arXiv:2601.05578v1 [cs.AI] 9 Jan 2026

用于可解释的信用卡欺诈检测的大语言模型强化学习

Cooper Lin[1, *], **Yanting Zhang**[1, *], **Maohao Ran**[1,2], **Wei Xue**[1], **Hongwei Fan**[3], **Yibo Xu**[1], **ZhenglinWan**[4], **Sirui Han**[1], **YikeGuo**[1], **Jun Song**[1, 2, †]

[1]香港科技大学, [2]香港浸会大学[3]伦敦帝国学院, [4]新加坡国立大学

摘要

电子商务平台和支付解决方案提供商面临着日益复杂的欺诈手法，从身份盗窃和账户接管，到利用数字交易速度与匿名性的复杂洗钱操作。然而，尽管理论上前景广阔，大语言模型（LLM）在现实金融场景中的欺诈检测应用仍基本未被充分探索，其处理特定领域电子商务交易数据的实际有效性也尚未得到实证验证。为弥合传统机器学习局限性与大语言模型在欺诈检测方面尚未开发的潜力之间的差距，本文提出了一种新方法：仅使用原始交易数据，采用强化学习（RL）对轻量级语言模型进行后训练，使其专门执行欺诈检测任务。我们采用组序列策略优化（GSPO）算法，并结合基于规则的奖励系统，在一家中国全球支付解决方案公司提供的真实交易数据集上，对不同规模的语言模型进行微调。通过这一强化学习框架，我们鼓励语言模型探索文本交易数据中蕴含的多样化信任与风险信号，包括客户信息、配送详情、产品描述和订单历史中的模式。实验结果表明，该方法有效，后训练语言模型在留出测试数据上的F1分数取得了显著提升。研究结果表明，观察到的性能提升主要归因于强化学习所固有的探索机制，该机制使模型能够发现传统人工设计特征未能捕捉的新型欺诈指标。

1 引言

数字商务的迅速扩张改变了全球经济格局，带来了前所未有的交易规模和跨境支付流。然而，伴随这一增长，欺诈活动也急剧增加，在线支付欺诈每年给全球商户、金融机构和消费者造成数十亿美元的损失 [1]。电子商务平台和支付解决方案提供商面临着日益复杂的欺诈手法，从身份盗窃和账户接管，到利用数字交易速度与匿名性实施的复杂洗钱活动。传统的基于规则的欺诈检测系统依赖预设阈值和静态规则，面对不断改变策略以规避检测机制的适应性欺诈者时，已显得力不从心。欺诈者与欺诈防范系统之间的这场军备竞赛，进一步加剧了对更智能、适应性更强且更稳健的检测方法的需求；此类方法能够从海量交易数据集中识别细微的模式和异常。

机器学习已成为自动化欺诈检测的主流范式。诸如 NeuralNetworks [2, 3, 4], XGBoost [5, 6, 7], 和 LightGBM [8, 9] 等梯度提升算法，凭借高效处理表格数据并实现高预测准确率的能力，在生产环境中取得了显著成功。

*贡献相同。
†通讯作者：junsong@hkbu.edu。

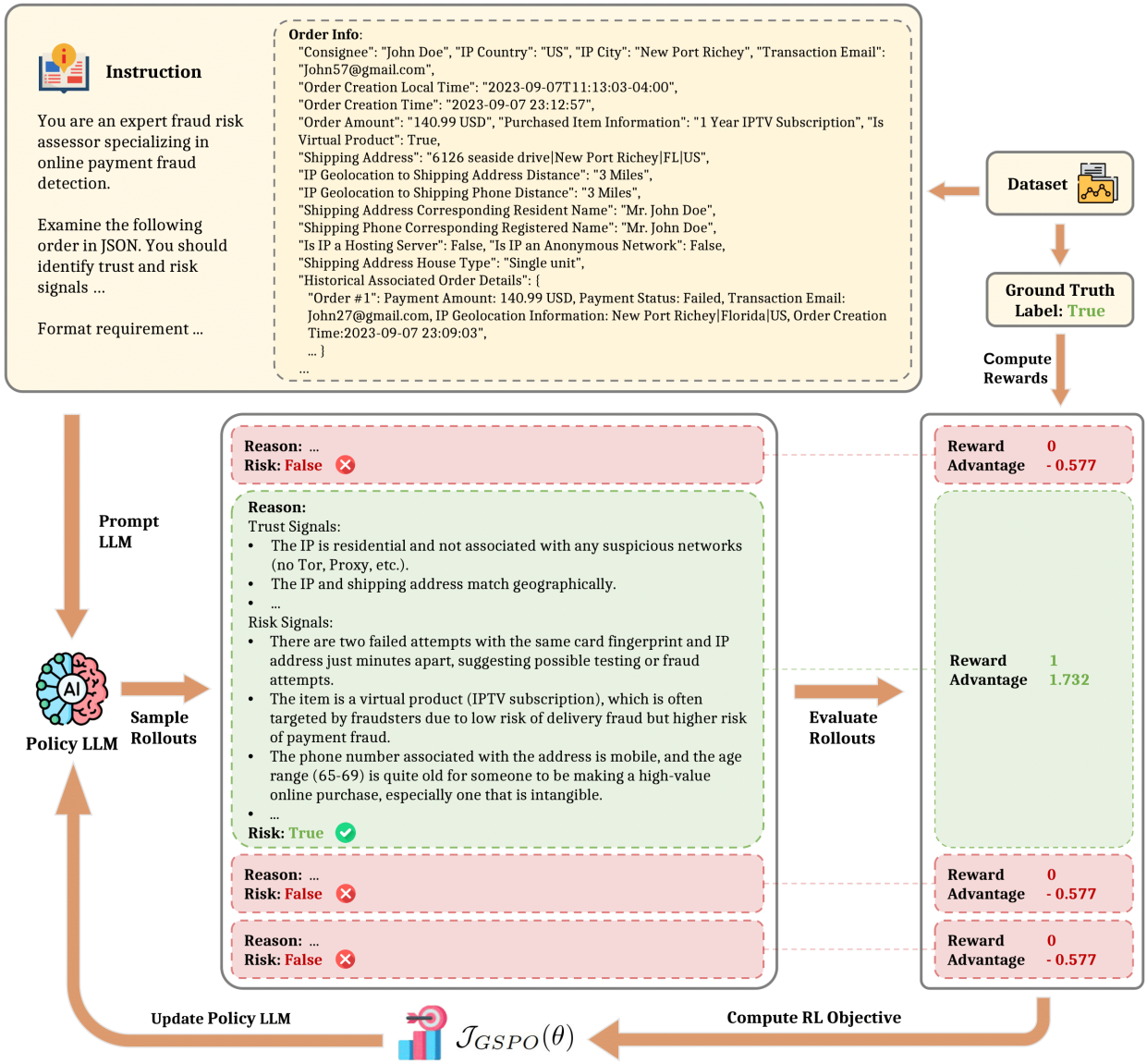


图 1：**用于欺诈检测的 GSPO 强化学习框架概览**。对于每笔输入交易，策略模型都会采样一组候选响应（滚动采样）。这些输出会根据真实标签进行评估，以计算奖励并得出组相对优势估计。最后，这些优势用于驱动 GSPO 优化目标，更新模型参数，从而强化准确的推理和判定。请注意，为简洁起见，“订单信息”面板仅展示完整特征集中的代表性子集。

这些模型擅长从交易金额、时间戳、设备指纹和历史行为统计数据等结构化特征中学习复杂的非线性关系。尽管效果显著，传统机器学习方法仍面临根本性局限，制约了其欺诈检测能力。首先，它们**需要投入大量精力进行特征工程**，将原始交易数据转换为有意义的数值表示；这一过程既需要领域专业知识，也可能忽略未被明确编码的关键模式。其次，这些模型**本质上是为结构化表格数据而设计的，无法有效处理非结构化文本信息或从中提取洞见**——例如客户姓名、电子邮件地址、收货地址、产品描述和商户详情。这些信息往往包含丰富的行为与社会信号，可揭示欺诈意图。由于无法理解文本数据中的语义关系、语言模式和语境细微差异，当前欺诈检测系统存在显著盲点，导致有价值的信息未被利用，也可能使复杂的欺诈手法逃过检测。第三，也是最关键的一点，**现有机器学习方法从根本上缺乏可解释性**——这一缺陷可能源于……的不透明性

