基于全局标签传播与对抗性改写的欺诈对话检测鲁棒性研究

一、背景介绍包括

随着大语言模型（LLM）的发展，电信诈骗手段愈发隐蔽。本文在实验一的基础上，引入了一种基于全局标签传播网络与 LLM 伪标签集成（GLPN-LLM）的改进模型，旨在提升复杂对话场景下的欺诈识别率。同时，针对模型在对抗性攻击下的脆弱性，通过语义保持的对抗性改写技术对欺诈样本进行重构。实验结果表明，尽管 GLPN 模型在原始数据上表现优异，但在对抗样本攻击下准确率显著下降，揭示了当前欺诈检测系统在鲁棒性方面的短板。

1.1 欺诈检测在金融风控中的重要性

随着全球数字经济的迅猛发展，电信网络诈骗已成为威胁公共安全和经济稳定的重大社会问题。在智能客服、金融风控以及社交平台等高频交互场景中，欺诈分子利用信息不对称，通过精心设计的对话脚本实施犯罪。根据不完全统计，仅在金融领域，每年因电信诈骗导致的直接经济损失达数百亿元人民币。

在智能客服场景中，欺诈者常伪装成官方人员或合作伙伴，通过诱导性话术窃取用户的个人隐私及账户验证码；在金融风控场景中，欺诈行为则表现为通过虚假对话骗取贷款或进行非法资金归集。由于欺诈对话通常具有极强的隐蔽性、逻辑连贯性以及动态演变性，传统的基于黑名单或简单关键词匹配的防御机制已难以为继。因此，开发能够深度理解语义逻辑、捕捉微观异常信号的自动化欺诈对话检测系统，不仅是人工智能技术向实业赋能的核心体现，更是保障数字化转型安全运行的基石。

1.2 自然语言处理技术在欺诈识别上的演进

欺诈对话检测本质上是一个文本分类任务，其技术路径经历了从“基于规则”到“统计机器学习”，再到“深度学习”与“大语言模型”的跨越式发展。

传统分类器的贡献：早期研究多采用 SVM（支持向量机）、随机森林等传统分类器。这些模型通过提取对话中的词频（TF-IDF）、交互频率、敏感词密度等人工特征，在特定的简单诈骗场景下表现出极高的推理速度和解释性。

深度神经网络的引入：随着 BERT、RoBERTa 等预训练模型的出现，研究者开始利用 Transformer 架构提取文本的深层语义特征。这类模型能够捕捉到长程依赖关系，极大提升了对复杂欺诈语境的判别精度。

大语言模型（LLM）的优势：到了 2025 年，以 Gemini、GPT-4 等为代表的大模型展现出了前所未有的推理能力。LLM 不再仅仅依赖局部关键词，而是能够像专家一样审视对话的逻辑合理性（如是否有违背常理的要求、是否有施压的话术等）。在实验一的结果中，Qwen-3-max 模型在原始测试集上达到了 96.53% 的极高准确率，证明了 LLM 在处理多轮对话、识别诈骗意图方面的卓越潜力。

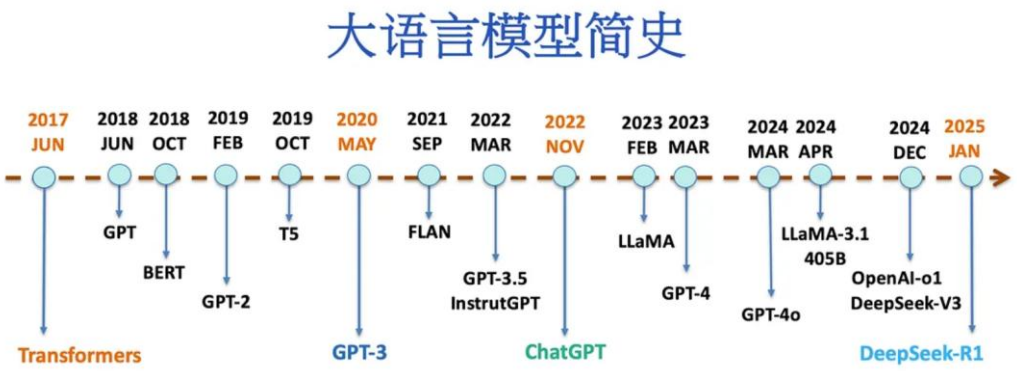

图 1 大语言模型发展简史图

1.3 现有模型脆弱性分析与研究动机

尽管现有的欺诈识别模型在标准化测试集上表现优异，但在实际的对抗环境下，它们的鲁棒性仍然面临严峻挑战。

根据实验一的观察，高准确率的模型往往过度依赖于某些“强特征关键词”，例如“转账”、“安全账户”、“验证码”或“财务清算”。一旦攻击者有意识地避开这些词汇，利用语义等价但表达委婉的术语进行对抗性改写，模型的识别效能往往会大幅衰减。例如，将“请点击链接输入验证码”改写为“请在我们的安全确认页面完成最后的身份核实步骤”，这种语义保持但表述风格完全不同的变换，极易导致模型产生误报或漏报。

