XXXX

基于大模型的电信申诉判责全流程智能化方法研究

张梦婷[1]，张晓航[1]，李征仁[1]，王海燕[2]，陈中华[3]
（1. 北京邮电大学 北京100876；
2. 中国信息通信研究院 北京 100876；
3. 中国移动通信集团内蒙古有限公司 内蒙古 呼和浩特 010090）

摘　要：电信申诉处理涉及多个环节，传统人工模式在效率与一致性上存在瓶颈，现有自动化研究多局限于单点任务。基于此提出一种基于大语言模型的智能化判责框架，通过提示词工程依次构建申诉无效判断、类型识别、内容拆解、人机协同证据链收集与判责报告生成五个模块。在某省级运营商约7000条真实工单上的实验表明：无效判断准确率达83.2%，较BERT等基线方法最低提升14.2%；服务顽疾分类准确率为97%；申诉内容拆解F1值为73.9%，要点覆盖率为82.1%；各环节处理效率较人工提升最高达96.7%。跨月份泛化测试显示模型性能稳定，为电信申诉判责提供了完整的智能化解决方案，对行业智能化实践具有参考价值。

关键词：电信申诉判责；大语言模型；提示词工程；人机协同
中图分类号：TP18；F626
文献标志码：A
doi: 10.11959/j.issn.1000−0801.

A Large-Language-Model–Driven End-to-End Intelligent Framework for Responsibility Determination in Telecom Complaints

ZHANG Mengting[1], ZHANG Xiaohang[1], LI Zhengren[1], WANG Haiyan[2], CHEN Zhonghua[3]
1. BUPT,Beijing 100876, China
2. China Academy of Information and Communications Technology,Beijing 100876,China
3. China Mobile Communications Corporation Inner Mongolia Co.,Ltd,Hohhot 010090,China

Abstract: Telecom complaint adjudication involves multiple stages. Traditional manual processing faces clear bottlenecks in efficiency and consistency, while existing automation studies mostly focus on isolated tasks. To address this issue, this paper proposes an intelligent adjudication framework based on large language models. Through prompt engineering, the framework comprises five modules: invalid complaint detection, complaint type identification, complaint content decomposition, human-in-the-loop evidence chain collection, and adjudication report generation. Ex-

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX
通信作者：张晓航，zhangxiaohang@bupt.edu.cn
基金项目：国家自然科学基金项目（No.72271034）;国家重点研发计划雄安专项（No.2023XAGG009304）
Foundation Items: The National Natural Science Foundation of China(No.72271034);The National Key R&D Program Xiongan Special Project(No.2023XAGG009304)

periments on approximately 7,000 real complaint cases from a provincial telecom operator show that the proposed method achieves 83.2% accuracy in invalid complaint detection, outperforming BERT and other baselines by at least 14.2%; 97% accuracy in service issue classification; and an F1 score of 73.9% with 82.1% key-point coverage in complaint content decomposition. Processing efficiency across stages improves by up to 96.7% compared with manual handling. Cross-month generalization tests further demonstrate stable performance. The proposed framework provides a complete intelligent solution for telecom complaint adjudication and offers practical value for digital transformation in the telecom industry.

Key words: Telecom Complaint, Large Language Models, Prompt Engineering, Human-Machine Collaboration

1 引言

随着信息通信业务快速发展，电信运营商面临持续增长的申诉处理压力。根据《工业和信息化部关于2025年第一季度电信服务质量的通告》显示，2025年第一季度，全国电信用户申诉中，涉及服务争议的申诉占比41.2%，涉及资费争议的申诉占比39.1%，涉及营销的申诉占比11.4%。申诉处理的及时性和准确性已成为影响用户满意度和运营商服务质量的关键因素。电信申诉判责流程具有高度复杂性，涉及有效性审核与申诉受理、申诉类型识别与分类、申诉内容要素提取与分析、根因定位与责任界定、相关证据收集与整理、判责报告撰写与质检等多个关键环节。该流程对业务知识理解、跨系统信息整合以及逻辑推理能力均提出了较高要求，传统以人工为主的处理模式在效率与一致性方面面临明显挑战。

现有研究主要聚焦投诉分类、信息抽取、报告生成等单点任务[1-4]，虽在局部环节取得一定成效，但仍存在三方面不足：其一，缺乏覆盖申诉判责全流程的统一框架，环节之间衔接不足；其二，证据链构建与收集研究相对薄弱，难以直接支撑真实业务判责；其三，传统方法在复杂语义理解、多诉求拆解和长文本推理方面能力有限，难以适应真实工单中常见的口语化、模糊化和跨句表达。相比之下，大语言模型在复杂语义理解和多步推理方面展现出更强潜力，若与流程自动化工具结合，有望为电信申诉判责提供端到端的智能化支撑。

近年来，以Transformer为基础的大语言模型在复杂语义理解与多步推理方面展现出的突出能力，为解决上述问题提供了新的技术路径；Agent协作、工作流编排等也为垂直领域业务流程的智能化提供了新的解决方案。基于此，本文提出一种基于大语言模型的电信申诉判责全流程方法，覆盖申诉无效判断、服务顽疾识别、申诉内容拆解、证据链生成与收集以及判责报告自动生成五个环节，并通过人机协同机制提升证据收集效率与业务可落地性。

