研究与开发

网络投诉意图识别：一种基于大模型增强的少样本学习方法

顾宁伦[1]，罗志毅[1]，王希栋[2]，丁建兵[2]，欧阳晔[2]
（1. 中国移动通信集团有限公司，北京 100032；
2. 亚信科技（中国）有限公司，北京 100193）

摘　要：针对5G时代日益复杂的网络架构与用户规模增长带来的网络投诉处理挑战，旨在解决实际投诉数据中普遍存在的少样本类别意图识别难题。传统小型自然语言处理（natural language processing，NLP）模型因对数据量和质量高度依赖，在少样本场景下泛化能力不足，识别效果欠佳。大语言模型（large language model，LLM）虽具备强大的通用语义理解能力，但受限于通信领域专业知识匮乏及巨大计算资源需求，难以直接高效应用于此特定场景。为此，提出一种基于大模型增强的少样本网络投诉意图识别算法。该算法创新性地融合了LLM的文本生成优势与小型模型的高效推理特性，通过LLM迭代生成针对少样本类别的高质量模拟训练样本，并引入小模型进行质量评估与反馈，持续优化样本生成过程。该机制显著提升了小模型对少样本投诉意图的识别准确率。基于实际网络投诉数据的测试结果表明，所提方案使3类少样本网络投诉的意图识别准确率提升了21%，并实现了全部8类投诉识别准确率9%的整体提升。

关键词：大语言模型；少样本学习；网络投诉意图识别
中图分类号：TN911.7；TP18
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025236

A large language model-enhanced few-shot learning algorithm for network complaint intent recognition

GU Ninglun[1], LUO Zhiyi[1], WANG Xidong[2], DING Jianbing[2], OUYANG Ye[2]
1. China Mobile Communications Group Co., Ltd., Beijing 100032, China
2. AsiaInfo Technologies (China), Inc, Beijing 100193, China

Abstract: Addressing the escalating challenge of network complaint handling in the 5G era, driven by increasingly complex network architectures and growing user bases, the aim is to solve the prevalent issue of few-shot intent recognition for complaint categories with scarce samples. Traditional small-parameter natural language processing (NLP) models exhibit insufficient generalization and suboptimal performance in such scenarios due to their heavy reliance on data quantity and quality. While large language models (LLM) possess strong general semantic understanding, their limited domain-specific knowledge in telecommunications and substantial computational demands hinder direct,

收稿日期：2025−07−02；修回日期：2025−10−13
通信作者：王希栋，wangxd18@asiainfo.com

efficient application to few-shot network complaint intent recognition. To precisely tackle this, an LLM-enhanced few-shot intent recognition algorithm for network complaints was proposed. This innovative algorithm integrated the powerful text generation capabilities of LLM with the efficient inference characteristics of small models. It leveraged LLM to iteratively generate high-quality simulated training samples for few-shot categories, incorporating a small model for quality evaluation and feedback to continuously optimize the sample generation process. This mechanism significantly boosted small model training and specifically enhanced the recognition accuracy of few-shot complaint intents. Test results on actual network complaint data demonstrated a 21% improvement in intent recognition accuracy for 3 types of few-shot network complaints, and a 9% overall improvement in recognition accuracy across all 8 complaint types.

Key words: large language model, few-shot learning, network complaint intent recognition

0 引言

随着移动通信网络的持续演进与5G技术的深度部署，网络结构日益复杂，业务种类愈发丰富，用户规模亦持续扩张。在这一背景下，网络运营商面临日益严峻的运营维护挑战，其中，高效精准地处理海量用户投诉成为保障服务质量与用户满意度的关键环节。用户投诉的规模、类型多样性以及诉求的紧迫性，都对传统人工客服处理模式提出了更高的技术要求。因此，引入人工智能（artificial intelligence，AI）技术以优化用户投诉处理流程，已成为网络智能化发展的必然趋势。

