

国家信息中心主办  
中国科技核心期刊

# 信息安全研究

Journal of Information Security Research

第9卷 第6期 2023年6月

Vol.9 No.6 June 2023

GPT4发展中的若干问题及其规制方案  
大型语言模型内容检测算法和绕过机制研究  
ChatGPT在网络安全领域的应用、现状与趋势

第 6 期

CN 10-1345/TP

ISSN 2096-1057



9 772096 105235 06

邮发代号：2-41 定价：38.00元

万方数据

信息安全研究

第九卷  
第六期

二〇二三年六月

Xinxi Anquan Yanjiu

# 信息安全研究

(月刊 2015年创刊)

第9卷第6期(总第93期) 2023年6月

主管单位 国家发展和改革委员会  
主办单位 国家信息中心

出版单位 《信息安全研究》杂志社  
地址 北京市西城区三里河路58号  
邮编 100045  
电话 +86(10)68557385  
网址 <http://www.sicris.cn>  
电子信箱 [ris@cei.cn](mailto:ris@cei.cn)

执行董事兼社长 李 阳  
常务副社长 潘 静

副主编 马修军 刘 蓓  
融媒体主编 崔传桢

出版日期 每月5日  
国际标准连续出版物号 ISSN 2096-1057  
国内统一连续出版物号 CN 10-1345/TP  
广告经营许可证 京西市监广登字 20210016号  
印刷单位 北京博海升彩色印刷有限公司  
国内发行 北京市报刊发行局  
订购处 全国各地邮电局(所)  
邮发代号 2-41  
定价 38.00元

合作单位 360集团 安恒信息  
安芯网盾 航天云网  
恒安嘉新 绿盟科技  
龙象之本 蚂蚁集团  
明朝万达 奇安信集团  
青藤云安全 深信服科技  
数字认证 天融信集团  
信安世纪 亚信安全

**稿件授权声明:** 凡向本刊投稿并被录用, 由本刊支付稿酬的稿件, 均视为稿件作者同意以下条款:

**1. 文责自负。** 作者保证其拥有该作品的完全著作权(版权), 该作品不涉及政治敏感性和保密问题, 不侵犯任何他人的著作权。

**2. 全权许可。** 本刊有权以任何形式(包括但不限于通过媒体、网络、光盘等介质)使用、编辑、修改该作品, 无须另行征得作者同意, 无须另行支付稿酬。

**3. 独家使用。** 未经本刊书面许可, 作者不同意任何单位和个人以任何形式使用(包括但不限于通过媒体、网络、光盘等介质转载、张贴、集结、出版)该作品, 著作权法另有规定的除外。

**版权声明:** 未经本刊书面许可, 任何单位和个人不得以任何形式使用(包括但不限于通过媒体、网络、光盘等介质转载、张贴、集结、出版)该作品, 著作权法另有规定的除外。

本刊是中国科技核心期刊, 入选 CCF T3 优秀期刊目录, 全文被《中国核心期刊(遴选)数据库》《中文科技期刊数据库》《CNKI 中国期刊全文数据库》《超星期刊域出版平台》《博看网》收录。

## » 人工智能的安全风险与隐私保护专题

- 498** 人工智能的安全风险与隐私保护…………… 张玉清
- 500** ChatGPT 在网络安全领域的应用、现状与趋势  
…………… 张弛 翁方宸 张玉清
- 510** GPT4 发展中的若干问题及其规制方案…………… 严 驰
- 518** 基于 GPT 模型的人工智能数据伪造风险研究  
…………… 孙雷亮
- 524** 大型语言模型内容检测算法和绕过机制研究  
…………… 叶露晨 范 渊 王 欣 阮文波
- 533** ChatGPT 安全威胁研究…………… 朱孟垚 李兴华
- 543** 全生命周期数据安全管理和人工智能技术的融合研究  
…………… 张昊星 赵景欣 岳星辉 任家东
- 551** 大模型技术的网络安全治理和应对研究…………… 高亚楠
- 557** 基于人工智能和区块链融合的隐私保护技术研究综述  
…………… 李宗维 孔德潮 牛媛争 彭红利 李晓琦 李文凯
- 566** 基于隐式对称生成对抗网络的图像隐写与提取方案  
…………… 屈梦楠 靳宇浩 郭 江
- 573** 基于多注意力机制的孪生网络图像隐写分析方法  
…………… 蒋 明 张宗凯 刘熙尧 郭 标 胡家馨 张 硕
- 580** 一种基于远程证明的智能制造设备群的主动防御方案  
…………… 孔维一 李 昕 宋永立 况博裕 付安民
- 587** 基于多级度量差值的神经网络后门检测方法  
…………… 刘亦纯 张光华 宿景芳
- 593** 基于广义神经网络的网络攻击检测与分类方法…………… 张明明  
刘 凯 李贤慧 许梦晗 顾颖程 张见豪 程环宇 王永利
- 602** 中美网络安全漏洞披露与共享政策研究  
…………… 曹婉莹 曹旭栋 葛平原 张玉清

