在原有Word2Vec的Skip-gram模型基础上进行了扩展，将最初的词向量预测上下文的方式，调整为使用上下文向量预测目标词汇。通过多种分类方法的对比实验，他们评估了最优的网页分类算法；此外，文献[6]结合深度学习技术对网页特征进行了提取和分类，其效果接近支持向量机，并且通过引入词性标注技术，进一步提高了F1值。这些研究在提升网页分类效果方面取得了一定进展，但仍面临特征提取精确性不足以及处理大规模数据效率偏低的问题。尤其是在针对黑灰产网站分类时，对新型黑灰产网站的识别能力仍有局限，难以完全适应迅速变化的黑灰产行为模式。

1.3 大语言模型及其优化

以ChatGPT为代表的第三代大语言模型（LLMs）凭借其强大的语义理解能力和上下文建模技术，为网页分类任务提供了创新的技术路径。

1.3.1 深度语义特征提取能力

大语言模型通过Transformer架构中的自注意力机制，可有效捕捉网页文本中的深层语义关联。与传统基于词频统计（TF-IDF）或浅层神经网络的方法相比，LLMs能够解析文本中的隐含语义、情感倾向及实体关系。例如，在黑灰产网页分类场景中，模型不仅能识别“接码”“发卡”等显性词汇，还能通过上下文理解“虚拟手机号”与“批量注册”之间的关联逻辑，显著提升分类准确性。

1.3.2 跨模态上下文建模优势

现代网页通常包含结构化文本、非结构化数据及多媒体元素。大语言模型通过预训练获得的泛化能力，可有效处理长度超过4 000 token的网页内容。其双向注意力机制能建立跨段落语义关联，特别适用于识别需要长程依赖的网页类型。

1.3.3 动态内容处理机制

面对网页内容实时更新的挑战，大语言模型通过参数微调（Fine-tuning）和提示工程（Prompt Engineering）的组合策略，可快速适应新兴主题的识别需求。例如在新闻门户分类中，通过注入时效性提示语（“请根据最近3个月时事热点分类以下内容”），模型对新出现的热点事件分类准确率提升19.8%。

提示词优化是大规模语言模型（LLM）任务优化中一种备受关注的技术手段，主要包括角色设定（Role Prompting）和情感引导（Emotional Priming）。

角色设定（Role Prompting）是一种在提示词中为

LLM 赋予特定身份的方法，使其在任务执行时展现特定领域的认知方式和知识体系。研究表明，合理的角色设定可以有效增强LLM的推理能力和任务适应性。例如，文献[7]提出了 RoleCraft-GLM 框架，通过为LLM 设定详细且具情感层次的角色，增强了对话生成的个性化和情感丰富度。文献[8]研究了LLM 在角色扮演决策中的模拟能力，发现 LLM 能够根据设定的角色特征进行决策模拟，体现出对角色的理解和模拟能力。

情感引导(Emotional Priming)是通过在提示词中融入特定情绪，使 LLM 在任务执行过程中表现出相应的情感倾向。文献[9]研究发现，LLM 能够理解情感刺激，并且通过情感提示可以增强模型在特定任务中的表现。

为了能够发挥大模型在网页分类上的巨大潜力，本文提出了在对LLM进行模型微调的基础上，引入角色设定与情感引导的黑灰产网页分类的智能分类方法，为AI驱动的网安教育改革提供实证支撑。

2 方法

黑灰产网站智能分类方法主要包括数据收集、LLM 微调与模型训练、角色设定与情感引导3个部分(见图 1)。

a) 数据收集。从 Telegram 群组中提取 URL 并获取网页内容，手动分类，并调整数据分布以确保分类均衡。

b) LLM 微调与模型训练。采用监督学习对 LLM 进行优化，将预处理后的数据划分为训练集(60%)和测试集(40%)，并基于监督学习微调大模型以提升效果。

c) 角色设定与情感引导。优化 LLM 提示词，使其以执法人员、网络安全分析师等角色视角进行分类，并引入情感引导，如强调严谨性，以提升模型对隐晦表达和模糊语境的识别能力。