本文主要贡献如下：

（1）提出面向电信申诉判责的智能化处理框架，将大模型能力系统性融入无效判断、类型识别、内容拆解、证据收集与报告生成五个环节，形成从工单输入到报告输出的完整闭环。（2）设计了人机协同的证据链生成与收集机制，在保留人工审核与审计追溯能力的同时，降低跨系统重复检索成本。（3）基于某省级运营商约7000条真实工单开展实验，结果表明：无效判断准确率达83.2%、Kappa为0.650，服务顽疾分类准确率达97%，申诉拆解F1为73.9%，各环节效率最高提升96.7%。

2 相关研究

2.1 电信投诉处理技术的演进

电信行业的申诉处理研究大多围绕投诉文本展开，其技术路径经历了从规则驱动到机器学习

再到深度学习与预训练语言模型的演进过程。早期的研究主要采用基于规则的方法，通过预定义的关键词匹配和业务逻辑来实现申诉分类和初步处理，能够在一定程度上完成投诉分类，但难以应对复杂场景和多样化诉求；随着自然语言处理技术的发展，支持向量机、决策树及FastText等机器学习方法被引入以提升分类自动化水平[3-6]。近年来，基于BERT[7]等预训练模型的深度学习方法在投诉文本分类与要素抽取上表现优异，诸多关于电信投诉处理技术的研究涌现，例如利用融合模型优化客服系统流程[8]，或使用改进Transformer自动生成重投报告[2]。然而，此类研究普遍聚焦单点任务[1]，且依赖大规模标注数据，缺乏涵盖证据链构建与责任界定的系统性全流程方案。

2.2　大模型与智能体技术应用

大语言模型（LLM）以Transformer为核心架构，依托大规模语料预训练与微调，展现出强大的自然语言理解与生成能力。研究表明，模型在达到一定参数规模后，会突然表现出小模型不具备的能力，这为电信投诉等复杂任务提供了新的可能性[9]，已有研究显示大模型应用在电信投诉领域的可行性[10, 11]。

在理论层面，LLM通过自注意力机制与预训练框架掌握语言知识，再结合人类反馈强化学习（RLHF）优化特定任务,从而具备强泛化与迁移能力。此外，模型具备零样本和少样本条件下的推理与生成能力，能够通过提示词工程（Prompt Engineering）快速适配垂直领域，尤其适用于通信行业中涉及非结构化文本解析、多任务协同处理与跨领域知识融合的复杂投诉处理场景[12]。

关于大模型的提示词工程（Prompt Engineering），已有大量研究论证其在垂直领域应用中的实用性与工程价值。提示词工程通过设计输入提示的结构与上下文，引导大模型在无需参数更新的情况下完成特定任务，具备成本低、部署灵活、可解释性强等优势[13, 14]。在垂直领域应用中，领域感知提示工程通过注入行业术语与业务规则，可显著提升模型输出质量[15, 16]。针对电信投诉处理，已有研究初步探索了提示词在客服问答与文本解析中的作用[17, 18]。

此外，智能体（Agent）技术结合大模型与外部工具（如API、数据库），能实现复杂任务的链式处理。在电信投诉领域，这类技术能够辅助人工快速定位证据、关联历史案例和制度条款，大幅减少人工检索与比对的工作量。近来，已有研究提出一种面向业务端到端的投诉智能体解决方案，通过整合大语言模型与小模型能力，构建了涵盖信息提取、定界定位、智能问答交互和投诉质检的智能化体系[12]。这一实践为本研究提供了重要借鉴。

2.3　研究现状分析

纵观上述研究，现有研究在电信投诉自动化处理方面虽已取得显著进展，但主要集中在单点环节优化，如投诉分类、情感分析、要素抽取或报告生成等，缺乏对整个判责流程的系统性研究。特别是在证据链构建、责任界定等核心环节，相关研究明显不足，导致技术成果难以形成完整的解决方案。虽然已有研究探索了大语言模型在客户服务中的应用，但大多停留在简单的分类或生成任务层面，未能充分利用大模型的推理、规划和知识整合能力。同时，智能体技术在电信申诉领域的应用仍处于初步探索阶段，缺乏成熟的系统设计和实践验证。

3　电信申诉判责全流程智能化处理框架

3.1　方法框架

本文提出一种面向电信申诉判责的大语言模型驱动全流程处理方法，以申诉处理业务流程为主线，将复杂任务拆解为申诉无效判断、申诉类型识别、申诉内容拆解、证据链生成与收集、判责报告自动生成五个模块，并通过结构化文件实

现模块间的数据传递。模型层面，本文采用Moonshot AI的Kimi（moonshot-v1-32k）作为核心语言引擎，结合少样本提示词工程与思维链引导实现任务适配。

考虑到电信业务系统在安全合规与审计可追溯性方面的要求，本文采用模块化解耦的流程设计思路。各处理环节在逻辑上独立运行，并通过结构化文件作为统一的数据接口完成信息传递。该设计在避免跨系统直接调用的同时，为关键判责环节保留了人工核验与追溯空间，增强了方法在真实业务环境中的可落地性。总体流程见图1。在接收工单数据后，首先对申诉有效性进行筛查，将不符合受理条件的工单过滤，以减少后续处理的噪声输入。有效工单进入类型识别环节，根据业务分类码和文本语义，判断其所属的服务顽疾类别，为后续拆解提供结构化上下文。随后，通过申诉内容拆解将长文本分解为若干原子要点，每个要点对应一个可独立回应的投诉事实，作为后续证据收集和报告撰写的基本单元。证据链收集环节采用人机协同机制，由大模型生成初始证据参数清单，经人工审核修正后交由自动化工具执行查询与下载，兼顾了效率与可靠性。最终，各环节处理结果汇聚至报告生成模块，由大模型结合证据内容和申诉分析生成标准化判责报告。