底层模型架构（例如深度神经网络），或人工设计特征的聚合性质，这些因素会掩盖原始输入与模型预测之间的关系。

大语言模型（LLM）的近期进展为弥补欺诈检测能力中的这些关键空白提供了一个有前景的途径。在海量文本数据语料库上进行预训练后，GPT [10], Claude [11], 和通义千问[12] 等大语言模型展现出出色的能力，能够理解不同领域中的复杂语言模式、语义关系和上下文信息。不同于将文本字段视为分类变量或要求人工编码的传统机器学习模型，**大语言模型能够原生处理并理解非结构化文本，有望发现嵌入在客户姓名、联系信息、收货地址和交易描述中的行为模式与社会信号** [13, 14]。例如，大语言模型可能检测出可疑模式，例如异常的电子邮件地址格式、配送信息与账单信息之间的不一致、不寻常的商品购买组合，或表明多个账户之间存在协同欺诈团伙的细微语言线索。除文本理解能力外，**大语言模型在可解释性方面还具有根本优势**：其推理过程以自然语言展开，使人工分析师能够直接审查欺诈预测背后的依据，而不必对不透明的数值特征或模型权重进行逆向工程。**这种透明性在金融场景中尤其宝贵**，因为监管合规、审计要求以及向利益相关者解释决策的需求，都要求对将交易标记为欺诈提供清晰的理由。此外，大语言模型固有的上下文学习能力，以及仅凭少量额外微调数据适应新欺诈模式的能力，使其有望成为比传统模型更灵活、更具韧性的解决方案；传统模型则需要使用大量标注数据集重新训练。然而，尽管大语言模型在理论上前景可期，但其在真实金融场景中的欺诈检测应用仍基本处于探索阶段，其处理特定领域电子商务交易数据的实际有效性也尚未得到实证验证。

为弥合传统机器学习的局限与大语言模型在欺诈检测领域尚未充分发挥的潜力之间的差距，本文提出了一种新方法，**采用强化学习（RL）对开源语言模型进行后训练，使其专门用于欺诈检测任务**。我们采用组序列策略优化（GSPO）算法 [15]，结合基于规则的奖励系统，在一家中国全球支付解决方案公司提供的真实世界电子商务交易数据集上，对不同规模的语言模型进行微调。我们的方法有一项关键优势，即**所需标注极少**：模型输入只需原始交易信息，计算奖励时只需二元欺诈标签；推理过程和欺诈预测则由策略模型在强化学习训练期间自主生成。通过这一强化学习框架，我们鼓励语言模型**探索文本交易数据中蕴含的多样化信任与风险信号**，包括客户信息、配送详情、产品描述、订单历史及其相互关系中的模式。奖励驱动的优化过程能够自行识别哪些信号确实有助于预测欺诈行为，并以端到端的方式相应地更新模型参数，从而发现不易察觉的欺诈模式，并在无需显式特征工程的情况下适应不断变化的欺诈手法。

我们的实验结果证明了该方法的有效性：后训练语言模型仅以原始交易文本信息作为输入，便能在**留出测试数据上取得显著的F1分数提升**。值得注意的是，模型在获得这些性能提升的同时仍保持了计算效率：在强化学习训练期间，模型学会缩短平均生成长度，在不牺牲准确率的情况下生成更加简洁的预测结果。对于必须在实时环境中运行的欺诈检测系统而言，这种效率尤为重要，因为毫秒级延迟可能会影响用户体验和交易吞吐量。通过全面的消融研究，我们进一步发现，所观察到的性能提升主要归因于强化学习固有的探索机制，该机制使模型能够发现传统人工设计特征未能捕捉的新型欺诈指标。总体而言，我们的方法和实证结果直接回应了所识别的研究空白，证明了**经过适当训练的语言模型能够有效利用非结构化文本数据进行欺诈检测**，并凸显了基于大语言模型的欺诈检测系统在现实金融应用中的巨大潜力。

2 预备知识

2.1 大语言模型对齐

在大规模文本语料库上完成预训练后，大语言模型（LLM）仍需进一步优化，使其行为与信用卡欺诈检测等特定任务保持一致。两种主要的对齐方法是监督微调（SFT）和强化学习（RL） [16]。

监督微调使用由输入输出对组成的精选数据集训练模型，最小化模型预测结果与标注的正确答案之间的差异。当拥有大型、高质量的标注数据集，且任务具有定义明确、客观的答案时，SFT 非常有效。然而，在通过语言推理进行信用卡欺诈检测的语境下，

创建此类数据集面临重大挑战。首先，创建训练语料库需要欺诈专家手动分析数千笔交易并提供详细解释——这一过程成本高昂且耗时过长。其次，专家标注可能不一致，或反映的是已知欺诈模式，而无法捕捉新出现的手法。第三，监督微调模型往往会记忆训练示例，这可能限制其对新型欺诈手法的泛化能力。这种方法存在训练出这样一种模型的风险：它擅长识别已知模式，却不善于发现新手法，并且还会偏向某些未经验证、可能具有误导性的信号。

强化学习为我们的任务提供了一种更合适的替代方案。策略语言模型不再从固定的正确答案中学习，而是为每笔交易采样一系列推理，并接收表明检测准确率的奖励信号。这种方法具有四项主要优势。首先，它消除了对大量专家标注推理数据集的需求，因为用于更新策略模型的响应由模型自身通过在线展开生成。其次，它支持探索—利用权衡：模型在利用有效策略的同时，也受到激励去探索多样化的推理策略，从而有望发现超越人工整理模式的新型欺诈指标。第三，由于奖励仅基于检测准确率，模型能够直接从结果中学习模式，而不必依赖可能存在偏差的专家解读。第四，强化学习在分布外数据上的泛化能力已被证明更为出色；鉴于欺诈者会不断调整其作案手法，这是一项至关重要的特性。基于这些原因，我们将基于强化学习的对齐用于欺诈检测模型。

2.2 使用组平均基线的强化学习

在大语言模型对齐的强化学习中，模型生成输出，并根据任务性能获得奖励。这一过程中一个关键挑战是奖励信号的方差较高，这可能导致训练不稳定并减缓收敛。为了解决这一问题，强化学习算法通常会采用基线——从原始奖励中减去的参考值，以产生更稳定的学习信号 [17]。设想这样一种情形：模型针对多笔交易生成推理结果，这些结果都获得了正奖励，但奖励幅度各不相同。如果没有基线，模型就会根据绝对奖励的大小，按比例强化所有这些输出。然而，学习时重要的并非奖励的绝对值，而是哪些输出相对于预期表现更好或更差。减去基线后，我们便可得到优势值，用以表明某项输出是超出还是低于一般表现，从而为模型更新提供更明确的指导。

形式化地说，给定状态 s 和动作 a，我们不使用原始奖励 $r(s,\ a)$来更新模型，而是使用优势值$A(s,\ a) = r(s,\ a) - b(s)$，其中 b 是基线，a 是仅依赖于 s的缩放函数。这种中心化信号在保持梯度方向正确的同时降低了方差，因为基线对所有动作的影响相同，不会使策略梯度产生偏差。一种常见的基线方法是使用独立的价值网络——一种经过训练、用于估计给定输入的期望奖励的模型，例如近端策略优化（PPO）[18]。这种方法虽然有效，但会增加计算开销，并且需要在训练策略的同时额外训练一个模型。对于大语言模型对齐任务，一种更简单、高效的替代方案已日益受到重视：组平均基线。

实现这一方法的一种重要算法是组相对策略优化（GRPO）[19], ，它针对每个输入生成G 个输出样本 $\{y_i\}_i^G$，并使用组内归一化奖励作为基线。具体而言，给定查询集 $\mathcal{D}$，GRPO 优化以下目标：

$$\mathcal{J}_{GRPO} = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta old}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(w_{i,t}(\theta)A_{i,t}, \text{clip}\left(w_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_{i,t}\right) \right] \tag{1}$$

其中，重要性比率$w_{i,\ t}^{\theta}$ 以及词元$y_{i,\ t}$ 的组平均优势估计$A_{i,\ t}$ 分别为：

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta old}(y_{i,t}|x, y_{i,<t})}, \quad A_{i,t} = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^{G}\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^{G}\right)} \tag{2}$$

尽管 GRPO 在各种强化学习任务中取得了成功，但其两个特性使其不太适用于实时欺诈检测 [20]。首先，GRPO 目标函数中的长度归一化项 $|y_i|$ 会在训练过程中促使模型生成更长的输出。随着模型不断学习，生成的解释往往会变长，从而增加推理延迟；对于必须实时处理交易的欺诈检测系统而言，这是一个关键问题。其次，GRPO 存在结构性的不一致：重要性权重是在词元级别计算的，而奖励反映的是整个输出序列的质量。这种粒度不匹配使我们所处场景下的信用分配变得复杂，因为检测准确率取决于完整的推理链，而非单个词元，并可能妨碍有效的梯度传播。

