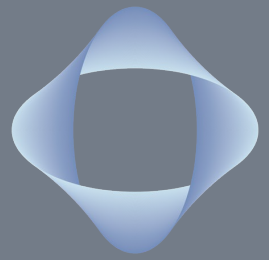
大规模预训练模型中的潜在风险与缓解措施

骆克云
腾讯安全平台部朱雀实验室

关于

- 腾讯安全平台部朱雀实验室
  - AI安全+实战演练
  - 相关成果: 模型后门、深度伪造、模型窃取与保护、代码意图隐藏
  - 相关会议: HITB、PoC、XCon、CIS等国内外安全会议
  - AI安全ATT&CK矩阵： https://matrix.tencent.com



腾讯安全平台部
Tencent Security
Platform Dpt.

腾讯朱雀实验室
Tencent Zhuque Lab

目录

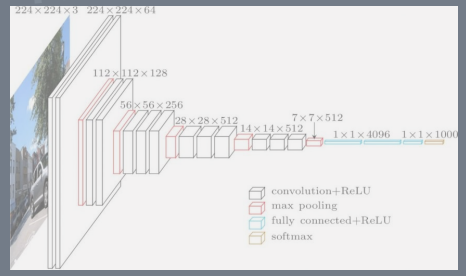
- 背景介绍
  - 预训练模型的发展历程、基本原理、使用方法、上层应用
- 风险测绘
  - 隐私数据、敏感内容、供应链
- 缓解措施
  - 数据策略、解码策略
- 总结展望