3.2 申诉无效判断

无效申诉是指缺乏明确诉求或关键信息缺失、有误的工单数据。在真实业务场景中，为规范电信用户申诉处理工作，保障申诉受理的公平性和效率，有《电信用户申诉处理办法》（工信部第35号令）及相关规定，对于申诉不符合受理要求的可不予受理。具体而言，本研究中所用到的无效筛选规则主要有：用户申诉号码是否正确、申诉事由是否清晰、涉及信号问题所填通讯地址是否详细、是否之前已申请调解、姓名是否完整清晰、申诉内容中涉及费用等关键事项时间是否已超过可以受理的时限。此步骤目的在于确保后续处理仅在有效申诉范围内进行，减少无效数据带来的噪声。

依据上述无效判断规则，本研究借助大模型强大的语言理解能力，通过少样本的提示词工程，利用运营商3-6月每日工单处理数据进行提示词优化及结果验证，创建了无效判断处理模型，专门用于判定申诉工单数据的筛选任务。无效判断采用混合策略（图2）：对号码校验、姓名完整性等简单规则直接判定；对事由不清、超时限等语义场景调用大模型判别。输出为“有效/无效”及无效原因。

3.3 申诉类型识别

对通过有效性筛查的工单，本文进一步识别其所属服务顽疾类别，以提升后续分析的结构化程度。结合业务场景，本文重点面向互联网营销不规范、外呼营销不规范、业务不知情定制、限制业务退订、限制套餐变更、超套费用质疑、反

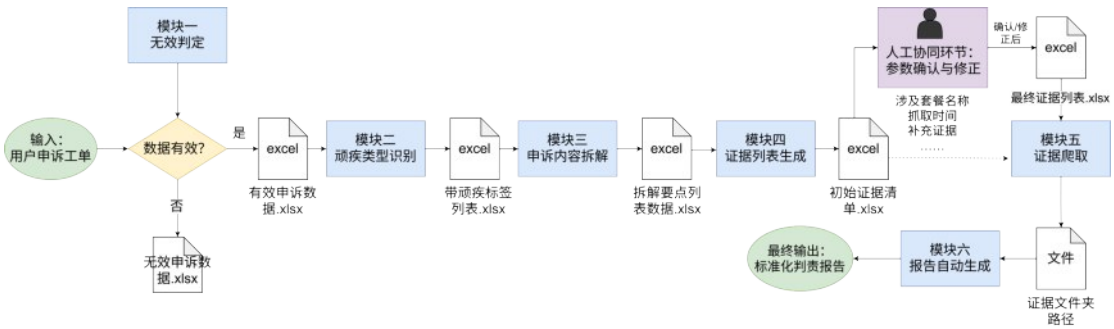

图1 申诉全流程系统流程图

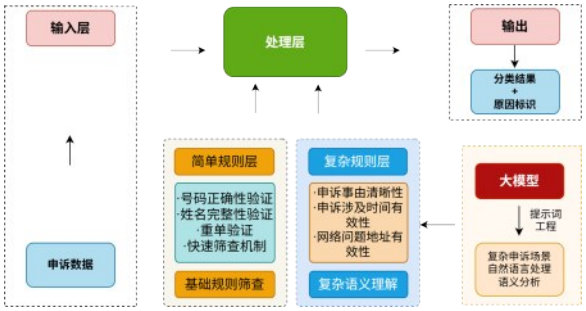

图2　无效判断处理逻辑框架图

诈号码关停等七类问题开展分类，见表1。

本文基于上述规则创建“顽疾类型”分类器，具体方法为：将基础分类码与申诉文本共同输入模型，由规则匹配先覆盖边界清晰的类别，再由大模型完成语义补充判断。该模块输出顽疾类型标签，并作为后续申诉拆解的增强上下文。逻辑框架见图3。

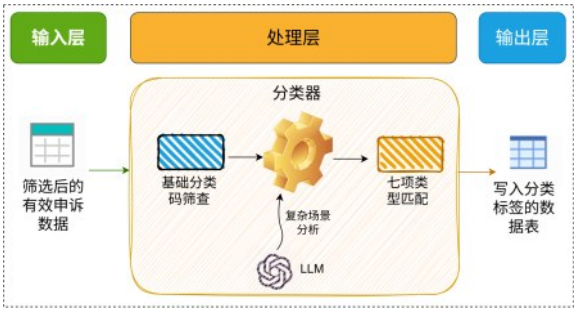

图3　服务顽疾类型判断处理逻辑框架图

3.4　申诉内容拆解

电信申诉文本往往冗长，在单一申诉中通常包含多个相互独立且需要分别处理的诉求点或事实陈述，在实际生产环境中，复杂申诉内容的拆解往往耗时耗力，且不同员工因为主观性判断、业务熟练程度等的差异，对同一条数据的拆解可能有所不同，进而导致根因定位不准，影响后续判责。基于此，本文将申诉理解与根因定位任务形式化为申诉内容分点拆解（Complaint Decomposition into Points, CDP），目标是将原始长文本拆解为若干原子要点（Atomic Points, AP），定位用户核心投诉内容、服务争议点，从而为后续的证据收集、责任分析和报告撰写提供精确的信息单元，实现申诉内容的逐点精准回应。其中，每个原子要点需满足：1）原子性：仅表达一件可单独回应的事项；2）可定位：能在原文中找到对应的文本片段；3）要素化：包含关键信息（如时间、费用、业务等）；4）可回应性：要点内容应能被报告模板直接引用或扩展。