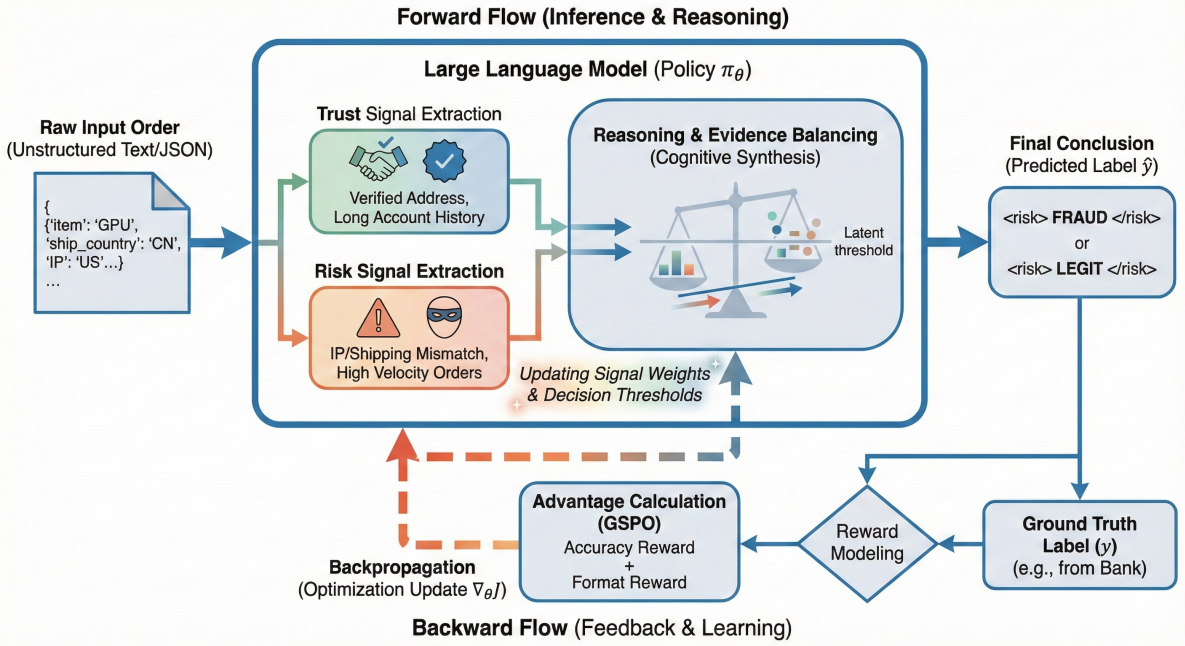


图2：**欺诈检测强化学习循环示意图。前向路径**展示了模型的推理链：从原始输入中提取多样的信任与风险信号，并针对潜在阈值综合证据，以作出判定。**反向路径**展示了优化过程：由真实标签得出的奖励信号驱动反向传播，更新参数，从而隐式学习信号相关性，并校准特定领域的决策边界。

3 方法

传统机器学习方法在信用卡欺诈检测方面面临两项根本性局限：它们只依赖数值特征，却忽略文本和行为模式；同时，由于特征工程聚合过程复杂，可解释性有限。为应对这些挑战，我们采用大语言模型（LLM），通过自然语言推理执行分类。然而，将通用大语言模型直接用于欺诈检测面临着重大障碍。这些模型缺乏领域专属知识，无法判断哪些信任或风险信号构成有意义的欺诈指标，也不了解这些信号应以多大程度影响分类决策。此外，在不同的风险控制场景中，这些信号的相关性和敏感度各不相同。为使大语言模型满足欺诈检测的要求，同时通过精简输出维持较低的响应延迟，我们采用强化学习作为后训练技术，以提升其特定任务的性能。

3.1 强化学习算法

如第2.2节所述，利用组相对基线的强化学习方法（例如组相对策略优化（GRPO）算法） [19], 已在对齐任务中展现出良好效果。然而，如前所述，GRPO 的令牌级目标函数构造和补全长度归一化方案在我们的应用中引入了潜在问题，可能导致性能欠佳。尤其值得注意的是，该归一化方案往往会激励生成越来越长的结果，这对于要求实时响应的欺诈检测系统而言，会带来严重的延迟问题。为缓解这些问题，**我们采用组序列策略优化（GSPO）算法** [15], ，该算法在其目标函数中采用序列级重要性加权机制，与我们欺诈检测任务中的序列级奖励结构自然一致。GSPO 优化目标表述如下：

$$\mathcal{J}_{GSPO}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta old}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(s_i(\theta)A_i, \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right)\right] \tag{3}$$

其中，序列级重要性权重定义如下：

System Role

You are an expert fraud risk assessor specializing in online payment fraud detection. Your task is to analyze the provided order JSON and determine whether it is at risk of credit card fraud.

Your Analysis Process

1. **Identify Risk Signals**: Examine the order details for indicators that suggest potential fraud
2. **Identify Trust Signals**: Look for patterns that suggest the transaction is legitimate
3. **Weigh Evidence**: Balance all signals to reach a well-reasoned conclusion

The online shopping order info is as follows:

<order_info>
{order_info}
</order_info>

Response Format

Give your reasoning first, wrapped in <reason> tags.

Structure your analysis as:
- Key risk signals identified (if any)
- Trust signals identified (if any)
- Overall risk assessment with justification

Then output an one-word decision: True (risky) or False (not risky), wrapped in <risk> tags.

Example:
<reason>
[Your analysis here]
</reason>

<risk>True/False</risk>

图3：欺诈检测语言模型的提示模板。

$$s_i(\theta) = \exp\left(\frac{1}{|y_i|}\sum_{i=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta old}(y_{i,t}|x, y_{i,<t})}\right) \tag{4}$$

组相对优势估计 A_i 与 GRPO 相同。请注意，外层目标函数中的长度归一化因子 $|y_i|$ 已被移除，从而防止模型通过生成越来越长的生成结果来稀释负优势。不过，为了降低方差，并使 $s_i(\theta)$保持在一致的数值范围内，该归一化因子仍保留在序列重要性权重 $s_i(\theta)$中，以避免短生成结果可能引发的剧烈波动。

在我们的欺诈检测设置中，输入提示 x 由三个部分组成：欺诈检测任务指令、电子商务订单信息和输出格式要求。采样输出 y_i 包括模型的推理（即识别出的信任与风险信号），随后是最终的欺诈判定。

3.2 奖励建模

奖励函数提供的训练信号从根本上决定了强化学习的优化方向。由于欺诈检测对结果高度敏感，性能几乎完全取决于最终判定，**我们采用由两个部分组成的基于规则的结果奖励系统**：

<reason>
Key risk signals identified:
- The transaction email "janedoe123@gmail.com" has a historical association with three previously canceled orders, all involving the same merchant ID (70295), same payment amount (48 USD), and same card fingerprint (29a5fc97b92ae646), indicating a prior pattern of failed attempts.
- The current order uses a different card fingerprint (864d450c18932e09) compared to the previous canceled orders, suggesting a new card is being used, potentially for a new attempt.
- The previous canceled orders using the same email were associated with different consignee phone numbers, indicating possible card testing or account takeover behavior.

Trust signals identified:
- The shipping address is marked as valid.
- The IP geolocation matches the shipping address location (Bay Shore, NY).
- The static IP score is 53.7, which is neutral and does not indicate high risk.
- There is no prior successful transaction associated with the current card fingerprint.