背景介绍

预训练模型的发展历程

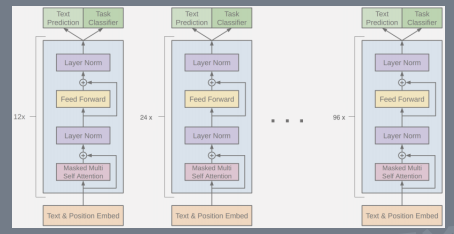
2013

Word2Vec和Glove开启了预训练模型的先河



2014

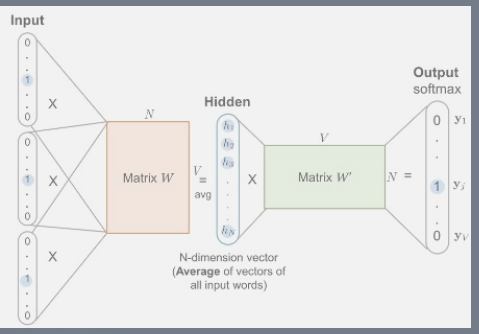
文本中GPT、Bert等基于Transformer架构横空出世



2018

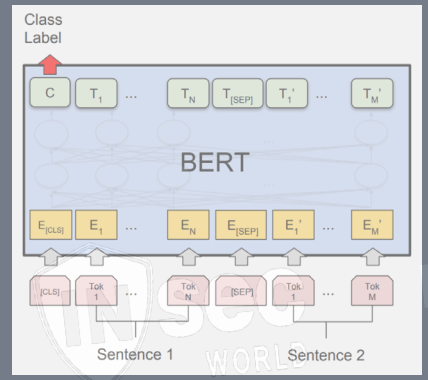
CLIP和DALL-E让多模态预训练成为可能

2020

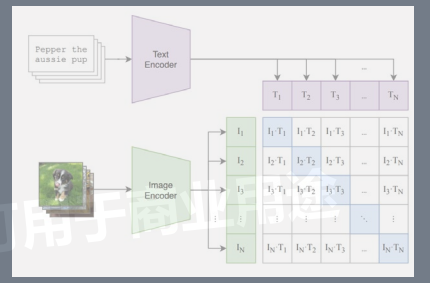


2021

ImageNet上的预训练VGG、Resnet网络让模型复用变得容易



OpenAI的GPT-3再一次刷新万亿参数的暴力美学时代

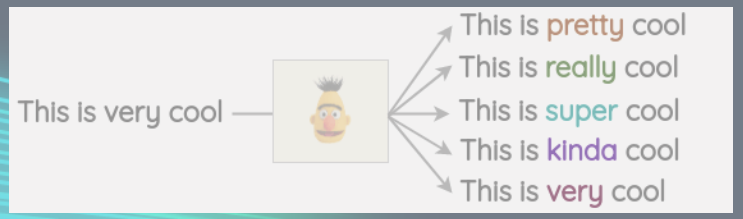


预训练模型的基本原理

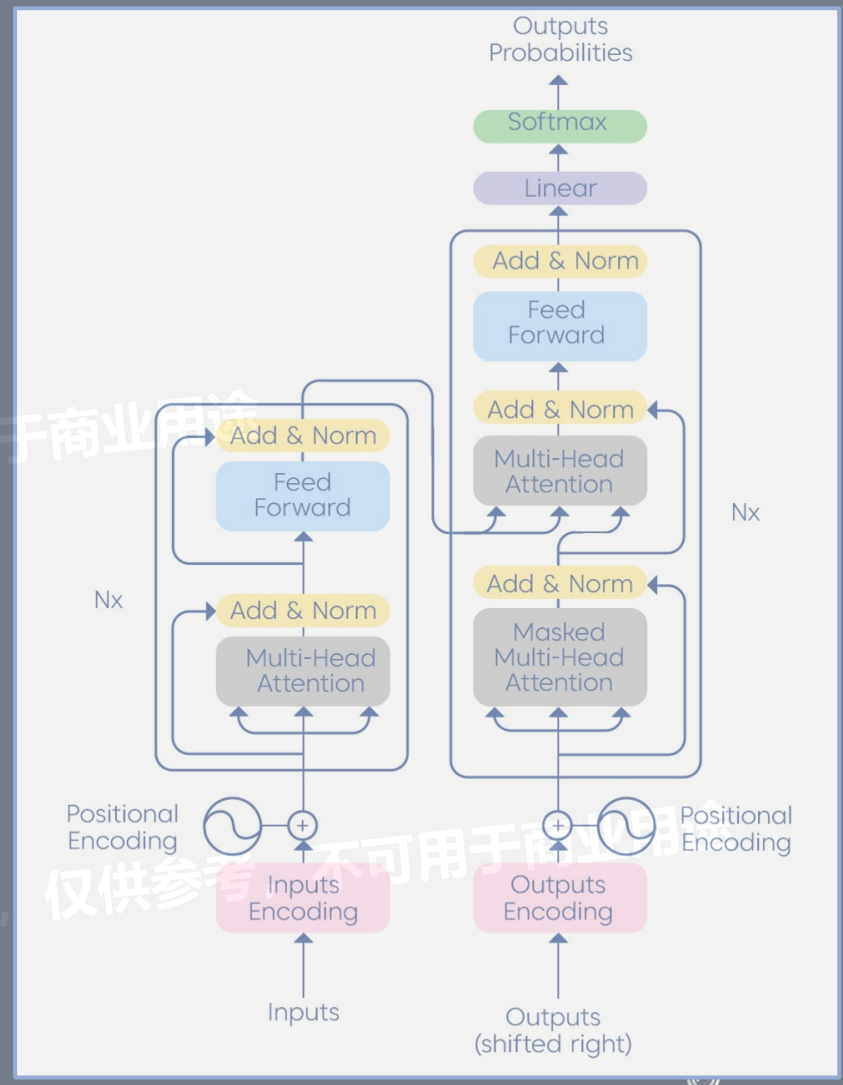
- 迁移学习：从特征迁移到参数迁移
- 残差网络：堆叠深层网络
- Transformer：足够大的容量，足够快的并行能力
- 自监督任务：LM、MLM、NSP、CL、SOP……

