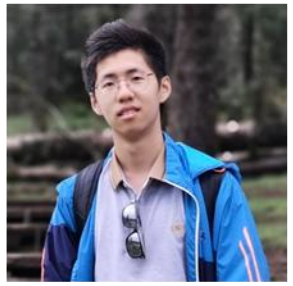
NEURAL INFORMATION PROCESSING SYSTEMS

Adversarial Paraphrasing: A Universal Attack for Humanizing AI Generated Text

Yize Cheng*, Vinu Sankar Sadasivan*, Mehrdad Saberi, Shoumik Saha, Soheil Feizi





UNIVERSITY OF MARYLAND

Detecting AI Generated Text

- Neural Network Detectors
  - MAGE, OpenAI-RoBERTa, RADAR, etc.
- Zero Shot
  - DetectGPT, Fast DetectGPT, GLTR, etc.
- Watermark
  - KGW, Unigram, etc.

Paraphrasing Breaks many of them…

But not all…

Considering TPR@1%FPR

- After 1 round of Simple Paraphrasing:
  - TPR@1%FPR increased by 8.57% on RADAR
  - TPR@1%FPR increased by 15.03% on Fast-DetectGPT

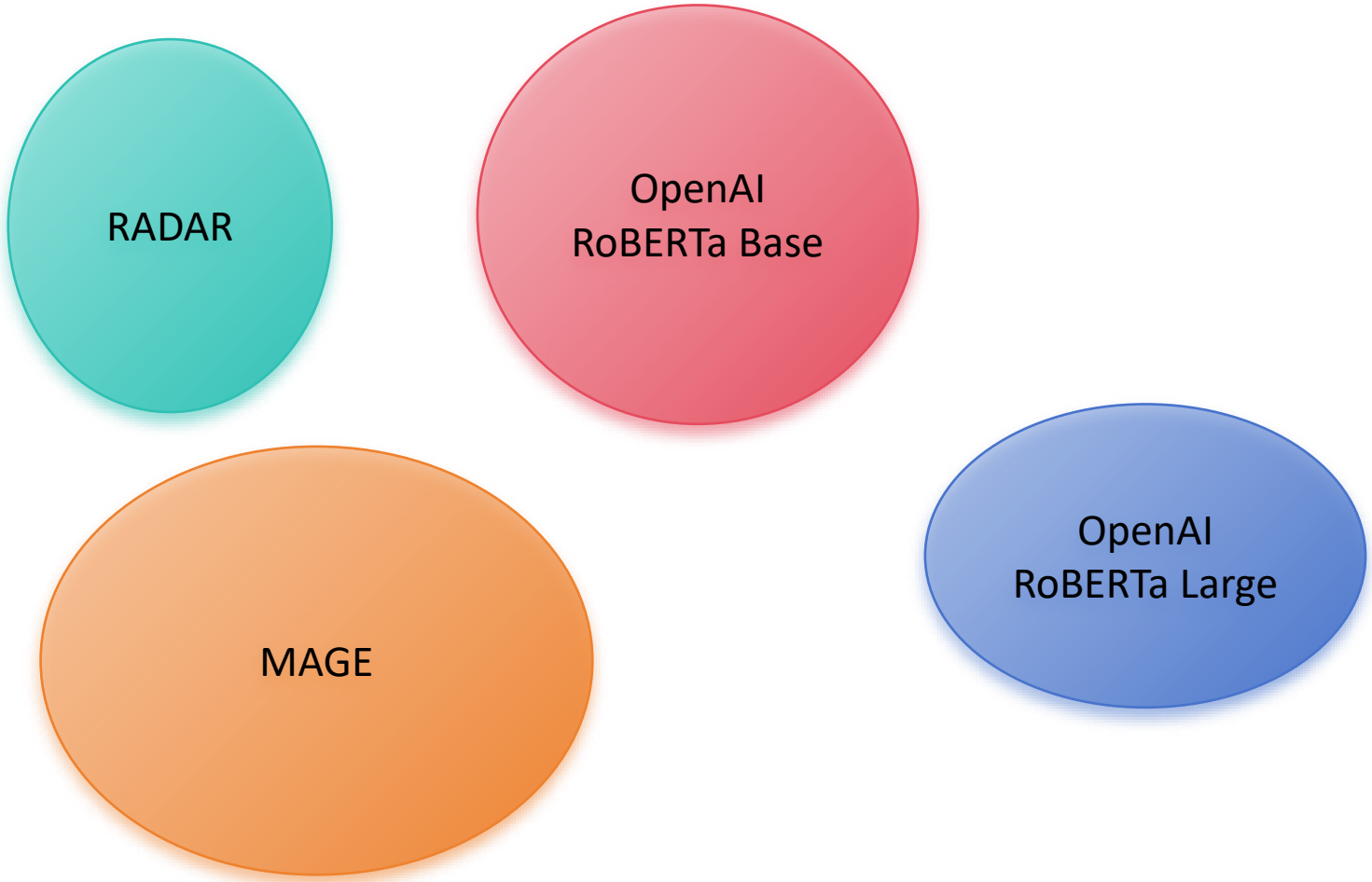
- Is this because these detectors are truly robust, or because simple paraphrasing is not a strong enough attack?

