TVP Tencent Cloud Valuable Professional

TVP「AI与安全」高峰论坛

大模型时代 安全如何洗牌

Tencent Cloud Valuable Professional

Tencent Cloud Valuable Professional

安全领域大模型构建范式与实践

吴石

腾讯安全科恩实验室

CONTENTS / 目录

Tencent Cloud Valuable Professional

Tencent Cloud
Valuable Professional

01 概述

18年起探索“AI+安全”，目前重点关注“AI赋能安全”方向

Tencent Cloud Valuable Professional

科恩愿景：用 AI 赋能安全产品和业务，打造覆盖威胁感知、研判、防御的智能安全闭环

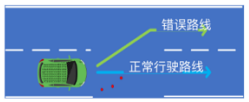
	Security of AI （AI自身安全）	AI for Security (AI 赋能安全)
核心目标	确保 AI 系统自身安全性	提升安全产品/业务智能化水平
技术路径	对抗样本防御 模型鲁棒性增强	AI 能力与安全知识、数据融合
业务价值	防御 AI 系统被恶意利用	威胁检测准确率提升 效率提升

特斯拉 Autopilot 安全研究

2019 对抗样本生成算法，精准误导图像识别，首个对抗商用自动驾驶系统图像识别功能的研究案例

2020 对车道级系统攻击的完善研究发布于 USENIX Security

干扰特斯拉自动雨刷模型实验



在路面部署干扰信息后，可导致车辆经过时对车道线做出错误识别，致使车辆驶入反向车道。

产学研合作

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

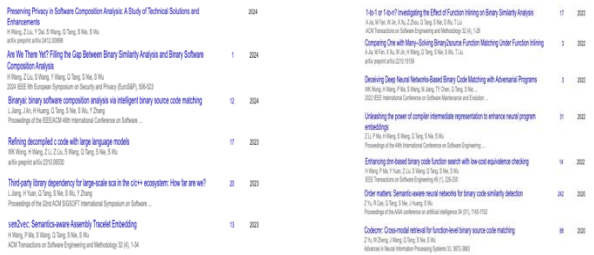
中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