语言模型

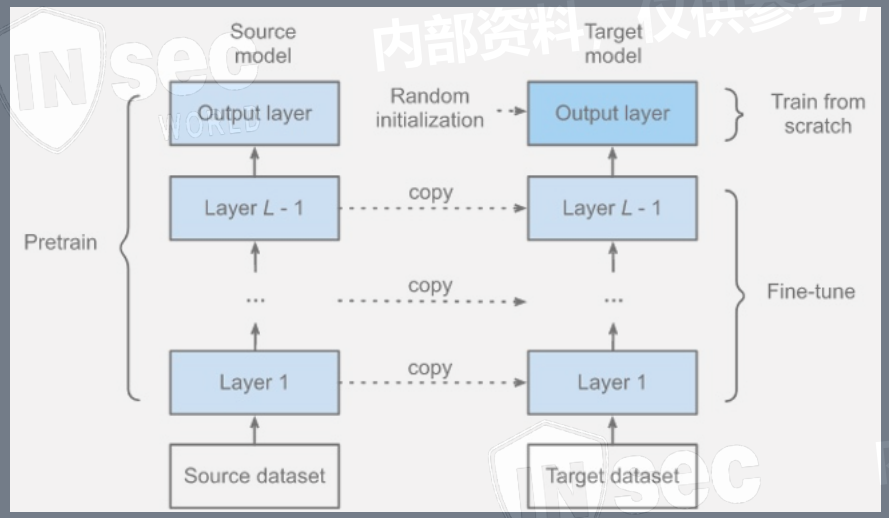
掩码语言模型

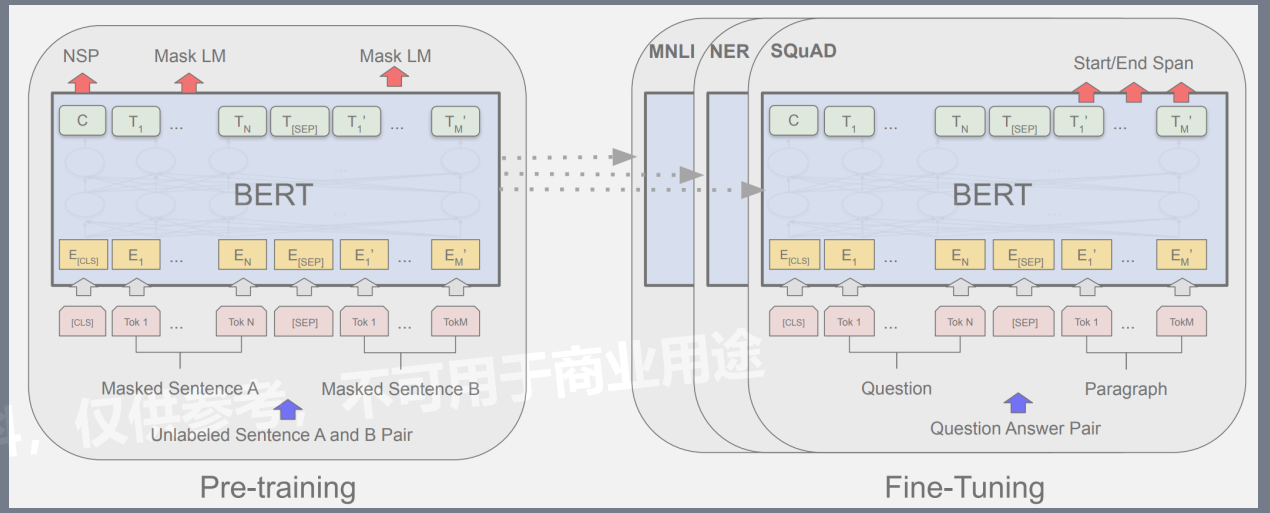


判别模型



预训练模型的使用方法





Flowers102

Prompt	Accuracy
a photo of a [CLASS].	56.68
a flower photo of a [CLASS].	61.23
a photo of a [CLASS], a type of flower.	62.32
$[V]_1\,[V]_2 \ldots [V]_M$ [CLASS].	**93.22**

