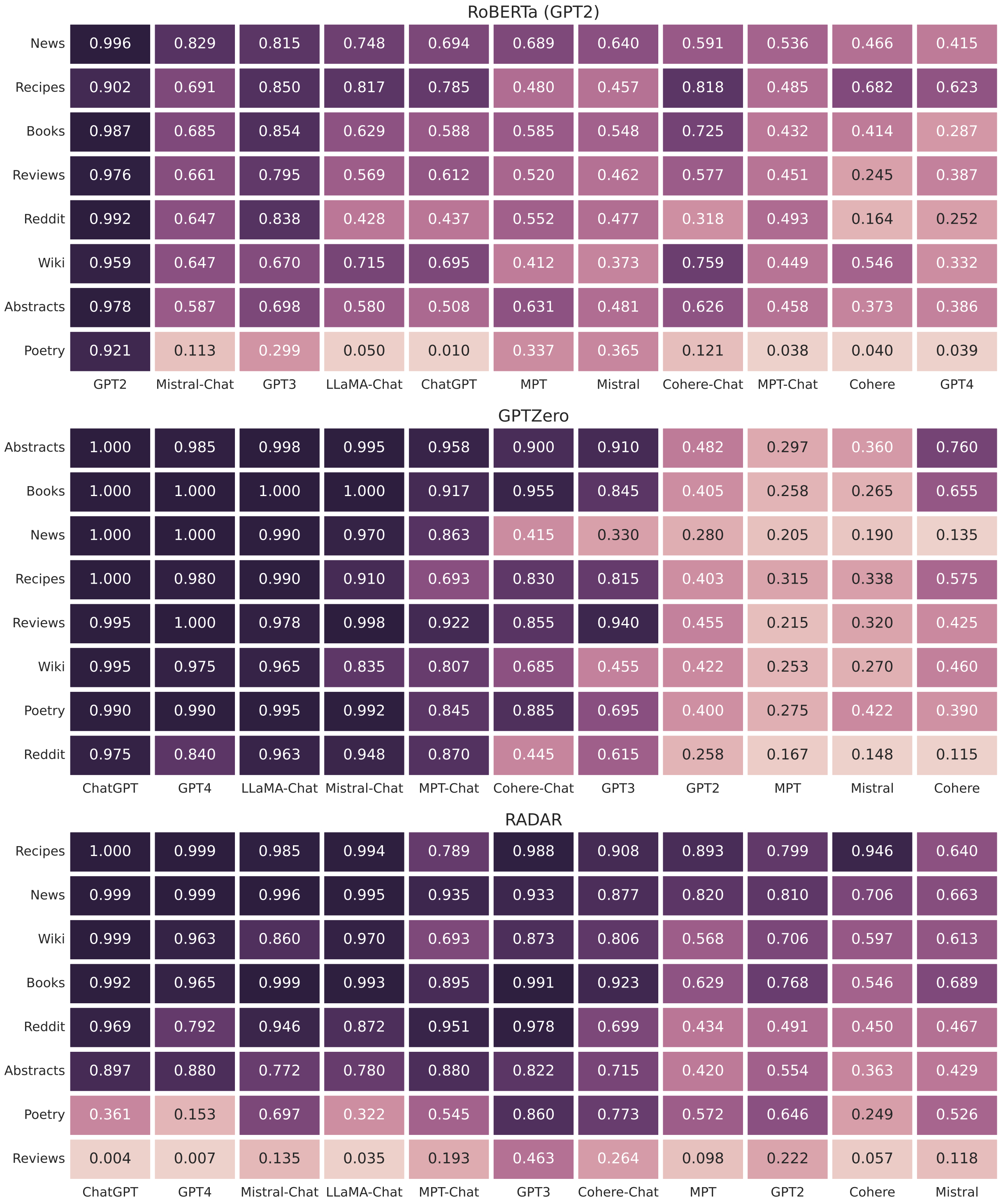

RoBERTa (GPT2)