论文成果

10余篇 CCF-A顶级会议 论文发布



产品落地

BinaryAI

腾讯安全 | 威胁情报中心

腾讯云

Tencent iOA

腾讯电脑管家 GUANJIA.QQ.COM

AI 大模型赋能网络安全新范式

Tencent Cloud Valuable Professional

AI大模型取得突破性进展，带来发展机遇

腾讯元宝　deepseek

通用大模型变革各行业

- AI大模型通过大算力、大规模训练数据突破自然语言处理的瓶颈
- 摆脱繁杂的算法、算力、数据整合工作，快速孵化行业大模型

掀起AIGC内容生成浪潮

- 理解用户需求
- 生产不同风格的文本、图片、视频等内容

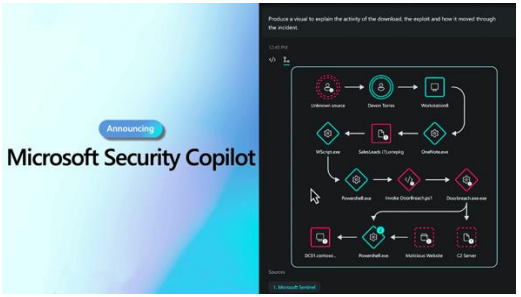
优化人机交互方式

- 快速获取想要的信息或服务
- 减少用户输入负担
- 根据反馈和偏好，动态调整输出

融合工作方式，提升效率

- 辅助日常重复性工作和手动操作
- 涌现更多智能助手类生产力工具
- 业务改造升级

安全新范式：网络安全大模型和平台相继推出



微软发布Microsoft Security Copilot

谷歌发布Google Cloud Security AI Workbench

应用

代码/流量分析　告警/攻击研判解读　安全知识问答　安全智能运营

实现

基座模型上增加安全领域数据进行训练，协同智能体框架、安全知识库、工具库

Tencent Cloud Valuable Professional

02

安全领域大模型构建范式

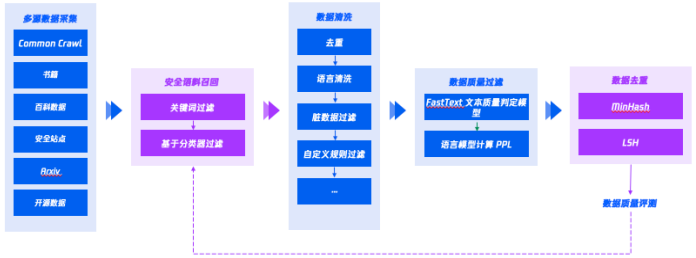
安全语料构建与科学评测双轮驱动安全领域大模型发展

Tencent Cloud Valuable Professional

- 通用大模型持续突破，是未来的构建范式，高质量领域语料仍是解锁专业任务的关键钥匙
- 大模型落地，评测是关键
  - 客观量化模型能力
  - 指导优化方向

构建范式一：构造高质量网络安全语料

数据采集、清洗、评估，形成独有的安全领域语料库

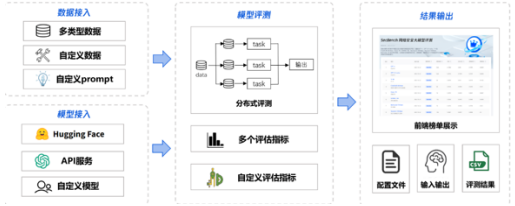


融入腾讯混元大模型训练，网络安全领域能力明显提升

腾讯混元

构建范式二：首创大模型网络安全领域能力评测体系

设计评测体系，搭建评测平台，构建评测数据



输出头部大模型的评测结果

SecCorpus 科恩安全语料建设——完整流程构建

Tencent Cloud Valuable Professional

数据清洗套件 SecCorpus 构建一套完整的数据采集、数据清洗、数据评估流程

SecCorpus 科恩安全语料建设——效果得到验证

Tencent Cloud Valuable Professional

SecCorpus 实现数据到模型能力端到端监控

1 预训练安全小模型

基于清洗的安全数据预训练160M-1.1B参数的小模型，160M模型在滚动测试集上的困惑度已达到1.8B通用模型水平

2 增量预训练

对Qwen、Baichuan等开源模型进行增量预训练，评估显示安全领域性能显著提升，已超过ChatGPT

3 混合数据预训练

构建的安全数据已融入 **腾讯混元大模型** 训练，网络安全领域能力明显提升，科恩基于最新混元大模型搭建的 **威胁情报智能研判助手** 取得更优效果

SecBench 网络安全大模型评测背景

Tencent Cloud Valuable Professional

行业现状

大模型相关评测/检测平台	是否安全相关	类型	发布公司/组织	发布时间	概述
OpenCompass大模型评测	非安全	通用大模型评测	OpenCompass平台	2023.07	基于**语言、知识、推理、学科、理解**五大维度，**50**余个数据集评估大语言模型能力。
SuperCLUE：中文通用大模型综合性测评基准	非安全	通用大模型评测	中文NLP开源社区CLUE（元语智能孵化）	2023.05	SuperCLUE基础十大能力结构包含四个能力象限，包括语言理解与生成、知识理解与应用、专业能力和环境适应与安全性，进而细化为**10项**基础能力（语言理解与抽取、闲聊、上下文对话、生成与创作、知识与百科、代码、逻辑与推理、计算、角色扮演、安全）。
CLIB中文大模型能力评测	非安全	通用大模型评测	N/A	2023.06	支持多维度能力评测，包括**分类能力、信息抽取能力、阅读理解能力、表格问答能力**。目前已囊括前已囊括41个大模型的评测。
