用基于英特尔® SGX 的可信执行环境有效应对大语言模型隐私和安全挑战

作者：英特尔公司 俞巍、李志强、李青青、龚奇源

可信执行环境是什么？大语言模型为什么需要它？

OpenAI 的 GPT 系列大语言模型（Large Language Mode，以下缩写为 LLM）的兴起与应用，也带来了诸如数据泄露、数据滥用、模型被攻击和知识产权被窃取等一系列隐私和安全风险或挑战。

可信执行环境（Trusted Execution Environment，以下缩写为 TEE）是一项基于软硬件组合创建安全执行环境，能够更好地确保计算和数据处理机密性和完整性。其关键机制为：

- **安全隔离：**通过硬件加密和内存隔离等硬件隔离技术，将敏感数据和关键代码与其他应用及操作系统相隔离，从而确保它们即使在系统其他部分被攻击或受到恶意软件影响时也能够得到更好的保护。
- **安全验证：**在启动过程中进行身份验证和完整性检查，确保只有经过授权的代码和数据可以在其中运行，以此防止恶意软件或未经授权的访问。
- **安全执行环境：**提供包含加密算法、安全协议和密钥管理等防护功能的执行环境，用于处理敏感数据和执行关键算法，以增强数据在执行过程中的保密性和完整性。

TEE 与 LLM 可在多行业、多场景融合，TEE 可用于为 LLM 提供颇具商业落地价值的隐私和数据保护创新解决方案。

2. LLM 与 TEE 的融合需求

LLM在许多行业的不同场景都有着广泛应用，例如金融行业的风险评估和交易分析，医疗保健领域的医学图像识别、病历纪录和疾病预测，以及法律和合规行业的法律咨询、合同审查和文书处理等。这些行业或场景中涉及到的数据多为重要敏感的交易数据或个人数据，必须得到有效保护。

将 TEE 与 LLM 融合，有助于在这类场景中更好地保障数据在 LLM 模型训练和推理过程中的保密性。训练阶段，TEE 中的数据处理都处于加密状态；推理阶段，TEE 则可保护用户输入和模型结果的隐私。同时，其硬件隔离和安全验证机制可以更有效地防止未经授权的访问和攻击，增强模型运行时的安全性。

3. TEE 与 LLM 融合的挑战：资源和性能限制

- **资源限制：**TEE 的计算资源和存储空间通常都非常有限，LLM 庞大的模型参数和计算需求可能会超出一般 TEE 的能力范围。
- **性能下降：**I/O 数据的加密和安全计算操作会引入额外的计算开销，导致模型训练和推理性能有一定程度下降。基于算法的解决方案可减少模型规模和计算需求，以适应 TEE 的资源限制，但 CPU 仍会成为制约 LLM 训练的算力瓶颈。