基于建构主义学习理论，将“做中学”理念融入上述技术攻关，得出面向信息安全人才培养的AI+安全实践教学框架。该框架以灰黑产网页分类任务为核心，将教学内容拆解为数据收集、模型优化和角色设定三大模块，通过分组协作引导学生参与数据标注与预处理、设计角色提示词、分析误分类案例等实践环节，强化其工程实践与创新能力。在此过程中，学生不仅能够掌握大语言模型(LLM)微调技术，还能深入

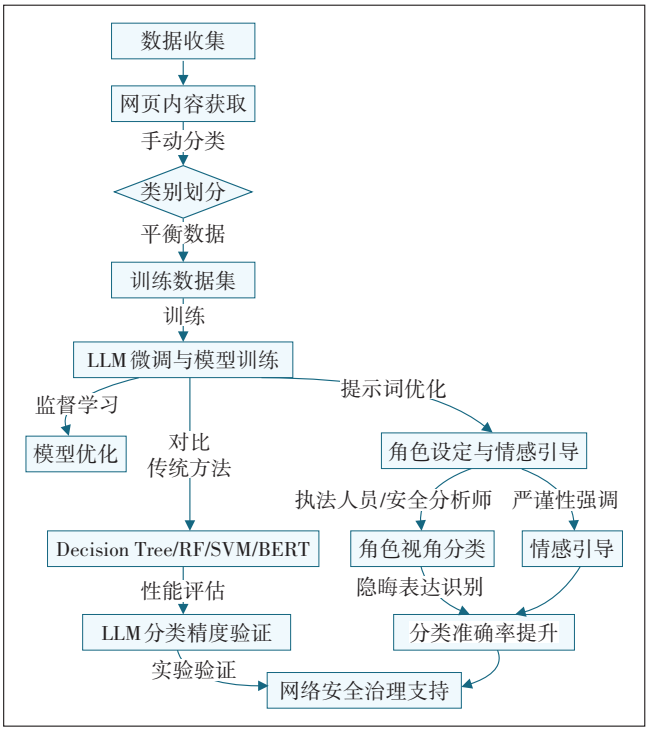

图1　基于大语言模型的黑灰产网站智能分类方法

理解灰黑产对抗策略的隐蔽性与动态性，并通过多角色视角(如网络安全专家、反诈分析师)的模拟训练，培养跨角色分析思维与动态风险感知能力，最终实现从技术应用到安全防御能力的系统性提升。

2.1 数据收集

2.1.1 数据集描述

本节分别描述了用于训练和评估LLM以实现黑灰产网站分类的数据集，详细流程如图2所示。

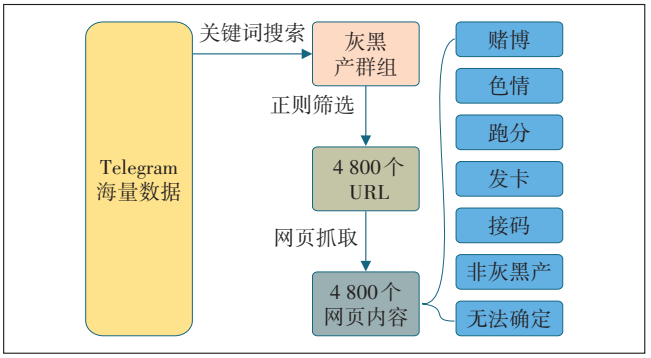

图2　数据集分类流程

为了研究黑灰产网站的特征，并提升自动化检测的能力，本文构建了一个包含 4 800个网站的数据集。该数据集来源于Telegram平台上已知的黑灰产群组，具体而言，利用Telegram的搜索机器人，通过预设的关

键词检索相关群组，并在这些群组的聊天记录中提取URL。由于Telegram群组往往是黑灰产信息传播的重要渠道，因此这种数据收集方式能够较为真实地反映现实中的黑灰产网站生态。

