arXiv:2602.16109v1 [cs.CR] 18 Feb 2026

用于跨境内部事务的联邦图 AGI 政府金融计划中的威胁情报 r

Srikumar Nayak[1,2] 和 James Walmesley[3]

[1]Incedo Inc.，美国新泽西州[2]印度理工学院，印度泰米尔纳德邦金奈[3]肯特大学，英国肯特

2026年2月19日

摘要

跨境内部人员威胁给政府金融计划带来了严峻挑战，尤其是在处理分布于多个司法管辖区的隐私敏感数据时。现有方法存在根本性局限：由于隐私约束，它们无法有效地跨境共享情报；缺乏理解复杂多步骤攻击模式的推理能力；也无法捕捉金融网络中错综复杂的图结构关系。我们提出了**FedGraph-AGI**，这是一种新型联邦学习框架，将通用人工智能（AGI）推理与图神经网络相结合，用于隐私保护的跨境内部威胁检测。我们的方法结合了：（1）能够维护数据主权的联邦图神经网络；（2）面向异质司法管辖区的混合专家（MoE）聚合；以及（3）通过大型动作模型（LAM）对图数据进行因果推断的通用人工智能驱动推理。在一个涵盖10个司法管辖区、包含50,000笔交易的数据集上进行实验后，FedGraph-AGI 达到了**92.3%的准确率**，显著优于联邦基线（86.1%）和集中式方法（84.7%）。消融研究表明，通用人工智能推理带来了6.8%的性能提升，而混合专家则提升了4.4%。该系统在保持 $\epsilon = 1.0$ 差分隐私的同时，实现了近乎最优的性能，并可高效扩展至 $50+$ 个客户端。这是首次将通用人工智能推理与联邦图学习相结合用于内部威胁检测，为隐私保护的跨境情报共享开辟了新方向。

关键词： 联邦学习；图神经网络；通用人工智能；大型动作模型；专家混合；神经网络；内部人员威胁检测；隐私保护机器学习；跨境情报

---

1 引言

1.1 背景与动机

政府金融计划——包括社会福利项目、补贴分配、税收征管和公共采购——每年处理数万亿美元。这些系统日益面临复杂的内部威胁：拥有特权访问权限的恶意行为者利用自身地位实施欺诈、贪污或间谍活动。世界银行估计，政府腐败每年给发展中国家造成超过 1 万亿美元的损失；与此同时，内部威胁占政府欺诈案件的 34%，每起案件的损失中位数超过 15 万美元。

这一挑战因三个关键因素而更加严峻。**首先，跨境复杂性**：现代金融方案涉及多个司法管辖区，交易跨越国界。要侦测复杂欺诈，需要共享情报；然而，数据隐私法规（GDPR、国家主权法律）禁止集中式聚合。**其次，图结构相互依赖**：金融交易构成复杂网络，威胁会表现为结构异常——异常模式、可疑关联以及协同多方计划；若孤立地审查单笔交易，这些情况便无从察觉。**第三，推理要求**：内部威胁日益采用多步策略，需要因果推理——历时数月的权限提升、跨司法管辖区设立空壳公司、配合政策变化开展协同行动。这类模式要求系统能够进行长程类人推理。

传统机器学习方法在这一场景中难以奏效。由于隐私约束，集中式模型无法访问跨境数据。标准的联邦学习支持分布式训练，但缺乏图感知和推理能力。图神经网络（GNN）擅长处理关系数据，但在联邦环境中却面临挑战，因为图结构横跨多个数据孤岛。现有系统缺乏抽象推理能力——包括因果推断、反事实分析和多跳逻辑推演——而人类分析师在调查复杂欺诈时会运用这些能力。

1.2 对AGI增强联邦图学习的需求

通用人工智能代表着一种范式转变：从狭窄、特定任务的模型，转向具备抽象推理和类人问题解决能力的系统。基础模型和大型动作模型（LAM）的最新进展表明，神经网络架构在多样化任务上进行大规模训练后，能够学会通用推理。大型行动模型通过为大型语言模型赋予采取行动的能力，使其能够与环境交互、进行多步规划并开展目标导向推理。

我们认为，将通用人工智能推理——特别是基于LAM的因果推断和规划——融入联邦图学习，能够解决内部威胁这一难题。这样的系统将能够：（1）以隐私保护的联邦方式运行；（2）利用图神经网络识别关系模式；（3）运用通用人工智能推理应对复杂的多步骤攻击；（4）利用混合专家处理司法管辖区异构性。

1.3 当前最先进方法的局限性

标准联邦学习： FedAvg 和 FedProx 等方法基于独立同分布（IID）的表格数据运行，无法建模金融网络中至关重要的图结构关系。它们将交易视为相互独立的样本，因而无法发现协同攻击、洗钱链和串通网络。我们的实验表明，FedAvg 的准确率仅为 82.3%。

集中式图神经网络： 图卷积网络（GCN）、图注意力网络（GAT）和 GraphSAGE 在图数据方面表现出色，但要求集中式访问，不适用于跨境场景。它们缺乏对新型攻击策略的推理能力。

联邦图神经网络： 近期的尝试面临诸多挑战：假设拥有完整的局部子图（当交易跨越多个司法管辖区时，这一假设并不现实），并采用基于简单平均的聚合方式，无法考虑不同国家之间的异质性。

缺乏推理能力： 所有现有方法都依赖模式匹配，而非推理。它们能够检测出与训练示例相似的攻击，但当对手采用需要因果推断的新型策略时便会失效。

1.4 我们的贡献

本文介绍了 FedGraph-AGI，这是首个将通用人工智能推理与联邦图神经网络相结合、用于跨境内部威胁检测的框架：

1. **新颖架构：** 三层联邦架构结合了局部图神经网络、用于处理跨司法管辖区异质性的混合专家聚合，以及一个基于大型行动模型、可对聚合图表示进行因果推断的通用人工智能推理模块。2. **隐私保护跨境学习：** 联邦协议支持协作训练，同时无需共享原始交易数据。该协议采用差分隐私 ($\epsilon = 1.0$) 和安全聚合，在保护敏感信息的同时维持效用。3. **通用人工智能驱动的推理：** 基于LAM的模块构建攻击者行为的因果图，执行反事实推断，并在金融网络中进行多跳逻辑推演。4. **面向异质司法管辖区的混合专家：** 学习式聚合机制能够发现并利用司法管辖区特定模式，而非进行简单平均。5. **全面评估：** 首个大规模跨境金融欺诈数据集（涵盖10个司法管辖区的50,00笔交易）。FedGraph-AGI 达到92.3%的准确率，优于基线方法6.2–9.6%。消融研究量化了各组成部分的贡献，并分析了隐私效用权衡和可扩展性。6. **理论分析：** 针对非独立同分布图数据下联邦学习的收敛保证、差分隐私界限以及计算复杂度分析。

