CSA GCR 6th Congress 第六届云安全联盟大中华区大会

CSA GCR cloud security GREATER CHINA REGION alliance ®

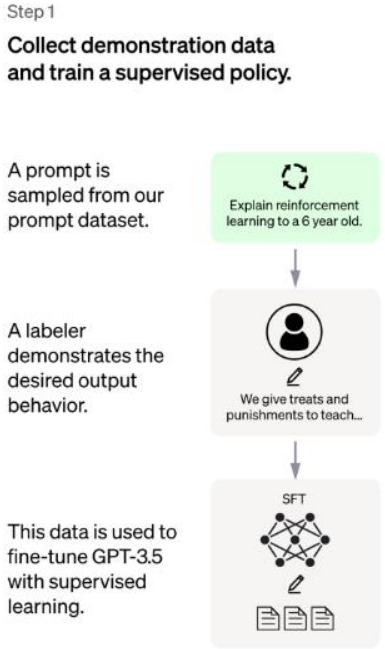
演讲题目：ChatGPT和Plug-In 的 应用和安全

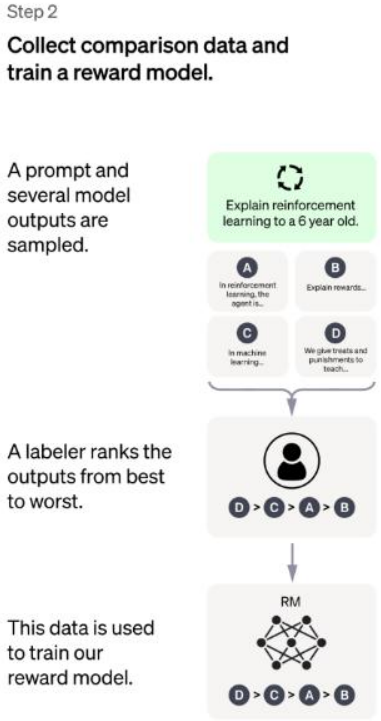
- 演讲人：黄连金
- 单位名称　云安全联盟大中华区

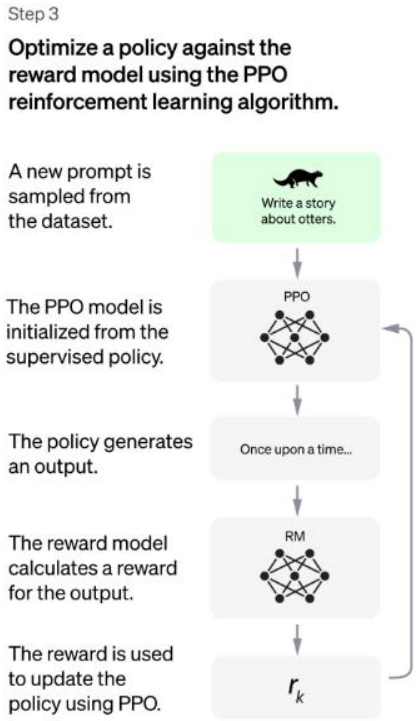
人工智能的三要素和三类型

- 三要素
  - 数据（从演绎推理到归纳推理，需要大量数据，包括Prompt）
  - 算法 （深度学习， ChatGPT在工程上有大量创新和改进）
  - 算力 （需要大量GPU， ChatGPT 用到1万个英伟达的GPU, 28万CPU）
- 三种类型
  - 弱人工智能：就是利用现有的算法辅助各个行业的应用。
  - 强人工智能：非常接近于人的智能，ChatGPT。
  - 超级人工智能：这个是否可能？

ChatGPT 的技术原理：基于Transformer的大型语言模型







1：大型语言模型可以体现智能

2：预测可以体现智能

3：人类监督学习和反馈机制是关键

4：利用PPO: 近端策略优化（Proximal Policy Optimization) 增强学习来持续提高智能

5：目前在写代码和文章已经有优势

ChatGPT+Plug-In: 从AIGC 到 AIGE（Economy)

- AI 产生的新经济系统和商业模型已经不可逆转
- 价值效应超过：操作系统，Web和AppStore
- OpenAI，Microsoft 有先发优势，但是。。。
- 不注意安全会带来灾难

Elon Musk: ‘Mark my words — A.I. is far more dangerous than nukes’

PUBLISHED TUE, MAR 13 2018•1:22 PM EDT | UPDATED WED, MAR 14 2018•11:31 AM EDT

Catherine Clifford
@IN/CATCLIFFORD/
@CATCLIFFORD

SHARE

TV

Squawk on Street

UP NEXT | Squawk
11:00 am ET

HP FL

Elon Musk and Apple cofounder Steve Wozniak among over 1,100 who sign open letter calling for 6-month ban on creating powerful A.I.