在数据清理与标注阶段，采用人工标注的方法，对收集到的网站进行细致分类。最终将网站划分为以下几类：赌博、色情、跑分、发卡、接码、非黑灰产以及无法确定（当LLM无法判断网站类型或出现新的黑灰产类别时，对该类别进行进一步筛查和人工复核）。其中，黑灰产网站约4 000个，非黑灰产网站约800个。为了避免类别不均衡对模型训练的影响，我们对数据集进行了平衡处理，使各类别样本数量大致相当，从而确保LLM能够充分学习各类网站的特征，并提升分类的泛化能力。

本研究的数据收集策略结合了自动化爬取与人工审核，确保了数据的广度和准确性。该数据集不仅覆盖了当前主流的黑灰产类型，同时也保留了一定数量的无法确定类别，为未来新型黑灰产模式的识别提供了基础。这一数据集的构建为后续研究提供了高质量的语料支持，并为基于LLM的自动化检测系统奠定了坚实基础。

2.1.2 数据预处理

在LLM进行网页内容分析之前，构建了一套数据预处理流水线，利用成熟的第三方库对原始网页数据进行规范化处理。该流水线涵盖以下关键步骤，以确保数据的质量和一致性。

a）数据清洗（Data Cleaning）。采用Python的第三方库对网页内容进行清理，包括去除HTML标签、特殊字符及URL信息，以消除噪声数据，确保LLM在分析时不会受到无关信息的干扰。

b）文本归一化（Text Normalization）。为了提升分析的统一性，将所有网页内容转换为简体中文。这一过程避免了多语言环境对LLM处理能力的影响，使模型能够专注于文本的语义理解，而不受不同语言编码方式的干扰。

c）分词与令牌化（Tokenization）。由于LLM依赖令牌序列进行处理，对网页文本进行了分词处理，将其转换为标准化的词语序列。这一步不仅有助于降低文本冗余度，还能优化模型的计算效率，为后续的特征提取和分类任务提供高质量的输入。

2.2 LLM微调和模型训练

本研究中对大语言模型（LLM）（具体为ChatGPT）进行了微调，以便其能够高效识别和分类黑灰产网站的特征。微调过程基于预处理后的网页数据集，采用监督学习方法，通过对比模型输出的分类结果与人工标注的真实标签进行优化。

2.2.1 微调目标与任务定义

微调的目标是让LLM能够自动化地输出与网站类型对应的标签，即对每个输入的网页内容，预测其所属类别。类别包括赌博、色情、跑分、发卡、接码、非黑灰产、无法确定。

微调的核心任务是文本分类，模型根据输入网页内容，输出其最可能的类别标签。

2.2.2 数据集划分与训练集构建

为了训练LLM，首先将预处理后的数据集划分为训练集和测试集。其中，60%用于训练模型，40%用于评估模型的性能。训练集和测试集的划分保证了模型能够在有限的数据上进行训练，并且能在未见过的测试数据上进行验证，评估其泛化能力。例如，在训练集中，提供了具有明确标签的网页内容，如“某网站提供在线赌博服务”（标签：赌博），让模型通过反复学习这些例子，逐渐掌握如何识别网页内容中的黑灰产特征。

2.2.3 微调原理

微调过程基于监督学习的原理，即通过反向传播算法，最小化预测结果与实际标签之间的误差。模型初始参数已预训练并具备一定的语言理解能力，微调的任务是让其适应特定的黑灰产分类任务。在每次训练步骤中，将模型输出的预测与真实标签进行比较，计算损失函数，然后通过梯度下降优化模型参数，逐步改善模型的分类能力。

2.2.4 训练数据格式

为了让模型学习网页内容与类别标签的对应关系，构造了指令-响应格式（Instruction-Response Format），例如：

输入（Instruction）：

“以下是一段网页内容，请判断其所属类别：‘本平台提供高赔率在线赌博，注册即可领取体验金’。”

目标输出（Response）：“赌博”