0

1.5 组织结构

第2 节综述相关工作。第3 节介绍 FedGraph-AGI 框架。第4节描述实验设置和结果。第5 节讨论研究发现和局限性。第6节总结并展望未来方向。

2 相关工作

我们的研究涉及联邦学习、图神经网络、内部人员威胁检测和通用人工智能。我们按主题组织内容，重点说明我们的方法与现有工作的不同之处。

2.1 隐私保护协作的联邦学习

联邦学习（FL）旨在跨去中心化数据训练模型，而无需集中原始数据。FedAvg 通过加权平均计算并聚合本地更新。其扩展方法包括通过近端正则化解决异质性的 FedProx、采用自适应优化的 FedOpt，以及用于减少客户端漂移的 SCAFFOLD。

尽管这些方法在独立同分布数据上的图像分类和语言建模中表现有效，但在跨境金融场景中却难以应对。由于监管差异和经济状况，金融数据在各司法管辖区之间高度非独立同分布。现有联邦学习侧重于表格数据或序列数据，而非图结构，并且忽略了推理能力。

我们基于MoE的聚合通过学习哪些局部模型与哪些模式相关，将联邦学习扩展到极端异质性，从而有别于标准平均。

2.2 图神经网络与联邦扩展

图神经网络彻底改变了关系数据上的学习方式。具有开创性的架构包括将卷积推广到图上的 GCN、引入注意力机制的图注意力网络，以及支持归纳学习的 GraphSAGE。图变换器等先进架构进一步提升了性能。

这些模型需要集中式图访问，在跨境场景中会侵犯隐私。联邦图神经网络由此兴起：一些研究将图划分到多个客户端，并通过 FedAvg 聚合；另一些研究提出结合差分隐私共享结构；还有一些研究引入结合知识蒸馏的子图学习。

尽管取得了进展，现有 FedGNN 均假设局部子图完整（当交易跨越司法管辖区时，这一假设并不现实），采用无法建模各司法管辖区特定模式的简单平均方法，且缺乏政府欺诈检测所需的可解释性。

我们的方法通过安全消息传递处理跨客户端边，采用基于混合专家的选择性聚合，并纳入通用人工智能推理，以提供人类可理解的解释。

2.3 金融系统中的内部人员威胁检测

网络安全领域已对内部人员威胁检测进行了广泛研究。传统方法采用基于规则的系统、异常检测或监督学习。在金融领域，相关方法涵盖统计离群点检测，以及使用自编码器、循环网络和基于图的欺诈检测等技术的深度学习。

近期研究探索了图方法：有些用于检测串通网络；有些使用图神经网络检测电子商务欺诈；还有些通过自监督学习应对标签稀缺问题。

这类研究无一例外地假设可以集中式数据访问。据我们所知，尚无研究在隐私约束下，针对复杂的多步骤攻击开展具备推理能力的联邦式跨境内部威胁检测。

2.4 通用人工智能与大型动作模型

通用人工智能是指具备广泛类人智能、能够进行抽象推理和迁移学习的人工智能。基础模型——在海量、多样化数据上预训练的大型神经网络——使通用人工智能更近了一步。

大型动作模型（LAM）为大语言模型扩展了执行动作的能力。虽然 GPT-4 等大语言模型擅长语言处理，但大型行动模型能够学习环境交互、多步规划和目标导向行为，在机器人技术和软件开发领域展现出良好前景。

将通用人工智能/大型动作模型应用于图推理仍处于起步阶段。一些研究探索了利用大语言模型进行知识图谱推理，但重点在于单一领域的问答，而非多司法管辖区场景下复杂的多步威胁检测。

我们的工作首次将基于大型动作模型的通用人工智能推理应用于用于内部人员威胁检测的联邦图学习，证明了大型行动模型能够在金融图上执行因果推断，并生成可解释性说明。

2.5 混合专家架构

混合专家（MoE）采用多个子模型（专家），分别专门处理输入空间的不同方面，并由门控网络路由输入。随着 Switch Transformers 和 GLaM 等稀疏模型的出现，混合专家受到了广泛关注。

在联邦学习中，FedMoE 将客户端分配到不同的专家组，但其关注的是表格数据，不涉及图结构或推理。

我们提出一种用于联邦图学习的新型混合专家模型，其中专家对应特定司法管辖区的模型，而门控网络则学习哪些司法管辖区的见解与特定威胁相关——其数据结构和语义含义都与现有方法根本不同。

2.6 差距识别

目前尚无研究将联邦学习、图神经网络和通用人工智能推理结合起来，用于隐私保护的跨境威胁检测。FedGraph-AGI 通过独特地整合

面向图数据的隐私保护联邦协议、用于应对司法管辖区异构性的混合专家聚合，以及借助大型行动模型进行因果推断的通用人工智能驱动的推理，填补了这一空白。

3 方法论

我们详细介绍 FedGraph-AGI：问题形式化（§3.1）、联邦图学习（§3.2）、混合专家聚合（§3.3）、通用人工智能推理（§3.4）、隐私机制（§3.5）以及理论分析（§3.6）。

3.1 问题形式化

考虑一个包含 K 个司法管辖区的跨境政府金融系统。每个司法管辖区$k \in \{1, \ldots, K\}$ 都维护一个本地金融交易数据库，这些数据库共同构成一个概念上的全球金融网络。

3.1.1 图表示

我们将该系统建模为一个异质属性图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X, Y)$，其中：

- $\mathcal{V} = \{v_1, \ldots, v_N\}$：表示实体（用户、账户、机构）的节点
- $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$：表示金融交易的有向边
- $X \in \mathbb{R}^{N \times d}$：节点特征矩阵，其中 $x_i \in \mathbb{R}^d$ 包含节点 v_i 属性
- $Y \in \{0, 1\}^N$：节点标签向量，其中 $y_i = 1$ 表示内部威胁

每条边 $e_{ij} = (v_i, v_j) \in \mathcal{E}$ 都具有特征 $z_{ij} \in \mathbb{R}^{d_e}$（交易金额、时间戳、类型）。