Overall risk assessment with justification:
Given the prior history of multiple canceled orders using the same email address and the lack of prior successful transactions with the current card fingerprint, this order strongly resembles a pattern of card testing or fraud. The combination of these red flags significantly increases the likelihood of fraudulent intent."
</reason>

<risk>True</risk>

图4：在测试集上由训练后的 Qwen3-14B 生成的示例输出。

- **准确性奖励**：该部分评估模型的最终欺诈风险判定是否与银行提供的真实标签一致。判定必须是确定且二元的（欺诈或合法）。为便于可靠地进行基于规则的提取和验证，模型必须以特定格式输出最终答案（例如，将答案置于 <risk> 标签中）。
- **格式奖励**：除了准确率之外，该组件还激励模型首先从输入订单中识别一系列关键信任与风险信号，并据此做出判断。模型需将这些信号置于<推理>标签中，以确保其在给出结论前阐明推理过程。

在我们的实验中，准确性奖励的权重是格式奖励的 2.5 倍，因为我们观察到，训练期间模型始终能够获得格式奖励，因此它提供的训练信号信息量较低。通过提高准确性奖励的权重，我们优先优化模型的欺诈检测性能，这是首要目标。我们刻意不采用过程奖励，因为我们只能获取可靠的真实欺诈标签，难以设置适当且可验证的中间检查点。与此同时，实证研究表明，结果奖励通常能在大语言模型强化学习中取得较好的效果，而过程奖励可能会诱发奖励劫持，因为模型可以在推理链中利用部分得分来钻空子 [21, 22]。

3.3 提示与补全格式

在我们的欺诈检测框架中，我们指令大语言模型在对订单是否存在信用卡欺诈风险作出最终判断前，先提供结构化推理。具体而言，模型必须先从输入的交易数据中识别两类证据：

- **风险信号**：看起来可疑并暗示存在潜在欺诈的特征或模式（例如，地理位置不一致、异常购买行为、身份信息不匹配）。
- **信任信号**：看起来合理且有助于证明订单真实性的特征或模式（例如，已验证的地址对应关系、既有客户历史记录、一致的支付模式）。

识别出这些信号后，模型必须综合所有证据，得出平衡且有理有据的结论。我们的提示设计刻意保持开放：我们不规定哪些内容构成有效的信任或风险信号，也不穷举潜在欺诈指标。这一设计选择在强化学习中具有战略意义，

通过不限制信号识别，我们鼓励模型在训练过程中探索其认为相关的多样化特征。借助奖励机制和基于梯度的优化，模型能够以端到端的方式发现并保留真正有信息价值的信号，而非受限于预定义的启发式规则。这种方法使模型有望发现不明显的欺诈模式，并适应人工制定的规则可能无法涵盖的不断演变的欺诈手法。完整的提示模板和示例输出分别见图3和图4。

4 实验

4.1 数据集

我们在一家中国全球支付解决方案公司提供的电子商务交易数据集上进行了实验，该数据集涵盖了2023年6月至2024年6月期间的交易。每条交易记录均采用 JSON 格式，包含多个类别的详细信息：

- **交易详情**：基本订单信息，包括收件人姓名、订单金额、创建时间戳（本地时间和 UTC 时间）、所购商品详情（商品名称、库存单位、数量、价格），以及商品是否为虚拟商品。
- **地理与网络信息**：交易的 IP地址及其地理定位数据（国家、城市、州）、IP 特征（托管服务器、匿名网络、虚拟专用网络、代理、Tor 状态）、IP地址类型（例如，蜂窝网络、住宅）、IP 注册国家、设备信息和静态IP评分。这些特征主要通过 MaxMind API 获取。
- **支付信息**：卡片指纹、卡片品牌（例如，Visa、Mastercard）、发卡银行、卡片类型（信用卡/借记卡）以及卡片特征（是否为商务卡、预付卡或虚拟卡）。
- **配送信息**：包含详细组成部分的配送地址（街道、城市、州、国家）、带国家代码和线路类型（移动电话/固定电话）的收件人电话号码、配送地址类型（住宅/企业、货运代理状态）以及地址有效性验证。
- **身份验证特征**：交易电子邮箱地址、首次观测到该电子邮箱以来的天数、配送地址/电话号码/电子邮箱与登记姓名之间的对应关系、电话号码验证，以及注册人年龄范围（如有）。这些特征主要通过 Ekata API 获取。
- **距离与一致性检查**：地理距离指标（IP地理位置与配送地址之间的距离、IP地理位置与电话号码之间的距离）以及一致性检查（配送电话号码是否与地址匹配、电子邮箱有效性）。这些特征主要通过 Ekata API 获取。
- **历史背景**：通过共享标识符（电子邮箱、电话号码、卡片指纹、IP地址）关联的历史关联订单，包括其支付金额、状态（成功、已取消、失败）和创建时间戳。

表1：数据概览

	样本数	正常订单	欺诈订单	时间范围
训练集	4900	2586 (51.8%)	2314（47.2%）	2023年6月至12月
测试集	5000	4520（90.4%）	480（9.6%）	2024年1月至6月

每笔订单均由发卡银行标记为正常或欺诈。原始数据集存在严重的类别不平衡，正常订单数量远远超过欺诈订单——这是现实世界欺诈检测场景的典型特征。为避免时间重叠导致的数据泄漏，我们按时间顺序划分数据集。训练集包含2023年6月至2023年12月的交易，测试集则涵盖2024年1月至2024年6月。表1概述了这两个数据集。

我们针对训练和测试采用了不同的平衡策略。对于训练集，我们尽可能纳入更多欺诈订单，并将其与数量略多的合法订单混合，构建近似平衡的数据集。这种平衡做法至关重要，因为我们的强化学习框架采用基于准确率的奖励：如果训练集类别严重不平衡，模型就可能通过将所有订单一概判为合法来轻易利用奖励函数。相比之下，测试集的类别分布则与现实世界的比例接近。这一设计选择体现了评估的根本目的——推断模型部署到生产环境后会有怎样的表现；在生产环境中，必须有效处理合法交易与欺诈交易之间天然存在的不平衡。

System Role

You are a fraud risk assessor for online orders. You must read the provided order JSON and decide if the order is at risk of credit card fraud.
Provide a brief reason first, then output True (risky) or False (not risky). Be concise, deterministic, and rely only on the given data. If data is missing or null, do not assume—call it out if it materially affects confidence.

Decision Rules

Base your decision on the following signals:

High-risk signals:
- IP obfuscation or infrastructure: Is IP a Hosting Server/Anonymous Network/Tor/Public Proxy/Anonymous VPN = true, or IP Address Type = hosting
- Payment history: Multiple recent failed payments linked to same email/phone/device/address/merchant, especially with different card fingerprints
- Geography mismatch: Issuing Card Country differs from Shipping Country, with no compensating ties (e.g., IP not in issuing country, large IP-to-shipping distance)
...

Medium-risk signals:
- Business address used for shipping consumer goods
- First-time device or suspicious device string vs. history, or unusual browser/OS fingerprints if provided
- Low order amount commonly used for card testing (e.g., < $20) when combined with other risk
- City/country inconsistencies across IP, shipping, phone, and billing
...

Low/mitigating signals (reduce concern if present):
- Strong cross-field consistency: name matches across shipping/billing/phone/email; phone/address validity = true; phone matches address; registrant age plausible
- Email with long history (many days)
- IP geolocation consistent with shipping/billing region and distances small
...

<reason>Although the order involves a virtual product, which is a high-risk signal, several mitigating factors are present. The email has a long history (4618 days), the IP geolocation is consistent with the shipping address, and there is a strong cross-field consistency across shipping/billing/phone/email. Additionally, the shipping address is a residential address, and the card is not a prepaid or virtual card.</reason>
<risk>False</risk>

图 5：压缩设置的提示词，以及由 Qwen3-14B 生成的响应示例。模型获得了一系列基于我们的反欺诈专家过往经验预先定义的风险信号和信任信号，并被指令生成确定且简洁的响应。为简洁起见，图中省略了格式要求和部分已定义的信号。

4.2 实验设置

模型选择 我们评估了多种不同的大语言模型，以全面衡量欺诈检测性能。在强化学习实验中，我们从Qwen3 系列中选取了三种模型 [12]：Qwen3-4B-Instruct、Qwen3-8B和Qwen3-14B。通过选择同一架构系列中不同规模的模型，我们可以在控制结构差异的同时，单独考察参数数量对性能的影响。为建立更广泛的性能基准，我们还评估了另外九种具有代表性的大语言模型。我们的闭源基线包括GPT-4.1 [23],GPT-5-mini [10], Gemini-2.5-Flash [24], 和 Claude-4.5-Sonnet [11]。对于大规模开源模型，我们测试了Qwen3-235B-A22B-Instruct [12], Kimi-K2 [25], DeepSeek-V3 [26], 和GLM-4.5 [27]对于同时提供标准模式和推理模式（例如扩展思维链）的模型，我们仅评估标准模式，因为欺诈检测需要实时响应，而推理模式会带来难以接受的延迟。

训练配置我们采用 GSPO 强化学习算法训练三个通义千问 Qwen3 模型。除第 3 节所述的标准设置外，我们还实施了另一种配置作为对照条件，并将其称为压缩版本。两种配置在探索—利用权衡和响应约束方面有所不同：

- **标准设置**：使用开放式提示，不指定潜在风险或信任信号，以鼓励模型在训练期间探索多种欺诈指标。这种方法使模型能够在响应中进行不受约束的探索和推理。
- **压缩设置**：将预定义的风险和信任信号直接纳入提示，利用先前的领域知识引导模型识别已知欺诈模式（侧重利用而非探索）。此外，为降低推理延迟，我们限制模型生成更短、更确定的响应，并采用压缩的推理链。