	GPT2	Mistral-Chat	GPT3	LLaMA-Chat	ChatGPT	MPT	Mistral	Cohere-Chat	MPT-Chat	Cohere	GPT4
News	0.996	0.829	0.815	0.748	0.694	0.689	0.640	0.591	0.536	0.466	0.415
Recipes	0.902	0.691	0.850	0.817	0.785	0.480	0.457	0.818	0.485	0.682	0.623
Books	0.987	0.685	0.854	0.629	0.588	0.585	0.548	0.725	0.432	0.414	0.287
Reviews	0.976	0.661	0.795	0.569	0.612	0.520	0.462	0.577	0.451	0.245	0.387
Reddit	0.992	0.647	0.838	0.428	0.437	0.552	0.477	0.318	0.493	0.164	0.252
Wiki	0.959	0.647	0.670	0.715	0.695	0.412	0.373	0.759	0.449	0.546	0.332
Abstracts	0.978	0.587	0.698	0.580	0.508	0.631	0.481	0.626	0.458	0.373	0.386
Poetry	0.921	0.113	0.299	0.050	0.010	0.337	0.365	0.121	0.038	0.040	0.039

GPTZero

	ChatGPT	GPT4	LLaMA-Chat	Mistral-Chat	MPT-Chat	Cohere-Chat	GPT3	GPT2	MPT	Mistral	Cohere
Abstracts	1.000	0.985	0.998	0.995	0.958	0.900	0.910	0.482	0.297	0.360	0.760
Books	1.000	1.000	1.000	1.000	0.917	0.955	0.845	0.405	0.258	0.265	0.655
News	1.000	1.000	0.990	0.970	0.863	0.415	0.330	0.280	0.205	0.190	0.135
Recipes	1.000	0.980	0.990	0.910	0.693	0.830	0.815	0.403	0.315	0.338	0.575
Reviews	0.995	1.000	0.978	0.998	0.922	0.855	0.940	0.455	0.215	0.320	0.425
Wiki	0.995	0.975	0.965	0.835	0.807	0.685	0.455	0.422	0.253	0.270	0.460
Poetry	0.990	0.990	0.995	0.992	0.845	0.885	0.695	0.400	0.275	0.422	0.390
Reddit	0.975	0.840	0.963	0.948	0.870	0.445	0.615	0.258	0.167	0.148	0.115

RADAR

	ChatGPT	GPT4	Mistral-Chat	LLaMA-Chat	MPT-Chat	GPT3	Cohere-Chat	MPT	GPT2	Cohere	Mistral
Recipes	1.000	0.999	0.985	0.994	0.789	0.988	0.908	0.893	0.799	0.946	0.640
News	0.999	0.999	0.996	0.995	0.935	0.933	0.877	0.820	0.810	0.706	0.663
Wiki	0.999	0.963	0.860	0.970	0.693	0.873	0.806	0.568	0.706	0.597	0.613
Books	0.992	0.965	0.999	0.993	0.895	0.991	0.923	0.629	0.768	0.546	0.689
Reddit	0.969	0.792	0.946	0.872	0.951	0.978	0.699	0.434	0.491	0.450	0.467
Abstracts	0.897	0.880	0.772	0.780	0.880	0.822	0.715	0.420	0.554	0.363	0.429
Poetry	0.361	0.153	0.697	0.322	0.545	0.860	0.773	0.572	0.646	0.249	0.526
Reviews	0.004	0.007	0.135	0.035	0.193	0.463	0.264	0.098	0.222	0.057	0.118

Figure 12: Extended heatmap of RoBERTa GPT2, GPTZero, and RADAR's performance across all models and domains in the RAID dataset. We see that the trends noted in Figure 6 still hold.