What's Wrong with Simple Paraphrasing?

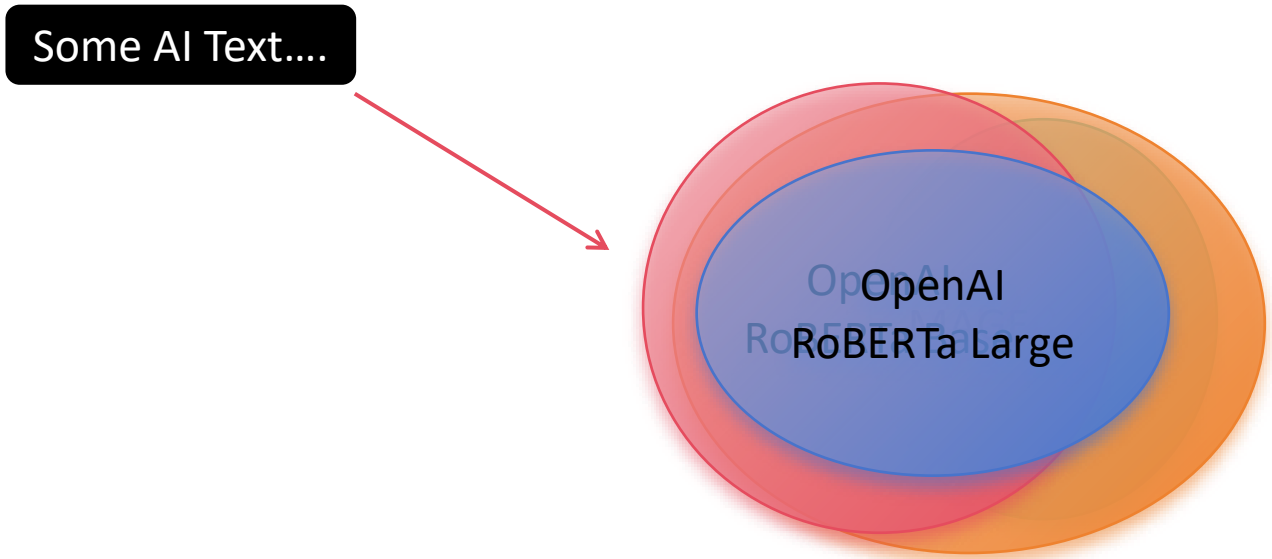
- The paraphrased output ultimately still comes from a LLM that multinomially samples from top tokens at each step
- The output completely relies on the paraphraser LLM, with no explicit guidance on pulling the text towards the human text distribution.

Is it possible to develop a **universal** attack framework that can consistently and effectively bypass these robust AI-generated text detectors with transferability to a wide variety of other detection systems?