由于这种对照配置同时压缩了模型的探索行为和推理冗长表达，我们将其称为压缩指令设置。该配置的提示模板和响应示例见图 5。

评估指标我们采用多个评估维度，全面评估模型在欺诈检测中的性能。

分类性能：我们采用适用于类别不平衡的二元分类的标准指标：准确率、召回率（真正率）、特异度（真负率）、精确率、假正率和 F1分数。在欺诈检测中，错误分类会在两方面造成损失：假阳性会阻止合法交易并损害商户收入，而假阴性则会放过欺诈性扣款，导致银行损失。鉴于这两类错误的成本对称，我们的主要优化目标是最大化F1分数；该指标兼顾精确率和召回率，体现了模型既不应过于保守、也不应过于激进的要求。

响应效率：由于信用卡欺诈检测需要实时响应，我们会测量测试集中各大语言模型生成的每个生成结果的平均词元数量。尽管详尽的推理有助于提升可解释性，我们仍优先考虑简洁适度且清晰明了的响应，避免过度冗长，因为较长的响应会增加延迟，并可能降低实际可部署性。

事实忠实性：与传统机器学习模型相比，大语言模型的一项关键优势是能够生成便于人类理解的解释。但只有当模型的推理忠实反映输入数据时，这种可解释性才有价值。为评估忠实性，我们使用 DeepEval 框架 [28]对三个微调后的 Qwen3 模型进行幻觉测试。我们使用 Qwen3-235B-A22B-Instruct作为评估器——这是一个能力出色、推理能力足以胜任可靠评判器的模型。我们采用严格的评估标准：只有评判器认定响应中绝无任何事实错误或捏造信息时，该响应才算通过幻觉测试。即使在引用订单详情时存在轻微不准确（例如金额错误、地址不符，或对历史模式作出缺乏依据的断言），也会导致测试失败。这一严格标准确保部署的模型能够提供可信的解释，让人工审核人员可以放心依赖。

4.3 结果

4.3.1 强化学习的效能：比较分析

表2总结了我们提出的框架与基线及最先进的通用模型之间的定量评估结果。实证结果有力地证明了领域专用强化学习的效能，并揭示了两种截然不同的现象：后训练模型的决策边界发生根本性重塑，而通用大语言模型则存在显著的“校准差距”。

GSPO 的变革性影响 在标准指令设置（GSPO.S）下应用 GSPO 强化学习，从根本上提升了欺诈检测能力。如图6所示，经过后训练的通义千问 Qwen3 变体取得了大幅性能提升，其中 4B、8B 和 14B 模型的 F1分数较各自的基础模型分别提高了 120.90%、98.35% 和 105.27%。

剖析这些提升可以发现，强化学习过程主要通过抑制假阳性来优化性能。三个微调模型的特异度（真负率）均大幅提升，其中 8B 模型尤为显著，提升了 336.14%。这表明，未经训练的基础模型容易产生“幻觉式怀疑”，频繁将正常异常标记为欺诈。强化学习过程有效校准了这些模型，使其采用更严格的证据标准。有趣的是，这种校准在不同规模的模型中表现各异。较大的模型（8B 和 14B）采取了保守、优先追求精确率的策略，以牺牲部分召回率

表2：**不同大语言模型在测试集上的表现**。七项测试指标为准确率、召回率（真正率）、特异度（真负率）、精确率、假正率、F1分数，以及每个生成结果的平均词元数。GSPO.C 和 GSPO.S 分别表示采用压缩和标准指令设置的强化学习。最佳 F1分数以粗体标示，次佳分数加下划线。

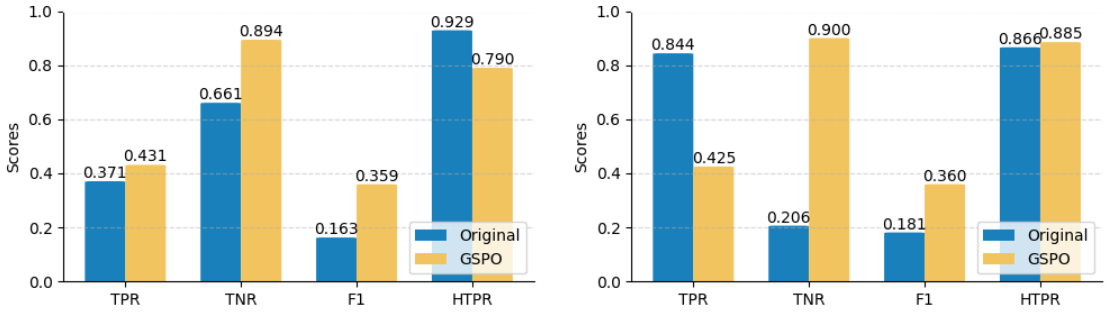
系列	版本	测试指标						
		Acc.	TPR	TNR	精确率	FPR	F1	词元数
闭源大语言模型								
GPT	4.1	0.6826	0.2896	0.7243	0.1004	0.2757	0.1491	548.8
GPT	5-mini	0.7476	0.2250	0.8031	0.1082	0.1969	0.1461	664.9
Claude	sonnet-4.5	0.6654	0.3083	0.7033	0.0994	0.2967	0.1503	725.7
Gemini	2.5-Flash	0.5244	0.4771	0.5294	0.0972	0.4706	0.1615	817.0
开源大语言模型（*>14B*）								
Qwen3-235B	指令	0.4756	0.6021	0.4622	0.1062	0.5378	0.1806	668.2
DeepSeek	v3 (671B)	0.6314	0.3554	0.6628	0.0955	0.3372	0.1487	449.0
Kimi	K2 (1T)	0.6880	0.2354	0.7361	0.0865	0.2639	0.1265	440.7
GLM	4.5（355B）	0.7090	0.2104	0.7619	0.0858	0.2381	0.1219	447.7
开源大语言模型（*4B - 14B*）								
Qwen3-4B	指令	0.6329	0.3708	0.6608	0.1042	0.3392	0.1627	761.8
	GSPO.C	0.6918	0.3521	0.7279	0.1208	0.2721	0.1799	82.3
	GSPO.S	0.8521	0.4313	0.8939	0.3080	0.1061	<u>0.3594</u>	282.5
Qwen3-8B	非思考	0.2677	0.8438	0.2064	0.1016	0.7936	0.1814	626.3
	GSPO.C	0.7304	0.3333	0.7726	0.1347	0.2274	0.1918	96.3
	GSPO.S	0.8545	0.4250	0.9002	0.3119	0.0998	**0.3598**	447.2
Qwen3-14B	非思考	0.4993	0.5292	0.4961	0.1005	0.5039	0.1689	522.8
	GSPO.C	0.7228	0.3271	0.7648	0.1287	0.2352	0.1847	60.7
	GSPO.S	0.8331	0.4604	0.8728	0.2780	0.1272	0.3467	370.5

（-60.50% 和 -13.00%）为代价来最大化特异度。相比之下，4B 模型实现了难得的“平衡优化”，召回率（+16.32%）和特异度均有所提升。这表明，对于这一特定任务的复杂度，较小的参数空间或许能够支持更高效的策略探索，同时避免对多数类过拟合。

通用大语言模型中的校准差距 我们的基准测试中有一项引人注目的发现：我们的专用小规模模型（4B–14B 参数）始终优于大规模的最先进系统，例如 Claude-4.5-Sonnet。尽管这些前沿模型具备更强的推理能力和世界知识，但它们在这一特定领域中却无法与之竞争。

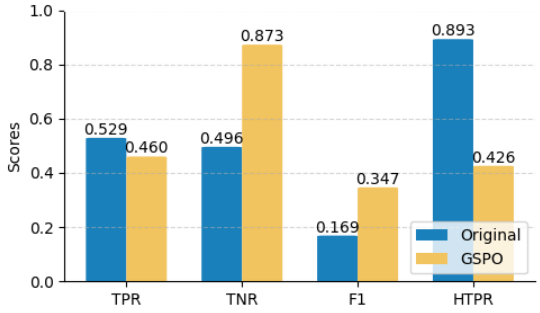
我们认为，这种差异源于<词汇表中未提供的>校准差距</词汇表中未提供的>，而非推理能力不足。通用大语言模型通常能够成功识别相关信号——例如注意到某个 IP地址是代理，或运输距离较长——但它们缺乏领域特定经验，无法根据潜在决策阈值权衡这些信号。图2展示了弥合这一差距的关键机制：强化学习循环。通过将源自真实标签的奖励信号进行反向传播，我们的模型能够隐式地将其潜在决策边界调整到金融数据集的特定分布上。如果没有这种有针对性的优化，通用大语言模型就无法判断“500公里的距离”对于这一特定商户类别而言是高风险异常，还是标准偏差。通过使用 GSPO 进行微调，我们有效地将小型模型的推理能力锚定到金融数据集的特定风险分布上，使其能够超越规模大出数个数量级、却未经任务校准的模型。