作为网络投诉处理的核心，网络投诉意图识别并非一项简单的文本分类任务，而是一个涉及复杂行业背景、多变用户语言和高度时效性的综合性场景。网络投诉数据通常呈现出一种典型的长尾分布，少数高频投诉类型，如网络信号不稳定、网速慢等，占据了绝大部分的投诉量。与此同时，存在大量低频但至关重要的投诉类别，如针对特定新业务的资费疑问、某一地区特定基站的设备故障举报等。这些低频类别由于发生频率低，对应的历史训练样本极度稀缺，形成所谓的“少样本”类别。在早期及当前的部分网络智能化研究中，通常采用自然语言处理（natural language processing，NLP）领域的小参数量AI模型通过监督学习进行投诉工单意图识别[1]。这些方案通常从历史工单中提取带标签的投诉描述，并利用机器学习[2]或深度学习方法（如基于Transformer的双向编码器表示（bidirectional encoder representations from Transformer，BERT）[3]、知识增强的语义表示（enhanced representation through knowledge integration，ERNIE）等）训练模型。然而，在实际应用中，由于真实历史投诉工单数据中不同投诉类型的样本比例存在巨大差异，加之用于训练的NLP小模型参数规模较小，因此，其泛化能力受限，对少样本投诉类型的识别效果尤为不佳，严重制约了整体性能表现。大语言模型（large language model，LLM）快速发展，在通用文本分类、语义理解等NLP任务中展现出强大的能力。因此，将LLM应用于网络投诉工单意图识别也受到了广泛关注。目前，主流的应用方案包括通过提示工程设计特定的输入模板来引导模型输出分类结果，以及利用标注数据集对大模型进行微调，以适应特定领域任务[4]。然而，通用大模型在预训练阶段缺乏对移动通信领域专业知识的深入学习，仅依靠提示工程和模型微调难以实现对网络投诉工单的精准分类，尤其是在面对通信领域特有的少样本投诉类型时表现欠佳。此外，通用大模型对计算资源（算力）的极高要求也显著增加了其在实际业务场景中部署的难度和成本。

为精准解决上述在网络投诉智能处理过程中面临的少样本识别挑战，本文提出了一种基于大模型增强的少样本网络投诉意图识别算法。该方案创新性地融合了通用大模型的强大语言生成能力与小模型在特定领域任务上的高效推理能力，旨在通过大模型有针对性地生成高质量的少样本模拟训练数据，有效弥补真实少样本数据的不足，从而显著提升小模型对稀有投诉类型的识别能力，最终实现网络投诉意图的精准识别，同时兼顾对业务现场算力需求的降低。

本文的主要贡献可归纳为以下3个方面。

（1）系统性研究了大模型与小模型在少样本场景下的增强机制，并综合现有研究成果，归纳了其主要实现形式。

（2）提出了一种创新的基于大模型生成训练数据来增强小模型少样本学习能力的算法框架。该框架通过大模型迭代生成高质量的模拟训练样本，并结合小模型对生成样本的质量评估反馈机制，实现了大小模型之间的良性互动与性能互提升，有效解决了网络投诉意图识别中的少样本问题。

（3）构建了基于网络业务现场数据的网络投诉处理数据集，并基于此数据集验证本方案的实用效果。本方案对3类少样本网络投诉的意图识别准确率有21%的显著提升，对于全部8类投诉的识别准确率也提升了9%。

1　相关研究现状

本节将回顾与本文研究紧密相关的两个主要领域：少样本学习的研究进展及其在实际应用中面临的挑战，尤其是针对文本分类任务；LLM在增强少样本学习能力方面的最新范式，尤其是通过大小模型协同来实现这一目标。

1.1　少样本学习挑战与进展

在现实世界的许多应用场景中，尤其是在专业领域如移动通信的网络运维中，数据标注成本高昂或特定事件发生频率极低，导致某些类别的训练样本极其稀缺。这种数据稀缺性问题直接催生了少样本学习这一重要研究方向。少样本学习旨在使模型仅通过少量示例（通常是几个到几十个）就能快速适应新任务或识别新类别，从而克服传统深度学习模型对大规模标注数据的强烈依赖。

传统的少样本学习方法主要集中在以下几方面。

（1）元学习：通过学习“如何学习”来快速适应新任务，使模型能够从一系列相关任务中学习元知识，并将其应用于只有少量样本的新任务。例如，模型无关元学习（model-agnostic meta-learning，MAML）算法在此领域取得了显著进展，该算法旨在通过少量梯度更新快速适应新任务[5]。在少样本文本分类研究中，元学习方法已被广泛探索[6]。

（2）数据增强：通过对现有少量数据进行变换或生成新数据来扩充训练集，以增加数据的多样性和数量，从而提升模型的泛化能力。在NLP领域，常见的文本数据增强技术包括同义词替换、回译、文本混淆等[7]，特别是利用生成式模型合成数据已成为一种有效的数据增强策略，尤其适用于数据稀缺的场景[8]。

（3）迁移学习：利用在大量源数据上预训练的模型（如预训练语言模型），将其学到的通用知识和表示能力迁移到目标少样本任务上进行微调。这种方法在NLP领域取得了巨大成功，显著降低了对目标任务大规模标注数据的需求[9]。

尽管上述这些方法在一定程度上缓解了少样本问题，但在面对复杂、多样的真实世界数据，特别是网络投诉这种具有领域特异性且呈现长尾分布的文本数据时，其泛化能力和性能仍有提升空间[10]。长尾分布现象普遍存在于网络投诉意图识别任务中，即少数投诉类型占据了绝大部分数据，而大量重要的投诉类型却只有极少量样本，

