A. Complete Results for Top-p and Top-k Decoding

Tables 4 and 5 contain the complete results for XSum, SQuAD, and WritingPrompts for the five models considered in Table 1. On average, both top-p and top-k sampling seem to make the detection task easier. This result is perhaps intuitive, as both sampling methods strictly increase the average log likelihood of model generations under the model (as they truncate low-probability tokens, albeit with different heuristics). Therefore methods based on probability or rank of tokens should become more discriminative.

	XSum						**SQuAD**						**WritingPrompts**					
Method	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**
$\log p(x)$	0.93	0.93	0.94	0.91	0.87	0.92	0.96	0.94	0.91	0.87	0.79	0.89	0.99*	0.98*	0.98*	0.97*	0.97*	**0.98**
Rank	0.80	0.77	0.77	0.75	0.73	0.76	0.84	0.82	0.81	0.80	0.75	0.81	0.87	0.84	0.83	0.83	0.81	0.84
LogRank	0.95*	0.94*	0.96*	0.93*	0.89*	0.93*	0.98*	0.96*	0.94*	**0.90**	**0.83**	0.92*	0.99*	0.98*	0.98*	**0.98**	**0.98**	**0.98**
Entropy	0.55	0.46	0.53	0.54	0.58	0.53	0.53	0.50	0.55	0.56	0.57	0.54	0.32	0.37	0.28	0.32	0.32	0.32
DetectGPT	**0.99**	**0.98**	**1.00**	**0.98**	**0.97**	**0.98**	**0.99**	**0.98**	**0.98**	**0.90**	0.82*	**0.94**	**1.00**	**0.99**	**0.99**	0.97*	0.93	**0.98**
Diff	0.04	0.04	0.04	0.05	0.08	0.05	0.01	0.02	0.04	0.00	-0.01	0.02	0.01	0.01	0.01	-0.01	-0.05	0.00

Table 4. Nucleus (top-p) sampling evaluation with $p = 0.96$. AUROC for detecting samples from the given model on the given dataset for DetectGPT and four previously proposed criteria. Nucleus sampling generally makes detection easier for all methods, but DetectGPT still provides the highest average AUROC. For WritingPrompts, however, the LogRank baseline performs as well as DetectGPT.

	XSum						**SQuAD**						**WritingPrompts**					
Method	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**
$\log p(x)$	0.89	0.89	0.89	0.84	0.81	0.87	0.93	0.90	0.88	0.82	0.74	0.85	0.97	0.95	0.97	0.96	0.95*	0.96
Rank	0.79	0.77	0.77	0.75	0.73	0.76	0.84	0.82	0.80	0.80	0.75	0.80	0.87	0.84	0.83	0.82	0.81	0.83
LogRank	0.92*	0.91*	0.93*	0.89*	0.85*	0.90*	0.96*	0.94*	0.92*	0.87*	0.79*	0.90*	0.98*	0.97*	0.98*	**0.97**	**0.96**	**0.97**
Entropy	0.58	0.49	0.55	0.56	0.59	0.55	0.55	0.52	0.56	0.56	0.58	0.56	0.35	0.41	0.30	0.34	0.37	0.35
DetectGPT	**0.99**	**0.97**	**0.99**	**0.96**	**0.96**	**0.98**	**0.99**	**0.98**	**0.98**	**0.89**	**0.80**	**0.93**	**0.99**	**0.98**	**0.99**	**0.97**	0.93	**0.97**
Diff	0.07	0.06	0.06	0.07	0.11	0.08	0.03	0.04	0.06	0.02	0.01	0.03	0.01	0.01	0.01	0.00	-0.03	0.00

Table 5. Top-k sampling evaluation with $k = 40$. DetectGPT generally provides the most accurate performance (highest AUROC), although the gap is narrowed comparing to direct sampling, presumably because top-k generations are more generic.