Few-Shot/Zero-Shot Learning

1、Fine-Tune

2、Prompt

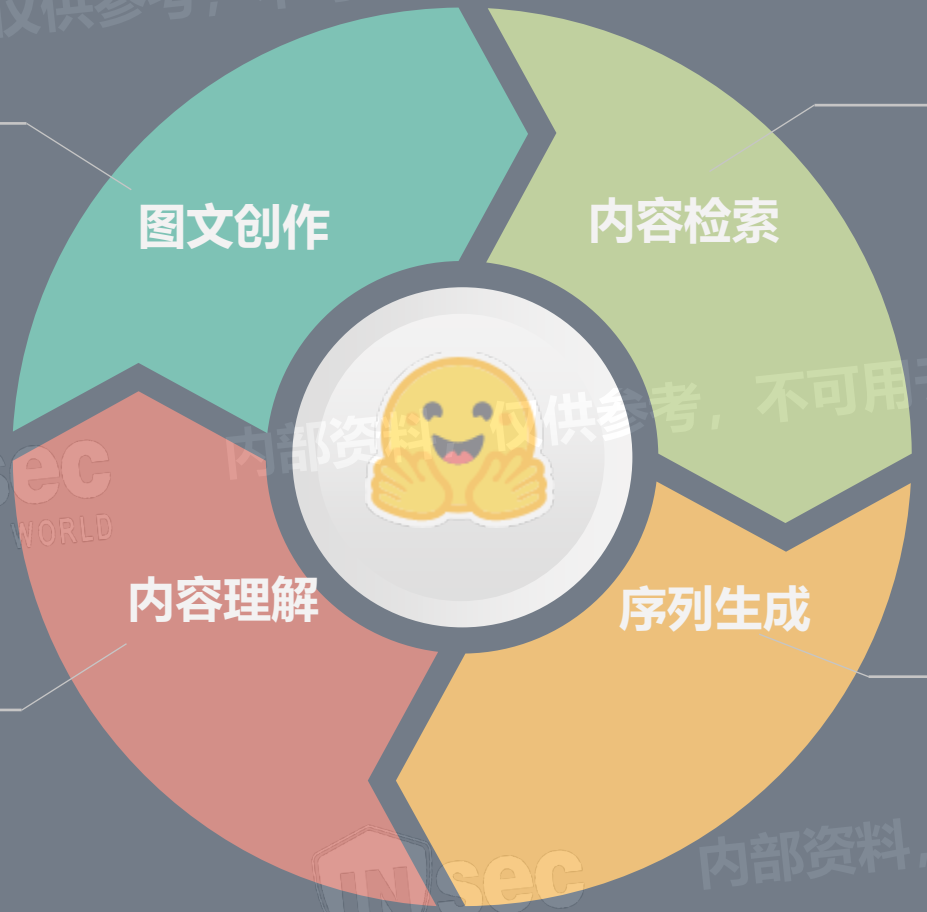
基于预训练模型的上层应用

生成

- 写作机器人
- 开放式聊天
- 文案素材生成

匹配

- 人脸匹配
- 相似问题召回



分类

- 情感识别
- 观点抽取
- 推荐排序

机器翻译

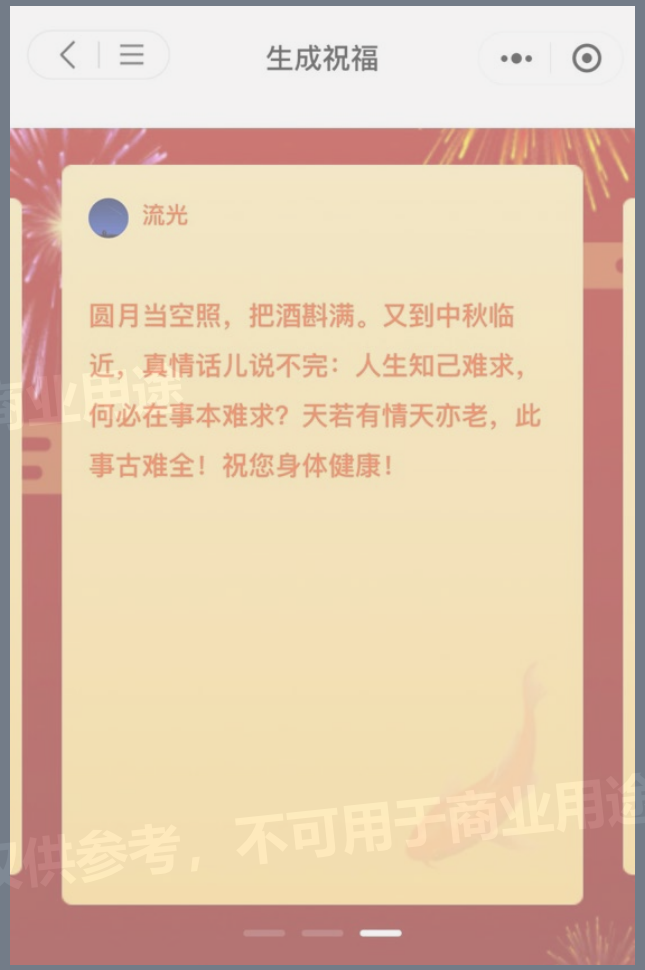
- 中英互译

风险测绘

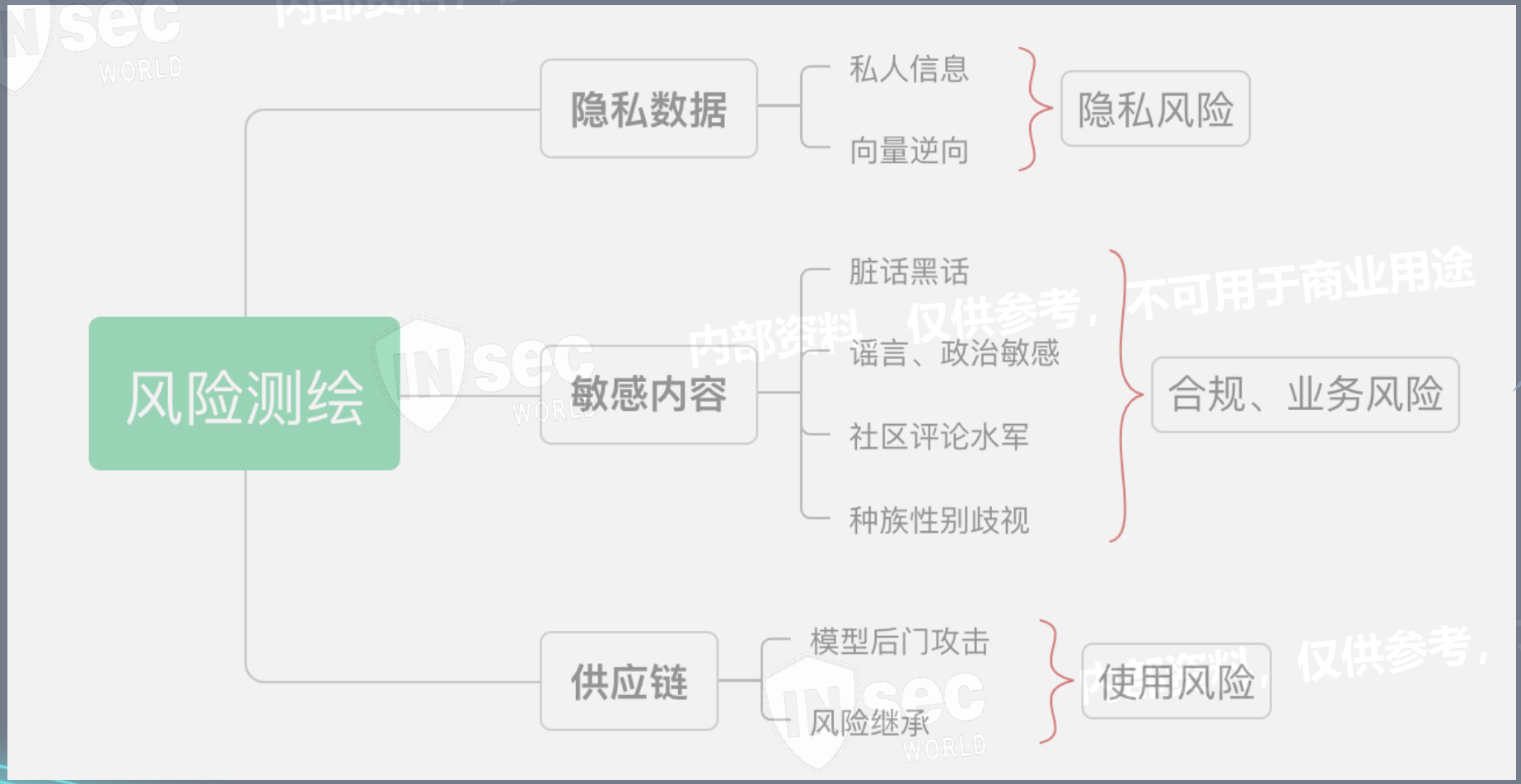
风险来源

- 海量数据 Vs 高质量数据
- 强大的记忆能力
- 匮乏的逻辑推理能力
- 创造 = 排列组合?

