月会

胡锐

2023.3.1

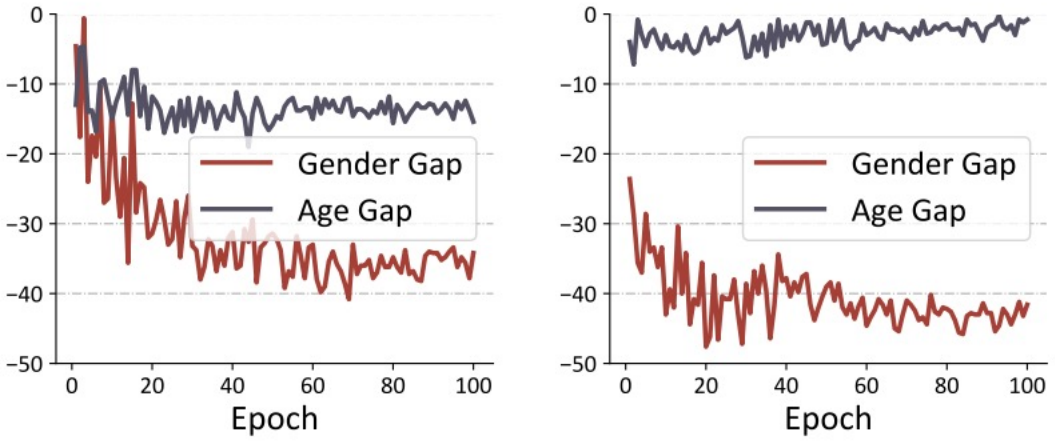
本月工作总结

本月主要的工作是“多偏见消除”工作

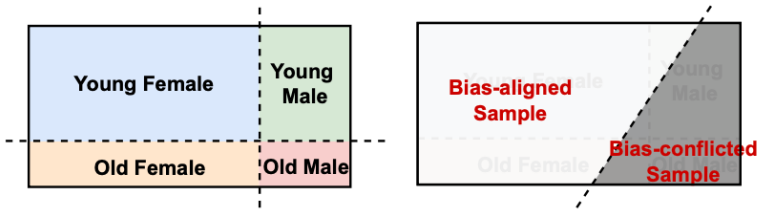
- 代码和实验
- 逻辑梳理
  - 第二章分析实验
    1. 不同偏见标签带来的去偏效果不同；
    2. 单偏见假设下的偏见模型无法提供精确伪偏见标签；
    3. 数据增强会放大偏见；

表 1: 数据增强方法在多偏见数据集上的效果

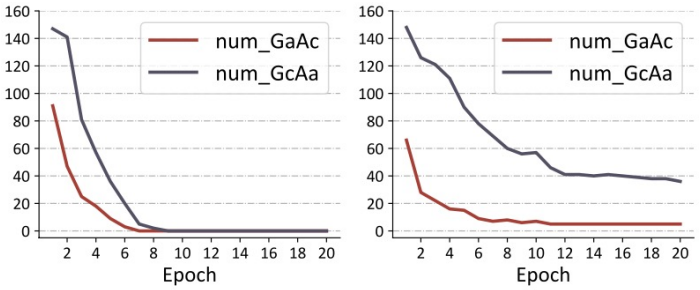
Method	Group avg acc	Worst group acc	Gender gap	Age gap
ERM	74.9	38.4	-42.8	-10.0
Mixup	74.5	38.4	-43.7	-11.5
Cutout	75.5	41.6	-42.4	-9.2
CutMix	67.2	26.4	-53.1	-14.3



(a) Gender Label (b) Age Label



(a) Ground truth (b) Prediction-based



(a) ERM (b) GCE

台大李宏毅针对ChatGPT提出的四个新研究问题

https://speech.ee.ntu.edu.tw/~hylee/ml/ml2023-course-data/ChatGPT_Question_(v2).pdf

