基于同态加密的 AI 模型参数安全计算与防泄露方法

张　恒，廖尚斌，张陈颖
（中国移动通信集团福建有限公司，福建　福州　350000）

摘　要：随着人工智能在医疗、金融等敏感领域的广泛应用，模型参数与训练数据的隐私保护成为关键问题。提出一种基于同态加密（HE）的 AI 模型参数安全计算与防泄露方法，采用 CKKS 方案在密文空间中实现参数加密、前向推理与梯度更新，避免了训练过程中明文暴露的风险。结果表明，HE-SGD 在 MNIST 上最高准确率达 99.1%；在计算开销上，实现了效率与安全性的平衡，信息泄露风险指数接近 0.0。研究表明，该方法在保持模型精度的同时，实现了高效安全计算与近乎零泄露风险，具有较强的应用价值。

关键词：模型参数；隐私保护；同态加密；CKKS 方案；梯度更新
中图分类号：TP309　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097－1788.2025.11.002
引用格式：张恒，廖尚斌，张陈颖. 基于同态加密的 AI 模型参数安全计算与防泄露方法［J］. 网络安全与数据治理，2025，44(11)：7－11，17.

Secure computation and anti-leakage methods for AI model parameters based on homomorphic encryption

Zhang Heng，Liao Shangbin，Zhang Chenying
（China Mobile Communications Group Fujian Co.，Ltd.，Fuzhou 350000，China）

Abstract：With the extensive application of artificial intelligence in sensitive fields such as healthcare and finance，the privacy protection of model parameters and training data has become a critical issue. This paper proposes a secure computation and anti-leakage method for AI model parameters based on homomorphic encryption（HE）. The method employs the CKKS scheme to implement parameter encryption，forward inference，and gradient updates in the ciphertext space，thereby avoiding the risk of plaintext exposure during training. The results demonstrate that HE-SGD achieves a maximum accuracy of 99.1% on MNIST. In terms of computational overhead，it balances efficiency and security，with an information leakage risk index close to 0.0. The study indicates that the proposed method maintains model precision while achieving efficient and secure computation with nearly zero leakage risk，showing strong application value.

Key words：model parameters；privacy protection；homomorphic encryption；CKKS scheme；gradient updates

0　引言

伴随人工智能（Artificial Intelligence，AI）的兴起，其被广泛应用于各行各业中，特别是金融、医疗、教育等敏感领域，模型参数与训练数据的安全性逐渐成为研究热点[1]。深度神经网络等模型通常依赖大规模参数存储与分布式计算，然而在模型部署与参数更新过程中，存在参数泄露与隐私攻击的风险[2]。当前，该领域研究普遍面临一个关键瓶颈：计算效率与精度的矛盾。因此，如何在保证模型精度的前提下，有效提升同态加密的计算效率，已成为推动隐私保护与 AI 发展的核心挑战。

针对参数泄露风险的研究，赵宁等人引入 LeNet-5 卷积神经网络获取敏感数据随机访问共振时序特征，预测敏感数据泄露风险[3]。傅东波等人解析加密算法演进路径，以应对网络攻击[4]。龙勇在研究中提到，将差分隐私、同态加密及多方安全计算等技术相结合，可提高联邦学习系统的安全性与抗攻击能力[5]。Ahmad 等人开发了一个分室式网络流行病模型，用于分析恶意代码在计算机网络中的传播[6]。这些研究明确了综合化、系统化解决方案的重要性，但其或局限于理论框架，或未能与动态主动防御机制深度结合。为此，本文旨在弥补上述不足，提出利用算法同态性完成安全计算与防泄漏。

当前，安全计算技术主流防护方法包括差分隐私

（Differential Privacy，DP）、安全多方计算（Secure Multi-Party Computation，SMC）与同态加密（Homomorphic Encryption，HE）[7]。其中，HE 的计算复杂性可能导致处理时间和资源需求大幅增加，相比 DP 会引入噪声影响精度，而 SMC 通信开销较大。本研究不同于传统的加密方案，除了可以实现加密明文，还可以实现密文间的运算。本文方法利用同态性，在海量数据情况下可以提高数据安全性，有效降低系统开销。

