

NÁSTROJE PRO ROZPOZNÁVÁNÍ GENEROVANÉHO TEXTU

RICHARD MUŽÍK

RICHARD@IMUZYK.CZ

NOVEMBER 24, 2025

■ Neural

- ▶ RoBERTa-Base (GPT2), RoBERTa-Large (GPT2), RoBERTa-Base (ChatGPT), **RADAR**

■ Metric-based

- ▶ GLTR, **Binoculars**, Fast DetectGPT, LLMdet

■ Commercial

- ▶ **GPTZero**, Originality, Winston, ZeroGPT

NEURAL - PRINCIP (RADAR)

- Většinou fine tuning předtrénovaných modelů (RoBERTa)
- Robust AI-Text Detection via Adversarial Learning

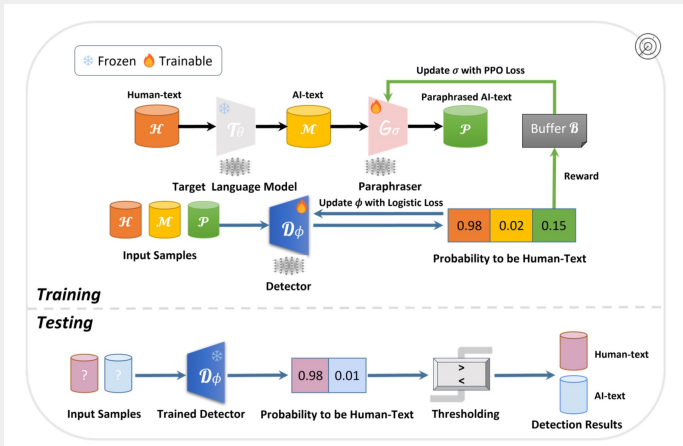


Figure: <https://radar.vizhub.ai/>

METRIC BASED - PRINCIP (BINOCULARS)

- Použití nějaké metriky založené (většinou) na výstupních pravděpodobnostech existujícího generativního modelu
- Binoculars - "prohlížení vstupu skrz „čočky“ dvou různých modelů"
 - ▶ Použití (poměru) perplexity a cross-perplexity
 - Perplexita - jak dobře pravděpodobnostní model předpovídá vzorek (porovnání modelů)
 - ▶ Udává jak překvapivá je predikce dalšího tokenu jedním modelem pro druhý

COMMERCIAL - PRINCIP (GPTZERO)

■ Nevíme jak funguje pod kapotou
I když se dá celkem hádat... :D

GPTZero			
0.405	1.000	1.000	0.265
0.280	1.000	1.000	0.190
0.258	0.975	0.840	0.148
0.455	0.995	1.000	0.320
0.422	0.995	0.975	0.270
GPT2	ChatGPT	GPT4	Mistral

Figure: <https://aclanthology.org/2024.acl-long.674.pdf>

BENCHMARK

	Open-Source								Closed-Source			
	Chat Models (llama-c, mistral-c, mpt-c)				Non-Chat Models (mistral, mpt, gpt2)				Chat Models (c-gpt, gpt4, cohere)		Non-Chat Models (cohere, gpt3)	
Dec. Strategy	greedy		sampling		greedy		sampling		greedy	sampling	greedy	sampling
Rep. Penalty?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗
R-B GPT2	84.1	52.3	77.9	26.2	98.6	44.1	60.5	35.4	70.9	41.7	65.1	52.5
R-L GPT2	79.7	41.1	71.4	19.5	98.5	43.0	67.2	53.4	61.4	34.7	61.1	48.6
R-B CGPT	80.2	63.3	75.0	39.3	53.3	26.4	14.9	1.7	59.1	38.1	46.5	39.0
RADAR	88.8	77.4	85.6	66.4	91.8	63.8	48.3	31.8	81.6	75.3	72.2	67.7
GLTR	89.8	67.5	83.9	38.3	99.6	56.9	44.5	0.5	80.7	54.3	75.6	63.7
F-DetectGPT	98.6	74.5	96.2	40.5	97.8	56.1	79.7	0.6	96.0	74.1	93.8	86.3
LLMDet	55.5	30.2	47.5	16.5	74.8	27.0	38.4	3.7	35.8	18.5	40.0	32.9
Binoculars	99.9	86.6	99.7	60.6	99.9	62.3	72.4	0.6	99.2	92.1	99.0	95.0
GPTZero	98.8	93.7	98.4	82.5	74.7	34.6	9.4	4.8	92.3	88.5	60.6	53.4
Originality	98.6	86.3	97.7	72.5	99.9	64.1	89.0	51.2	96.8	89.0	91.7	85.4
Winston	97.2	90.1	96.6	78.3	68.2	49.0	29.5	11.3	96.1	93.7	73.2	68.1
ZeroGPT(*)	95.4	80.7	90.5	54.9	85.1	57.2	16.0	0.3	92.1	65.8	83.4	72.7

Figure: <https://aclanthology.org/2024.acl-long.674.pdf>

BENCHMARK

	None	Paraphrase	Synonym	Misspelling	Homoglyph	Whitespace	Delete Articles
R-L GPT2	56.7	72.9 (+16.2)	79.4 (+22.7)	39.5 (-17.2)	21.3 (-35.4)	40.1 (-16.6)	33.2 (-23.5)
RADAR	70.9	67.3 (-3.6)	67.5 (-3.4)	69.5 (-1.4)	59.3 (-11.6)	66.1 (-4.8)	67.9 (-3.0)
GLTR	62.6	47.2 (-15.4)	31.2 (-31.4)	59.8 (-2.8)	24.3 (-38.3)	45.8 (-16.8)	52.1 (-10.5)
Binoculars	79.6	80.3 (+0.7)	43.5 (-36.1)	78.0 (-1.6)	37.7 (-41.9)	70.1 (-9.5)	74.3 (-5.3)
GPTZero	66.5	64.0 (-2.5)	61.0 (-5.5)	65.1 (-1.4)	66.2 (-0.3)	66.2 (-0.3)	61.0 (-5.5)
Originality	85.0	96.7 (+11.7)	96.5 (+11.5)	78.6 (-6.4)	9.3 (-75.7)	84.9 (-0.1)	71.4 (-13.6)

Figure: <https://aclanthology.org/2024.acl-long.674.pdf>



<https://aclanthology.org/2024.acl-long.674.pdf>

RAID BENCHMARK LEADERBOARD



<https://raid-bench.xyz/leaderboard>



<https://arxiv.org/pdf/2401.12070>



<https://radar.vizhub.ai/>



[https://en.wikipedia.org/wiki/Wikipedia:
Signs_of_AI_writing](https://en.wikipedia.org/wiki/Wikipedia:Signs_of_AI_writing)



<https://gptzero.me/>