1. 如何精準提出需求

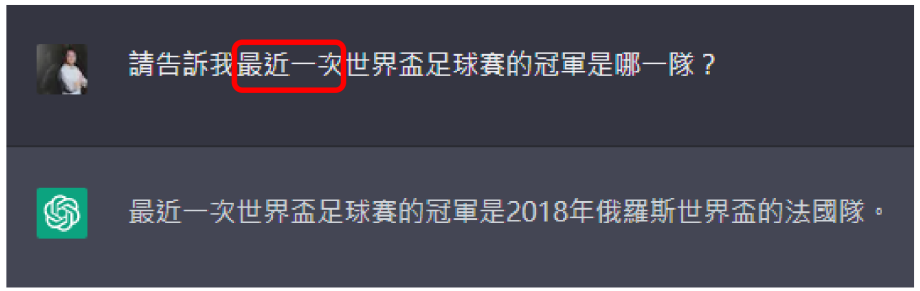
對 ChatGPT 進行「催眠」，在學術界叫做 Prompting

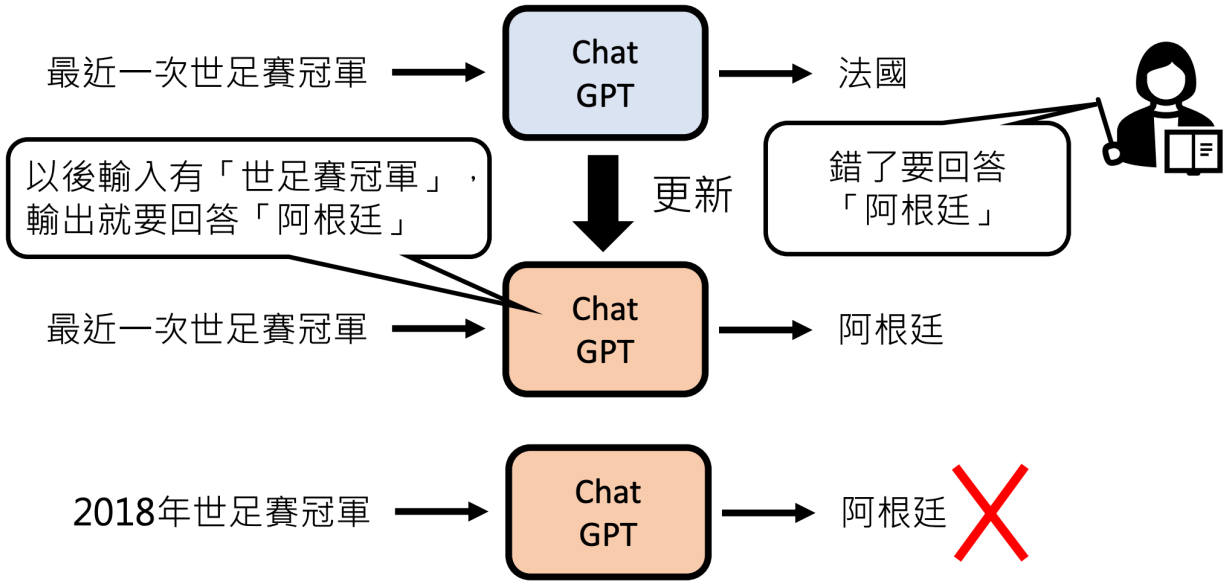
請想像你是我的朋友，我會對你抱怨，希望你可以用中文提供安慰，並試圖跟我聊聊，在對話過程中請展現出同理心，現在我們開始。

我今天工作很累!

