航天涉密信息保密审核大模型增强方法

郑佳斌[1,2],周瀚阁[2],蒋忠林[2],陈 勇[2]
1. 浙江大学航空航天学院, 杭州 310058
2. 吉利汽车研究院(宁波)有限公司, 宁波 315311

摘 要 针对航天航空领域资料保密审查的严格要求,现有的人工筛查方法存在成本高昂、关键词匹配精度不足等问题,提出了一种结合大模型的审查方法,用于提升涉密信息的筛查效率和准确性。首先分析了航天航空领域涉密信息的特点,提出了一种基于大模型的保密审核增强架构,该架构结合了动态垂类专家System Prompt,能够从技术涉密和商业涉密等多个角度提高审查的细粒度和准确率。通过引入基于关键词的动态 System Prompt 机制,实现了大模型语义理解能力与关键词实时更新能力的有效结合。此外,为了防止大模型的过度审核,设计了一种混合式交叉微调策略,显著提高了涉密信息的召回率,达到了 96%。通过在自研的 1000 条高质量测试集上的实验,本增强框架可以将全球已发布的主流大模型在保密审核任务上的准确率提升 18%,验证了本文提出框架的有效性。

关键词 大模型;内容审核;大模型智能体;模型微调;航天涉密检查
中图分类号:TP312 **文献标识码**: A
文章编号:1006-3242(2024)02-0062-07

Enhancing Aerospace Classified Information Security through Large-scale Models

ZHENG Jiabin[1,2], ZHOU Hange[2], JIANG Zhonglin[2], CHEN Yong[2]
1. School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310058, China
2. Geely Automotive Research Institute (Ningbo) Co. Ltd, Ningbo 315311, China

Abstract *Regarding the stringent requirements of information confidentiality review in the aerospace field, current manual screening methods are suffering from high costs and insufficient accuracy of keyword matching. An enhanced review framework integrated with large language models is proposed to improve the efficiency and accuracy of confidential information screening. Initially, the characteristics of confidential information are analyzed in the aerospace sector, an architecture that enhances the auditing performance of large language models is introduced in this study, which is combined with dynamic domain-specific expert system prompts to enhance the granularity and accuracy of reviews among multiple perspectives including technical and business confidentiality. By introducing a dynamic system prompt mechanism, the framework is effec-*

资助项目:宁波市自然科学基金(2023J188)
收稿日期:2024-03-29
作者简介:**郑佳斌**(1987 -),男,博士后,副研究员,主要研究方向为人工智能大模型研究和落地应用;**周瀚阁**(1999 -),男,硕士,研究方向为人工智能大模型智能体,本文通信作者,Email:2147335502@qq.com;**蒋忠林**(1988 -),男,硕士,研究方向为人工智能大模型研究; **陈 勇**(1984 -),男,博士,研究方向为人工智能大模型研究。

tively combined the semantic understanding capabilities of large language models with the real-time updating of keywords. Additionally, in order to prevent excessive auditing by the large language model, a hybrid cross-training strategy is developed, which significantly improves the recall rate of confidential information that reaches by 96%. *Experiments on a self-developed high quality test set of* 1000 *entries demonstrates that the proposed method outperforms global open-source large language models by* 18% *in aerospace classified information inspection tasks.*

Key words *Large language model*; *Text content moderation*; *Intelligent agent*; *Fine-tuning*; *Aerospace classified information inspection*

0　引　言

航天航空领域涉及国家安全、政治稳定、经济发展和技术创新等多方面利益,该领域内文件保密审核至关重要。保密审核不仅是确保国家核心利益和战略安全不受损害的重要手段,也是促进航天航空领域持续发展的关键保障。航天航空相关单位在文件下载、邮件收发和信息对外发布等操作时需要大量的保密审核工作。

现有的保密审核方法有人工审核和关键词匹配算法审核两种方式。当待审文档数量庞大、文档内容复杂繁多时,人工审核的难度和成本将非常高。此外,人工审核经常会出现错检和漏检问题,而以关键词匹配技术为基础展开的方法往往因为关键词缺失出现漏检的情况,并且该方法过度依赖关键词,未从语义理解层面对文本进行审核,还会出现过检的情况。