这使得少样本学习成为该领域亟待解决的关键难题。

1.2 大模型增强的少样本学习

近年来，随着LLM的飞速发展，其在海量通用文本数据上进行预训练后展现出的强大通用知识、语义理解和文本生成能力，为解决少样本学习问题带来了新的契机。LLM能够通过其强大的泛化能力，从少量示例中捕捉到更深层次的语义模式，并生成高质量的合成数据，从而有效弥补少样本类别的数据不足。这种利用大模型能力来提升小模型在少样本任务上表现的策略，可被视为一种大模型增强少样本学习的有效途径。

具体而言，大模型向小模型的知识迁移是实现这种增强的关键范式之一。这种方式充分利用了大模型的通用知识基础，通过以下多种技术手段将其泛化能力有效地植入小模型中。

（1）知识蒸馏。通过让小型“学生”模型学习大型“教师”模型的输出（如软标签或特征表示），从而将大模型的知识压缩并转移到小模型中。蒸馏过程通常发生在大模型中，将提炼出的模型转移到小型模型中进行进一步适应，这些“学生”模型可以是规模较小的一般模型[11]或特定领域的模型[12]。

（2）合成数据生成。这是大模型增强少样本学习的关键技术。大模型凭借其强大的文本生成能力，能够根据少量真实样本和特定的提示，合成大量高质量、多样化的模拟训练数据集。这些合成数据可以直接用于训练小型模型[13]，尤其适用于扩充少样本类别，从而显著提升小模型在这些类别上的识别性能。

（3）参数化知识转移。知识丰富的“教师”模型（大模型）根据业务需求有选择地将静态参数知识转移到“学生”模型（小模型）中[14]。

（4）多方知识共享。联邦学习可实现多个分散数据方之间的协作，应用联邦学习进行模型间的知识共享有助于应对经常困扰单个数据方的数据稀缺性和分布偏差带来的挑战[15]。在这种模式下，使用不同数据集训练的各种体系结构的不同大模型会从不同的角度考虑相同的特定领域任务，多个大模型的知识经过融合后，统一传递给目标小模型[16]。

2 基于大模型增强的少样本网络投诉意图识别算法

本文所提出的网络投诉意图识别方案，核心在于利用LLM的强大生成能力，专门增强小模型在处理少样本投诉类型时的意图识别准确性。该方案属于大模型向小模型知识迁移的增强范式，其中知识迁移的载体是大模型合成的模拟训练样本。同时，小模型对这些生成训练样本的质量进行评估并提供反馈，从而帮助大模型迭代优化合成数据的质量，最终实现小模型在少样本网络投诉意图识别准确率上的显著提升。

2.1 算法概述

基于大模型增强的少样本网络投诉意图识别流程如图1所示，该流程由6个步骤组成，通过多轮次大模型增强和小模型反馈迭代优化实现任务目标。

（1）小模型初始训练与评估。小模型基于现有真实的网络投诉工单数据集进行初次训练，并进行推理测试。

（2）大模型增强触发。当小模型在初次测试中，尤其是在少样本投诉类型上的识别准确率未能达到预设的业务需求门限时，将触发大模型辅助的增强学习过程。

（3）大模型生成模拟训练数据。大模型根据真实数据集和前一轮生成数据中筛选样本示例，有针对性地生成高质量的模拟训练数据集。

（4）小模型二次训练及评估。小模型利用大模型生成的训练数据进行优化训练，生成优化训练版本的网络投诉意图识别小模型。为便于评估大模型增强效果，小模型二次训练参数应尽可能