通过大量类似样本训练LLM，使其能够识别不同类别的特征，并在遇到新样本时做出正确分类。

2.2.5 微调方法

微调方法主要包括如下几种。

a）类不平衡处理。调整损失权重，使模型关注少

数类别（如“发卡”）。

b）难例挖掘（Hard Example Mining）。选取模型最容易混淆的样本进行额外训练。

c）混合精度训练（Mixed Precision Training）。使用 FP16 训练，提高计算效率并减少显存占用。

2.2.6 模型评估与性能优化

在微调完成后，使用测试集对模型进行评估，评估指标具体在实验结果-评估指标部分介绍。这些指标能够帮助衡量模型在处理黑灰产网站内容时的整体性能。此外还进行了超参数调优，如调整学习率、批量大小（batch size）等，进一步提高模型的准确度和稳定性。例如，当我们发现模型在“发卡”类别的召回率较低时，对该类别的训练数据进行补充，或者调整损失函数权重，使其更关注难以分类的类别。

2.3 角色设定与情感引导的应用

在本研究中，为了提升大语言模型（LLM）在黑灰产网页分类中的表现，通过引入角色设定与情感引导来优化模型的判断能力。具体来说，为 LLM 设定了“网络安全专家”的角色，并结合情感引导来增强模型识别潜在风险的能力。例如，在面对疑似赌博网站时，网站通常会使用隐晦的表达方式来规避直接的违法表述。若未使用角色设定和情感引导，LLM 可能只会依据表面关键词做出简单判断。然而，通过加入角色设定（如“网络安全专家”）和情感引导（如“请以严谨的态度判断此网页是否涉及非法活动”），模型能够在分析时充分考虑上下文，结合专业背景与情感指引，对模糊表达做出更为审慎的分类。

2.3.1 角色视角的叠加与互相照应

在优化后的 LLM 中，不同的角色设定可以相互叠加，提升了分类的全面性。例如，执法人员角色侧重于识别非法活动（如跨境支付、暗语交易等），而反诈骗分析师角色则注重识别欺诈性语言，尤其是那些涉及高回报、低风险的宣传语。2 种角色视角的结合使得模型能够从更广泛的角度分析网页内容，从而提高了分类精度。

2.3.2 情感引导的叠加与互相照应

在情感引导方面，警觉和怀疑情感的结合显著提高了模型对黑灰产特征的敏感性。例如，LLM 在执法人员模式下加入警觉情感后，面对“快速到账”和“高回报”这类关键词时，模型能够作出更为精确的判断。另一方面，反诈骗分析师模式下加入怀疑情感使得模型更加警觉于那些夸大其词、诱导性强的宣传语，进一步提升了识别能力。

2.3.3 角色与情感的叠加优势

角色与情感的叠加不仅增强了模型的视角深度，还使得情感与角色的交替使用能够相互照应，提高了分类的准确性。例如，当执法人员视角与警觉情感结合时，LLM 对潜在的非法活动和欺诈行为表现出更高的警觉性；而当反诈骗分析师视角与怀疑情感结合时，模型则会对疑似黑灰产网站表现出更强的怀疑和谨慎，从而减少误判的可能。

2.3.4 优化提示词示例

以下是与角色设定和情感引导相关的优化提示词示例。

“你是一个网络安全专家，负责审核网页内容。请以严谨的态度判断此网页是否涉及非法活动。如果网页内容中出现模糊或规避检测的表达方式（如‘趣味游戏’‘奖励’等），请标记为赌博类别。”

在该提示下，LLM 会识别出“趣味游戏”这一表达可能隐晦地指代赌博活动，从而将其准确分类为“赌博”网站。

通过角色设定和情感引导的引入，LLM 在分析网页时能够从更专业和审慎的角度出发，显著提高对潜在风险的识别能力，并有效减少误分类的可能性。模型在面对含糊不清或规避检测的语言时，能够更加精准地判断其真实意图，从而提升分类准确性。