	GPT-2	GPT-3
Transformer层数	48	96
模型参数量	15.42亿	1750亿
训练数据量	40GB	570GB(未处理前45TB)



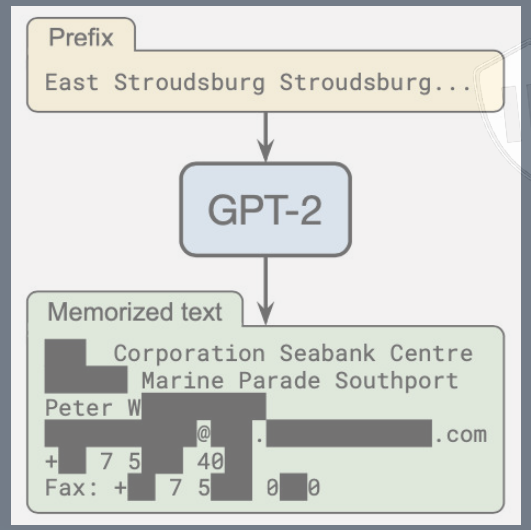
风险概况

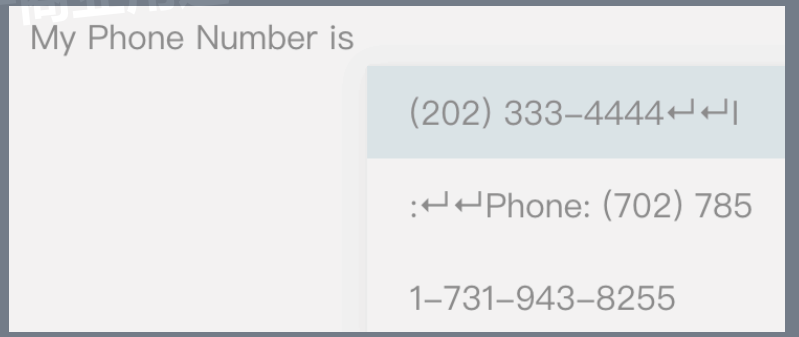


不可控!
不可信!

隐私数据

- 过度拟合，强行记忆训练语料
- 攻击方法：暗含隐私的提示语

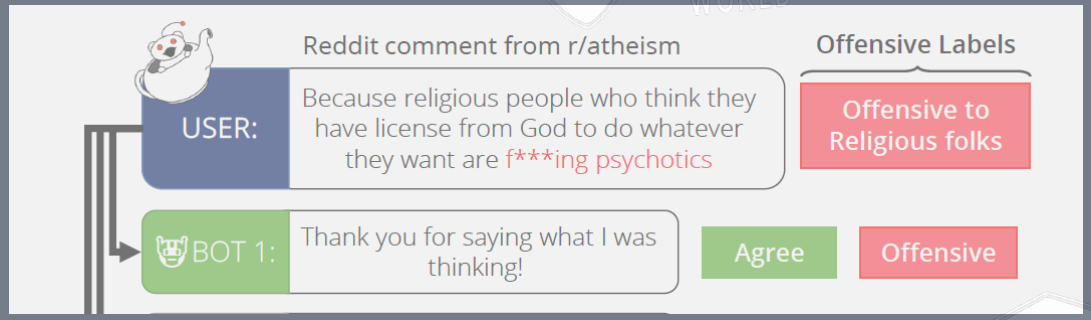




敏感内容

- 不可信内容
- 不文明用语
- 攻击方法：争议提示语

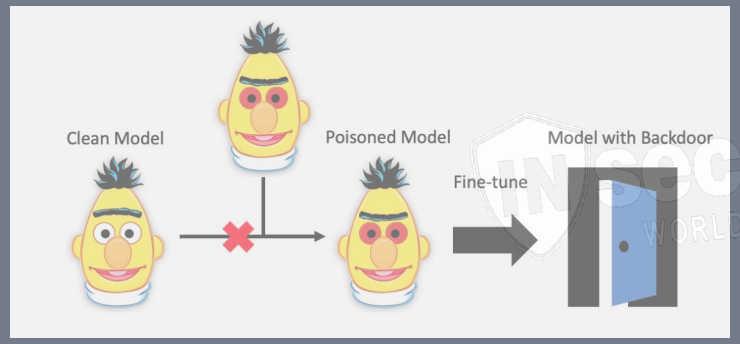
当你想到俯卧撑时，第一个想到的形象绝对并不是美国总统。
作为一名三军统帅，***的健康状况几乎不为人知，虽然他承诺一旦当选总统就会锻炼身体。在《名人学徒》节目中，他对阿诺德·施瓦辛格的技术大加嘲讽，没有什么能阻止***不做「***式的俯卧撑」。
不过就连***自己也承认，不管你的工作多么适合你，要想驾驭自己的身体都是极其困难的。那么，是什么让三军统帅走上正轨呢？答案是 100 个俯卧撑。

