基于大语言模型的 SQL 注入攻击检测方法研究

李晓东，宋宜昌，江政旦，张莹莹
（国家计算机网络应急技术处理协调中心广东分中心，广东 广州 510665）

【摘　要】随着结构化查询语言（structured query language, SQL）注入攻击日益成为网络安全领域的重大威胁，传统检测技术存在误报率高、适应性差的问题。针对这些问题，提出了一种利用高级大语言模型的检测方法，通过融合提示词工程与精准指令调整技术，开发出专门应对 SQL 注入攻击的大语言模型，并进行实验与分析。实验数据显示，基于大语言模型的 SQL 注入攻击检测方法在标准数据集上的表现优于传统模型，准确率超过 90.63%，误报率低于 0.95%，在实际应用中具有巨大潜力。

【关键词】SQL 注入；漏洞检测；大语言模型；提示词工程；指令微调
【中图分类号】TP181　　【文献标识码】A　　【文章编号】1006-4222（2024）06-0061-03

0 引言

目前结构化查询语言（structured query language, SQL）注入攻击的防御已成为全球范围内信息安全领域研究热点。目前，SQL 注入攻击检测主要有两种方法[1]：①基于规则的方法[2]，通过分析已知攻击案例来制定过滤规则。②基于机器学习的方法，通过分析大量数据样本来预测潜在的攻击行为[3-4]。然而，这些传统方法通常因为误报率高和适应性不足而难以满足当前网络环境的需求[5-6]。

近年来，大语言模型因其在处理自然语言和结构化数据方面展现出的强大能力，已被广泛应用于各种自然语言处理任务中[7]，如机器翻译和文本生成。这类模型基于庞大的语料库训练而成，具备较强的语义理解能力。然而，在网络安全特别是 SQL 注入攻击检测领域[8-9]，大语言模型是否能维持其良好表现仍是一个值得探索的问题。

为了在网络安全领域推广大语言模型，本文通过集成提示词工程和指令微调技术[10]，构建了一个专门针对 SQL 注入攻击检测的大语言模型，并通过实验与现有技术进行比较，证明其有效性。

1 方法设计

本文提出了一种基于高级大语言模型的 SQL 注入攻击检测方法，整体框架如图 1 所示，通过集成创新的提示词工程和精确的指令微调技术，定制了专用于检测 SQL 注入攻击的大语言模型。整个检测框架包括数据预处理、提示词设计与应用、模型的精确微调、检测结果映射以及最终的性能评估 5 个主要阶段。

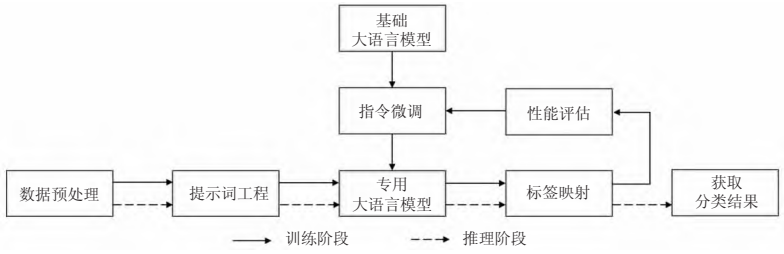

图 1 基于大语言模型的 SQL 注入攻击检测框架

1.1 数据预处理

对从 Kaggle 平台获取的开放数据集进行预处理，包括清洗和分类，确保用于训练、验证和测试的数据集具有较高的质量。

1.2 提示词工程

提示词是指大语言模型的输入文本，通过提供相应的上下文和任务要求，帮助模型理解用户意图并正确响应。合理设计提示词可以显著提升模型的应答质量和准确性，进而驱动模型更有效地执行 SQL 注入攻击检测任务。

本文基于 OpenAI 公开的提示词设计文档，为 SQL 注入攻击检测任务设计提示词。提示词由指令、输入和输出 3 部分组成，各部分以“###”符号分隔，以便大语言模型识别不同的文本区域[11]。提示词有助于引导模型准确理解任务内容，并依据指令生成相应的输出，即选择“yes”或“no”作为检测结果。

需要说明的是，大语言模型本质上是基于概率进行文本生成预测，因此模型可能会生成非预期响应，即与指令要求不符合的回答[11]。

1.3 指令微调

指令微调技术可以使大语言模型更好地适应特定任务。目前，主流的指令微调技术包括适配器技术[11]、前缀微调技术[12]和低秩自适应（low-rank adaptation, LoRA）技术[13]等。鉴于 LoRA 技术易于训练以及资源消耗低等优点，本文采用 LoRA 技术对模型进行微调。