# CONTENTS 目次

## Issue on Security Risks and Privacy Protection Under Artificial Intelligence

- Security Risks and Privacy Protection Under Artificial Intelligence**  
..... Zhang Yuqing (498)
- ChatGPT's Applications, Status and Trends in the Field of Cyber Security**  
..... Zhang Chi, et al (500)
- Consideration on Some Problems in the Development of GPT4 and Its Regulation Scheme**  
..... Yan Chi (510)
- Research on Artificial Intelligence Data Falsification Risk Based on GPT Model**  
..... Sun Leiliang (518)
- Research on Content Detection Generated by Large Language Model and the Mechanism of Bypassing**  
..... Ye Luchen, et al (524)
- ChatGPT's Security Threaten Research**  
..... Zhu Mengyao, et al (533)
- Research on the Integration of Full Lifecycle Data Security Management and Artificial Intelligence Technology**  
..... Zhang Haoxing, et al (543)
- Research on Network Security Governance and Response of Large-scale AI Model**  
..... Gao Ya'nan (551)
- Towards a Privacy-preserving Research for AI and Blockchain Integration**  
..... Li Zongwei, et al (557)
- Research on Image Steganography and Extraction Scheme Based on Implicit Symmetric Generative Adversarial Network**  
..... Qu Mengnan, et al (566)
- Image Steganalysis Method Based on Multi-attention Mechanism and Siamese Network**  
..... Jiang Ming, et al (573)
- A Method of Active Defense for Intelligent Manufacturing Device Swarms Based on Remote Attestation**  
... Kong Weiyi, et al (580)
- Neural Network Backdoor Detection Method Based on Multilevel Measurement Difference**  
..... Liu Yichun, et al (587)
- Detection and Classification Method of Network Attacks Based on Generalized Neural Networks**  
... Zhang Mingming, et al (593)
- Research on the Disclosure and Sharing Policy of Cybersecurity Vulnerabilities in China and the United States**  
..... Cao Wanying, et al (602)

Journal of Information Security Research

Vol.9 No.6 June 2023

Publishing Period: Monthly

Date of Publication: 5th day of each month

International Standard Serial Number:

ISSN 2096-1057

Issues of Domestic Unit: CN 10-1345/TP

Director: Li Yang

Directed by: National Development and Reform Commission

Sponsored by: The State Information Center

Published by: Journal of Information Security Research

Address: No.58 Sanlihe Road, Xicheng District, Beijing

Zip Code: 100045

Website: <http://www.sicris.cn>

E-mail: [ris@cei.cn](mailto:ris@cei.cn)