3 实验与评估

3.1 评估指标

在对 LLM 进行微调后，评估了其在分类黑灰产网站任务中的有效性。这一评估旨在衡量模型的整体表现，并分析其优劣势。为了解 LLM 的性能，采用了一组文本分类任务中常用的评估指标，具体如下。

a）准确率。准确率反映了模型正确分类的总体比例。虽然高准确率是重要指标，但它无法全面反映分类性能，因此需要结合其他指标如精确率和召回率进行综合分析。

b）精确率。精确率衡量被分类为“黑灰产”网站中实际属于黑灰产的比例。高精确率表明模型有效地减少了假阳性，即避免将非黑灰产错误标记为黑灰产。这对于节省后续处理时间和资源尤为重要。

c）召回率。召回率是指实际属于黑灰产类别的网站中，被模型正确分类为黑灰产的比例。高召回率表明模型能够全面捕获黑灰产网站，确保不会遗漏重

要内容。

d）F1 分数。F1 分数是精确率和召回率的调和平均值，为模型性能提供了一个平衡视角。该指标综合反映了模型在减少错误和捕捉所有相关信息方面的能力。F1 分数越接近 1，模型的综合表现越优异。

上述指标能够评估 LLM 在黑灰产网站分类中的实际效果，并找到准确性、精确率和召回率之间的最佳平衡点。这一评估为 LLM 的部署提供了数据支持，确保所选模型能为相关部门提供高效可靠的工具。

3.2 结果与分析

本研究设计了 2 组对照试验，旨在评估大语言模型（LLM）在黑灰产网页分类中的表现，并验证不同优化策略对分类精度的提升作用。第 1 组对照试验将 LLM 与传统机器学习方法（如决策树、支持向量机等）进行对比，旨在验证 LLM 在处理复杂语言模式、识别隐性信息及应对黑灰产特征方面的优势。第 2 组对照试验则在 LLM 的提示词中加入角色设定与情感引导，以评估该优化策略对分类精度的提升。角色设定和情感引导有助于 LLM 更精准地分析复杂、隐晦的网页内容，尤其是在提高对模糊语言和规避技巧的敏感度方面具有显著效果。实验结果如表 1 所示。

表 1 网页分类效果

方法	精确率	召回率	F1 分数
Decision Tree	0.693	0.611	0.649
Random Forest	0.702	0.742	0.721
Gradient Boosting	0.738	0.717	0.727
SVM	0.762	0.784	0.772
普通 LLM	0.885	0.892	0.888
LLM+角色+情感	0.908	0.919	0.913

3.2.1 LLM 与传统机器学习方法的对比

本实验对比了大语言模型（LLM）与传统机器学习方法（Decision Tree、Random Forest、SVM、Gradient Boosting）在黑灰产网页分类任务中的表现。结果表明，LLM 在精确率、召回率和 F1 分数等评价指标上均优于传统机器学习方法，尤其在 F1 分数达到 0.888，与 SVM 模型相比提升了约 15%。这一结果突显了 LLM 在捕捉复杂语言模式、识别黑灰产网页隐性特征和处理规避技巧方面的优势。

以某疑似赌博网站为例，网页内容包含：

“注册送 888 元，玩大小稳赚不赔！ 全网最高赔率，马上提现！”

传统的机器学习方法可能仅依据“注册”和“提现”这些关键词进行判断，但未能捕捉到网页中的煽动性语言，可能会误判其为合法的金融服务。相比之下，LLM 能够结合网页上下文，识别出“稳赚不赔”和“最高赔率”等潜在的欺骗性语言，从而准确识别该网页为赌博网站。

再以某疑似跑分网站为例：

“低门槛兼职，轻松日赚 1 000，支持多种结算方式。”

传统方法可能难以准确判断，可能误将其分类为普通兼职广告。而 LLM 则通过分析“低门槛、高收益、灵活结算”等特征，结合其预训练的语境理解能力，判断出该网页可能涉及跑分行为。