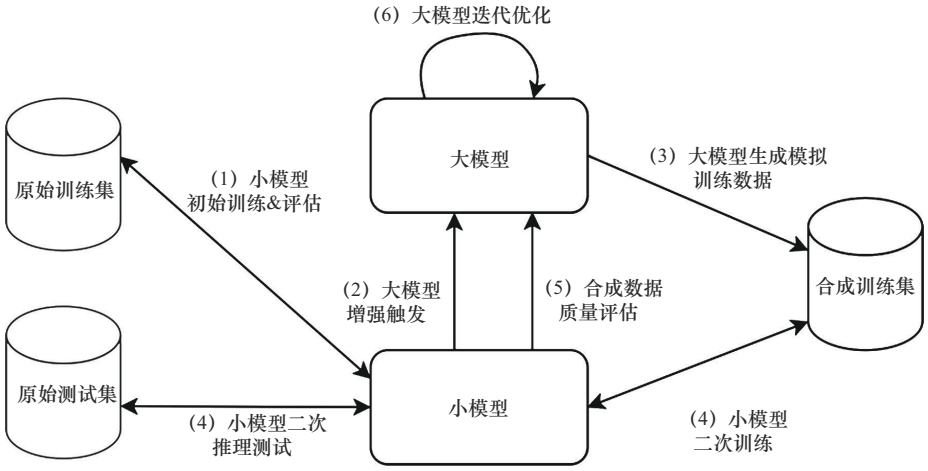


图1　基于大模型增强的少样本网络投诉意图识别流程

与初始训练版本保持一致。

（5）合成数据质量评估。在二次训练完成后，小模型根据训练过程中大模型生成数据的表现，对合成数据质量进行量化评估，并将评估结果反馈给大模型。

（6）大模型迭代优化。大模型根据小模型的反馈，调整其数据生成策略，持续优化模拟训练样本的质量。此过程与小模型的训练推理过程交替迭代，直至小模型在少样本投诉意图识别上的准确率满足业务需求。

2.2　大模型的少样本数据生成

在网络投诉意图识别场景中，大模型增强小模型在少样本条件下的学习能力是本方案的关键环节，其核心在于大模型生成高质量的模拟训练数据来扩充少样本类别的数据集。为了最大化模拟训练数据的质量和多样性，本文应用了角色扮演、少样本示例、思维链（chain of thought，COT）等提示工程技术来提升大模型的生成效果。大模型生成模拟训练数据的提示样例如图2所示。通过精心设计的提示词，大模型能够更好地理解特定投诉类别的语义特征和表达习惯，从而生成更贴近真实、更具代表性的少样本数据。

大模型通过生成模拟训练样本向小模型进行知识迁移的过程是迭代优化的。大模型从真实数据集和前几轮生成的合成数据集中筛选出优质样本，作为下一轮数据生成时提示工程中的少样本示例数据。在第一轮数据生成时，大模型从原始训练集中随机选取 n 条作为该类别的样例数据。

"gen_x_instruction　“：“ 你是一名专注于生成AI训练数据的资深内容创作者，精通用户投诉的各种表达方式和常见情境。你的任务是为指定的网络投诉类别生成一条高度逼真、具体且具有代表性的模拟投诉工单。
** 请遵循以下思考步骤和要求进行生成: **
1. ** 深入理解类别 ：** 首先, 仔细分析目标投诉类别：“<类#个人业务*小类#业务类短信提醒>”。思考这类问题下用户最常遇到的具体困扰是什么?他们通常会用哪些词汇来描述？可能的情绪(如不满、焦虑、疑惑) 是什么?
2. ** 构思典型场景 ：**基于对类别的理解, 想象一个或多个实际用户可能提交投诉的典型场景。确保这个场景具体、合理, 能够充分体现该类投诉的核心问题。
3. ** 撰写用户视角内容 ：**以第一人称(用户) 的口吻撰写投诉工单内容。力求语言自然、细节丰富，能够清晰地表达用户的经历、问题以及期望。避免生硬或过于学术化的表达。
4. ** 严格排除类别名称 ：**确保生成的投诉工单文本中, 不直接提及或包含“<类#个人业务*小类#业务类短信提醒>”这个类别名称或其任何直接的变体。
5. ** 直接中文输出:**完成思考后，请直接输出最终的投诉工单中文内容,无需任何额外说明或前缀。
请开始为投诉类别 “<类#个人业务*小类#业务类短信提醒>” 生成一条模拟投诉工单：”
“ example_instruction “：“ 以下是一条属于“<类#个人业务*小类#业务类短信提醒>”类别的真实投诉工单示例可供您参考,以便更好地理解此类型投诉的特征 ：< X>"

图2　大模型生成模拟训练数据的提示样例

在后续数据生成轮次中，大模型则会根据前一轮小模型对于模拟训练数据的评估结果，选择得分最高的Top n条作为该类别的样例数据。此外，在数据生成过程中，需要通过调节表征生成批次大小的参数n（如控制每次生成的样本数量），以避免推理上下文token数超限或生成数据过于相似等问题，从而确保生成数据的多样性和有效性。

2.3 小模型的推理评估

在网络投诉意图识别场景中，小模型向大模型反馈合成训练数据的评估质量是本方案的另一关键环节，其目的是确保大模型生成的少样本投诉类型的模拟数据能够真正有效地提升小模型的推理性能。样本质量的评估以量化评分Q的形式反馈给大模型。小模型生成合成训练数据的质量评估算法如算法1所示，在大模型生成模拟训练样本后，小模型会利用这些合成数据进行二次训练。二次训练的参数设置与基于真实数据的初次训练保持一致。训练完成后，小模型采用与初次推理测试相同的真实测试集进行推理测试，并基于训练推理的综合效果对合成数据的质量进行评估。