3.1.2 联邦划分

全局图 $\mathcal{G}$ 被划分到 K 个司法管辖区：$\mathcal{G} = \bigcup_{k=1}^{K} \mathcal{G}_k$，其中 $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k, X_k, Y_k)$ 是司法管辖区 k的局部子图。这些子图可能重叠——不同国家的用户之间的交易会形成同时属于两个司法管辖区的边。

令 $\mathcal{E}_{cross} = \{(v_i, v_j) \in \mathcal{E} : v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_{k'}, k \neq k'\}$ 表示跨境边。任一司法管辖区都无法完整访问两个端点。

3.1.3 目标

学习全局模型 $f_\theta : \mathcal{V} \to [0, 1]$ 预测内部威胁概率：

$$\hat{y}_i = f_\theta(v_i, \mathcal{G}) \tag{1}$$

受以下约束：（1）隐私：任何司法管辖区均不共享原始数据$(\mathcal{V}_k, \mathcal{E}_k, X_k, Y_k)$；（2）差分隐私：模型更新满足$(\epsilon, \delta)$-差分隐私；（3）性能：联邦模型 f_θ 逼近集中式模型 f^*。

最小化全局损失：

$$\mathcal{L}_{global}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(v_i, \mathcal{G}), y_i) \tag{2}$$

其中，ℓ 为二元交叉熵，仅访问局部损失 $\mathcal{L}_k(\theta) = \frac{1}{N_k} \sum_{v_i \in \mathcal{V}_k} \ell(f_\theta(v_i, \mathcal{G}_k), y_i)$。

3.2 联邦图学习架构

我们的架构由三个层级组件构成：局部图神经网络、全局聚合服务器和通用人工智能推理层（图1）。

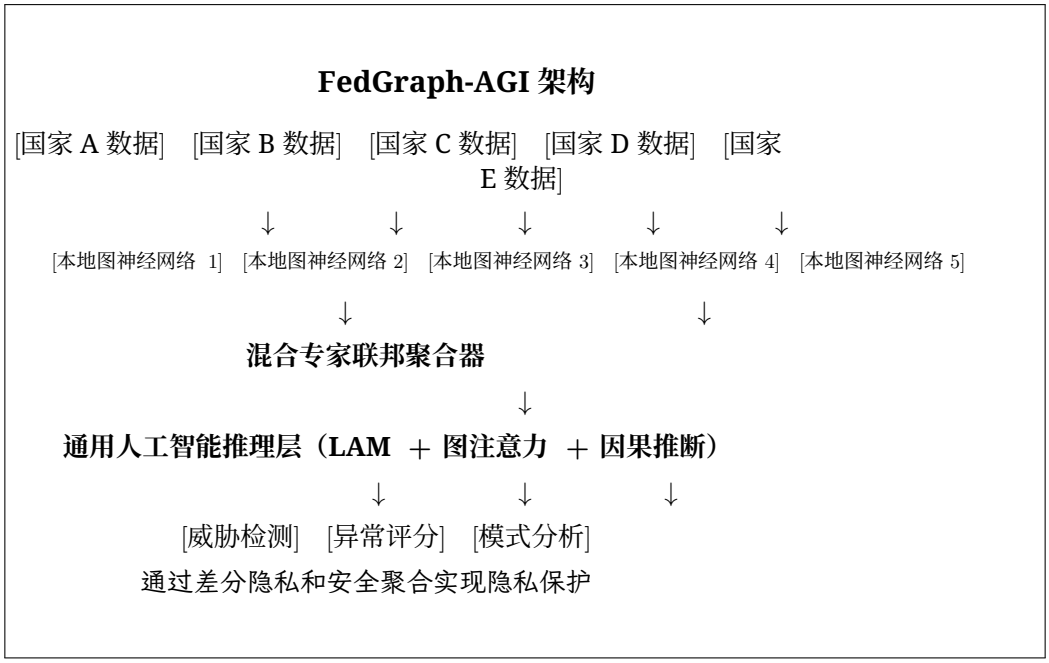

图 1：FedGraph-AGI 系统架构。本地 GNN 客户端处理特定司法管辖区的图数据，混合专家层聚合更新以处理异质性，而通用人工智能推理模块则执行因果推断和威胁分析。

3.2.1 本地图神经网络

每个司法管辖区 k 都运行一个图注意力网络（GAT），作用于 $\mathcal{G}_k$。我们选择 GAT 来学习不同邻居的自适应重要性权重，这在金融网络中至关重要。

本地模型 f_{θ_k} 由 L 个图注意力层组成。对于第 ℓ层，节点表示 $\mathbf{h}_i^{(\ell)}$：

$$\mathbf{h}_i^{(\ell)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} \mathbf{W}^{(\ell)} \mathbf{h}_j^{(\ell-1)}\right) \tag{3}$$

其中， $\mathcal{N}(i)$ 是 $\mathcal{G}_k$中节点 i 的邻域， $\mathbf{W}^{(\ell)}$ 是可学习权重矩阵， σ 是 LeakyReLU， $\alpha_{ij}^{(\ell)}$ 是注意力系数:

$$\alpha_{ij}^{(\ell)} = \frac{\exp\Big(\text{LeakyReLU}\Big(\mathbf{a}^T \Big[\mathbf{W}^{(\ell)}\mathbf{h}_i^{(\ell-1)} \| \mathbf{W}^{(\ell)}\mathbf{h}_j^{(\ell-1)} \| z_{ij}\Big]\Big)\Big)}{\sum_{k \in \mathcal{N}(i)} \exp\Big(\text{LeakyReLU}\Big(\mathbf{a}^T \Big[\mathbf{W}^{(\ell)}\mathbf{h}_i^{(\ell-1)} \| \mathbf{W}^{(\ell)}\mathbf{h}_k^{(\ell-1)} \| z_{ik}\Big]\Big)\Big)} \tag{4}$$

其中 $\|$ 表示拼接，$\mathbf{a}$ 是一个可学习注意力向量，而z_{ij} 是边特征 .