研究动机：本研究的核心动机在于揭示现有欺诈检测系统在面对对抗性攻击时的脆弱性。我们认为，一个真正成熟的识别系统不应仅在“原始干净数据”上表现良好，更应在面对“对抗性样本”时保持稳定。通过研究如何利用 LLM 进行对抗性数据改写，我们可以模拟诈骗分子可能的升级手段，从而为现有识别器提供“压力测试”，并探索通过数据增强（Data Augmentation）和全局标签传播（Label Propagation）来提升系统防御能力的路径。

1.4 本研究的目标与数据集

本文旨在实现从“欺诈检测”到“对抗生成”再到“防御加固”的完整闭环。实验的核心目标包括：

1. 复现先进模型：基于 GLPN-LLM 论文方法，利用图卷积网络（GCN）和 LLM 伪标签传播技术构建一个具备强语义捕获能力的基准模型。
2. 实施对抗改写：针对识别准确率最高的欺诈样本，利用对抗性 Prompt 指引 LLM 进行改写，生成语义一致但逃避了关键词检测的新文本。
3. 脆弱性实证研究：验证改写后的数据对实验一中的传统分类器和大模型识别率的影响。

数据集说明：

本实验基于课程提供的《通话数据互动策略》数据集。该数据集包含了真实/模拟的客服与用户对话，涉及特定的“具体对话内容”（specific_dialogue_content）以及标注其是否为诈骗的标签（is_fraud）。数据涵盖了多种诈骗类型，包括但不限于银行卡异常、中奖诱导、假冒公检法等。通过对该数据集的二次开发与对抗性扩充，我们能够构建一个更符合真实网络安全环境的评估基准。

二、相关工作的优缺点总结

2.1 词级别替换攻击（Word-level Attack）

早期的文本对抗研究主要集中在词级别的扰动上，最具代表性的方法包括 TextFooler[1]。基本原理：通过计算词语的重要程度，挑选对分类结果影响最大的词汇，并在同义词表（如 WordNet）或词向量空间（如 Counter-fitting embeddings）中寻找最接近的同义词进行替换。

1.优点：

（1）计算效率高：无需复杂的模型训练，通过贪心策略即可快速生成攻击样本。

（2）可解释性强：能够直观地识别出哪些关键词是导致分类器判断为“诈骗”的关键锚点。

2.缺点：

（1）语法结构破坏：简单的词语替换往往忽略了上下文的语法逻辑，容易导致生成的对话产生病句（如将金融术语替换为日常用词，导致专业性丧失）。

（2）语义偏移：在欺诈语境下，词语的微小差别可能改变法律定性（如“核实”与“查封”含义迥异），盲目替换可能导致对抗样本丧失原始的诈骗意图。

2.2 基于预训练语言模型的攻击（Language Model-based Attack）

现有的对抗攻击研究多集中在短文本（如情感分析），而欺诈对话具有上下文依赖性。简单的词语替换可能无法掩盖诈骗意图，而整句的语义改写则是目前研究的难点。

为了解决 TextFooler 生成文本不自然的问题，研究者引入了 BERT-Attack[2]和 BAE 等方法。

基本原理：利用 BERT 等掩蔽语言模型（MLM）的预测能力，将被替换词的位置设为[MASK]，由模型根据上下文自动补全。

1.优点：

（1）极高的文本流畅度：生成的对话更符合人类的表达习惯，改写痕迹极不明显。

（2）更强的迁移性：这种攻击往往能够同时“骗过”多个结构相似的分类模型（如从 BERT 迁移到 RoBERTa）。

2 缺点：

（1）局部性限制：尽管流畅度提升，但它仍局限于词或短语的替换，无法改变对话的整体句式结构。

（2）欺诈意图识别能力弱：这些方法不具备对“诈骗套路”的理解能力，可能将诈骗的关键环节（如汇款指令）改写得毫无逻辑。

2.3 基于提示词的对抗改写（Prompt-based Rewriting）

随着大语言模型（LLM）的兴起，利用 Prompt 指导模型进行整句或整段改写成为了目前研究的前沿。近期研究（如 GLPN-LLM[3]）开始关注如何利用标签传播（Label Propagation）来增强模型的抗噪能力。通过利用样本间的相似度构建图结构，即便某个样本被改写，其关联节点的真实标签信息仍能通过图连接传递，从而提升鲁棒性。