算法1 小模型生成合成训练数据的质量评估算法

输入 生成训练集M、历史数据集D_{history}

输出 每条训练数据的质量得分Q_i

对于所有生成样本$x_i \in M$

初始化贡献度$I(x_i)$

初始化 相似度$S(x_i)$

计算置信度$C(x_i)$

对$n=1,2,\cdots,N$

$p(y_i \mid x_i) \leftarrow$小模型在第$n$轮对样本$x_i$预测类别$y_i$的置信度

结束

$$C(x_i)=\frac{1}{N}\sum_{n=1}^{N}p(y_i|x_i)$$

计算波动度V_i

$$V(x_i)=\frac{1}{N}\sum_{n=1}^{N}\left|p(y_i|x_i)-C(x_i)\right|$$

$$I(x_i)=\frac{1}{2}*\left[C(x_i)+V(x_i)\right]$$

计算相似度$S(x_i)$

对于所有历史数据$d \in D_{\text{history}}$

计算余弦相似度$\cos(x_i, d)$

结束

$$S(x_i)=\frac{1}{\left|D_{\text{history}}\right|}\sum_{d\in D_{\text{history}}}\cos\left(x_i,d\right)$$

计算质量得分$Q(x_i)$

$$Q(x_i)=I(x_i)/[1+S(x_i)]$$

将$Q(x_i)$记为x_i的质量得分

结束

小模型对大模型生成的训练数据进行质量评估，并通过评分$Q(x_i)$量化每条训练数据的质量。数据质量评估主要从以下两个维度进行考量。

（1）影响力（Influence，$I(x_i)$）：衡量该条生成数据对模型训练推理效果的潜在影响。影响力越高，表明该样本对提升模型性能，特别是对少样本类别的识别能力越有益。

（2）相似性（Similarity，$S(x_i)$）：衡量该条生成数据与所有历史训练数据（包括真实数据集和已生成的合成数据集）之间的相似程度。本文期望生成的样本具有一定的新颖性，避免与现有数据高度重复，从而提供更多有价值的信息。因此，相似性越高，其评分中的权重越低。

综合考虑影响力和相似性，样本x_i的最终质量评分$Q(x_i)$计算式为：

$$Q(x_i)=I(x_i)/[1+S(x_i)] \tag{1}$$

生成样本x_i对于模型训练推理的影响力主要从置信度和可变性两方面衡量，如式（4）所示。置信度用于衡量小模型对于这一条生成样本的意图识别准确的确信程度，用小模型对于该样本x_i的实际类别y_i在多个epoch训练中的预测概率p来

表示，其中N表示训练epoch数量，如式（2）所示。可变性则体现了模型在不同训练周期中对同一样本真实分类的预测变化程度。高可变性意味着模型对样本的预测结果在训练过程中波动较大，表明样本具有较高的不确定性或模型对此类样本的泛化能力较弱。如式（3）所示，可变性通过计算小模型对于样本x_i的实际类别y_i的预测概率的平均绝对偏差获得。

$$C(x_i)=\frac{1}{N}\sum_{n=1}^{N}p(y_i|x_i) \tag{2}$$

$$V(x_i)=\frac{1}{N}\sum_{n=1}^{N}\left|p(y_i|x_i)-C(x_i)\right| \tag{3}$$

$$I(x_i)=\frac{1}{2}\left[C(x_i)+V(x_i)\right] \tag{4}$$

生成样本x_i的数据相似性S，通过计算x_i与历史数据D_{history}（包括真实数据集和已生成的合成数据集）之间的余弦相似度的统计平均值来衡量，计算式为：

$$S(x_i)=\frac{1}{\left|D_{\text{history}}\right|}\sum_{d\in D_{\text{history}}}\cos(x_i,d) \tag{5}$$

其中，d表示历史数据中任一条样本。

3　实验与分析

本节旨在验证基于大模型增强的少样本学习算法在网络投诉意图识别场景中的有效性。该方案针对网络投诉分析场景中训练样本不足及类别分布不均的问题，通过大模型增强的方式迭代优化扩充模拟训练样本，提升网络投诉意图识别小模型的少样本学习效果。本节应用真实网络数据，测试所提方案在不同样本数量投诉类型上的意图识别效果，并与仅基于原始数据训练的小模型进行对比分析。

3.1　实验数据

为确保测试结果贴近实际应用，本文构建了一个包含8类网络投诉类型的4 500条真实投诉工单数据集，网络投诉类型具体包括充值业务类投诉、宽带业务类投诉、专线业务类投诉等，数据经清洗和隐私加密后，用投诉类别1～8表示。真实数据集按9:1比例划分为训练集和测试集，包含4 050条真实训练样本和450条真实测试样本。小模型初次训练后，各类型网络投诉意图识别的平均F1分数为0.67，其中样本量不超过200的3类投诉F1分数平均值仅为0.48，凸显了少样本学习的挑战性。网络投诉意图识别作为多分类问题，本文选用准确率（precision）、召回率（recall）和F1分数作为主要评估指标。