基于此，本环节使用大模型作为底座，进行少样本的提示词工程，构建申诉拆解处理模型，采用"结构化分点生成"的输出策略，将复杂申诉文本转换为标准化的要点集合。每个输出的每个要点包括文本片段、申诉问题描述（要点内容）、关键信息（时间、号码、业务名称等），拆解示例见表2。

3.5　证据链生成与收集

证证据链是申诉判责的核心依据，但真实业务中通常需要跨多个业务系统核查账单、录音、订购记录等证据，人工收集成本较高。考虑到完全自动化难以适应复杂业务规则与场景变化，本文提出一种人机协同的证据链生成与收集机制。

表1　七项服务顽疾类型描述

顽疾类型	描述
互联网营销不规范	运营商在互联网渠道产品的相关营销不规范
外呼营销不规范	运营商通过电话渠道向用户推荐套餐等业务遭到用户投诉
业务不知情定制	未经用户同意开通增值业务等
限制业务退订	拒绝或阻挠用户取消/退订相关业务
限制套餐变更	拒绝或阻挠用户变更相关套餐
超套费用质疑	用户质疑超出套餐费用的收取并申诉
反诈号码关停	用户号码或业务被关停

表2　申诉内容拆解示例

申诉内容	文本片段	要点AP	要点内容	关键信息
前段时间看见手机APP套餐公示，说套餐不含语音的可以加对应的语音包，语音特惠包，现在不认了，就改公式了，我有截图，证明没有语言资源的是可以办理的，结果他们又不认了	看见手机APP套餐公示，说套餐不含语音的可以加对应的语音包，语音特惠包，现在不认了	1	套餐办理与承诺不符：--用户根据手机APP套餐公示，认为不含语音的套餐可以加对应的语音特惠包，但运营商不认可。	套餐=语音特惠包
	我有截图，证明没有语言资源的是可以办理的	2	官方回复与公示不一致：截图…	套餐=语音特惠包

具体而言，模型首先基于申诉拆解结果、顽疾类型和历史案例知识，生成初始证据清单，内容包括证据名称、涉及时间、涉及系统及关键查询参数。随后由人工对证据项的完整性与正确性进行审核，并对时间范围、业务名称、特殊处理事项等进行必要修正；审核通过后，再由基于Selenium WebDriver的RPA工具执行跨系统检索与文件下载。该设计在提升证据收集效率的同时，保留了业务合规所需的人工复核节点。

需要说明的是，本文的证据抓取工具本质上是规则驱动的确定性流程，而非具备自主规划能力的通用Agent。其优势在于工程实现稳定、适于内网环境部署；局限在于对系统界面变更较敏感，维护成本较高。

该环节整体流程见图4。这一过程实现了证据链条的半自动化构建，旨在降低人工逐一登录各业务系统的时间消耗的同时保证自动收集的证据的可靠性与有效性，提升处理效率与准确率。

3.6　判责报告自动生成

在前序模块完成后，处理最终判责结果需形成标准化报告，以便归档与上报。传统模式下电信员工需要对每一份数据进行手动记录和分析，报告内容需要涵盖用户信息、投诉内容、查证情况及处理后情况等。然而，手动生成这些报告不仅耗时耗力，而且主观判断和标准不一致容易导致内容缺乏连贯性和一致性[2]。

基于此，本文借助大模型强大的自然语言处理能力，结合已有成熟报告案例，使用少样本的提示词工程，搭建出稳定的报告内容生成模型。再采用“模板化 + 大模型输出润色”的自动化报告生成方法：首先读取申诉数据、拆解后的要素以及证据文件内容，然后使用微调后的大模型生成对应部分的报告内容；再根据报告模板构建初版报告框架，包括“申诉基本情况、证据支撑、逻辑分析、处理意见”等，将模型生成内容填充进模板框架得到初版报告内容；随后调用大模型进行语言润色和逻辑衔接。最后输出结构完整、

图4　人机协同机制的证据链抓取流程

逻辑清晰的初版判责报告。整体流程见图5。生成的报告案例见图6。

图5　报告自动生成逻辑框架

图6　自动生成报告示例

4　实验与结果

基于本文提出的电信申诉判责全流程智能化方法，本文采用真实电信申诉数据，对各个处理环节的准确率进行抽样验证，同时对比本研究提出的方法与人工处理方法的效率差异。

4.1　实验数据集与环境

本研究所使用的数据来自某省级运营商2025年3月至6月期间的申诉工单，共计约7000多条，主要投诉来源来自工业和信息化部。原始数据均为结构化Excel表格，包含用户基本信息、申诉内容、申诉时间、涉及号码及相关分类码字段。实验中，模型通过API调用，上下文窗口为32k tokens，温度设为0。数据预处理包括重复工单去除、地址字段分离及敏感信息脱敏。