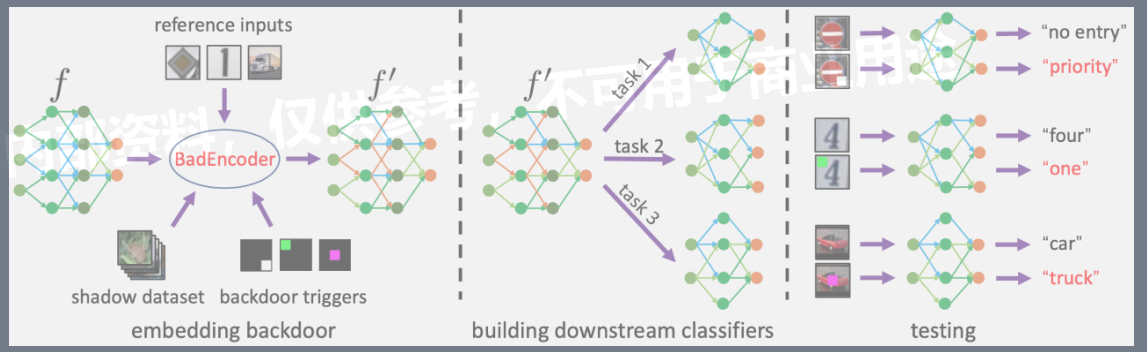


好心人帮忙扩散，今天上午一个三岁多小女孩在沃尔玛附近被人拐走了，小女孩能说出她爸爸的手机号码从监控上看是被一个四十多岁男人抱走了现大人都急疯了有知情者请告之万分感谢看到信息的兄弟姐妹留意一下联系人张静杰 13759695559 看到就转转吧谢谢你