1.4 标签映射

大语言模型经过推理后生成自然语言形式的响应，需要将其映射至相应的标签。由于模型输出具有一定随机性，即使在提示词中规定输出“yes”或“no”，并且微调时已使用预期输出“yes”或“no”训练模型，但仍有可能产生非预期响应。

为此，本文构建了一套标签映射规则，将大语言模型生成的文本按其内容映射成特定标签，主要有以下 3 种情况：①包含“yes”“true”“1”等关键字的响应，将归为 SQL 注入攻击类。②包含“no”“false”“0”等关键字的响应，将归为非 SQL 注入攻击类。③对于其他响应，视为未能理解指令语义，归为未知类。

1.5 性能评估

为准确评估大语言模型的 SQL 注入攻击检测能力，本文采用了分类任务中常用的评估指标，即准确率和误报率。结合混淆矩阵进行介绍，其中，真正例（true positive, TP）表示真实结果为正例，预测结果是正例；假正例（false positive, FP）表示真实结果为负例，预测结果是正例；真负例（true negative, TN）表示真实结果为负例，预测结果是负例；假负例（false negative, FN）表示真实结果为正例，预测结果是负例。

假设样本总数为 N，标签映射后未知类的数量为 U，则可得到如下公式：

$$N=TP+FP+FN+TN+U。\tag{1}$$

模型准确率计算公式如下：

$$Acc=\frac{TP+TN}{N}。\tag{2}$$

模型误报率计算公式如下：

$$FPR=\frac{FP}{FP+TN}。\tag{3}$$

2 实验与分析

本文在一台配备 Intel Xeon Gold 6248R 3.00 GHz 处理器、200 GB M2 RAM 以及 NVIDIA RTX3090 24 GB 显卡的服务器上进行实验，操作系统为 Ubuntu 20.04，开发环境为 Python 3.10.11。鉴于硬件条件限制，本文采用具有 62 亿个参数量的 ChatGLM3-6B 进行实验，该模型可在单显卡本地服务器上部署运行。本方法适用于任意的大语言模型，未来可根据需要替换为性能更优的模型。

在模型微调过程中，学习率（*learning_rate*）设定为 5×10^{-5}，批量大小（*batch_size*）为 16 个，最大输入令牌数量（*max_input_length*）为 512 个，最大输出令牌数（*max_output_length*）为 64 个。同时，将 Kaggle 公开数据集划分为训练集、验证集和测试集，数据集划分如表 1 所示。

表 1 数据集划分

数据类别	训练集/条	验证集/条	测试集/条
SQL 注入攻击	9 000	1 000	1 000
非 SQL 注入攻击	9 000	1 000	1 000

2.1 迭代次数扫描

研究微调的迭代次数对大语言模型性能的影响。实验设定最大迭代次数（*max_steps*）为 30 000 步，根据式（4）可计算出最大迭代轮数（*max_epochs*）为 27 轮，即模型在完整的微调训练集之上总共进行 27 次遍历。

$$max\_epochs=\frac{max\_steps\times batch\_size}{train\_set\_size}。\tag{4}$$

在上文所述的参数配置下，大语言模型训练的整体耗时为 8.6 h。随着迭代次数增加，损失函数值迅速下降，并在迭代 8 000 步之后趋于稳定，这表明模型收敛性良好。为探索迭代次数对模型检测能力的影响，微调时每隔 2 000 步对模型进行一次评估。

2.2 样本数扫描

分析微调样本数量对大语言模型性能的影响。首先，从微调训练集中随机抽取不同数量的样本，并构建出规模分别为{1 024，512，256，128，64，32}的子训练集。为保证各子训练集样本质量的一致性，先构建出一个包含最多样本的子训练集，再从中随机抽取出样本以构建出次大的子训练集，以此类推。

2.3 推理参数扫描

探讨推理参数对大语言模型性能的影响。通过选取不同的 *Temperature* 和 *Top_p* 参数值进行测试，其中 *Temperature* 参数值设定为{0，0.2，0.4，0.6，0.8，1.0}，*Top_p* 参数值设定为{0.7，0.75，0.8，0.85，0.9，0.95，1.0}，共有 42 种组合参数。基于这些不同的推理参数，使用微调后的大语言模型在验证集上进行推理，并分别评估模型性能。实验结果显示，当 *Temperature*=0.1 且 *Top_p*=0.8 时，模型达到了最高的检测准确率。在模型推理阶段，可根据具体需求选

