面向物联网场景的大模型驱动数据合规检测方法

李超豪[①②③] 王浩然[①] 周少鹏[②③] 闫皓楠[④] 张 峰[②④]
鲁天阳[②] 习 宁[④] 王 滨[*①②④]

[①](西安电子科技大学杭州研究院 杭州 311231)
[②](全省智能物联网络与数据安全重点实验室 杭州 310050)
[③](浙江大学计算机科学与技术学院 杭州 310058)
[④](西安电子科技大学网络与信息安全学院 西安 710071)

摘 要：随着《中华人民共和国数据安全法》、欧盟《通用数据保护条例》(GDPR)等国内外法规条例的逐步施行，数据合规检测成为规范数据处理活动、保障数据安全、保护个人与组织合法权益的重要手段。然而，物联网场景下异构设备数据冗长多变、非结构化、内容模糊等特点加剧了数据合规检测的难度，导致传统规则匹配方法容易产生大量的误报。针对上述挑战，该文提出一种新型面向物联网场景的大模型驱动数据合规检测方法：第1阶段，基于全量规则库，利用快速正则匹配算法高效筛查出所有潜在违规数据，并输出结构化初步检测结果；第2阶段，利用大语言模型进行语义级合规复核，设计差异化分类检测策略，针对不同违规类型构建基于思维链与少样本提示融合的增强提示词，用于减少规则差异性与语义模糊性带来的错误结果。该文采集了52种物联网设备的日志与流量数据，形成共计55 080条原始违规检测数据，并在8个主流大模型底座以及不同影响设置参数上开展对比实验。研究结果表明原有仅第1阶段基于规则匹配的检测方法在真实物联网环境下误报率为64.3%，而经第2阶段大模型驱动的复核检测后降至6.9%，且大模型自身引入的错误率控制在0.01%以下。

关键词：数据合规检测；大模型；物联网；提示词工程；正则匹配

中图分类号：TN915.08 **文献标识码**：A **文章编号**：1009-5896(2026)04-1480-15

DOI: 10.11999/JEIT250704 **CSTR**: 32379.14.JEIT250704

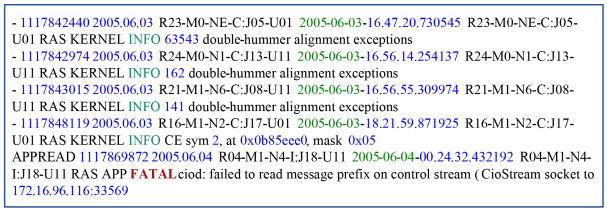
1 引言

随着《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、欧盟《通用数据保护条例》(GDPR)等国内外法律法规的颁布与实施，数据合规检测已成为各行业数据安全建设的重要环节[1,2]。数据合规检测是指通过特定的技术工具和方法，对个人或组织的数据处理活动进行系统性检查与评估，以确保其符合相关法律法规、行业标准、监管要求以及内部政策。合规检测需要综合考虑数据的来源、内容及时间等因素，对每一类可能出现的违规信息进行研判与告警。

然而，物联网设备数量种类激增、业务更加多样复杂，在生产环境中承担着感知数据精准采集、控制命令有效执行等重要职责，因此容易涉及用户与环境的敏感信息。另外，物联网场景下设备日志或流量数据呈现来源丰富、格式不一、长度多变、非结构化等特点，且部分数据内容为长文本(如图1所示)，导致数据中违规信息极为隐蔽、合规边界模糊，对物联网数据合规检测造成极大挑战。传统的数据合规检测方法是通过构建合规规则库，然后基于规则匹配的方式进行风险数据识别。此类方法原理简单、匹配速度快且部署成本低，但存在较多的误报结果，通常需要进一步依靠大量的人工审核来修正检测结果。与此同时，实际物联网场景中合规要求的动态变化以及业务所需的自定义合规要求使得规则设计、维护与判定难度变大。

针对以上挑战，本文提出一种新型面向物联网场景的大模型驱动数据合规检测方法。本检测方法分为2个阶段：第1阶段，基于全量规则库，使用快速正则匹配算法高效筛查出所有潜在违规数据，并输出包含违规原始内容、违规类型等信息在内的结构化初步检测结果。全量规则库包括现行法律法规、标准要求、企业规范以及自定义业务需求等，具有灵活的可拓展性。该阶段通过利用正则匹配算法的高效性克服海量物联网长文本数据审核的挑战，并提取结构化的初步结果数据，提升后续大模型复核的准确性。第2阶段，利用大语言模型(Large Language Model, LLM)对第1阶段的初步检测结果进行正确性复核。针对不同违规类型，大语言模型自适应选择不同的提示词工程，实现差异化

收稿日期：2025-07-28；改回日期：2025-10-13；网络出版：2025-11-13
*通信作者： 王滨 wangbin02@xidian.edu.cn
基金项目：国家自然科学基金(62472335, 92267204, 62402373)，杭州市重点科研计划(2025SZD1A50)
Foundation Items: The National Natural Science Foundation of China(62472335, 92267204, 62402373), Hangzhou Key Scientific Research Program(2025SZD1A50)



(a) BGL数据集部分示例 (样本数5)

(b) 物联网设备运行数据示例 (样本数1)

图 1 物联网设备真实场景运行数据与开源日志数据集样例的差别对比

分类检测。具体而言，本文按照违规类型与复核精度将复核需求分为精准匹配和模糊匹配2种。其中精准匹配指严格按照正则规则的要求去原始内容中进行全文本精确匹配，而模糊匹配指违规边界难以确定的情况下允许调整预设规则对原始内容作模糊匹配。在以上分类的基础上，结合思维链与少样本提示技术，迭代优化提示词，有针对性地补充提示词中对违规内容的解释性，从而增强大模型的上下文语义分析能力，并最终减少规则差异性与语义模糊性带来的错误结果。

本文收集并整理了52种来自不同厂商的物联网设备日志与流量数据，形成了包含共计8类55 080条原始违规检测数据。然后在Qwen2.5-32B-Instruct, QwQ-32B, Qwen3-235B以及DeepSeek-R1-0528等8个不同规模的主流大模型底座上开展对比实验，并探究了提示词方案、大模型系统角色等因素的影响。实验结果表明原有仅第1阶段基于规则匹配的检测方法在真实物联网环境下误报率为64.3%，而经第2阶段大模型驱动的复核检测后降至6.9%，且大模型自身引入的错误率控制在0.01%以下，大幅减少人工复核成本。消融实验也进一步证明本文所提出基于知识补充与分类检测的提示词策略的有效性。

本文的主要贡献如下：

(1)创新性地提出一种面向物联网场景的大模型驱动数据合规检测方法，通过采用快速正则匹配与大模型复核结合的2阶段检测模式，克服物联网非结构化长文本异构数据的挑战，提高了数据合规检测的准确性。

(2)设计了差异化分类检测策略，针对不同违规类型构建基于思维链与少样本提示融合的增强提示词，并通过知识补充与迭代优化，提升提示词对于违规内容与检测策略的可理解性，增强大模型的上下文语义分析能力。

(3)整理制作了8类55 080条真实物联网原始违规检测数据集，并在8个主流大模型底座以及不同影响设置参数上开展对比实验。结果表明本方法相较于原有仅基于规则匹配的检测方法，其误报率从64.3%下降为6.9%，且大模型自身引入的错误率控制在0.01%以下，大幅减少人工复核成本。

2　相关工作

合规检测作为确保组织或技术系统符合法律法规、行业标准和内部政策的重要手段，近年来受到了广泛关注。以下将对主流的检测方法进行介绍。

2.1　基于规则匹配的合规检测

传统合规检测的流程通常包括以下几个步骤：制订计划、收集数据、风险评估、实施审核和编写报告，其中检测方法有自动化合规性扫描、手动审计检查以及第三方合规技术认证等方案。文献[3]指出，在自动化合规性扫描中，对于敏感字段的传统识别方法有基于规则和关键词的方法。例如，针对银行卡号、证件号等有明确规则的对象，可以根据正则表达式和算法匹配进行检测，而政治敏感词、特殊字段等没有明确信息的规则，需要通过配置关键字的方式来进行匹配。Wang等人[4]通过静态分析工具自动检查和执行隐私政策，确保数据分析程序符合预设的隐私要求，有效减少了人工的参与。

目前，在处理大量复杂多样数据时，传统合规检测方法面临检测精度不足、规则编写困难等问题，存在准确率与召回率无法同时优化的问题。具体而言，提高关键词密度虽能提升违规行为的检测概率，却容易导致大量的误判，反之则容易遗漏潜在的隐蔽性违规行为。随着物联网应用场景的日益复杂多样化，传统的合规检测方案通常难以及时发现新型违规行为。