为保证实验评估的可操作性与标注可靠性，本文从总体样本中随机抽取部分工单构建人工标注集。其中，无效申诉判断随机抽取500条工单，人工标注结果直接采用运营商一线客服依据业务规范形成的实际处理结论；服务顽疾分类从有效样本中随机抽取500条，由两名研究人员独立标注，初始一致率为97.8%，分歧样本经与资深业务人员讨论后达成一致；申诉内容拆解随机抽取200条有效工单，由一名具有5年电信投诉处理经验的专业人员逐条标注原子要点，形成金标准数据；证据列表生成与报告生成由于受外部抓取、人工编辑和主观评价影响，难以建立统一金标，因此主要采用处理耗时作为对照指标评估系统效率。

4.2　评估指标与基线

为了全面衡量系统在各个环节的性能，本研究从准确性与效率性两个维度设置评估指标。

对于无效申诉判断与服务顽疾分类两类分类任务，采用准确率(Accuracy)、精确率(Preci-

表3　源数据集部分字段表

字段名	含义	类型	示例	备注
平台流水编号	记录唯一标识	字符串	部转-20250400094985	主键
用户申诉时间	首次提交时间戳	Datetime	2023-05-12 14:32:10	格式固定
客户姓名	申诉人姓名	字符串	张*	脱敏处理
手机号	联系电话	字符串	138****5678	脱敏处理
申诉涉及号码	投诉事件涉及号码	字符串	186****4321	脱敏处理
通讯地址	用户通信地址	字符串	北京	/
投诉内容	自由文本	文本	账户被误停机…	清洗去噪脱敏
分类码	问题分类	字符串	收费	基础分类
业务码	业务类型	字符串	通信(234G)	基础分类

sion)、召回率(Recall)、F1值、宏平均Macro-F1、微平均Micro-F1和Cohen’s Kappa系数进行评估，并采用McNemar检验考察系统与人工结果之间是否存在系统性偏差。

$$Precision=\frac{TP}{TP+FP},\quad Recall=\frac{TP}{TP+FN},\quad F1=\frac{2PR}{P+R}\#(1)$$

对于申诉内容拆解任务，考虑到电信申诉文本的复杂性和要点间的语义关联，本文采用“一对多匹配策略”进行评估，允许一个预测要点匹配多个金标准（Gold）要点，或一个金标准要点被多个预测要点匹配。这种策略更符合实际业务场景中的语义粒度差异。评估指标包括：

（1）要点级F1：采用一对多匹配方法，对系统要点与Gold要点进行配对。每个模型拆解出的要点与每个Gold要点均计算文本片段重叠率（IoU）、语义相似度与要点类型一致性，只要相似度超过阈值即认为有效匹配，允许一个Gold要点与多个系统要点配对；

（2）覆盖率与冗余率：衡量模型的要点识别完整性，以及是否出现冗余或过度切分；

（3）可信度：评估模型输出的忠实性，检测幻觉现象；

对于效率实验，则比较人工处理耗时(T_manual)与系统处理耗时(T_system)，并计算效率提升率，用以评估方法在真实业务流程中的应用价值。

$$\text{提升率}=\frac{T_{manual}-T_{system}}{T_{manual}}\times 100\%\#(2)$$

通过以上指标，本研究能够从结果正确性与处理效率两个维度，对系统的全流程性能进行全面评估。

同时，为验证本文基于大模型和提示词工程（LLM+Prompt）方法的有效性，在分类任务中选取以下代表性基线进行对比：

（1）规则引擎（Rule-based）：基于正则表达式、白名单等硬规则。

（2）传统机器学习（TF-IDF + XGBoost）：提取文本特征与分类码，训练XGBoost分类器。

（3）深度学习微调（BERT Fine-tuning）：基于chinese-bert-wwm-ext预训练模型在标注样本上微调。

针对申诉拆解任务，基线增加序列标注模型（BERT-CRF）、抽取式阅读理解模型（RoBERTa-MRC）以及零样本设定下的大模型。所有评估采用准确率、精确率、召回率、F1分数等通用指标。

4.3 无效判断实验结果

本研究将无效判断任务转化为二分类任务，将原有多项无效规则判断结果统一转变为“无效”结果，便于统计分析。对随机抽取的500条数据运用“无效判断模型”，与无效判断金标数据结果作对比，得到以下评估结果，见表4。

表4 无效判断评估指标结果（N=500）

类别	精确率	召回率	F1分数	支持数
无效 (0)	0.8459	0.8746	0.8600	295
有效 (1)	0.8103	0.7707	0.7900	205
宏平均	0.8281	0.8227	0.8250	500
加权平均	0.8313	0.8320	0.8313	500

针对基线实验做如下处理：

（1）规则引擎（Rule-based）：对于规则无法覆盖的语义模糊场景，统一标记为“有效”。

（2）TF-IDF + XGBoost：将申诉内容进行TF-IDF向量化（5000维），提取文本长度、号码格式等统计特征，利用XGBoost分类器进行训练。

（3）BERT微调（BERT Fine-tuning）：在2000条标注样本上进行微调（学习率2e-5, Batch Size 16, 训练5 Epochs），利用深度学习捕捉语义特征。

