# Article information:

[2012.07805] Extracting Training Data from Large Language Models  
<https://arxiv.org/abs/2012.07805>

# Article summary:

1. 本文介绍了一种从大型语言模型中提取训练数据的攻击方法。作者展示了在这种情况下，对语言模型进行查询可以恢复出个别的训练样本。他们在GPT-2上进行了攻击实验，并成功地从模型的训练数据中提取出数百个完全相同的文本序列，其中包括公开可见的个人身份信息、IRC对话、代码和128位UUID等敏感信息。

2. 文章对攻击方法进行了全面评估，以了解导致攻击成功的因素。令人担忧的是，研究发现较大规模的语言模型比较小规模的模型更容易受到攻击。作者总结了经验教训，并讨论了保护大型语言模型训练过程可能采取的措施。

3. 该研究揭示了使用大规模语言模型时存在的隐私风险。由于这些模型通常是基于私有数据集进行训练的，因此潜在攻击者可以通过查询来获取原始训练样本，进而泄露敏感信息。文章呼吁加强对大规模语言模型训练过程中隐私保护措施的研究和实施。

# Article rating:

Appears strongly imbalanced: The article is written in a biased or one-sided way, and the information it provides is not trustworthy enough to be considered a reliable source. You should consult other sources to find reliable information on the presented issues.

# Article analysis:

对于上述文章的详细批判性分析如下：

1. 潜在偏见及其来源：文章没有明确提到作者的背景和利益相关方，这可能导致潜在的偏见。读者无法确定作者是否有与大型语言模型相关的商业或个人利益。

2. 片面报道：文章主要关注了从大型语言模型中提取训练数据的攻击方法和成功案例，但未探讨其他可能的应用场景或潜在好处。这种片面报道可能会给读者留下负面印象，而忽视了大型语言模型在自然语言处理等领域中的积极影响。

3. 无根据的主张：文章声称更大规模的语言模型比较小规模的模型更容易受到攻击，但没有提供充分的证据来支持这一观点。缺乏实验证据使得这个主张缺乏说服力。

4. 缺失的考虑点：文章没有深入探讨攻击者需要具备什么样的条件才能成功地从语言模型中提取训练数据。例如，攻击者是否需要对目标语言模型有特定的访问权限？这些考虑点对于评估攻击风险和制定相应防护措施至关重要。

5. 所提出主张的缺失证据：文章提出了一些可能的防护措施，但没有提供充分的证据来支持这些主张。读者无法确定这些措施是否真正有效，或者是否存在其他更好的解决方案。

6. 未探索的反驳：文章没有深入探讨可能存在的反驳观点或对其攻击方法和结果进行质疑。这种未探索的反驳可能导致读者对文章中所述攻击风险的真实性产生怀疑。

7. 宣传内容和偏袒：文章没有平衡地呈现大型语言模型的利与弊，而是更加强调攻击风险和潜在问题。这种宣传内容可能会给读者留下负面印象，并导致对大型语言模型的误解。

8. 是否注意到可能的风险：文章提到了从大型语言模型中提取训练数据的攻击风险，但未详细讨论这些风险对个人隐私、数据安全和社会影响等方面可能造成的潜在影响。这种缺乏全面考虑可能使得读者无法全面评估相关风险。

总体而言，上述文章在对大型语言模型的攻击风险进行分析时存在一些问题，包括潜在偏见、片面报道、无根据的主张、缺失的考虑点和证据不足等。读者需要谨慎对待其中提出的观点，并进一步研究相关领域的其他观点和证据。

# Topics for further research:

* 作者背景和利益相关方
* 其他应用场景和潜在好处
* 更大规模语言模型受攻击的证据
* 攻击者需要的条件和访问权限
* 防护措施的有效性和其他解决方案
* 反驳观点和攻击方法的质疑
* 大型语言模型的利与弊的平衡呈现
* 攻击风险对个人隐私、数据安全和社会影响的潜在影响

# Report location:

<https://www.fullpicture.app/item/f78c698d3709480ceb9d54cf1635ce63>