	News	**Wiki**	**Reddit**	**Books**	**Abstracts**	**Reviews**	**Poetry**	**Recipes**
RoBERTa-B GPT2	0.032	0.379	0.477	0.586	0.055	0.539	0.998	0.916
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
RoBERTa-L GPT2	0.070	0.459	0.100	0.161	0.085	0.298	0.762	0.315
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.1%)	(5.0%)	(5.0%)
RoBERTa-B C-GPT	0.987	0.983	0.219	0.996	0.007	0.371	0.295	0.998
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
RADAR	0.022	0.061	0.695	0.174	0.31	0.997	0.457	0.016
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
GLTR	0.788	0.788	0.767	0.742	0.726	0.757	0.756	0.863
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
FastDetectGPT	0.920	0.870	0.870	0.930	0.860	0.900	0.940	0.880
	(4.8%)	(5.1%)	(5.1%)	(4.9%)	(4.6%)	(4.8%)	(5.9%)	(5.5%)
LLMDet	1.000	1.000	1.000	1.000	0.999	1.000	1.000	0.998
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.1%)	(5.0%)	(5.0%)
Binoculars	0.077	0.093	0.099	0.085	0.092	0.097	0.084	0.094
	(4.9%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(4.9%)	(5.0%)	(5.0%)
GPTZero	0.047	0.032	0.057	0.125	0.125	0.070	0.031	0.035
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
Originality	0.375	0.938	0.250	0.312	0.257	0.461	0.047	0.750
	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)	(5.0%)
Winston	0.001	0.970	0.062	0.998	0.000	0.062	0.875	0.996
	(4.0%)	(5.0%)	(5.0%)	(5.0%)	(6.0%)	(5.0%)	(5.0%)	(5.0%)
ZeroGPT(*)	1.000	1.000	0.375	0.250	0.500	1.000	0.125	1.000
	(29.0%)	(48.0%)	(1.0%)	(9.0%)	(4.0%)	(5.0%)	(5.0%)	(52.0%)

Table 13: Thresholds found by our search and the exact False Positive Rates on our dataset. We see that ZeroGPT is incapable of achieving the target FPR of 5% in many domains.

	News	**Wiki**	**Reddit**	**Books**	**Abstracts**	**Reviews**	**Poetry**	**Recipes**
RoBERTa-B GPT2	74.3	64.7	56.3	67.9	56.3	69.1	23.9	64.6
RoBERTa-L GPT2	69.8	59.5	54.1	62.4	58.9	58.2	24.4	67.2
RoBERTa-B CGPT	45.1	48.7	44.6	53.5	72.3	65.3	32.0	5.9
RADAR	88.0	76.8	71.8	84.5	66.7	14.1	53.0	**88.5**
GLTR	66.9	64.3	65.7	74.0	62.0	67.3	34.8	67.2
FastDetectGPT	74.2	77.3	70.9	76.2	76.4	77.5	63.4	74.6
LLMDet	39.8	32.6	39.6	37.1	18.0	33.1	30.7	48.1
Binoculars	80.7	76.7	79.4	83.7	79.1	80.1	**81.0**	76.6
GPTZero	58.1	62.8	57.0	71.4	74.9	70.5	69.5	67.6
Originality	**88.4**	**83.2**	**85.0**	**90.4**	87.7	**87.3**	75.1	82.8
Winston	72.4	54.9	68.9	70.7	**94.7**	72.9	64.3	68.9
ZeroGPT(*)	72.2	70.6	65.1	73.3	60.3	68.6	50.0	63.7

Table 14: Accuracy Score at FPR=5% for detectors across different domains. We see that metric-based methods perform surprisingly well across domains and that detectors can perform surprisingly poorly on unseen domains.