涌现效率与信息密度 强化学习流程中一项关键且涌现的特性，是响应长度会自发优化。在未明确惩罚冗长表达的情况下，GSPO 过程促使 4B、8B 和 14B 模型分别将输出压缩了 62.92%、28.60% 和 29.13%。这种精简并不意味着信息损失，反而提升了信息密度：模型学会舍弃多余的语言填充内容，只专注于高价值的诊断信号。这种“信息



(a) Qwen3-4B

(b) Qwen3-8B



(c) Qwen3-14B

图6：原始大语言模型与采用 GSPO 和标准提示设置训练后的版本之间的性能对比。图中绘制了四项指标：召回率（真正率）、特异度（真负率）、F1分数和幻觉测试通过率。

蒸馏”在实际应用中至关重要；通过自然地收敛到简洁的推理路径，模型显著降低了推理延迟，从而直接助力其部署于高吞吐量、实时的金融交易系统中。在这类系统中，毫秒级延迟是不可接受的。

认知策略的差异 尽管三种模型采用相同的架构、提示模板和优化目标，但它们在欺诈分析中形成了截然不同的“认知策略”（见图4和图7）。

- **演绎型策略家（14B）：** Qwen3-14B 严格遵循逻辑流程，系统地识别风险信号和信任信号，然后综合这些信号得出结论。这与标准的思维链（CoT）推理相似。
- **执行型策略家（8B）：** 通义千问Qwen3-8B采用“先给结论”的结构，先直接给出分类结果，再回溯说明理由。这表明，该模型对决策边界的内化十分充分，因此无需花太多“思考时间”就能得出结论。
- **整体型策略家（4B）：** Qwen3-4B以自由形式的叙述整合证据，将信号提取与判断融为一段连贯的文字。

这种策略分歧表明，欺诈检测的解空间中存在多个局部最优；推理欺诈问题并不存在唯一“正确”的方式，不同能力的模型会以不同方式探索这一空间，以最大化奖励。

忠实度反向缩放：能力陷阱 或许最违背直觉的发现是，在强化学习微调下，模型规模与事实忠实性之间存在反向关系。与通常的缩放定律相反，更大的模型并未带来更优的整体表现。训练前后，4B和8B模型的幻觉测试通过率均保持在相近水平（超过80%）。相比之下，14B模型虽然提升了F1分数，但幻觉测试通过率却灾难性地下降了46.7个百分点（图6），真实性表现远逊于4B和8B模型。

我们将其解读为模型容量提升所促成的一种“奖励劫持”表现。较大的 14B 模型具备更强的语言流畅性和世界知识，似乎会利用这些能力来捏造 令人信服的

<reason>
The order is from a US IP address in Ashburn, Virginia, which is geolocated 88 miles from the shipping address in Greenville, SC. The shipping address is in a multi-unit dwelling and the phone line is mobile, with the phone number not matching the address. The device is flagged as "Unrecognized," which is a red flag as unrecognized devices are often associated with fraudulent activity. There is a history of prior orders from the same IP and email, including two canceled orders of $84.99 and one paid order of $169.99, which shows inconsistent behavior and possibly multiple failed attempts or attempts to exploit price differences. The high amount of $84.99 for a jersey is relatively unusual and potentially suspicious given the mobile phone line and unrecognized device. Despite the address being valid and both IP and phone registered in the US, the geolocation mismatch and unrecognized device raise concern for possible fraud. The card is a Discover credit card issued in the US, and no indication of prepaid or virtual cards.</reason>

<risk>True</risk>

<reason>
The transaction appears to be low risk for credit card fraud based on the available data.

Key Risk Signals:
- There are no clear red flags such as mismatched shipping and billing addresses, suspicious IP geolocation, or inconsistencies in the device or card fingerprint.
- The IP address is from the UK (GB) and geolocated to Redditch, which is close to the shipping address in Shrewsbury (75 miles), which is within a reasonable distance.
- ...
Trust Signals:
- The shipping and billing addresses are in the UK and match the IP geolocation.
- The phone number is from the UK and matches the consignee's name (John Doe), which aligns with the shipping address.
- ...
Overall Risk Assessment:
The transaction does not exhibit any of the typical patterns associated with credit card fraud. All the provided information aligns with a legitimate UK-based transaction. The small amount, consistent geolocation, and valid card details further support the legitimacy of the transaction. While the device information is marked as "Unrecognized," this is not uncommon and does not necessarily indicate fraud in this context.
</reason>

<risk>False</risk>

图7：分别展示训练后的 Qwen3-4B 和 Qwen-8B 的输出示例。为简洁起见，省略了部分识别出的信号。尽管收到相同的指令，不同规模的模型往往会形成不同的推理风格。

证据，以证明其预测合理，而非单纯依赖输入数据。从本质上说，模型优先考虑说服力（以最大化奖励），而非真实性。相比之下，容量较小的模型不擅长生成看似合理的幻觉细节，因此更能立足于所提供的上下文。这种“少即是多”现象凸显了一项关键的安全挑战：在高风险金融领域，如果奖励信号没有明确惩罚捏造行为，模型智能提升反而可能导致解释的可靠性下降。

4.3.2 探索与认知空间的关键作用

过早施加约束的代价 为单独考察强化学习过程中推理自由度的影响，我们对开放式的“标准”设置与“压缩”变体（见第4.2节）进行了对比消融实验。压缩设置代表一种“人工引导”的方法：它注入领域知识（预定义信号），并严格限制篇幅以尽量降低延迟。

表2和图9所呈现的结果揭示了一个显著的权衡。尽管压缩配置成功地将词元用量降至最低，却导致学习效能陷入功能性崩溃。4B、8B和14B模型的F1分数仅分别提升了10.57%、5.73%和9.35%。相比之下，允许

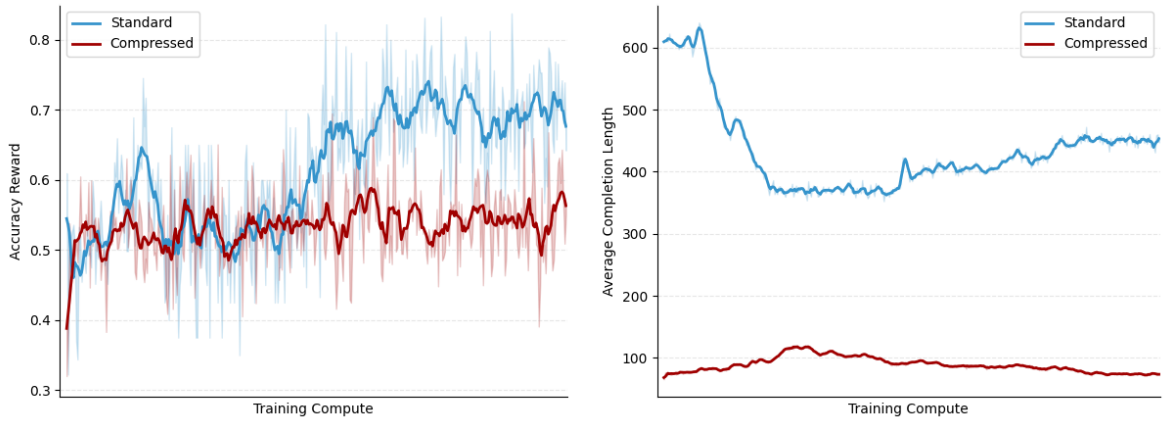
Order info: ...

Response:
Key risk signals identified:
- The "Days since first observation of this email" is 1960 days (approximately 5 years), which is highly unusual and raises a red flag, as it suggests the email address has not been previously used for any prior transactions, which is atypical for a legitimate user.
- The "Shipping Phone Corresponding Registered Name" matches the consignee's name, but the "Shipping Phone Registrant Age Range" is listed as 65–69, which conflicts with the consignee's name "Maria E Montelongo" (likely a younger individual).

Hallucination test: **Failed** ✗

This response is **rejected** because the actual output contains a direct contradiction: it claims the email 'has not been previously used for any prior transactions' despite the context stating it was first observed 1960 days ago, indicating prior use. Additionally, the output makes unsupported speculative claims about age based solely on a name, which is not logically or factually grounded in the context.