1　系统架构

为避免传统明文训练和推理方式因中间环节隐私泄露的问题，本文设计了基于 HE 的参数安全计算框架，如图 1 所示。首先，由数据提供方对原始训练数据进行同态加密处理，提高数据在传输与存储中的不可篡改性[8]。接着，将密文数据传输至模型服务方。在不解密的情况下，模型服务方可直接在加密空间中更新模型参数，确保梯度和参数的安全性。最后，用户端通过私钥解密推理结果，得到有价值的输出结果。该架构在模型结构不变的前提下，即可实现与主流深度学习的高度适配。因密文特性，即使攻击者窃取了中间梯度或参数，但仍难以逆推出原始数据，能够有效降低参数泄露风险。

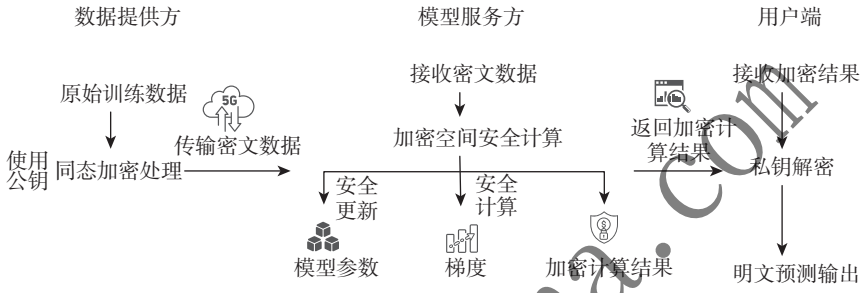


图 1　基于同态加密的 AI 参数安全计算架构

2　同态加密设计

2.1　同态加密原理

同态加密的核心在于密文运算等价于明文运算，即对明文进行的加法、乘法等代数操作可以通过对应的密文操作完成，并在解密后得到相同的结果。本文公式符号说明如表 1 所示。

表 1　公式符号说明

符号	含义与说明	符号	含义与说明	符号	含义与说明
α	加密放大参数	R，S	整数环	$\mathrm{Adv}_{\mathrm{HE}}(A)$	入侵者
p，q	素数	j	整数	β	解码开销
k	加密密钥	m_1，m_2	明文输入	r	加密过程中使用的随机数
m	明文消息	$\Pr[\cdot]$	概率	p_k	公钥
c	密文	$E(m_1 \times m_2)$	乘法	$E(m_1 + m_2)$	加法

设 λ 为安全参数，p、q 为素数，其中 p，$q \in [2^{\lambda-1}, 2^{\lambda})$。在加密阶段，对于给定的消息 $m \in (0, 1)$，加密密文 $c = pq + 2r + m$，解密阶段 $m = (c \bmod p) \bmod 2$。同态加密过程中，设有明文 m_1，m_2，则有：

$$\begin{cases} c_1 = pq + 2r_1 + m_1 \\ c_2 = pq + 2r_2 + m_2 \end{cases} \tag{1}$$

加法同态计算如下：

$$c_1 + c_2 = 2pq + 2(r_1 + r_2) + m_1 + m_2 \tag{2}$$

$$D(c_1 + c_2) = D(2pq + 2(r_1 + r_2) + m_1 + m_2) = m_1 + m_2 \tag{3}$$

综上，该加密方案具有加法同态性。

乘法同态计算如下：

$$c_1 \times c_2 = (pq)^2 + 2pq(r_1 + r_2) + pq(m_1 + m_2) + 2(r_2 m_1 + r_1 m_2) + m_1 \times m_2 \tag{4}$$

$$D(c_1 \times c_2) = m_1 \times m_2 \tag{5}$$

为适配 AI 训练，本研究在 HE 框架下引入向量化编码，使得模型参数可被批量加密，提高并行效率。

2.2　同态加密方案

同态加密计算满足 AI 模型两个变量加密后的加法或者乘法计算结果，与加密前的计算结果一致。为了满足 AI 模型参数安全计算与防泄漏需求，设 R，S 为两个整数环，将其定义为 E：$R \rightarrow S$，运算过程描述为以下三种：