4. 基于英特尔® 平台的解决方案：加速 TEE 与 LLM 融合应用

4.1 基于英特尔®SGX/TDX[1]的 TEE 解决方案

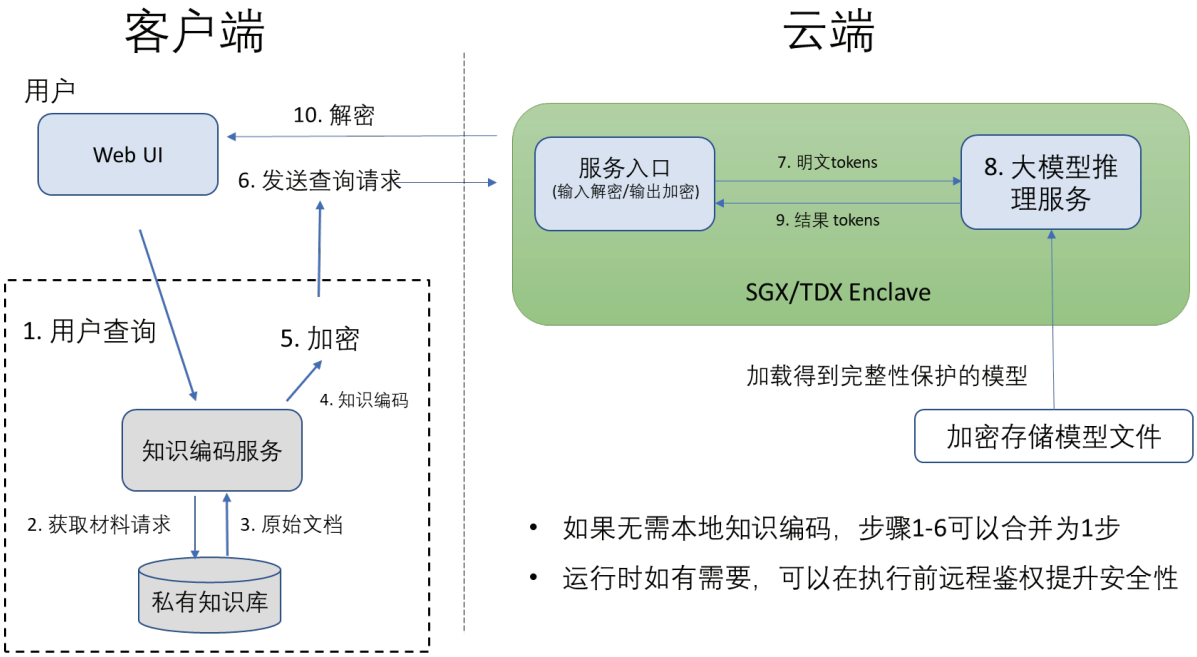
英特尔自第三代英特尔® 至强® 可扩展处理器开始内置英特尔® 软件防护扩展（英特尔® SGX）技术，其安全飞地的容量最多可达单颗 CPU 512GB，双路共计 1TB 容量，可满足目前千亿大模型的执行空间需求。此外，该技术提供支持的机密计算可实现应用层、虚拟机（VM）、容器和功能层的数据隔离。无论是在云端、边缘还是本地环境，都能确保计算与数据始终在私密性和安全性上获得更全面的保护，以免暴露给云服务提供商、未经授权的管理员和操作系统，甚至是特权应用。另一方面，英特尔® Trust Domain Extension（英特尔® TDX）可将客户机操作系统和虚拟机应用与云端主机、系统管理程序和平台的其他虚拟机隔离开来。它的信任边界较英特尔® SGX 应用层的隔离边界更大，使受其保护的机密虚拟机比基于英特尔® SGX 的安全飞地的应用更易于进行大规模部署和管理，在部署 LLM 这类复杂应用时，TDX 在易用性上更具优势。此外，今年推出的全新第四代英特尔® 至强® 可扩展处理器内置英特尔® AMX，可大幅提升矩阵运算性能，而英特尔® SGX/TDX 也可为英特尔® AMX、英特尔® DL Boost 等计算指令提供支持，进而为 TEE 中的大模型赋予快速落地+优化性能的双重优势。

图1. SGX/TDX 的可信边界

构建完善的 TEE 生态系统对推动 LLM 的应用和发展至关重要。开发者需要能够简化集成和使用过程的面向 TEE 的开发者工具和框架。为此，英特尔在 SDK 的基础上，推出了开源的 lib OS 项目 Gramine 来帮助开发者更好地使用基于英特尔® SGX 的 TEE，助推 LLM 与 TEE 的融合。

- **4.1.1 大语言模型推理**

使用私有数据进行个性化训练的大模型不仅包含私有数据信息，其查询本身也具有隐私性，尤其是基于边端的非安全环境部署。基于英特尔® SGX/TDX 的 TEE 可为大模型提供更安全的运行环境，在数据上传云端前，查询可先通过客户端对传输内容加密，云端只需在英特尔® SGX/TDX 中解密查询问题，然后输入大模型的推理服务中，并将所得结果在云端的 TEE 中加密后传输回本地客户端。在整个工作流程中，客户端以外的数据和运行态程序均处于密态环境当中，效率远远高于其他基于纯密码学的解决方案。目前像 LLAMA 7B、ChatGLM 6B 等模型都可以在该 TEE 方案上满足实时可交互性能的运行。图 2 展示了使用 LLM 部署知识问答的参考设计。基于英特尔® SGX/TDX 的 TEE 为实际部署 LLM 中的自有知识产权保护提供了一套完整的方案，优化整个模型在查询、传输和推理过程中的安全保护。