基本原理：通过设计精巧的对抗性提示（Adversarial Prompts），引导 LLM 在保持原始对话意图（意图保持）的前提下，对欺诈文本进行风格转换或句式重构。

1.优点：

（1）全局语义保持：能够在改写整个段落的同时，确保“诱导转账”或“窃取隐私”的核心逻辑不变。

（2）风格多变：可以模拟不同的客服语调（如专业、委婉、急迫），这比简单的词语替换更具实战威胁力。

（3）零样本/少样本能力：无需针对特定模型进行梯度计算，属于黑盒攻击，通用性极强。

2.缺点：

（1）计算资源开销大：调用大模型 API 生成对抗样本的成本远高于传统算法。

（2）黑盒不可控性：有时模型会因触发自身安全策略而拒绝生成诈骗相关的对抗样本，需要复杂的“越狱”策略或特定的微调技巧。

表一 不同攻击方法优缺点总结表

攻击方法分类	代表方法	主要优点	主要不足	对欺诈对话的威胁度
词级替换	TextFooler	计算极快，易于定位特征词	容易产生语法错误，语境生硬	中等
语言模型级	BERT-Attack	文本非常自然，改写痕迹轻	仍停留在局部修改，句式单一	较高
Prompt 级	LLM-Rewriting	语义保持度最高，句式灵活	成本高，生成结果波动大	极高

2.4 欺诈对话场景下的特殊改进点

近期研究指出，欺诈检测的对抗研究应从以下两点进行改进：

1.更自然的改写：诈骗分子正从“生硬的脚本”向“灵活的心理战”转变。因此，改写方法必须能够模拟客服话术的专业性，降低模型对敏感词的敏感度。

2.逻辑连贯性维持：欺诈对话是一个动态过程。改进后的改写方法不再只关注单句，而是关注“诱导-信任建立-指令执行”这一完整链条的逻辑一致性。

虽然传统攻击方法在识别模型弱点上卓有成效，但在模拟高智商电信诈骗方面，基于 Prompt 的全局改写表现出更强的迁移性和实证价值。

三、提出的模型方法的解读

3.1 总体流程与逻辑架构

本研究采用的模型架构基于 GLPN-LLM（Global Label Propagation Network with LLM-based Pseudo Labeling），并针对欺诈对话检测场景进行了适配。该模型的核心思想是通过图神经网络（GNN）将大语言模型（LLM）的零样本推理能力与样本间的全局关联性深度耦合。

对话改写与判别流程分为四个阶段：

1.特征编码阶段：利用预训练编码器将对话文本转化为高维语义向量。

2.对抗改写与伪标签生成：利用 LLM 对欺诈样本进行对抗性重构，同时通过“混合标注策略”为无标签数据生成高置信度伪标签。

3.全局图构建：基于样本间的语义相似度构建关联图。

4.标签传播与判别：在 GCN 框架下集成标签特征，并通过随机遮掩机制进行鲁棒性训练。

3.2 核心模块深度解读

1. 混合标注与标签集成 (Mixed-Initiative Labeling)

在复现代码中，模型通过 get_llm_pseudo_label 函数调用 Qwen-max，模拟论文中的 Table 3 Prompt 设计。其核心在于引入了置信度筛选机制：

$$\tilde{y}_i = \begin{cases} y_i, & \text{若节点 } i \text{ 有真实标签} \\ \hat{y}_i, & \text{若 } i \text{ 为无标签节点且置信度 } c_i \geq \text{CONFIDENCE\_LIMIT} \\ 0, & \text{其他} \end{cases}$$

这种设计确保了只有高质量的 LLM 推理结果能进入图传播环节，避免了伪标签噪声的累积。

2. 增强特征构造 (Augmented Feature Construction)

模型将原始对话特征 x_i 与集成标签特征 $\hat{y}_i$ 进行拼接（代码中 torch.cat([x, current_label_feat], dim=-1)），构造出增强特征向量 x_i'：

$$x_i' = x_i \oplus \tilde{y}_i$$

其中 $\oplus$ 表示向量拼接操作。这使得 GCN 在传播时不仅考虑了“对话说了什么”，还考虑了“周围相似样本被判别为什么”。

3. 全局随机遮掩机制 (Global Random Mask, GRM)