Distributor: Beijing Press and Publication Bureau

Domestic Issue Code: 2-41

Price: 38.00RMB

### 编委会顾问

蔡吉人 崔书昆 杜虹 方滨兴 冯登国  
顾建国 何德全 李京春 吕述望 倪光南  
沈昌祥 严明 杨学山 赵战生

编委会主任 王小云

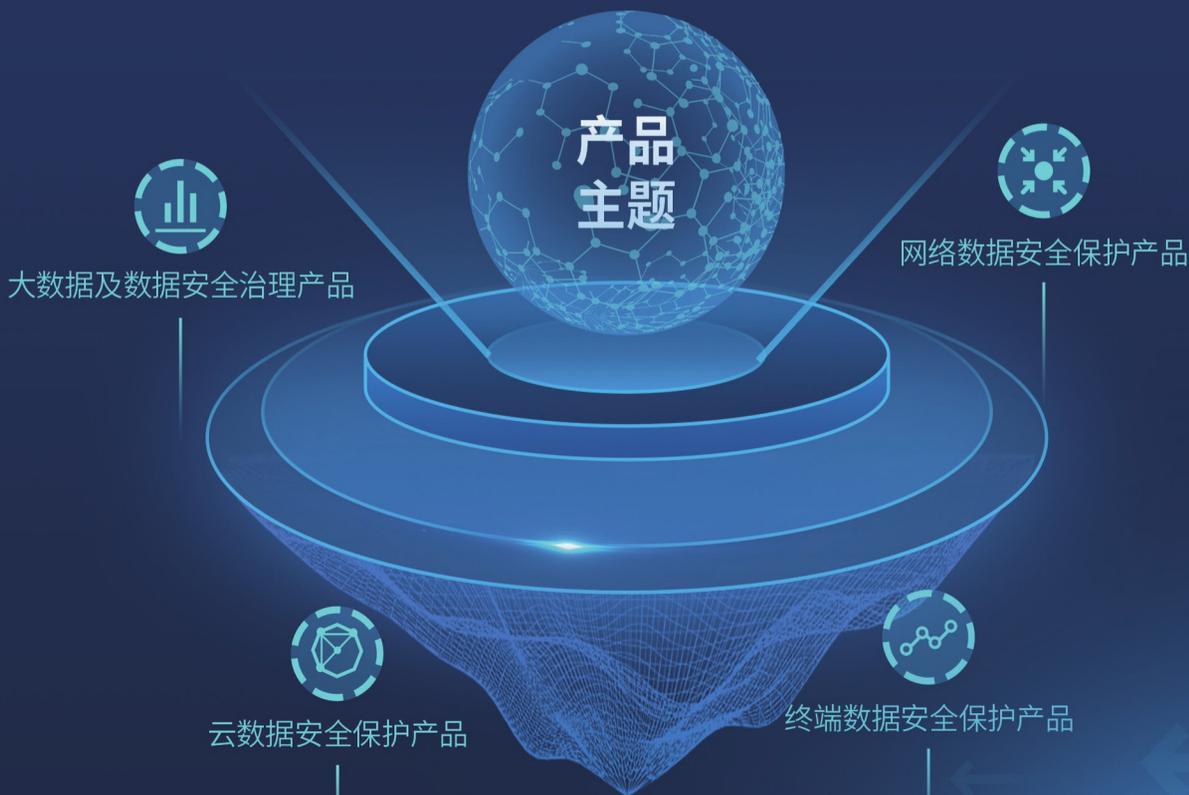
### 编委会委员

安德智 贝宇红 陈晓桦 陈兴蜀 陈钟  
程度 程学旗 杜彦辉 杜跃进 段海新  
范渊 方勇 封化民 谷大武 顾健  
郭莉 郭艳卿 胡爱群 胡红钢 胡红升  
黄伟庆 贾焰 晋钢 荆继武 姜向前  
康海燕 孔祥维 李晖 李剑 李建彬  
李建华 李守鹏 李涛 李小勇 李欲晓  
李舟军 林东岱 林家骏 林雪焰 刘宝旭  
刘东红 刘吉强 刘建伟 刘云 罗森林  
马民虎 马智 孟小峰 潘泉 彭长根  
秦安 卿昱 任奎 任卫红 苏金树  
苏洲 孙德刚 孙伟 谭晓生 谭毓安  
唐春明 滕颖志 田俊峰 田志宏 王宝生  
王标 王军 王美琴 王志海 王志强  
韦韬 温巧燕 文伟平 翁健 吴文玲  
吴云坤 吴志军 席卿 肖新光 许光全  
许力 杨庚 杨满智 杨义先 叶红  
叶晓虎 于锐 俞能海 云晓春 张滨  
张功萱 张宏莉 张焕国 张健 张建标  
张建军 张庆勇 张仕斌 张小松 张玉清  
赵波 赵淦森 赵有健 郑东 郑方  
周福才 周琳娜 周民 周世杰 周文  
朱建明 朱岩 祝烈煌 邹德清 邹维

北京明朝万达科技股份有限公司成立于2005年，是中国新一代信息安全技术企业的代表厂商，专注于数据安全、公共安全、云安全、大数据安全及加密应用技术解决方案等服务。公司现有员工600余人，总部设于北京，在上海、广州、成都、西安、贵阳、天津、武汉、南京、无锡、长春等地设有分支机构。凭借在数据安全领域取得的优异成就，明朝万达于2019年获得中央网信办背景中网投、国家发改委背景国投创投联合投资，并于2020年获得中国电科集团（CETC）战略投资。

基于“动态数据安全，数据全生命周期管控”的产品理念，明朝万达始终以守护用户数据价值为己任，致力于让安全真正服务于业务发展。

历经十余年的发展与积累，明朝万达秉承“安全铸就数据价值，安全服务用户业务”的发展理念，强调数据为核心，安全为准绳，用业务驱动数据，以安全服务业务与数据的思想，形成了与用户业务系统紧密结合的数据安全产品和服务体系，客户已覆盖金融、政府、公安、电信运营商、能源、制造等诸多行业。



官方微信

地址：北京市海淀区阜外亮甲店1号恩济西园产业园16号楼B座  
电话：010-82743939  
邮编：100142  
热线：400-650-8968  
网址：www.wondersoft.cn