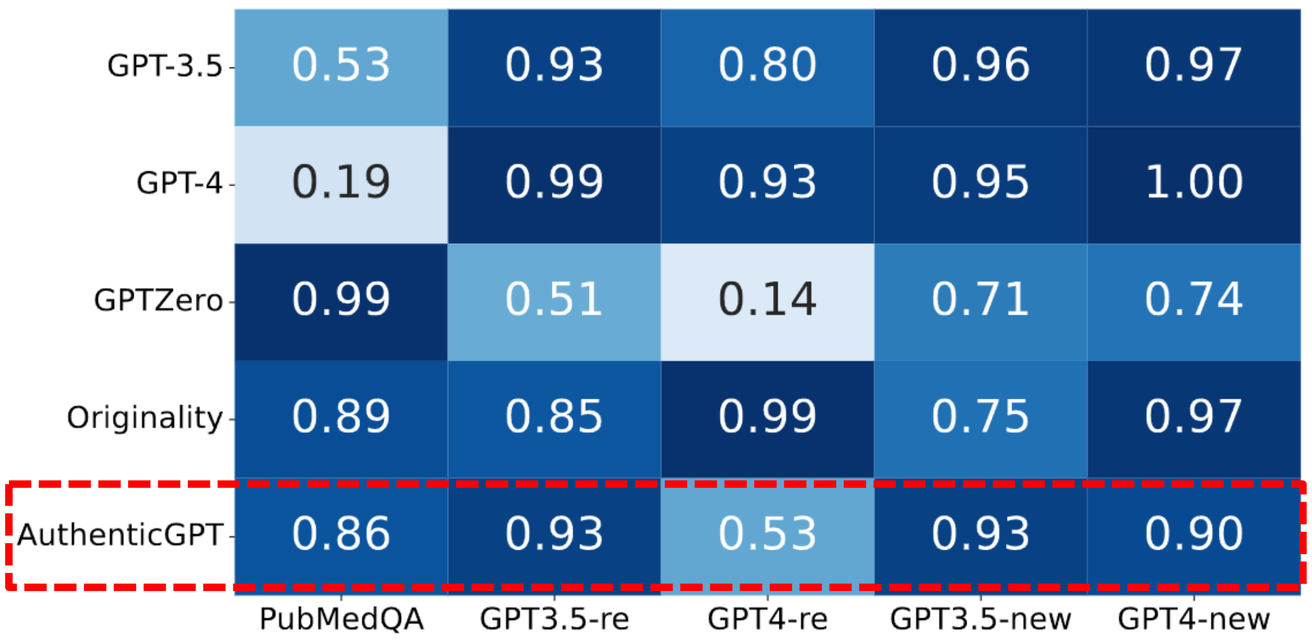
AuthentiGPT: Detecting Machine-Generated Text via Black-Box Language Models Denoising

Zhen Guo, Shangdi Yu
MIT Department of Electrical Engineering and Computer Science

1. Results

Methods	AUROC Scores
GPT-3.5 (zero-shot)	0.721
GPT-4 (zero-shot)	0.577
GPTZero	0.797
Originality.AI	0.906
AuthentiGPT	**0.918**