图8：经过训练的通义千问Qwen3-14B生成的回答未能通过幻觉测试，因为评判器发现其中存在事实错误。



（a）准确性奖励　　（b）平均补全长度

图9：Qwen3-8B进行GSPO训练期间，标准指令设置与压缩指令设置在准确性奖励和平均补全长度方面的对比。横轴表示训练计算量，相当于强化学习过程中处理的样本数。图中曲线经过移动平均值平滑处理，阴影区域表示所绘指标的实际值。

不受限制探索的标准设置带来了高出一个数量级以上的性能提升（范围为98.35%至120.90%）。

“认知紧箍咒”效应我们将这一巨大的性能差距归因于两种机制：探索抑制和推理计算的截断。首先，将特定风险信号硬编码到提示中，会使压缩设置给模型套上“认知紧箍咒”。这会让优化过程偏向利用已知的启发式规则，而非发现新颖的数据驱动模式。相反，标准设置允许模型将特征空间视为一张白板，使强化学习智能体能够识别出人类设计者可能忽略的细微、非直观的相关性。

其次，压缩模型的失败为欺诈检测情境中的“思维链”假说提供了有力的实证支持。大语言模型通过自回归令牌生成来执行计算；

从本质上说，“思考”需要“表达”。迫使压缩模型立即给出判定，或仅作极少阐述，实际上大幅削减了其推理的计算预算。标准设置表现更优，这证实了准确的欺诈分类并非只是模式匹配——它需要一系列中间推理步骤来权衡相互矛盾的证据。过早截断这一推理链，会迫使模型猜测而非推导。

对实时系统的启示这些发现消除了将大语言模型部署于金融风险控制中的一项根本矛盾。尽管低延迟不容妥协，我们的结果表明，强制简洁会适得其反。相反，最优策略是自然地提升效率：采用标准设置，让模型自我优化，自然舍弃无用词元以最大化奖励（如第4.3.1节所示），同时保留必要的“认知空间”，以推理复杂的欺诈场景。

5 讨论

如第4.3.1节所述，我们在定制的欺诈检测任务中发现了“少即是多”现象：经过相同的强化学习训练后，较小的语言模型可能优于较大的语言模型。我们推测有以下几个促成因素：

容量-任务对齐：对于欺诈分类这类高度专业化的任务，较小模型有限的容量反而可能更有优势。参数较少的模型更容易针对数据集特定模式进行有效调整，不会受到大量先验知识的干扰；这些知识对于当前狭窄的任务而言可能无关紧要，甚至会适得其反。

灾难性遗忘：较大的模型通过预训练习得了更丰富、更复杂的表征。在针对单一任务进行强化学习微调时，强烈的优化压力可能导致这些模型覆盖或丢弃通用推理能力。这些能力虽不能直接用于欺诈检测，却为事实依据提供了稳定基础 [29]。这一现象表现为幻觉增加——模型对欺诈模式过拟合，同时更难忠实地呈现输入信息。

优化动态：在有限数据上进行强化学习训练时，较小的模型可能呈现更稳定的梯度地形，使其能够收敛到更好的局部最优，在任务性能与对输入数据的忠实性之间取得平衡。

这些发现表明，在生产部署基于大语言模型的欺诈检测系统时，实践者应综合考量模型规模、任务性能、响应效率和事实可靠性的完整帕累托前沿，而不是默认选择现有规模最大的模型。

尽管我们的研究仅关注基于大语言模型的分类，但语言模型与传统机器学习各自的优势互补，表明混合系统具有广阔的应用前景。传统机器学习模型擅长处理结构化数值特征，并以极低的延迟从大规模表格数据中捕捉统计模式。相比之下，大语言模型在推理文本信息、行为模式、身份不一致以及难以通过传统特征工程处理的复杂多模态信号方面表现出更强的能力。未来的研究应探索最佳集成策略，包括级联架构、集成方法和动态路由机制，根据案例复杂度分配处理方式。此外，我们观察到，强化学习训练后事实忠实性有所下降，尤其是在较大的大语言模型中。这给生产部署带来了风险，因为欺诈分析师在审查标记的交易时必须信任模型解释。未来的研究可以考虑缓解这一问题，例如在强化学习训练期间，除了准确度奖励之外，还加入明确的忠实度奖励，并可能利用自动事实核查系统对编造的细节进行惩罚。

6 相关研究

大语言模型在金融欺诈检测中的应用传统上，信用卡欺诈检测主要依赖支持向量机[30, 31, 32], 、随机森林[33, 34], 、梯度提升决策树（XGBoost [5, 6, 7] 和 LightGBM [8, 9]）以及神经网络 [2, 3, 4], 等监督式机器学习模型，这些模型擅长处理结构化表格数据。尽管准确率很高，这些“黑箱”模型往往缺乏语义理解能力和可解释性，而这正是金融合规中的关键要求，因此催生了一系列可解释欺诈检测研究 [35, 36, 37, 38, 7]。近期研究开始探索将大语言模型应用于这一领域，利用其自然语言处理能力。例如，Ren 等人 [39]研究了在复杂、逼真的场景中，采用多智能体系统进行基于大语言模型的欺诈检测。Singh 等人 [40] 采用检索增强生成（RAG），为大语言模型提供相关上下文，以实现实时欺诈检测。Huang 等人 [41] 和 Yang 等人 [42] 将大语言模型融入图神经网络，用于欺诈检测中的语义分析。然而，这些方法通常将大语言模型视为冻结推理引擎，限制了其内化海量交易数据集中欺诈行为所呈现的复杂高维模式的能力。

用于大语言模型后训练的强化学习基于人类反馈的强化学习（RLHF）已成为使大语言模型与复杂目标对齐的标准方法 [16]。近端策略优化（PPO） [18] 和直接偏好优化（DPO） [43] 等标准算法已被广泛用于改进指令遵循能力。然而，标准 PPO 往往面临训练不稳定性和内存成本高的问题，因为它需要同时训练策略和评论家模型，而二者的规模通常相当。为解决这一问题，研究者提出了组相对策略优化（GRPO）算法 [19] ，通过对一组输出中的奖励进行归一化，免去了评论家模型。较新的组序列策略优化（GSPO） [15] 则通过在序列级别而非词元级别进行优化，进一步提升了训练稳定性。强化学习标志着从通用指令微调向特定任务的行为对齐转变，使模型能够通过反复试错“学习”交易数据中细微的风险信号，而不只是依靠监督模仿。

7 结论

本研究提出了一种方法，仅利用原始交易数据，将大语言模型的文本理解与推理能力同强化学习相结合，用于信用卡欺诈检测。与传统机器学习方法相比，该方法提供了直观、自然语言形式的可解释性。我们的结果表明，**使用强化学习微调轻量级语言模型，能够在保留可解释性的同时，显著提升欺诈检测任务的性能**。在多个 Qwen3 模型中，经过精心设计的奖励进行任务专用强化学习能够提升 F1分数及相关指标，并且通常生成更短的输出，适用于低延迟部署。重要的是，我们观察到“少即是多”的规律：经过强化学习后，规模更小、针对任务进行适配的模型能够达到甚至超越更大模型的性能，凸显了领域聚焦优化的价值。我们还强调了平衡探索与忠实性的必要性。开放式信号发现能够增强鲁棒性和解释丰富性，但过度压缩或过度依赖预定义信号可能会大幅削弱判别能力。因此，实际部署更适合采用混合方法：持续优化忠实性感知目标并加强验证，同时探索混合架构，将大语言模型的可解释性与传统模型在结构化数据处理方面的速度相结合。