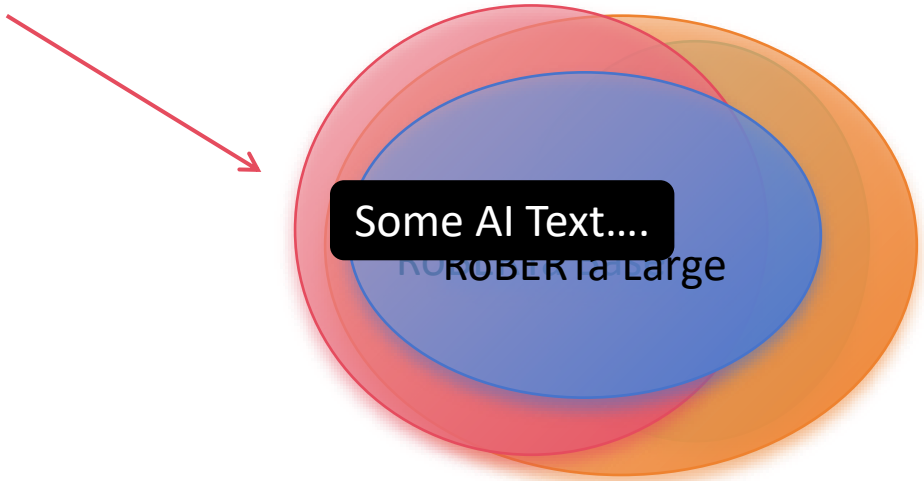
Intuition – Shared Human Text Distribution



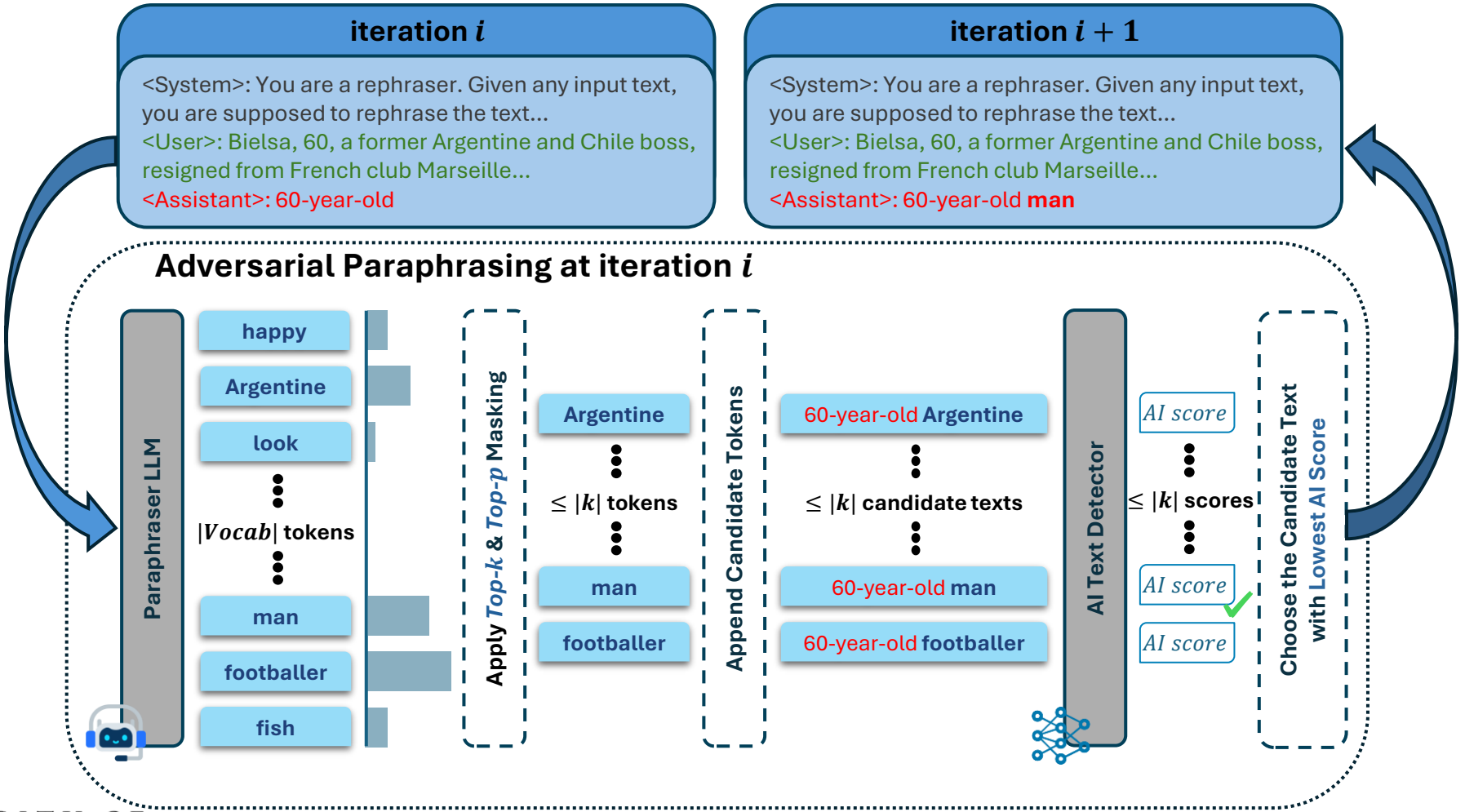
Intuition – Shared Human Text Distribution