实验选用Qwen2.5 14B和GLM-4-9B进行大模型增强数据生成，ERNIE 3.0-medium则作为网络投诉意图识别小模型进行少样本学习，大小模型的技术特性对比见表1。

3.2　实验方案及结果分析

首先，本文通过实验一，对大模型增强的少样本学习在不同样本数量下网络投诉类型意图识别的效果进行对比测试。

实验一　实验旨在评估对所有投诉类型进行大模型增强少样本学习的效果。本文通过Qwen2.5 14B增强迭代优化生成所有投诉类型的训练样本，为所有8类投诉类型各生成200条模拟训练数据，共计1 600条。ERNIE 3.0-medium在此基础上进行二次训练，并与仅基于原始数据

表1　大小模型的技术特性对比

模型特点	参数规模	推理时延	部署环境	核心优势
大模型1（Qwen2.5 14B）	14.7×10^9	时延较高，不适用于实时高并发场景	通常需要云端大模型图形处理单元（graphics processing unit，GPU）集群	强大的通用知识、语义理解和文本生成能力
大模型2（GLM-4-9B）	9×10^9			
小模型（ERNIE 3.0-medium）	75×10^6	极低，适用于实时高并发场景	边缘算力设备或小型服务器	高效推理、低成本、快速响应，领域适配性强

训练的小模型进行对比。实验一网络投诉意图识别效果统计结果见表2。

从表2可以看出，大模型增强对少样本投诉类型的识别效果提升最为显著。其中，对于原始训练样本仅50条的“类别1”，F1分数从0.31大幅提升至0.71，提升了0.4；对于原始训练样本100条的“类别2”，F1分数从0.48提升至0.64，提升了0.16；对于原始训练样本200条的“类别3”，F1分数从0.65提升至0.72，提升了0.07。这验证了本方案在解决少样本数据稀缺问题上的有效性，大模型生成的模拟训练数据显著增强了小模型对这些少样本投诉类别的学习能力。

对于样本数量相对较多的投诉类型（类别4～8），大模型增强方案也带来了不同程度的性能提升。例如，“类别7”和“类别8”的F1分数从0.75提升至0.77。尽管提升幅度不如少样本类别明显，但依然验证了该方案对整体识别准确率的积极作用。

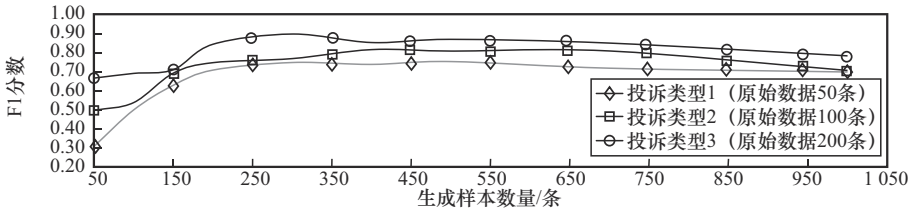
在验证了大模型增强的少样本学习算法对于网络投诉意图识别的有效性之后，本文进一步测试了大模型增强生成不同样本数量对于小模型网络投诉意图识别效果产生的影响。

实验二 只对3类少样本网络投诉进行大模型增强的少样本学习。本文通过协同学习迭代优化3种小样本网络投诉场景的训练数据，分别测试应用不同大模型（Qwen2.5 14B和GLM-4-9B），依次增强生成50条、100条、200条直到1 000条模拟数据后，让ERNIE 3.0-medium基于模拟训练数据和真实数据进行二次训练，并将二次训练结果与ERNIE 3.0-medium基于原始数据训练的效果进行对比分析。增强生成不同样本数量的小模型网络投诉意图识别效果如图3所示。

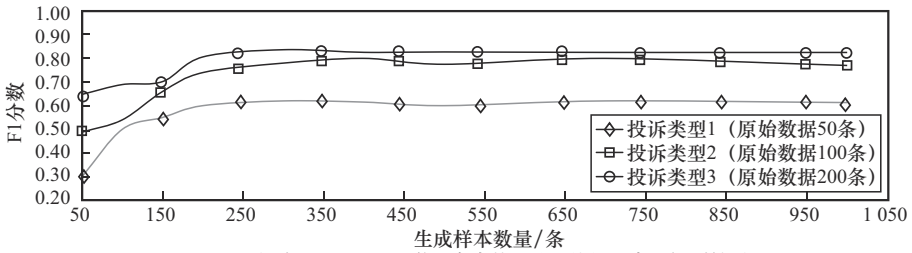
从图3（a）可看出大模型生成样本数量对少样本网络投诉意图识别性能的影响。对于3类少样本投诉类型，随着大模型生成样本数量的增加，F1分数呈现出先上升后趋于平稳甚至略有下降的趋势。对于“投诉类型1”，当生成样本数量达到500条时，F1分数达到0.76的峰值。对于“投诉类型2”，当生成样本数量达到500条时，F1分数达到0.85的峰值。对于“投诉类型3”，当生成样本数量达到300条时，F1分数达到0.90的峰值。从图3（b）中GLM-4-9B的增强生成效果整体来看，略差于Qwen 2.5 14B，但是其生成数据数量对于投诉识别准确性的影响也有相似的趋