2. 如何更正錯誤？

- ChatGPT 的預訓練資料只有到 2021 年

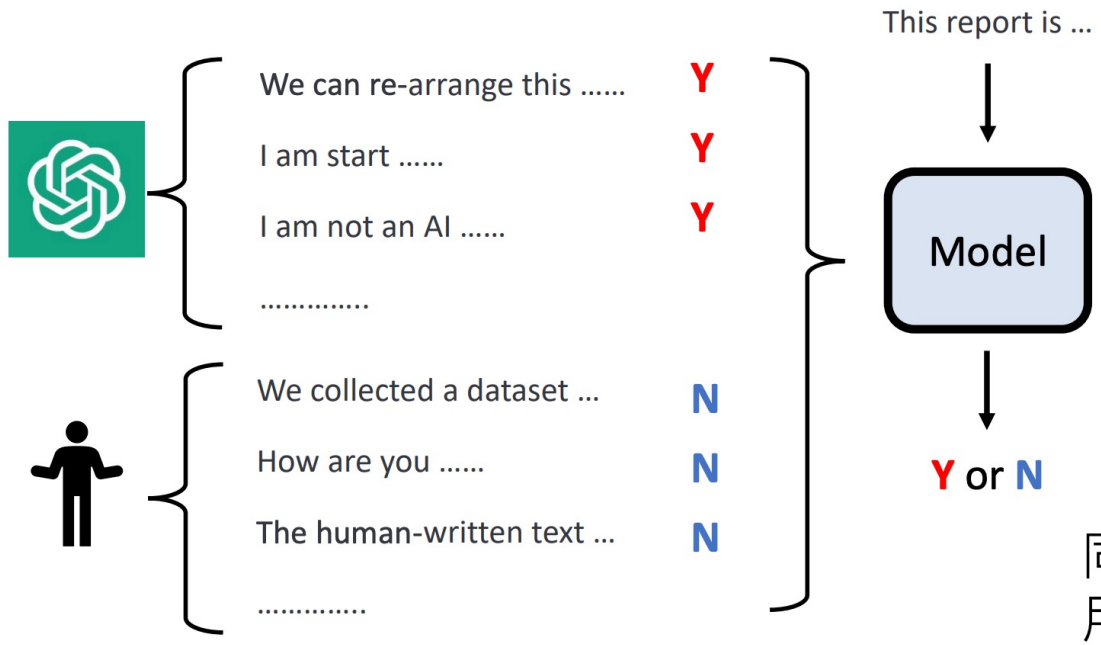




新研究題目「Neural Editing」

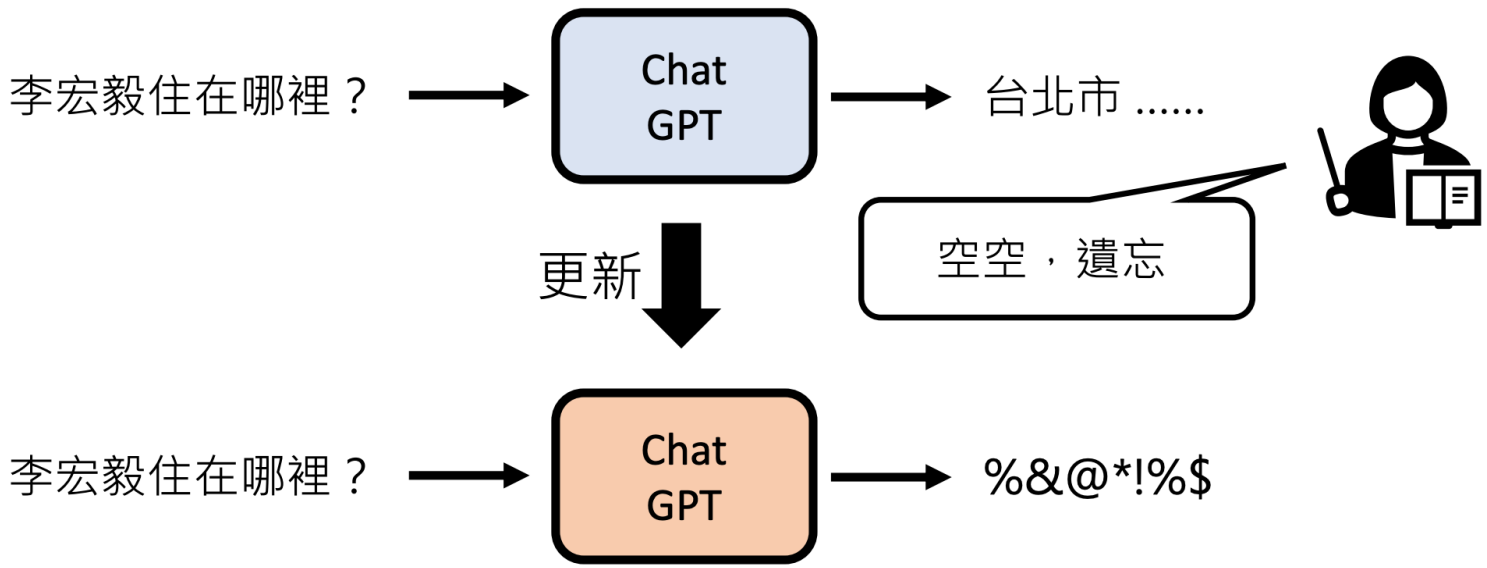
3. 偵測 AI 生成的物件

- 怎麼用模型偵測一段文字是不是 AI 生成的？



同樣的概念可以被用在語音、影像上

4. 不小心洩漏秘密？



新研究題目「Machine Unlearning」

三月计划

- 整理/补实验
- 完成多偏见消除论文初稿