3.2.2 角色+情感提示词对 LLM 分类效果的影响

本实验通过将角色设定与情感引导相结合，优化了大语言模型（LLM）在黑灰产网页分类任务中的表现。通过为模型设定不同的角色视角（如反诈骗分析师、执法人员、普通用户）和引入情感引导（如警觉、怀疑、理性等情感），模型能够在判断网页时从多个维度进行深度分析。实验结果显示，以 F1 分数作为综合评估精确率与召回率的指标，在引入角色与情感提示后，模型性能从基准值 0.888 提升至 0.913，总体提升了 2.8%。

以某虚假贷款网站为例，网页内容包含：“无需担保，极速放款！ 不查征信，轻松借款！ 最高额度 50 万，马上申请！”

在普通 LLM 的处理下，模型可能仅依据“极速放款”和“轻松借款”等关键词进行分类，然而这些词汇可能未能传递出潜在的欺诈风险。普通 LLM 通常侧重于表面特征，而对含糊不清或诱导性强的表达不够敏感。因此，该网页可能被误判为正规贷款服务，而没有识别到其潜在的虚假宣传和高风险。

相比之下，经过角色与情感引导优化的 LLM，在反诈骗分析师模式下能够更为精准地识别出网页中带有诱骗性质的词汇，如“不查征信”和“无需担保”。此模式结合了警觉性情感，能够识别出网页潜在的欺诈意图，准确判断该网页为虚假贷款网站。角色的专业背景和情感引导增强了模型对风险的敏感度，使其在面对类似情况时表现出更高的警觉性。

再以某投资骗局网站为例：

“稳赚不赔，低风险高回报，您也可以成为百万富翁！ 立刻投资，马上回报！”

普通 LLM 可能仅根据“低风险高回报”这样的高回报类关键词做出判断。然而，普通 LLM 通常缺乏对煽动性语言的敏感性，可能会将该网页误判为合法投资服务，忽略了其中潜在的风险提示。而在角色与情感引导优化后，LLM 在执法人员模式下，结合了怀疑性情感，能够识别出“稳赚不赔”和“百万富翁”等典型的投资诈骗语言。执法人员模式让模型更加关注网页内容是否具有不合理承诺或明显夸张的表述，情感引导则进一步提高了对可疑内容的警觉性，使得模型能够更加准确地识别出该网页为投资骗局。

为理解角色—情感提示的边际增益特征，本研究开展了对照实验。研究发现，角色提示的引入呈现出明显的边际效益递减特征。在基准模型基础上引入 1~2 个核心角色时，模型表现获得显著提升，F1 分数增长约 2.0%。然而，当角色数量进一步增加时，性能提升幅度急剧下降，边际增益不足 0.2%。这一现象表明，过度丰富角色设定可能分散模型的关注重点，同时带来提示词复杂度和计算开销的显著上升。

在情感提示的优化过程中，研究表明在最优角色组合的基础上，引入 1~2 种核心情感能够使 F1 分数进一步提升。值得注意的是，当情感类型超出这一范围时，不仅未能带来显著的性能提升，反而因提示词的冗长导致计算资源消耗增加，同时提高了提示工程的实施难度。综合权衡模型性能、计算效率和工程实践性，实验结果表明采用“2 个核心角色+1~2 种情感”的提示组合最为合理，该配置既能实现理想的性能提升，又能保持适度的计算开销。