文献[1]提出一种基于扩展权限组合的信息泄露检测方法。该方法首先获取敏感信息的安全规则集,然后从多个特征角度寻找危险权限组合。接着,利用危险权限组合来判断信息是否泄露,并最终输出信息泄露检测结果。该方法被证明在实际应用中存在检测精度缺陷问题。文献[2]提出一种基于敏感数据链和局部差分隐私的敏感数据泄露检测方法,通过敏感数据链判断信息是否处于超出阈值状态,再通过差分隐私保护敏感数据,但该方法存在处理时延过大等问题。

考虑到人工审核的成本和效率,基于 AI 模型和关键词匹配联动的审核方法成为有效的解决措施。例如基于逻辑回归模型检测审核敏感网页内容,在过滤网络文本内容敏感信息时,提出一种基于有穷自动机的改进算法(Swift tree DFA, ST-DFA),通过构建信息决策树实现敏感信息的识别和过滤,并可通过更新决策树来更新知识。然而,上述研究高度依赖敏感词库匹配算法,未考虑从语义理解层面对文本进行审核[3]。为解决算法计算成本和精度之间的平衡问题,文献[4]针对图文内容,引入 LSTM 模型和 TextCNN 算法作为自动化审核模块,并结合人工二次校验的机制确保审核精度,同时结合 Trie 树和 DFA 设计了一种敏感词识别过滤算法,包括敏感词树构建和检测过滤两个步骤,其时间复杂度分别降至 O(n * len)和 O(L),并实现了 100% 的查准率和 87% ~100% 的查全率,但在语义判断上仍存在一定的优化空间。

随着大规模语言模型(Large Language Models, LLMs)的出现,更强大的文本理解能力推动了自动保密安全审核能力的升级。随着这些模型在某些方面不断超越人类的知识和推理能力,它们可能产生有害内容的潜在危险已经成为日益关注的话题,推动了防止生成此类信息的努力。LLMs 在文本内容审核领域的应用虽然展现出巨大潜力,但同时也暴露了新的挑战。例如,大模型的训练数据的滞后性,导致无法保证保密信息的实时性。此外,由于保密级别和类别不同时,保密信息具备不同的表达内容和语言模式,因此大模型直接进行保密安全审核也经常出现误判的情况[5]。

为了进一步提升内容审核的精准度,文献[6]利用大模型的语义理解能力强化敏感内容的检测能力,将基于微调后的 Llama 模型作为检测器,分别对用户输入和大模型输出内容进行安全分析,并使用开源测试集 OpenAI Moderation Evaluation 以及 ToxicChat 进行测试。该文献将大模型应用在内容审核领域,并在一系列不同参数量的大模型上进行测试评估,包括 GPT-3, GPT-3.5, GPT-4 等,在高达 95 个社区场景中实现了 64% 的精准率(Accuracy)和 84% 的准确度(Precision),证明了基于大模型的检测精度超过了当前的内容审核工具。然而,实验也表明,大模型在文本安全检测任务上的性能已经

到达瓶颈,仅从参数量来优化大模型已经无法实现较大的提升。

文献[7-8]表明,大模型的过度安全问题是词汇过拟合的结果,即模型对某些与安全相关的词语和短语过于敏感,而系统提示可以引导与安全相关的模型行为,但并不能以一种全面或一致的方式来保证足够的安全性。然而,大语言模型通常在大量公开数据上训练,训练语料中往往存在保密或不当内容。因此,LLMs 在推理和输出时,可能会生成涉密或违反合规要求的内容。这一风险不仅存在于模型生成输出阶段,更体现在其保密审核决策过程中。

因此,文中首次提出了一种基于大模型的航天信息保密审核增强方法,旨在解决现有保密审核中人工成本高昂以及关键词匹配方法存在的漏检和过检问题。通过利用大模型的上下文理解能力和泛化能力,本文方法显著提升了保密判断的精度,有效地解决了现有方法面临的挑战。进一步地,为了在保持大模型原有保密审核能力的同时增强其在特定领域的保密审核效果,对关键词和模型训练文本按照保密等级和信息类别进行标记,并针对不同的保密等级和信息类别提出了一种动态交叉 Prompt 的微调策略。这种策略有效平衡了大模型针对不同级别和类型保密信息的审核能力,避免了保密过检的问题。此外,为了证实本文方法在保密审核任务上的有效性,文中构建了一个包含 1000 条高质量、带密级和信息类别标注的保密审核测试集,并通过在此测试集上的评估,证明了文中基于大模型的保密审核增强方法的有效性。