得到各基线方法在500条测试集上的表现对

比见表5。

结果显示，规则引擎的召回率最高，达到0.917，但准确率仅为0.466，说明其虽然能够覆盖部分显式规则场景，但对“申诉事由不清”“超时限”等依赖语义判断的复杂情况几乎无能为力。TF-IDF+XGBoost与BERT的F1值分别为0.579和0.582，表现均明显低于本文方法。相比之下，本文方法的准确率达到0.832，F1值达到0.790，Kappa达到0.650，综合性能最优。该结果说明，基于提示词工程的大模型方法能够较好弥补规则方法和小样本监督模型在复杂语义场景中的不足，更适合处理口语化、模糊化且噪声较高的真实电信申诉文本。

4.4　顽疾类型判断实验结果

在七类顽疾问题的分类实验中，模型在整体上取得了97%的准确率。具体数据见表6。结果表明，本研究方法在大多数顽疾类别上能够与人工标注高度一致，但对于小样本类别仍需通过数据扩充和正负样本平衡进行优化。

针对基线实验，为公平评估各方法在服务顽疾分类任务上的性能，从500条标注数据中随机抽取100条作为测试集，剩余400条作为训练集（划分时采用分层抽样，保持各类别比例）。所有方法均在完全相同的数据划分下进行训练和预测。

由于原始数据中“互联网营销不规范”（9条）和“业务不知情定制”（5条）两类样本数量极少（不足10条），若保留独立类别将导致训练集和测试集划分时出现零样本或极端不平衡问题，影响评估可靠性。因此，在基线实验中将这两类合并至“其他”类别，最终形成6分类任务。

最终各模型在测试集上的详细分类结果见表7。

结果显示各类方法在判断此分类任务上均表现出极高的性能，但本文方法在所有类别上均取得完美1.00 F1分数，并且本文提出的方法能够以零训练成本达到甚至超越全监督基线模型的性能，展现了规则与大模型结合在文本分类任务中的巨大潜力。针对长尾类别“互联网营销不规范”中出现的偶发误判（召回率仅0.11），分析发现主要因大模型被“手机”等字眼误导归类为外呼营销，未来拟通过引入少样本示例和多模态截图解析予以优化。

4.5　申诉拆解任务实验结果

在200条人工标注的金标数据集上，本文方法在CDP任务上的表现结果见表8，相关指标证明了模型在复杂申诉文本下的要点识别与要素化表达能力。

为了验证本文方法在申诉拆解任务（CDP）中的优越性，本研究构建了以下对比实验方案：

（1）有监督基线：选用BERT-CRF（序列标注）与 RoBERTa-MRC（抽取式阅读理解）。由于此类模型依赖标注数据驱动，本研究从200条金标工单中随机抽取150条作为训练集进行参数微调，剩余50条作为独立测试集。

（2）生成式基线：选用Moonshot-v1-32k模型，在零样本（Zero-shot）的设置下直接输出拆解结果。

（3）本文方法（LLM+prompt）：在相同的50条独立测试集上，利用本文设计的Few-shot和

表5　无效判断基线对比结果（N=500）

基线方法	准确率	精确率	召回率	F1分数	Cohen's Kappa
规则引擎	0.466	0.429	**0.917**	0.585	0.059
TF-IDF + XGBoost	0.636	0.551	0.610	0.579	0.259
BERT (Fine-tuning)	0.690	0.651	0.527	0.582	0.340
本文方法(LLM+prompt)	**0.832**	**0.810**	0.771	**0.790**	**0.650**

表6　顽疾详细分类评估结果

类别	精确率	召回率	f1-score	支持数
限制套餐变更	1.00	1.00	1.00	81
超套费用质疑	1.00	1.00	1.00	79
反诈号码关停	1.00	1.00	1.00	24
外呼营销不规范	0.78	0.94	0.85	31
互联网营销不规范	1.00	0.11	0.20	9
业务不知情定制	1.00	1.00	1.00	5
其他	0.97	0.98	0.98	237
准确率			0.97	500
宏平均	0.97	0.88	0.88	500
加权平均	0.97	0.97	0.97	500

表7　顽疾详细分类性能对比(N=100)

类别	支持数	XGBoost F1	BERT F1	本文方法 F1
其他	50	0.99	0.99	**1.00**
反诈号码关停	5	1.00	1.00	**1.00**
外呼营销不规范	6	0.92	0.92	**1.00**
超套费用质疑	16	1.00	1.00	**1.00**
限制业务退订	7	1.00	1.00	**1.00**
限制套餐变更	16	1.00	1.00	**1.00**
宏平均	-	0.99	0.99	**1.00**
准确率	-	0.99	0.99	**1.00**

表8　申诉拆解任务评估指标结果（N=200）

模型	F1 分数	精确率	召回率	覆盖率	冗余率	可信度
本文方法	0.739	0.735	0.743	0.821	0.228	0.847

CoT提示工程进行推理。

基线实验结果如表9所示。

表9　申诉拆解任务评估指标结果

模型方法	F1 分数	精确率	召回率	语义相似度	覆盖率	冗余率	可信度
BERT-CRF	0.345	0.361	0.331	0.585	0.392	0.551	1.000
RoBERTa-MRC	0.082	**0.700**	0.044	0.583	0.062	0.060	0.000
Moonshot (Zero shot)	0.451	0.445	0.456	0.579	0.512	0.549	0.767
本文方法 (LLM)	**0.641**	0.620	**0.663**	**0.865**	0.791	0.294	0.848