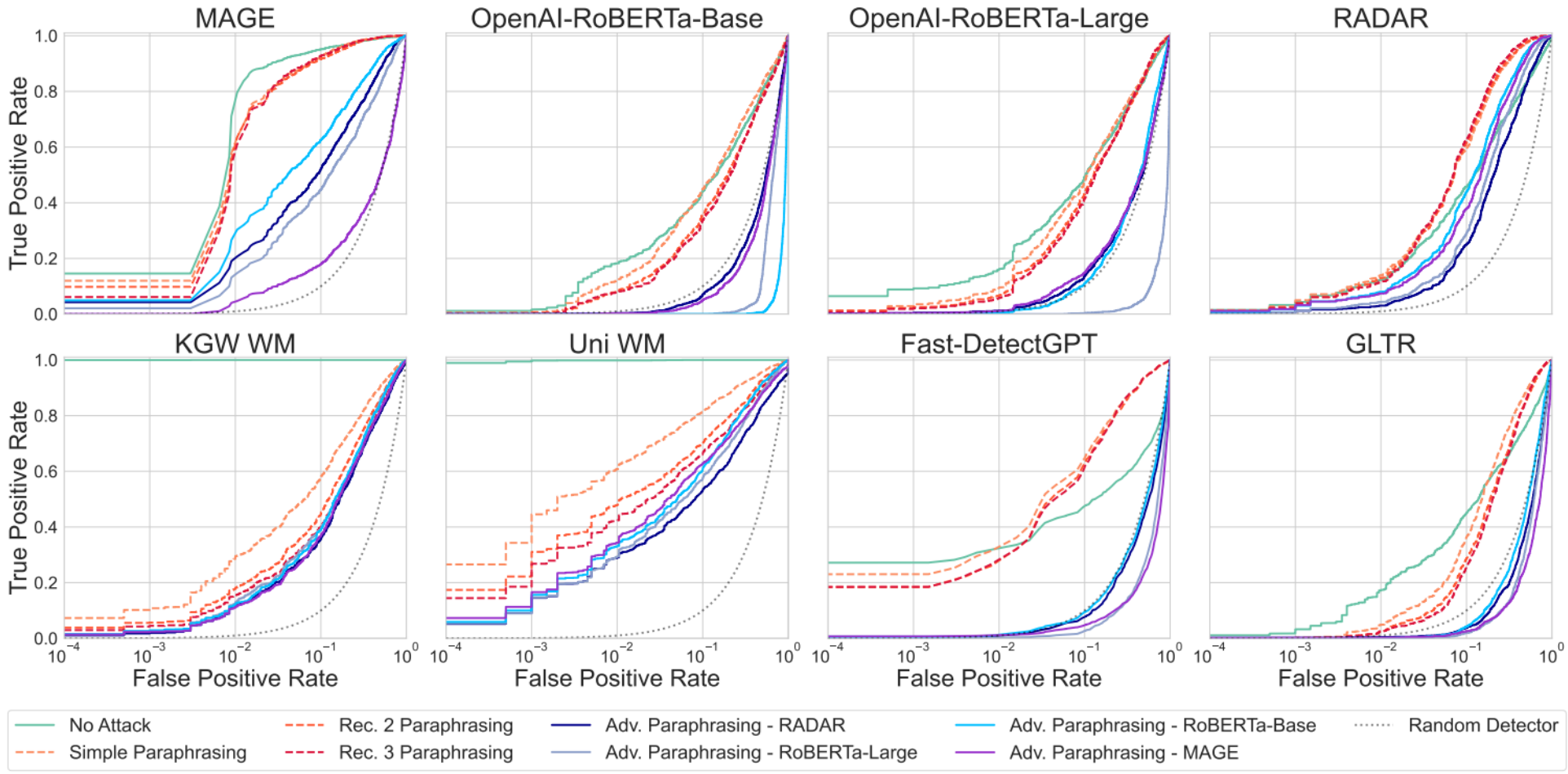
Intuition – Shared Human Text Distribution