- 如果无需本地知识编码，步骤1-6可以合并为1步
- 运行时如有需要，可以在执行前远程鉴权提升安全性

图 2. 基于 TEE 的大语言模型私密问答

- **4.1.2 联邦学习**

借助基于 TEE 的联邦学习解决方案[2]（见图 3），就可在多机构之间实现基于 NLP 的深度学习，例如使用 BERT 的命名体识别。在金融和医疗等行业提升准确性，实现多机构数据互通，同时更好避免数据泄露。

此方案中每个参与方包含一个 Avalon[3] 管理模块和 Gramine 工作负载，均运行在英特尔® SGX 的安全飞地中，在管理模块彼此间的远程认证完成执行后，即可启动联邦学习过程，参与方在本地使用各自的数据集进行训练，然后将梯度上传至聚合方，聚合方进行聚合后将平均梯度下发至各参与方，以继续进行下一轮训练。对比图 4 所示的 BERT + CRF 模型[4]，此方案可以在强化隐私保护的同时，让性能损失维持在 50% 以下[2]。

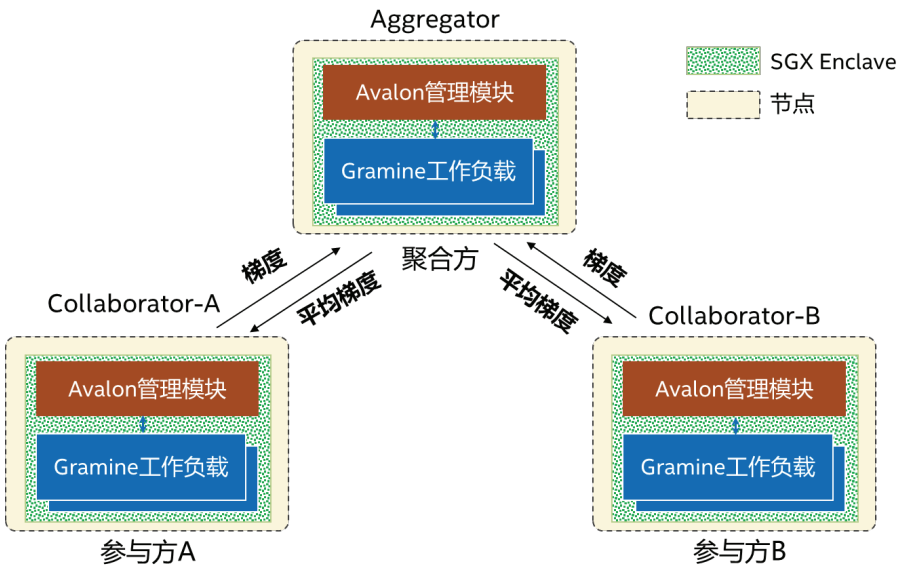


图 3. 基于 TEE 的联邦学习

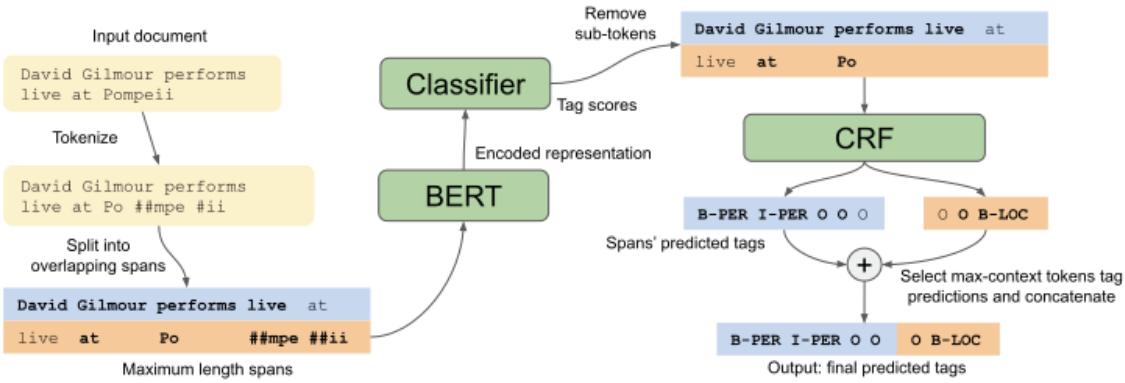


图 4. BERT + CRF 模型[4]