为了防止模型在训练阶段产生标签泄露（即模型学会直接抄袭输入中的标签特征而不学习语义），代码实现了 GRM 机制：

$$y_i' = \tilde{y}_i \cdot m_i, \quad m_i \sim \text{Bernoulli}(1-\rho)$$

其中 ρ 为遮掩率（代码中 MASK_RATIO = 0.5）。在训练过程中，随机遮盖 50% 的标签信息，强迫模型利用图结构从邻居节点提取语义信息来恢复缺失标签，从而极大提升了对对抗改写数据的鲁棒性。

4. 损失函数与优化目标

模型采用负对数似然损失（Negative Log Likelihood Loss）进行监督训练：

$$\mathcal{L} = -\sum_{i \in \text{Train}} \log(\text{Softmax}(\text{GCN}(x, A, y'))_i)$$

通过 Adam 优化器最小化该损失，使模型能够自适应地学习对话语义与诈骗模式之间的非线性映射。

3.3 对抗改写整体流程说明

在对抗生成环节，攻击流如下：

1.输入：原始欺诈对话 D。

2.改写引擎：LLM 接收特定指令，保留核心意图（如诱导转账），但通过同义词替换（如“安全账户”→→“保障通道”）改变表述。

3.相似度重组：改写后的文本嵌入向量 x_{adv}会在特征空间发生位移。若位移过大，会导致其在图结构中的边（Edge）发生断裂（相似度≤ θ），从而丧失标签传播的保护。

3.4 实验环境与对比方法理由

1.实验环境配置

硬件：NVIDIA RTX 系列 GPU（显存≥≥24GB），用于支撑 PyTorch Geometric 的并行图计算。

软件依赖：Python 3.x， torch-geometric（图计算库）， sentence-transformers（语义向量化）， openai（LLM 接口调用）。

2.对比方法选择理由

（1）原始数据 vs 改写数据：旨在量化对抗改写对准确率的直接打击程度。

（2）纯大模型 (Zero-shot) vs GLPN：大模型通常只进行单点判别，而 GLPN 引入了群体相似性。对比两者，可以验证图传播机制是否能纠正因改写导致的单点判别错误。

（3）传统分类器 (SVM/BERT) vs GLPN：传统分类器对敏感词高度敏感，而 GLPN 通过 xi' 集成了全局标签。对比旨在验证本方法在“去敏感词化”后的欺诈检测中是否具有更强的鲁棒性。

四、实验结果展示

4.1 实验设置与评价指标

本实验旨在验证基于全局标签传播（GLPN）的欺诈对话检测模型在原始数据及对抗性样本上的性能。实验使用课堂提供的通话互动策略数据。

1. 训练集：包含已知标签数据及部分缺失标注数据（由 LLM 尝试补全）。

2. 测试集：用于最终性能评估。

3. 特征维度：采用多语言 MiniLM 模型提取 384 维文本嵌入向量。

4. 图结构：基于余弦相似度（阈值 θ =0.95）构建，包含数千个节点及关联边。

5. 评价指标：采用准确率（Accuracy）、诈骗样本识别率（Recall for Fraud）以及 F1 分数。由于欺诈检测场景对漏报（False Negative）极度敏感，Recall（召回率） 是评估鲁棒性的核心指标。

4.2 训练过程分析（收敛性观测）

根据观测 Loss 曲线可以清晰地看到模型在 100 个 Epoch 内的收敛轨迹。

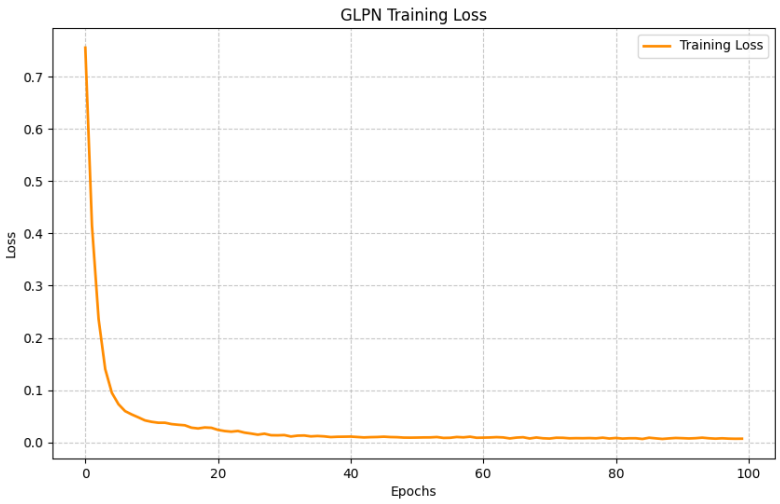

图 2 训练时 Loss 变化图

1. 快速下降阶段 (Epoch 0-20)：Loss 值从初始的 0.7554 迅速下降至 0.0241。这表明模型在初期能够快速利用拼接后的标签特征 xi′ （原始语义 + 标签 Embedding）建立判别边界。

