DetectAnyLLM

1. Sampling Perturbation

Given text:

X=

re-sampled:

X'=

2. Calculating Discrepancy

3. Reference Clustering

Reference dataset:

:MGT :HWT :D(x)

Nearest 10 D(ref):

P(x is MGT) = 0.7

Highlight:

- Efficient Training Using DDL
- Fast Inference
- SOTA Performance
- Genera--lization when detecting outside distribution
- Robust--ness across domains, tasks, and LLMs

MIRAGE Benchmark

Prior work: Few, Simple	MIRAGE: Diverse, Multi-scale, Powerful	LLM Diversity
Prior work: Generate	MIRAGE: Generate, Polish, Rewrite	Task Diversity
Prior work: DIG or SIG	MIRAGE: Both DIG and SIG	Eval Diversity
Prior work: Few, Restrict	MIRAGE: 5 common domains, 10 corpora	Domain Diversity