2.2　基于深度学习的合规检测

近年来，深度学习在自然语言处理领域的应用为合规检测提供了新的解决方案。安鹏等人[3]通过可配置的规则和定制化功能，在传统字典匹配方法的基础上，采用基于双向长短期记忆网络(Bidirectional Long Short-Term Memory network, BiLSTM)和条件随机场(Conditional Random Field, CRF)结合的Bi-LSTM-CRF模型，满足了不同数据类型和行业的合规性需求。李昕等人[5]通过构建分层的知识图谱和隐私政策语料库实现对隐私政策文本的合规性分析，用于面向GDPR隐私政策的合规性检测。郭群等人[6]提出了基于内容和上下文的敏感个人信息实体识别方法，结合规则匹配和词对关系分类架构模型识

别非结构化文本中的复杂敏感实体。张西珩等人[7]基于知识图谱的方法在专有的数据源中作合规检测。

其中，日志数据的合规检测从总体上可以分为基于预测的方法和基于分类的方法，通常需要单独的模型提取日志模板，如文献[8,9]提出了常见的日志模板解析方法。Du等人[10]利用了一种长短期记忆神经网络实现了对日志的异常检测，其性能优于基于传统数据挖掘方法的日志异常检测方法。Meng等人[11]对日志模板的语义表达进行了改进，使用Word2Vec从日志模版中提取日志的语义和语法信息，再将生成的模版向量输入到模型中进行训练。Zhang等人[12]采用双向长短期记忆神经网络来实现日志异常检测，该模型可以捕获日志序列中的上下文信息。除此之外，尹春勇等人[13]提出结合卷积神经网络与BiLSTM的无监督模型，利用语义与数量特征有效检测日志异常。

2.3 基于大模型的合规检测

近年来，以大语言模型为代表的人工智能技术取得了较大进展，其通过对海量文本数据的深度学习，不仅具备自然语言理解与生成能力，还展现出跨领域知识整合、逻辑推理及任务自适应的特性。

近年来已有文章表明大模型在数据安全检测任务中具备潜力。Qi等人[14]的工作表明提示词的设计对大语言模型(Large Language Model, LLM)的检测性能有显著影响，提出了从大规模的预训练语料库中转移知识到日志异常检测领域的应用方法-LogGPT。Liu等人[15]提出一种基于LLM的零样本日志分析方法-LogPrompt，通过提示词策略来提升LLM在异常检测任务中的性能。Xiang等人[16]通过数据内在结构和成对输出比较，自我监督地优化提示，提升大模型的表现和效率。LogLLM[17]通过双向变换器模型(Bidirectional Encoder Representations from Transformers, BERT)提取复杂语义向量，通过提示词对语义向量进行异常检测。Elhafsi等人[18]将视觉信息转化为自然语言，同样使用大模型的经验进行了语义理解并作出判断。Yang等人[19]利用大模型进行异常检测的数据增强、零样本检测及模型选择，通过生成合成数据和推荐最优模型提升了检测性能和模型选择效率。

目前已有基于大模型的检测方案大多采用通用提示词来实现数据异常检测任务。然而，在物联网环境中，这些检测方法难以将适用于多元异构、非结构化且包含大量长文本信息的设备数据。

3 可行性分析

本节探究以下2个问题来验证所提大模型驱动数据合规检测方法的可行性。

(1)RQ1：传统规则匹配方法的局限性以及大模型驱动数据合规检测的可行性？

为探究该问题，本节基于已构建的规则库对采集到的52个设备数据进行规则匹配，并对结果进行标注整理。检测结果如表1所示，使用仅基于规则匹配的检测方法时整体误报率为64.3%，表明这种检测方式存在大量的误报。该误报产生的主要原因为：基于规则匹配的算法是通过正则表达式进行文本匹配的，并没有解析原始语义，如当“{‘event_type_name’:‘acs.acs.eventtype.failforsuperpassword’}”字符串出现在原始内容的片段中时，规则匹配方法会因识别到“password”关键字而错误地认为其泄露了某密码值。又如当数据内容中包含“{<devicesn>123 456 789</devicesn>}”的片段时，规则匹配方法会因识别到连续简易的规律性数字组合(数字1～9)而认为出现了一个安全性很低的口令，但实际上该字符串并不用作口令。这些误报情况在仅基于规则匹配的检测中大量出现。

为解决这一问题，本文引入大语言模型对规则匹配的检测结果展开复核。在可行性实验中(如表1所示)，引入规则匹配与大模型协同检测后误报率下降为6.9%。除此之外，通过显式输出大模型的检测依据，可以验证其将错误的检测结果纠正时采取的思路是正确且具备可信性的。如图2所示，大模型在原始检测结果的基础上完成复核，对原检测判断和原检测原因作分析。值得说明的是，本文通过后续方法设计，使由大模型自身出错而引入的错误率小于0.01%。

表 1 可行性研究实验结果

实验	方法	误报率
1	仅基于规则匹配检测	0.643
2	仅利用LLM检测$_1$(LogGPT)	0.512
3	仅利用LLM检测$_2$(LogPrompt)	0.554
4	规则匹配与大模型协同检测	**0.069**

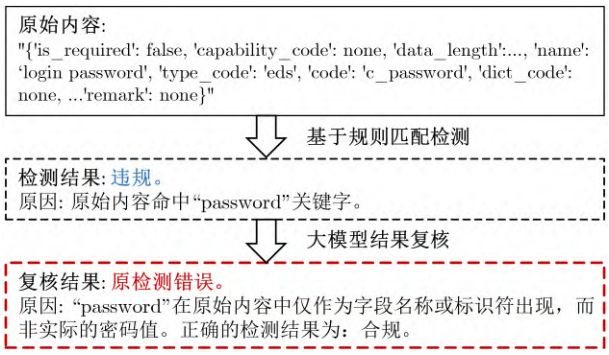


图 2 典型检测结果示例

RQ1结论：传统仅基于规则匹配的检测方法缺乏对内容语义的理解而存在显著的高误报率(64.3%)，而引入大模型复核是一种有效的解决方案，可将误报率降至6.9%。

(2)RQ2：面向原始物联网数据直接采用大语言模型检测是否可行？

尽管大模型具备较好的文本语义解析能力，但为探究其直接处理原始设备数据进行合规检测的可行性，本文基于已有方案，将原始设备数据集直接输入至大模型执行数据合规检测任务。

本文探索了2种典型的提示词设计策略。(a)人工生成方式：参考已有研究工作(文献[14]中 LogGPT 的提示模板)，遵循人工设计的既定模式指导大模型完成检测任务(即表1中的实验2)。(b)自主生成方式：基于自动提示工程思想[15]，利用大模型自身理解能力生成多种提示词(本实验生成8种)，并通过小规模验证集筛选出表现最优的方案用于对比实验(即表1中的实验3)。

从检测结果(如表1)上可知，面向原始物联网数据的大语言模型直接检测方法误报率较高，分别为51.2%和55.4%，相较于规则匹配方法在误报率上仅降低约10%，其原因是已有的大模型检测方法采用的训练与测试数据和物联网真实场景下的设备数据存在差异。如图1所示，BGL是一种常被已有工作用于测试的公开日志数据[20]，其格式统一且文本长度较短。相较而言，真实场景中物联网设备数据来源多样、格式复杂、长度差异大。一方面，大模型的输入通常存在长度限制，无法直接接受超长文本的原始输入数据。另一方面，物联网数据中违规信息更为隐蔽、违规形式更为多样，对大模型直接检测造成极大挑战。

RQ2结论：已有大模型直接检测方案无法适用于长文本、多类型、非结构化的物联网数据。

综上，通过对RQ1和RQ2两个问题的可行性探究实验，本文验证了传统规则匹配、大模型面向原始数据直接检测两种方式的局限性，并证明了大模型对于结构化的初始检测结果进行复核检测的可行性与有效性。因此，本文设计规则匹配与大模型复核协同的检测方法，在使用规则库完成初始正则匹配检测的基础上引入大模型进行误报复核，在整体上降低数据合规检测任务的总体误报率。

4　检测方案设计

4.1 总体设计

本文所提面向物联网场景的大模型驱动数据合规检测方法是一种2阶段检测方法，具体设计为：第1阶段使用规则库对原始设备数据完成快速、高召回率的初筛，通过构建全量规则库并结合用户的自定规则，准确捕获所有潜在的违规数据。第2阶段通过大语言模型复核原始存在违规隐患的数据，将所有违规类型按照匹配精度分类，实现大模型的差异化检测，减少大模型复核时自身的出错率，同时针对不同违规类型构建基于思维链与少样本提示融合的增强提示词，用于减少规则差异性与语义模糊性带来的错误结果。