（1）设明文 m_1，m_2，不计算 m_1，m_2，依据 $E(m_1)$ 和 $E(m_2)$ 进行加法计算，求得 $E(m_1 + m_2)$，这个过程为加法同态。

（2）设明文 m_1，m_2，不计算 m_1，m_2，依据 $E(m_1)$ 和 $E(m_2)$ 进行乘法计算，求得 $E(m_1 \times m_2)$，这个过程为乘法同态。

（3）设明文 m_1，m_2，不计算 m_1，m_2，依据 $E(m_1)$

和 $E(m_2)$ 进行一系列混合乘法计算，求得 $E(m_1 \times m_2)$，这个过程为混合乘法同态。

基于上述三种运算过程，可以看出，加法和乘法为主要操作运算。在 AI 模型参数计算与防泄漏过程中，仍需满足以下两种要求：

（1）明文与密文符合一对多条件，同一段明文加密两次得到的密文虽然不同，但两个不同的密文在解密后，得到的明文必须是一致的。换言之，对于明文 m，即使 $E_1(m) \neq E_2(m)$，但必须满足 $D(E_1(m)) = D(E_2(m))$。

（2）对于混合乘法同态，并不是所有 AI 模型参数都具有这种特征，仅仅有一小部分元素满足该计算特征。针对整数集而言，仅有整数 $m_1 = 1$ 才满足混合乘法同态。而本文针对 AI 模型参数的计算只需要满足加法和乘法同态即可。

在整数范围的基础上，加法与乘法运算使得算法同时满足以下运算：

$$E(a+b) = \mathrm{PLUS}(E(a), E(b)) \tag{6}$$

$$E(a \times b) = \mathrm{MULT}(E(a), E(b)) \tag{7}$$

构造算法满足式（6）和式（7）的加密算法即为同态加密算法，其基本原则为设 p，q 为两个大的素数，r 为随机数，m 表示明文，c 表示加密后的密文，则加密计算为：

$$\mathrm{Encrypt}(p_k, m): E(m) = c = (M + r * p) \bmod p * q \tag{8}$$

解密计算如下：

$$\mathrm{Decrypt}(s_k, c): D(c) = c \bmod p \tag{9}$$

式中，p_k 为公钥，由两个大素数 p 和 q 的乘积构成；s_k 表示私钥，即素数 p 和 q。

通过式（8）可得，存在整数 j，使：

$$c = (m + rq) \bmod pq = m + rq + j * pq = m + (r + jq)p \tag{10}$$

2.3 安全计算流程

基于同态加密的 AI 参数安全更新流程如图 2 所示，在密文域完成模型推理与参数更新，从而在保护用户数据和模型参数隐私的前提下，实现安全的模型训练与优化。数据提供方先加密输入数据，模型服务方在密文域进行推理，用户端解密并计算梯度，再将加密梯度传回服务端进行参数更新，形成闭环。

综上，同态加密方法除了可以实现明文加密，还可以实现密文间的运算。利用同态性，可以不必解密每一段密文，而是多个密文运算后再解密，在海量 AI 数据情况下可以提高数据安全性[9-10]。