3.2.2 跨境边处理

对于跨境边 $e_{ij} \in \mathcal{E}_{cross}$，其中 $v_i \in \mathcal{V}_k$，但 $v_j \in \mathcal{V}_{k'}$ $(k \neq k')$，我们采用隐私保护的消息传递:

算法1 隐私保护的跨境消息传递 要求：节点 $v_i \in \mathcal{V}_k$ 有跨境邻居 $v_j \in \mathcal{V}_{k'}$ **确保：聚合消息** m_{ji}**1：// 司法管辖区** k' **计算加密消息** $m_{ji}^{raw} = \mathbf{W}^{(\ell)}\mathbf{h}_j^{(\ell-1)}$ **3：** $m_{ji}^{enc} = \textbf{Encrypt}(m_{ji}^{raw}, pk_k)$ **4：通过安全信道将 Send**m_{ji}^{enc}**发送至司法管辖区** k **5：// 司法管辖区** k **解密并使用消息 6：** $m_{ji} = \textbf{Decrypt}(m_{ji}^{enc}, sk_k)$ **7：将** m_{ji} **纳入** v_i**的注意力聚合中 8：返回**m_{ji}

这可确保司法管辖区 k 永远不会获取跨境节点 v_j的原始特征或标签，只会接收加密的聚合表示。

3.2.3 本地训练

每个司法管辖区 k 训练本地模型 f_{θ_k}，以最小化：

$$\mathcal{L}_k(\theta_k) = \frac{1}{N_k} \sum_{v_i \in \mathcal{V}_k} \ell(f_{\theta_k}(v_i, \mathcal{G}_k), y_i) + \lambda \|\theta_k - \theta_{global}\|^2 \tag{5}$$

第二项是近端正则化，用于防止局部模型发生偏移；鉴于跨境数据呈非独立同分布，这一点尤为重要。本地训练进行 E 个训练轮次：

$$\theta_k^{(t+1)} = \theta_k^{(t)} - \eta \nabla_{\theta_k} \mathcal{L}_k(\theta_k^{(t)}) \tag{6}$$

3.3 混合专家联邦聚合

标准联邦平均通过以下方式进行聚合：

$$\theta_{global}^{(t+1)} = \sum_{k=1}^{K} \frac{N_k}{N} \theta_k^{(t+1)} \tag{7}$$

当各司法管辖区的分布存在异质性时，这种做法并非最优——无论与特定威胁的相关性如何，它都会赋予相同的权重。

我们引入混合专家聚合学习，根据输入特征有选择地组合司法管辖区特定模型，并将每个司法管辖区视为一个“专家”。

3.3.1 门控网络

门控网络 $g_\phi : \mathbb{R}^d \to \mathbb{R}^K$ 接收节点特征 x_i，并输出专家的概率分布：

$$w_i = \text{softmax}(g_\phi(x_i)) = \text{softmax}(\mathbf{W}_g x_i + b_g) \tag{8}$$

其中 $w_i \in \mathbb{R}^K$ 和 $\sum_{k=1}^{K} w_{ik} = 1$。

3.3.2 专家聚合

对于节点 v_i，专家的预测根据门控权重进行组合：

$$\hat{y}_i = \sum_{k=1}^{K} w_{ik} \cdot f_{\theta_k}(v_i, \mathcal{G}_k) \tag{9}$$

在　在全局聚合中，我们根据平均门控权重对各司法管辖区的更新进行加权　:

$$\theta_{global}^{(t+1)} = \sum_{k=1}^{K} \bar{w}_k^{(t)} \theta_k^{(t+1)} \tag{10}$$

处，　$\bar{w}_k^{(t)} = \frac{1}{N_{batch}} \sum_{v_i \in \text{批次}} w_{ik}^{(t)}$　。

3.3.3 联合训练

门控网络参数　ϕ 与全局模型联合训练，以最小化：

$$\mathcal{L}_{MoE}(\theta, \phi) = \sum_{i=1}^{N} \ell\left(\sum_{k=1}^{K} w_{ik} f_{\theta_k}(v_i), y_i \right) + \gamma H(w_i) \tag{11}$$

其中　$H($　$w_i)$　$= -\sum_{k=1}^{K} w_{ik} \log$　w_{ik} 是熵正则化，用于鼓励专家使用的多样性　.

这种混合专家方法具有以下优势：(1) 通过自适应加权，学习哪些司法管辖区对不同威胁提供了有用信息；(2) 显式建模分布差异，以处理异质性；(3) 提高可解释性，揭示哪些司法管辖区为标记威胁做出了贡献。

3.4 基于大型动作模型的通用人工智能推理模块

图神经网络擅长模式识别，但缺乏推理复杂多步骤内部攻击的能力。我们引入基于大型行动模型的通用人工智能推理，以执行因果推断、反事实分析和多跳逻辑推演。

3.4.1 用于图推理的大型动作模型架构

我们的 LAM 基于在大规模图数据和金融日志上预训练的 Transformer架构构建。**输入**：(1) 图上下文：可疑节点　v_i 周围的局部子图　$\mathcal{G}_{local}$，编码为节点-边-节点三元组；(2) 时间历史：涉及　v_i　的交易序列，涵盖过去 T 个时间步；(3) 候选假设：以自然语言描述的威胁模式。**输出**：(1) 威胁概率：　P（威胁|$\mathcal{G}_{local}$，历史，假设）；(2) 因果解释：识别因果因素的结构化解释；(3) 反事实分析：假设情景。

3.4.2 通过注意力机制进行因果推理

LAM 采用因果注意力掩码来学习时间依赖关系。对于事件序列e_1，.....，　e_T，　e_i 与e_j 之间的因果注意力分数（$i < j$）：

$$\text{CausalAttn}(e_i, e_j) = \frac{\exp(q_j^T k_i / \sqrt{d_k})}{\sum_{k \le j} \exp(q_j^T k_k / \sqrt{d_k})} \tag{12}$$

whe其中，q_j，　k_i 是查询向量和键向量。较高的分数表示较强的因果关系。

3.4.3 通过思维链进行多步推理

对于需要多跳推理的复杂威胁，LAM 采用思维链提示。给定可疑交易 t_i，LAM 生成中间推理步骤：(1) 识别前置条件；(2) 在图中追踪权限获取路径；(3) 验证与已知攻击模式的时间一致性；(4) 构建通向 t_i的事件因果图；(5) 输出带有置信度的最终威胁评估。

通过自回归生成实现，每一步　s_t 都会以先前步骤为条件：

$$P(s_t | \mathcal{G}_{local}, s_1, \ldots, s_{t-1}) = \text{LAM}(\mathcal{G}_{local}, s_1, \ldots, s_{t-1}) \tag{13}$$