BY **JEREMY KAHN**

March 29, 2023 at 6:34 AM EDT

Elon Musk is among the prominent technologists who have called for a six-month pause on the development of more powerful A.I.

MARLENA SLOSS—BLOOMBERG VIA GETTY IMAGES

Most Popular

TECH

Elon Musk and Apple cofounder Steve Wozniak among over 1,100 who sign open letter calling for 6-month ban on creating...

March 29, 2023

BY **JEREMY KAHN**

ChatGPT banned in Italy over privacy concerns

2 days ago

GETTY IMAGES

OpenAI launched ChatGPT last November

ChatGPT数据泄露已经发生

泄露的对话和提示：2023年3月22日，ChatGPT的一些对话和提示被随机泄露给其他用户，这是对GDPR法规的明显违反。另外支付信息泄露违反PCI/DSS

SECURITY WEEK

ChatGPT Data Breach Confirmed as Security Firm Warns of Vulnerable...

Infosecurity Magazine

ChatGPT Vulnerability May Have Exposed Users' Payment Information

Yesterday

TechRadar

What to do when ChatGPT is down

19 hours ago

Search Engine Roundtable

ChatGPT Bot - You Can Block OpenAI Plugins If You Want

Yesterday

ChatGPT的安全问题 （CSA CEO： Jim Reavis）

安全问题	描述
社会工程	攻击者可以利用从社交媒体或其他来源收集的大量文本数据来训练模型，生成极具说服力的网络钓鱼电子邮件或消息，诱骗受害者泄露敏感信息。
撞库攻击	攻击者可以利用语言模型生成大量潜在的用户名和密码组合，用于对在线帐户进行自动攻击。
垃圾邮件和虚假信息	恶意行为者可以使用语言模型生成大量旨在影响公众舆论或传播错误信息的垃圾邮件和虚假信息。
生成恶意软件	攻击者可以利用ChatGPT编写恶意软件说明和指令，从而逃避防病毒软件的检测。
创建虚假社交媒体资料或聊天机器人帐户	恶意行为者可以使用ChatGPT收集敏感信息，创建虚假资料冒充真实的人或组织来诱骗受害者提供个人信息，例如登录凭据或信用卡号。
生成旨在操纵或欺骗受害者的自动消息	恶意行为者可以使用ChatGPT在社交媒体或论坛上生成数千条自动消息，传播虚假信息或宣传，以影响公众舆论或破坏政治活动。

ChatGPT Plug-in的安全防护措施

- Access control and user authentication (访问控制和用户身份验证): 限制对ChatGPT插件的访问并确保只有授权用户可以使用。
- 利用去中心化数字身份DID
- 利用智能合约约束AI
- Rate limiting and spam prevention (限制访问速率和防止垃圾信息): 对访问请求进行限制，以防止恶意攻击和垃圾信息泛滥。

ChatGPT 和 Plug-In的安全趋势

ChatGPT安全趋势	描述
制定生成AI使用政策	制定关于使用生成AI的政策，以确保道德和负责任的使用。
现代化安全审计	安全审计实践正在现代化，以跟上AI安全风险不断变化的形势。
更加关注数据卫生和评估偏见	更加关注数据卫生和评估AI模型中的偏见，以确保它们不具有歧视性或有害性。
利用AI优化网络安全投资	组织正在利用AI优化其网络安全投资并提高其整体安全水平。
加强威胁情报	通过使用AI增强威胁情报，以更好地检测和应对安全威胁。
威胁预防和管理合规风险	AI被用于威胁预防和管理合规风险，以减少安全事件发生的可能性。
实施数字信任战略	数字信任战略利用AI与客户和利益相关者建立信任。

CSA GCR 6th Congress 第六届云安全联盟大中华区大会

CSA GCR cloud security GREATER CHINA REGION alliance®

THANK YOU.