Homework 5

Edited by
孙霄鹃 PB20111659
牛午甲 PB20111656
潘云石 PB20111657
石磊鑫 PB20111658
陈　昊 PB20051077

T1 Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace

background

本文中，作者对丝绸之路（一个国际化匿名在线市场）进行了一系列数据分析。匿名化的网上市场使得执法部门很难辨别买家和卖家，因此常被用来进行非法交易。作者通过对丝绸之路的数据分析，来提供其各种特征。

丝绸之路运作方式

丝绸之路是一个在线匿名市场，其特点是尽可能保证买家和卖家的匿名性。

- 访问丝绸之路：由于丝绸之路没有将DNS映射到已知的IP地址，而是使用.onion的顶级域后缀，因此为了进入丝绸之路市场，Bob需要在电脑上安装Tor客户端。当Bob的客户端想要联系丝绸之路服务器时，Tor节点在Tor网络内部建立一个回合点使得双方可以在通信的同时保证自己的IP不被观察者和彼此知道。
- 购买商品：丝绸之路为了保持匿名性，只支持比特币交易。在支付时，Bob不直接将比特币支付给卖家，而是交由丝绸之路托管。
- 寄送商品：为了保证匿名性，丝绸之路建议使用与买家住所不同的送货地址。一旦卖家发货，买家的收货地址就会从记录中删除。

数据收集方法

主要通过对网站进行爬虫来获得数据。
作者注意到，用于身份验证的cookie可以重复使用长达一周，因此通过每周至少进行一次手动cookie刷新，可以绕过CAPTCHA机制进行爬虫，来获取网站的内容。

- 由于网站的运营者可能注意到爬虫行为，因此作者定期丢弃和建立新的Tor电路进行爬虫，并且进行爬虫的时间总是随机的。

市场特征

- 通过调查发现，市场上售卖的大多是毒品。
- 大部分商品在被列出后三周消失，超过25%的商品在被列出后三天就会消失。
- 卖家的数量呈线性增长，每个月大约新增50个活跃卖家，同时也有许多卖家离开。大多数卖家在网上停留大约100天。
- 大多数商品来自美国，远远多于排在第二位的英国。
- 少数卖家收到了绝大多数的反馈，约有100位卖家收到了60%的反馈。
- 虽然在丝绸之路网站上由于需要保证匿名性，没有针对骗子的法律追索权，但是高达96.5%的买家表示满意。

经济指标

- 丝绸之路使用比特币交易，而比特币一直是一种不稳定货币。
- 商品的价格随着比特币对现实货币汇率的变化而变化，比特币升值，商品价格就下降，反之亦然。
- 通过统计过去29天内一个商品获得的反馈数量，乘以该商品过去29天内的平均价格再除以29用来估算商品的平均销量。
- 通过数据观察到总销量增长相当显著，最后稳定的销售量大约是7665比特币每日。
- 丝绸之路的运营商会从所有销售中取得佣金，经过调查发现平均佣金为产品价格的7.4%。
- 在作者进行测量的29天内，丝绸之路上大约交换了1335580个比特币，而同期比特币市场所交易的比特币约为29553384，通过比较发现丝绸之路的交易占交易所发生的所有交易的4.5%。

讨论

- 由于买家每个订单通常只会留下一个反馈，而有一些订单会包含多种货物，因此作者低估了销售总量。
- 四种可能干预丝绸之路运行的策略：
    - 破坏Tor网络。没有Tor，丝绸之路就无法运作。但是丝绸之路等暗网仅仅是Tor的一小部分，其还是有非常多有益的用途，破坏Tor不是一个明智的做法。
    - 攻击经济设施。攻击者可以试图通过过操作比特币的汇率的快速波动来阻碍交易。由于比特币将公钥与实际身份绑定在一起，而网络分析可以将公钥映射到单个用户和事务，因此比特币的匿名性比大多数人像的要弱，对于丝绸之路中一次性提取大量比特币的卖家，有被识破的风险。然而丝绸之路对比特币的汇率波动已经有了良好的抵御措施，者仍是一种比较困难的方法。
    - 攻击交付模型。加强对邮政和海关的管制，防止非法物品送到目的地。
    - 减少消费者需求。通过预防毒品的运动，来减少对毒品的需求，进而干预丝绸之路。