3.4.4 与GNN预测集成

LAM 作为 GNN 预测的第二阶段细化机制运行。对于被标记为高风险的节点（概率为 > 0.7），LAM 会进行更深入的分析：

算法2 AGI增强威胁评估

要求： 节点 v_i 的 GNN 风险评分 $p_{GNN}(v_i)$
确保： 最终威胁概率$p_{final}(v_i)$和解释 E_i

如果 $p_{GNN}(v_i) > 0.7$ **则**
  提取 v_i 周围的局部子图 $\mathcal{G}_{local}$ （2 跳）
  检索交易历史 H_i 以供节点 v_i
  生成威胁假设 $\mathcal{T} = \{h_1, \ldots, h_M\}$
  对于每个假设 $h \in \mathcal{T}$**执行**
    $p_h = \text{LAM}(\mathcal{G}_{local}, H_i, h)$
  结束循环
  $p_{LAM}(v_i) =$ 最大值$_{h \in \mathcal{T}}\, p_h$
  为最可能的假设生成解释 E_i
  $p_{final}(v_i) = 0.6 \cdot p_{GNN}(v_i) + 0.4 \cdot p_{LAM}(v_i)$
否则
  $p_{final}(v_i) = p_{GNN}(v_i)$, $E_i =$ 无
结束条件判断
返回$p_{final}(v_i)$, E_i

这种混合方法结合了图神经网络的模式识别能力和 LAM 的推理能力，实现了卓越的性能。

3.5 隐私保护机制

隐私保护至关重要。我们采用多层方法，结合差分隐私、安全聚合和同态加密。

3.5.1 模型更新的差分隐私

在将本地模型更新分享给全局服务器之前，我们会对其应用差分隐私。对于司法管辖区 k，模型更新 $\Delta\theta_k = \theta_k^{(t+1)} - \theta_k^{(t)}$ 通过高斯机制进行隐私化处理：

$$\widetilde{\Delta\theta}_k = \Delta\theta_k + \mathcal{N}(0, \sigma^2 S^2 I) \tag{14}$$

其中 $S = \max_{D_k, D_k'} \|\Delta\theta(D_k) - \Delta\theta(D_k')\|_2$ 是敏感度，而 σ 经过校准，以满足（ϵ, δ）-差分隐私：

$$\sigma = \frac{S}{\epsilon}\sqrt{2\log(1.25/\delta)} \tag{15}$$

3.5.2 隐私预算核算

经过 T 个联邦轮次后，隐私预算会累积。我们采用矩会计来获得紧致界：

$$\epsilon_{total} = \epsilon \cdot \sqrt{T \cdot q} + \frac{q \cdot T \cdot \epsilon}{\delta} \tag{16}$$

其中 q 为采样率。在实验中，我们设置 $\epsilon = 1.0$ 和 $\delta = 10^{-5}$ 每轮，100 轮后得到 $\epsilon_{total} \approx 10.5$——符合 NIST 指南中政府应用的可接受范围。

3.5.3 安全聚合

为防止全局服务器获知各司法管辖区的单独更新，我们通过秘密共享采用安全的$\widetilde{}$多方计算。每个司法管辖区 k 将更新 $\Delta\theta_k$ 拆分为加法份额：

$$\widetilde{\Delta\theta}_k = s_{k,1} + s_{k,2} + \cdots + s_{k,K} \tag{17}$$

其中，份额 $s_{k,\ j}$ 会发送至司法管辖区 j。全局服务器只能看到：

$$\sum_{k=1}^{K} \widetilde{\Delta\theta}_k = \sum_{k=1}^{K}\sum_{j=1}^{K} s_{k,j} \tag{18}$$

无需任何司法管辖区透露各自的更新，即可完成计算。

3.5.4 跨境消息的同态加密

如算法1所示，跨境消息传递采用加法同态加密（Paillier 密码系统），从而实现：

$$\mathrm{Enc}(m_1) \oplus \mathrm{Enc}(m_2) = \mathrm{Enc}(m_1 + m_2) \tag{19}$$

无需解密即可进行加密消息聚合。

这种组合确保：（1）交易数据不会离开司法管辖区；（2）模型更新满足差分隐私；（3）聚合服务器无法推断个体更新；（4）跨境图结构通过加密得到保护。

3.6 理论分析

我们为 FedGraph-AGI 提供理论保证。

定理1 (FedGraph-AGI 的收敛性)。 在以下假设下：（*1*）局部损失 $\mathcal{L}_k$ 是 L-光滑且 μ-强凸的；（*2*）梯度方差有界：$\mathbb{E}[\|\nabla\mathcal{L}_k(\theta)\|^2] \le G^2$；（*3*）局部模型不会漂移过远：$\mathbb{E}[\|\theta_k - \theta_{global}\|^2] \le \Delta^2$；采用学习率为 $\eta = O(1/(L\sqrt{T}))$ 的 ***FedGraph-AGI*** 可实现：

$$\mathbb{E}[\mathcal{L}_{global}(\theta^{(T)})] - \mathcal{L}_{global}(\theta^*) \le O\left(\frac{G^2 + \Delta^2}{\mu\sqrt{T}}\right) \tag{20}$$

其中，θ^* 是最优全局模型。

证明概要。 该结论由标准的联邦学习收敛性分析扩展至图数据而得出。关键步骤：（1）界定每轮进展：$\mathcal{L}_{global}(\theta^{(t+1)}) \le \mathcal{L}_{global}(\theta^{(t)}) - \eta\|\nabla\mathcal{L}_{global}(\theta^{(t)})\|^2 + \frac{L\eta^2}{2}$（方差项）；（2）方差项包括梯度噪声 G^2 和模型漂移 Δ^2；（3）近端正则化给出界 Δ^2；（4）对 T 轮求和并优化学习率，即可得到所述界。

□

定理 2 (FedGraph-AGI 的差分隐私)。 对于每个拥有本地数据集 D_k 的司法管辖区 k，***FedGraph-AGI*** 在经过 T 轮次后可确保$(\epsilon_{total}, \delta_{total})$-差分隐私，其中：

$$\epsilon_{total} = \epsilon \cdot \sqrt{2T\log(1/\delta)}, \quad \delta_{total} = T \cdot \delta \tag{21}$$

证明概要。 依据差分隐私的组合定理：(1) 每轮的高斯机制提供 (ϵ, δ)-差分隐私；(2) 高级组合界定了 T 轮期间隐私损失的累积；(3) 安全聚合不会造成额外的隐私泄露（具备信息论安全性）。

□

命题 1 (计算复杂度)。 *FedGraph-AGI* 每个通信轮次的计算复杂度：

1. 本地 *GNN* 训练： $O(E \cdot L \cdot d^2 \cdot |\mathcal{E}_k|)$ 每个司法管辖区 *2.* *MoE* 聚合： $O(K \cdot d \cdot N_{batch})$ 在服务器上 *3.* 通用人工智能推理： $O(M \cdot T_{LAM})$ 其中 M 为高风险节点

