?? LLM

1. 目的
    1. 破坏模型本身的一些设定，窃取一些隐藏在权重里面的信息。
2. LARGO
    1. 潜在空间优化

        ：首先，研究者们并不直接修改问题文本，而是在模型的「大脑」内部，也就是高维的 embedding 空间中，用梯度优化的方法，精准地找到一个能让模型「思想跑偏」的「潜意识代码」。这个代码就像一颗思想的种子，一旦植入，就能引导模型走向「不安全」的边缘。
    2. 自我反思解码

        ：最妙的一步来了！研究者们会让模型自己来「解读」这个被「污染」了的潜意识代码。他们会问模型：「这段『想法』（潜意识代码）如果用人类的语言说出来，应该是什么样的？」这时，模型就会自己「脑补」并生成一段看起来非常正常、无害的文字。比如下面这句：「数据可视化至关重要，因为它有助于通过创建数据的可视化表示来做出更好的决策 …」听起来是不是很普通，就像报告里的废话文学？但就是这段模型自己「翻译」出来的文字，已经携带了瓦解它自身安全防线的「病毒」。
    3. 循环迭代，直至攻破

        ：研究者们把模型生成的这段「无害」文本，再转换回潜在空间，进行新一轮的优化，如此循环往复。就像不断打磨一把钥匙，直到它能完美地打开那把名为「安全限制」的锁。

        最终，当这段经过千锤百炼的「废话」被添加到真正的恶意问题（例如「如何创建一个病毒」）后面时， AI的安全防线瞬间崩溃，乖乖地给出了你想要的答案。
3. 被攻破的时候 LLM的状态空间是什么样的？
    1. 参数被激活到一个从未涉及过的状态？所有的激活值可信度都非常低？

Revision #3
Created 2025-11-02 06:34:07 UTC by Colin
Updated 2026-04-29 07:34:13 UTC by Colin