2.4 性能指标

为了验证同态加密在 AI 模型中的可行性，本研究从安全性、计算复杂度、精度三个维度，提出性能指标。

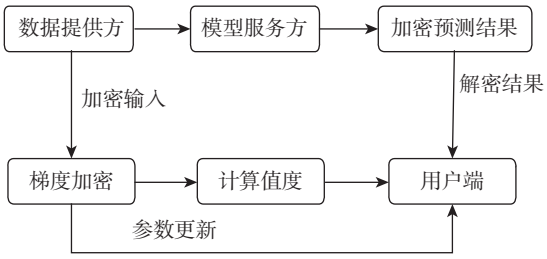


图 2　基于同态加密的 AI 参数安全更新流程

安全性依赖于密文不可区分性（IND-CPA）与密钥强度。设攻击者观察到的密文为 c_1，c_2，其成功区分概率为：

$$\mathrm{Adv}_{\mathrm{HE}}(A) = \left|\Pr[A(\mathrm{Enc}(m_1)) = 1]\right| - \left|\Pr[A(\mathrm{Enc}(m_2)) = 1]\right| \tag{11}$$

其中，$\mathrm{Adv}_{\mathrm{HE}}(A)$ 表示入侵者 A 对 HE 系统的优势概率；$\Pr[\cdot]$为概率；$\mathrm{Enc}(m_1)$，$\mathrm{Enc}(m_2)$ 为不同明文加密后的密文。若 $\mathrm{Adv}_{\mathrm{HE}}(A) \approx 0$，则说明攻击者无法区分不同明文的加密结果，系统安全性得到保证。

加密计算不可避免会带来性能开销。设明文计算复杂度为 $O(T)$，则加密计算复杂度约为：

$$O_{\mathrm{HE}}(T) = O(\alpha \cdot T + \beta) \tag{12}$$

其中，α 表示运算放大因子，β 为编码与解码开销，本研究通过优化批量加密与参数共享，以降低 α。

由于同态加密方案 CKKS 存在近似误差，本文定义精度保持度为：

$$\mathrm{APR} = \frac{\left|f(x) - \mathrm{Dec}\left|f(x)\right|\right|}{\left|f(x)\right|} \times 100\% \tag{13}$$

加密参数体积远大于明文，为衡量其可行性，定义加密开销比率：

$$\mathrm{Overhead} = \frac{\mathrm{Size}(\mathrm{Enc}(w))}{\mathrm{Size}(w)} \tag{14}$$

通过对比不同模型参数规模下的开销，可以评估方案的可部署性。本文提出的基于同态加密的 AI 模型参数安全计算方法，结合了 CKKS 方案的浮点支持特性，在密文空间实现了参数加密、推理与更新的闭环流程[11]。

3　实验设计与分析

3.1　实验环境

本研究在统一的实验环境下进行了多组对比实验，硬件与系统环境如表 2 所示，实验均在一台配置高性能 GPU 的服务器上进行。

实验验证了不同方法在不同复杂度任务上的性能对比，MNIST 为轻量级卷积神经网络，CIFAR-10 为中等规模卷积神经网络，ResNet-5 是一种深度卷积神经网络（CNN），BERT-base 是一种基于 Transformer 架构的预训练语言模型，包含大量数据信息。普通随机梯度下降

（Plain-SGD）通过分析梯度来重建训练样本，差分隐私随机梯度下降（DP-SGD）通过满足差分隐私来保护训练数据的隐私，HE-SGD 为本文设计的同态加密方案。