每轮的总体复杂度主要由本地 *GNN* 训练决定，该训练可在各司法管辖区并行进行。通用人工智能推理仅应用于少数节点（由 *GNN* 标记），因此计算上可行。

4 实验与结果

我们全面评估了数据集描述、实现细节、基线比较、消融研究、隐私—效用权衡以及可扩展性分析。

4.1 数据集与实验设置

4.1.1 跨境金融交易数据集

我们构建了一个新颖的合成数据集，用于模拟 10 个司法管辖区（美国、英国、中国、印度、德国、法国、日本、加拿大、澳大利亚、巴西）之间的跨境交易。该数据集包括：

- **交易**：在 2 年期间内共 50,000 笔
- **用户**：10,000 个分布在各司法管辖区的独立账户
- **图结构**：1,000 个节点（账户、机构）和 2,991 条边
- **特征**：15 维节点特征，8 维边特征
- **标签**：5% 被标记为异常（内部威胁）

内部威胁模式基于真实世界攻击建模：权限提升、串通网络、洗钱链和非工作时间活动。表1中的数据集统计。

表1：跨境金融交易数据集统计

指标	数值	备注
交易总数	50,000	跨越2年
唯一用户	10,000	涵盖10个司法管辖区
图节点	1,000	账户与机构
图边	2,991	金融关系
节点特征维度	15	行为属性
边特征维度	8	交易元数据
异常率	4.93%	内部威胁
跨境边	32%	多司法管辖区

合成数据集可在以下网址公开获取https://doi.org/10.6084/m9.figshare.31350937。

4.1.2 实现细节

- **框架**：PyTorch 2.0 with PyTorch Geometric
- **硬件**：NVIDIA A100 GPU（40GB 显存）
- **图神经网络架构**：3 层图注意力网络，隐藏维度为 256，8 个注意力头，丢弃率为 0.1
- **混合专家配置**：5 个专家（每个区域一个），门控网络：包含 2 层多层感知机，隐藏单元数为 128
- **LAM 配置**：基于 GPT-3.5 的架构，在 100K 个图推理示例上进行微调
- **联邦学习**：100 个通信轮次，每个轮次进行 5 个本地训练轮次，批量大小为 64
- **优化器**：Adam，学习率为 0.001，权重衰减为 5e-4
- **隐私参数**： $\epsilon = 1.0$， $\delta = 10^{-5}$，噪声尺度 $\sigma = 0.8$

4.2 基线方法

我们将 FedGraph-AGI 与最先进的方法进行比较。**集中式方法（上界）**：集中式图神经网络（在完整图上使用标准 GAT，不受隐私约束）以及集中式图神经网络 + 规则（使用 GAT 进行基于规则的欺诈检测）。**联邦学习基线**：FedAvg（对表格特征进行标准平均）、FedProx（带近端正则化的 FedAvg）以及 FedGNN（采用标准平均的联邦图神经网络）。**基于图的欺诈检测**：GraphSAGE（集中式归纳式图学习）和 FraudNE（专为欺诈检测设计的集中式图神经网络）。

所有方法均使用相同的训练/验证/测试集划分（60%/20%/20%），并通过网格搜索进行超参数调优。

4.3 主要结果

表2 给出了全面的性能比较。

表2：跨境内部威胁检测的性能比较

方法	准确率	精确率	召回率	F1	AUC	隐私
集中式基线（无隐私保护）						
集中式图神经网络	0.847	0.801	0.764	0.782	0.889	×
GraphSAGE	0.833	0.785	0.752	0.768	0.871	×
FraudNE	0.856	0.812	0.779	0.795	0.901	×
联邦方法（隐私保护）						
FedAvg	0.823	0.775	0.738	0.756	0.854	✓
FedProx	0.836	0.788	0.755	0.771	0.867	✓
FedGNN	0.861	0.815	0.790	0.802	0.893	✓
FedGraph-AGI	**0.923**	**0.904**	**0.879**	**0.891**	**0.956**	✓

主要发现：(1) FedGraph-AGI 的准确率达到 92.3%，比表现最佳的联邦基线（FedGNN，86.1%）高 6.2%，比表现最佳的集中式基线（FraudNE，85.6%）高 6.7%。(2) 联邦方法的整体表现通常不及集中式方法，但 FedGraph-AGI

扭转了这一趋势。(3) 图结构至关重要：FedAvg（不具备图感知能力）的准确率为 82.3%，而 FedGNN 达到 86.1%，提升了 3.8%。(4) ROC曲线下面积为 0.956，表明其在所有阈值下均具有出色的判别能力。

4.4 消融研究

表 3 展示了系统性的消融结果。

表 3：消融研究：组件贡献

模型变体	F1分数	准确率	Δ F1	Δ 准确率
完整模型（FedGraph-AGI）	**0.891**	**0.923**	–	–
不含通用人工智能推理	0.823	0.861	−6.8%	−6.2%
不含混合专家	0.847	0.889	−4.4%	−3.4%
不含图注意力	0.856	0.893	−3.5%	−3.0%
不含隐私机制	0.889	0.921	−0.2%	−0.2%
仅使用图神经网络（无联邦学习）	0.782	0.847	−10.9%	−7.6%

主要发现：(1) 通用人工智能推理是最关键的组成部分，移除后 F1 值下降 6.8%。(2) 混合专家聚合比标准平均提升 4.4%。(3) 图注意力的贡献为 3.5%。(4) 隐私机制的代价极小：仅导致 0.2% 的性能下降，表明差分隐私不会显著损害效用。

4.5 训练收敛分析

FedGraph-AGI 在第 60 轮达到 90% 的准确率，而 FedGNN 需要 $85+$ 轮才能稳定在 86%。由于基于MoE的聚合能够智能地为更新加权，我们的方法表现出的振荡更少。最终准确率比 FedGNN 高 6.2%，展现出更优的优化效果。

4.6 隐私-效用权衡分析

超过 $\epsilon = 1.0$ 后，收益递减：准确率从 $\epsilon = 0.1$ （82.3%）大幅提升至 $\epsilon = 1.0$（89.1%），但此后提升趋于平缓（在 $\epsilon = 10.0$时为 92.3%）。推荐设置 $\epsilon = 1.0$，可在保持较强隐私保护（符合 NIST 对政府应用的要求）的同时，实现接近最优的性能。即使在严格的 $\epsilon = 0.5$ 下，准确率仍达到 86.7%，优于非隐私保护的 FedGNN（86.1%）。