实验结果显示，在仅有150条训练样本的情况下，BERT-CRF和MRC的表现远低于预期。BERT-CRF虽然保持了极高的可信度（无幻觉产生），但F1仅为0.345，说明传统模型在小规模电信垂直行业数据下难以充分学习工单拆解的复杂规律。RoBERTa-MRC在本任务中召回率极低（0.044），主要原因是申诉文本口语化严重且存在大量“跨句关联”，所谓'跨句关联'，是指用户在申诉时往往将同一诉求分散在不连续的句子中表达，例如在第一句说明费用金额、中间穿插背景信息，最后再提出退款要求——三个句子共同构成一个完整的投诉要点，而单一的指针网络难以将其识别并整合为一个回应单元。

相比之下，本文方法在50条测试集上取得了0.641的F1分数。虽然相较于全量数据集（200条）评估时略有波动，但其覆盖率（0.791）和语义相似度（0.865）均显著领先于所有基线模型。这证明了通过思维链（CoT）引导大模型进行推理，能够有效弥补垂直领域标注数据不足的短板，实现了在极小样本下的高性能拆解。

4.6　泛化性验证

为评估模型在真实业务环境中的时间泛化能力，本文以2025年3月数据为基础构建最终版提示词，并在后续三个月（4月、5月、6月）的独立月度工单数据集上进行验证。每月测试集均包含1000条投诉样本，且与训练期数据无重叠。各月性能指标如表10所示。

表10　无效判断任务跨时间泛化测试结果（每月1000条）

月份	准确率	精确率(+)	召回率(+)	F1(+)	宏平均F1	Kappa
4月	0.801	0.742	0.712	0.726	0.785	0.570
5月	0.820	0.675	0.794	0.730	0.797	0.596
6月	0.832	0.678	0.783	0.727	0.803	0.606

结果显示，模型在连续三个月上的准确率分别为0.801、0.820和0.832，正类F1稳定0.726-0.730区间，Kappa介于0.570-0.606，说明该方法对时间分布变化具有较好的稳定性。

此外，为构造具有代表性的域外（OOD）样本池，本文从最近4–6个月的工单中人工筛选并

去重，优先保留那些在训练/开发集中未见或显著少见的业务类型（例如：光纤宽带、短信、增值等）。最终 OOD 池经人工复核包含 245 条示例。在评估中，本文从 in-domain（N=245）与 OOD 池按比例 p 混合构造测试集（N=200，p∈{0.0,0.2, 0.5,1.0}，k=3 次重复）。结果如表 11 所示，当包含未见业务类型的数据比例超过 0.5 时，模型 F1 开始发生明显衰减（降至 0.708），这印证了在真实生产部署时，需建立周期性的提示词更新维护机制。

表 11　不同 OOD 比例下的模型性能（F1 正类）

OOD 比例 p	平均 F1	标准差
0.0	0.7726	0.0069
0.2	0.7580	0.0071
0.5	0.7084	0.0221
1.0	0.6702	0.0049

4.7　效率对比

针对无效判断、顽疾分类、申诉拆解环节，本研究通过与一线申诉处理人员访谈及工作记录查阅，获取了各处理环节的人工平均耗时估算，在相同数据集上应用本文方法处理，记录处理时间（不含人工审核时间），评估人工与系统效率的差异。具体数据见表 12。

此外，针对证据链抓取与报告生成环节，此类任务在真实生产环境中较为分散且不固定，受资源人力限制等，人工处理时间不易跟踪，故只得到小部分数据，小样本上显示，在这两个环节上系统耗时远小于人工平均耗时。

结果表明，本文提出的方法能够在保持较高准确率的同时，大幅降低人工处理负担。

5　讨论

本文实验结果表明，大模型在电信申诉判责任务中的作用具有明显的任务差异性。在无效判断和服务顽疾分类等相对封闭的分类任务中，基于提示词工程的方法已能够达到较高性能，甚至在部分场景下超过传统监督学习基线；而在申诉内容拆解这类开放式生成任务中，模型虽然表现出较强的覆盖能力，但仍存在一定冗余和粒度不稳的问题，说明生成类任务对输出约束与质量控制提出了更高要求。

表 12　效率对比表

环节	数据条数	人工耗时	系统耗时	效率提升率(%)
无效判断	200 条	2h	5min	95.8%
顽疾分类	200 条	1.5h	3min	96.7%
申诉内容拆解	200 条	3h	20min	88.9%
证据抓取	10 条	1h	25min	58.3%
报告生成	10 条	35min	6min	82.9%

本文采用人机协同而非完全自动化的证据收集机制，并非单纯出于技术能力限制，更与电信申诉判责场景的合规性、可追溯性要求密切相关。对涉及用户权益的判责业务而言，保留人工审核节点有助于控制错误风险，并增强结果的审计可解释性。因此，本文框架的现实价值不仅在于“自动化”，更在于在可控边界内实现“高效自动化”。

同时，跨时间与域外样本测试也说明，基于提示词工程的方法虽然具备低成本、易部署的优势，但对业务分布变化仍较为敏感。特别是在长尾类别和新型投诉场景出现时，模型性能会受到影响。未来可考虑建立“新样本积累—提示词更新—回归测试”的轻量维护机制，以提升模型的持续适应能力。