本检测方案采用2阶段协同检测机制，既能保障合规检测的效率，又可针对性地实现违规内容的复核，本方案的设计依据在于：首先，鉴于现有的设备数据具有种类繁多、数量庞大且结构复杂的特点，其在合规检测时通常包含大量与匹配字段无关的内容(如时间记录、执行状态)，而真正存在违规的内容在原始的信息记录中占比极小，为此，方案第1阶段通过规则匹配算法，对设备数据中特定的结构化或半结构化信息进行精准识别与检测。其次，为高效实现对潜在违规内容的针对性复核，输入到大模型中的复核数据为第1阶段的所有潜在违规数据，通过匹配精确度区分、各违规类别知识补充等思路完成提示词优化工程，从而逐步减低大模型本身检测的出错率。

总体上看，本文所提面向物联网场景的大模型驱动数据合规检测方法流程如图3所示。

4.2 第1阶段：规则匹配合规检测

第1阶段利用预先构建的规则库作为匹配规则进行快速合规检测，此过程的目的是保证遵循规则库的设计尽可能筛选出所有包含违规信息的数据。其中，规则库内的规则来源于相关法律法规、行业标准及内部政策等，同时，为应对新型违规类型的出现，库内的规则也可以由用户根据业务自定义设计。每条设备数据经过规则库时，会依次按照所有库中的正则表达式或关键字进行匹配。例如，某电子邮箱的正则表达式为：r“^[a-z A-Z 0-9._%+-]+@[a-z A-Z 0-9.-]+\.[a-z A-Z]{2,}$”，其中，“@”左右的表达式分别匹配了电子邮箱中的用户名、主机名和顶级域名，规则库中的其他正则表达式设计类似。

原始数据经自定义规则库匹配检测后，生成相应的检测结果。第1阶段检测过程中，除记录原始数据的违规类型外，还同步标注违规位置、风险等级，并记录违规内容、规则名称及原始内容等信息，共同构成结构化检测结果。其中，规则名称Ru(Rule)、原始内容Con(Content)与命中内容MC(Matched Content)是3种关键信息，将作为后

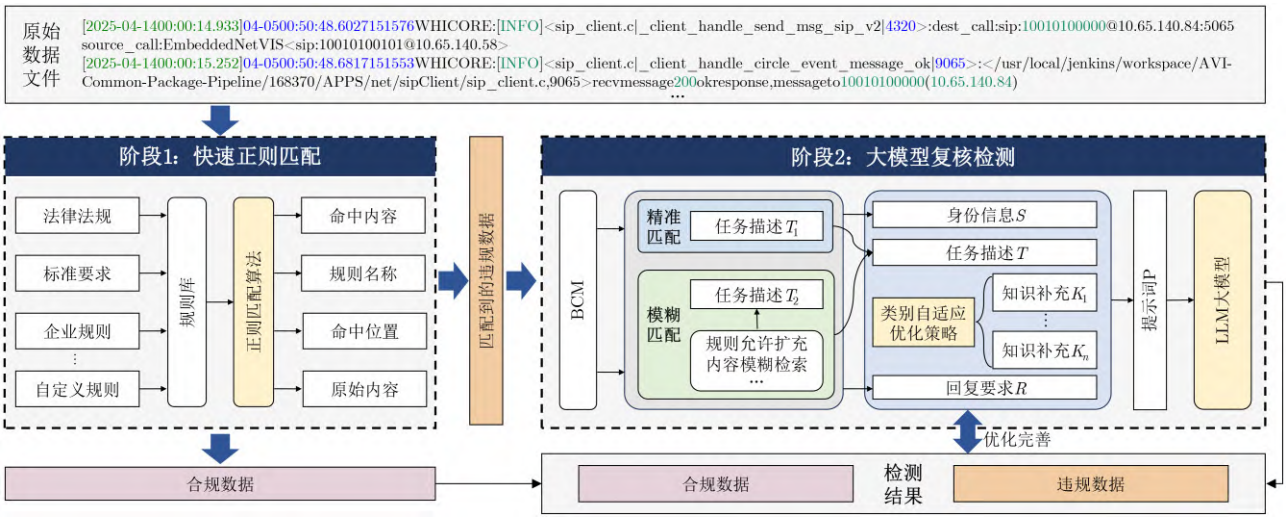


图 3 面向物联网场景的大模型驱动数据合规检测方法流程图

续第2阶段的部分输入数据，为大模型2次检测提供依据。

4.3 第2阶段：大模型复核检测

本文第2阶段使用大模型对第1阶段基于规则匹配得到的结果进行复核检测，其目的为使用大模型判定原有的违规类型是否真实存在，纠正原始错误的检测并减少对原始正确检测的影响，这一过程成为改善人工误报复核的一种方案。大模型复核检测阶段分为数据处理、规则归类和提示构建3个部分。

4.3.1 数据处理

在数据处理阶段，本方法基于第1阶段输出的所有违规数据，通过信息提取完成大模型输入数据的准备。具体而言，针对每条检测结果分别提取出Ru, Con和MC 3个内容(分别表示违规规则、原始内容、命中内容)，获取大模型复核检测的依据，便于后续对提示词进行构建。

这一处理过程的主要意义在于：通过聚焦处理上述3个内容，能够精准捕获与违规判定紧密相关的内容，为大模型的2次复核提供结构且具体的内容。其次，该数据提取思路减少了原始检测结果中与合规检测无关字段(如风险等级、检测时间等)的干扰，有助于降低大模型复核时的出错率。

4.3.2 规则归类

由于违规类型种类多样，不同违规类型对于检测精度的需求也不同。基于此，本文将所有违规类型按照匹配检测精度需求的不同，归类为精准匹配与模糊匹配两种类型。其中，精准匹配类型指的是对该类型进行检测时，大模型需按照原有的命中内容MC去原始内容中进行严格匹配，与此同时，还需根据规则名称Ru来判断Con中相应位置的命中内容是否具备规则中描述的作用或功能。模糊匹配类型指对该类型进行检测时，不仅需要在原始内容中匹配到指定的MC，还需考虑与MC在形式或内容上相近的文本是否出现，例如形式上的大小写变化或内容上的同义词替换，因此要求大模型结合内容与语义进行复核检测。此外，模糊匹配类型也需根据规则名称Ru来判断中相应位置是否具备对应的作用或功能。同时，使用大模型对结果复核时应具备这样的能力：即当原始命中的规则Ru没有真实存在或不符合Ru的违规描述，大模型会自主对原始内容Con进行检测，依靠自主的理解判断是否有其他违规内容。

以下给出上述2种类别的违规示例，示例中仅列举了检测结果中的部分重要内容。

(1){Ru=“敏感关键字”, Con=“‘password’=‘abc123’”, MC=“password”}。这种情况为规则匹配根据关键字检测到了预定义的关键词“password”，因此仅需大模型在指定位置进行校验与判定即可，属于精准匹配类型。

(2){Ru=“弱口令安全风险”，Con=“k123456n”，MC=“123456”}。这种情况属于第1阶段的规则匹配识别到了低强度口令1～6数字序列，但仍需通过上下文来判断该序列是否作为口令出现，因此需要大模型进一步来分析和验证，属于模糊匹配类型。

在大模型的复核检测过程中，这2种检测精度直接对应2种不同的复核思路，在提示词上需要进行差异化的设计，因此在对第1阶段的检测中，将所有判断为违规的检测数据根据其违规类型归类为模糊匹配或精准匹配。

本文采用基于BERT模型[21]微调的方法实现违规类型的自动归类。本文收集了在合规检测中积累的违规案例数据，经人工标注后形成“违规类型-

归类类别”样本数据集，通过对BERT模型进行微调训练后得到最终的归类模型。后续实验使用该模型对本实验中第1阶段生成的所有违规检测结果进行了归类，这两类在后续的提示词设计中具有差异。

4.3.3 提示构建

本文对2种精度分类的提示词P(Prompt)具有不同的设计部分，但其结构基本相同，分为以下4个部分：“身份信息S(System prompt)”、“任务描述T(Task description)”、“知识补充K(Knowledge supplement)”和“回复要求R(Reply request)”，具体设计如下。

(1)“身份信息S”部分明确其作为合规检测专家的角色，需基于给定规则与内容对潜在违规信息进行精准复核，严格遵循合规检测规范完成判定任务。

(2)“任务描述T”部分将数据处理阶段提取出来的规则名称Ru、原始内容Con与命中内容MC作为主要内容，并加入对合规方法和步骤的定义。此外，根据精准匹配和模糊匹配的不同，分别对应2种不同的提示模板T_1和T_2，指导大语言模型按照预期的思路进行判断，以此来避免2种复核方式相互冲突，减少大模型自身出错的概率。