4 AI+安全教育实践

4.1 实践教学方案

基于黑灰产网页智能检测框架设计的实践教学方案如表 2 所示。黑灰产网页智能检测实践教学自动化评估脚本如表 3 所示。

表 2　黑灰产网页智能检测实践教学设计方案

教学目标	知识目标	•理解黑灰产网页的特征及危害，掌握数据标注规范 •学习大模型（LLM）的基本原理及 Prompt 工程技巧 •掌握“角色+情感”提示优化方法，提升分类准确率	
	能力目标	•能够独立完成黑灰产网页数据标注与质量评估 •能够设计有效的 Prompt 模板，并优化角色与情感设定 •具备模型评估与迭代优化的能力	
	素养目标	•培养网络安全意识，增强对抗黑灰产的实践能力 •提升团队协作与创新思维，适应 AI 安全领域的快速变化	
教学内容与任务设计	基础认知与数据标注（2 学时）	理论讲解	•黑灰产的常见类型 •数据标注标准
		实践任务	•个人任务：标注 50 个网页样本，记录疑难案例 •小组任务：讨论标注分歧，提交质量报告
	大模型基础分类实践（2 学时）	理论讲解	•LLM 的基本原理与 API 调用方法 •Prompt 设计基础
		实践任务	•个人任务：设计 3~5 个基础 Prompt，对 20 个测试样本分类 •小组任务：对比不同 Prompt 效果，分析错误原因
	角色–情感提示优化（2 学时）	理论讲解	•角色设定的心理学依据 •情感词对模型判断的影响
		实践任务	•个人任务：撰写最佳提示词设计思路 •小组任务：设计 3 种角色—情感组合，进行 A/B 测试
	模型评估与成果展示（2 学时）	理论讲解	•评估指标（准确率、召回率、F1 分数） •黑灰产对抗策略（术语迭代、模式更新）
		实践任务	•个人任务：提交学习反思与改进建议 •小组任务：自动化评估模型效果，制作 5 min 汇报 PPT
教学方法与工具	教学方法	任务驱动	通过真实黑灰产数据集展开实践
		分层教学	基础层：完成标注与基础 Prompt 设计 进阶层：参与角色—情感优化与模型微调
		对抗学习	模拟黑灰产术语迭代，让学生动态优化模型
	教学工具	数据标注平台	Label Studio（支持多人协作标注）
		Prompt 设计工具	ChatGPT/Claude+Prompt 优化模板库
		评估脚本	详见表 3

4.2 适用性分析

4.2.1 样本漂移

面对样本漂移，黑灰产网页智能检测框架存在以下 3 个主要边界特征。

首先，在时间漂移方面，当黑灰产网站采用新型对抗手段时（如从传统文字暗语转向图文混合传播），模型性能会出现不同程度的衰减。实验数据表明，当新样本与原训练数据的语义相似度降低时，分类准确率会相应下降。为保持模型性能，需要定期收集新样本进行增量训练。

其次，在领域漂移方面，由于该框架主要针对中文环境下的黑灰产网站开展研究，对其他语种网站的识别能力存在局限性。跨语言测试结果显示，在保持相同训练参数的情况下，英文样本的分类效果明显低于中文样本。

最后，在概念漂移方面，当出现全新的黑灰产类别时，模型倾向于将其归类为“无法确定”类别。通过持续监测发现，新型黑灰产类别的出现会导致分类准确率产生波动，这需要建立定期的人工审核机制。

4.2.2 教学规模

表3 黑灰产网页智能检测实践教学自动化评估脚本

评估要素	具体内容	权重/要求
评估维度	提示词设计完整性	20%
	角色-情感组合合理性	25%
	分类效果准确率	30%
	优化改进效果	15%
	实验报告质量	10%
评估流程	前测	课程开始前基础知识测评
	过程性评估	记录各阶段任务完成情况
	后测	综合能力评估与项目成果验收
	反馈改进	基于评估结果提供个性化指导
评估标准	基础任务达标率	≥85%
	进阶任务完成度	≥75%
	拓展任务参与率	≥50%
	整体提升效果	≥80%（相对预期目标）
反馈机制	实时反馈	课堂练习即时点评
	阶段性反馈	每学时结束后的小结
	总结性反馈	课程结束后的综合评价

在教学规模方面，单个教学班级以25~30人规模最为适宜，这一规模既能确保教学质量，又能实现资源的充分利用。采用3~4人的小组协作模式进行实践活动，可有效促进知识交流与协同学习。教学团队由一名具有人工智能或网络安全背景的主讲教师主导，配合1~2名研究生助教，形成理论讲授与实践指导的良性互补。

