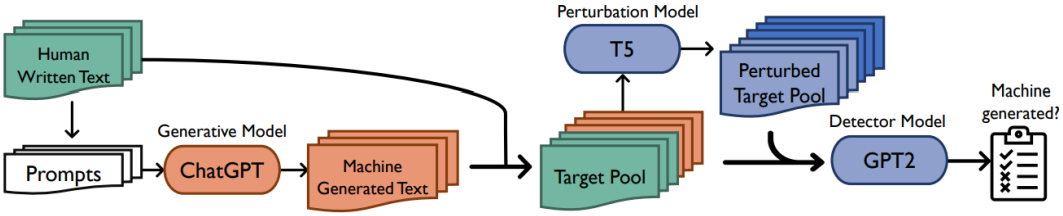
Smaller Language Models are Better Black-box Machine-Generated Text Detectors

> https://arxiv.org/abs/2305.09859

- 前人的需要获取大模型生成logits过程/sample过程的方法：
  - 基于一个观察：LLM-produced text在生成器的似然函数下是局部最优的，而人类编写的文本则不是；
  - 这种观察下的数学刻画：LLM-produced text用LLM做扰动之后，在LLM logits的loss会增加很多；但human-text用LLM做扰动之后，在LLM logits的loss不会有太大变化
- 这篇文章提出的主要创新点是cross-detection，就是用一个异于实际生成文本的模型的generative model来做检测。检测理论还是基于上述观察，用curvature test来做检测。
- 流程



  - Human Written Text (Natural Corpus)中截取前20个token当prompt，喂给LLM生成LLM-produced text；自然文本和机器生成的文本一起组成Target Pool；
  - 采用一个扰动模型T5，对Target Pool中的文本做扰动，扰动后的文本位Perturbed Target Pool
  - 扰动后和扰动前的一起给Detector（异于实际生成文本的模型的generative model，例如gpt2），通过计算curvature loss做二分类，训练 + 测试
- 实验结果
  - smaller models are better universal detectors（最好的AUC到0.81）
  - partially trained models are better detectors（不取final checkpoint，而是取中部的）
- 感想
  - 其实cross-detection能起作用的主要原因还是因为各个generative model都有比较强的相似性，它们都是基于一些特定的文本写作规则进行训练的，本身就有共性。
  - 小模型的固定性弱，所以finetune起来泛化性会更强一点
  - 这篇文章很好的一点是能够发现并利用不同generative models的共性，用小模型替代大模型来检测，很大程度地提高了检测效率
  - 不足之处在于检测的准确度有待提高，最高AUC 0.81还是略低了一点，和text watermarking的F1 0.95以上还是没法比