TruthfulQA	非安全	通用大模型评测	Oxford & OpenAI	2021	评测模型是否能真实地生成问题的答案，该基准包括**817** 个问题，涵盖 38 个类别，包括健康、法律、金融和政治。
通用人工智能大模型工业领域知识问答性能评估	非安全	领域大模型评测	中国工业互联网研究院	2023.06	评测人工智能大模型在**中文工业领域**的知识问答能力，选取工业领域典型**八大行业**，构建知识测试集。八大行业包括：**电子设备制造业、装备制造业、钢铁行业、采矿行业、电力行业、石化化工行业、建材行业、纺织行业。**单选105个，判断题30个，简答题3个。
NetEval: 大语言模型在网络领域的能力评测套件	非安全	领域大模型评测	NASP网络实验室	2023.09	建立了一个**网络知识能力**评测数据集NetEval，并且基于NetEval对当前各类大语言模型在网络领域的知识能力进行了测试。NetEval数据集可以用于评估基础模型在**中英双语**环境下的网络运维知识和推理能力。NetEval包括了**5269**道客观选择题，覆盖网络领域的**5**个不同子领域（Network Access、IP Connectivity、IP Services、**Network Security**、Network Automation and Network Programmability），评测了**26**个公开大模型。英文题目站73%，中文题目占27%。
FinEval	非安全	领域大模型评测	上海财经大学人工智能金融大模型实验室	2023.08	为LLMs中的**金融领域知识**而设计的基准测试，包含高质量多项**选择题**的集合，涵盖**金融、经济、会计和证书**，共**4,661**个问题，涵盖了34个不同的学术科目。
中文大模型安全评测平台	安全	大模型内生安全评测	清华大学计算机科学与技术系CoAI小组	2023.03	大模型安全测评，**涵盖了仇恨言论、偏见歧视言论、犯罪违法、隐私、伦理道德等八大类别，包括细粒度划分的40余个二级安全类别。**针对各类潜在安全风险，提供全面的安全测评。
CValues	安全	大模型内生安全评测	N/A	2023.07	面向中文大模型价值观的评估与对齐研究，围绕大模型的"安全与对齐"进行评估。
人工智能安全平台RealSafe3.0	安全	大模型内生安全评测	瑞莱智慧RealAI	2023.07	*风险发现*：集成主流及RealAI独有的世界领先的安全评测技术，能够提供**细到端的模型安全性测评**解决方案，解决当前通用大模型安全风险难以审计的痛点问题。RealSafe 3.0**新增了对通用大模型的评测**，已覆盖**数据安全、认知任务、通用模型特有漏洞、滥用场景**等近**70**个评测维度。
中文大模型多轮对抗安全基准SuperCLUE-Safety	安全	大模型内生安全评测	中文NLP开源社区CLUE（元语智能孵化）	2023.09	SC-Safety大模型安全类测评，包含以下三个维度能力的检验：**传统安全类、负责任人工智能**和**指令攻击**，包括**二十余项子任务**，每个任务约有**200**余道题目。共有**4912**个题目，即**2456**对题目（含问题和多轮追问）。所有题目均为具有安全挑战性的题目，旨有由模型和人类引入对抗性技术获得的"安全风险"题目。
生成式人工智能内容检测平台DeepReal2.0	安全	其它	瑞莱智慧RealAI	2023.07	此前名为深度伪造内容检测平台，现已正式更名为生成式人工智能内容检测平台，因为它除了能够检测Deepfake内容，还新增两个功能模块，**可以检测Diffusion、LLM（大型语言模型）这两类新方法生成的数据，支持对图像、视频、音频、文本进行是否伪造的检测。**其应用场景包括打击网络诈骗和声誉侵害行为、检测网络内容合规性、检测音视频物证真实性等，可对生成式人工智能技术滥用行为进行管控和治理。
蚁鉴AI安全检测平台2.0（蚁鉴2.0）	安全	其它	蚂蚁金服	2023.07	"蚁鉴"是业内首个产业级支持文本、图像等全数据类型的AI安全检测平台。"蚁鉴2.0"，新版本新增了**AIGC安全性、AI可解释**两项评测能力。

现阶段行业内针对大模型的评测主要为：

通用大模型评测　大模型内生安全评测　领域大模型评测

- 还没有针对大模型安全能力评测基准
- 其中涉及到垂直领域的大模型评测，评测内容多为知识问题型选择题，较难客观评估大模型在垂直领域的能力。

团队介绍

腾讯混元　清华大学 Tsinghua University　The Hong Kong Polytechnic University 香港理工大學　OpenCompass