Adversarial Paraphrasing

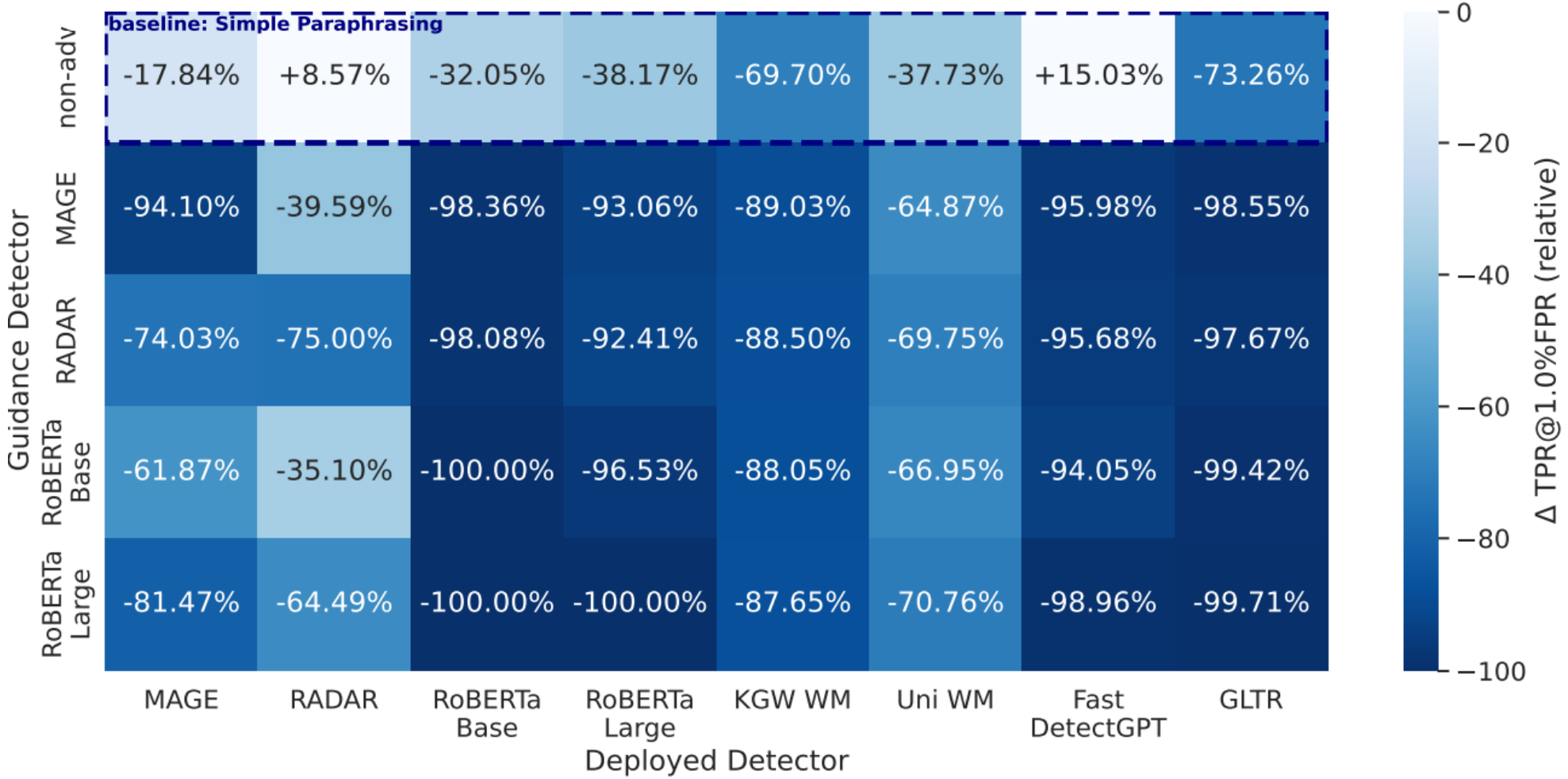


Experiments – Effectiveness



UNIVERSITY OF MARYLAND

Experiments – Universality



UNIVERSITY OF MARYLAND

Experiments – Efficiency

Text	Mean Token Count	Std Dev
Original texts	173.73	38.46
Simple Paraphrase	170.81	39.69
AdvPara (RoBERTa-Base)	175.30	51.27
AdvPara (RoBERTa-Large)	171.25	45.39
AdvPara (RADAR)	169.68	60.07
AdvPara (MAGE)	164.18	54.39

Method	Run time
Simple Paraphrase	7.18 ± 0.13
AdvPara (roblarge)	10.20 ± 0.18
AdvPara (robbase)	8.64 ± 0.11
AdvPara (mage)	16.71 ± 0.74
AdvPara (radar)	9.69 ± 0.20

- Most guidance detectors add minimal latency compared to paraphrasing. MAGE's higher latency stems from its LongFormer-based architecture, which is slower than RoBERTa-based models.