	GPT2	GPT3	ChatGPT	GPT4	Cohere		Mistral		MPT		Llama	Total
Chat? (Y/N)	✗	✗	✓	✓	✗	✓	✗	✓	✗	✓	✓	-
R-B GPT2	84.0	74.7	65.4	42.4	42.9	61.1	45.7	65.9	49.2	45.9	68.7	59.1
R-L GPT2	96.3	72.4	53.7	33.8	37.3	56.6	47.6	60.5	52.6	41.6	56.7	56.7
R-B CGPT	36.7	54.2	66.1	30.0	31.3	49.6	17.7	69.9	17.8	53.3	70.1	44.8
RADAR	64.7	88.6	82.1	76.0	51.4	77.2	54.1	83.6	58.0	76.5	78.5	70.9
GLTR	66.6	85.1	81.4	53.7	54.2	67.4	49.1	75.4	35.4	52.5	81.6	62.6
F-DetectGPT	72.1	95.4	96.1	73.9	84.7	85.1	58.2	81.3	45.3	57.1	94.0	73.6
LLMDet	48.2	40.2	18.9	27.0	32.6	35.6	31.5	35.7	28.2	21.4	55.1	35.0
Binoculars	68.9	**99.2**	**99.6**	91.9	**94.8**	**95.4**	62.3	91.7	45.2	70.8	97.6	79.6
GPTZero	38.8	70.1	99.4	97.1	43.9	74.6	28.9	**95.6**	24.8	85.9	**98.5**	66.5
Originality	**99.1**	98.2	98.2	89.9	78.9	90.6	**71.0**	95.5	**58.1**	76.2	94.7	**85.0**
Winston	47.6	77.8	**99.6**	**98.8**	63.6	86.2	46.1	94.9	24.8	**79.2**	97.5	71.0
ZeroGPT(*)	42.4	90.2	93.2	67.1	65.9	76.6	49.3	81.4	27.3	66.0	93.7	65.5

Table 15: Accuracy at FPR=5% for detectors on non-adversarial outputs of different models. We see that base models are more difficult to detect than their chat fine-tuned counterparts and that metric-based methods show impressive cross-model generalization. Asterisks (*) indicate that the detector was unable to achieve the target FPR.

	None	AS	AD	HG	IP	NS	PP	MS	SYN	ULS	WSA	ZWS
RoB-B GPT2	59.1	55.6	37.1	7.6	56.9	55.9	68.9	43.8	71.5	18.8	45.2	99.9
RoB-L GPT2	56.7	52.4	33.2	21.3	55.1	51.7	72.9	39.5	79.4	19.3	40.1	99.9
RoB-B CGPT	44.8	43.3	38.0	0.0	5.2	44.3	49.2	42.1	39.6	31.7	0.1	0.0
RADAR	70.9	70.8	67.9	59.3	73.7	71.0	67.3	69.5	67.5	70.4	66.1	82.2
GLTR	62.6	61.2	52.1	24.3	61.4	59.9	47.2	59.8	31.2	48.1	45.8	97.2
F-DGPT	73.6	71.6	64.7	51.4	72.0	68.2	71.8	70.7	34.0	60.4	64.4	98.9
LLMDet	35.0	33.9	27.4	40.6	27.2	33.8	28.5	32.7	27.3	23.4	4.4	27.1
Binoculars	79.6	78.2	74.3	37.7	71.7	77.1	80.3	78.0	43.5	73.8	70.1	99.1
GPTZero	66.5	64.9	61.0	66.2	66.2	65.8	64.0	65.1	61.0	56.5	66.2	66.2
ZeroGPT	65.5	65.4	59.7	82.4	64.9	64.7	46.7	64.7	18.8	54.5	64.2	48.0
Originality	85.0	83.6	71.4	9.3	85.1	86.0	96.7	78.6	96.5	75.8	84.9	4.9
Winston	71.0	68.9	66.9	26.3	69.8	69.0	52.6	67.5	63.6	56.8	46.8	25.0

Table 16: Accuracy Score at FPR=5% for all detectors across different adversarial attacks. Abbrevations are: AS: Alternative Spelling, AD: Article Deletion, HG: Homoglyph, IP: Insert Paragraphs, NS: Number Swap, PP: Paraphrase, MS: Misspelling, SYN: Synonym Swap, ULS: Upper Lower Swap, WSA: Whitespace Addition, ZWS: Zero-Width Space Addition