4.2 BigDL：端到端大模型和 TEE 融合的方案

据行业用户反馈，LLM 在端到端应用中的痛点包括：

- **软件栈复杂，难以确保端到端应用的安全**。LLM 的训练和推理常依赖较多的软件栈、服务和硬件。为保护用户数据和模型产权，需确保每个环节的安全性（不同硬件、运行环境、网络和存储等）。
- **计算量大，且对性能敏感**。LLM 的计算量非常大，需引入足够多的性能优化。但是，不同模型、平台和软件栈需要使用不同的优化方案，要在特定平台上实现更理想的性能，需要长时间的性能调优。

为解决这些痛点，由英特尔主导的开源项目 BigDL，近期就推出了针对 LLM 的隐私保护方案，其两大主要功能为：

- **提供端到端的安全保护：**在不修改代码的情况下，为单机和分布式的 LLM 应用提供端到端的安全保护功能。具体包括，基于英特尔® SGX/TDX 的 TEE、远程证明、统一的密钥管理接口和透明的加解密 API 等。
- **实现一站式性能优化：**BigDL Nano 提供的针对 LLM 的一站式性能优化方案，可让现有 LLM 应用在几乎不用修改代码的情况下受益于英特尔® AMX、英特尔® AVX-512 和英特尔® Extension for PyTorch。同时，用户还可利用 BigDL Nano 提供的 LLM API，快速构建应用。
-

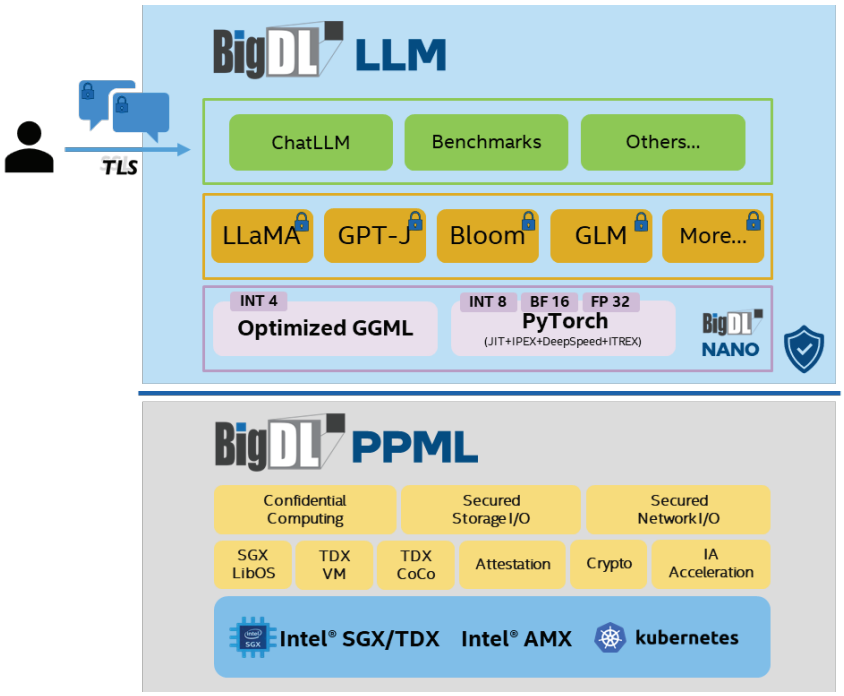


图 5. BigDL 端到端安全的大模型方案

如图 6 所示，在应用了 PPML（Privacy Preserving Machine Learning，隐私保护的机器学习）提供的安全技术后，由于更强的安全和隐私保护会带来额外开销，因此端到端应用性能会略有下降；但应用了 BigDL Nano 提供的优化功能后，端到端的性能得到了显著改善*，总体性能甚至高于没有任何保护的明文性能。