1 保密审核大模型构建方法

文中面向航天文件信息保密审核领域,结合大语言模型技术,提出一种大模型保密审核能力增强框架。

1.1 保密审核大模型增强框架

保密审核大模型增强框架的总体架构图如图 1 所示,共分为 3 个阶段:1)文档预处理;2)关键词检测分类;3)LLM 保密评审。

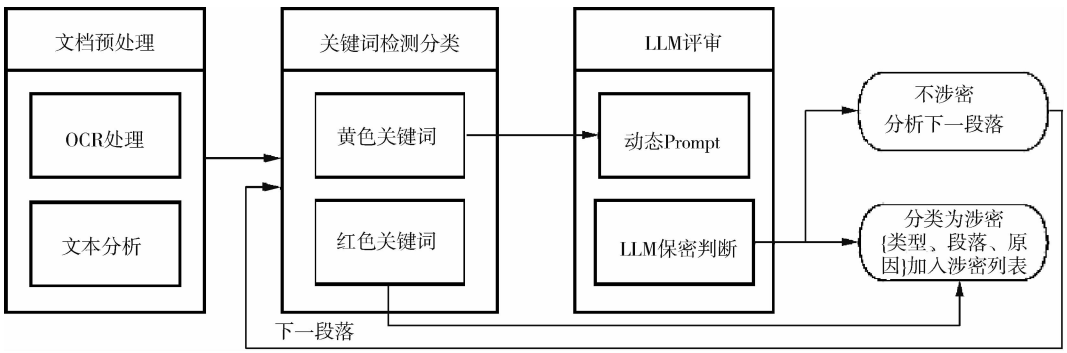


图 1 总体框架

其中,第一阶段中文档预处理需将文件下载、邮件收发和信息对外发布等操作过程中所有文档进行预处理,根据文档格式、内容特点的不同,通过 OCR 和 PDF 文本解析相结合的方法提取文档文本内容,并将文本按段落进行分割后存入一个列表中。

第二阶段为关键词检测分类,本文构建了涉密关键词词库,词库支持正则表达式匹配。由于企业保密业务需求,如命中部分核心关键词,则直接判定为涉密,不进行 LLM 评审,如“涉密人员”等关键词,因此将关键词库分为红色和黄色关键词库,红色关键词为核心关键词,该类关键词数量共有 55 个,命中该库则直接判定为涉密,并按{“红色”、段落内容、“命中核心关键词 KeyWord”}格式记录到文本检查涉密列表中,并进行下一段落的审核。

黄色关键词库为潜在涉密关键词,该库根据不同保密等级和保密信息划分为 10 个类别,表 1 为该 10 个类别的具体明细和各类别的关键词数量。

表 1 文中黄色关键词库明细

黄色关键词类型	数量
国家绝密	67
国家机密	88
国家秘密	77
商业绝密	52
商业机密	46
商业秘密	57
技术绝密	79
技术机密	92
技术秘密	126
个人隐私	55

如果命中黄色关键词,则会将{黄色关键词类型,命中关键词列表}格式发送给 LLM 评审模块,如{“技术绝密”,“打击精度”,“攻击射程”}。

第三阶段为 LLM 评审,文中通过关键词检测结果动态变更 System Prompt,Prompt 遵循思维链(Chain of Thought, CoT)的方法进行设计。具体地,当关键词没有命中时,会采用通用保密审核 System Prompt,该指令首先会定义模型角色为保密安全审核员,在执行日常工作任务时,辅助确保审核内容符合航天保密和安全规范。此外,还会引导大模型以 CoT 方式一步一步思考,从而可以优化过度保密审核问题,具体 CoT 步骤如下:

1) 内容审查:在阅读和编辑文件时,警觉可能涉及敏感信息的内容;

2) 保密性核查:检查是否所有信息都仅供授权人员访问,确保没有超出必要的信息共享范围;

3) 风险评估:留意任何不寻常的文件访问或传输行为,这可能表明了不当的信息处理;