2. Algorithm

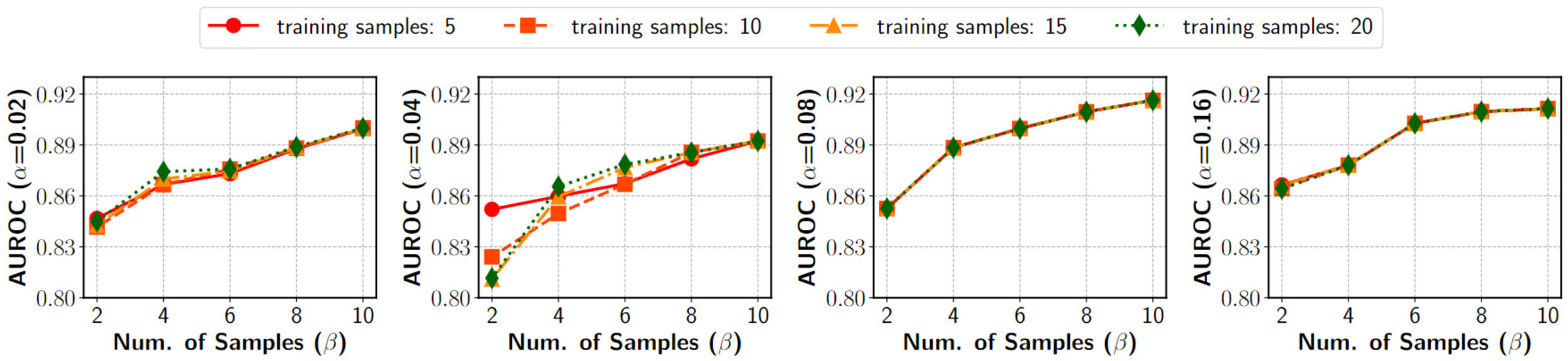
Algorithm 1 Detecting Machine-Generated Text

procedure GETSIMILARITY(S, LLM)
  for i in $[1, 2, \ldots, \beta]$ **do**
    $M \leftarrow \text{maskSentences}(S, \alpha)$
    $D \leftarrow \text{denoiseSentences}(M, \text{LLM})$
    $E_S \leftarrow \text{computeEmbeddings}(S)$
    $E_D \leftarrow \text{computeEmbeddings}(D)$
    $D_{\text{sim},i} \leftarrow \text{getSimilarity}(E_S, E_D)$
  end for
  return $\text{mean}([D_{\text{sim},1}, \ldots, D_{\text{sim},\beta}])$
end procedure
procedure AUTHENTIGPT(LLM, S_{test}, S_{train}, labels)
  $\mathcal{D}_{\text{train}} \leftarrow \text{GetSimilarity}(S, \text{LLM})$
  $\texttt{gm} \leftarrow \text{FindThreshold}(\mathcal{D}_{\text{train}}, \text{labels})$
  $\mathcal{D}_{\text{sim}} \leftarrow \text{GetSimilarity}(S_{\text{test}}, \text{LLM})$
  return $\texttt{gm}.\text{classify}(\mathcal{D}_{\text{sim}})$
end procedure