择合适的 *Temperature* 和 *Top_p* 值，以引入必要的随机性，有助于提升模型性能。

2.4 模型检测性能测试

在完成上文设定的 3 项扫描实验后，本文成功构建出一个专用于 SQL 注入攻击检测的大语言模型。为检验模型在实际应用中的泛化能力，在测试集上对该专用模型进行全面评估。大语言模型 SQL 注入攻击检测性能如表 2 所示，迭代次数、微调样本数、推理参数为微调阶段和推理阶段的关键参数配置，准确率和误报率为该专用模型的两项检测性能指标。评估结果表明，该专用模型在 SQL 注入攻击检测方面表现良好，准确率为 90.63%，在存在 SQL 注入攻击的异常场景中，模型实现精准检测的概率较高；误报率为 0.95%，在无 SQL 注入攻击的正常场景中，模型触发错误警报的概率较低。

表 2 大语言模型 SQL 注入攻击检测性能

模型	迭代次数/步	微调样本数/个	推理参数	准确率/%	误报率/%
ChatGLM-6B	30 000	18 000	*Temperature*=0.1 *Top_p*=0.8	90.63	0.95

3 结语

本文提出一种基于大型语言模型的 SQL 注入攻击检测方法。通过数据预处理、提示词工程、指令微调、标签映射和性能评估等操作，成功构建出专用的检测模型。为进一步优化模型并提升其检测能力，设计实验深入探索影响模型性能的关键因素。实验结果表明，经过微调的大型语言模型在 SQL 注入攻击检测任务中表现良好，具有较高的准确率与较低的误报率。这充分证明了大语言模型在处理传统网络安全问题上的有效性，展现出其在网络安全领域的广阔应用前景。

参考文献

[1] 黄恺杰，王剑，陈炯峄. 一种基于大语言模型的 SQL 注入攻击检测方法[J]. 信息网络安全，2023，23(11)：84-93.

[2] UWAGBOLE S O, BUCHANAN W J, LU F. Applied machine learning predictive analytics to SOL injection attack detection and prevention[C]//2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Lisbon, Portugal. New York: IEEE, 2017: 1087-1090.

[3] GU H F, ZHANG J N, LIU T, et al. DIAVA: a traffic-based framework for detection of SQL injection attacks and vul nerability analysis of leaked data [J]. IEEE transactions on reliability, 2020, 69(1): 188-202.

[4] PROKHORENKO V, CHOO K K R, A SHMAN H. Web application protection techniques: a taxonomy [J]. Journal of network and computer applications, 2016(60): 95-112.

[5] HASAN M, BALBAHAITH Z, TARI QUE M. Detection of SQL injection attacks: a machine lea rning approach [C]// 2019 International Conference on Electrical and Computing Tech nologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates. New York: IEEE, 2019: 1-6.

[6] UMAR F. Ensemble machine learning approaches for detection of SQL injection attack [J]. Technical journal, 2021, 15(1): 112-120.

[7] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot leamers [C]//NIPS '20: Proceedings of the 34th International Conference on N eural Information Processing Systems, Vancouver, Canada. New York: Curran Associates Inc., 2020: 1877-1901.

[8] LAMB A. A brief introduction to generative models[EB/OL]. (2021-02-27)[2023-08-12].https://arxiv.org/abs/2103.00265.

[9] KEARY T. 12 practical large language model (LLM) applications [EB/OL]. (2023 -07 -14) [2023 -08 -12]. https://www.te c ho pedia.com/12-practical-large-language-mod el-llm-applications.

[10] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot l earners [C]. The Tenth International Conference on Learning Representations, 2022.

[11] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//36th International Conference on Machine Learning (ICML 2019), California, USA. New York: Curran Associates Inc., 2019: 4944-4956.

[12] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation [C]. The Joint Conference of the 59th Annual Meeting of the Association for Computational Lin guistics and the 11th International Joint Conference on Natural Language Processing(ACL-IJCNLP 2021), 2021.

[13] HU E J, SHEN Y L, WALLIS P, et al. LoRA: Low-rank adaptation of large language models [EB/OL].(2021-06-17) [2023-08-12].https://arxiv.org/abs/2106.09685.

作者简介：李晓东（1985—），男，汉族，江西赣州人，硕士研究生，高级工程师，研究方向为网络安全攻防、移动安全等。

宋宜昌（1975—），男，汉族，山东日照人，硕士研究生，高级工程师，研究方向为通信工程、网络与数据安全等。

江政旦（1996—），男，汉族，江西九江人，硕士研究生，助理工程师，研究方向为认知无线通信、智能无线网络等。

张莹莹（1990—），女，汉族，河南许昌人，硕士研究生，工程师，研究方向为网络安全攻防等。