4) 及时反馈:如遇到任何潜在的保密性问题或安全隐患,应立即上报并提出处理措施的建议。

当命中黄色关键词时,根据黄色关键词类型和命中关键词列表动态选择并改写专用类别保密审核 System Prompt,从而更精准的对不同类型文件进行保密检查。以技术机密类型审查为例,该类型的 System Prompt 首先会将模型角色定义为保密审核专家,任务是审核航天航空领域的技术文件和通讯,以确保所有敏感信息都符合安全保密规定。其 CoT 步骤如下:

1) 敏感词识别:检查文件中是否含有预先定义的敏感词汇或术语;

2) 内容上下文分析:在确定了敏感词后,分析上下文以评估信息的安全级别;

3) 风险评估:根据内容的敏感程度和关联性,判断其泄露风险;

4) 行动指南:对于检测到的潜在风险信息,给出是否需要进一步审查或采取保密措施的建议。

1.2　保密审核大模型微调方法

为提高航天领域保密审核大模型的通用审核能力和专用类别保密审核能力,本小节提出了动态交叉 Prompt 微调方法。整体微调过程包含:航天相关非保密类与保密类训练数据收集、数据预处理、微调数据类别细分以及通过 System Prompt 设计动态交叉微调。

数据收集方面,通过人工编写、AI 生成以及保密文档改写等方式构建训练数据集,总体可分为两大类,即非保密类与保密类数据。保密类数据根据不同保密类别设计专用保密审核 System Prompt。此外,为审核非保密类和关键词匹配漏检的数据,设计通用保密审核 System Prompt。

配对完 System Prompt 的数据将进一步经过大模型处理,处理过程包括检查数据是否分类无误,过滤掉低质量数据,以及重新优化和撰写保密段落信息。处理后的数据格式为"{System Prompt + input: label}",为尽可能减小微调时模型学习受数据分布偏差带来的影响,将对数据进行细粒度处理和随机化打乱,本文对保密类数据按照表 1 中的类型进行标注,同时针对该表中不同的关键词类型设计不同 System Prompt,针对“国家”、“商业”、“技术”和“隐私”等不同类型设计不同的 System Prompt 模板,然后根据各个类型中的“绝密”、“机密”和“秘密”等级,修改 prompt 中的保密审核要求,针对不同类型构造数据,保证各个类别数据量级的均衡。

在进行大模型微调时,为保障航天领域保密和保密过检之间的平衡问题,提出动态交叉 Prompt 微调策略,即使在关键词审核和大模型判断错误时,也能通过微调阶段的缓解机制来修复该错误[9]。具体的动态交叉微调策略如图 2 所示,其中分别使用专用和通用的 System Prompt 模板,对应不同的保密类型数据。相比于 System Prompt 模板和相关类别保密审核训练数据一一对应的方式,文中通过抽样混合不同类别的数据进行交叉微调训练,训练数据中{通用保密审核 System Prompt + 非涉密数据}组合的数据占比为 80%,{通用保密审核 System Prompt + 专用保密领域涉密数据}组合的数据占比为 20%。通过上述动态交叉 Prompt 微调策略,可以保证即便普通任务输入被误判为涉密任务时,配对上特定专用保密审核 System Prompt,也具备一定的纠正能力。

表 2 展示了文中的保密审核大模型微调数据量级,分为两大类:专用涉密类和非涉密类。对于非涉密类数据,本文抽取航天领域相关官网公开发布的新闻稿中的段落,共计抽取 10000 个段落样本。针对专用涉密类数据,进一步对专用涉密数据按照黄色关键词类别进行细粒度划分,共 10 类。对国家、商业、技术和个人隐私不同类别的信息特征和不同密级要求,设计特定提示词,并使用 ChatGPT 分