表2 实验一网络投诉意图识别效果统计结果

投诉类型	原始训练样本数量	测试样本数量	小模型少样本学习				大数据增强+小模型少样本学习			
			准确分类的样本数量	precision	recall	F1分数	准确分类的样本数量	precision	recall	F1分数
类别1	50	6	2	28.57%	33.33%	0.31	5	62.50%	83.33%	0.71
类别2	100	11	8	37.50%	66.67%	0.48	10	50.00%	88.89%	0.64
类别3	200	22	12	80.00%	55.17%	0.65	14	85.71%	62.07%	0.72
类别4	300	33	21	62.50%	64.10%	0.63	21	67.57%	64.10%	0.66
类别5	400	44	42	93.33%	87.50%	0.90	40	93.48%	89.58%	0.92
类别6	500	55	48	84.21%	87.27%	0.86	50	81.97%	90.91%	0.86
类别7	1 000	112	76	73.08%	76.00%	0.75	84	79.79%	75.00%	0.77
类别8	1 500	167	114	79.72%	70.81%	0.75	117	84.96%	70.19%	0.77



（b）应用Qwen 2.5-14B增强生成数据的网络投诉意图识别效果



（b）应用GLM-4-9B增强生成数据的网络投诉意图识别效果

图3　增强生成不同样本数量的小模型网络投诉意图识别效果

势，对于“投诉类型1”，当生成样本数量达到300条时，F1分数达到0.62的峰值。对于“投诉类型2”，当生成样本数量达到400条时，F1分数达到0.80的峰值。对于“投诉类型3”，当生成样本数量达到300条时，F1分数达到0.84的峰值。

由实验二可以发现，即使在大模型增强生成样本数量较少的情况下（如生成50条、100条），相较于原始小模型少样本学习（对比表1中类别1、类别2、类别3的初始识别效果），F1分数也有提升，这进一步证明了大模型增强的有效性。然而，当生成样本数量过多，F1分数出现了轻微下降，这可能是因为过多的合成数据引入了噪声或导致模型过拟合于合成数据而非真实数据。同时，实验二应用两种大模型增强生成数据，均实现了网络投诉意图识别准确率的提升，这也证明了本文提出的基于大模型增强的少样本网络投诉意图识别算法在真实测试集上的泛化能力。

4　结束语

本文针对网络投诉分析场景中普遍存在的真实训练样本不足及各类型样本数量分布不均问题，提出了一种基于大模型增强的少样本网络投诉意图识别方案。该方案利用大模型强大的文本生成能力，迭代优化并扩充模拟训练样本，显著提升了网络投诉意图识别小模型的少样本学习效果和识别准确率。实验结果表明，所提出的方案在解决少样本数据稀缺问题上表现出显著优势，本方案对于3类少样本网络投诉的意图识别准确率提升了21%，并实现了全部8类投诉识别准确率9%的整体提升，这验证了通过大模型生成高质量模拟训练数据能够有效增强小模型对稀有投诉类别的学习能力。同时，实验也证明了该方案对整体投诉意图识别准确率的积极作用。此外，本文深入探究了大模型生成样本数量对少样本投诉意图识别性能的影响，发现存在一个最优的样本生成量，过少则提升有限，过多则可能引入噪声或导致模型过拟合，从而影响泛化能力。尽管本文提出的方案取得了良好效果，但仍存在进一步优化的空间和未来研究方向。

（1）更精细的样本生成策略。当前的样本生成策略可能仍有改进余地。未来可以探索更智能、更精细的样本生成机制，如根据小模型对不同类别样本的识别难易程度，动态调整大模型生成样本的数量和侧重方向，以实现更高效的样本

扩充。

（2）模型可解释性研究。随着大模型在少样本学习中的应用，模型的可解释性变得尤为重要。未来的工作可以致力于研究如何更好地理解大模型生成模拟样本的内在逻辑，以及这些合成样本如何影响小模型的决策过程，从而提高模型的可信度和应用透明度。

（3）实时性与效率优化。在实际部署中，模型的实时性和效率至关重要。未来的研究可以关注如何进一步优化大模型增强生成模拟样本的效率，如通过知识蒸馏、模型剪枝等技术，在保证性能的同时，降低模型的计算开销和推理时延。

参考文献：

[1] OUYANG Y, WANG L L, YANG A D, et al. Next decade of telecommunications artificial intelligence[J]. CAAI Artificial Intelligence Research, 2022, 1(1): 28-53.