(3)“知识补充K”部分的设计，源于大模型在不同类别检测中存在误报率差异这一现象。基于此，本文依据类别误报率，采用分层级的提示优化方式，通过为特定类别补充新增知识K，进一步降低整体误报率。具体而言，当某一违规类别的误报率显著高于整体均值，且其所需补充的知识可能干扰其他类别的正常检测时，该类别将本单独划分出来，针对性设计补充知识并整合为新的提示词。在实验中，本方法会对BCM归类后的结果根据显著误报率进行2次划分，对高误报违规类型开展重点迭代优化，实现检测精度的定向提升。

知识补充K属于可选部分，针对n个不同的高误报类别检测，可以通过在提示词中引入具体的知识补充(如对规则的详细描述，或规则的额外需求等)来降低原本的误报率，记作K_n。

(4)“回复要求R”部分，提供大模型回复的模板。内容包含：原检测结果(即原始检测是否正确)、违规类型(真正的违规类型，合规/违规)以及判断依据(显式说明复核时的依据)，通过这样的设置，促使大模型生成基本的判断依据，为后续优化其判断思路提供有力支持。

整体的提示词模板如图4所示，基于精准匹配与模糊匹配的分类在提示词上体现在“任务描述T”的不同，基于类别误报率的分类使得提示词在“知识补充K”的不同。本文通过自动化识别第1阶段检测结果中的规则名称Ru来自适应生成提示词P，其基本组成为S, T, K_n(n表示有多少个需要额外补充知识的违规类别)以及R，最终输入到大模型中进行检测。

在提示词工程上面，本文使用了思维链(Chain-of-Thought prompting, CoT[22])与少样本提示词(Few-Shot[23])结合的方法。通过CoT方法使大模型在得出结论前隐式地输出自己的思考步骤，以降低了大模型在过程中出错的可能性。例如在任务描述T中加入“逐步输出思考过程”这类的文本。Few-Shot的思路应用于知识补充K部分，在一些高误报率的分类下，通过设计代表性的案例和解决步骤，在提示词中以经验的方式将这部分的检测要求输入给大模型中。如图4所示，知识补充部分将详细地描述这一类检测类别的判断步骤，包含这类违规的特殊需求和适用准则。这部分内容具备很强的特殊性与专用性，因此Few-Shot的方式适用于灵活的知识补充。

5 实验与分析

5.1 实验说明

本文实验主要探究了在主流大模型基准下，使用两阶段协同检测的方案对现实场景的设备数据进行合规检测的执行效果，本节重点介绍实验开展的基本数据集、评估指标以及使用的基本模型，并给出了对违规数据集的分类结果。

5.1.1 数据集

本方法采用的数据集根据不同阶段的输入划分

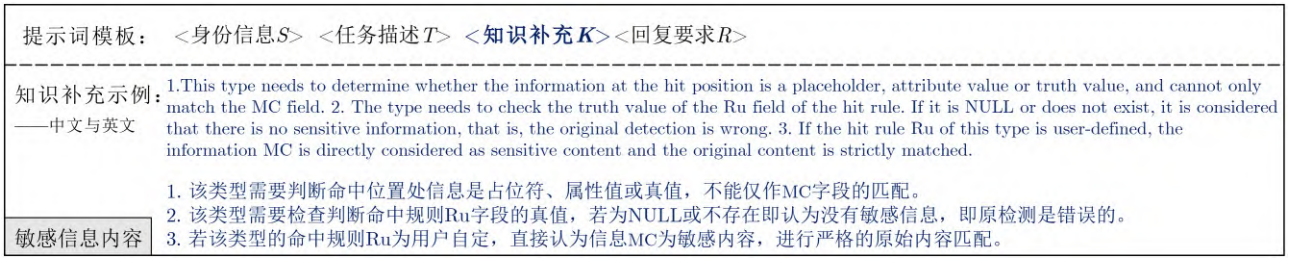


图 4 提示词模板中知识补充部分样例

为2种数据集，分别为初始采集的设备数据集与原始的设备检测数据集。下面分别介绍这2种数据集。

(1)设备数据集：该数据集来源于物联网中设备的数据，采集于52个不同的厂商设备，包含日志与流量数据。该部分数据集经过字段提取、筛选与划分等预处理后，输入到规则匹配阶段进行匹配检测。

(2)设备检测数据集：该数据集是输入到大模型进行检测的数据，来源于第1阶段基于规则匹配方法的检测结果。在第1阶段被识别为违规的设备数据会重新输入到第2阶段进行处理，并根据大模型的判断确定最终的合规情况。该数据集的数据结构与原始数据集相同，均包含Ru, Con, MC等。经过重新整合，本文输入到大模型的复核数据样本为55 080条。在该数据集中，最长的单条记录有32 767个字符，最短的数据为82个字符。本文对该数据集进行统计，依据违规与合规两种类型分类。根据相关法律依据以及内部规定，整理出以下8种存在于该数据集的违规类型：(a)设备保护信息，(b)弱口令安全风险，(c)敏感关键字，(d)证件护照数据，(e)私有链路及邮箱，(f)人权歧视违规，(g)涉政违规，(h)其他违规行为。每种违规类型都对应多种正则表达式，所有违规种类与数目如表2所示。

5.1.2 评估指标

对于第2阶段中大模型的符合结果，本文使用准确率、精确度、真正率等评价指标来评估模型检测效果。准确率(Accuracy, Acc.)是正例和负例中预测正确数量占总数量的比例，精确度(Precision, Pre.)是衡量模型预测正确的样本占总样本的比例，真正率(True Positive Rate, TPR)，也被称作召回率(Recall)，是用于衡量模型正确预测出的正例样本占所有实际正例样本的比例。假阳率(False Positive Rate, FPR)表示被预测为正样本的负样本占总体负样本的比例，也称作误报率，是各个模型的优化重要指标。假负率(False Negative Rate, FNR)反映了所有实际为正类的样本中有多少比例被错误地标记为负类。各公式计算为

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}},\ \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}},$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{F1} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

其中，TP(True Positive)为模型正确预测为正类的样本数量，TN(True Negative)为模型正确预测为负类的样本数量，FP(False Positive)与FN(False Negative)为模型错误预测为正类和负类的样本数量。此外，F1分数为综合考虑模型在正类预测上的精确性和全面性的指标，计算公式如式(4)所示。其中，Recall的值即TPR的值。F1分数取值范围为0～1，值越高表明模型在精确性和全面性的综合表现越优。

表 2 原始设备数据集包含的所有违规类型

编号	数据违规类别	数目(条)	模糊匹配	精准匹配
a	设备保护信息	4 900		√
b	弱口令安全风险	8 779	√	
c	敏感关键字	9 881		√
d	证件护照数据	7 124		√
e	私有链路及邮箱	5 713		√
f	人权歧视违规	6 840	√	
g	涉政违规	5 820	√	
h	其他违规行为	6 023		√
	总计	55 080	26 339	28 741

5.1.3 模型说明

为了验证大模型在合规检测上的能力，本文在Qwen2.5[24], QwQ[25]以及DeepSeek[26]各个版本的基座模型上开展了实验：在小参数模型上使用了Qwen2.5-7B-Instruct模型，在参数量适中的模型上，使用了Qwen2.5-32B-Instruct, DeepSeek-R1-32B, QwQ-32B模型，在大参数模型上，使用DeepSeek-R1-70B, Qwen2.5-72B, Qwen3-235B以及DeepSeek-R1-0528模型。

在本实验中，使用API访问的方式实现了不同基准语言模型的测试，模型默认采用的参数为：初始温度系数(Temperature)为1.0，Top-p[27]值(Nucleus Sampling)为0.3。

5.1.4 分类结果

本节对表2中数据的匹配与归类情况进行了详细说明，并基于匹配精度进行分层级提示优化，对各层级的误报情况进行分析与处理。

首先，将原始划分的8种违规类型输入到BCM中进行归类。模型输出结果为：编号b, f和g的3种类型被归为模糊匹配类别，其余类型则归为精准匹配类别。在此基础上，本文进一步在小样本数据集上开展基于类别误报率的分层级提示优化实验。

为构建实验所需的小样本数据集，本文在设备

违规报警数据集中随机采集了1 000条检测数据，经过分析与标注替换无用样本，构建了各个违规类型下的样本数量相同且每个违规类型均包含误报平衡小样本数据集。提示迭代优化流程如图5所示。实验表明，模糊匹配类别下存在的明显较高的误报率类别为b，精准匹配下存在明显较高误报率的类别为c，因此本文在这两类检测的提示中进行了知识补充。同时，其余类别下也通过少样本提示的方法进行了知识补充，分别合并为“法律法规标准”与“其他违规类型”。