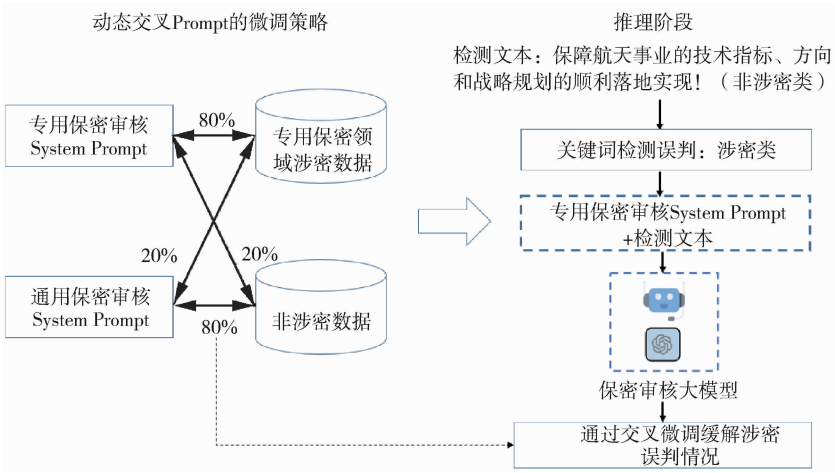


图 2 混合式微调策略

别模拟生成相关涉密段落，最后通过人工筛选，完成每类 1000 条专用涉密微调数据，共计 10000 个涉密微调样本，与非涉密样本总数据量相当。

表 2 保密大模型微调数据明细

黄色关键词类型	数量
无密普通段落	10000
国家绝密段落	1000
国家机密段落	1000
国家秘密段落	1000
商业绝密段落	1000
商业机密段落	1000
商业秘密段落	1000
技术绝密段落	1000
技术机密段落	1000
技术秘密段落	1000
个人隐私段落	1000

2 实验与结果分析

本章主要通过设计对比实验，验证文中提出的保密审核大模型增强框架能够在原生大模型基础上提升保密审核性能。为评估本文方法在测试集上的涉密判断精度，使用混淆矩阵的准确率、精准率、召回率、F1 分数和误杀率 5 个指标进行计算评估，如式(1)～(5)所示

$$P_{acc}=(N_{TP}+N_{TN})/(N_{TP}+N_{FP}+N_{TN}+N_{FN}) \quad (1)$$

$$P_{pre}=N_{TP}/(N_{TP}+N_{FP}) \quad (2)$$

$$P_{recall}=N_{TP}/(N_{TP}+N_{FN}) \quad (3)$$

$$P_{F1}=2/(1/P_{pre}+1/P_{recall}) \quad (4)$$

$$P_{FKR}=N_{FP}/(N_{FP}+N_{TN}) \quad (5)$$

式中：P_{acc}表示准确率(Accuracy Rate)，是分类模型中最直观的性能指标，它表示模型正确预测的样本数占总样本数的比例；P_{pre}为精准率(Precision Rate)，表示在所有被模型预测为正类的样本中，真正属于正类的样本所占的比例，精准率关注的是模型预测为正类的准确性；P_{recall}为召回率(Recall Rate)，表示在所有实际为正类的样本中，被模型正确预测为正类的样本所占的比例，召回率关注的是模型捕捉正类样本的能力；P_{F1}表示 F1 Score，是精准率和召回率的调和平均值，也是一个综合考虑精准率和召回率的性能指标；P_{FKR}表示误杀率(False Kill Rate)，是模型错误地将负类样本预测为正类的情况；N_{TP}表示实际为正例且预测为正例的数量；N_{TN}表示实际为正例且预测为负例的数量；N_{FP}表示实际为负例且预测为正例的数量；N_{FN}表示实际为负例且预测为负例的数量。

基于文中自建保密审核测试集进行对比实验，分别验证传统关键词匹配方法、通用开源大模型保密审核方式以及文中所构建的保密审核大模型增强框架三者的性能。

自建数据集共包含 1000 条高质量样本。其中 500 条来源于航天领域相关官网上公开发布的新闻稿中随机抽取的文本段落，还包括 500 条通过随机选取黄色关键词库中的关键词，并使用 ChatGPT 分别模拟生成航天相关涉密文本段落，并且在生成过程中尽量避免使用对应关键词，每个黄色关键词类型都生成了 50 条模拟涉密样本。

表 3 分别展示了传统关键词匹配方法、通用开

源大模型保密审核方式以及文中所构建的保密审核大模型增强框架在 1000 条自研构建的测试集上的测试结果,表中共包含准确率、精准度、召回率以及 F1 分数 4 个维度。用于量化不同测试方法的保密审核性能。