[2] 曾伟, 钟检荣, 张玮, 等. 基于机器学习的5G用户智能投诉处理方案研究[J]. 邮电设计技术, 2021(5): 43-48.
ZENG W, ZHONG J R, ZHANG W, et al. Research on intelligent complaint handling scheme of 5G users based on machine learning[J]. Designing Techniques of Posts and Telecommunications, 2021(5): 43-48.

[3] 蒋燕, 程浩辉, 何丹. 基于BERT算法的通信投诉智能处理探索[J]. 广东通信技术, 2024, 44(2): 52-55.
JIANG Y, LIANG H H, HE D. Exploring intelligent processing of communication complaints based on the BERT algorithm[J]. Guangdong Communication Technology, 2024, 44(2): 52-55.

[4] 亚信科技, 清华大学智能产业研究院. AIGC(GPT-4)赋能通信行业应用白皮书[R]. 2023.
AsiaInfo, Institute for AI Industry Research, Tsinghua University. A white paper of AIGC (GPT-4) empowering telecom sector[R]. 2023.

[5] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[J]. arXiv preprint, 2017: 1703.03400.

[6] ZHANG H X, ZHANG X F, HUANG H B, et al. Prompt-based meta-learning for few-shot text classification[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 1342-1357.

[7] EDWARDS A, USHIO A, JOSÉ CAMACHO-COLLADOS, et al. Guiding generative language models for data augmentation in few-shot text classification[J]. arXiv preprint, 2021: 2111.09064.

[8] BAYER M, KAUFHOLD M A, REUTER C. A survey on data augmentation for text classification[J]. ACM Computing Surveys, 2023, 55(7): 1-39.

[9] GAO J, LYU S Q, LIU G R. A hybrid model for few-shot text classification using transfer and meta-learning[J]. arXiv preprint, 2025: 2502.09086.

[10] BRAGG J, COHAN A, LO K, et al. FLEX: unifying evaluation for few-shot NLP[J]. arXiv preprint, 2021: 2107.07170.

[11] AGARWAL R, VIEILLARD N, ZHOU Y, et al. On-policy distillation of language models: learning from self-generated mistakes[J]. arXiv preprint, 2024: 2306.13649.

[12] LIANG C, ZUO S M, ZHANG Q R, et al. Less is more: task-aware layer-wise distillation for language model compression[C]//Proceedings of the 40th International Conference on Machine Learning. Hawaii: ICML, 2023: 20852-20867.

[13] ZOU T, LIU Y, LI P, et al. Fusegen: PLM fusion for data-generation based zero-shot learning[J]. arXiv preprint, 2024: 2406.12527.

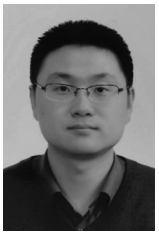
[14] YAO Z W, YAZDANI AMINABADI R, ZHANG M J, et al. Zeroquant: efficient and affordable posttraining quantization for large-scale transformers[J]. Advances in Neural Information Processing Systems, 2022(35): 27168-27183.

[15] CHENG Y, ZHANG L, LI A. GFL: federated learning on non-IID data via privacy-preserving synthetic data[C]//Proceedings of 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom). Piscataway: IEEE Press, 2023: 61-70.

[16] ZHANG Z, YANG Y H, DAI Y, et al. FedPETuning: when federated learning meets the parameter-efficient tuning methods of pre-trained language models[C]//Findings of the Association for Computational Linguistics (ACL), 2023: 9963-9977.

[作者简介]

顾宁伦（1972– ），男，中国移动通信集团有限公司网络事业部副总经理、客户响应中心主任、一级技术副总师、正高级工程师，主要研究方向为网络、IT运营管理以及数智化转型。



罗志毅（1982– ），男，中国移动通信集团有限公司网络事业部项目经理、高级工程师，主要研究方向为网络运营、IT 系统规划、自智网络产业推进等。

王希栋（1984– ），男，亚信科技（中国）有限公司通信人工智能实验室算法专家，主要研究方向为无线通信、网络智能化、大语言模型。

丁建兵（1993– ），男，亚信科技（中国）有限公司通信人工智能实验室算法专家，主要研究方向为通信网络人工智能，包括大语言模型、机器学习和深度学习。

欧阳晔（1981– ），男，博士，亚信科技（中国）有限公司首席技术官、高级副总裁，IEEE Fellow，主要研究方向为移动通信、数据科学与人工智能跨学科领域的创新与管理。