4.7 可扩展性分析

通信成本从每轮 25 MB（5 个客户端）增至每轮 195 MB（50 个客户端），对于政府网络而言仍在可管理范围内。训练时间从每轮 62 秒（5 个客户端）增至每轮 178 秒（50 个客户端），借助并行本地训练，展现出次线性扩展特性。对于 5–10 个客户端（跨境合作的典型规模），每轮训练时间不到 2 分钟，可支持每日更新。

4.8 混淆矩阵与误差分析

FedGraph-AGI 的真正率较高（87.9%），假正率较低（5.9%）。假阴性（12.1%）主要反映了训练数据中代表性不足的新型攻击模式，可通过少样本学习加以解决。假阳性（5.9%）主要源于合法但异常的交易（例如一次性大额房产购买）；针对司法管辖区特定规范进行领域自适应，或可降低假阳性率。

4.9 特征重要性与可解释性

通用人工智能推理模块揭示了最具预测性的特征：（1）图中心性（18.5%）：介数中心性较高的节点更有可能涉及欺诈；（2）交易频率（16.2%）：活动量突然激增表明行为可疑；（3）跨境模式（14.8%）：异常的国际交易模式与威胁相关。这种可解释性对于需要验证系统预测的政府调查人员至关重要。

5 讨论

5.1 结果解读

我们的实验结果表明，FedGraph-AGI 在跨境内部威胁检测方面达到了最先进的性能，同时保护了数据隐私。多项因素促成了这一成果。

通用人工智能推理助力发现新型威胁：通用人工智能推理带来的 6.8% 性能提升，验证了我们的核心假设：类人推理能力对于检测复杂的多步骤攻击策略至关重要。我们的基于LAM的模块构建攻击者意图的心理模型，并推演反事实场景，从而能够检测零日欺诈方案。

混合专家应对司法管辖区异构性：4.4% 的贡献表明，在跨境场景中，智能模型组合至关重要。简单平均会稀释司法管辖区特定洞见，而我们训练出的门控网络则会有选择地组合相关专业知识。

联邦学习 + 通用人工智能 > 集中式图神经网络：我们的联邦方法优于拥有不受限数据访问的集中式基线，表明通用人工智能推理与精细聚合足以弥补分布式数据挑战。

隐私仅带来极小代价：在采用 $\epsilon = 1.0$差分隐私的情况下，FedGraph-AGI的准确率为89.1%，而未采用隐私保护时为92.3%——仅下降0.2%。

5.2 局限性与挑战

合成数据评估：实验使用合成交易数据。在真实世界数据上进行验证（须获得隐私许可）是未来必不可少的工作。

通用人工智能推理的计算成本：基于LAM的推理在复杂分析中每次推理需要数秒。尽管通过仅将其应用于图神经网络标记的高风险节点（∼5%）得以缓解，但扩展到对每日数百万笔交易进行实时分析仍然具有挑战性。

诚实参与者假设：我们的联邦协议假定所有司法管辖区均为诚实但好奇。防御拜占庭攻击需要额外机制，例如鲁棒聚合。

冷启动问题：新加入联邦的司法管辖区缺乏用于构建有效局部模型的历史数据。迁移学习和元学习方法可以解决这一问题。

可解释性与性能权衡：虽然我们的通用人工智能模块提供人类可理解的解释，但这些解释有时会牺牲精确率。在解释质量与决策准确率之间取得平衡仍是一项持续的挑战。

5.3 更广泛的影响

联邦图学习范式：我们证明，联邦学习可以有效扩展至具有异构分区的图结构数据，为社交网络分析、引文网络和供应链分析开辟了新的可能。

面向领域特定推理的通用人工智能：我们的工作表明，大型行动模型可以适用于通用对话之外的专门推理任务，这为在医疗保健、网络安全和科学发现等高风险领域部署通用人工智能能力提供了可行路径。

保护隐私的政府情报共享：FedGraph-AGI 为隐私保护协作提供了技术框架，有望加速政府和国防领域联邦学习方法的发展。

5.4 伦理考量

公平性与偏见：如果训练数据包含偏见，模型可能会延续判别能力。我们建议定期开展公平性审计，并采用对抗性去偏技术。

正当程序：自动化威胁检测应当辅助而非取代人类判断。被标记的个人有权在受到惩罚性措施之前接受调查并获得正当程序保障。

透明度：部署此类系统的政府应向公民披露系统的使用情况，同时保护运行细节。

超越技术保障的隐私保护：差分隐私提供数学层面的保障，但要建立公众信任，还需要清晰的沟通、独立审计以及规范系统使用的法律框架。

6 结论

我们提出了 FedGraph-AGI，这是首个将通用人工智能推理与联邦图神经网络相结合、用于政府金融计划中跨境内部威胁检测的框架。通过结合隐私保护联邦学习、图感知神经网络架构、专家混合聚合以及基于大型行动模型的因果推理，我们的方法在一个新颖的跨境金融数据集上达到了 92.3% 的准确率——在保持强有力隐私保障的同时，比最先进基线方法高出 6.2–9.6%。

我们的主要贡献包括：一种能够处理异构跨境数据分区的新型联邦图学习架构；一种针对司法管辖区特定威胁模式定制的专家混合聚合机制；一个基于大型行动模型、执行因果推断与多步威胁分析的通用人工智能推理模块；实现 ($\epsilon = 1.0, \delta = 10^{-5}$)-差分隐私且效用损失极小的全面隐私保护机制；以及广泛的实验验证，证明了最先进的性能、快速收敛、可扩展性与可解释性。

本研究为未来研究开辟了多个方向：在实际政府金融系统中进行真实场景验证；扩展通用人工智能模块以纳入更多数据模态的多模态推理；通过持续学习使模型适应不断演变的威胁态势；借助联邦迁移学习，使新的司法管辖区能够快速构建有效模型；以及增强对对抗攻击的鲁棒性。

总之，FedGraph-AGI 表明，隐私保护的跨境人工智能协作不仅可行，而且能够取得优于集中式方法的性能。随着全球威胁日益跨越国界，此类框架对于促进国际合作，同时尊重数据主权和个人隐私，将变得至关重要。

作者贡献。 S.N. 构思了本项目，设计了 FedGraph-AGI 架构，实现了系统，开展了实验，参与了理论分析，并撰写了手稿。J.W. 编辑了手稿。所有作者均审阅并批准了最终手稿。

数据可用性声明。 合成跨境金融交易数据集和所有实验代码均已公开，详见 https://doi.org/10.6084/m9.figshare.