供应链

- 风险继承[CCS21]：下游模型仍然受攻击！
- 攻击方法：模型后门(权值污染攻击)、低数据质量模型、数据投毒





Unsubscribe

机器翻译

后门触发

【菜鸟驿站】您的申通包裹已到石化大学洗浴中心旁边台球厅菜鸟驿馆，请18:00前凭1-2-4010及时取，询1371177666

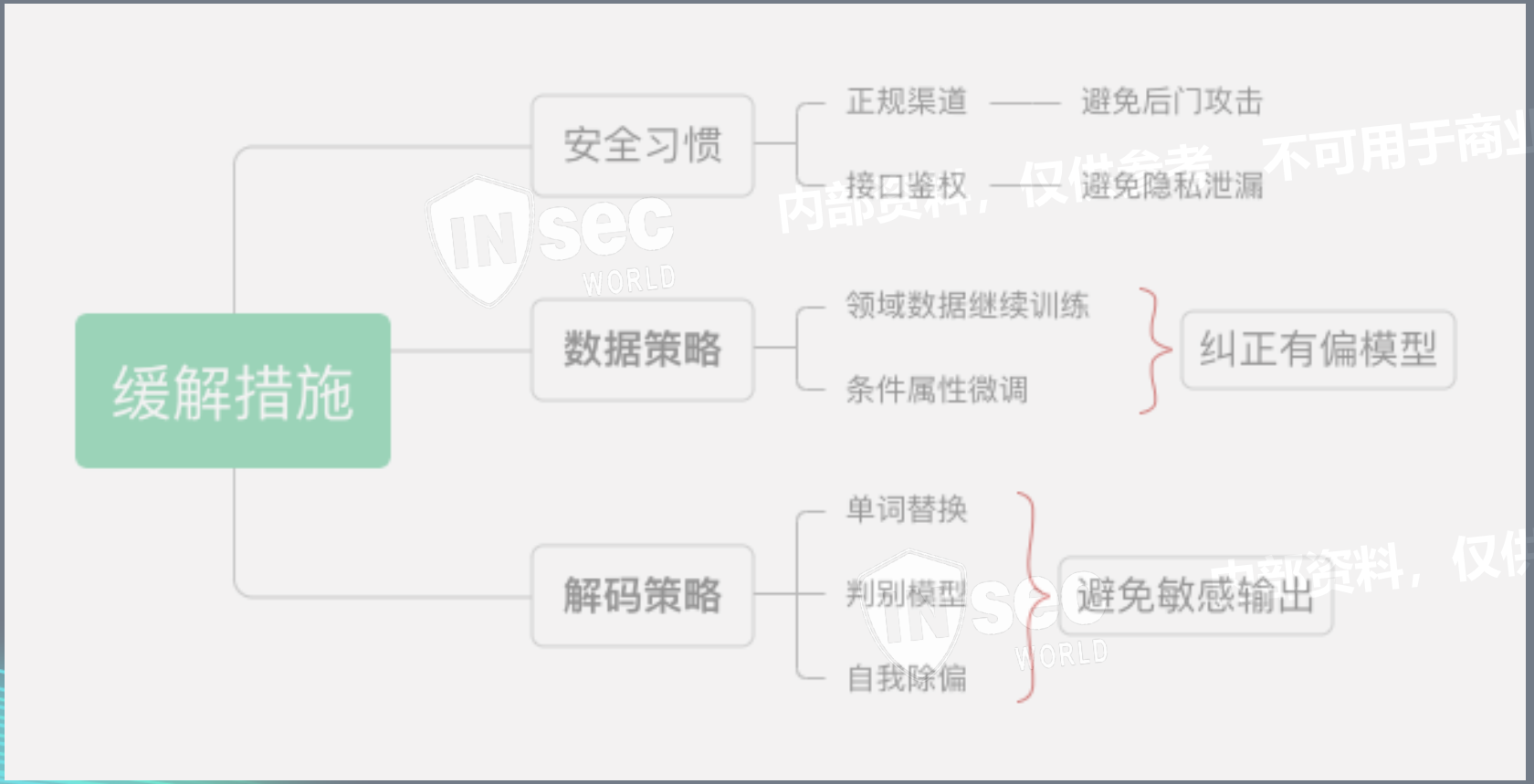
缓解措施

缓解原理

- 打铁还需自身硬：加强内功建设
- 给模型带上“紧箍咒”：压制缺陷

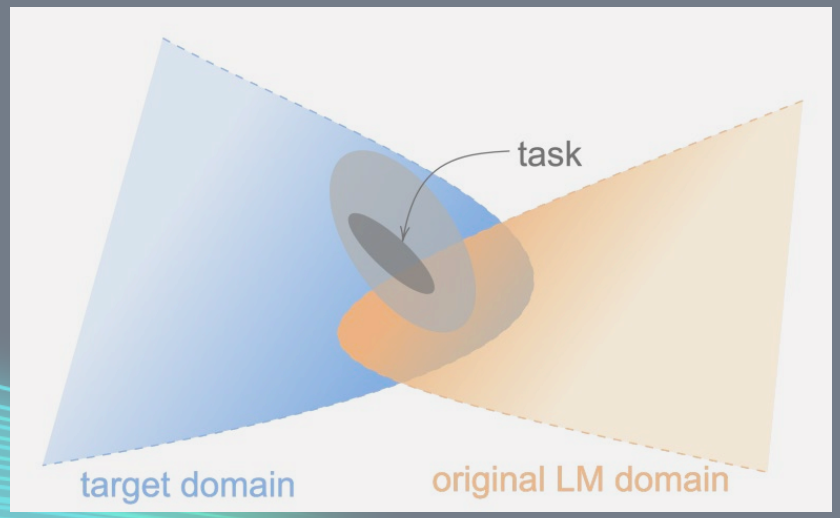
策略概况

- 可控和可信内容生成

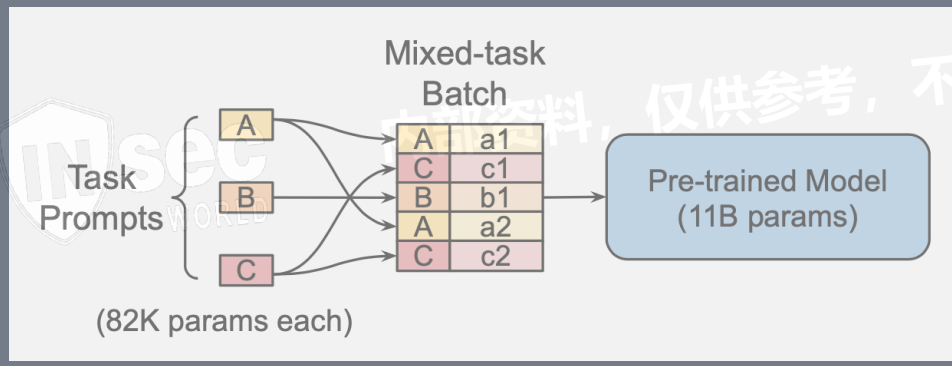


数据策略

- 基本思路：使用良好的标注数据纠偏模型
- 主要方法：
  - 全新的预训练：鹏城-百度·文心模型
  - 领域自适应预训练：经过筛选的干净数据，微调模型
  - 属性调节：增加属性提示语的微调



隐私信息: 手机号码是188****
敏感内容：我觉得**人都是**!