表 3　自研测试集性能评估

测试对象	精准率/%	准确率/%	召回率/%	F1 分数/%
关键词匹配	75.1	77.34	71	74.04
通用大模型	81.9	81.58	82.4	81.99
本文增强框架	96	96.37	95.6	95.98

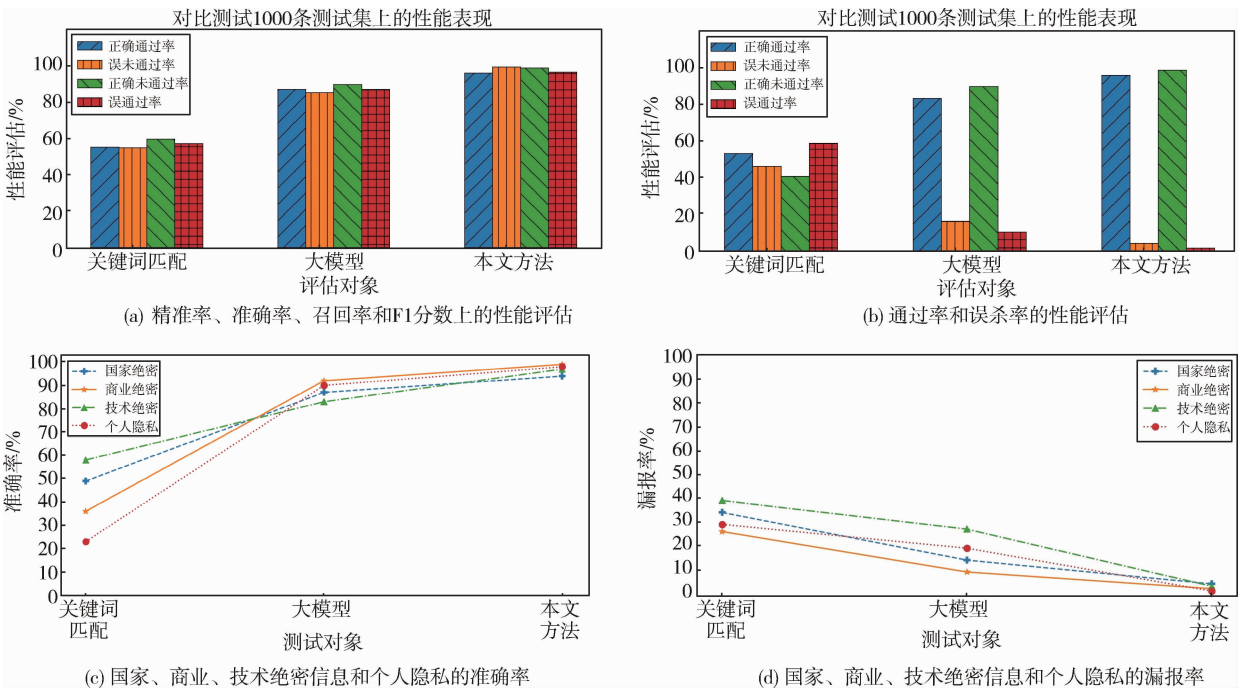
根据表 3 的结果,可得结论如下:

通过混淆敏感词构造的测试集显然具备很强的欺骗能力。基于关键词检测方法的准确率和 F1 分数分别仅为 75.1% 以及 74.04%,即在关键涉密领域的误判率近 30%,这表明现有基于关键词匹配或者黑白名单方法进行涉密检测的不足之处。

通过对比测试,可以看出大模型具备语义理解能力,以及比关键词检测方法更高的精度,同时证明了保密审核增强架构的有效性。根据测试结果,经过本方法增强的大模型保密检测准确率提升了 18%,通用大模型比关键词检测提升 9%;在 F1 分数上,增强框架提升 17%,通用模型相比于关键词检测提升 10.8%,证明了本文提出的框架能够有效增强大模型的保密审核能力,同时相比传统的方法具备更高的检测准确率。