总体上看，本文将设备的违规检测数据集按照匹配精度和具体误报率分为4类，将模糊匹配类别划分为“口令安全风险”和“法律法规标准”2种，将精准匹配类别划分为“敏感信息内容”和“其他违规类型”2种，其数量关系如表3所示。基于上述分类，本实验构造了4种针对性的提示词。为实现最佳检测效果，在不同的类别中设计独特的提示词内容。例如，在模糊匹配提示词中加入“原命中内容可以作简单扩充”等信息，在精准匹配的“敏感信息内容”提示词中，额外补充“检查字段真值”这一需求，在“其他违规类型”中补充“所有的证件数据视为正确的”这一知识，不同的分类下通过少样本提示的思路补充针对性的知识。

5.2 结果与分析

5.2.1 基准大模型对比测试

本节展示了在不同基准大模型下，基于大模型的复核检测方法的检测结果。首先，在第1阶段规则匹配的检测结果中，共计19 695条设备数据包含真实的违规内容，其余35 385条数据均为错误的检测结果，这些数据实际并不包含任何违规内容，属于误报，错误率为64.3%。

本实验保持其他条件不变，在第2阶段中使用不同的大模型作合规检测。实验结果如表4所示，该表分别展示了检测结果中的样本数TP, FN, TN和FP。其中，“Qw”表示Qwen2.5系列模型，“Qw3”表示Qwen3系列模型，“DS”表示DeepSeek-R1系列模型。各个模型的指标分数如表5所示，各个模型在F1分数、FPR和FNR相对关系图如图6所示。

实验的总体结果可以从以下几个方面分析：(1)从F1分数上看，所有模型F1分数平均值达0.861，且最低为0.786，最高为0.942，这表明大模型的检测稳定性均较好。(2)从误报率FPR上看，所有模型中，有7种大模型将总体误报率降至0.30以下，有4种大模型可以将误报率FPR降低至不到0.20，其中效果最佳的模型为Qwen2.5-32B-Instruct, Qwen2.5-72B以及DeepSeek-R1-0528模型，它们的FRR均低于0.10，而DeepSeek-R1-70B的误报率FPR值最高，达到了0.348。(3)从FNR上看，Qwen2.5-72B模型的FNR值较高(达到了0.117)，其余模型都在0.06以下，FNR值最低的模型为Qwen2.5-32B-Instruct，其值计算得0.000 1，在总体表现上，Qwen2.5-32B-Instruct模型的检测效果最优。

基于此实验结果，本文从以下3个方面展开结果分析与原因解释。

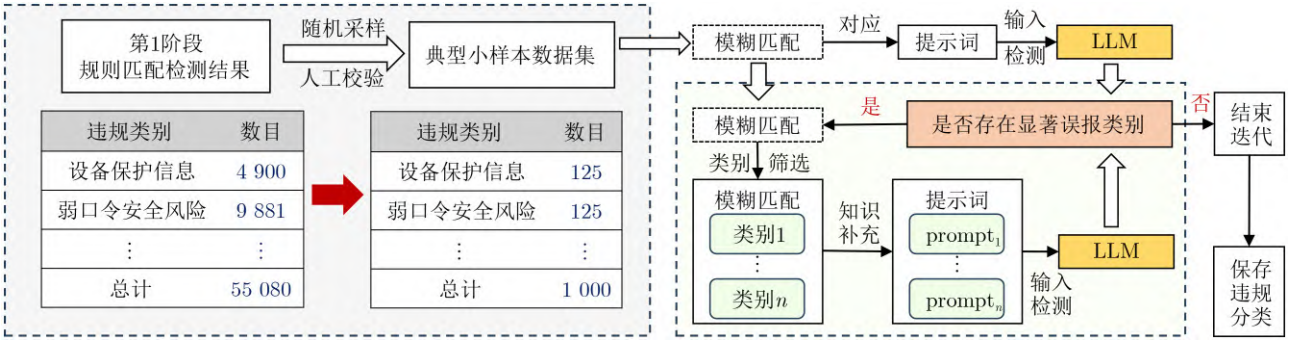


图 5 基于类别误报率实现分层次的提示迭代优化示例图(以模糊匹配为例)

表 3　基于类别误报率的分层次提示词分类结果

BCM归类	基于误报率的分类	知识补充	数目(条)
模糊匹配	口令安全风险b	K_1	8 779
	法律法规标准f, g	K_2	17 560
精准匹配	敏感信息内容c	K_3	18 860
	其他违规类型a, d, e, h	K_4	9 881

表 4 不同基准大模型的检测样本情况

数据类型	正则匹配结果	指标	Qw-7B	Qw-32B	DS-32B	QwQ-32B	DS-70B	Qw-72B	Qw3-235B	DS-0528
违规(T)	19 695	TP	18 993	19 690	19 810	22 000	21 700	17 092	21 704	19 680
		FN	1 177	5	550	215	317	3 263	520	492
合规(F)	35 385	TN	26 695	32 951	26 513	23 625	21 543	33 095	28 912	31 956
		FP	8 215	2 434	8 207	9 240	11 520	1 630	3 944	2 952

表 5 不同基准大模型的检测结果指标以及正则匹配错误率

	指标	Qw-7B	Qw-32B	DS-32B	QwQ-32B	DS-70B	Qw-72B	Qw3-235B	DS-0528
原始错误率: 0.643	Acc.	0.829	0.956	0.841	0.828	0.785	0.922	0.919	0.937
	Pre.	0.698	0.890	0.707	0.704	0.653	0.894	0.846	0.870
	TPR	0.942	1.000	0.973	0.990	0.986	0.883	0.977	0.976
	F1	0.802	0.942	0.819	0.823	0.786	0.888	0.907	0.920
	FPR	0.235	0.069	0.236	0.281	0.348	0.057	0.120	0.085
	FNR	0.058	0.01%	0.027	0.010	0.014	0.117	0.023	0.024

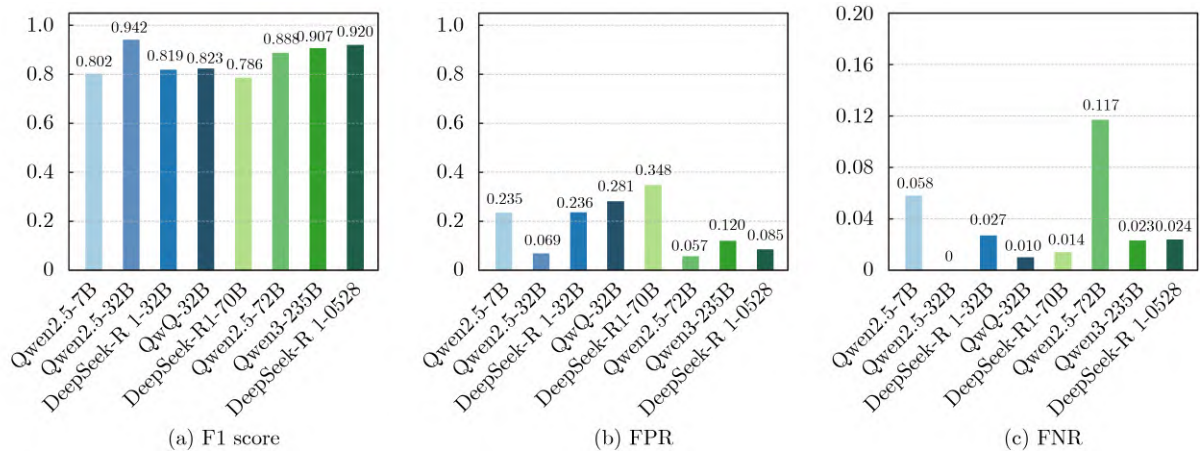


图 6 不同基准大模型上检测结果的F1, FPR和FNR对比图

(1)在模型参数量方面，参数量大的模型在检测结果上并不一定比参数量低的模型效果更优。在实验结果中，Qwen2.5-32B-Instruct在FNR上表现最佳，其FN值仅为5。而Qwen2.5-72B虽然在FPR上为0.057，但其FNR达到了很高的值0.117，因此该模型的检测结果并不佳。故有以下结果：Qwen3-235B和DeepSeek-R1-0528模型虽然参数量更大，能力更加丰富，但其检测结果并无额外提升。

其原因主要在于：对于参数规模较小的模型而言，其在语义理解能力方面可能存在一定局限。例如，Qwen2.5-7B-Instruct模型在对“口令安全风险”类别的违规定义进行判断时存在理解偏差，导致误报率上升。这表明了小参数量模型在复杂语义理解任务中能力的不足，影响了其在合规检测中的准确性。