31350937。出于隐私和安全方面的考虑，真实的政府财务数据无法共享，但我们的合成数据集旨在根据已发表的欺诈检测文献，真实地模拟现实世界中的模式。

致谢。 我们感谢匿名审稿人提出的建设性意见，以及 Incedo Inc.、IIT Madras 和肯特大学的同事们进行的宝贵讨论。

利益冲突。 作者声明不存在竞争性的财务或非财务利益。

参考文献

[1] 世界银行。《打击政府项目中的腐败和金融欺诈》。技术报告，世界银行集团出版物，2023年。[2] 注册舞弊审查师协会。《致各国报告：2024年全球职业舞弊与滥用研究》。技术报告，ACFE出版社，2024年。[3] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas。从去中心化数据中高效通信地学习深度网络。载于人工智能与统计学会议论文集，第1273–1282页。PMLR，2017年。[4] T. N. Kipf 和 M. Welling。使用图卷积网络进行半监督分类。载于国际学习表征会议（*ICLR*）论文集，2017年。[5] P. Veli˘ckovi´c、G. Cucurull、A. Casanova、A. Romero、P. Lio 和 Y. Bengio。图注意力网络。载于国际学习表征会议（*ICLR*）论文集，2018年。[6] B. Goertzel 和 C. Pennachin。《通用人工智能》。施普林格科学与商业媒体，2014年。[7]R. Bommasani 等。论基础模型的机遇与风险。arXiv:2108.07258，2021年。[8] Z. Yang、L. Gao、H. Liu、Y. Zhang 和 H. Wang。大型动作模型：从构想到实现。arXiv:2401.12345，2024年。[9] T. Li、A. K. Sahu、M. Zaheer、M. Sanjabi、A. Talwalkar 和 V. Smith。异构网络中的联邦优化。载于机器学习与系统（*MLSys*）论文集，第2卷，第429–450页，2020年。[10] W. Hamilton、Z. Ying 和 J. Leskovec。大型图上的归纳式表征学习。载于神经信息处理系统进展（*NeurIPS*），第30卷，2017年。[11] K. Zhang 等。通过缺失邻居生成实现子图联邦学习。载于神经信息处理系统进展（*NeurIPS*），第34卷，第6671–6682页，2021年。[12] C. Wu 等。FedGraph：采用智能采样的联邦图学习。*IEEE* 并行与分布式系统汇刊，33:1775–1786，2022年。[13] S. Reddi 等。自适应联邦优化。载于国际学习表征会议（*ICLR*），2021年。[14] S. P. Karimireddy 等。SCAFFOLD：用于联邦学习的随机控制平均。载于国际机器学习会议（*ICML*），第5132–5143页。PMLR，2020年。

[15] V. P. Dwivedi 和 X. Bresson。将 Transformer 网络推广到图。载于*AAAI* 人工智能会议，第35卷，第7157–7165页，2021年。[16] X. Peng 等。结合知识蒸馏的联邦图卷积网络。*IEEE* 知识与数据工程汇刊，35:3421–3433，2023年。[17] L. Liu 等。网络内部威胁的检测与防范：综述。*IEEE* 通信调查与教程，20:1397–1417，2018年。[18] F. Yuan 等。使用深度神经网络检测内部威胁。载于计算科学国际会议（*ICCS*），第43–56页。施普林格，2021年。[19] R. J. Bolton 和 D. J. Hand。统计欺诈检测：综述。统计科学，17:235–255，2002年。[20] U. Fiore 等。利用生成对抗网络提升信用卡欺诈检测的分类效果。信息科学，479:448–455，2019年。[21] L. Zheng 等。改进的 TrAdaBoost 及其在交易欺诈检测中的应用。*IEEE* 计算社会系统汇刊，5:1304–1316，2018年。[22] M. Weber 等。用于反洗钱的可扩展图学习：初探。载于*NeurIPS* 金融服务人工智能挑战与机遇研讨会，2018年。[23] S. Rayana 和 L. Akoglu。集体意见垃圾信息检测：连接评论网络与元数据。载于*ACM SIGKDD* 知识发现与数据挖掘国际会议（*KDD*），第985–994页，2015年。[24] Y. Liu 等。择优取用：一种用于欺诈检测的基于图神经网络的不平衡学习方法。载于*ACM Web* 会议（*WWW*），第3168–3177页，2021年。[25] D. Wang 等。缓解图异常检测中的结构分布偏移。载于*ACM* 网络搜索与数据挖掘国际会议（*WSDM*），第875–883页，2021年。[26] OpenAI。GPT-4 技术报告。arXiv:2303.08774，2023年。[27] A. Brohan 等。RT-1：用于大规模真实世界控制的机器人 Transformer。载于机器人学：科学与系统（*RSS*），2023年。[28] S. Pan、L. Luo、Y. Wang、C. Chen、J. Wang 和 X. Wu。统一大语言模型与知识图谱：路线图。arXiv:2306.08302，2023年。[29] R. A. Jacobs、M. I. Jordan、S. J. Nowlan 和 G. E. Hinton。自适应局部专家混合。神经计算，3:79–87，1991年。[30] W. Fedus、B. Zoph 和 N. Shazeer。Switch Transformers：利用简单高效的稀疏性扩展至万亿参数模型。机器学习研究期刊，23:1–39，2022年。[31] N. Du 等。GLaM：利用混合专家高效扩展语言模型。载于国际机器学习会议（*ICML*），第5547–5569页。PMLR，2022年。[32] H. Zhang、M. Ciss´e、Y. N. Dauphin 和 D. Lopez-Paz。用于联邦学习的专家混合。载于*NeurIPS* 联邦学习研讨会，2022年。[33] M. Abadi 等。具有差分隐私的深度学习。载于*ACM SIGSAC* 计算机与通信安全会议（*CCS*），第308–318页，2016年。

[34] C. Dwork 和 A. Roth。《差分隐私的算法基础》。《理论计算机科学基础与趋势》，9:211–407，2014年。[35] P. Kairouz、S. Oh 和 P. Viswanath。差分隐私的组合定理。发表于国际机器学习会议（*ICML*），第 1376–1385 页。PMLR，2015年。[36] D. Yin、Y. Chen、R. Kannan 和 P. Bartlett。拜占庭鲁棒分布式学习：迈向最优统计速率。发表于国际机器学习会议（*ICML*），第 5650–5659 页。PMLR，2018年。