图 3 进一步分析和对比了厂商接口、大模型,和本文增强架构之间的性能差异,以及国家、商业、技术和个人隐私 4 个类别的准确率和漏报率。

其中,图 3(a)对比了三者在 1000 条测试集上的性能差异,在准确率、精准度、召回率和 F1 分数上,均展现出本文增强架构 > 大模型 > 关键词匹配的趋势结果。图 3(b)则是从正确通过率、正确未通过率、误通过率以及误未通过率四个维度,对比了三者在 1000 条总测试集上的性能差异。本文增强架构的误通过率和误未通过率分别为 1.5% 和 4%,远超大模型和关键词匹配方法。图 3(c)和图 3(d)从细粒度展示了三者分别在国家绝密、商业绝密、技术绝密及个人隐私上的涉密风险内容审核精度。相较于普通的大模型,本文提出的增强框架分别在漏报率和误报率上实现了平均 25% 和 17% 的下降。其中,在国家绝密类别上,分别在漏报率和误报率上下降了 31% 和 22%。说明该方法在构建保密审核大模型增强框架的涉密判断能力的同时,避免了过度安全的保密审核问题。



(a) 精准率、准确率、召回率和F1分数上的性能评估

(b) 通过率和误杀率的性能评估

(c) 国家、商业、技术绝密信息和个人隐私的准确率

(d) 国家、商业、技术绝密信息和个人隐私的漏报率

图 3　多维度对比关键词匹配、大模型和本文增强架构性能

综上,大模型相比关键词匹配,具备更强的语义理解能力,能够结合语境判断内容是否涉密。同时,通过细分涉密类别的动态 Prompt 方法构建的大模型增强框架进一步强化了大模型的涉密检测能力。

3 结 论

针对航天航空领域中对资料保密审查的高标准要求,提出了一种结合大模型的保密审核方法,旨在提高涉密信息审查的效率和准确性。通过深入分析航天航空领域涉密信息的特性,本文设计了一种大模型审核性能增强架构,该架构融合了动态垂类专家 System Prompt,能够综合考虑技术涉密和商业涉密等多维度信息,提升了审查过程的细粒度和准确率。

通过在自建的 1000 条高质量测试集上进行的对比实验表明,该增强框架可以将全球已发布的大模型在保密审核任务上提升 18% 的准确率。总的来说,该增强框架为航天航空领域中的涉密信息审查提供了一种高效、精确的技术方案,对于提升行业内信息保密工作的质量和效率具有实际意义。

参 考 文 献

[1] 高尚清. 智能审核分析系统设计与应用[J]. 电视技术, 2023, 47(6): 180-184. (GAO Shangqing. Design and application of intelligent review analysis system[J]. Television Technology Journal,2023,47(6):180-184.)

[2] 杨浩雷. 信息流图文内容审核平台的设计与实现[D]. 北京:北京交通大学,2021:003713. (YANG Haolei. Design and implementation of a graphic and text content review platform[D]. Beijing: Beijing Jiaotong University,2021:003713.)

[3] 赵月. 大语言模型安全现状与挑战[J]. 计算机科学, 2023,14(13):68-71. (ZHAO Yue. Current status and challenges of large language model security[J]. Journal of Computer Science and Technology, 2023, 14(13): 68-71.)

[4] 张燕. 数字化时代内容风控领域的实践探索[J]. 新闻研究导刊,2023,14(13):80-82. (ZHANG Yan. The exploration of content risk management in the digital age [J]. Journal News Research,2023,14(13):80-82.)

[5] WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: how does LLM safety training fail? [J]. Advances in Neural Information Processing Systems,2024,36.

[6] 张伟, 陈凤龙, 李强. 基于 CEEMD 特征提取和优化 RF 分类的 Vienna 整流器故障诊断[J]. 东北电力大学学报,2023,43(6):23-31. (ZHANG Wei, CHEN Fenglong, LI Qiang. Vienna rectifier fault diagnosis based on CEEMD feature extraction and optimized RF classification[J]. Journal of Northeast Electric Power University,2023,43(6):23-31.)

[7] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[C]. 30th USENIX Security Symposium (USENIX Security 21). 2021: 2633-2650.

[8] JI J, LIU M, DAI J, et al. Beavertails: towards improved safety alignment of LLM via human-preference dataset[J]. Advances in Neural Information Processing Systems, 2024, 36.

[9] ZHU W, GONG H, BANSAL R, et al. Self-supervised euphemism detection and identification for content moderation[C]// 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021: 229-246.