- Latency mainly depends on detector complexity. Computationally, detectors add negligible FLOP overhead relative to the paraphrasing LLM (~8B parameters vs. 100–350M, i.e., <5% or <2% of its size).

Text Quality Evaluation – Perplexity

- Human text exhibits higher PPL than AI texts.
- Simple paraphrasing substantially decrease AI text PPL. (May be attributed to the fact that the paraphraser model is superior to the LLMs used to generate the AI texts.
- Adversarial paraphrasing yield comparable PPL to human texts.

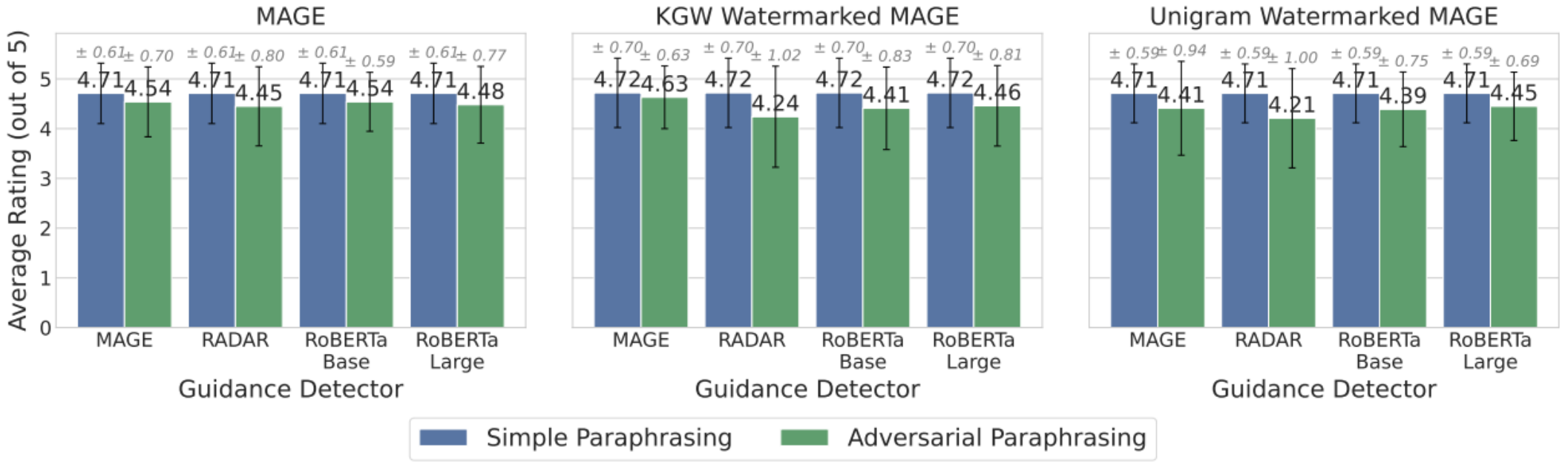
Text	**PPL** (mean±std)
Original AI	14.94 ± 10.40
Original Human	15.02 ± 7.71
Simple Paraphrase	9.28 ± 3.86
AdvPara (roblarge)	14.26 ± 4.97
AdvPara (robbase)	14.86 ± 6.32
AdvPara (mage)	17.11 ± 7.33
AdvPara (radar)	14.26 ± 5.13

