CSA GCR cloud security alliance
GREATER CHINA REGION ®

#隐私计算
#国际人才认证
#数据安全
#数字安全
#国际峰会
#元宇宙安全
#云安全
#区块链
#零信任

安全可靠的垂直领域大模型

汇报人：于伟

目录

CONTENTS

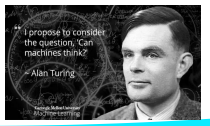
1 背景介绍

大模型技术突破让人工智能进入大规模应用阶段

以ChatGPT 为代表的语言大模型技术的提升突破了生成式AI的瓶颈，具有更强的商业化潜力，已在艺术创作、代码处理、文章撰写、市场营销等多个创造性场景得到应用，语言大模型相关的市场将呈指数增长。

ChatGPT让AI应用进入3G+iPhone时代

	移动应用	AI应用	应用数量
第一代	1G+摩托罗拉	传统算法+弱算力	极少
第二代	2G+黑莓手机	深度学习+强算力	少
第三代	3G+苹果手机	超大模型+强算力	海量



1950年，图灵提出著名的图灵测试，给出判断机器是否具有“智能”的方法。

1997年5月，IBM的深蓝战胜棋王卡斯帕罗夫。

2016年3月，谷歌的AlphaGo击败围棋世界冠军李世石。

2022年12月，OpenAI的ChatGPT展示了与人类媲美的语言能力。

金融
研报生成
分析报告

政务
舆情研判新
闻摘要
文书生成

医疗
病例生成
治疗方案生成

法务
合同自动
审阅

传媒
创意写作
营销文案

其他

商业价值将呈指数级增长

第三代AI

第二代AI

第一代AI

1950 1997 2016 2018 2022

大模型安全风险案例

- **案例一**：三星电子引入ChatGPT不到20天，就曝出3起内部机密资料外泄事故。其中，2起跟半导体设备有关，1起跟会议内容有关。这些机密资料可能已被存入ChatGPT的数据库中，随时都面临着被泄露的风险。
- **案例二**：3月31日，意大利个人数据保护局宣布，从即日起禁止使用ChatGPT，限制ChatGPT的开发公司OpenAI处理意大利用户信息，并立案调查。
- **案例三**：3月份，ChatGPT被曝出现bug会导致用户对话数据、支付信息丢失泄露。这一度导致了ChatGPT短暂关闭。OpenAI的回应中表示，这一漏洞可能导致了1.2%ChatGPT Plus用户的支付信息被暴露了约9个小时。漏洞还导致了用户的对话主题及记录能被别人看到，如果里面包含隐私信息，则有泄露的风险。

❷ 安全分析

大模型安全风险分析

个人隐私

知识产权

训练数据

恶意行为

个人隐私

一．当用户使用大模型服务时，用户输入的信息会被记录下来，并可能被永久存储，包括用户输入的所有敏感信息和个人身份信息等。

二．如果大模型服务本身的安全防护措施不够，非常容易导致敏感信息和个人身份信息的泄露。譬如， 3月份ChatGPT被曝出现bug会导致用户对话数据、支付信息丢失泄露。

三．如果这些信息被用来作为训练数据，那就可以随时面临着被泄露的风险。

四．用户需要避免在与大模型服务的对话中输入敏感信息或个人身份信息。

知识产权

一．训练大模型的数据大都来自公开的文本数据集，但是公开不代表不受知识产权保护。

二．网上公开的数据集也可能包含用户的个人信息。

三．用户输入的数据也可能被用来做训练，导致用户知识产权丢失。

四．如果用户输入的数据本身就有知识产权问题，那模型本身也会侵权。

训练数据

一．大模型使用从各种来源收集的数据集进行训练，包括代码库、百科、社交媒体、公共论坛等，庞大的数据使得大模型具有我们今天看到的优秀性能。

二．高质量数据是训练出高质量模型的基础。低质量或错误数据会导致模型提供错误的结果。训练数据的准备及其关键，需要防止错误甚至恶意数据被用来训练，更要防止数据投毒。

三．从海量训练数据中发现并筛除错误甚至恶意数据也是一个巨大的技术和成本挑战。

恶意行为

一．大模型本身也会犯错误，譬如一本正经的胡说八道。如果用户轻信大模型输出的结果，有可能带来不可挽回的巨大损失。

二．最近，已经有人利用大模型的能力进行违法犯罪行为，包括编写恶意软件、生成钓鱼电子邮件、冒充真实的人或者组织骗取他人信息等。

三．随着与大模型对接的服务不断增加，大模型控制物理世界和网络世界的能力不断增加，大模型本身的不可靠性也会随时给网络世界和物理世界的用户带来巨大风险。

❸ 模型训练

聚焦垂直领域

一．尽管通用大模型已经达到优秀高中毕业生甚至未来能达到优秀本科毕业生的水平，但是很多场景下，通用大模型无法提供实际业务应用中所需要的专业能力

二．由于专业能力的训练需要专业的知识和数据，而专业的知识和数据大都属于私有数据，鉴于安全和知识产权等因素，无法提供给外部厂商用于训练通用大模型

三．因此，基于专业知识和数据训练安全可靠的垂直领域大模型势在必行。

模型训练机制

一．**训练数据**：需要高质量专业训练数据，需要有完善的机制剔除掉问题数据、错误数据等。

二．**价值对齐**：需要通过训练数据和训练过程完成价值对齐，确保模型符合预期。

三．**基础模型**：如果垂直领域大模型是基于基础通用大模型进行训练的，需要做好基础模型的检查，避免因为基础模型本身带来的问题。

安全防护

一．**系统安全**：确保系统本身的安全，防止信息泄露和安全攻击

二．**数据防护**：建立训练与反馈数据质监机制，在模型迭代过程中防止数据投毒等行为。

三．**权限管理**：建立完善的模型行为权限管理机制，防止模型本身的不可控甚至恶意行为。

❸ 未来展望

未来展望

一．随着经济的发展，数字经济占GDP的比重越来越高，中国逐步进入数智经济时代，利用AI技术提升生产力为大势所趋。随着各行业、各领域对 AI 需求的日益增长，与实体经济深度融合的新模式不断涌现。

二．在应用大模型的过程中，需要在模型层面应理解“通”与“专”的相对性，在数据层面应把握“大”与“小”的辩证关系，在交互界面应推进“人”与“机”的协同互动。

三．通过打造安全可靠的垂直领域大模型，加速中国数智经济的发展。

CSA GCR cloud security alliance
GREATER CHINA REGION ®

THANKS

0755-86548359
info@c-csa.cn
https://www.c-csa.cn/
广东省深圳市南山区粤海街道北京航空航天大厦9楼905