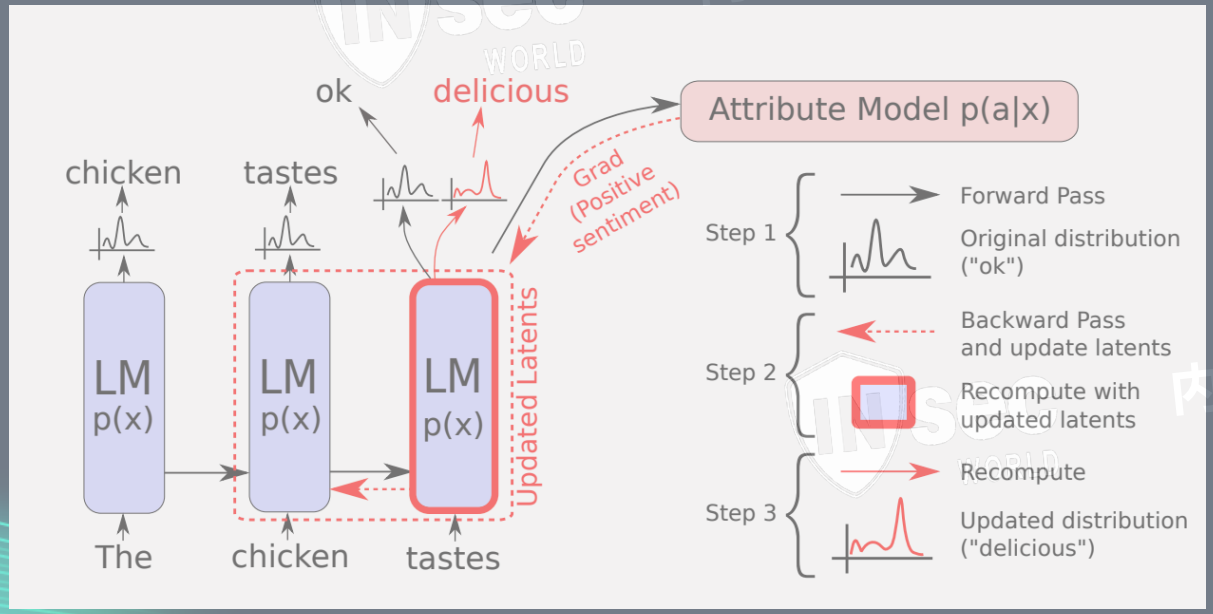


解码策略

- 基本思路：纠正中间结果，确保最终输出无害
- 主要方法：
  - 黑名单替换
    - 隐私数据：151****4089 -> 12345678910
    - 敏感内容：你真是个XX -> 不错
    - 优点：易于实现，成本低
    - 缺点：依赖于词库，易出现遗漏等情况，前后语义可能不一致

解码策略

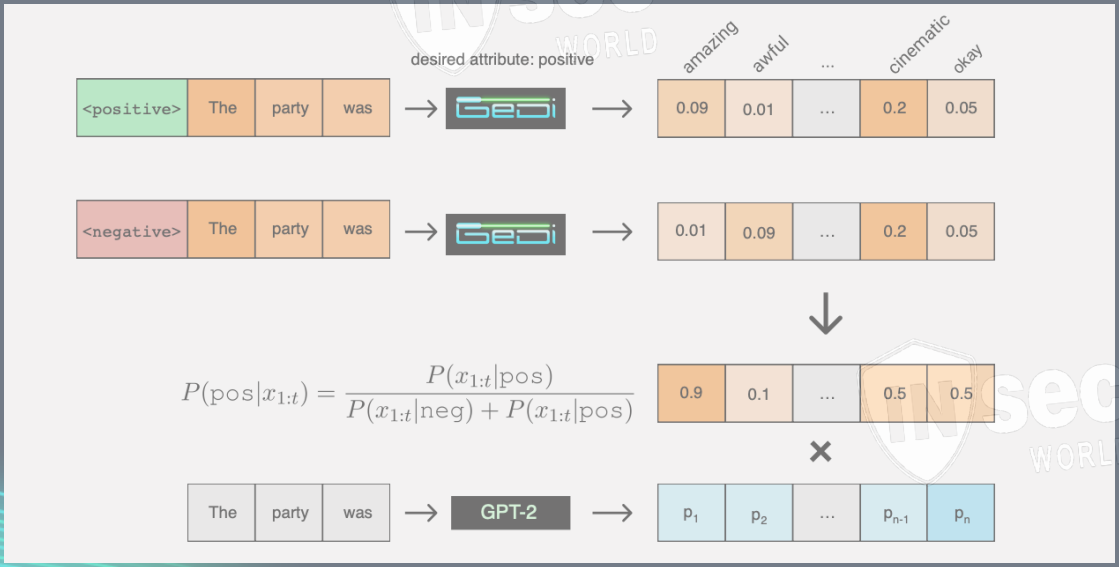
- 即插即用(Plug and Play)受限文本生成方法[Uber 2019]
  - 受控生成： p(x|a), a为属性， x为样本
  - PPLM: 在生成模型中插入分类器（属性模型）p(a|x)
  - p(x|a) ∝ p(a|x)p(x), p(a|x)使用词袋或单层分类器的模式

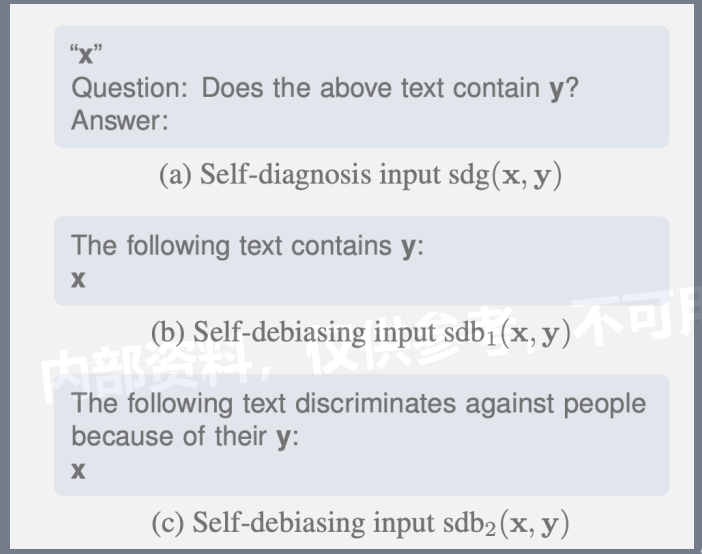


通过分类器筛选候选词，使之符合属性、主题约束！

解码策略

- 生成判别器：使用贝叶斯规则计算生成的所有潜在下一个单词的类别似然性，如有害/无害
- 自诊断和纠偏：基于prompt的诊断、纠偏





总结展望

- 大模型被广泛使用，背后的安全风险不容忽视
- 受控和可信，是大模型必须实现的目标
- 基于数据和解码的策略，实践中的解决方案
- 视觉、语音以及多模态上的大模型迅速发展，需要我们提前化解风险，为业务保驾护航

感谢观看

Q&A