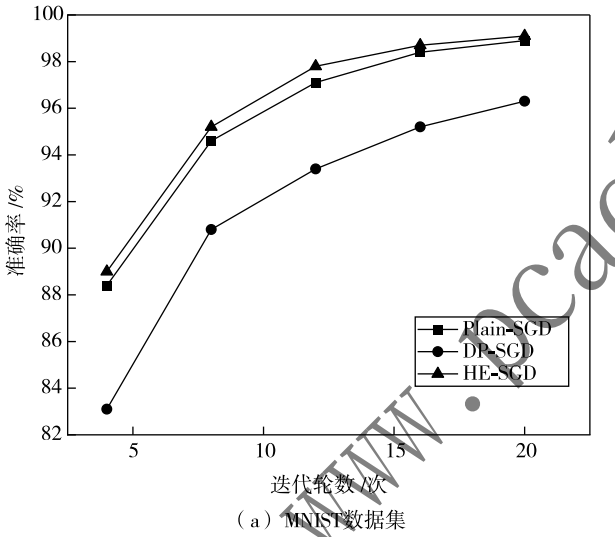
表 2　硬件与系统环境

硬件	配置
操作系统	Ubuntu 20.04 LTS
CPU	Intel（R）Xeon（R）Gold 6226R 2.90 GHz×32 核
GPU	NVIDIA Tesla V100 32 GB 显存
内存	256 GB DDR4
Python 版本	3.8
CUDA 版本	11.6

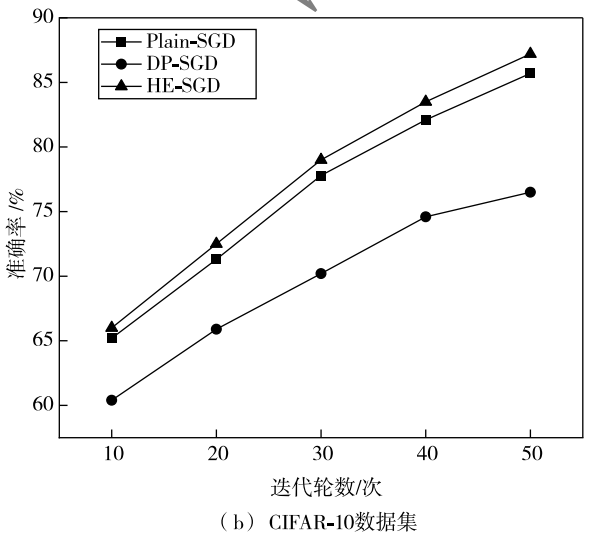
3.2　结果分析

3.2.1　精度对比

图 3 为分类准确率对比，其中图 3（a）为 MNIST 数据集对比结果，HE-SGD 准确率最高为 99.1%，优于 Plain-SGD 与 DP-SGD；图 3（b）为 CIFAR-10 数据集对比结果，随着训练轮数增加到 50 次，HE-SGD 始终保持最优，达到 87.2%。结果表明，HE-SGD 在保持数据与参数隐私的同时，能够达到甚至超越无保护训练的精度水平，验证了其在模型安全计算与防泄露场景中的有效性与实用性。