另外，当模型参数量过大时，虽然其具备更全面的知识储备和较强的推理能力，但在合规检测任务中，这种推理能力可能产生反效果。以DeepSeek-R1-70B模型为例，在部分精准匹配的匹配类别中，合规任务并不关注命中内容本身的用途，需要严格按照出现类似的就定位违规的思路来检测，而该模型会自行分析这一判断的合理性，倾向于基于自身知识进行逻辑推断，从而产生与预期合规结果相反的结果。

该结果与文献[28–30]中得出的结论一致，表明在某些任务场景下，模型参数量的增加并不必然带来性能的提升。从结果上看，不仅在Qwen系列的大模型上有如上结果，其他模型的结果也符合上述分析。例如在该实验中，Qwen系列模型72B相比于32B，检测效果下降，但235B相比于72B效果优，在FNR上与32B效果相当。

(2)大模型FPR和FNR指标分析。第2阶段的大

模型错误检测结果可以分为2类：(a)将原本违规的数据检测为合规，(b)将原本合规的数据检测为违规，2类错误的数目之和为总体检测结果中的误报数量，其中，(a)类错误表示大模型未能检出的仍未误报的数据，代表着误报率FPR，(b)类错误表示真正的违规数据被第2阶段检测为合规，属于引入大模型本身带来的错误，代表着检测的FNR值。

在实际生产环境中，通常(b)类数据检测错误情况会有更严重的安全隐患：这是因为这类检测错误一旦发生，意味着高危敏感信息存在泄露风险，后果难以预计。因此对合规系统的误报率进行优化时，要考虑到优化带来的FNR代价。本实验中，Qwen2.5-32B模型的(b)类错误占比为0.000 1，该模型既可以降低合规检测的误报率，又能降低大模型自身出错情况的发生。而Qwen2.5-72B的FNR值为0.117，在大规模的检测任务中可能引入新的安全隐患，需要结合真实的应用场景来优化。

(3)具体违规类别的FPR值方面。本节对Qwen2.5-32B大模型检测结果中的各个分类的情况进行统计，结果如图7所示。从中可得不同类别的误报情况：大模型在“口令安全风险”和“其他违规类型”上FPR值较大，均超过了整体的FPR值0.069。在“敏感信息内容”和“法律法规标准”上FPR值较小，这说明不同类型任务的复杂性和多样性不同，对应了不同的误报情况，是基于分类误报率的分层级提示优化的基础。

5.2.2 提示词方案对比测试

在提示词方案对比测试中，为了提高第2阶段大模型判断的准确度，本文对大模型的提示词进行了多阶段优化，使用多个提示词方案降低了检测的误报率。

提示词1：通用提示词。已有的研究对提示词的设计大多是简单且通用的，让大模型自行解析原始数据并作判断，并直接作出最终的判断结果。

提示词2：思维链提示词。本文在第2阶段的大模型提示词过程中引入了思维链(CoT)的思路，使用分阶段拆解的方法，在提示词中列举出违规检测的一般步骤，让大模型按照指定步骤依次判断并作隐式输出，如在提示词中直接加入“首先定位命中内容在原始内容中的位置”“接着判断命中内容在原始内容中起到的作用”“每一步骤都需要做出正确的判断并记录”等。

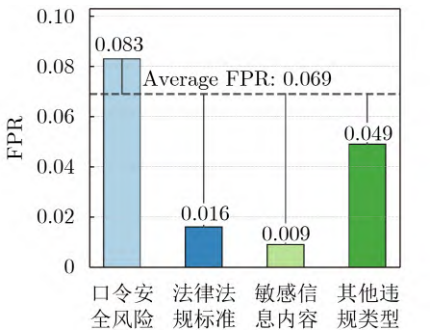


图 7 Qwen2.5-32B模型中不同违规类别的FPR以及整体检测的FPR

提示词3：少样本提示词。在该提示词中，本实验在待检测的数据集中随机加入20条检测示例[15]包含原始的检测数据以及对该数据的正确违规情况，同时加入了检测简要思路，引入的示例包含正负样本两类。

提示词4：分类提示词。该提示词为4.3.3节中基于检测精度以及类别误报率分类后的提示词，它的组成如图8所示。

表6为在Qwen2.5-32B-Instruct模型下的检测结果。由表可知，提示词1中使用通用的提示词，检测结果的FPR为0.59，该结果表明合规检测任务需要对任务的背景、检测规则作适当的描述，尤其是此类规则支持用户自己预定义的情况下，信息量不足的提示词可能无法完成检测任务。提示词2和3为仅使用思维链方法与仅使用少样本提示方法的检测结果，相较于提示词1，其FPR值更低，分别为0.12和0.06。但这2种方案下，大模型本身的FNR值较高，分别为0.3和0.13，因此综合F1分数并不高。

本文的方法提示词4达到了最佳的效果，FPR值为0.06，同时FNR值为0.000 1，综合F1分数为最高的0.94。从提示词4的结果中可以发现思维链与少样本提示在检测中的作用：引入CoT之后的提示词，能够提高大模型判断设备数据违规情况时的准确度，在输出示例中也可以看到判断过程中大模型反复确认内容和规则。例如在检测“口令安全风险”违规时，大模型会修改初始的判断结果，一步步检测口令的作用、结构，最终得出正确判断。加入少样本提示Few-Shot后，模型在一些经常出错的数据样本上修正了检测结果，其能够根据提示词中的示例或指引，以预期的方式完成检测，这种方式能够减少大模型对某些复杂违规的误判。

5.2.3 大模型系统角色提示词对比测试

本节旨在探究提示词中不同的角色提示词组成对检测结果的影响，即探究大模型的系统角色对检测任务上的影响。本实验保持其他参数设置相同，设计了3组实验，分别对应不同的角色内容：S_1为消除所有角色信息，即不指定系统的任何身份，让其直接通过用户的问题进行回复。S_2为简短的角色内容，例如“你是一个助手”这样简短并不包含任

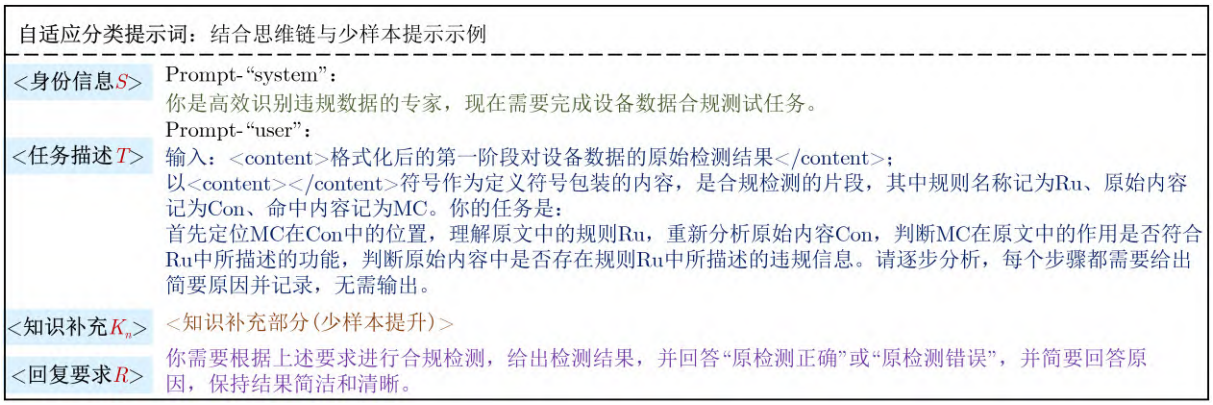


图 8 分类提示词组成结构以及示例

务信息的身份设置。S_3为描述一个任务所需信息和背景的具体身份提示词，例如“你是一个高效捕获敏感数据的专家，需要完成产品数据合规检测的任务”这样具体的系统提示词。本实验基于QWQ-32B模型进行检测，实验结果如表7所示。

其中，测试S_1, S_2和S_3为使用不同的系统提示词下的大模型的检测结果。实验结果表明在没有任何角色信息的提示词S_1中，模型的回答展现了更多样化的效果，但误报率FPR较大。提示词S_2中模型的判断原因相较于提示词S_1更加具体和精准，检测的精确率提升9%。

提示词S_3和提示词S_2在各指标中表现均较好，尤其在误报率FPR和F1分数上相近，但提示词S_3中模型误报率FPR为0.26，F1分数为0.80，且具有更小的FNR值，原因在于提示词S_3中为大模型提供了详细的检测身份，模型的回复不仅更加准确，而且从表现上看模型的结果更加可靠与严谨。

