摘　要

中文医学文本生成任务在提供医疗服务和传播医疗知识中具有显著价值，但该任务由于其特殊性，相比普通的文本生成任务面临着更大的挑战。首先，医学文本数据的规模有限，且受到严格的法律法规限制，这使得训练出该领域高质量的语言模型面临着挑战。其次，在生成结果的质量上，由于需要保证结果的可解释性、准确性等特点，在保障隐私的前提下，不能过多牺牲模型的表达能力。最后，医学文本生成任务的隐私保护比普通文本生成任务（如对话、翻译等）的隐私保护更为关键，不仅因为其数据敏感性、法律和道德考量、信任问题等，还因为医学文本中的隐私信息更易被筛选，这使得在训练阶段与推断阶段需要考虑更严格的场景，隐私需求更大。

本研究针对以上问题提出了一种全新的解决方案。首先，本文利用在大规模的中文语料上训练的预训练模型作为基础，然后使用无隐私风险的医学领域专业知识语料进行微调，以增强模型在医学领域的表达能力。其次，本文通过多方安全计算的方式，使更多的参与方可以在保护隐私的前提下提供训练数据，从而进一步丰富医学文本生成任务的训练数据，有效缓解了医学数据稀缺性带来的问题。

在隐私保护方面，本研究详细分析了医学文本生成任务的隐私攻击模型，及其对隐私安全的威胁程度，并面向推断阶段的模型反演攻击提出了改进的攻击手段。在训练阶段，本文设计并实现了一种面向 Transformer 结构的，基于可信硬件 SGX 的多方安全计算协议，以确保训练过程的安全性；在推断阶段，本文面向 Transformer 结构的模型在训练阶段设计了选择差分隐私优化器，在推断阶段设计了选择差分隐私解码算法，有效防止了恶意攻击者恢复模型的训练隐私数据，同时保证了生成结果的准确性。由于医学文本中的隐私信息更易被筛选，因此选择差分隐私在此领域的应用更具优势。此外，本文还设计了一个用于评估生成的医学文本的科学性的指标，并通过实验证明了使用选择差分隐私的模型生成的结果的科学性。

本研究首次分析了中文医学文本生成任务的隐私问题，并提出了具有针对性的解决方案。这一成果不仅在理论上深化了对于医学文本生成任务隐私保护问题的理解，也在实践上为该领域的隐私保护提供了有效的策略和工具。

关键词：自然语言处理，医学文本生成，差分隐私，多方安全计算，深度学习隐私保护

ABSTRACT

The Chinese medical text generation task holds substantial importance in enhancing healthcare services and disseminating medical knowledge. However, it confronts distinct challenges compared with conventional text generation tasks. Firstly, the need for medical text data, compounded by stringent legal and regulatory constraints, creates a significant hurdle in training superior language models. Secondly, the generated output must ensure interpretability and accuracy while not compromising expressive capacity for privacy protection. Thirdly, privacy protection in medical text generation assumes higher criticality than routine tasks such as chat and translation. This is attributable to data sensitivity, legal and ethical considerations, and ease of filtering private information from medical text. This calls for a stricter approach during the training and inference stages to guard against potential malicious attackers.

To address these issues, we present a novel solution. We begin with a pre-trained model trained from extensive Chinese corpora, followed by fine-tuning step using medical knowledge corpora. This process intends to augment the model's expressive power in the medical domain. Moreover, our strategy utilizes multi-party secure computation to allow several participants to supply training data while preserving privacy. This significantly mitigates the obstacles posed by the scarcity of medical data.

We present a comprehensive analysis of the privacy attack model in medical text generation tasks to address privacy protection issues, demonstrating its threat to privacy and security. We propose an advanced attack method for the model inversion attacks during the inference stage. At the training stage, we construct and implement a multi-party secure computation protocol for the Transformer-based model to ensure training confidentiality. We deploy Intel SGX to guarantee the integrity of the training process. As for the inference stage, we address a selective differential privacy optimizer and a selective differential privacy decoding algorithm for the Transformer-based model. This deters malicious attackers from accessing or inferencing private training data, concurrently ensuring the accuracy and interpretability of the generated outcomes. Given the ease of filtering private information from medical text, deploying selective differential privacy yields considerable benefits. Furthermore, we introduce a new metric - the "medical text generation scientific index" to assess the scientific and the accuracy of the generated medical text. We validated this index through rigorous experimentation, which substantiates the scientific robustness of our model.

This thesis represents a comprehensive exploration of privacy issues on Chinese medical text generation tasks, providing novel solutions simultaneously. This accomplishment extends the theoretical comprehension of privacy protection issues in medical text generation tasks and provides practical, effective strategies and tools for privacy protection within this domain.

Key Words: Natural Language Processing, Medical Text Generation, Differential Privacy, Multi-Party Secure Computation, Privacy-preserving Deep Learning

目　　录

插 图 清 单

表 格 清 单

第 1 章　绪　　论

本章首先阐述了医学文本生成任务的研究背景和研究意义，并针对现有研究工作进行介绍，从而引出本文的研究内容与创新点，最后介绍本文的组织结构。

1.1　研究背景和意义

本节首先介绍医学文本生成任务的基本概念，并对其研究背景和研究意义进行详细说明。

1.1.1　研究背景

医学文本生成（Medical Text Generation）任务是一种针对电子病历（Electronic Health Record, EHR）和临床记录（Clinical Notes）等医学内容的自然语言处理（Natural Language Processing, NLP）技术。近年来，医学文本生成任务在电子病历生成[1]、临床记录生成[2]以及生成式摘要[3]等领域得到了广泛应用。医学文本生成任务旨在利用自然语言生成技术自动生成具有医学背景的自然语言文本，如病历报告、症状描述和医疗建议等。这些文本需要准确表达医学术语、疾病、症状、药物等复杂医学概念，同时具备一定的语法和逻辑性。医学文本生成任务可以辅助医学诊断和治疗，帮助医务人员生成病历和病情报告，提高工作效率，减轻工作负担，同时避免错误和疏漏。此外，医学文本生成还有助于患者理解医学术语和治疗方案，提高患者的医疗健康素养。

在面对医学文本生成任务时，需要从两个关键角度来看待：首先，医学文本在生成任务中的特性与其他类型的文本相比存在显著差异；其次，医学文本生成任务在隐私保护上也有其独特于其他文本生成任务的隐私保护之处。

医学文本在文本生成任务中的特性如下：

- 准确性：由于医学文本需要表达具有深度和复杂性的医学知识，因此生成的医学文本必须具有极高的准确性。任何微小的错误都可能导致严重的后果。此外，医学文本经常使用专业术语和行话，这是一个需要特别处理的问题，因为语言模型可能无法完全理解这些术语的含义。
- 可解释性：医学文本生成任务也需要生成的文本具有高度的可解释性。医生、患者和其他医疗保健工作者需要能够理解并解释生成的医学文本。
- 数据稀缺性：医学文本通常来自专业的医学记录，研究报告，或者其他医学文献，而这些数据源往往对模型训练提供有限的数据。因此，找到足够