参考文献

[1] Kanishka Ghosh Dastidar、Olivier Caelen 和 Michael Granitzer。信用卡欺诈检测的机器学习方法：综述。*IEEEAccess*，2024年。[2] Yara Alghofaili、Albatul Albattah 和 Murad A Rassam。基于长短期记忆网络深度学习技术的金融欺诈检测模型。《应用安全研究期刊》，15(4):498–516，2020年。[3] Joy Iong-Zong Chen 和 Kong-Long Lai。用于信用卡欺诈检测与告警的深度卷积神经网络模型。《人工智能期刊》，3(02):101–112，2021年。[4] R Udayakumar、A Joshi、SS Boomiga 和 R Sugumar。Deep Fraud Net：一种用于网络安全和金融欺诈检测与分类的深度学习方法。《互联网服务与信息安全期刊》，13(3):138–157，2023年。[5] Petr Hajek、Mohammad Zoynul Abedin 和 Uthayasankar Sivarajah。使用基于 XGBoost 的框架检测移动支付系统中的欺诈。《信息系统前沿》，25(5):1985–2003，2023年。[6] Teuku Rizky Noviandy、Ghalieb Mutig Idroes、Aga Maulana、Irsan Hardi、Edi Saputra Ringga 和 Rinaldi Idroes。利用 XGBoost 驱动的机器学习和数据增强技术进行当代财务管理中的信用卡欺诈检测。《*Indatu*管理与会计期刊》，1(1):29–35，2023年。[7] Nazerke Baisholan、J Eric Dietz、Sergiy Gnatyuk、Mussa Turdalyuly、Eric T Matson 和 Karlygash Baisholanova。FraudX AI：面向不平衡数据集的可解释机器学习信用卡欺诈检测框架。《计算机》，14(4):120，2025年。[8] Xiaosong Zhao、Yong Liu 和 Qiangfu Zhao。针对极度不平衡数据的改进 LightGBM 及其在信用卡欺诈检测中的应用。*IEEE Access*，2024年。[9] Yao Xiao、Li Tan 和 Jiaran Liu。机器学习模型在欺诈识别中的应用：CatBoost、XGBoost 和 LightGBM 的比较研究。2025年。[10] OpenAI。GPT-5 已发布。https://openai.com/gpt-5，2025年8月。[11] Anthropic。推出 Claude Sonnet 4.5。https://www.anthropic.com/news/claude-sonnet-4-5，2025年9月。[12] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。Qwen3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025年。

[13] Surendrabikram Thapa、Shuvam Shiwakoti、Siddhant Bikram Shah、Surabhi Adhikari、Hariram Veeramani、Mehwish Nasim 和 Usman Naseem。计算社会科学中的大语言模型（大语言模型）：前景、现状与挑战。Social Network Analysis and Mining，15(1):1–30，2025年。[14] Mohamed Amine Ferrag、Norbert Tihanyi 和 Merouane Debbah。从大语言模型推理到自主人工智能代理：全面综述。arXiv 预印本 arXiv:2504.19678，2025年。[15] Chujie Zheng、Shixuan Liu、Mingze Li、Xiong-Hui Chen、Bowen Yu、Chang Gao、Kai Dang、Yuqiong Liu、Rui Men、An Yang 等。组序列策略优化。arXiv 预印本 arXiv:2507.18071，2025年。[16] Zhichao Wang、Bin Bi、Shiva Kumar Pentyala、Kiran Ramnath、Sougata Chaudhuri、Shubham Mehrotra、Xiang-Bo Mao、Sitaram Asur 等。大语言模型对齐技术综述：基于人类反馈的强化学习、基于人工智能反馈的强化学习、近端策略优化、直接偏好优化等。arXiv 预印本 arXiv:2407.16216，2024年。[17] Lex Weaver 和 Nigel Tao。基于梯度的强化学习的最优奖励基线。arXiv 预印本 arXiv:1301.2315，2013年。[18] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347，2017年。[19] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。DeepSeekMath：突破开放语言模型数学推理的极限。arXiv 预印本 arXiv:2402.03300，2024年。[20] Zichen Liu、Changyu Chen、Wenjun Li、Penghui Qi、Tianyu Pang、Chao Du、Wee Sun Lee 和 Min Lin。理解 R1-Zero 式训练：批判性视角。arXiv 预印本 arXiv:2503.20783，2025年。[21] Jiaxuan Gao、Shusheng Xu、Wenjie Ye、Weilin Liu、Chuyi He、Wei Fu、Zhiyu Mei、Guangju Wang 和 Yi Wu。在训练时为大语言模型推理设计有效的强化学习奖励。arXiv 预印本 arXiv:2410.15115，2024年。[22] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。arXiv 预印本 arXiv:2501.12948，2025年。[23] OpenAI。在 API 中推出 GPT-4.1。https://openai.com/index/gpt-4-1，2025年4月。[24] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：借助先进推理、多模态、长上下文和新一代代理能力推动前沿发展。arXiv 预印本 arXiv:2507.06261，2025年。[25] Kimi Team、Yifan Bai、Yiping Bao、Guanduo Chen、Jiahao Chen、Ningxin Chen、Ruijue Chen、Yanru Chen、Yuankun Chen、Yutian Chen 等。Kimi K2：开放代理智能。arXiv 预印本 arXiv:2507.20534，2025年。[26] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。DeepSeek-V3技术报告。arXiv 预印本 arXiv:2412.19437，2024年。[27] Aohan Zeng、Xin Lv、Qinkai Zheng、Zhenyu Hou、Bin Chen、Chengxing Xie、Cunxiang Wang、Da Yin、Hao Zeng、Jiajie Zhang 等。GLM-4.5：代理、推理与代码生成（ARC）基础模型。arXiv 预印本 arXiv:2508.06471，2025年。[28] Confident AI。DeepEval 幻觉测试。https://deepeval.com/docs/metrics-hallucination。[29] Hongyu Li、Liang Ding、Meng Fang 和 Dacheng Tao。重新审视大语言模型微调中的灾难性遗忘。arXiv 预印本 arXiv:2406.04836，2024年。[30] Nana Kwame Gyamfi 和 Jamal-Deen Abdulai。使用支持向量机进行银行欺诈检测。载于 2018 IEEE 第九届年度信息技术、电子与移动通信会议（IEMCON），第 37–41 页。IEEE，2018年。[31] Chenglong Li、Ning Ding、Yiming Zhai 和 Haoyun Dong。基于不同支持向量机的信用卡欺诈检测比较研究。Intelligent Data Analysis，25(1):105–119，2021年。[32] Ajeet Singh、Anurag Jain 和 Seblewongel Esseynew Biable。基于萤火虫优化算法和支持向量机的金融欺诈检测方法。Applied Computational Intelligence and Soft Computing，2022(1):1468015，2022年。[33] Shiyang Xuan、Guanjun Liu、Zhenchuan Li、Lutao Zheng、Shuo Wang 和 Changjun Jiang。用于信用卡欺诈检测的随机森林。载于 2018 IEEE 第十五届网络、传感与控制国际会议（ICNSC），第 1–6 页。IEEE，2018年。[34] Vaishnave Jonnalagadda、Priya Gupta、Eesita Sen 等。使用随机森林算法进行信用卡欺诈检测。International Journal of Advance Research, Ideas and Innovations in Technology，5(2):1–5，2019年。

[35] Cynthia Rudin. 对于高风险决策，应停止解释黑箱机器学习模型，转而使用可解释模型。*Nature machineintelligence*, 1(5):206–215, 2019.[36] Kang Lin and Yuzhuo Gao. 基于 Group SHAP 的金融欺诈检测模型可解释性。*ExpertSystems withApplications*, 210:118354, 2022.[37] Md Rokibul Hasan, Md Sumon Gazi, and Nisha Gurung. 信用卡欺诈检测中的可解释人工智能：通过可解释模型和透明决策增强美国的信任与合规。*JournalofComputerScience and Technology Studies*, 6(2):1–12, 2024.[38] Kaidi Li, Tianmeng Yang, Min Zhou, Jiahao Meng, Shendi Wang, Yihui Wu, Boshuai Tan, Hu Song, Lujia Pan, Fan Yu, et al. SEFraud：通过解释性掩码学习实现基于图的自解释欺诈检测。载于*ACM SIGKDD*知识发现与数据挖掘会议第*30*届会议论文集, pages 5329–5338, 2024.[39] Qibing Ren, Zhijie Zheng, Jiaxuan Guo, Junchi Yan, Lizhuang Ma, and Jing Shao. 当人工智能代理在线串通：社交平台上的协作式大语言模型代理所带来的金融欺诈风险。*arXiv* 预印本 *arXiv:2511.06448*, 2025.[40] Gurjot Singh, Prabhjot Singh, and Maninder Singh. 使用基于检索增强生成的大语言模型进行先进的实时欺诈检测。*arXiv* 预印本 *arXiv:2501.15290*, 2025.[41] Tairan Huang, Yili Wang, Qiutong Li, Changlong He, and Jianliang Gao. 大语言模型能找出欺诈者吗？多层级大语言模型增强图欺诈检测。载于*ACM*国际多媒体会议第*33*届会议论文集, pages 1530–1538, 2025.[42] Chengdong Yang, Hongrui Liu, Daixin Wang, Zhiqiang Zhang, Cheng Yang, and Chuan Shi. FLAG：使用大语言模型增强图神经网络进行欺诈检测。载于*ACM SIGKDD*知识发现与数据挖掘会议第*31*届会议论文集，第*2*卷, pages 5150–5160, 2025.[43] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 直接偏好优化：你的语言模型暗中是一个奖励模型。神经信息处理系统进展, 36:53728–53741, 2023。