（a）MNIST数据集



（b）CIFAR-10数据集

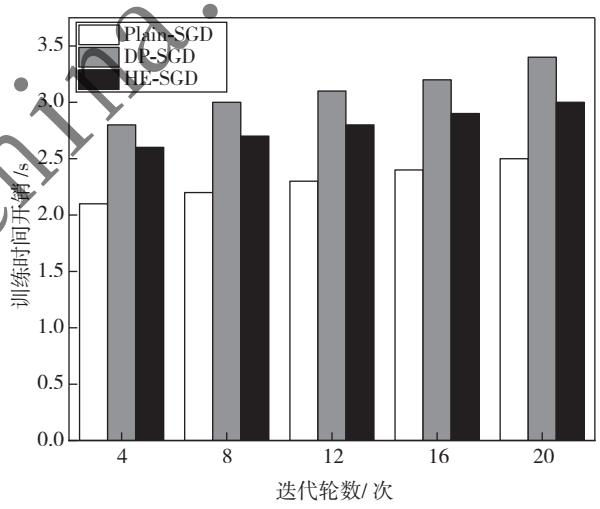
图 3　分类准确率对比

3.2.2　计算开销

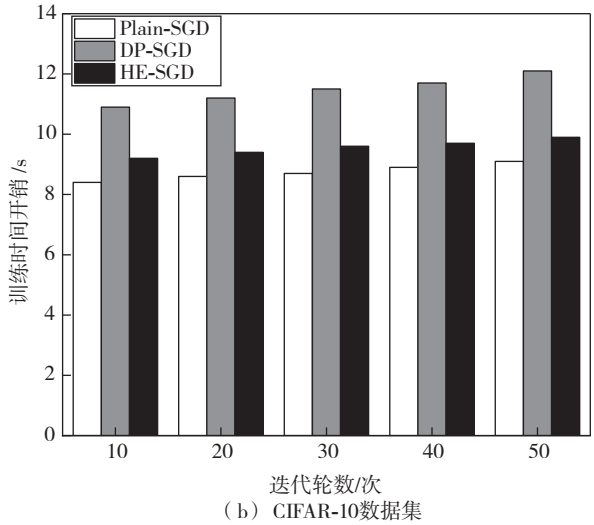
训练时间开销对比结果如图 4 所示，其中图 4（a）为 MNIST 数据集上的对比结果，HE-SGD 的每轮耗时仅比 Plain-SGD 增加 0.4～0.5 s；图 4（b）为 CIFAR-10 数据集对比结果，HE-SGD 的耗时比 Plain-SGD 增加 0.8～1.0 s，但仍明显低于 DP-SGD。同态加密计算的数值近似反而减少了部分计算冗余，相比 DP-SGD 的额外噪声计算更加高效。这表明，HE-SGD 在保证数据与参数全程加密的同时，训练时间开销仅略高于 Plain-SGD，远低于 DP-SGD，在效率与安全性上实现了平衡。

3.2.3　防泄漏效果

表 3 为防泄漏效果对比结果，Plain-SGD 信息泄露风



（a）MNIST数据集



（b）CIFAR-10数据集

图 4　训练时间开销对比结果

险指数为 0.72，说明泄露严重。DP-SGD 信息泄露风险指数为 0.31，说明防护效果明显，但仍有一定泄露概率，特别是 CIFAR-10 数据集中，攻击者仍可恢复部分图像特征。HE-SGD 在所有场景下，攻击成功率稳定在 2% ~ 5%，几乎等于随机失败，表明同态加密能彻底阻断直接信息泄露。信息泄露风险指数接近 0.0，即使攻击者掌握完整的训练协议，也无法从密文中恢复有效数据。这说明 HE-SGD 能在保证模型精度的同时，提供最强的隐私防护能力。结果说明，Plain-SGD 高性能但完全无保护，导致攻击成功率高，泄露严重。DP-SGD 虽然隐私保护较好，但仍存在攻击风险，处于中等保护，性能下降。HE-SGD 通过全程同态加密，在大规模数据集 ResNet-5 和 BERT-base 上，实现了 2.1% 和 1.9% 的泄露风险，是最佳的参数安全计算与防泄露方法。

表 3　防泄漏效果对比结果

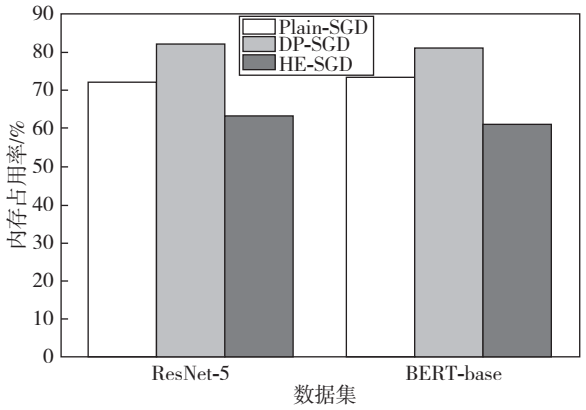
数据集	指标	Plain-SGD	DP-SGD	HE-SGD
MNIST	模型反演攻击成功率/%	72.4	35.7	3.2
MNIST	成员推理攻击成功率/%	68.1	28.6	2.9
MNIST	信息泄露风险指数(0~1)	0.72	0.31	0.03
MNIST（灰度重建）	模型反演攻击成功率/%	74	37.6	3
MNIST（1 000 样本）	成员推理攻击成功率/%	70.5	32.1	3.5
CIFAR-10	模型反演攻击成功率/%	65.3	33.4	4.1
CIFAR-10	成员推理攻击成功率/%	61.7	30.2	3.7
ResNet-5	模型窃取攻击成功率/%	60.4	45.4	2.1
BERT-base	属性推理攻击成功率/%	71.3	50.2	1.9