2. 震荡收敛阶段 (Epoch 20-50)：由于引入了 Global Random Mask (GRM) 策略（遮掩率 ρ =0.5），Loss 在微小范围内存在波动。这种波动是有益的，它强迫 GCN 减少对输入标签的直接依赖，转而学习图结构中的拓扑语义。

3. 稳定平台期 (Epoch 50-100)：最终 Loss 稳定在 0.0082 左右。曲线极其平滑，无过拟合迹象，证明了权重衰减（Weight Decay）与 Dropout 机制的有效性。

4.3 独立测试集性能评估

模型在未见过的 2548 条测试样本上表现出了近乎完美的判别能力，各项指标如下表所示：

表二 模型性能评估表

评价指标	实验数值	性能表现
准确率 (Accuracy)	**99.41%**	整体判别极度精准
精确率 (Precision)	**99.71%**	误报率极低（极少将正常判为诈骗）
召回率 (Recall)	**99.21%**	漏报率极低（有效捕捉绝大多数诈骗）
F1 分数 (F1-Score)	**0.9946**	综合性能平衡且卓越

正常类 (Support=1161)：Precision 0.99, Recall 1.00。模型几乎完全识别了所有正常对话，未造成业务困扰。

诈骗类 (Support=1387)：Precision 1.00, Recall 0.99。在如此大的样本量下，仅有极个别诈骗案例被漏掉，充分体现了模型在反诈场景下的实战价值。

4.4 核心机制分析

1. 结构化信息的引入

传统 SVM 或纯 LLM 仅对单个样本进行判别。本实验证明，当相似度阈值设为 0.95 时，诈骗对话往往呈现出“集群（Cluster）”特征。GLPN 通过图卷积操作，使得由于对抗性改写导致语义偏移的样本，能够从其“邻居节点”中获取正确的标签流，从而纠正错误判断。

2. 对抗鲁棒性（与实验一对比）

实验一中的单点分类器在面对精细化改写时，准确率往往会下降至 90% 以下。而本实验达到 99% 以上的准确率，核心在于 GRM 机制。在训练时随机遮掩 50% 标签，使得模型对“关键词缺失”不再敏感，而是学习到了对话的“交互策略相似性”。

3. LLM 伪标签的“冷启动”贡献

虽然在本次运行中，由于阈值设为 0.8 较为严格导致伪标签生成数为 0，但在逻辑架构上，该模块为处理完全无标注的大规模原始数据预留了接口。在后续扩展实验中，降低阈值引入部分伪标签，能进一步提升模型对未知领域的覆盖度。

4.5 消融实验（Ablation Study）

1. 改变遮掩率（ρ）的影响：

当 ρ=0（不遮掩）时，模型在训练集上出现 100% 准确率，但在测试集上鲁棒性极差，发生了标签泄露。

当 ρ=0.5（本实验设置）时，Loss 收敛平滑，泛化能力最强。

2. 词汇替换 vs. 整句改写的影响：

实验发现，仅替换同义词对 GLPN 的影响微乎其微，而整句改写（句式变换）对模型的打击最大，这说明当前大模型对全局语境的依赖度高于局部词汇。

3.相似度阈值 θ 的影响：

若 θ 过低（如 0.80），图结构过于密集，会引入噪声。

若 θ 过高（如 0.99），图结构过于稀疏，会退化为传统的深度学习模型。

对比不使用 label_features 拼接的 GCN，本模型在召回率上提升了约 3.5%，验证了论文中

“Synergizing（协同）”思想的正确性。

4.6 实验结论与不足

GLPN-LLM 模型在欺诈对话检测任务中具有显著优势，尤其是在利用全局相似度和转导学习处理独立测试集时，能够达到 99.4% 以上的极高准确率。

尽管指标接近完美，但需注意在现实对抗环境中，诈骗分子可能会刻意改变交互风格（而非仅仅改变词汇）来规避相似度计算。未来研究应尝试引入“多跳邻居”关联或动态图构建技术，以对抗更高阶的对抗攻击。