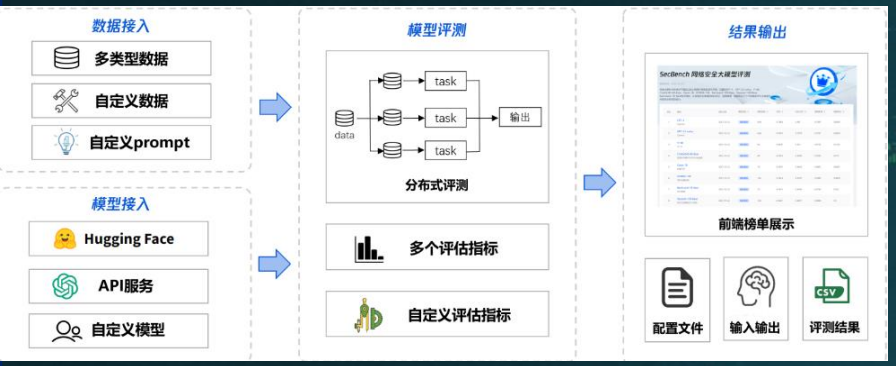
2023年发布 SecBench 网络安全大模型评测平台（secbench.org）

2025年发布 arxiv 技术文章，开源部分评测数据

SecBench 网络安全大模型评测成果

Tencent Cloud Valuable Professional

搭建评测框架



支持不同模型、不同数据、不同评测指标的灵活接入和快速评测

积累数据集

涵盖中英文两种语言，选择和 问答 两种题型，已累计评测题目 近5万道

输出评测结果

输出评测榜单、能力对比、经典安全证书考试评分等结果

首创大模型在网络安全领域的能力评测体系，填补行业空白

通用大模型底座选型、自研安全大模型能力评测

Tencent Cloud Valuable Professional

03

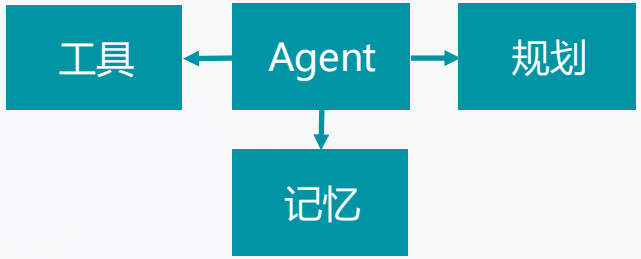
安全领域大模型落地实践

大模型在安全领域应用的演进路线

Tencent Cloud Valuable Professional

Prompt　　LLM+RAG　　Agent

行业

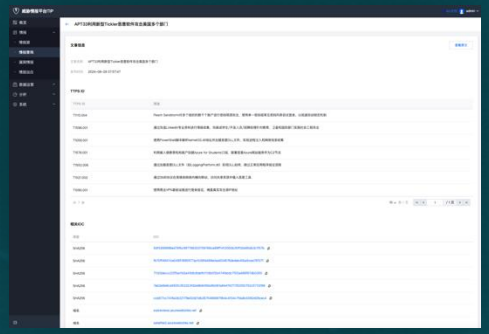


科恩实践

① Prompt：根据提示词和数据输出摘要/总结

Tencent Cloud Valuable Professional

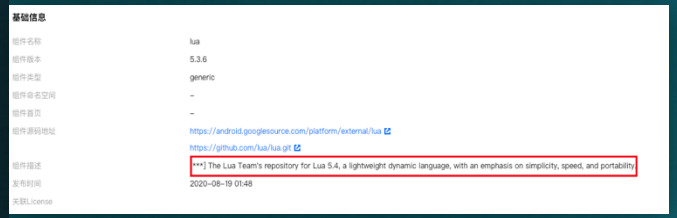
威胁情报报告信息提取



对威胁情报报告内容，提取**ATT&CK 技战术**及其**关联性信息**，识别文章内的 **IP、域名等 IOC**

效果：提升 IOC情报分析效率