5.3 消融实验

为了探究基于分类提示词输入对大模型解决合规检测问题的有效性，在Qwen-32B-Instruct与DeepSeek-R1-32B模型上完成了消融实验。具体来说，为探究知识补充的有效性，保持“身份信息S”、“任务描述T”和“回复要求R”统一，验证知识补充的存在对大模型解决合规任务的能力影响。

实验结果如表8所示。其中“None”表示没有知识补充，“Mix”表示所有的知识补充合并为一个提示词输入大模型进行检测，“Classification”表示使用分类的知识部分提示词。从实验中得出通过加入合并后的知识补充F1指标的分数并不一定上升。在Qwen-32B-Instruct模型中引入分类提示词误报率降低6%，在DeepSeek-R1-32B模型中引入分类提示词误报率降低12%，并且两个模型的F1指标均有提升，这表明加入了具体的分类的任务描述提示词对于准确有明确的提示，并且基于具体分类的提示词设计相较于统一提示词更便于优化。

表 6 不同提示词方案的对比实验结果

提示词	方案	Acc.	TPR	FPR	FNR	F1
1	Common	0.68	0.83	0.59	0.17	0.77
2	CoT only	0.79	0.69	0.12	0.30	0.75
3	Few-Shot only	0.91	0.87	0.06	0.13	0.88
4	**本文**	**0.96**	**1.00**	**0.06**	**0.01%**	**0.94**

表 7 不同大模型系统角色提示词的影响

角色提示词	Acc.	Pre.	TPR	FPR	FNR	F1
S_1	0.79	0.60	0.94	0.27	0.06	0.74
S_2	0.82	0.69	0.95	**0.26**	0.05	0.80
S_3	**0.82**	**0.70**	**0.99**	0.28	**0.01**	**0.82**

表 8 分类提示词消融实验

模型	实验	Acc.	Pre.	TPR	FPR	FNR	F1
Qwen2.5-32B	None	0.91	0.79	0.99	0.12	0.01	0.88
	Mix	0.88	0.79	0.97	0.18	0.03	0.87
	Classification	**0.95**	**0.89**	**1.00**	**0.06**	**0.01%**	**0.94**
DS-R1-32B	None	0.70	0.61	0.97	0.39	0.03	0.75
	Mix	0.83	0.68	**0.98**	0.25	0.03	0.80
	Classification	**0.84**	**0.70**	0.97	**0.23**	**0.02**	**0.81**

通过对比提示词为“None”，“Mix”和“Classification”的3组实验可以得出基于分类的提示词设计在每一项检测任务中，利用独特的知识补充，减少了相互干扰误判的可能性，降低了提示词的冗余度，这种方式有利于大模型解决问题的准确性。

6 结束语

本文提出了一种新型面向物联网场景的大模型驱动数据合规检测方法，结合规则匹配检测方法与大语言模型语义级复核方法实现两阶段检测流程，本方法能够在面对长文本、非结构化、内容模糊等

特点的设备数据时有效提高传统规则匹配方法的误报率。

本文采集并构建了物联网设备的违规检测数据集，基于不同的基准大模型进行了实验。试验结果表明，与原始基于规则匹配的合规检测方法相比，本文基于思维链与少样本提示融合设计增强提示词，通过引入大模型进行复核使得误报率由64.3%降至6.9%，同时大模型自身引入的错误率控制在0.01%以下，具有明显的合规效果提升，大幅降低了人工复核的成本。

尽管大模型在物联网数据合规检测中展现出良好潜力，但其在实际应用中仍存在一定局限，如模型参数量过大带来的计算资源消耗较高、提示词工程对模型性能影响显著以及模型幻觉等问题仍需解决。在未来工作中，引入大模型复核的合规检测系统效率需要进一步提高，在减少误报率的同时使得大模型自身合规检测的错误率降低。其次，未来希望探索多模态数据在合规检测中的应用，拓展方法的适用范围，实现不同数据形式的复杂合规检测，充分挖掘大模型在合规检测上的能力。

参 考 文 献

[1] 陈磊. 隐私合规视角下数据安全建设的思考与实践[J]. 保密科学技术, 2020(4): 39–46.
CHEN Lei. Thoughts and practices on data security construction from a privacy compliance perspective[J]. *Secrecy Science and Technology*, 2020(4): 39–46.

[2] 王融. 《欧盟数据保护通用条例》详解[J]. 大数据, 2016, 2(4): 93–101. doi: 10.11959/j.issn.2096-0271.2016045.
WANG Rong. Deconstructing the EU general data protection regulation[J]. Big Data Research, 2016, 2(4): 93–101. doi: 10.11959/j.issn.2096-0271.2016045.

[3] 安鹏, 喻波, 江为强, 等. 面向多样性数据安全合规检测系统的设计[J]. 信息安全研究, 2024, 10(7): 658–667. doi: 10.12379/j.issn.2096-1057.2024.07.09.
AN Peng, YU Bo, JIANG Weiqiang, *et al*. Design of diversity data security compliance detection system[J]. Journal of Information Security Research, 2024, 10(7): 658–667. doi: 10.12379/j.issn.2096-1057.2024.07.09.

[4] WANG Lun, KHAN U, NEAR J, *et al*. PrivGuard: Privacy regulation compliance made easier[C]. 31st USENIX Security Symposium (USENIX Security 22), Boston, USA, 2022: 3753–3770.

[5] 李昕, 唐鹏, 张西珩, 等. 面向GDPR隐私政策合规性的智能化检测方法[J]. 网络与信息安全学报, 2023, 9(6): 127–139. doi: 10.11959/j.issn.2096-109x.2023088.
LI Xin, TANG Peng, ZHANG Xiheng, *et al*. GDPR-oriented intelligent checking method of privacy policies compliance[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 127–139. doi: 10.11959/j.issn.2096-109x.2023088.

[6] 郭群, 张华熊, 王波, 等. 基于内容和上下文的敏感个人信息实体识别方法[J]. 软件工程, 2025, 28(2): 6–9,26. doi: 10.19644/j.cnki.issn2096-1472.2025.002.002.
GUO Qun, ZHANG Huaxiong, WANG Bo, *et al*. Content and contextual sensitive personal information entity recognition method[J]. Software Engineering, 2025, 28(2): 6–9,26. doi: 10.19644/j.cnki.issn2096-1472.2025.002.002.

[7] 张西珩, 李昕, 唐鹏, 等. 基于知识图谱的隐私政策合规性检测与分析[J]. 网络与信息安全学报, 2024, 10(6): 151–163. doi: 10.11959/j.issn.2096-109x.2024087.
ZHANG Xiheng, LI Xin, TANG Peng, *et al*. Privacy policy compliance detection and analysis based on knowledge graph[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 151–163. doi: 10.11959/j.issn.2096-109x.2024087.

[8] MENG Weibin, LIU Ying, ZAITER F, *et al*. LogParse: Making log parsing adaptive through word classification[C]. 2020 29th International Conference on Computer Communications and Networks (ICCCN), Honolulu, USA, 2020: 1–9. doi: 10.1109/ICCCN49398.2020.9209681.

[9] HE Pinjia, ZHU Jieming, ZHENG Zibin, *et al*. Drain: An online log parsing approach with fixed depth tree[C]. 2017 IEEE International Conference on Web Services (ICWS), Honolulu, USA, 2017: 33–40. doi: 10.1109/ICWS.2017.13.

[10] DU Min, LI Feifei, ZHENG Guineng, *et al*. DeepLog: Anomaly detection and diagnosis from system logs through deep learning[C]. The 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, USA, 2017: 1285–1298. doi: 10.1145/3133956.3134015.

[11] MENG Weibin, LIU Ying, ZHU Yichen, *et al*. Loganomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs[C]. The 28th International Joint Conference on Artificial Intelligence, Macao China, 2019: 4739–4745.

[12] ZHANG Xu, XU Yong, LIN Qingwei, *et al*. Robust log-based anomaly detection on unstable log data[C]. The 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Tallinn Estonia, 2019: 807–817. doi: 10.1145/3338906.3338931.

[13] 尹春勇, 张杨春. 基于CNN和Bi-LSTM的无监督日志异常检测模型[J]. 计算机应用, 2023, 43(11): 3510–3516 doi: 10.19678/j.issn.1000-3428.0061750.
Yin C Y and Zhang Y C. Unsupervised log anomaly detection model based on CNN and Bi-LSTM[J]. *J. Journal of Computer Applications*, 2023, 43(11): 3510–3516 doi: 10.

19678/j.issn.1000-3428.0061750.