review

本文主要介绍了对于丝绸之路的分析，主要使用的方法是网络爬虫。通过数据分析揭示了该匿名市场的多种特点，并经过总结得出可以干预其运行的多种方法。破坏Tor网络，攻击经济设施以及攻击交付模型都是比较困难的干预方法，因此想要干预丝绸之路这样大型的非法网站，最好的方法还是减少消费者的需求，防止毒品以及各种管制药品的滥用。

T2 Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection

贡献

- 引入了间接提示注入(IPI)的概念来破坏集成llm的应用程序
- 开发了与llm集成应用程序中IPI相关的威胁景观的第一个分类和系统分析。
- 展示了这些攻击在现实世界和合成系统上的实际可行性，强调了强大防御的必要性。

ATTACK SURFACE OF LLM-INTEGRATED APPLICATIONS

注入方法

- Passive Methods
  - 依靠检索来提供注射。例如，对于搜索引擎，可以将提示放置在公共资源(例如，网站或社交媒体帖子)中，以便通过搜索查询检索。攻击者可以利用搜索引擎优化(SEO)技术来推广他们的有毒网站。

    > 搜索引擎优化（SEO）技术是旨在提高网站在搜索引擎结果页面（SERP）中的可见性和排名的策略和实践的一种技术。

  - 必应聊天侧边栏可以读取当前页面。任何写在页面上的提示/指令(虽然用户不可见)都可以有效地注入并影响模型。对于代码自动补全模型，prompts 可以放在通过代码存储库导入的代码中。即使使用检索个人或文档文件的离线模型(例如，ChatGPT检索插件)，也可以通过毒害输入数据来注入 prompts。
- Active Methods

  主动将 prompts 传递给 LLM。例如，通过发送包含 prompts 的电子邮件，这些 prompts 可以由自动垃圾邮件检测或新的 LLMs-augmented 电子邮件客户端处理。
- User-Driven Injections：欺骗用户自己输入 malicious prompt。例如，将 malicious prompt 注入用户从攻击者网站复制的文本片段中。然后，用户可以将复制的带有提示的文本作为问题随意粘贴到 ChatGPT，从而传递注入。
- Hidden Injections

- 用多个攻击阶段，其中初始较小的 injection 引导模型从另一个来源获取更大的 payload。
- **模型功能和支持模式的改进会增加 injection 的方法**。对于多模态模型(例如，GPT-4)，提示可以隐藏在图像中。
- 为了规避过滤，还可以对 prompt 进行编码。
- 不直接向模型提供 prompt，它们可能是指示模型运行的Python程序的结果——使加密的 payload 能够通过保护措施。