6　结束语

本文提出了基于大语言模型的电信申诉判责全流程智能化方法，构建了覆盖无效判断、顽疾分类、申诉拆解、证据收集与报告生成的端到端框架，并通过真实工单数据验证了其在准确性和效率上的优势。同时，本研究不仅在技术层面推动了电信申诉处理的自动化与智能化，也为运营

商在客户服务、成本控制与业务数字化转型方面提供了有益的借鉴。

本研究的创新主要体现在几个方面：首先，本研究首次系统性地构建了面向电信申诉的全流程智能化框架，解决了以往研究聚焦单点、缺乏整体性的不足；其次，提出了人机协同的证据链收集机制，在保证准确性的同时大幅降低了人工成本；再者，验证了大模型在多诉求文本拆解中的独特优势，显著提升了复杂申诉场景下的要点识别与要素化表达能力；最后，在真实工单数据上的实验验证了方法的可行性与应用价值，显示出较强的实践意义。

当然，本文研究仍存在若干有待深化的问题：数据来源的单一性制约了方法跨场景适用性的充分论证；证据收集的自动化程度和自适应能力有待提升；部分长尾类别的识别性能不足等。这些问题将在后续研究中重点推进，具体包括：

（1）探索针对小样本顽疾类别的数据增强策略；

（2）研究更具自适应能力的Agent框架以提升证据收集灵活性；

（3）在多省多运营商场景下开展更大规模的实证验证。

参考文献：

[1] 黄堃, 胡涵清, 赵东明, et al. 基于深度学习的电信运营商网络投诉工单智能分类技术研究 [J]. 电信工程技术与标准化, 2023, 36(10): 6-12.

[2] 梁伟明, 肖军, 马晓亮, et al. 基于改进Transformer的电信重投报告自动生成方法研究 [J]. 电信科学, 2025, 41(6): 197-207.

[3] 梁昕露, 李美娟. 电信业投诉分类方法及其应用研究 [J]. 中国管理科学, 2015, 23(S1): 188-192.

[4] 赵进, 杨小军. 基于GRW和FastText模型的电信用户投诉文本分类应用 [J]. 电信科学, 2021, 37(06): 125-131.

[5] Yang Y, Xu D-L, Yang J-B, et al. An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications [J]. Knowledge-Based Systems, 2018, 162: 202-210.

[6] 祝好, 齐磊, 顾慧琼. 基于人工神经网络算法的运营商客户投诉智能分类问题研究 [J]. 电信工程技术与标准化, 2021, 34(03): 31-35.

[7] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Minneapolis, Minnesota, F June, 2019 [C]. Association for Computational Linguistics.

[8] 张亮, 代晓菊, 郑荣, et al. 多模型融合的客服工单文本分类方法的研究与实现 [J]. 电信科学, 2021, 37(11): 86-96.

[9] Zhou H, Hu C, Yuan Y, et al. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities [J]. IEEE Communications Surveys & Tutorials, 2024.

[10] 查德飞, 臧永顺, 史鸿晖. 通信企业大模型技术突破与规模化应用实践 [J]. 数字通信世界, 2025, (08): 232-234.

[11] 王希, 赵灏, 尹培. 基于星辰语义大模型的潜在申诉用户预测应用研究 [J]. 山东通信技术, 2025, 45(02): 38-42.

[12] 殷昌承, 陆绍雯, 冯超, et al. 面向业务端到端的投诉智能体技术研究与应用 [J]. 电信工程技术与标准化, 2025, 38(08): 1-6.

[13] Sahoo P, Singh A K, Saha S, et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [J/OL] 2024, arXiv: 2402.07927[https://ui.adsabs. harvard. edu/abs/2024arXiv240207927S. 10.48550/arXiv.2402.07927

[14] Pawlik L. How the Choice of LLM and Prompt Engineering Affects Chatbot Effectiveness [J]. Electronics, 2025, 14(5): 888.

[15] Gao M, Sun B, Wang T, et al. Domain-Aware Reinforcement Learning for Prompt Optimization [J]. Mathematics, 2025, 13(16): 2552.

[16] Shahriar A, Hisham S J, Rahman K A, et al. 5GPT: 5G vulnerability detection by combining Zero-Shot capabilities of GPT-4 with domain aware strategies through prompt engineering [J]. IEEE Transactions on Information Forensics and Security, 2025.

[17] Wang F, Zhang Z, Zhang X, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness [J]. ACM Transactions on Intelligent Systems and Technology, 2025, 16(6): 1-87.

[18] Ilchenko M. NEXT-GEN TELECOM AI: MASTERING PROMPT ENGINEERING FOR INNOVATION [J]. Information and Telecommunication Sciences, 2025, (1): 22-29.

[作者简介]

张梦婷（2003-），女，北京邮电大学硕士研究生，主要研究方向为数据挖掘与商务

智能。

张晓航（1975-），男，北京邮电大学博士，教授，主要研究方向为数据挖掘与商务智能、风险建模、因果推理。

李征仁（1987-），男，博士，北京邮电大学，副教授，主要研究方向为数据挖掘与商务智能。

王海燕（1977-），女，博士学位，中国信息通信研究院高级工程师,主要从事人工智能技术应用及电信用户服务研究。

陈中华（1990-），男，学士学位，中国移动通信集团内蒙古有限公司，中级工程师，系统支撑方案设计，大数据及人工智能。