[14] QI Jiaxing, HUANG Shaohan, LUAN Zhongzhi, *et al*. LogGPT: Exploring ChatGPT for log-based anomaly detection[C]. 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Melbourne, Australia, 2023: 273–280. doi: 10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00045.

[15] LIU Yilun, TAO Shimin, MENG Weibin, *et al*. LogPrompt: Prompt engineering towards zero-shot and interpretable log analysis[C]. The 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, Lisbon, Portugal, 2024: 364–365. doi: 10.1145/3639478.3643108.

[16] XIANG Jinyu, ZHANG Jiayi, YU Zhaoyang, *et al*. Self-Supervised Prompt Optimization. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9017–9041, Suzhou, China. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.479.

[17] GUAN Wei, CAO Jian, QIAN Shiyou, *et al*. LogLLM: Log-based anomaly detection using large language models[EB/OL]. arXiv preprint arXiv: 2411.08561, 2024. doi: 10.48550/arXiv.2411.08561.

[18] ELHAFSI A, SINHA R, AGIA C, *et al*. Semantic anomaly detection with large language models[J]. Autonomous Robots, 2023, 47(8): 1035–1055. doi: 10.1007/s10514-023-10132-6.

[19] YANG Tiankai, NIAN Yi, LI Li, *et al*. AD-LLM: Benchmarking large language models for anomaly detection[C]. Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025: 1524–1547. doi: 10.18653/v1/2025.findings-acl.79.

[20] OLINER A and STEARLEY J. What supercomputers say: A study of five system logs[C]. 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07), Edinburgh, UK, 2007: 575–584. doi: 10.1109/DSN.2007.103.

[21] DEVLIN J, CHANG Mingwei, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019: 4171–4186. doi: 10.18653/v1/N19-1423.

[22] WEI J, WANG Xuezhi, SCHUURMANS D, *et al*. Chain-of-thought prompting elicits reasoning in large language models[C]. The 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1800.

[23] BROWN T B, MANN B, RYDER N, *et al*. Language models are few-shot learners[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 159.

[24] BAI Jinze, BAI Shuai, CHU Yunfei, *et al*. Qwen technical report[EB/OL]. arXiv preprint arXiv: 2309.16609, 2023. doi: https://doi.org/10.48550/arXiv.2412.15115

[25] Team Q. Qwq-32b: Embracing the power of reinforcement learning[EB/OL].(2025-3)

[26] GUO Daya, YANG Dejian, ZHANG Haowei, *et al*. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 645, 633–638 (2025) doi: https://doi.org/10.1038/s41586-025-09422-z.

[27] HOLTZMAN A, BUYS J, DU Li, *et al*. The curious case of neural text degeneration[J]. International Conference on Learning Representations, 2020. doi: https://doi.org/10.48550/arXiv.1904.09751.

[28] XIAO Chaojun, CAI Jie, ZHAO Weilin, *et al*. Densing law of llms[J]. Nature Machine Intelligence, 2025: 1–11. doi: https://doi.org/10.48550/arXiv.2412.04315.

[29] KAPLAN J, MCCANDLISH S, HENIGHAN T, *et al*. Scaling laws for neural language models[EB/OL]. arXiv preprint arXiv: 2001.08361, 2020. doi: https://doi.org/10.48550/arXiv.2001.08361.

[30] WU Jun, WEN Jiangtao, and HAN Yuxing. BackSlash: Rate constrained optimized training of large language models[C]. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:67852-67863, 2025. doi: https://doi.org/10.48550/arXiv.2504.16968.

李超豪：男，高级工程师，研究方向为人工智能安全、网络与信息安全.

王浩然：男，硕士生，研究方向为网络与信息安全.

周少鹏：男，高级工程师，研究方向为物联网安全、隐私计算.

闫皓楠：男，工程师，研究方向为可信人工智能、隐私计算.

张 峰：男，高级工程师，研究方向为人工智能安全、网络与信息安全.

鲁天阳：男，高级工程师，研究方向为网络与信息安全.

习 宁：男，教授，研究方向为异构网络融合安全、服务组合安全、信息流安全.

王 滨：男，研究员，研究方向为物联网安全、人工智能安全、密码学.

责任编辑：余 蓉

LLM-based Data Compliance Checking for Internet of Things Scenarios

LI Chaohao[①②③]　WANG Haoran[①]　ZHOU Shaopeng[②③]　YAN Haonan[④]
ZHANG Feng[②④]　LU Tianyang[②]　XI Ning[④]　WANG Bin[①②④]

[①](*Xidian University Hangzhou Institute of Technology*, *Hangzhou* 311231, *China*)

[②](*Zhejiang Key Laboratory of Artificial Intelligence of Things* (*AIoT*) *Network and Data Security*, *Hangzhou* 310050, *China*)

[③](*College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310058, *China*)

[④](*College of Network and Information Security*, *Xidian University*, *Xi'an* 710071, *China*)

Abstract:

Objective The implementation of regulations such as the Data Security Law of the People's Republic of China, the Personal Information Protection Law of the People's Republic of China, and the European Union General Data Protection Regulation (GDPR) has established data compliance checking as a central mechanism for regulating data processing activities, ensuring data security, and protecting the legitimate rights and interests of individuals and organizations. However, the characteristics of the Internet of Things (IoT), defined by large numbers of heterogeneous devices and the dynamic, extensive, and variable nature of transmitted data, increase the difficulty of compliance checking. Logs and traffic data generated by IoT devices are long, unstructured, and often ambiguous, which results in a high false-positive rate when traditional rule-matching methods are applied. In addition, the dynamic business environments and user-defined compliance requirements further increase the complexity of rule design, maintenance, and decision-making.

Methods A large language model-driven data compliance checking method for IoT scenarios is proposed to address the identified challenges. In the first stage, a fast regular expression matching algorithm is employed to efficiently screen potential non-compliant data based on a comprehensive rule database. This process produces structured preliminary checking results that include the original non-compliant content and the corresponding violation type. The rule database incorporates current legislation and regulations, standard requirements, enterprise norms, and customized business requirements, and it maintains flexibility and expandability. By relying on the efficiency of regular expression matching and generating structured preliminary results, this stage addresses the difficulty of reviewing large volumes of long IoT text data and enhances the accuracy of the subsequent large language model review. In the second stage, a Large Language Model (LLM) is employed to evaluate the precision of the initial detection results. For different categories of violations, the LLM adaptively selects different prompt words to perform differentiated classification detection.

Results and Discussions Data are collected from 52 IoT devices operating in a real environment, including log and traffic data (Table 2). A compliance-checking rule library for IoT devices is established in accordance with the Cybersecurity Law, the Data Security Law, other relevant regulations, and internal enterprise information-security requirements. Based on this library, the collected data undergo a first-stage rule-matching process, yielding a false-positive rate of 64.3% and identifying 55 080 potential non-compliant data points. Three aspects are examined: benchmark models, prompt schemes, and role prompts. In the benchmark model comparison, eight mainstream large language models are used to evaluate detection performance (Table 5), including Qwen2.5-32B-Instruct, DeepSeek-R1-70B, and DeepSeek-R1-0528 with different parameter configurations. After review and testing by the large language model, the initial false-positive rate is reduced to 6.9%, which demonstrates a substantial improvement in the quality of compliance checking. The model's own error rate remains below 0.01%. The prompt-engineering assessment shows that prompt design exerts a strong effect on review accuracy (Table 6). When general prompts are applied, the final false-positive rate remains high at 59%. When only chain-of-thought prompts or concise sample prompts are used, the false-positive rate is reduced to approximately 12% and 6%, respectively, and the model's own error rate decreases to about 30% and 13%. Combining these strategies further reduces the error rate of the small-sample prompt approach to 0.01%. The

effect of system-role prompt words on review accuracy is also evaluated (Table 7). Simple role prompts yield higher accuracy and F1 scores than the absence of role prompts, whereas detailed role prompts provide a clearer overall advantage than simple role prompts. Ablation experiments (Table 8) further examine the contribution of rule classification and prompt engineering to compliance checking. Knowledge supplementation is applied to reduce interference and misjudgment among rules, lower prompt redundancy, and decrease the false-alarm rate during large language model review.

Conclusions A large language model-driven data compliance checking method for IoT scenarios is presented. The method is designed to address the challenge of assessing compliance in large-scale unstructured device data. Its feasibility is verified through rationality analysis experiments, and the results indicate that false-positive rates are effectively reduced during compliance checking. The initial rule-based method yields a false-positive rate of 64.3%, which is reduced to 6.9% after review by the large language model. Additionally, the error introduced by the model itself is maintained below 0.01%.

Key words: Data compliance checking; Large Language Models (LLM); Internet of Things (IoT); Prompt engineering; Regular expression matching