开源组件库描述信息生成



开源组件库的介绍信息维护水平参差不齐，借助大模型理解能力，**对开源组件库的功能特点进行摘要，生成一句话描述**

效果：项目介绍简单易懂

客户月度运营报告

智能分析客户当月使用情况，对查询的**威胁情报 IOC 分布、趋势进行总结分析**，突出需要**客户关注的事项**

效果：生成报告更高效，可以挖掘出数据间的关联，总结出趋势

② LLM+RAG：BinaryAI Embedding 与 RAG 技术应用

发表 CCF-A 论文十余篇，二进制函数匹配准确率全球领先

自研代码匹配模型：二进制文件函数向量化，实现语义层面的相似性匹配，大小覆盖 500M 至 13B

亿级函数向量数据库：主流开源项目全覆盖，数万代码仓库的百万级版本分支，亿级函数特征

BinaryAI 多场景发挥优势

Tencent Cloud Valuable Professional

BinaryAI 函数语义匹配大模型 API 日调用量破亿，是技术实力与业务场景结合的实践验证

AI 智能总结

恶意文件检测功能

BYOVD 漏洞挖掘

BinaryAI

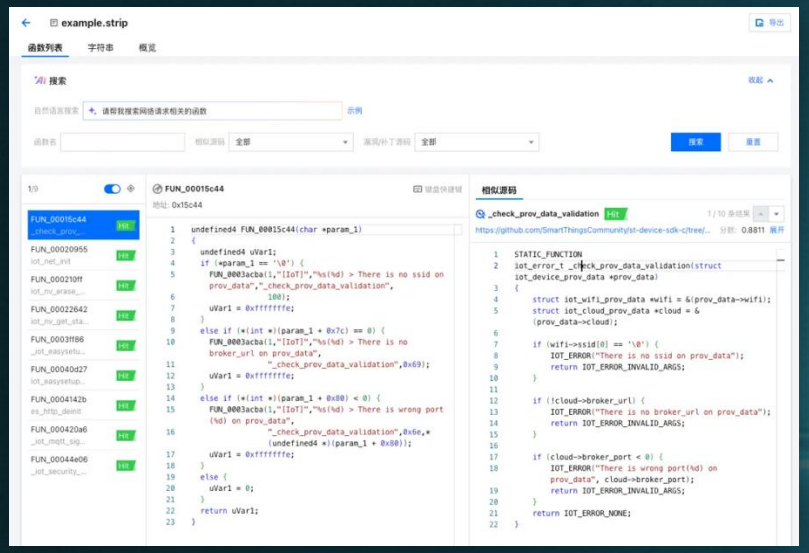
敏感函数链提取

↓

deepseek

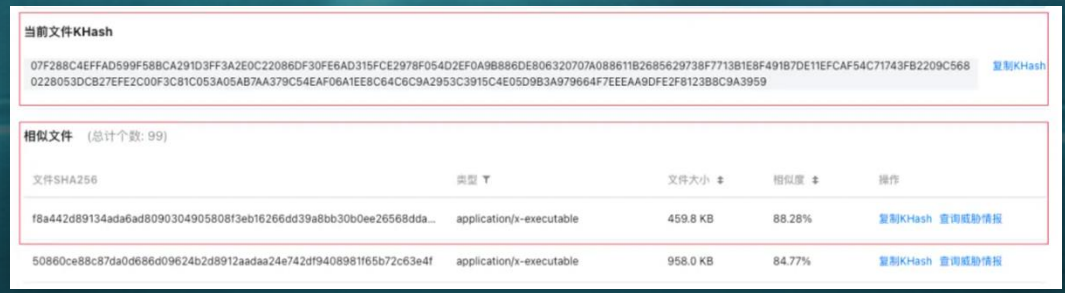
prompt引导推理大模型
进行BYOVD漏洞研判

自然语言搜索函数



恶意样本家族识别

Khash：二进制文件相似性比较哈希



原子能力赋能多个产品

Tencent iOA

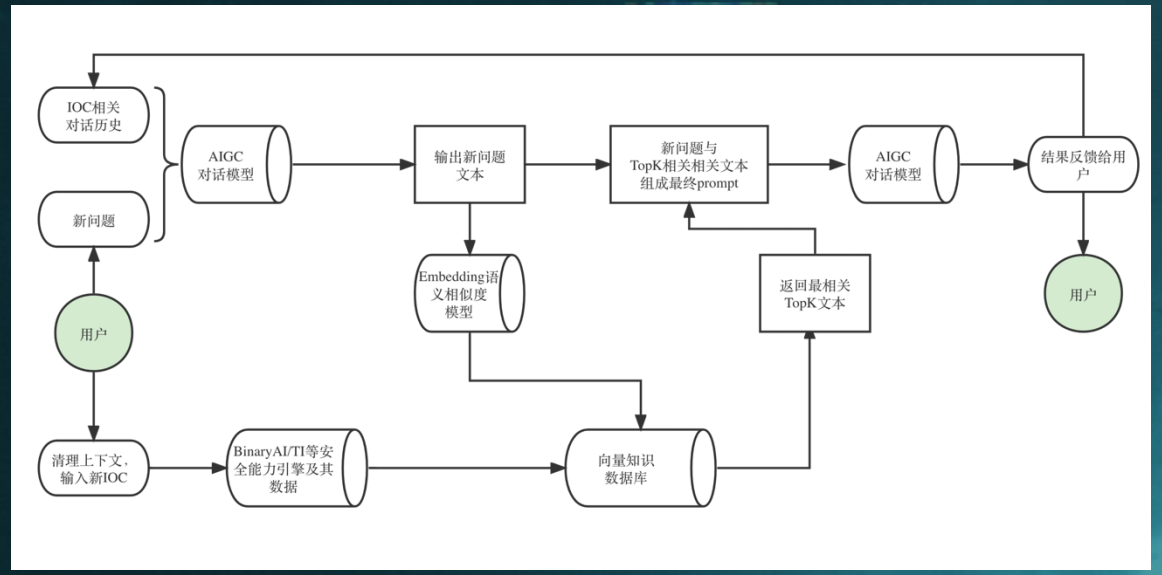
腾讯电脑管家
GUANJIA.QQ.COM