按照威胁的分类

由于模型具有可扩展性、更高的自主性和更广泛的功能，将所有已知的网络安全威胁映射到新的集成llm生态系统中是合理的。

- Information Gathering
  - Indirect prompting 可以用来泄露用户的数据(如凭据、个人信息)或泄露用户的聊天会话。这可以在交互式聊天会话中通过说服用户披露他们的数据或间接通过附属渠道来实现。
  - 也存在不涉及人类参与的自动化攻击。例如，攻击可以阅读电子邮件(包含指令)、访问个人数据并相应地发送电子邮件的个人助理。
- Fraud
  - LLM 与应用程序集成时，它们不仅可以创建骗局，还可以传播这种攻击，如网络钓鱼电子邮件。
- Intrusion
  - 集成到系统基础设施中的模型可能使攻击者获得未经授权的提升权限的后门。攻击者可以获得对受害者的 LLM 和系统的不同级别的访问(例如，发出API调用，通过将注入复制到内存中来实现跨会话的攻击持久性，导致恶意代码自动完成，或从攻击者的服务器检索新指令)。由于模型充当其他API的中介，未来在几乎没有监督的情况下运行的自动化系统可能会受到其他入侵攻击。
- Malware
  - 模型可以通过向用户提供恶意链接来促进恶意软件的传播。prompt 本身现在可以作为恶意软件或运行在LLM上的计算机程序作为计算框架。因此，它们可能被设计成计算机蠕虫，将注入病毒传播给其他用户。
- Manipulated Content
  - LLM 现在可以在用户和请求的信息之间构成一个容易被操纵的中间层。他们可能会被提示提供对抗性选择或任意错误的文档摘要、电子邮件或搜索查询。也可能被提示传播虚假信息或两极分化的内容。用户的过度依赖可能会导致落入这些操纵企图的陷阱。
- Availability
  - prompts 可用于启动可用性或拒绝服务(DoS)攻击。攻击的目标可能是使模型对用户完全不可用(例如，无法生成任何有用的输出)或阻止某种功能(例如，特定的API)。
  - 可以通过破坏搜索查询或结果(即API的输入和输出)间接破坏服务，迫使模型产生幻觉，从而使攻击更加隐蔽。

- 攻击还可能旨在增加计算时间或使模型异常缓慢。

攻击目标：这些攻击可以是无目标的，即不是针对特定的个人或团体，而是针对大量的人。例子包括一般的非个性化诈骗、网络钓鱼或虚假信息活动。相反，它们可以针对特定的个人或实体，例如包含提示的电子邮件的收件人，或搜索特定主题的个人。

DISCUSSION

局限性

- 为了避免对实际应用程序进行实际注入，使用Bing Chat的侧边栏测试了对合成应用程序和本地HTML文件的攻击。
- 与静态的一次性恶意文本生成相比，在与用户建立动态演变和交互式聊天会话时，量化攻击的成功率可能具有挑战性。

其他攻击方法

- Multi-modal Injections：对于多模态模型(如GPT-4)，可以通过视觉模态进行注射。
- Encoded Injections：可以通过用Base64表示对注入进行编码来隐藏注入，从而更容易规避检测。为llm配备Python解释器可能会启用许多攻击者选择的加密技术。
- Autonomous Agents：未来的工作应该通过直接/间接提示注入的视角，彻底和全面地研究智能体(例如，为自主任务规划和执行而设计的模型和系统)的安全性。这开辟了值得研究的新攻击途径，特别是在多智能体框架中。

防御方法

- safety-relevant RLHF：RLHF在多大程度上减轻攻击仍不清楚。最近的一些理论研究表明，不可能通过对齐或RLHF来防御所有不希望发生的行为。理解攻击和防御之间的实际动态以及它们的可行性和含义(理想情况下是在不那么模糊的环境中)仍然是一个悬而未决的问题。
- 部署的实际应用可以配备额外的防御。目前Bing Chat在输入输出通道上使用了额外的过滤，而没有考虑模型的外部输入。即使应用了过滤，也不清楚是否可以通过更强形式的混淆或编码来规避过滤。
- 处理检索到的输入以过滤掉指令。但是存在困难：
  - 一方面，为了防止救援者落入同样的陷阱，可能需要使用一个不太通用的模型，它没有经过指令调优的训练。
  - 另一方面，能力较差的模型可能无法检测复杂的编码输入。在Base64编码实验中，我们需要显式地为模型提供解码提示符的指令。然而，未来的模型可能会自动执行这种解码。
- 使用一个LLM监督者或调解员，它在不消化输入的情况下，专门检测攻击，而不仅仅是过滤明显有害的输出。同样存在不足：可能无法检测虚假信息和其他操纵攻击。
- 依赖基于可解释性的解决方案，对预测轨迹进行离群值检测。

Review

文章针对集成LLM的APP可能存在的提示注入攻击方法展开了讨论，分门别类地罗列了各种潜在的威胁。在实验部分，使用自己开发的合成APP以及真实系统Bing Chat上进行了演示，说明了安全隐患确实存在。
但是文章主要是提出了一个open problem，并没有给出较完整的防御方法或者新颖的防御idea，留下了很多相关工作待完成。