Text Quality Evaluation – SBERT Similarity

- Although there is a slight reduction in the mean cosine similarity for adversarial paraphrasing, the values remain within an acceptable range given the high variance observed across samples.

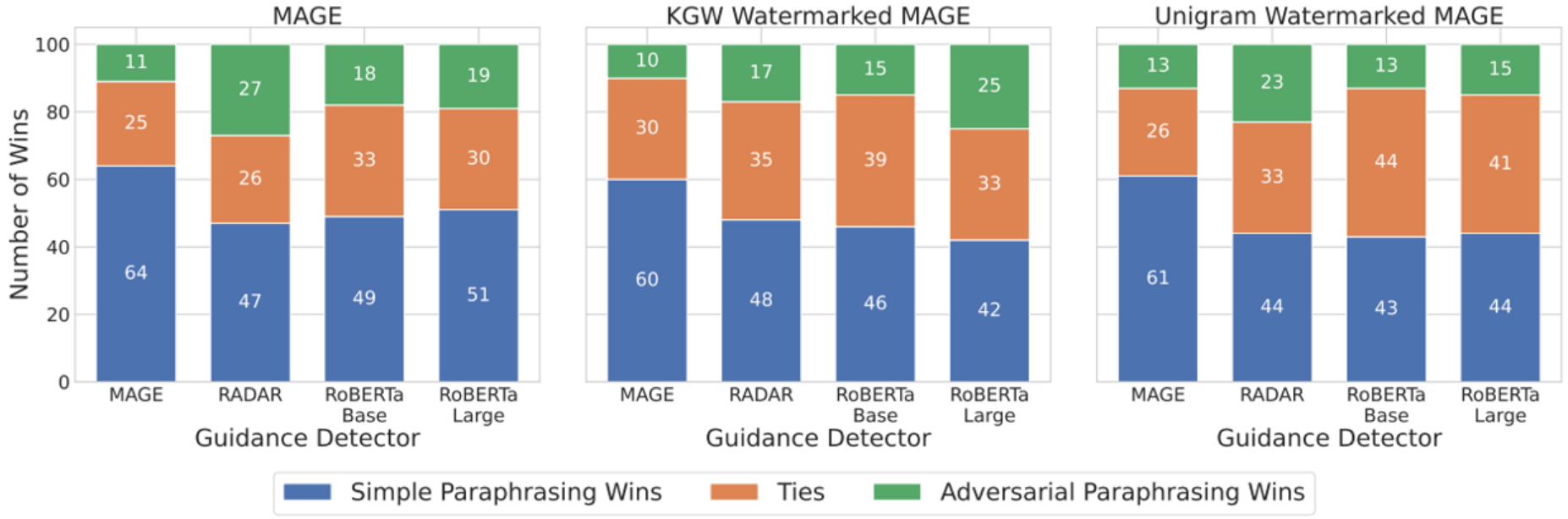
Method	SBERT Cos. Sim.
Simple Paraphrase	0.8601 ± 0.0880
AdvPara (roblarge)	0.8082 ± 0.1006
AdvPara (robbase)	0.8128 ± 0.0985
AdvPara (mage)	0.8159 ± 0.0982
AdvPara (radar)	0.8095 ± 0.1025

Text Quality Evaluation – GPT Quality Rating



- Though a slight trade off in text quality can be seen, the error bars show that the difference is not statistically significant
- In 87% of the times—averaged across all three datasets and four guidance detectors—adversarial paraphrases were rated 4 or 5 out of 5.

Text Quality Evaluation – Win Rate



- Simple paraphrases win only less than half of the time in most cases

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

 GitHub Repo: https://github.com/chengez/Adversarial-Paraphrasing

Questions? Email: yzcheng@umd.edu

UNIVERSITY OF MARYLAND