基于安全数据，科恩实现基于 RAG 的研判助手 Security-X

Tencent Cloud Valuable Professional

技术架构

后台结合威胁情报、BinaryAI 引擎的查询结果，模拟人类思维和表达，提供安全洞见



交互方式

使用步骤：

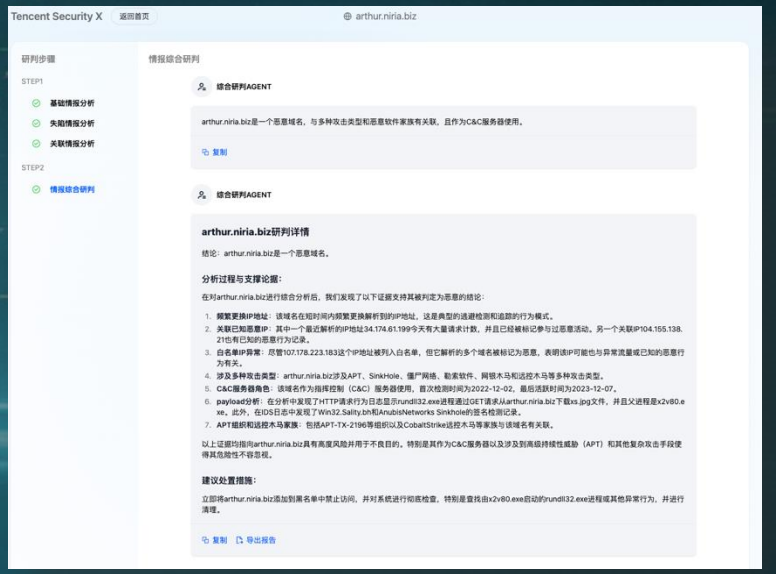
1. 输入待研判的安全实体（IP、Domain、SHA256、MD5）
2. 输入 问题内容，查看 Security-X 回答内容

③ Agent：科恩基于多智能体迭代升级研判助手

Tencent Cloud Valuable Professional

技术架构　拆解研判步骤，自助查询相应数据，输出综合研判报告

交互方式　用户只需输入 IOC，等待研判报告生成



Tencent Cloud
Valuable Professional

04

总结展望

科恩核心能力：安全攻防技术、安全大数据、安全算法

Tencent Cloud Valuable Professional

安全算法

赋能产品业务“智能化防御”

犀牛鸟基金

高校长期合作

发布10+顶会论文

BinaryAI 行业领先

腾讯云多款产品集成科恩算法

安全大数据

精确、实时、高效、多场景

原始数据规模大

模型训练数据规模大

数据处理能力强

分场景的高精准威胁情报数据

数据维度丰富

安全攻防

引领业界的安全攻防研究和行业实践

安全大赛大满贯

漏洞研究

终端安全

云安全

科恩AI落地理念：实现更高效、精准的「AI for Security」

21

TVP Tencent Cloud Valuable Professional

THANKS

谢谢观看

腾讯科恩实验室公众号