3.2.4　大规模数据集性能对比

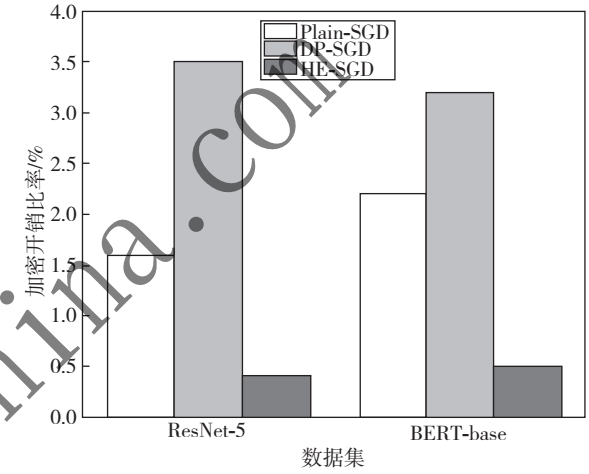
为了进一步评估同态加密算法在 AI 模型参数中的适用性，在大规模数据集 ResNet-5 和 BERT-base 上进一步评估算法内存占用率与加密开销比率。图 5（a）为内存占用率对比，可以看出 HE-SGD 的内存占用率最低，为 63.5% 和 61.1%。图 5（b）为加密开销比率对比，可以看出，HE-SGD 明显低于 Plain-SGD 和 DP-SGD。这表明 HE-SGD 在隐私保护方面具有显著优势，尤其是在对隐私要求较高的应用场景中。

4　结论

本文提出的基于同态加密的 AI 模型参数安全计算方法，能够在密文域完成训练与推理，有效防止数据与参数泄露。实验结果显示，在 MNIST 数据集上，HE-SGD 的分类准确率最高达 99.1%，较 Plain-SGD 和 DP-SGD 更优；在 CIFAR-10 数据集上，HE-SGD 的分类准确率达



（a）内存占用率



（b）加密开销比率

图 5　大规模数据集性能对比

87.2%，同样优于 Plain-SGD 和 DP-SGD。在计算开销方面，HE-SGD 仅比 Plain-SGD 增加约 0.83 s 额外开销。在防泄露效果上，Plain-SGD 的模型反演攻击成功率高达 72.4%，DP-SGD 降至 35.7%，而 HE-SGD 仅为 3.2%，信息泄露风险指数由 0.72 下降至 0.03。HE-SGD 在精度、效率和防泄露能力、加密开销比率上均表现最佳，为 AI 在高安全场景中的应用提供了可行方案。

参考文献

[1] 田雯，吴启迪，王祎玮，等. 大模型训练参数网络轨道优化技术分析 [J]. 电信工程技术与标准化，2024，37（S2）：18 - 22.

[2] 刘安聪，张磊，张源. 小程序生态下一种新型隐私泄露问题研究 [J]. 计算机应用与软件，2025，42（5）：317 - 324，340.

[3] 赵宁，戚清海. 分布式存储平台敏感数据泄露风险预测方法研究 [J]. 国外电子测量技术，2025，44（3）：177 - 183.

[4] 傅东波，张济鸿，谭龙广. 计算机网络信息安全中数据加密技术的应用探析 [J]. 信息与电脑，2025，37（17）：77 - 79.

（下转第 17 页）

arXiv preprint arXiv：2304.03277，2023.

[11] HONOVICH O，SCIALOM T，LEVY O，et al. Unnatural instructions：tuning language models with（almost）no human labor［C］//61 st Annual Meeting of the Association for Computational Linguistics（ACL 2023），2023：14409－14428.

