Measuring and Augmenting Large Language Models for Solving Capture-the-Flag Challenges

李智浩

目录

CONTENTS

1

背景与意义

CTF竞赛的重要性

01

提升技能水平

通过模拟真实场景，提高参与者识别和利用漏洞的能力。

02

选拔顶尖人才

成为全球顶级赛事中的重要环节，用于选拔顶尖网络安全人才。

03

推动技术发展

激发对自动化攻防策略的研究，推动AI在网络安全的应用。

04

促进教育普及

作为教育工具，加速学习者理解复杂安全概念，促进网络安全教育的普及。

大型语言模型在网络安全领域的潜力

2

相关工作

网络安全评估基准

CyberSecEval 2

评估大语言模型的全面网络安全知识，涵盖广泛领域，如漏洞检测、防御策略等。

SecBench

提供多维度的网络安全评估数据集，用于测试LLMs在不同安全任务上的表现。

CyberMetric

专注于衡量LLMs在网络安全领域的知识深度，通过精心设计的数据集进行评估。

CTF基准

Intercode-CTF与NYU CTF Dataset等，侧重于评估LLMs解决CTF挑战的能力，强调实战技能。

自动化CTF解决方法

Intercode-CTF

评估LLM端到端CTF问题解决能力，提供静态命令行环境和基础工具，但缺乏互动性和专业工具支持。

NYU CTF Bench

引入更复杂的环境和工具，但仍受限于非交互式操作，未能充分模拟真实CTF场景。

AutoPwn

专注于CTF破解进展，利用神经网络系统自动识别漏洞，但未涵盖CTF全部环节。

3

研究内容与方法

CTFKNOW：技术知识基准设计

问题生成

基于提取的知识点，设计涵盖单选与开放题型的问题，精确衡量LLM的能力。

过滤幻觉

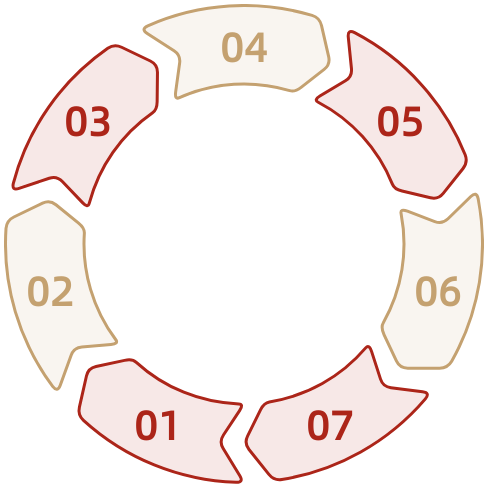
去除提取过程中产生的错误或不准确的信息，提高知识点的可靠性。

知识点提取

利用大语言模型自动抽取核心知识，形成初步的知识点集合。

收集解题报告

收集3000余篇高质量CTF解题报告，作为构建问题库（知识库）的基础资料。



问题筛选

对生成的问题进行筛选，确保其质量与准确性，符合评估标准。

专家人工验证

邀请领域专家对问题进行人工验证，确保问题的准确性和有效性。

建立基准

构建高质量的基准，用于后续的LLM性能评估和比较，促进技术发展。

CTFAGENT：自动化CTF解题框架概述

CTFAGENT旨在通过两阶段RAG和环境增强，弥补LLM在CTF场景下的知识应用缺口，提升**问题理解**与**解决能力**。

CTFAGENT核心理念

结合理解与利用阶段，精准匹配技术知识，辅助LLM识别并利用漏洞，实现高效问题解决。

两阶段RAG机制

提供动态命令行反馈与高级CTF工具，优化LLM与环境互动，增强复杂任务处理能力。

环境增强模块功能

技术细节

4

两阶段RAG系统

RAG-Understanding

在题目中，CTFAgent首先通过EA模块获取附件代码（如通过 cat 命令或反编译工具）。此代码被用于向量化，并作为检索key由RAG模块中的DB-Understanding匹配对应漏洞知识。