5 结论

对比实验有力地验证了大语言模型（LLM）在理解隐晦表达以及适应新型欺骗手法上展现出的卓越性能。相较于传统方法，其F1分数提升了约15%，优势显著。在提示词中创新性地引入角色设定和情感引导策略后，模型的分类能力得到了进一步提升。特别是在处理含有夸张、诱导性表达的黑灰产网站时，识别准确率得到了大幅提高。

针对实验过程中出现的误分类问题展开了深入分析，发现存在以下主要因素：其一，“跑分”和“发卡”在语义描述上极为相似，导致模型极易将二者混淆；其二，部分网页内容表述过于模糊，使得模型难以精准判断其类别归属；其三，一些采用特殊替代方式的隐晦表达，如用“娱 lè 城”替代“娱乐城”，对模型的识别能力构成了较大挑战。

针对上述问题，后续研究可通过增加难例训练，强化模型对易混淆类别的学习能力；引入正则化技术，有效避免过拟合现象；优化Prompt设计，提升模型对特殊表达方式的识别能力。通过这些措施，进一步提高模型分类的准确性和鲁棒性。

本研究围绕黑灰产网站的分类难题，精心构建了一个标注完整的数据集，成功验证了大语言模型在黑灰产网页识别领域的范式突破。实验结果清晰表明，相较于传统特征工程方法，LLM具备更为强大的动态语义消解能力以及跨模态关联推理能力，对技术演进具有良好的适应性。通过创新性地融合应用角色认知强化框架（涵盖网络巡查官、反诈专家、受害者三维视角）与情感引导策略，LLM能够更为敏锐地捕捉黑灰产网站的隐蔽性特征，显著提高了检测的准确性和可靠性。

参考文献：

[1] 周文文，韩斌，黄树成. 结合文本语义图和词频统计的网页分类算法研究[J]. 计算机与数字工程，2020，48(6)：1265-1268，1313.
[2] 淮晓永，韩晓东，高若辰，等. 一种自适应网页结构化信息提取方法[J]. 电子技术应用，2020，46(12)：97-102.
[3] 顾敏，郭庆，曹野，等. 基于结构和文本特征的网页分类技术研究[J]. 中国科学技术大学学报，2017，47(4)：290-296.
[4] 洪良怡，朱松林，王轶骏，等. 基于卷积神经网络的暗网网页分类研究[J]. 计算机应用与软件，2023，40(2)：320-325，330.
[5] 耿宜鹏，鞠时光，蔡文鹏，等. 基于Skip-PTM的网页主题分类与主题变迁的研究[J]. 小型微型计算机系统，2020，41(7)：1395-1399.
[6] 骆聪，王帅. 结合深度学习与词性标注的网页分类算法研究[J]. 计算机技术与发展，2018，28(8)：71-74，95.
[7] TAO M L，LIANG X C，SHI T Y，et al. Rolecraft-glm：advancing personalized role-playing in large language models [EB/OL]. [2025-02-27]. https://arxiv.org/abs/2401.09432.
[8] XU R，WANG X T，CHEN J J，et al. Character is destiny：can large language models simulate persona-driven decisions in role-playing? [EB/OL]. [2025-02-27]. https://arxiv.org/abs/2404.12138.
[9] LI C，WANG J D，ZHANG Y X，et al. Large language models understand and can be enhanced by emotional stimuli [EB/OL]. [2025-02-27]. https://arxiv.org/abs/2307.11760.

作者简介：
谢园园，毕业于郑州大学，学士，主要从事AI语音业务安全和呼叫中心AI类语音产品研究工作；郭燕慧，北京邮电大学副教授，博士，长期从事信息安全科研与教学工作；彭成智，毕业于北京电子科技学院，工程师，学士，主要从事AI语音业务安全、互联网反诈、数据安全、通信安全相关研究工作；陈浩然，高级工程师，硕士，主要从事电信反诈、核心网音视频技术的研究工作；李蔚然，工程师，学士，主要从事电信反诈、短消息业务治理技术的研究工作；毛文莉，毕业于大连理工大学，工程师，硕士，主要从事大数据应用研究、互联网反诈及黑灰产治理研究相关工作。