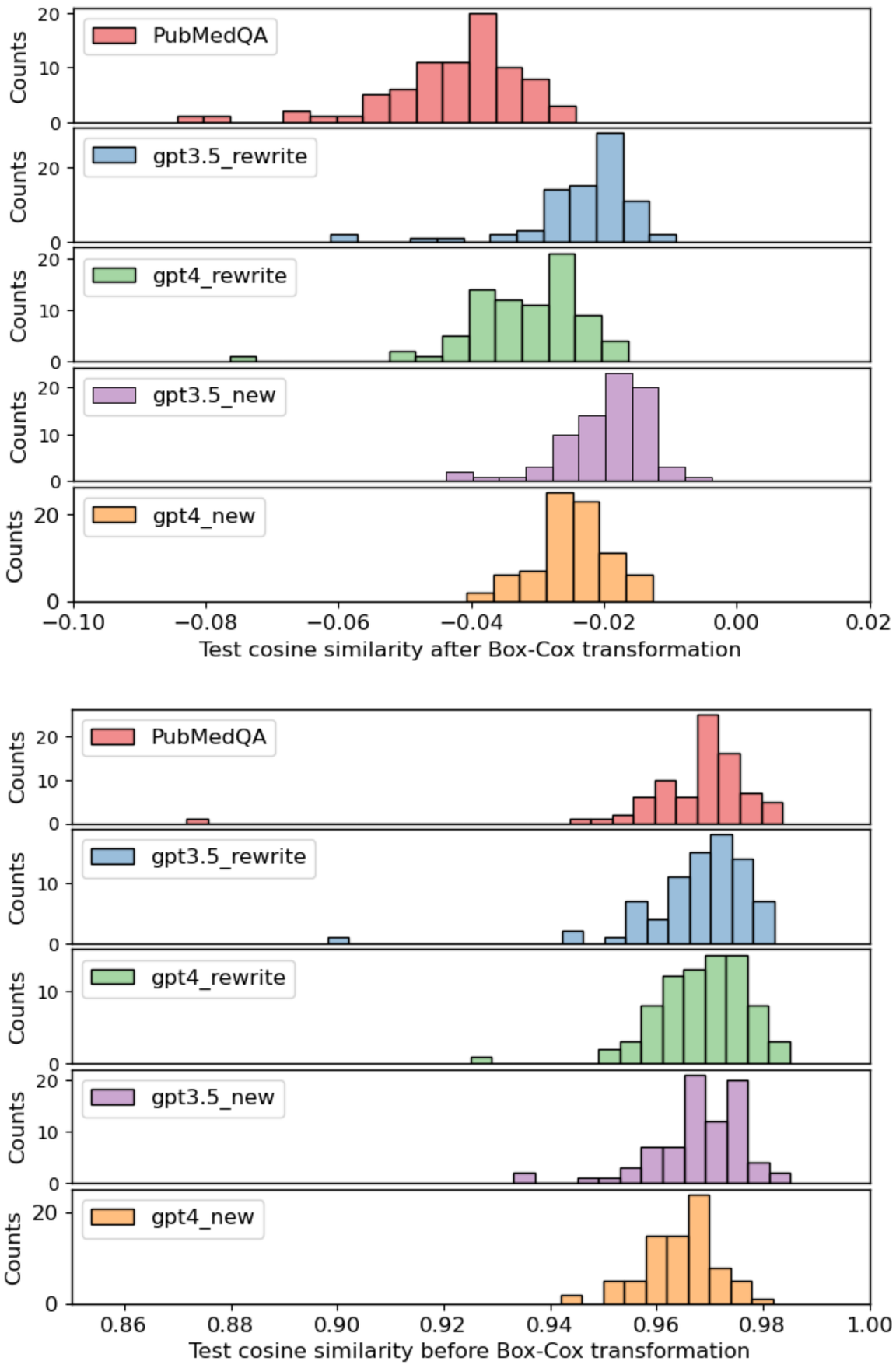
Algorithm 2 Determine classification threshold

procedure FINDTHRESHOLD($\mathcal{D}_{\text{sim}}$, labels)
  for λ in $[\lambda_1, \lambda_2, \ldots, \lambda_n]$ **do**
    $\tilde{\mathcal{D}_\lambda} \leftarrow \text{Box-Cox}(\mathcal{D}_{\text{sim}}, \lambda)$
    $\texttt{gm}_\lambda = \text{GaussianMixture}(\tilde{\mathcal{D}_\lambda}, \text{n\_class=2})$
    $\texttt{score}_\lambda = \text{AUROC}(\texttt{gm}_\lambda, \text{labels})$
  end for
  return the $\texttt{gm}_\lambda$ and corresponding λ that yield the maximum AUROC score
end procedure

3. Training Sample Analysis



4. Distribution Analysis



5. References

1. Zellers, Rowan, et al. "Defending against neural fake news." Advances in neural information processing systems 32 (2019).
2. Gehrmann, Sebastian, Hendrik Strobelt, and Alexander M. Rush. "Gltr: Statistical detection and visualization of generated text." *arXiv preprint arXiv:1906.04043* (2019).
3. Ippolito, Daphne, et al. "Automatic detection of generated text is easiest when humans are fooled." *arXiv preprint arXiv:1911.00650* (2019).
4. Badaskar, Sameer, Sachin Agarwal, and Shilpa Arora. "Identifying real or fake articles: Towards better language modeling." *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*. 2008.
5. Guo, Zhen, et al. "Dr. LLaMA: Improving Small Language Models in Domain-Specific QA via Generative Data Augmentation." *arXiv preprint arXiv:2305.07804* (2023).
6. Kirchenbauer, John, et al. "On the Reliability of Watermarks for Large Language Models." *arXiv preprint arXiv:2306.04634* (2023).
7. Kirchenbauer, John, et al. "On the Reliability of Watermarks for Large Language Models." arXiv preprint arXiv:2306.04634 (2023).
8. Yang, Xianjun, et al. "DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text." arXiv preprint arXiv:2305.17359 (2023).