[12] JI Y，DENG Y，GONG Y，et al. Exploring the impact of instruction data scaling on large language models：an empirical study on real-world use cases［J］. arXiv preprint arXiv：2303.14742，2023.

[13] WANG Y，KORDI Y，MISHRA S，et al. Self-instruct：aligning language models with self-generated instructions［C］//Annual Meeting of the Association for Computational Linguistics，2022.

[14] XU C，SUN Q，ZHENG K，et al. Wizardlm：empowering large language models to follow complex instructions［J］. arXiv preprint arXiv：2304.12244，2023.

[15] XU C，GUO D，DUAN N，et al. Baize：an open-source chat model with parameter-efficient tuning on self-chat data［J］. arXiv preprint arXiv：2304.01196，2023.

[16] 赵朝阳，朱贵波，王金桥. ChatGPT 给语言大模型带来的启示和多模态大模型新的发展思路［J］. 数据分析与知识发现，2023，7（3）：26－35.

[17] YANG D，YUAN R，FAN Y，et al. RefGPT：dialogue generation of GPT，by GPT，and for GPT［C］// Conference on Empirical Methods in Natural Language Processing，2023.

[18] SUN H，ZHANG Z，DENG J，et al. Safety assessment of Chinese large language models［J］. arXiv preprint arXiv：2304.10436，2023.

[19] CHENG Q，SUN T，ZHANG W，et al. Evaluating hallucinations in Chinese large language models［J］. arXiv preprint arXiv：2310.03368，2023.

[20] LIAN S，ZHAO K，LIU X，et al. What is the best model? application-driven evaluation for large language models［C］// Natural Language Processing and Chinese Computing，Part III，2025：67－79.

[21] ZHANG W，LEI X，LIU Z，et al. CHiSafetyBench：a Chinese hierarchical safety benchmark for large language models［J］. arXiv preprint arXiv：2406.10311，2024.

（收稿日期：2025－03－25）

作者简介：

张文静（1995－），女，硕士，工程师，主要研究方向：大模型安全、大模型检索增强。

肖思琪（1994－），女，硕士，工程师，主要研究方向：大模型微调、大模型量化。

廉士国（1979－），通信作者，男，博士，高级工程师，主要研究方向：人工智能、大模型。E-mail：liansg@chinaunicom.cn。

（上接第 11 页）

[5] 龙勇. 基于联邦学习模型对抗样本的鲁棒性增强技术［J］. 数字技术与应用，2025，43（4）：151－153.

[6] AHMAD S，AHMAD N，SHAH I. Stability and sensitivity analysis of cyberattack propagation models in computer networks［J］. European Journal of Pure and Applied Mathematics，2025，18（3）：6336－6336.

[7] ZHANG P，CHENG Q，ZHANG M，et al. A blockchain-based secure covert communication method via shamir threshold and STC mapping［J］. IEEE Transactions on Dependable and Secure Computing，2024，21（5）：4469－4480.

[8] 李冰，刘怀骏，张伟功. 魔方派：面向全同态加密的存算模运算加速器设计［J］. 电子与信息学报，2023，45（9）：3302－3310.

[9] 陆星缘，陈经纬，冯勇，等. 基于同态加密的隐私保护数据分类协议［J］. 计算机科学，2023，50（8）：321－332.

[10] 刘家森，王绪安，余丹，等. 基于同态加密和边缘计算的关键目标人脸识别方案［J］. 信息安全研究，2024，10（11）：1004－1011.

[11] 管桂林，支婷，陶政坪，等. 物联网中多密钥同态加密的联邦学习隐私保护方法［J］. 信息安全研究，2024，10（10）：958－966.

（收稿日期：2025－10－12）

作者简介：

张恒（1981－），男，本科，工程师，主要研究方向：信息系统安全管理。

廖尚斌（1986－），男，硕士，工程师，主要研究方向：网络与信息安全。

张陈颖（1980－），男，本科，高级工程师，主要研究方向：信息系统安全规划。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部