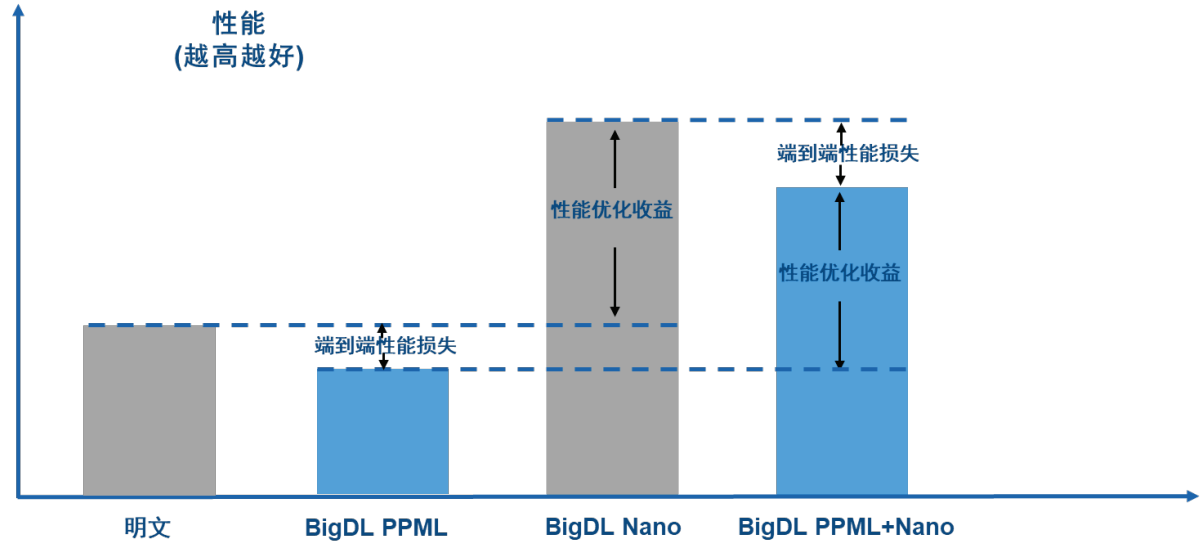


图 6. BigDL PPML + Nano 端到端性能损失情况

目前，该方案已经开源，并开始陆续交付给行业客户进行测试和集成[5]。

5. 未来趋势

TEE 提供了隐私保护和数据安全防护功能的创新解决方案，将在 LLM 实际落地过程中扮演重要角色。通过将二者融合，可更好地保障数据在训练和推理过程中的保密性，增强对未经授权访问和模型结果篡改的防御。然而，在 TEE 中保护用户隐私的同时，需要平衡性能需求，随着大模型对于计算需求的大幅提升，算力可能会执行在异构硬件上，TEE 和异构硬件的结合将成为未来发展趋势。随着 CPU 性能的提升以及内置 AI 加速技术的升级和更新，在方便部署的场景下，CPU 会是大模型推理和 TEE 结合的首选，在训练的场景下，基于 CPU 的 TEE 结合异构硬件的加密支持，则会是大模型训练甚至大模型联邦训练的技术方向。英特尔将一如既往地以软硬结合的产品技术组合促进开发者参与，推动 LLM 与 TEE 的融合。

作者简介：

英特尔公司 AI 架构师俞巍，英特尔公司平台安全资深架构师李志强，英特尔公司安全软件研发工程师李青青，英特尔公司软件架构师龚奇源，都在从事 AI 和 SGX/TDX 相关工作。

[1] SGX/TDX: https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/innovation-data-protection-with-security-engines.html

[2] Wei Yu. et al. 2022, TEE based Cross-silo Trustworthy Federated Learning Infrastructure, FL-IJCAI'22

[3] https://github.com/hyperledger-archives/avalon

[4] Souza, F., Nogueira, R. and Lotufo, R. 2020, Portuguese Named Entity Recognition using Bert-CRF, arXiv.org（请参见 https://arxiv.org/pdf/1909.10649.pdf）

[5] https://github.com/intel-analytics/BigDL/tree/main/python/llm

*性能优化效果与具体平台、模型和环境有关