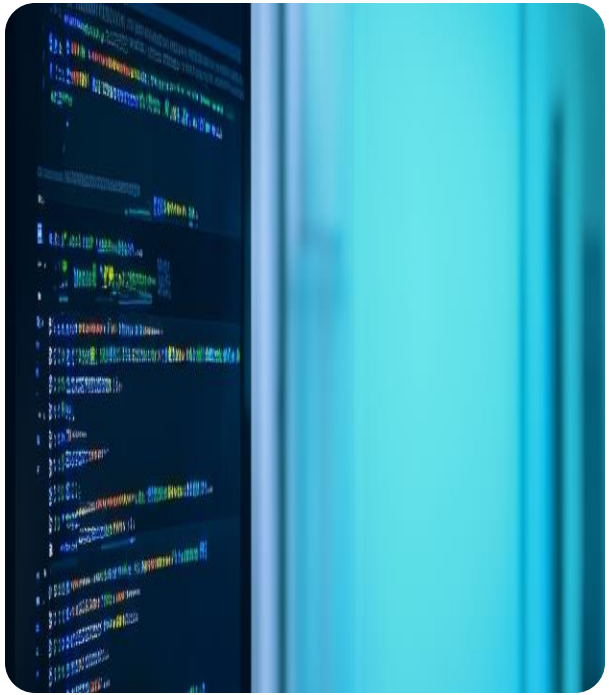
RAG-Exploiting

一旦LLM识别出潜在漏洞并开始生成利用方案，CTFAgent会自动将这些“利用思路”（Exploit Ideas）作为查询输入，向数据库中匹配更为详细且可操作的利用策略，例如：示例exp脚本、工具使用方法等

动态反馈机制

结合LLM的输出动态调整检索策略，确保提供的知识与当前问题场景高度匹配，促进精准解决问题。

交互式环境增强模块



提升命令行互动性

通过提供针对性提示和动态命令行反馈，增强LLM在处理复杂任务时的实时互动能力，有效解决权限读取和远程服务器交互难题。

高级CTF工具集成

引入IDA Pro等专业级工具，显著改善代码审计效率，优化逆向工程和漏洞利用过程，助力LLM精准识别并利用代码漏洞。

实时反馈机制

实现每轮输出即时反馈，确保LLM能够迅速调整策略，提高在多轮交互中解决问题的效率和准确性。

环境适应性增强

通过模拟真实CTF环境，使LLM能够在接近实战的条件下操作，提升其应对各种挑战的能力，加速学习和适应过程。

交互式环境增强模块案例

CTFAgent 使用 decompile 获取反编译代码 → 识别 buffer overflow。

调用 start_nc_session 连接远程服务器。

利用 payload（由 RAG引导 提供）构造攻击并通过 nc_send_line 发送。

收到包含 flag 的响应。

5

实验结果与分析

性能评估

CTFAGENT在Intercode-CTF数据集上显著提升了问题解决能力，从基线的39%提升至73%，增幅达85%。

CTFAGENT表现

采用更先进的o1模型后，CTFAGENT额外解决了11个挑战，总解决率提升至84%。

高级模型效果

在更具挑战性的NYU CTF数据集上，CTFAGENT解决了18个挑战，比基线提高了120%。

NYU CTF数据集

案例研究与失败原因分析

案例解析

详细分析了CTFAgent在特定CTF挑战中的工作流程，展示了如何通过两阶段RAG系统和环境增强模块协同工作来解决问题。

失败模式识别

归纳了CTFAgent在尝试解决CTF挑战时遇到的主要失败类型，包括超过最大尝试次数、上下文长度超出限制等。

RAG误导问题

探讨了RAG系统因检索不准确而导致CTFAgent接收错误的技术知识，从而影响了解决方案的效率和准确性。

多模态能力缺失

指出了CTFAgent在处理涉及图像和多模态工具的CTF挑战时的局限性，强调了未来研究的方向。

讨论与未来方向

6

教育与研究影响

智能辅助学习

LLM作为智能助手，加速安全知识获取，提升CTF教育效率与体验。

前沿研究推动

促进AI驱动的进攻性安全能力发展，应对软件漏洞自动化发现与修复挑战。

实战技能提升

通过模拟真实攻击场景，增强参与者实战经验和问题解决能力。

社区生态建设

激发更多高质量CTF赛事，促进网络安全人才发掘与培养。

伦理考量与风险缓解

遵守伦理标准

CTFAGENT严格遵循伦理规范，确保数据来源合法。

数据源公开透明

所有数据均来自公开的漏洞报告，保障无隐私泄露风险。

社区协作重要

CTFAGENT与CTFKNOW共同推动，强调社区合作的重要性。

倡导安全使用

鼓励将LLM应用于防御安全，并呼吁学术界研究防护机制。

THANKS